AN INTRODUCTION TO FUNCTIONS OF SEVERAL REAL VARIABLES

Robert C. Gunning

Introduction

I have taught the honors version of calculus of several variables at Princeton University from time to time over a number of years. In the past I used Michael Spivak's path breaking and highly original *Calculus on Manifolds* as the course textbook, supplemented by one or another of the available excellent textbooks covering the standard approaches to calculus in several variables and including a wide range of examples and problems. As usual though, for anyone who has taught a course from the same textbook more than a couple of times, there are some parts of Spivak's treatment that I would like to expand, and some parts that I would like to contract or handle differently. During the 2007-2008 academic year Lillian Pierce and I cotaught the course, while she was a graduate student here, and reorganized the presentation of the material and the problem assignments; her suggestions were substantial and she played a significant part in working through the revisions that year. The goal was to continue the rather abstract treatment of the underlying mathematical topics but to supplement it by more emphasis on using the theoretical material to solve problems and to see some applications. The problems were divided into two categories: a first group of problems testing a basic understanding of the essential topics discussed and their applications, problems that almost all serious students should be able to solve without too much difficulty; and a second group of problems covering more theoretical aspects and tougher calculations, challenging the students but still not particularly difficult. The temptation to include a third category of optional very challenging problems, to introduce interested students to a wider range of other theoretical and practical aspects, was frustrated by a resounding lack of interest on the part of the students. A team of Princeon undergraduate mathematics majors consisting of Robert Haraway, Adam Hesterberg, Jay Holt and Alex Schiller took careful notes of the lectures that both Lillian and I gave. In the subsequent years I have used these notes when teaching the course again, modified each year with revisions and additions using the very helpful corrections and suggestions from the students taking the course and the graduate assistants coteaching and grading the course. The resulting notes still follow to an extent the pattern and outlook pioneered in Michael Spivak's book. The principal differences are more emphasis on differentiation and the inverse mapping and related theorems; a more extensive treatment of orientation, which can be a rather confusing concept; an introduction of differential forms rather more simply and primitively, rather than as duals to differentiations, which seems sufficient when the interest is really on their role in Euclidean spaces; and more emphasis on the classical interpretations and analytic aspects of differential forms. The course based on these notes covers a good deal of material, so it meets four hours each week, one of which is devoted principally to examples and illustrative calculations, while the discussion of the problems is left to office hours outside the regular class schedule. It has been possible at least to discuss all the topics covered in general term, focusing on the most difficult proofs and leaving it to the students to read the details of some of the proofs in the notes.

I would like to express here my sincere thanks to the students who compiled the lecture notes that were the basis for these notes, Robert Haraway, Adam Hesterberg, Jay Holt and Alex Schiller; to Lillian Pierce for her suggestions and great help in reorganizing the course; to Tadahiro Oh for a very thorough review of the notes and a great many very valuable suggestions; to the graduate students who have cotaught the course and done the major share of grading the assignments; and to the students who have taken the course since its reorganization for a number of corrections and suggestions for greater clarity in the notes. The remaining errors and confusion are my own responsibility though.

Robert C. Gunning
Fine Hall, Princeton University, September 2012

Contents

Chapter 1

Background

1.1 Norms

Some fairly standard set-theoretic notation and terminology will be used consistently. In particular, $a \in A$ indicates that a is an element of, or a point in, the set A, while $A \subset B$ indicates that A is a subset of B, possibly equal to B. The union $A \cup B$ of sets A and B consists of all elements that belong to either A or B or both A and B, while the intersection $A \cap B$ consists of all elements that belong to both A and B. The complement $A \sim B$ consists of all elements of A that do not belong to B, whether or not B is a subset of A. If the set A is understood from context it may be omitted; so $\sim B$ consists of all elements in a set A that are not in B, where the set A is understood. A mapping $f : A \longrightarrow B$ associates to each element $a \in A$ an element $f(a) \in B$, the image of the point a. This mapping f is injective if distinct elements of A have distinct images in B, is surjective if every element of B is the image of some element of A, and is bijective if it is both injective and surjective; thus f is bijective if and only if is a one-to-one mapping between the sets A and B, and consequently has an inverse mapping $f^{-1} : B \longrightarrow A$ that is also bijective.

The n-dimensional real vector space will be denoted by $\mathbb{R}^n$, following the Bourbaki convention. A vector $\mathbf{x} \in \mathbb{R}^n$ always will be viewed as a column vector of length n, so as given by

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

where $x_i \in \mathbb{R}$. For notational convenience a vector $\mathbf{x} \in \mathbb{R}^n$ sometimes will be indicated by listing its coordinates in the form $\mathbf{x} = \{x_j\} = \{x_1, \ldots, x_n\}$, although it still will be viewed as a column vector; and the image of a mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ often will be denoted by $f(x_1, \ldots, x_n)$, viewing the mapping f as a function of the individual coordinates of the column vector $\mathbf{x}$. The origin in $\mathbb{R}^n$ is the vector $\mathbf{0} = \{0\} = \{0, 0, \ldots, 0\}$.

Vectors are added by adding their coordinates and multiplied by a real number a by multiplying their coordinates by a, so that

$$\mathbf{x}+\mathbf{y} = \begin{pmatrix} x_1+y_1 \\ \vdots \\ x_n+y_n \end{pmatrix} \quad \text{and} \quad a\,\mathbf{x} = \begin{pmatrix} a\,x_1 \\ \vdots \\ a\,x_n \end{pmatrix} \quad \text{for any} \quad a \in \mathbb{R}.$$

Linear mappings between vector spaces are described by matrix multiplication; for example, a 2×3 matrix $A=\{a_{ij}\}$ describes the mapping $A:\mathbb{R}^3 \longrightarrow \mathbb{R}^2$ that takes a vector $\mathbf{x}=\{x_j\}\in\mathbb{R}^3$ to the vector

$$A\mathbf{x} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} a_{11}x_1+a_{12}x_2+a_{13}x_3 \\ a_{21}x_1+a_{22}x_2+a_{23}x_3 \end{pmatrix} \in \mathbb{R}^2.$$

There are a various norms measuring the sizes or lengths of vectors in $\mathbb{R}^n$, only two of which will be considered here:

- the **Euclidean norm** or **Cartesian norm** of a vector $\mathbf{x}=\{x_j\}$ is defined by
 $\|\mathbf{x}\|_2 = \sqrt{\sum_{j=1}^n x_j^2} = \sqrt{x_1^2+\cdots+x_n^2}$ (with the non-negative square root);

- the **supremum norm** or **sup norm** of a vector $\mathbf{x}=\{x_j\}$ is defined by $\|\mathbf{x}\|_\infty = \max_{1\le j\le n}|x_j|$.

In general, a **norm** on the vector space $\mathbb{R}^n$ is a mapping $\mathbb{R}^n \longrightarrow \mathbb{R}$ that associates to any $\mathbf{x}\in\mathbb{R}^n$ a real number $\|\mathbf{x}\|\in\mathbb{R}$ and that has the following properties:

1. *positivity* : $\|\mathbf{x}\|\ge 0$ and $\|\mathbf{x}\|=0$ if and only if $\mathbf{x}=0$;
2. *homogeneity*: $\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\|$ for any $c\in\mathbb{R}$;
3. *the triangle inequality*: $\|\mathbf{x}+\mathbf{y}\|\le\|\mathbf{x}\|+\|\mathbf{y}\|$.

That the supremum norm satisfies these three properties is obvious, except perhaps for the triangle inequality; to verify that, if $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ then since $|x_j|\le\|\mathbf{x}\|_\infty$ and $|y_j|\le\|\mathbf{y}\|_\infty$ for $1\le j\le n$ it follows that $|x_j+y_j|\le|x_j|+|y_j|\le\|\mathbf{x}\|_\infty+\|\mathbf{y}\|_\infty$ for $1\le j\le n$ and consequently that $\|\mathbf{x}+\mathbf{y}\|=\max_{1\le j\le n}|x_j+y_j|\le\|\mathbf{x}\|_\infty+\|\mathbf{y}\|_\infty$. That the Euclidean norm satisfies these three properties also is obvious, again except for the triangle inequality; but that is somewhat harder to prove in this case than in the case of the supremum norm. It is convenient first to demonstrate an inequality involving the **inner product** of two vectors, which is defined by

dot (1.1) $$(\mathbf{x},\mathbf{y}) = \sum_{j=1}^n x_j y_j \quad \text{for vectors } \mathbf{x}=\{x_j\} \text{ and } \mathbf{y}=\{y_j\}\in\mathbb{R}^n;$$

often this is written $(\mathbf{x},\mathbf{y}) = \mathbf{x}\cdot\mathbf{y}$ and called the **dot product** of the two vectors $\mathbf{x}$ and $\mathbf{y}$. It is clear from its definition that the inner or dot product has the following properties:

1. *linearity*: $(c_1\mathbf{x}_1 + c_2\mathbf{x}_2, \mathbf{y}) = c_1(\mathbf{x}_1,\mathbf{y}) + c_2(\mathbf{x}_2,\mathbf{y})$ for any $c \in \mathbb{R}$;
2. *symmetry*: $(\mathbf{x},\mathbf{y}) = (\mathbf{y},\mathbf{x})$;
3. *positivity*: $(\mathbf{x},\mathbf{x}) \geq 0$ and $(\mathbf{x},\mathbf{x}) = 0$ if and only if $\mathbf{x} = 0$.

It is apparent from symmetry that the inner product $(\mathbf{x},\mathbf{y})$ is also a linear function of the vector $\mathbf{y}$. The Euclidean norm can be defined in terms of the inner product by

norm1
$$\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x},\mathbf{x})} \quad \text{(with the non-negative square root)}, \tag{1.2}$$

as is also quite clear from the definition. Conversely the inner product can be defined in terms of the Euclidean norm by

norm2
$$(\mathbf{x},\mathbf{y}) = \tfrac{1}{4}\|\mathbf{x}+\mathbf{y}\|_2^2 - \tfrac{1}{4}\|\mathbf{x}-\mathbf{y}\|_2^2 \tag{1.3}$$

since

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2 &= (\mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}) - (\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}) \\ &= \Big((\mathbf{x},\mathbf{x}) + 2(\mathbf{x},\mathbf{y}) + (\mathbf{y},\mathbf{y})\Big) - \Big((\mathbf{x},\mathbf{x}) - 2(\mathbf{x},\mathbf{y}) + (\mathbf{y},\mathbf{y})\Big) \\ &= 4(\mathbf{x},\mathbf{y});\end{aligned}$$

equation (norm2 1.3) is called the **polarization identity**.

norm **Theorem 1.1** *For any* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$
(i) $|(\mathbf{x},\mathbf{y})| \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2$, *and this is an equality if and only if the two vectors are linearly dependent;*
(ii) $\|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$, *and this is an equality if and only if one of the two vectors is a non-negative multiple of the other.*

Proof: If $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ where $\mathbf{y} \neq \mathbf{0}$ and if $t \in \mathbb{R}$, introduce the continuous function $f(t)$ of the variable t defined by

aux1
$$\begin{aligned} f(t) = \|\mathbf{x}+t\mathbf{y}\|_2^2 &= \sum_{j=1}^n (x_j + ty_j)^2 \\ &= \sum_{j=1}^n x_j^2 + 2t\sum_{j=1}^n x_jy_j + t^2\sum_{j=1}^n y_j^2 \\ &= \|\mathbf{x}\|_2^2 + 2t(\mathbf{x},\mathbf{y}) + t^2\|\mathbf{y}\|_2^2. \end{aligned} \tag{1.4}$$

The function $f(t)$ becomes large for $|t|$ large, since $\|y\|_2 > 0$ by assumption, so it must take a minimum value at some point. Since

$$f'(t) = 2(\mathbf{x},\mathbf{y}) + 2t\|\mathbf{y}\|_2^2$$

there is actually just a single point at which $f'(t) = 0$, the point $t_0 = -\frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|_2^2}$, so this must be the point at which the function $f(t)$ takes its minimum value; and the minimum value is

aux1x
$$f(t_0) = \|\mathbf{x}\|_2^2 + 2\left(-\frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{y}\|_2^2}\right)(\mathbf{x},\mathbf{y}) + \frac{(\mathbf{x},\mathbf{y})^2}{\|\mathbf{y}\|_2^4}\|\mathbf{y}\|_2^2 = \frac{\|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2 - (\mathbf{x},\mathbf{y})^2}{\|\mathbf{y}\|_2^2}. \tag{1.5}$$

It is clear from (1.4) that $f(t) \ge 0$ at all points $t \in \mathbb{R}$, so in particular at the point t_0, and that yields the inequality (i). It is clear from (1.5) that this inequality is an equality if and only if $f(t_0) = 0$, hence if and only if $\mathbf{x} + t_0\mathbf{y} = 0$ since $f(t_0) = \|\mathbf{x} + t_0\mathbf{y}\|_2^2$; and since $\mathbf{y} \ne 0$ that is just the condition that the vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. Next from (1.4) with $t = 1$ and from the inequality (i) it follows that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_2^2 &= \|\mathbf{x}\|_2^2 + 2(\mathbf{x},\mathbf{y}) + \|\mathbf{y}\|_2^2\\ &\le \|\mathbf{x}\|_2^2 + 2\|\mathbf{x}\|_2\|\mathbf{y}\|_2 + \|\mathbf{y}\|_2^2 = (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2,\end{aligned}$$

which yields the inequality (ii). This inequality is an equality if and only if $(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\|_2\|\mathbf{y}\|_2$, or equivalently if and only if (i) is an equality and $(\mathbf{x},\mathbf{y}) \ge 0$; and since $\mathbf{y} \ne 0$ it follows from what has already been demonstrated that the inequality (i) is an equality if and only if $\mathbf{x} = c\,\mathbf{y}$ for some real number c, and then $(\mathbf{x},\mathbf{y}) = c\|\mathbf{y}\|_2^2 \ge 0$ if and only if $c \ge 0$. That suffices for the proof.

The very useful inequality (i) in the preceding theorem, called the **Cauchy-Schwarz inequality**, can be written

$$\left|\frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}\right| \le 1 \quad \text{if } \mathbf{x} \ne \mathbf{0}, \mathbf{y} \ne \mathbf{0};$$

as a consequence there is an angle θ, called the **angle** between the nonzero vectors $\mathbf{x}$ and $\mathbf{y}$, that is determined uniquely up to a multiple of 2π by

normq1
$$\cos\theta = \frac{(\mathbf{x},\mathbf{y})}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2}. \tag{1.6}$$

In particular if $\theta = 0$ or π so $\cos\theta = \pm 1$ then by Theorem 1.1 (i) the two vectors $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent; they are parallel and in the same direction if $\theta = 0$ so that $(\mathbf{x},\mathbf{y}) > 0$, or parallel and in the opposite direction if $\theta = \pi$ so that $(\mathbf{x},\mathbf{y}) < 0$. The two vectors are **orthogonal** or perpendicular to one another when the angle is either $\pi/2$ or $3\pi/2$ so that $(\mathbf{x},\mathbf{y}) = 0$. The geometrical interpretation of the norm $\|\mathbf{x}\|_2$ in terms of the inner product makes that norm particularly useful in many applications.

In general, two norms $\|\mathbf{x}\|_a$ and $\|\mathbf{x}\|_b$ on a vector space $\mathbb{R}^n$ are **equivalent** if there are nonzero constants c_a and c_b such that $\|\mathbf{x}\|_a \le c_a\,\|\mathbf{x}\|_b$ and $\|\mathbf{x}\|_b \le c_b\,\|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$. The Euclidean and supremum norms are equivalent, since they are related by the very useful inequalities

normcmp
$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty \quad \text{for any} \quad \mathbf{x} \in \mathbb{R}^n. \tag{1.7}$$

To verify these inequalities, if $\mathbf{x} = \{x_j\} \in \mathbb{R}^n$ then $|x_j| \le \|\mathbf{x}\|_\infty$ for $1 \le j \le n$ so

$$\|\mathbf{x}\|_2^2 = \sum_{j=1}^{n} x_j^2 \le \sum_{j=1}^{n} \|\mathbf{x}\|_\infty^2 = n\|\mathbf{x}\|_\infty^2 \quad \text{hence } \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty;$$

on the other hand

$$\|\mathbf{x}\|_2^2 = \sum_{j=1}^{n} x_j^2 \ge x_j^2 \quad \text{so } \|\mathbf{x}\|_2 \ge |x_j| \text{ for } 1 \le j \le n \text{ hence } \|\mathbf{x}\|_2 \ge \|\mathbf{x}\|_\infty.$$

It follows from (1.7) [normcmp] that whenever $\|\mathbf{x}\|_2$ is small then $\|\mathbf{x}\|_\infty$ is also small, and conversely; so for many purposes the two norms are interchangeable. If it is not really necessary to specify which norm is meant the notation $\|\mathbf{x}\|$ will be used, meaning either $\|\mathbf{x}\|_2$ or $\|\mathbf{x}\|_\infty$. Of course some care must be taken; in particular $\|\mathbf{x}\|$ should have the same meaning in any single equation, or usually throughout any single proof, for quite obvious reasons. Any norm can be used to define a **distance** function $d(\mathbf{x}, \mathbf{y})$ by

normdistdf (1.8)
$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

As an immediate consequence of the properties of norms it follows that the distance function has the following properties:

1. *positivity*: $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$,

2. *symmetry*: $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

3. *triangle inequality*: $d(\mathbf{x}, \mathbf{y}) \le d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$,

so the notion of distance does have the expected properties. Equivalent norms determine distance functions that are equivalent in the corresponding sense, as is quite evident from the definitions.

For some purposes it is convenient to view an $m \times n$ matrix

$$A = \left\{ a_{ij} \,\middle|\, 1 \le i \le m, 1 \le j \le n \right\} \in \mathbb{R}^{m \times n}$$

as a vector in $\mathbb{R}^{mn}$, and to consider its norm as a vector; thus

natnorm1 (1.9)
$$\|A\|_\infty = \max_{\substack{1\le i\le m\\ 1\le j\le n}} \left|a_{ij}\right| \text{ and } \|A\|_2 = \sqrt{\sum_{\substack{1\le i\le m\\ 1\le j\le n}} a_{ij}^2} \text{ (the nonnegative square root).}$$

If $\mathbf{x} \in \mathbb{R}^n$ then $A\mathbf{x} \in \mathbb{R}^m$, and since

$$\left|\sum_{j=1}^{n} a_{ij}x_j\right| \le \sum_{j=1}^{n} |a_{ij}||x_j| \le \sum_{j=1}^{n} \|A\|_\infty \|\mathbf{x}\|_\infty = n\,\|A\|_\infty \|\mathbf{x}\|_\infty$$

it follows that

natnorm2
$$\|A\mathbf{x}\|_\infty = \max_{1\le i\le m}\left|\sum_{j=1}^{n} a_{ij}x_j\right| \le n\|A\|_\infty\|\mathbf{x}\|_\infty. \tag{1.10}$$

An **ϵ-neighborhood** $\mathcal{N}_\epsilon(\mathbf{a})$ of a point $\mathbf{a}\in\mathbb{R}^n$, an **open ϵ-neighborhood** to be more specific, is a subset of $\mathbb{R}^n$ defined by

normdfn
$$\mathcal{N}_\epsilon(\mathbf{a}) = \left\{\ \mathbf{x}\in\mathbb{R}^n\ \middle|\ \|\mathbf{x}-\mathbf{a}\| < \epsilon\ \right\}; \tag{1.11}$$

and a **closed ϵ-neighborhood** is defined correspondingly by

normdfn1
$$\overline{\mathcal{N}}_\epsilon(\mathbf{a}) = \left\{\ \mathbf{x}\in\mathbb{R}^n\ \middle|\ \|\mathbf{x}-\mathbf{a}\| \le \epsilon\ \right\}. \tag{1.12}$$

The **boundary** of an open or closed ϵ-neighborhood is the closed set

nsormdfbd
$$\partial\mathcal{N}_\epsilon(\mathbf{a}) = \partial\overline{\mathcal{N}}_\epsilon(\mathbf{a}) = \left\{\ \mathbf{x}\in\mathbb{R}^n\ \middle|\ \|\mathbf{x}-\mathbf{a}\| = \epsilon\ \right\}. \tag{1.13}$$

If it is necessary or convenient to be specific, an open ϵ neighborhood in the Euclidean norm is denoted by $\mathcal{N}_{\epsilon,2}(\mathbf{a})$ while an open ϵ-neighborhood in the supremum norm is denoted by $\mathcal{N}_{\epsilon,\infty}(\mathbf{a})$. It is clear from the inequality (normcmp 1.7) that

normcmpa
$$\mathcal{N}_{\epsilon,2}(\mathbf{a}) \subset \mathcal{N}_{\epsilon,\infty}(\mathbf{a}) \subset \mathcal{N}_{\epsilon\sqrt{n},2}(\mathbf{a}) \quad\text{in}\quad \mathbb{R}^n. \tag{1.14}$$

The geometrical shapes of these neighborhoods depend on the norms used in their definitions; for example in the plane neighborhoods have the shapes as shown in Figure normfig1 1.1.

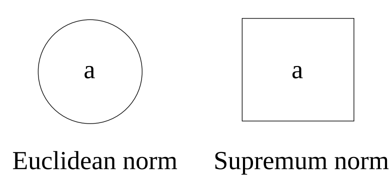

Figure 1.1: Neighborhoods $\mathcal{N}_\epsilon(\mathbf{a})\subset\mathbb{R}^2$. normfig1

Other useful subsets of $\mathbb{R}^n$ are **open cells**, subsets of the form

normclo
$$\Delta = \left\{\ \mathbf{x}=\{x_j\}\in\mathbb{R}^n\ \middle|\ a_j < x_j < b_j \ \text{ for }\ 1\le j\le n\ \right\} \tag{1.15}$$

for arbitrary real numbers $a_j < b_j$, and **closed cells**, subsets of the form

normclc
$$\overline{\Delta} = \left\{\ \mathbf{x}=\{x_j\}\in\mathbb{R}^n\ \middle|\ a_j \le x_j \le b_j \ \text{ for }\ 1\le j\le n\ \right\} \tag{1.16}$$

for arbitrary real numbers $a_j \le b_j$. Closed cells may have $a_j = b_j$ for some indices j, while of course open cells for which $a_j = b_j$ for any index j can only be the empty set. The **boundary** of an open or closed cell is the set

cellbd
$$\partial\Delta = \partial\overline{\Delta} = \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; \begin{array}{l} a_j \le x_j \le b_j \ \text{ for } 1 \le j \le n, \text{ and} \\ \text{at least one inequality is an equality} \end{array} \right\}. \tag{1.17}$$

A customary notation for cells in $\mathbb{R}^1$, which are also called **intervals**, is

int
$$(a,b) = \{\, x \in \mathbb{R}^1 \mid a < x < b \,\} \quad \text{and} \quad [a,b] = \{\, x \in \mathbb{R}^1 \mid a \le x \le b \,\}; \tag{1.18}$$

these can be mixed, as for instance $(a,b] = \{\, x \in \mathbb{R}^1 \mid a < x \le b \,\}$.

1.2 Topological Preliminaries

Some familiarity with the basic topological notions for subsets of the real line is assumed; but a brief review is included here, focusing on the extension of these notions to subsets of $\mathbb{R}^n$, and it may be a sufficient introduction even for those seeing these concepts for the first time. However if these notions are not entirely familiar it is advisable as an additional exercise to prove explicitly and in detail those statements that are labeled "clear" or "evident" in the discussion here.

The basic topological property of the real number system is its **completeness**, the property that any set S of real numbers that is bounded from above has a least upper bound or **supremum**, denoted by $\sup S$, or equivalently that any set S of real numbers that is bounded from below has a greatest lower bound or **infimum**, denoted by $\inf S$. An alternative statement of this property is that any Cauchy sequence of real numbers converges to a real limit, where a **Cauchy sequence** $\{a_i\}$ of real numbers consists of real numbers $a_i \in \mathbb{R}$ such that for any $\epsilon > 0$ there is an index N for which $|a_i - a_j| < \epsilon$ whenever $i, j > N$.

There are many topological notions and terms that are in common use. A subset $U \subset \mathbb{R}^n$ is said to be **open** if for each $\mathbf{a} \in U$ there is an $\epsilon > 0$ such that $\mathcal{N}_\epsilon(\mathbf{a}) \subset U$. It is clear from the inequalities (normcmp 1.7) that the condition that a set be open is independent of which of the two norms $\|\mathbf{x}\|_2$ or $\|\mathbf{x}\|_\infty$ is used to define the neighborhood $\mathcal{N}_\epsilon(\mathbf{a})$. Intuitively a set $U \subset \mathbb{R}^n$ is open if for any point $\mathbf{a} \in U$ all points that are near enough to $\mathbf{a}$ are also in the set U. For example an open ϵ-neighborhood $\mathcal{N}_\epsilon(\mathbf{a})$ of a point $\mathbf{a} \in \mathbb{R}^n$ and an open cell $\Delta \subset \mathbb{R}^n$ are open subsets. An open subset of $\mathbb{R}^n$ containing a point $\mathbf{a}$ is often called an **open neighborhood** of the point $\mathbf{a}$; an open ϵ-neighborhood $\mathcal{N}_\epsilon(\mathbf{a})$ is thus an open neighborhood of the point $\mathbf{a}$ in this sense as well. The collection $\mathcal{T}$ of open subsets of $\mathbb{R}^n$ has the following characteristic properties:

1. if $U_\alpha \in \mathcal{T}$ then $\bigcup_\alpha U_\alpha \in \mathcal{T}$;
2. if $U_i \in \mathcal{T}$ for $1 \le i \le N$ for some N then $\bigcap_{i=1}^N U_i \in \mathcal{T}$;
3. $\emptyset \in \mathcal{T}$ and $\mathbb{R}^n \in \mathcal{T}$.

These properties can be summarized in the statement that an arbitrary union of open sets is open, a finite intersection of open sets is open, and the empty set and the set $\mathbb{R}^n$ itself are open. That the open subsets of $\mathbb{R}^n$ satisfy these three conditions is quite clear. It is also clear that an infinite intersection of open sets is not necessarily open; for instance $\bigcap_{\nu=1}^{\infty}\mathcal{N}_{\frac{1}{\nu}}(a) = a$ and a single point is not open. A collection $\mathcal{T}$ of subsets of an arbitrary abstract set S having these three characteristic properties is said to be a **topology** on the set S, and the sets $U \in \mathcal{T}$ are said to be the **open sets** in this topology. For many purposes the relevant geometrical properties of a space can be described just in terms of the collection of open sets, the topology of that space, rather than in terms of a norm or distance function on the space. There are many topologies in addition to that defined by the Cartesian norm or the supremum norm or any norm that is equivalent to these norms, or indeed by any norm at all. Two particularly simple examples of topologies on a space that have occasional uses are the **discrete topology**, in which all subsets are open sets, and the **indiscrete topology**, in which the only open sets are the empty set and the entire set; the discrete topology is the largest possible topology, that with the greatest collection of open sets, while the indiscrete topology is the smallest possible topology, that with the least collection of open sets.

A point $\mathbf{a} \in \mathbb{R}^n$ is a **limit point** of a subset $E \subset \mathbb{R}^n$ if for any $\epsilon > 0$ the intersection $\mathcal{N}_\epsilon(\mathbf{a}) \cap E$ contains points of E other than $\mathbf{a}$; clearly that is equivalent to the condition that for all $\epsilon > 0$ the intersection $\mathcal{N}_\epsilon(\mathbf{a}) \cap E$ is an infinite set of points. The set of limit points of E is denoted by E' and is called the **derived set** of E. The set E is said to be **closed** if $E' \subset E$. For example a closed neighborhood $\overline{\mathcal{N}}_\epsilon(\mathbf{a})$ of a point $\mathbf{a} \in \mathbb{R}^n$ and a closed cell $\overline{\Delta}$ in $\mathbb{R}^n$ are closed subsets. The **closure** of a set E, denoted by $\overline{E}$, is defined to be the union $\overline{E} = E \cup E'$, and is readily seen to be a closed set; indeed it is easy to see that the closure of a set E is the smallest closed set containing E. A closed neighborhood $\overline{\mathcal{N}}_\epsilon(\mathbf{a})$ of a point $\mathbf{a} \in \mathbb{R}^n$ is the closure of the open neighborhood $\mathcal{N}_\epsilon(\mathbf{a})$ and a closed cell $\overline{\Delta}$ in $\mathbb{R}^n$ is the closure of the open cell Δ provided that $\Delta \neq \emptyset$. A basic result is that a subset $E \subset \mathbb{R}^n$ is closed if and only if its complement $F = \mathbb{R}^n \sim E$ is open. To verify that, if E is closed and $\mathbf{a} \notin E$ then $\mathbf{a} \notin E'$ so $\mathbf{a}$ is not a limit point of E hence there must be some $\epsilon > 0$ such that $\mathcal{N}_\epsilon(\mathbf{a}) \cap E = \emptyset$; consequently $\mathcal{N}_\epsilon(\mathbf{a}) \subset F$ so F is open. Conversely if E is open then no point of E can be a limit point of F, since for any point $\mathbf{a} \in E$ there is an $\epsilon > 0$ such that $\mathcal{N}_\epsilon(\mathbf{a}) \subset E$ hence that $\mathcal{N}_\epsilon(\mathbf{a}) \cap F = \emptyset$; consequently all the limit points of F are contained in F so F is closed. The complementarity between open and closed sets can be used to defined the **interior** of a subset $E \subset \mathbb{R}^n$, denoted by E^o, as the complement of the closure of the exterior of E, that is, $E^o = \mathbb{R}^n \sim \overline{(\mathbb{R}^n \sim E)}$; it is easy to see that the interior of E is the largest open set contained in E. It is also clear from the complementarity between open and closed sets that the collection $\mathcal{C}$ of closed subsets of $\mathbb{R}^n$ has the following characteristic properties:

1. if $U_\alpha \in \mathcal{C}$ then $\bigcap_\alpha \in \mathcal{C}$;
2. if $U_i \in \mathcal{C}$ for $1 \le i \le N$ for some N then $\bigcup_{i=1}^{N} \in \mathcal{C}$;

3. $\emptyset \in \mathcal{C}$ and $\mathbb{R}^n \in \mathcal{C}$.

These properties can be summarized in the statement that an arbitrary intersection of closed sets is closed, a finite union of closed sets is closed, and the empty set and the full set $\mathbb{R}^n$ are both closed; but an arbitrary union of closed sets is not necessarily closed, since any set can be written as a union of its points and each point is a closed set. A topology on a space can be defined equivalently by giving a collection of sets satisfying these three conditions and defining the open sets to be their complements, although it is customary to define topologies directly in terms of the open sets.

A subset $S \subset \mathbb{R}^n$ can be viewed as a set in itself, without regard to the ambient space $\mathbb{R}^n$ in which it is contained; in a sense that amounts to considering the set S itself as the entire universe, rather than viewing it as a subset contained in the universe $\mathbb{R}^n$. The restriction of a norm on $\mathbb{R}^n$ to the subset S can be used to describe a distance function on the set S itself, and open and closed subsets of S then can be defined in terms of this distance when an ϵ neighborhood of a point a in the set S is taken to be the intersection $\mathcal{N}_\epsilon(a) \cap S$. Thus a subset $E \subset S$ is said to be **relatively open** if for any point $a \in E$ there is an $\epsilon > 0$ such that $\mathcal{N}_\epsilon(a) \cap S \subset E$, where $\mathcal{N}_\epsilon(a) \cap S$ consists of those points of S that are distance at most ϵ from the point a. A point $a \in S$ is a **relative limit point** of E in S if for any $\epsilon > 0$ the intersection $\mathcal{N}_\epsilon(a) \cap S$ contains a point of E other than a, or equivalently, if for any $\epsilon > 0$ the intersection $\mathcal{N}_\epsilon(a) \cap S$ is infinite; and that is just the condition that $a \in E' \cap S$ where E' is the derived set of E as a subset of $\mathbb{R}^n$, so a relative limit point of E is a limit point of E in the ordinary sense but one that is also contained in S. A subset $E \subset S$ is said to be **relatively closed** if it contains all its relative limits points. Clearly a relatively open subset of S need not be an open subset of E; if S is a coordinate axis in the plane $\mathbb{R}^2$ an open interval of S is relatively open in S but is not open in $\mathbb{R}^2$. Equally clearly a relatively closed subset of S need not be a closed subset of $\mathbb{R}^n$; for $S \subset \mathbb{R}^n$ may not be closed in $\mathbb{R}^n$ it is still relatively closed in itself. However a subset $E \subset S$ is relatively open if and only if it is the intersection $E = U \cap S$ of S with an open subset $U \subset \mathbb{R}^n$, and is relatively closed if and only if it is the intersection $E = C \cap S$ of S with a closed subset $C \subset \mathbb{R}^n$. Indeed it is clear that if $E = U \cap S$ where $U \subset \mathbb{R}^n$ is open then E is relatively open; and if $E \subset S$ is relatively open then by definition for each point $a \in E$ there is an $\epsilon_a > 0$ such that $\mathcal{N}_\epsilon(a) \cap S \subset E$, and then $U = \bigcup_a \mathcal{N}_\epsilon(a)$ is an open subset of $\mathbb{R}^n$ such that $E = U \cap S$. Similarly if $E = C \cap S$ for some closed subset $C \subset \mathbb{R}^n$ then $E' \subset C' \subset C$ so $E' \cap S \subset C \cap S = E$, and if $E \subset S$ is a relatively closed subset of S then then $C = E \cup E'$ is a closed subset of $\mathbb{R}^n$ and $C \cap S = (E \cap S) \cup (E' \cap S) \subset E$. The relatively open subsets of a set S clearly satisfy the conditions that they form a topology on S, called the **relative topology** on S. This can be a somewhat confusing notion, and it is important to keep clearly in mind the distinction between open subsets of a set $S \subset \mathbb{R}^n$, those open subsets of $\mathbb{R}^n$ that are contained in S, and relatively open subsets of the set S, subsets of S that are open in the relative topology of S but are not necessarily open subsets of $\mathbb{R}^n$.

The **boundary** of a subset $E \subset \mathbb{R}^n$ is the closed set $\partial E = \overline{E} \cap \overline{(\sim E)}$. The boundaries of ϵ neighborhoods $\mathcal{N}_\epsilon(\mathbf{a})$ and of cells Δ as defined earlier are also their boundaries in this sense. For more general sets some caution is necessary, since the boundary may not always correspond to what naively might be viewed as the boundary of the set; for example if E is the set of all points in an open cell Δ that have rational coordinates then $\partial E = \overline{\Delta}$.

Two subsets $E, F \subset \mathbb{R}^n$ are **separated** if $\overline{E} \cap F = E \cap \overline{F} = \emptyset$. This is somewhat stronger than the condition that the two sets are disjoint; for example $E = [0, 1)$ and $F = [1, 2]$ are disjoint subsets of $\mathbb{R}$ but are not separated since $\overline{E} \cap F = 1$, but $E = [0, 1)$ and $G = (1, 2]$ are separated subsets of $\mathbb{R}$. A subset $S \subset \mathbb{R}^n$ is **connected** if it cannot be written as the union of two nonempty separated subsets. Alternatively a topological space S, such as a subset of $\mathbb{R}^n$ with the relative topology, is connected if it cannot be written as a disjoint union of two relatively open subsets of S, or equivalently if there is no subset of S that is both relatively open and relatively closed, other than the empty set and all of S. The equivalence of these conditions is quite obvious. It is easy to see that the entire space $\mathbb{R}^n$ is connected. Indeed if $U \subset \mathbb{R}^n$ is a subset other than the empty set or all of $\mathbb{R}^n$ that is both open and closed, choose a point $\mathbf{a} \in U$ and a point $\mathbf{b} \in \mathbb{R}^n \sim U$. The set of real numbers s such that $\mathbf{a} + t(\mathbf{b} - \mathbf{a}) \in U$ for $0 \le t < s$ is nonempty, since U is open, and it is bounded above since $\mathbf{b} \notin U$, so this set has a least upper bound s_0, one of the characteristic properties of the real number system. Since U is closed it must be the case that $\mathbf{a} + s_0\mathbf{b} \in U$, for this is the limit of the points $\mathbf{a} + s\mathbf{b} \in U$ as $s \to s_0$; but then since U is open it also must be the case that $\mathbf{a} + s\mathbf{b} \in U$ for some real numbers $s > s_0$, a contradiction. The same argument shows that an open neighborhood $\mathcal{N}_r(a)$ and an open cell in $\mathbb{R}^n$ are connected sets. On the other hand the subset $\mathbb{Q} \subset \mathbb{R}$ of rational points with the relative topology it inherits from $\mathbb{R}$ is not a connected set; indeed the interval $[\sqrt{2}, \sqrt{3}] \subset \mathbb{Q}$ is both open and closed.

An **open covering** of a subset $E \subset \mathbb{R}^n$ is a collection of open sets U_α, not necessarily a countable collection, such that $E \subset \bigcup_\alpha U_\alpha$. If some of the sets U_α are redundant they can be eliminated and the remaining sets are also an open covering of E, called a **subcovering** of the set E. A set E is said to be **compact** if every open covering of E has a finite subcovering, that is, if for any open covering $\{U_\alpha\}$ of E finitely many of the sets U_α actually cover all of E. This is a rather subtle notion, but it is very important and frequently used. To show that a set E is not compact it suffices to find a single open covering $\{U_\alpha\}$ of E such that no finite collection of the sets U_α can cover E; but to show that E is compact it is necessary to show that for *any* open covering $\{U_\alpha\}$ of E finitely many of the sets U_α actually will cover E. An example of a non-compact subset of $\mathbb{R}^n$ is a nonempty open neighborhood $\mathcal{N}_1(0)$ of the origin in $\mathbb{R}^n$; this set is covered by the open subsets $U_\nu = \mathcal{N}_{1-\frac{1}{\nu}}(0)$ for $\nu = 1, 2, 3, \ldots$, but the union of any finite collection of these subsets will be just the set U_ν for the largest ν in the collection, and that is a proper subset of $\mathcal{N}_1(0)$. An open cell is also noncompact, for essentially the same reason. A closed cell is an example of a compact subset; but the proof is rather more subtle and rests on the basic topological properties of the real number system, just as did the proof that $\mathbb{R}^n$

is connected. For the proof it is convenient to define the **edgesize** of an open or closed cell $\Delta = \{\, \mathbf{x} = \{x_j\} \in \mathbb{R}^n \,\big|\, a_j \le x_j \le b_j \,\} \subset \mathbb{R}^n$ to be the nonnegative number $e(\Delta) = \max_{1\le j\le n}(b_j - a_j)$.

complm **Lemma 1.1** *If $\overline{\Delta}_\nu \subset \mathbb{R}^n$ are closed cells in $\mathbb{R}^n$ for $\nu = 1, 2, \dots$ such that $\overline{\Delta}_{\nu+1} \subset \overline{\Delta}_\nu$ and $\lim_{\nu\to\infty} e(\overline{\Delta}_\nu) = 0$ then $\bigcap_\nu \overline{\Delta}_\nu$ is a single point of $\mathbb{R}^n$.*

Proof: If $\overline{\Delta}_\nu = \{\, \mathbf{x} = \{x_j\} \,\big|\, a_j^\nu \le x_j \le b_j^\nu \,\}$ then for each j clearly $a_j^{\nu+1} \ge a_j^\nu$ and $b_j^{\nu+1} \le b_j^\nu$; and since $\overline{\Delta}_\nu \subset \overline{\Delta}_1$ it is also the case that $a_j^\nu \le b_j^1$ and $b_j^\nu \ge a_j^1$. The basic completeness property of the real number system implies that any increasing sequence of real numbers bounded from above and any decreasing sequence of real numbers bounded from below have limiting values; therefore $\lim_{\nu\to\infty} a_j^\nu = a_j$ and $\lim_{\nu\to\infty} b_j^\nu = b_j$ for some uniquely determined real numbers a_j, b_j, and it is clear that the cell $\overline{\Delta} = \{\, \mathbf{x} = \{x_j\} \,\big|\, a_j \le x_j \le b_j \,\}$ is contained in the intersection $\bigcap_\nu \overline{\Delta}_\nu$. On the other hand since $(b_j - a_j) \le (b_j^\nu - a_j^\nu) \le e(\overline{\Delta}_\nu)$ and $\lim_{\nu\to\infty} e(\overline{\Delta}_\nu) = 0$ it must be the case that $b_j = a_j$ so the limiting cell $\overline{\Delta}$ is just a single point of $\mathbb{R}^n$, which concludes the proof.

cellcom **Theorem 1.2** *A closed cell in $\mathbb{R}^n$ is compact.*

Proof: If a closed cell $\overline{\Delta} = \{\, \mathbf{x} = \{x_j\} \,\big|\, a_j \le x_j \le b_j \,\}$ is not compact there is an open covering $\{U_\alpha\}$ of $\overline{\Delta}$ that does not admit any finite subcovering. The cell $\overline{\Delta}$ can be written as the union of the closed cells arising from bisecting each of its sides. If finitely many of the sets $\{U_\alpha\}$ covered each of the subcells then finitely many would cover the entire set $\overline{\Delta}$, which is not the case; hence at least one of the subcells cannot be covered by finitely many of the sets $\{U_\alpha\}$. Then bisect each of the sides of that subcell, and repeat the process. The result is that there is a collection of closed cells $\overline{\Delta}_\nu$ which cannot be covered by finitely many of the open sets $\{U_\alpha\}$ and for which $\overline{\Delta}_{\nu+1} \subset \overline{\Delta}_\nu \subset \overline{\Delta}$, and $\lim_{\nu\to\infty} e(\overline{\Delta}_\nu) \longrightarrow 0$. It then follows from the preceding lemma that $\bigcap_\nu \overline{\Delta}_\nu = \mathbf{a}$, a single point in $\mathbb{R}^n$. This point must be contained within one of the sets $\{U_{\alpha_0}\}$, and if ν is sufficiently large then $\overline{\Delta}_\nu \subset U_{\alpha_0}$ as well; but that is a contradiction, since the cells $\overline{\Delta}_\nu$ were chosen so that none of them could be covered by finitely many of the sets $\{U_\alpha\}$. Therefore the cell $\overline{\Delta}$ is compact, which concludes the proof.

clcomp **Theorem 1.3** *A closed subset of a compact set is compact.*

Proof: Suppose that $E \subset F$ where F is compact and E is closed, and that $\{U_\alpha\}$ is an open covering of the set E. The sets U_α together with the open set $\mathbb{R}^n \sim E$ form a covering of the compact set F, so finitely many of these sets cover F hence also cover E. Clearly the set $\mathbb{R}^n \sim E$ covers none of the points of E, so the remaining finitely many sets U_α necessarily cover E. Therefore E is compact, which suffices for the proof.

hb **Theorem 1.4 (Heine-Borel Theorem)** *A subset $E \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof: A bounded subset $E \subset \mathbb{R}^n$ is contained in a sufficiently large closed cell $\overline{\Delta} \subset \mathbb{R}^n$; the set $\overline{\Delta}$ is compact by Theorem cellcom 1.2, so if E is also closed then by Theorem clcomp 1.3 it is compact. Conversely suppose that E is a compact set. If E is not bounded it can be covered by the collection of open neighborhoods $\mathcal{N}_\nu(0)$ for $\nu = 1, 2, \ldots$, but it is not covered by any finite set of these neighborhoods; that contradicts the compactness of E, so a compact set is bounded. If E is not closed then there is a limit point $\mathbf{a} \in E'$ that is not contained in E. Each closed neighborhood $\overline{\mathcal{N}}_{1/\nu}(\mathbf{a})$ of the point $\mathbf{a}$ must contain a point of E, since $\mathbf{a} \in E'$, but the intersection of all of these neighborhoods consists of the point $\mathbf{a}$ itself, which is not contained in E. The complements $U_\nu = \mathbb{R}^n \sim \overline{\mathcal{N}}_{1/\nu}(\mathbf{a})$ then form an open covering of E, but no finite number of these sets suffice to cover E since the union of finitely many such sets must be one of the sets U_ν and hence does not cover the points of E contained in $\overline{\mathcal{N}}_{1/\nu}(\mathbf{a})$; and that again is a contradiction, which suffices to conclude the proof.

cas **Theorem 1.5 (Casorati-Weierstrass Theorem)** *A subset $E \subset \mathbb{R}^n$ is compact if and only if every sequence of distinct points in E has a limit point in E.*

Proof: Suppose that E is compact. If $\mathbf{a}_\nu \subset E$ is a collection of distinct points of E with no limit points in E then the points $\mathbf{a}_\nu$ can have no limit points in $\mathbb{R}^n$, for since E is closed by the Heine-Borel Theorem these limit points would necessarily lie in E; in particular the set $\bigcup_\nu a_\nu$ is a closed set. Each point $\mathbf{a}_\nu$ has an open neighborhood U_ν that contains none of the other points, since otherwise $\mathbf{a}_\nu$ would itself be a limit point of this collection of points. These open sets U_ν together with the open set $\mathbb{R}^n \sim (\bigcup_\nu a_\nu)$ form an open covering of E; and since E is compact finitely many of these sets already cover E. That is a contradiction, since the set $\mathbb{R}^n \sim (\bigcup_\nu a_\nu)$ does not cover any of the points a_ν and no finite collection of the sets U_ν cover all the points a_ν. On the other hand, suppose that $E \subset \mathbb{R}^n$ is not compact. Then by the Heine-Borel Theorem either E is not bounded or E is not closed. If E is not bounded it must contain a sequence of distinct points $\mathbf{a}_\nu$ such that $|\mathbf{a}_\nu|$ is a strictly increasing sequence of real numbers with no finite limit, and this sequence can have no limit point in E. On the other hand if E is not closed it contains a sequence of distinct points $\mathbf{a}_\nu$ with a limit point not contained in E. That suffices to conclude the proof.

As already noted, for any set $S \subset \mathbb{R}^n$ a subset $E \subset S$ that is relatively open in S need not be open in $\mathbb{R}^n$ and a subset $F \subset S$ that is relatively closed in S need not be closed in $\mathbb{R}^n$. However a subset $E \subset S$ is compact in the relative topology of S, in which the definition of compactness is expressed in terms of open coverings by relatively open sets in S, if and only E is compact as a subset of $\mathbb{R}^n$. Indeed that is quite clear, since the relatively open sets A_α in S are the intersections $A_\alpha = S \cap U_\alpha$ of S with open subsets U_α in $\mathbb{R}^n$ so that $E \subset \bigcup_\alpha A_\alpha$ if and only if $E \subset \bigcup_\alpha U_\alpha$. Thus the property that a set be compact is a much more intrinsic property of that set than that it be either open or closed. Note carefully though that the Heine-Borel Theorem only holds for subsets of the entire space $\mathbb{R}^n$; the proof quite explicitly rested on properties of closed cells in $\mathbb{R}^n$, and it is clear that a subset of an open cell $\Delta \subset \mathbb{R}^n$ can be relatively closed and bounded in Δ but not closed in $\mathbb{R}^n$ hence not compact. However it is clear that the Casorati-Weierstrass Theorem holds in any subset $S \subset \mathbb{R}^n$ in terms of the relative topology of S, since the limit points of a subset $E \subset S$ are just the points of the intersection $E' \cap S$.

1.3 Continuous Mappings

A **mapping f** from a subset $U \subset \mathbb{R}^m$ to a subset $V \subset \mathbb{R}^n$ associates to each point $\mathbf{x} = \{x_j\} \in U$ a point $\mathbf{f}(\mathbf{x}) = \mathbf{y} = \{y_j\} \in V$; the coordinates y_j depend on the point $\mathbf{x}$ so can be viewed as given by functions $y_j = f_j(\mathbf{x})$, which are the **coordinate functions** of the mapping $\mathbf{f}$. The mapping f is **continuous** at a point $\mathbf{a} \in U$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $||\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{x})|| \le \epsilon$ whenever $\mathbf{x} \in U$ and $||\mathbf{x}-\mathbf{a}|| \le \delta$, or equivalently, for every $\epsilon > 0$ there is a $\delta > 0$ such that $\mathbf{f}\big(U \cap \mathcal{N}_\delta(\mathbf{a})\big) \subset V \cap \mathcal{N}_\epsilon(\mathbf{f}(\mathbf{a}))$. It is clear from the inequalities (1.7) normcmp that in the definition of continuity the norm $||\mathbf{x}||$ can be either the Euclidean norm $||\mathbf{x}||_2$ or the supremum norm $||\mathbf{x}||_\infty$. It is also clear that the mapping $\mathbf{f}$ is continuous at a point $\mathbf{a}$ if and only if each of its coordinate functions f_j is continuous at that point. The mapping is said to be **continuous in the subset** U if it is continuous at each point of U.

congl **Theorem 1.6** *A mapping $\mathbf{f} : S \longrightarrow \mathbb{R}^n$ from a subset $S \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is continuous in S if and only if $f^{-1}(U)$ is a relatively open subset of S for any open subset $U \subset \mathbb{R}^n$.*

Proof: If $\mathbf{f} : S \longrightarrow \mathbb{R}^n$ is continuous, $U \subset \mathbb{R}^n$ is an open subset and $\mathbf{a} \in \mathbf{f}^{-1}(U)$ then $\mathbf{f}(\mathbf{a}) = \mathbf{b} \in U$; and since U is open there is an open neighborhood $\mathcal{N}_\epsilon(\mathbf{b})$ of the point $\mathbf{b}$ such that $\mathcal{N}_\epsilon(\mathbf{b}) \subset U$. Since $\mathbf{f}$ is continuous there is a $\delta > 0$ such that $\mathbf{f}\big(S \cap \mathcal{N}_\delta(\mathbf{a})\big) \subset \mathcal{N}_\epsilon(\mathbf{b})$, hence that $S \cap \mathcal{N}_\delta(\mathbf{a}) \subset \mathbf{f}^{-1}\Big(\mathcal{N}_\epsilon(b)\Big) \subset \mathbf{f}^{-1}(U)$; that is the case for any point $\mathbf{a} \in \mathbf{f}^{-1}(U)$, hence $\mathbf{f}^{-1}(U)$ is a relatively open subset of S. On the other hand if $\mathbf{f}^{-1}(U)$ is a relatively open subset of S for any open subset $U \subset \mathbb{R}^n$ then for any point $\mathbf{a} \in S$ with image $\mathbf{f}(\mathbf{a}) = \mathbf{b} \in \mathbb{R}^n$ and for any $\epsilon > 0$ sufficiently small that $\mathcal{N}_\epsilon(\mathbf{b}) \subset U$ the set $\mathbf{f}^{-1}(\mathcal{N}_\epsilon(\mathbf{b}))$ is a relatively open subset of S; since the point $\mathbf{a}$ is in this set there must be some $\delta > 0$ for which $S \cap \mathcal{N}_\delta(\mathbf{a}) \subset \mathbf{f}^{-1}(\mathcal{N}_\epsilon(\mathbf{b}))$ or equivalently for which $\mathbf{f}\big(S \cap \mathcal{N}_\delta(\mathbf{a})\big) \subset \mathcal{N}_\epsilon(\mathbf{b})$, showing that $\mathbf{f}$ is continuous at the point $\mathbf{a}$. That is the case for any point $\mathbf{a} \in S$, so $\mathbf{f}$ is continuous in S, which suffices for the proof.

conglx **Corollary 1.1** *A mapping $\mathbf{f} : S \longrightarrow \mathbb{R}^n$ from a subset $S \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is continuous in S if and only if $\mathbf{f}^{-1}(E)$ is a relatively closed subset of S for any closed subset $E \subset \mathbb{R}^n$.*

Proof: This is an immediate consequence of the preceding theorem, since a subset of S is closed in the relative topology of S if and only if its complement in S is relatively open and $\mathbf{f}^{-1}(\mathbb{R}^n \sim E) = \big(S \sim \mathbf{f}^{-1}(E)\big)$ for any subset $E \subset \mathbb{R}^n$. That suffices for the proof.

These results show that the continuity of a mapping in a set S really is a property of the topology of S. Indeed a mapping between two general topological spaces is defined to be continuous if and only if the inverse of any open subset is open. A simple consequence of this characterization of continuity is that if $\mathbf{g} : U \longrightarrow V$ and $\mathbf{f} : V \longrightarrow W$ are continuous mappings between subsets U, V, W of some Euclidean spaces then the composition $\mathbf{f} \circ \mathbf{g} : U \longrightarrow W$,the mapping that takes a point $\mathbf{x} \in U$ to the point $(\mathbf{f} \circ \mathbf{g})(\mathbf{x}) = \mathbf{f}\big(\mathbf{g}(\mathbf{x})\big)$, is also continuous; indeed if $E \subset W$ is open then $\mathbf{f}^{-1}(E)$ is open since $\mathbf{f}$ is continuous and $(\mathbf{f} \circ \mathbf{g})^{-1}(E) = \mathbf{g}^{-1}\big(\mathbf{f}^{-1}(E)\big)$ is open since $\mathbf{g}$ is continuous, and consequently $(\mathbf{f} \circ \mathbf{g})$ is continuous. The results in Theorem `congl` 1.6 and Corollary `conglx` 1.1 involve the inverse image of a set under a mapping f; the image of an open set under a continuous mapping is not necessarily open, and the image of a closed set under a continuous mapping is not necessarily closed. For instance if $f : \mathbb{R} \longrightarrow \mathbb{R}$ is the continuous mapping $f(x) = e^{-x^2}$ then $f(\mathbb{R}) = \{\, x \,\big|\, 0 < x \leq 1 \,\}$, which is neither open nor closed although $\mathbb{R}$ itself is both open and closed.

`concomim` **Theorem 1.7** *If $f : S \longrightarrow \mathbb{R}^n$ is a continuous mapping from a subset $S \subset \mathbb{R}^m$ into $\mathbb{R}^n$ then the image $f(E)$ of any compact subset $E \subset S$ is a compact subset of $\mathbb{R}^n$.*

Proof: If $E \subset S$ is compact and $f(E)$ is contained in a union of open sets U_α then E is contained in the union of the open sets $f^{-1}(U_\alpha)$; and since E is compact it is contained in a union of finitely many of the sets $f^{-1}(U_\alpha)$, so E is contained in the union of the inverse images of these finitely many open sets. That suffices for the proof.

Since a compact subset of $\mathbb{R}^n$ is closed, by the Heine-Borel Theorem, the inverse under a continuous mapping of a compact set is necessarily closed; but it is not necessarily compact. For example the inverse image of the set $[-1, 1] \subset \mathbb{R}$ under the mapping $f : \mathbb{R} \longrightarrow \mathbb{R}$ given by $f(x) = \sin x$ is an unbounded set hence is not compact.

`concomimcor` **Theorem 1.8** *If f is a continuous function on a compact set $U \subset \mathbb{R}^n$ then there are points $\mathbf{a}, \mathbf{b} \in U$ such that*

`concomimcorq1` (1.19)
$$f(\mathbf{a}) = \sup_{\mathbf{x} \in U} f(\mathbf{x}) \quad \textit{and} \quad f(\mathbf{b}) = \inf_{\mathbf{x} \in U} f(\mathbf{x}).$$

Proof: The image $f(U) \subset \mathbb{R}$ is compact by the preceding theorem, hence is a closed set; so if $\alpha = \sup_{\mathbf{x} \in U} f(\mathbf{x})$ then since α is a limit point of the set $f(U)$ it must be contained in the set $f(U)$ hence $\alpha = f(\mathbf{a})$ for some point $\mathbf{a} \in U$, and correspondingly for $\beta = \inf_{\mathbf{x} \in U} f(\mathbf{x})$. That suffices for the proof.

`conbij` **Theorem 1.9** *A one-to-one and continuous mapping from a compact subset $U \subset \mathbb{R}^m$ onto a subset $V \subset \mathbb{R}^n$ has a continuous inverse.*

Proof: If the mapping $\mathbf{f} : U \longrightarrow V$ is one-to-one it has a well defined inverse mapping $\mathbf{g} : V \longrightarrow U$. To show that $\mathbf{g}$ is continuous it suffices to show that $\mathbf{g}^{-1}(E)$ is closed for any closed subset $E \subset U$, in view of Corollary 1.1. If E is closed then by Theorem 1.3 it is compact, since U is compact; and then $\mathbf{g}^{-1}(E) = \mathbf{f}(E)$ is compact by Theorem 1.7, hence is closed by the Heine-Borel Theorem, and that suffices for the proof.

The assumption of compactness is essential in the preceding theorem. For example the mapping $\mathbf{f} : [0, 2\pi) \longrightarrow \mathbb{R}^2$ defined by

$$\mathbf{f}(t) = (\cos t, \sin t) \in \mathbb{R}^2$$

is clearly one-to-one and continuous; but the inverse mapping fails to be continuous at the point $\mathbf{f}(0) = (1, 0)$. Some properties of continuity are not purely topological properties, in the sense that they cannot be stated purely in terms of open and closed sets, but are metric properties, involving the norms used to define continuity. By definition a mapping $\mathbf{f} : U \longrightarrow W$ between two subsets of Euclidean spaces is continuous at a point $\mathbf{a} \in U$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ such that $||\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{x})|| \le \epsilon$ whenever $||\mathbf{x} - \mathbf{a}|| \le \delta$. The mapping $\mathbf{f}$ is continuous in the set U if for each point $\mathbf{a} \in U$ and any $\epsilon > 0$ there is a $\delta_{\mathbf{a}} > 0$, which may depend on the point $\mathbf{a}$, such that $||\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{x})|| \le \epsilon$ whenever $||\mathbf{x} - \mathbf{a}|| \le \delta_{\mathbf{a}}$. The mapping $\mathbf{f}$ is **uniformly continuous** in U if it is possible to find values $\delta_{\mathbf{a}}$ that are independent of the point $\mathbf{a} \in U$. Equivalently the mapping $\mathbf{f}$ is uniformly continuous in U if for any $\epsilon > 0$ there exists $\delta > 0$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < \epsilon$ whenever $\|\mathbf{x} - \mathbf{y}\| < \delta$. It should be kept in mind that the two norms are in different spaces; and it is evident from the inequalities (1.7) that they may be different norms as well. Not all continuous mappings are uniformly continuous, as for example the mapping $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $f(x) = x^2$; but in some circumstances continuous mappings are automatically uniformly continuous.

topunf **Theorem 1.10** *A continuous mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ from a compact subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is uniformly continuous.*

Proof: If $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ and $\epsilon > 0$ then for any point $\mathbf{a} \in U$ there is $\delta_{\mathbf{a}} > 0$ such that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})\| < \frac{1}{2}\epsilon$ whenever $\mathbf{x} \in \mathcal{N}_{\delta_{\mathbf{a}}}$. The collection of neighborhoods $\mathcal{N}_{\frac{1}{2}\delta_{\mathbf{a}}}(\mathbf{a})$ for all points $\mathbf{a} \in U$ are an open covering of U, and if U is compact finitely many of these neighborhoods cover all of U. If $\delta > 0$ is the minimum of the finitely many positive numbers $\delta_{\mathbf{a}}$ for this finite covering, and if $\mathbf{x}, \mathbf{y} \in U$ are any two points such that $\|\mathbf{x} - \mathbf{y}\| < \frac{1}{2}\delta$, then $\mathbf{x} \in \mathcal{N}_{\frac{1}{2}\delta}(\mathbf{a})$ for one of these neighborhoods, and since $\|\mathbf{x} - \mathbf{y}\| < \frac{1}{2}\delta$ it is also the case that $\mathbf{y} \in \mathcal{N}_{\delta}(\mathbf{a})$. It follows that $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \le \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a})\| + \|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{y})\| \le \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon$, so $\mathbf{f}$ is uniformly continuous, which concludes the proof.

PROBLEMS: GROUP 1

(1) Sketch the following subsets of $\mathbb{R}^2 = \{\, \mathbf{x} = (x_1, x_2) \,\}$, and for each subset indicate whether it is open, closed, compact, or connected and find its boundary.

(i) $\{\, \mathbf{x} \mid 0 < \|\mathbf{x}\|_\infty \le 1 \,\}$

(ii) $\{\, \mathbf{x} \mid 0 < x_1 \le 1,\ \ x_2 = \sin \frac{1}{x_1} \,\}$

(iii) $\{\, \mathbf{x} \mid \|\mathbf{x}\|_\infty \le 1 \,\} \sim \{\, \mathbf{x} \mid 0 < \|\mathbf{x}\|_2 < 1 \,\}$

(iv) $\{\, \mathbf{x} \mid |x_1| \le 1 \,\} \cap \{\, \mathbf{x} \mid x_2 \le 1 \,\}$

(2) Indicate the points at which each of the following functions in $\mathbb{R}^2$ is continuous.

(i) $f(\mathbf{x}) = \begin{cases} x_2 - \sin x_1 & \text{if} \quad x_1 \ge 0, \\ x_1 + x_2 & \text{if} \quad x_1 < 0; \end{cases}$ (ii) $g(\mathbf{x}) = \begin{cases} x_1 + x_2 & \text{if} \quad \|\mathbf{x}\|_\infty \le 1, \\ x_1^2 + x_2^2 & \text{if} \quad \|\mathbf{x}\|_\infty > 1. \end{cases}$

(3) Show that if f and g are continuous functions in an open subset $U \subset \mathbb{R}^n$ then $\max(f, g)$, $\min(f, g)$, and $|f|$ are also continuous; show that the converses do not hold, so if $\max(f, g)$ is continuous it is not necessarily the case that f and g are continuous, and similarly for $\min(f, g)$ and $|f|$.

(4) Show that the function $\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j|$ of a vector $\mathbf{x} = \{x_j\} \in \mathbb{R}^n$ is also a norm. What is the shape of an open neighborhood of the origin in the plane $\mathbb{R}^2$ defined by this norm?

(5) Show that the continuous image of a connected set is connected.

(6) Show that if $E_i \subset \mathbb{R}^n$ are nonempty compact subsets for $i = 1, 2, \ldots$ such that $E_{i+1} \subset E_i$ then $\bigcap_i E_i$ is also nonempty. Is that the case if the subsets E_i are closed but not necessarily compact?

PROBLEMS: GROUP 2

(7) Show that a norm $\|\mathbf{x}\|_0$ on vectors in $\mathbb{R}^n$ can be defined by an inner product $(\mathbf{x}_1, \mathbf{x}_2)_0$ on $\mathbb{R}^n$, in the sense that $\|\mathbf{x}\|_0^2 = (\mathbf{x}, \mathbf{x})_0$, if and only if the norm satisfies the *parallelogram law*:

$$\|\mathbf{x} + \mathbf{y}\|_0^2 + \|\mathbf{x} - \mathbf{y}\|_0^2 = 2\|\mathbf{x}\|_0^2 + 2\|\mathbf{y}\|_0^2.$$

As a consequence, show that the supremum norm cannot be defined by an inner product.

(8) The set of all real polynomials in a single variable x form a real vector space $\mathcal{P}$, but not a vector space of finite dimension; nonetheless it is possible

to define norms on this vector space. Show that the expressions $\|p(x)\|_\infty$ and $\|p(x)\|_c$ defined for a polynomial $p(x) = a_0 + a_1x + a_2x^2 + \cdots$ by

$$\|p(x)\|_\infty = \sup_{-\frac{1}{2}\le x\le\frac{1}{2}} |p(x)| \quad \text{and} \quad \|p(x)\|_c = \max\Big(|a_0|, |a_1|, |a_2|, \cdots\Big)$$

are norms on the vector space $\mathcal{P}$, but that these two norms are not equivalent norms.

(9) Show that $(E')' \subset E'$ for any subset $E \subset \mathbb{R}^n$, but that the containment is not necessarily an equality.

(10) There are at least 3 topologies that have been described on the sets $\mathbb{R}^n$: the ordinary topology, defined by the Euclidean or supremum norm; the discrete topology; and the indiscrete topology. For each pair of these topologies determine whether the identity mapping $\boldsymbol{\phi} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a continuous mapping when the first space $\mathbb{R}^n$ has one topology of the pair and the second space $\mathbb{R}^n$ has the other topology of the pair.

(11) Show that any two norms on $\mathbb{R}^n$ are equivalent.

(12) Show that if $f(x_1, x_2)$ is a continuous function in the unit cell $\overline{\Delta} \subset \mathbb{R}^2$ then the function $g(x_1) = \sup_{0\le x_2\le 1} f(x_1, x_2)$ is a continuous function on the unit interval $[0, 1]$.

(13) Let $\boldsymbol{\phi} : E \longrightarrow E$ be a mapping from a subset $E \subset \mathbb{R}^n$ into itself such that $||\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y})||_2 = ||\mathbf{x} - \mathbf{y}||_2$ for any points $\mathbf{x}, \mathbf{y} \in E$.
(i) Show that if E is compact the mapping $\boldsymbol{\phi} : E \longrightarrow E$ is a homeomorphism, a continuous one-to-one mapping from E onto E with a continuous inverse.
(ii) Is the mapping $\boldsymbol{\phi}$ necessarily a homeomorphism if the set E is not compact? Why?

(14) Show that if $E_\nu \subset \mathbb{R}^n$ are compact connected subsets such that $E_{\nu+1} \subset E_\nu$ then $E = \bigcap_{\nu=1}^{\infty} E_\nu$ is also connected. Is that the case if the subsets E_ν are not assumed to be compact? Why?

Chapter 2

Differentiable Mappings

2.1 The Derivative

A mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ from an open subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is **differentiable** at a point $\mathbf{a} \in U$ if there is a linear mapping $A : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ described by an $n \times m$ matrix A such that for all $\mathbf{h}$ in an open neighborhood of the origin in $\mathbb{R}^m$

$$\mathbf{f}(\mathbf{a}+\mathbf{h}) = \mathbf{f}(\mathbf{a}) + A\mathbf{h} + \boldsymbol{\epsilon}(\mathbf{h}) \quad \text{where } \lim_{\mathbf{h}\to 0} \tfrac{\|\boldsymbol{\epsilon}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0. \tag{2.1}$$

Here $\mathbf{h} \in \mathbb{R}^m$ so $A\mathbf{h} \in \mathbb{R}^n$, while $\mathbf{f}(\mathbf{a}),\ \mathbf{f}(\mathbf{a}+\mathbf{h}),\ \boldsymbol{\epsilon}(\mathbf{h}) \in \mathbb{R}^n$. This definition is independent of the norm chosen; for if $\lim_{\mathbf{h}\to 0} \frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2} = 0$ then from the inequalities (1.7) it follows that $\frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} \le \frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2}\sqrt{n}$ so $\lim_{\mathbf{h}\to 0} \frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} = 0$ as well, and the converse holds similarly. If $\mathbf{f}$ is differentiable at $\mathbf{a}$ it is continuous at $\mathbf{a}$, since $\lim_{\mathbf{h}\to 0} A\mathbf{h} = 0$ and $\lim_{\mathbf{h}\to 0} \boldsymbol{\epsilon}(\mathbf{h}) = 0$. For example, if $m = 3$ and $n = 2$ so that $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ then (2.1) takes the form

$$\begin{pmatrix} f_1(\mathbf{a}+\mathbf{h}) \\ f_2(\mathbf{a}+\mathbf{h}) \end{pmatrix} = \begin{pmatrix} f_1(\mathbf{a}) \\ f_2(\mathbf{a}) \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1(\mathbf{h}) \\ \epsilon_2(\mathbf{h}) \end{pmatrix}.$$

If $\mathbf{f}$ is a constant mapping then $\mathbf{f}(\mathbf{a}+\mathbf{h}) = \mathbf{f}(\mathbf{a})$, which is (2.1) for $A = 0$ and $\boldsymbol{\epsilon}(\mathbf{h}) = 0$; so any constant mapping is differentiable. If $\mathbf{f}$ is the linear mapping $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$ for a matrix A then $\mathbf{f}(\mathbf{a}+\mathbf{h}) = A(\mathbf{a}+\mathbf{h}) = \mathbf{f}(\mathbf{a}) + A\mathbf{h}$, which is (2.1) for the matrix A and $\boldsymbol{\epsilon}(\mathbf{h}) = 0$, so a linear mapping also is differentiable. If $m = n = 1$ it is possible to divide (2.1) by the real number h and to rewrite that equation in the form

$$\lim_{h\to 0} \frac{f(a+h) - f(a)}{h} - A = \lim_{h\to 0} \frac{\epsilon(h)}{h} = 0;$$

this is a form of the familiar definition that the real-valued function $f(x)$ of the variable $x \in \mathbb{R}$ is differentiable at the point a and that its derivative at that point is the real number A.

diffloc **Theorem 2.1** *A mapping* $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ *is differentiable at a point* $\mathbf{a}$ *if and only if each of the coordinate functions* f_i *of the mapping* $\mathbf{f}$ *is differentiable at that point.*

Proof: If $\mathbf{f}$ is a differentiable mapping it follows from (2.1) that each coordinate function f_i satisfies

diff2
$$f_i(\mathbf{a}+\mathbf{h}) = f_i(\mathbf{a}) + \sum_{j=1}^{m} a_{ij}h_j + \epsilon_i(\mathbf{h}) \ \text{ where } \lim_{\mathbf{h}\to 0} \frac{|\epsilon_i(\mathbf{h})|}{\|\mathbf{h}\|} = 0, \tag{2.2}$$

since $\frac{|\epsilon_i(\mathbf{h})|}{\|\mathbf{h}\|_\infty} \le \frac{\|\epsilon(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty}$; and this is just the condition that each of the coordinate functions f_i of the mapping $\mathbf{f}$ is differentiable at the point $\mathbf{a}$. Conversely if each of the coordinate mappings f_i is differentiable at the point $\mathbf{a}$ then (2.2) holds for $1 \le i \le n$. The collection of these n equations taken together form the equation (2.1) in which $\boldsymbol{\epsilon}(\mathbf{h}) = \{\epsilon_i(\mathbf{h})\}$; and since $\frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} = \max_{1\le i\le n} \frac{|\epsilon_i(\mathbf{h})|}{\|\mathbf{h}\|_\infty}$ it follows that $\lim_{\mathbf{h}\to 0} \frac{\|\boldsymbol{\epsilon}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} = 0$, so the mapping $\mathbf{f}$ is differentiable. That concludes the proof.

For the special case of a vector $\mathbf{h} = \{h_j\}$ for which $h_j = 0$ for $j \ne k$ for some index k, equation (2.2) takes the form

$$f_i(a_1, \ldots, a_k + h_k, \ldots, a_m) = f_i(a_1, \ldots, a_k, \ldots, a_m) + a_{ik}h_k + \epsilon_i(h_k)$$

where $\lim_{h_k\to 0} \frac{|\epsilon_i(h_k)|}{|h_k|} = 0$; that is just the condition that $f_i(\mathbf{x})$, viewed as a function of the variable x_k alone for fixed values $x_j = a_j$ for the remaining variables for $j \ne k$, is a differentiable function of that variable x_k and that its derivative at the point $x_k = a_k$ is the real number a_{ik}. The constant a_{ik} is called the **partial derivative** of the function f_i with respect to the variable x_k at the point $\mathbf{a}$, and is denoted by $a_{ik} = \partial_k f_i(\mathbf{a})$. It follows that the entries in the matrix $A = \{a_{ik}\}$ are the uniquely determined partial derivatives of the coordinate functions of the mapping; this matrix is called the **derivative** of the mapping $\mathbf{f}$ at the point $\mathbf{a}$ and is denoted by $\mathbf{f}'(\mathbf{a})$, so

dif1
$$\mathbf{f}'(\mathbf{a}) = \Big\{ \mathbf{f}'(\mathbf{a})_{ik} = \partial_k f_i(\mathbf{a}) \ \Big| \ 1 \le k \le m,\ 1 \le i \le n \Big\}. \tag{2.3}$$

Generally it is a fairly straightforward matter to calculate the partial derivatives of a function; merely consider all the variables except one of them as constants, and apply the familiar techniques for calculating derivatives of a function of one variable. That can be applied to each coordinate function of a mapping, to yield the derivative of that mapping. It is evident from (2.1) that a linear combination $c_1\mathbf{f}_1 + c_2\mathbf{f}_2$ of two differentiable mappings $\mathbf{f}_1, \mathbf{f}_2 : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ at a point $\mathbf{a}$ for any constants $c_1, c_2 \in \mathbb{R}$ is again a differentiable mapping at that point, and it follows from (2.3) that differentiation is linear in the sense that $(c_1\mathbf{f}_1 + c_2\mathbf{f}_2)'(\mathbf{a}) = c_1\mathbf{f}_1'(\mathbf{a}) + c_2\mathbf{f}_2'(\mathbf{a})$. There are various alternative notations for derivatives and partial derivatives of functions of several variables that are in common use. For instance $D\mathbf{f}(\mathbf{a})$ is often used for $f'(\mathbf{a})$, and $D_k f(\mathbf{a})$ or $\dfrac{\partial f}{\partial x_k}(\mathbf{a})$

for $\partial_k f(\mathbf{a})$ when $f(\mathbf{x})$ is a real-valued function of the variable $\mathbf{x} \in \mathbb{R}^n$; and when the mapping $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is viewed as giving the coordinates y_i of a point $\mathbf{y} \in \mathbb{R}^n$ as functions $y_i(\mathbf{x})$ of the coordinates x_j of points $\mathbf{x} \in \mathbb{R}^m$ the notation $\dfrac{\partial y_i}{\partial x_j}$ is quite commonly used for $\partial_j y_i$.

If a mapping $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is differentiable at a point $\mathbf{a} \in \mathbb{R}^m$ then it has partial derivatives $\partial_k f_i(\mathbf{a})$ with respect to each variable x_j at that point; but it is not true conversely that if the coordinate functions of a mapping $\mathbf{f}$ have partial derivatives at a point $\mathbf{a}$ with respect to each variable x_j then $\mathbf{f}$ is a differentiable mapping. For example the mapping $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ defined by

$$f(\mathbf{x}) = \begin{cases} \dfrac{x_1 x_2}{x_1^2 + x_2^2} & \text{if } \mathbf{x} \ne \mathbf{0}, \\ \\ 0 & \text{if } \mathbf{x} = \mathbf{0} \end{cases}$$

vanishes identically in the variable x_2 if $x_1 = 0$, so $\partial_2 f(0,0) = 0$, and similarly $\partial_1 f(0,0) = 0$. This function is not even continuous at the origin, since for instance it takes the value $\frac{1}{2}$ whenever $x_1 = x_2$ except at the origin where it takes the value 0; hence it is not differentiable at the origin. However if the partial derivatives of a mapping not only exist but also are continuous then the mapping is differentiable.

diffeach

Theorem 2.2 *If the partial derivatives of a mapping $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^n$ exist at all points near $\mathbf{a}$ and are continuous at the point $\mathbf{a}$ then the mapping $\mathbf{f}$ is differentiable at the point $\mathbf{a}$.*

Proof: In view of Theorem diffloc 2.1 it is enough to prove this for the special case that $n = 1$, in which case the mapping $f : \mathbb{R}^m \to \mathbb{R}$ is just a real valued function; and for convenience only the case $m = 2$ will be demonstrated in detail, since it is easier to follow the proof in the simpler case and all the essential ideas are present. Assume that the partial derivatives $\partial_k f_i(\mathbf{x})$ exist for all points $\mathbf{x}$ near $\mathbf{a}$ and are continuous at $\mathbf{a} = \{a_j\}$, and consider a fixed vector $\mathbf{h} = \{h_j\}$. When one of the variables is held fixed and f is viewed as a function of the remaining variable it is a differentiable function of a single variable. The mean value theorem for functions of a single variable asserts that if $f(x)$ is continuous in a closed interval $[a,b]$ and is differentiable at each point of the open interval (a,b) then $f(b)-f(a) = f'(c)(b-a)$ for some point $c \in (a,b)$; this can be applied to the function $f(x_1, a_2)$ of the single variable x_1 in the interval between a_1 and a_1+h_1, and to the function $f(a_1+h_1, x_2)$ of the single variable x_2 in the interval between a_2 and $a_2 + h_2$ if $\|\mathbf{h}\|$ is sufficiently small; as a consequence there exist values α_1 between a_1 and $a_1 + h_1$ and α_2 between a_2 and $a_2 + h_2$ such that

$$f(a_1 + h_1, a_2) - f(a_1, a_2) = h_1 \ \partial_1 f(\alpha_1, a_2) \quad \text{and}$$

$$f(a_1 + h_1, a_2 + h_2) - f(a_1 + h_1, a_2) = h_2 \ \partial_2 f(a_1 + h_1, \alpha_2).$$

Then

$$\begin{aligned}
f(\mathbf{a}+\mathbf{h})-f(\mathbf{a}) &= f(a_1+h_1,a_2+h_2)-f(a_1,a_2) = \\
&= \Big(f(a_1+h_1,a_2+h_2)-f(a_1+h_1,a_2)\Big)+\Big(f(a_1+h_1,a_2)-f(a_1,a_2)\Big) \\
&= h_2\ \partial_2 f(a_1+h_1,\alpha_2)+h_1\ \partial_1 f(\alpha_1,\alpha_2) \\
&= h_2\ \partial_2 f(a_1,a_2)+h_1\ \partial_1 f(a_1,a_2)+\epsilon(\mathbf{h})
\end{aligned}$$

where

$$\epsilon(\mathbf{h}) = h_2\ \Big(\partial_2 f(a_1+h_1,\alpha_2)-\partial_2 f(a_1,a_2)\Big)+h_1\ \Big(\partial_1 f(\alpha_1,a_2)-\partial_1 f(a_1,a_2)\Big).$$

By the triangle inequality

$$\frac{|\epsilon(\mathbf{h})|}{\|h\|_\infty} \le \Big|\partial_2 f(a_1+h_1,\alpha_2)-\partial_2 f(a_1,a_2)\Big| + \Big|\partial_1 f(\alpha_1,a_2)-\partial_1 f(a_1,a_2)\Big|$$

since $\frac{|h_1|}{\|\mathbf{h}\|_\infty} \le 1$ and $\frac{|h_2|}{\|\mathbf{h}\|_\infty} \le 1$; and since the partial derivatives are assumed to be continuous at the point $\mathbf{a}$ it follows that $\lim_{\mathbf{h}\to 0}\frac{|\epsilon(\mathbf{h})|}{\|\mathbf{h}\|_\infty} = 0$. That shows that the mapping $f:\mathbb{R}^2 \longrightarrow \mathbb{R}$ is differentiable at the point (a_1,a_2), which suffices to conclude the proof.

If a mapping $\mathbf{f}:\mathbb{R}^m \longrightarrow \mathbb{R}^n$ is differentiable at all points of an open subset $U \subset \mathbb{R}^m$ the partial derivatives of the coordinate functions of the mapping $\mathbf{f}$ exist at each point of U, but need not be continuous functions in U. The standard example for functions of a single variable is the function

diffexnc (2.4)
$$f(x) = \begin{cases} x^2\sin\dfrac{1}{x^2} & \text{if} \quad x \ne 0, \\ \\ 0 & \text{if} \quad x \ne 0, \end{cases} \tag{2.4}$$

which is differentiable at all points $x \in \mathbb{R}$ but for which the derivative is not continuous at the point 0. Indeed if $x \ne 0$ then

$$f'(x) = 2x\sin\frac{1}{x^2} - \frac{2}{x}\cos\frac{1}{x^2},$$

which is not bounded as x tends to zero; however

$$f'(0) = \lim_{h\to 0}\frac{h^2\sin\dfrac{1}{h^2}-0}{h-0} = \lim_{h\to 0} h\sin\frac{1}{h^2} = 0$$

since $|\sin\frac{1}{h}| \le 1$ for $h \ne 0$. Thus the hypothesis in the preceding theorem is strictly stronger than just the assumption that the partial derivatives of the coordinate functions of the mapping exist. A mapping $\mathbf{f}:U \longrightarrow \mathbb{R}^n$ defined in an open subset $U \subset \mathbb{R}^m$ is said to be **continuously differentiable** or of **class** $\mathcal{C}^1$ in an open subset $U \subset \mathbb{R}^m$ if the partial derivatives of its coordinate functions exist and are continuous throughout the set U. By the preceding theorem, that is precisely equivalent to the condition that the mapping $\mathbf{f}$ is differentiable at each point $\mathbf{a} \in U$ and the matrix function $\mathbf{f}'(\mathbf{a})$ is a continuous function of the variable $\mathbf{a} \in U$. There are thus two fairly natural but not entirely equivalent notions that relate to properties of differentiability of mappings: the formal definition (diff1 2.1) that a mapping is **differentiable** in an open set and the condition $\mathcal{C}^1$ that the component functions of a mapping have continuous partial derivatives in an open set. Different applications or considerations may lead to one or the other of these two properties being the primary one considered; that is just a matter of convenience, and you may expect to find different emphases in different situations. It is important though to recognize that these two notions are not entirely equivalent to one another and to remember how they are actually related.

2.2 The Chain Rule

If $\mathbf{g} : U \longrightarrow \mathbb{R}^m$ is a mapping defined in an open neighborhood U of a point $\mathbf{a} \in \mathbb{R}^l$ and $\mathbf{f} : V \longrightarrow \mathbb{R}^n$ is a mapping defined in an open neighborhood V of the point $\mathbf{b} = \mathbf{g}(\mathbf{a}) \in \mathbb{R}^m$, where $\mathbf{g}(U) \subset V$, the composition $\boldsymbol{\phi} = \mathbf{f} \circ \mathbf{g} : U \longrightarrow \mathbb{R}^n$ is the mapping defined by $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{f}\big(\mathbf{g}(\mathbf{x})\big)$ for any $\mathbf{x} \in U$; this situation is described in the following diagram.

difffg (2.5)
$$\begin{array}{ccccccc}
\mathbb{R}^l & & \mathbb{R}^m & & \mathbb{R}^n & & \\
\bigcup & & \bigcup & & \bigcup & & \\
U & \xrightarrow{\ \mathbf{g}\ } & V & \xrightarrow{\ \mathbf{f}\ } & W & & \boldsymbol{\phi} = \mathbf{f} \circ \mathbf{g} \\
& & & & & & \\
\mathbf{a} & \xrightarrow{\ \mathbf{g}\ } & \mathbf{b} = \mathbf{g}(\mathbf{a}) & \xrightarrow{\ \mathbf{f}\ } & \mathbf{f}(\mathbf{b}) = \boldsymbol{\phi}(\mathbf{a}) & &
\end{array}$$

chainrule **Theorem 2.3** *If the mapping* $\mathbf{g}$ *is differentiable at the point* $\mathbf{a}$ *and the mapping* $\mathbf{f}$ *is differentiable at the point* $\mathbf{b} = \mathbf{g}(\mathbf{a})$ *then the composite function* $\boldsymbol{\phi} = \mathbf{f} \circ \mathbf{g}$ *is differentiable at the point* $\mathbf{a}$ *and* $\boldsymbol{\phi}'(\mathbf{a}) = \mathbf{f}'\big(\mathbf{g}(\mathbf{a})\big) \cdot \mathbf{g}'(\mathbf{a})$.

Proof: Since the mapping $\mathbf{f}$ is differentiable at the point $\mathbf{b}$

comp1 (2.6)
$$\mathbf{f}(\mathbf{b}+\mathbf{h}) = \mathbf{f}(\mathbf{b}) + \mathbf{f}'(\mathbf{b})\mathbf{h} + \boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h}) \ \text{ where } \lim_{\mathbf{h}\to 0} \frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} = 0,$$

and since the mapping $\mathbf{g}$ is differentiable at the point $\mathbf{a}$

comp2 (2.7)
$$\mathbf{g}(\mathbf{a}+\mathbf{k}) = \underbrace{\mathbf{g}(\mathbf{a})}_{\mathbf{b}} + \underbrace{\mathbf{g}'(\mathbf{a})\mathbf{k} + \boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})}_{\mathbf{h}} \ \text{ where } \lim_{\mathbf{k}\to 0} \frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} = 0.$$

Substituting (2.7) into (2.6) where $\mathbf{b} = \mathbf{g}(\mathbf{a})$ and $\mathbf{h} = \mathbf{g}'(\mathbf{a})\mathbf{k} + \boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})$ leads to the result that

$$\begin{aligned}\boldsymbol{\phi}(\mathbf{a}+\mathbf{k}) &= \mathbf{f}\big(\mathbf{g}(\mathbf{a}+\mathbf{k})\big) = \mathbf{f}\big(\mathbf{b}+\mathbf{h}\big)\\ &= \mathbf{f}(\mathbf{b}) + \mathbf{f}'(\mathbf{b})\mathbf{h} + \boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\\ &= \boldsymbol{\phi}(\mathbf{a}) + \mathbf{f}'(\mathbf{b})\Big(\mathbf{g}'(\mathbf{a})\mathbf{k} + \boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\Big) + \boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\end{aligned}$$

hence that

$$\boldsymbol{\phi}(\mathbf{a}+\mathbf{k}) = \boldsymbol{\phi}(\mathbf{a}) + \mathbf{f}'(\mathbf{b})\mathbf{g}'(\mathbf{a})\mathbf{k} + \boldsymbol{\epsilon}(\mathbf{k}) \tag{2.8}$$

where $\boldsymbol{\epsilon}(\mathbf{k}) = \mathbf{f}'(\mathbf{b})\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k}) + \boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})$. From the inequality (1.10) and the triangle inequality it follows that

$$\begin{aligned}\frac{\|\boldsymbol{\epsilon}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} &\le \frac{\|\mathbf{f}'(\mathbf{b})\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} + \frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h}))\|_\infty}{\|\mathbf{h}\|_\infty}\cdot\frac{\|\mathbf{g}'(\mathbf{a})\mathbf{k}+\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty}\\ &\le n\|\mathbf{f}'(\mathbf{b})\|_\infty\frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} + \frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h}))\|_\infty}{\|\mathbf{h}\|_\infty}\cdot\left(n\|\ \mathbf{g}'(\mathbf{a})\|_\infty + \frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty}\right).\end{aligned}$$

Since $\lim_{\mathbf{k}\to 0}\frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} = 0$ and $\lim_{\mathbf{h}\to 0}\frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_\infty}{\|\mathbf{h}\|_\infty} = 0$ while $\lim_{\mathbf{h}\to 0}\mathbf{k} = 0$ it follows from the preceding equation that $\lim_{\mathbf{k}\to 0}\frac{\|\boldsymbol{\epsilon}(\mathbf{k})\|_\infty}{\|\mathbf{k}\|_\infty} = 0$, and it then follows from (2.8) that $\mathbf{h} = \mathbf{f}\circ\mathbf{g}$ is differentiable at the point $\mathbf{a}$ and that $\boldsymbol{\phi}'(\mathbf{a}) = \mathbf{f}'\big(\mathbf{g}(\mathbf{a})\big)\cdot\mathbf{g}'(\mathbf{a})$, which concludes the proof.

For a simple application of the chain rule, since the function $g(y_1, y_2) = y_1y_2$ is differentiable at any point $y_1, y_2 \in \mathbb{R}$ and $g'(y_1, y_2) = (y_2 \quad y_1)$ it follows that for any differentiable mapping $\mathbf{f} = \{f_1, f_2\} : U \longrightarrow \mathbb{R}^2$ in an open subset $U \subset \mathbb{R}^m$ the composition $\phi = g\circ\mathbf{f} : U \longrightarrow \mathbb{R}$ is a differentiable mapping and

$$\phi'(\mathbf{x}) = g'\big(\mathbf{f}(\mathbf{x})\big)\mathbf{f}'(\mathbf{x}) = \Big(f_2(\mathbf{x}) \ \ f_1(\mathbf{x})\Big)\cdot\begin{pmatrix} f_1'(\mathbf{x})\\ f_2'(\mathbf{x})\end{pmatrix} = f_2(\mathbf{x})f_1'(\mathbf{x}) + f_1(\mathbf{x})f_2'(\mathbf{x});$$

this is just an extension of the familiar product rule for differentiating functions from the case of functions of a single variable to the case of functions of several variables, since $\phi(\mathbf{x}) = f_1(\mathbf{x})f_2(\mathbf{x})$. The entries in the matrix $\phi'(\mathbf{x})$ thus have the form

$$\partial_k\phi(\mathbf{x}) = f_2(\mathbf{x})\partial_k f_1(\mathbf{x}) + f_1(\mathbf{x})\partial_k f_2(\mathbf{x}).$$

The corresponding argument shows that the quotient f_1/f_2 of two differentiable functions is differentiable at any point $\mathbf{x}$ at which $f_2(\mathbf{x}) \ne 0$, and that its derivative has the expected form. For another application of the chain rule, if $\mathbf{f} : U \longrightarrow V$ is a one-to-one mapping between two open subsets $U, V \subset \mathbb{R}^m$ and if $\mathbf{g} : V \longrightarrow U$ is the inverse mapping then $\mathbf{g}\circ\mathbf{f} : U \longrightarrow U$ is the identity mapping $\mathbf{g}\circ\mathbf{f}(\mathbf{x}) = \mathbf{x}$, so $(\mathbf{g}\circ\mathbf{f})'(\mathbf{x}) = \mathrm{I}$ where I is the $m\times m$ identity matrix. If the mappings $\mathbf{f}$ and $\mathbf{g}$ are continuously differentiable then by the chain rule $\mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\cdot\mathbf{f}'(\mathbf{x}) = (\mathbf{g}\circ\mathbf{f})'(\mathbf{x}) = \mathrm{I}$; thus the matrix $\mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)$ is the inverse of the matrix $\mathbf{f}'(\mathbf{x})$ at each point $\mathbf{x} \in U$, so both matrices are nonsingular matrices at each point $\mathbf{x} \in U$.

An alternative notation for the chain rule is suggestive and sometimes quite useful. When mappings $\mathbf{f} : \mathbb{R}^l \longrightarrow \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ are described in terms of the coordinates $\mathbf{t} = \{t_1, \ldots, t_l\} \in \mathbb{R}^l$, $\mathbf{x} = \{x_1, \ldots, x_m\} \in \mathbb{R}^m$ and $\mathbf{y} = \{y_1, \ldots, y_l\} \in \mathbb{R}^n$, the coordinate functions of the mappings $\mathbf{f}$, $\mathbf{g}$ and $\boldsymbol{\phi}$ have the form $y_i = f_i(\mathbf{x}) = \phi_i(\mathbf{t})$ and $x_j = g_j(\mathbf{t})$. The partial derivatives are sometimes denoted by

$$(\phi')_{ik} = \partial_k \phi_i = \frac{\partial y_i}{\partial t_k}, \quad (f')_{ij} = \partial_j f_i(\mathbf{x}) = \frac{\partial y_i}{\partial x_j}, \quad (g')_{jk} = \partial_k g_j(\mathbf{t}) = \frac{\partial x_j}{\partial t_k}.$$

By the preceding theorem the derivative of the composite function $\boldsymbol{\phi} = \mathbf{f} \circ \mathbf{g}$ is the matrix product $\boldsymbol{\phi}' = \mathbf{f}'\mathbf{g}'$, which in terms of the entries of these matrices is $(\phi')_{ik} = \sum_{j=1}^n (f')_{ij} (g')_{jk}$ or equivalently $\partial_k \phi_i = \sum_{j=1}^n \partial_j f_i \cdot \partial_k g_j$; and in the alternative notation this takes the form

ch1 (2.9)
$$\frac{\partial y_i}{\partial t_k} = \sum_{j=1}^m \frac{\partial y_i}{\partial x_j} \cdot \frac{\partial x_j}{\partial t_k}.$$

This is the extension to mappings in several variables of the traditional formulation of the chain rule for functions of a single variable as the identity $\frac{dy}{dt} = \frac{dy}{dx} \cdot \frac{dx}{dt}$; this form of the chain rule is in some ways easier to remember, and with some caution, easier to use, than the version of the chain rule in the preceding theorem. It is customary however to omit any explicit mention of the points at which the derivatives are taken; so some care must be taken to remember the the derivative $\frac{\partial y_i}{\partial x_j}$ is evaluated at the point $\mathbf{x}$ while the derivatives $\frac{\partial y_i}{\partial t_k}$ and $\frac{\partial x_j}{\partial t_k}$ are evaluated at the point $\mathbf{t}$. This lack of clarity means that some caution must be taken when this notation is used.

Some care also must be taken with the chain rule in those cases where the compositions are not quite so straightforward. For instance if $\phi(x_1, x_2) = f\big(x_1, x_2, g(x_1, x_2)\big)$ for a function $f(x_1, x_2, x_3)$ of three variables and a function $g(x_1, x_2)$ of two variables, the function ϕ is really the composition $\phi = f \circ G$ of the mappings $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^1$ given by the function f and the mapping $G : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ given by

$$G(x_1, x_2) = \begin{pmatrix} x_1 \\ x_2 \\ g(x_1, x_2) \end{pmatrix},$$

so if the function g is differentiable it follows from the chain rule that

$$\begin{aligned}
\phi'(\mathbf{x}) &= f'\big(G(\mathbf{x})\big)G'(\mathbf{x}) \\
&= \begin{pmatrix} \partial_1 f\big(G(\mathbf{x})\big) & \partial_2 f\big(G(\mathbf{x})\big) & \partial_3 f\big(G(\mathbf{x})\big) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \partial_1 g(\mathbf{x}) & \partial_2 g(\mathbf{x}) \end{pmatrix} \\
&= \begin{pmatrix} \partial_1 f\big(G(\mathbf{x})\big) + \partial_3 f\big(G(\mathbf{x})\big)\partial_1 g(\mathbf{x}) & \partial_2 f\big(G(\mathbf{x})\big) + \partial_3 f(\mathbf{x})\partial_2 g(\mathbf{x}) \end{pmatrix}
\end{aligned}$$

hence the coordinate functions of the matrix $\phi'(\mathbf{x})$ are

$$\begin{aligned}
\partial_1\phi(\mathbf{x}) &= \partial_1 f\big(G(\mathbf{x})\big) + \partial_3 f\big(G(\mathbf{x})\big)\partial_1 g(\mathbf{x}), \\
\partial_2\phi(\mathbf{x}) &= \partial_2 f\big(G(\mathbf{x})\big) + \partial_3 f\big(G(\mathbf{x})\big)\partial_2 g(\mathbf{x}).
\end{aligned}$$

This amounts to calculating the partial derivative $\partial_1\phi(\mathbf{x})$ as the sum of the partial derivatives of the function $f\big(x_1, x_2, g_2(x_1, x_2)\big)$ with respect to each of its three variables, and multiplying each of these derivatives by the derivative of what is in the place of that variable with respect to the variable x_1. Some care must be taken with this calculation, though. For example, if

$$f(x_1, x_2) = \sin x_1 + x_2^3$$

then

$$\partial_1 f(x_1, x_2) = \cos x_1 \quad \text{and} \quad \partial_2 f(x_1, x_2) = 3x_2^2$$

and consequently

$$\partial_1 f(e^{x_2}, x_1x_2) = \cos e^{x_2} \quad \text{and} \quad \partial_2 f(e^{x_2}, x_1x_2) = 3(x_1x_2)^2;$$

therefore if $g(x_1, x_2) = f(e^{x_2}, x_1x_2)$ then

$$\begin{aligned}
\partial_1 g(x_1, x_2) &= \partial_1 f(e^{x_2}, x_1x_2) \cdot \frac{\partial}{\partial x_1} e^{x_2} + \partial_2 f(e^{x_2}, x_1x_2) \cdot \frac{\partial}{\partial x_1}(x_1x_2) \\
&= 0 + 3(x_1x_2)^2 \cdot x_2 = 3x_1^2x_2^3
\end{aligned}$$

and

$$\begin{aligned}
\partial_2 g(x_1, x_2) &= \partial_1 f(e^{x_2}, x_1x_2) \cdot \frac{\partial}{\partial x_2} e^{x_2} + \partial_2 f(e^{x_2}, x_1x_2) \cdot \frac{\partial}{\partial x_2}(x_1x_2) \\
&= \cos(e^{x_2}) \cdot e^{x_2} + 3(x_1x_2)^2 \cdot x_1 = e^{x_2}\cos e^{x_2} + 3x_1^3x_2^2.
\end{aligned}$$

This can be checked in this case by an alternative approach, noting that $g(x_1, x_2) = \sin e^{x_2} + (x_1x_2)^3$ and applying the chain rule for functions of a single variable. Some practice, checked by going back to the form of the chain rule give in Theorem 2.3, may prove helpful; and if there are any doubts about an application of the chain rule they can be cleared up by identifying the function as an explicit composition of mappings. It should be noted that in this case the meaning of the expression

$$\frac{\partial}{\partial x_1} f\big(x_1, x_2, x_3\big) \quad \text{where} \quad x_3 = g(x_1, x_2)$$

is not clear; it may mean either the derivative of the function f with respect to its first variable or the derivative of the composite function of the two variables x_1, x_2 with respect to the variable x_1, while $\partial_1 f(x_1, x_2, x_3)$ where $x_3 = g(x_1, x_2)$ is less ambiguous.

The chain rule also is useful in deriving information about the derivatives of functions that are defined only implicitly. For example if a function $f(x_1, x_2)$ satisfies the equation

$$f(x_1, x_2)^5 + x_1 f(x_1, x_2) + f(x_1, x_2) = 2x_1 + 3x_2$$

and the initial condition that $f(0,0) = 0$ then the values of that function are determined implicitly but not explicitly by the preceding equation. This equation is the condition that the composition of the mapping $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ defined by $F(x_1, x_2) = \big(x_1, x_2, f(x_1, x_2)\big)$ and the mapping $G : \mathbb{R}^3 \longrightarrow \mathbb{R}$ defined by $G(x_1, x_2, y) = y^5 + x_1 y + y - 2x_1 - 3x_2$ is the trivial mapping $G \circ F(x_1, x_2) = 0$, so that $(G \circ F)'(0,0) = 0$; and if the function f is differentiable it follows from the chain rule that

$$\partial_1(G \circ F) = 5f(x_1, x_2)^4 \partial_1 f(x_1, x_2) + x_1 \partial_1 f(x_1, x_2) + f(x_1, x_2) + \partial_1 f(x_1, x_2) - 2,$$

so since $\partial_1(G \circ F) = 0$ and $f(0,0) = 0$ the preceding equation reduces to $\partial_1 f(0,0) = 2$. A similar calculation yields the value of $\partial_2 f(0,0)$.

2.3 Higher Derivatives

If $f : U \longrightarrow \mathbb{R}$ is a function defined in an open set $U \subset \mathbb{R}^2$ and if the partial derivative $\partial_{j_1} f(\mathbf{x})$ exists at all points $\mathbf{x} \in U$ the function $\partial_{j_1} f(\mathbf{x})$ may itself have partial derivatives, such as $\partial_{j_2}\big(\partial_{j_1} f(\mathbf{x})\big)$, which for convenience is shortened to $\partial_{j_2}\partial_{j_1} f(\mathbf{x})$; and the process may continue, leading to $\partial_{j_3}\partial_{j_2}\partial_{j_1} f(\mathbf{x})$ and so on. The order in which successive derivatives are taken may be significant; for example, a straightforward calculation for the function

$$f(x_1, x_2) = \begin{cases} \dfrac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \text{if} \quad (x_1, x_2) \ne (0,0) \\ 0 & \text{if} \quad (x_1, x_2) = 0 \end{cases}$$

shows that $\partial_1\partial_2 f(0,0) = 1$ but $\partial_2\partial_1 f(0,0) = -1$. However for sufficiently regular functions the order of differentiation is irrelevant.

multord **Theorem 2.4** *If $f : U \longrightarrow \mathbb{R}$ is a function in open subset $U \subset \mathbb{R}^2$, if the partial derivatives $\partial_1 f(\mathbf{x}), \partial_2 f(\mathbf{x}), \partial_1\partial_2 f(\mathbf{x}), \partial_2\partial_1 f(\mathbf{x})$ exist at all points $\mathbf{x} \in U$, and if the mixed partial derivatives $\partial_1\partial_2 f(\mathbf{x}), \partial_2\partial_1 f(\mathbf{x})$ exist and are continuous at a point $\mathbf{a} \in U$, then $\partial_1\partial_2 f(\mathbf{a}) = \partial_2\partial_1 f(\mathbf{a})$.*

Proof: By an application of the mean value theorem for functions of one variable it follows that

$$\begin{aligned}\Delta &= \Big(f(a_1+h_1,a_2+h_2)-f(a_1+h_1,a_2)\Big)-\Big(f(a_1,a_2+h_2)-f(a_1,a_2)\Big)\\ &= \phi(a_1+h_1)-\phi(a_1) \qquad \text{where } \phi(x_1)=f(x_1,a_2+h_2)-f(x_1,a_2)\\ &= \phi'(\alpha_1)h_1 \qquad\qquad\quad\ \ \text{where } \alpha_1 \text{ is between } a_1 \text{ and } a_1+h_1\end{aligned}$$

if $\|\mathbf{h}\|$ is sufficiently small, and by another application of the mean value theorem for functions of one variable it further follows that

$$\begin{aligned}\phi'(\alpha_1) &= \partial_1 f(\alpha_1,a_2+h_2)-\partial_1 f(\alpha_1,a_2)\\ &= \partial_2\partial_1 f(\alpha_1,\alpha_2)h_2 \quad \text{where } \alpha_2 \text{ is between } a_2 \text{ and } a_2+h_2\end{aligned}$$

if $\|\mathbf{h}\|$ is sufficiently small; consequently

$$\Delta = \partial_2\partial_1 f(\alpha_1,\alpha_2)h_2h_1.$$

On the other hand it is possible to group the terms in the equation for Δ in another way, so that

$$\Delta = \Big(f(a_1+h_1,a_2+h_2)-f(a_1,a_2+h_2)\Big)-\Big(f(a_1+h_1,a_2+h_2)-f(a_1,a_2)\Big).$$

The same argument used for the first grouping when applied to this grouping amounts to interchanging the roles of the two variables, which leads to the result that

$$\Delta = \partial_1\partial_2 f(\beta_1,\beta_2)h_1h_2$$

for some points β_1 between a_1 and a_1+h_1 and β_2 between a_2 and a_2+h_2 if $\|\mathbf{h}\|$ is sufficiently small; the points β_1 and β_2 are not necessarily the same as the points α_1 and α_2 since they are derived by different applications of the mean value theorem in one variable. Comparing the two expressions for Δ and dividing by h_1h_2 shows that

$$\partial_2\partial_1 f(\alpha_1,\alpha_2) = \partial_1\partial_2 f(\beta_1,\beta_2).$$

Since the functions $\partial_1\partial_2 f(x_1,x_2)$ and $\partial_2\partial_1 f(x_1,x_2)$ are both continuous and α_1 and β_1 both approach a_1 as h_1 tends to 0, and correspondingly for the points α_2 and β_2, it follows in the limit that $\partial_2\partial_1 f(a_1,a_2) = \partial 1\partial_2 f(a_1,a_2)$, which suffices to conclude the proof.

The preceding result shows that so long as the mixed partial derivatives exist and are continuous functions the order in which partial derivatives are taken is immaterial; in particular that is the case for $\mathcal{C}^2$ functions, those for which all the second partial derivatives exist and are continuous. In such cases the notation can be simplified, for example by writing $\partial_1^2\partial_2 f(\mathbf{x})$ in place of $\partial_1\partial_2\partial_1 f(\mathbf{x})$ or $\partial_2\partial_1\partial_1 f(\mathbf{x})$, or by writing $\partial_{j_1j_2} f(\mathbf{x})$ in place of $\partial_{j_1}\partial_{j_2} f(\mathbf{x})$. It is sometimes convenient to use the multi-index notation, in which $\partial^I f(\mathbf{x})$ where $I = (3,2,1)$ stands for $\partial_1^3\partial_2^2\partial_3 f(\mathbf{x})$ for instance. Another notation frequently used is

$$\frac{\partial^3 f(\mathbf{x})}{\partial x_1^2 \partial x_2} = \partial_1^2\partial_2 f(\mathbf{x}).$$

The definition (2.1) of differentiability involved an approximation of a function by a polynomial of degree 1 for sufficiently small values of the the auxiliary variable $\mathbf{h}$. The existence of higher order derivatives can be interpreted correspondingly as involving an approximation of a function by polynomials of higher degree for sufficiently small values of the the auxiliary variable $\mathbf{h}$. Since this topic is sometimes omitted in treatments of functions of a single variable the discussion here will begin with the results for functions of one variable, from which the general result follows readily.

taylor1 **Theorem 2.5 (Taylor expansion in one variable)** *If $f : U \longrightarrow \mathbb{R}^1$ has derivatives up to order $k+1$ in an open neighborhood $U \subset \mathbb{R}^1$ of a point $a \in \mathbb{R}^1$ then for any $h \in \mathbb{R}^1$ sufficiently small*

taylorexp
$$f(a+h) = f(a) + f'(a)h + \frac{1}{2!}f''(a)h^2 + \frac{1}{3!}f'''(a)h^3 + \dots \\ \dots + \frac{1}{k!}f^{(k)}(a)h^k + \frac{1}{(k+1)!}f^{(k+1)}(\alpha)h^{k+1} \tag{2.10}$$

where α is between a and $a+h$.

Proof: For a fixed $h \in \mathbb{R}$ sufficiently small that the closed interval from a to $a+h$ is contained in U let

taylorq1
$$R(x) = f(a+x) - f(a) - f'(a)x - \frac{1}{2!}f''(a)x^2 - \frac{1}{3!}f'''(a)x^3 \dots \\ \dots - \frac{1}{k!}f^{(k)}(a)x^k - \frac{1}{(k+1)!}cx^{k+1} \tag{2.11}$$

where $c \in \mathbb{R}$ is chosen so that $R(h) = 0$. Clearly $R(0) = 0$ as well, so it follows from an application of the mean value theorem for functions of a single variable that $R'(\alpha_1) = 0$ for some value α_1 between 0 and h. Note that

talorq2
$$R'(x) = f'(a+x) - f'(a) - f''(a)x - \frac{1}{2!}f'''(a)x^2 - \\ \dots - \frac{1}{(k-1)!}f^{(k-1)}(a)x^{k-1} - \frac{1}{k!}cx^k, \tag{2.12}$$

which is much the same as (2.11) but for the index $k-1$ in place of the index k. In particular $R'(0) = 0$, and since $R'(\alpha_1) = 0$ it follows from another application of the mean value theorem for functions of a single variable that $R''(\alpha_2) = 0$ for some α_2 between 0 and α_1. The process continues, with further differentiation of the expression (2.11), yielding a sequence of points $\{\alpha_1, \alpha_2, \dots, \alpha_{k+1}\}$ where

α_{i+1} is between 0 and α_i, hence is between 0 and $\mathbf{h}$, and $R^{(i)}(\alpha_i) = 0$. Finally for the case $i = k+1$ it follows that

$$R^{(k+1)}(x) = f^{(k+1)}(a+x) - c,$$

and since $R^{(k+1)}(\alpha_{k+1}) = 0$ it follows that $c = f^{(k+1)}(a + \alpha_{k+1})$. That suffices for the proof.

The corresponding result in several variables can be deduced from the result in a single variable by an application of the chain rule.

taylor2 **Theorem 2.6 (Taylor expansion in several variables)** *If $f : U \longrightarrow \mathbb{R}$ has continuous partial derivatives up to order $k+1$ in an open neighborhood $U \subset \mathbb{R}^m$ of a point $\mathbf{a} \in \mathbb{R}^m$ then for any $\mathbf{h} = \{h_j\} \in \mathbb{R}^m$ sufficiently small*

taylormult

$$\begin{aligned} f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) &+ \sum_{j=1}^{m} \partial_j f(\mathbf{a}) h_j + \frac{1}{2!} \sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\mathbf{a}) h_{j_1} h_{j_2} + \cdots \\ &\cdots + \frac{1}{k!} \sum_{j_1,\ldots,j_k=1}^{m} \partial_{j_1 \ldots j_k} f(\mathbf{a}) h_{j_1} \ldots h_{j_k} \\ &+ \frac{1}{(k+1)!} \sum_{j_1,\ldots,j_{k+1}=1}^{m} \partial_{j_1 \ldots j_{k+1}} f(\boldsymbol{\alpha}) h_{j_1} \ldots h_{j_{k+1}} \end{aligned} \tag{2.13}$$

where $\boldsymbol{\alpha}$ is between $\mathbf{a}$ and $\mathbf{a}+\mathbf{h}$ on the line segment connecting them.

Proof: Let $\phi(t) = \mathbf{a} + t\mathbf{h}$ for any $t \in \mathbb{R}$ and consider the function $g(t) = f(\phi(t))$, for which $g(0) = f\big(\phi(0)\big) = f(\mathbf{a})$ and $g(1) = f\big(\phi(1)\big) = f(\mathbf{a}+\mathbf{h})$. By the Taylor expansion (2.10) of the function $g(t)$

tayloraux

$$g(1) = g(0) + g'(0) + \frac{1}{2!} g''(0) + \ldots + \frac{1}{k!} g^{(k)}(0) + \frac{1}{(k+1)!} g^{(k+1)}(\tau) \tag{2.14}$$

for some $\tau \in (0,1)$. By repeated applications of the chain rule

$$\begin{aligned} g'(t) &= \sum_{j_1=1}^{m} \partial_{j_1} f(\mathbf{a}+t\mathbf{h}) h_{j_1}, \\ g''(t) &= \sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\mathbf{a}+t\mathbf{h}) h_{j_1} h_{j_2} \\ &\text{and in general} \\ g^{(\nu)}(t) &= \sum_{j_1,j_2,\ldots,j_{nu}=1}^{m} \partial_{j_1 j_2 \cdots j_\nu} f(\mathbf{a}+t\mathbf{h}) h_{j_1} h_{j_2} \cdots h_{j_\nu}; \end{aligned}$$

substituting these values into (2.14) for $t = 0$ yields (2.13) for $\boldsymbol{\alpha} = \mathbf{a} + \tau\mathbf{h}$ and thereby concludes the proof.

As a consequence, if a function $f(\mathbf{x})$ is twice differentiable near a point $\mathbf{a} \in \mathbb{R}^m$ and if $\mathbf{h} \in \mathbb{R}^m$ is sufficiently small then

taylorex
$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \sum_{j=1}^{m} \partial_j f(\mathbf{a}) h_j + \tfrac{1}{2} \textstyle\sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\boldsymbol{\alpha}) h_{j_1} h_{j_2} \tag{2.15}$$

for some point $\boldsymbol{\alpha}$ between $\mathbf{a}$ and $\mathbf{h}$ on the line segment connecting them. The expression $\sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\boldsymbol{\alpha}) h_{j_1} h_{j_2}$ is a quadratic form in the variables h_j defined by the matrix $\partial_{j_1}\partial_{j_2} f(\boldsymbol{\alpha})$; this matrix, which by Theorem 2.4 is a symmetric matrix, is called the **Hessian** of the function $f(\mathbf{x})$ at the point $\mathbf{a}$. The last term in the Taylor expansions (2.10) and (2.13) is called Lagrange's form of the remainder; there are various alternative forms for the last term that are often used, but they will not be needed in the discussion here.

2.4 Functions

fcurv

A common application of differentiation in the study of functions of several variables is to the determination of extremal values of a differentiable function f defined in an open set $U \subset \mathbb{R}^m$. A point $\mathbf{a} \in U$ is a **local maximum** (plural: local maxima) of f if $f(\mathbf{a}) \geq f(\mathbf{x})$ for all points $\mathbf{x}$ near $\mathbf{a}$, and is a **local minimum** (plural: local minima) of the function f if $f(\mathbf{a}) \leq f(\mathbf{x})$ for all points $\mathbf{x}$ near $\mathbf{a}$; in either case the point $\mathbf{a}$ is called a **local extremum** (plural: local extrema) of the function f. It is familiar that local extrema of differentiable functions $f(x)$ of a single variable occur among those points a at which $f'(a) = 0$. Local extrema of a differentiable function f of several variables occur similarly among those points $\mathbf{a}$ at which $f'(\mathbf{a}) = 0$, which points are called the **critical points** of the function f. That is quite clear, since if $\mathbf{a}$ is a local extremum then it is a local extremum of the function $f(\mathbf{x})$ as a function of any one coordinate x_j when the remaining coordinates are held fixed, so by the one-variable result $\partial_j f(\mathbf{a}) = 0$; and since that is the case for all indices j it follows that $\mathbf{f}'(\mathbf{a}) = \{\partial_j f_i(\mathbf{a})\} = 0$ hence the point $\mathbf{a}$ is a critical point. Not all critical points are local extrema though; the origin is a critical point of the function $f(x) = x^3$ of a single variable but is neither a local maximum nor a local minimum. The process of finding maxima or minima of functions defined in an open subset $U \subset \mathbb{R}^m$ begins with the identification of the critical points of the function, which are candidates for at least local extrema. In many cases it is fairly easy to tell by close inspection whether a local extremum is a local maximum or local minimum; and if that fails there is an analogue for functions of several variables of the second-derivative test familiar from the examination of extrema of functions of a single variable. If f is a twice continuously differentiable function with a critical point $\mathbf{a} \in \mathbb{R}^m$ then since $f'(\mathbf{a}) = 0$ the Taylor expansion(2.15) takes the form

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\boldsymbol{\alpha}) h_{j_1} h_{j_2}$$

where $\boldsymbol{\alpha}$ is a point between $\mathbf{a}$ and $\mathbf{a}+\mathbf{h}$ on the line joining these two points. If the Hessian $\partial_{j_1 j_2} f(\mathbf{a})$ is positive definite then by continuity the Hessian $\partial_{j_1 j_2} f(\boldsymbol{\alpha})$ is positive definite for all points $\boldsymbol{\alpha} \in U$ for an open neighborhood U of $\mathbf{a}$, so that

$$\sum_{j_1,j_2=1}^{m} \partial_{j_1 j_2} f(\boldsymbol{\alpha}) h_{j_1} h_{j_2} \ge 0 \quad \text{for all} \quad \boldsymbol{\alpha} \in U;$$

consequently $f(\mathbf{a}+\mathbf{h}) \ge f(\mathbf{a})$ so $\mathbf{a}$ is a local minimum of the function f. Similarly if the Hessian $\partial_{j_1 j_2} f(\mathbf{a})$ is negative definite then $\mathbf{a}$ is a local maximum. If the Hessian is neither positive definite nor negative definite the second derivative test generally provides no information; but if the Hessian has some strictly positive and some strictly negative eigenvalues then $\mathbf{a}$ is neither a local maximum nor local minimum, but is a point called a saddle point of the function. Geometrically the graph of a function of two variables near such a point looks something like an ordinary saddle; it is concave upwards in one direction and concave downwards in another direction. The difficulty lies in determining whether the Hessian is positive or negative definite. The simplest general result is that a symmetric matrix is positive definite if and only if the principal minors (the determinants of all submatrices formed from the first k rows and columns for $1 \le k \le n$) are all strictly positive, and the matrix is negative definite if and only if the principal minors alternate in sign beginning with a negative sign, so the 1×1 minor is negative, the 2×2 minor is positive, and so on. For the case of functions in $\mathbb{R}^m$ for sufficiently small m this is fairly easy to apply: for example if $m = 2$ the principal minors of a matrix $M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}$ are m_{11} and $m_{11}m_{22} - m_{12}m_{21}$, so it is fairly easy to see whether the critical point is a local maximum or local minimum, although it does involve calculating some second derivatives. It is sometimes tempting to think that a function f has a local minimum at $\mathbf{a} \in \mathbb{R}^m$ if $\partial_j^2 f(\mathbf{a}) \ge 0$ for $1 \le j \le m$; but that is not necessarily the case, since for example the matrix $\left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ has strictly positive diagonal entries but is not positive definite since the determinants of the two principal minors are 1 and -3.

As an example, consider the function

$$f(x_1, x_2) = x_1^4 x_2 + x_1 x_2^4 - x_1 x_2$$

in the plane $\mathbb{R}^2$. Since

$$\begin{aligned} f'(x_1, x_2) &= \big(4x_1^3 x_2 + x_2^4 - x_2, \quad x_1^4 + 4x_1 x_2^3 - x_1\big) \\ &= \big((4x_1^3 + x_2^3 - 1)x_2, \quad (x_1^3 + 4x_2^3 - 1)x_1\big) \end{aligned}$$

it is a straightforward calculation to see that there are just 4 critical points, points at which $f'(\mathbf{x}) = 0$, the points

$$(0,0),\ (1,0),\ (0,1)\ ,\left(\frac{1}{\sqrt[3]{5}}, \frac{1}{\sqrt[3]{5}}\right).$$

The function $f(x_1, x_2) = x_1 x_2(x_1^3 + x_2^3 - 1)$ is 0 at the origin $(0,0)$ but is clearly negative along the line $x_1 = x_2$ and positive along the line $x_1 = -x_2$, near the origin so the origin is not an extremum. It is possible to examine the other 3 critical points similarly, although perhaps with a bit more trouble; but in this case it is also easy to use the second derivative test. Since

$$\partial_1^2 f(\mathbf{x}) = 12x_1^2x_2,\ \partial_1\partial_2 f(\mathbf{x}) = \partial_1\partial_2 f(\mathbf{x}) = 4x_1^3 + 4x_2^3 - 1,\ \partial_2^2 f(\mathbf{x}) = 12x_1x_2^2$$

the Hessian matrix of the function $f(\mathbf{x})$, the matrix

$$H(\mathbf{x}) = \begin{pmatrix} \partial_1^2 f(\mathbf{x}) & \partial_1\partial_2 f(\mathbf{x}) \\ \partial_2\partial_1 f(\mathbf{x}) & \partial_2^2 f(\mathbf{x}) \end{pmatrix},$$

has the following values at the 4 critical points:

$$H(0,0) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}, \qquad H(1,0) = \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix},$$
$$H(0,1) = \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix}, \qquad H\left(\tfrac{1}{\sqrt[3]{5}}, \tfrac{1}{\sqrt[3]{5}}\right) = \tfrac{3}{5}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}.$$

The matrix $H(0,0)$ is neither positive nor negative definite, since the determinants of the principal minors are 0 and -1; indeed the usual calculation shows that its eigenvalues of the matrix $H(0,0)$ are ± 1, so the point $(0,0)$ is a saddle point. Similarly the two points $(1,0)$ and $(0,1)$ also are saddle points. However the determinants of the principal minors of the matrix $H(1,1)$ are $\frac{12}{5}$ and 9 so the matrix $H\left(\frac{1}{\sqrt[3]{5}}, \frac{1}{\sqrt[3]{5}}\right) = \frac{3}{5}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$ is positive definite and consequently the point $\left(\frac{1}{\sqrt[3]{5}}, \frac{1}{\sqrt[3]{5}}\right)$ is a local minimum of the function $f(\mathbf{x})$. It is clear that it is not a global minimum, since for instance $f(x,x) = 2x^5 - x^2$ takes arbitrarily large negative values.

There is a notational convention that is commonly used in considering the derivatives of functions and is quite convenient. The derivative $f'(\mathbf{x})$ of a function is a $1 \times m$ matrix, so it is not a vector in the sense used for points $\mathbf{x} \in \mathbb{R}^m$, which are always viewed as column vectors; but the transpose ${}^t f'(\mathbf{x})$ is a column vector, so it can be viewed as an ordinary vector in $\mathbb{R}^m$. To avoid confusion and keep these distinctions clear, the transpose of the derivative is often denoted by $\nabla f(\mathbf{x})$, although an earlier but still used alternative notation is $\mathbf{grad} f(\mathbf{x})$; and this vector is called the **gradient** of the function $f(\mathbf{x})$. Thus for a function $f(x_1, x_2)$ of two variables

$$f'(\mathbf{x}) = \begin{pmatrix} \partial_1 f(\mathbf{x}) & \partial_2 f(\mathbf{x}) \end{pmatrix} \quad \text{and} \quad \nabla f(\mathbf{x}) = \mathbf{grad} f(\mathbf{x}) = \begin{pmatrix} \partial_1 f(\mathbf{x}) \\ \\ \partial_2 f(\mathbf{x} \end{pmatrix}.$$

Although this is really a rather trivial point, the convention that the gradient of a function defined in a subset of a vector space $\mathbb{R}^m$ is a vector in the same sense as all other vectors in $\mathbb{R}^m$ is quite standard and quite useful; $\nabla f(\mathbf{x})$ is a vector that can be used in the same context as the vector $\mathbf{x}$, either in the addition of vectors or in the dot or inner product of two vectors. For example, a straight line through a point $\mathbf{a} \in \mathbb{R}^m$ in the direction of a unit vector $\mathbf{u}$ can be described parametrically as the set of points $\mathbf{x} = \boldsymbol{\phi}(t)$ for $t \in \mathbb{R}$, where $\boldsymbol{\phi} : \mathbb{R} \longrightarrow \mathbb{R}^m$ is the mapping $\boldsymbol{\phi}(t) = \mathbf{a} + t\mathbf{u}$. If $f : U \longrightarrow \mathbb{R}$ is a differentiable function in an open set $U \subset \mathbb{R}^m$ containing the point $\mathbf{a}$ the restriction of f to this straight line can be viewed as a function $(f \circ \boldsymbol{\phi})(t) = f(\mathbf{a} + t\mathbf{u})$ of the parameter $t \in \mathbb{R}$ near the origin. The derivative of this restriction is called the **directional derivative** of the function f at the point $\mathbf{a}$ in the direction of the unit vector $\mathbf{u}$, and is denoted by $\partial_{\mathbf{u}} f(\mathbf{a})$. It follows from the chain rule that $\partial_{\mathbf{u}} f(\mathbf{a}) = f'(\mathbf{a})\boldsymbol{\phi}'(0)$; the coordinate functions of the mapping $\boldsymbol{\phi}(t)$ are $\phi_j(t) = a_j + u_j t$ so the derivative $\boldsymbol{\phi}'(0)$ is just the $m \times 1$ matrix or vector $\boldsymbol{\phi}'(0) = \{u_j\}$, hence

der1

$$\partial_{\mathbf{u}} f(\mathbf{a}) = \sum_{j=1}^{m} \partial_j f(\mathbf{a}) u_j = \nabla f(\mathbf{a}) \cdot \mathbf{u}. \tag{2.16}$$

Thus the directional derivative of the function f in the direction of a unit vector $\mathbf{u}$ is the dot product of the gradient of the function f at the point $\mathbf{a}$ with the unit vector $\mathbf{u}$. The angle θ between the vectors $\nabla f(\mathbf{a})$ and u is determined by

nabdir

$$\cos\theta = \frac{\nabla f(\mathbf{a}) \cdot \mathbf{u}}{\|\nabla f(\mathbf{a})\|_2} = \frac{\partial_u f(\mathbf{a})}{\|\nabla f(\mathbf{a})\|_2} \tag{2.17}$$

since $\mathbf{u}$ is a unit vector; so it follows that the directional derivative is $\partial_{\mathbf{u}} f(\mathbf{a}) = \|\nabla f(\mathbf{a})\|_2 \cos\theta$, the length of the projection of the gradient vector $\nabla f(\mathbf{a})$ in the direction of the unit vector $\mathbf{u}$. Thus the maximal directional derivative of the function $f(\mathbf{x})$ at the point $\mathbf{a}$ is in the direction of the vector $\nabla f(\mathbf{a})$ and is equal to $\|\nabla f(\mathbf{a})\|_2$, while the minimal directional derivative is in the direction of the vector $-\nabla f(\mathbf{a})$ and is equal to $-\|\nabla f(\mathbf{a})\|_2$; and $\partial_u f(\mathbf{a}) = 0$ for any vector $\mathbf{u}$ orthogonal to the gradient $\nabla f(\mathbf{a})$. The gradient $\nabla f(\mathbf{a})$ of a function f at a point $\mathbf{a}$ is a vector in the direction in which f is increasing most rapidly; and the length $\|\nabla f(\mathbf{a})\|$ of the gradient vector is the maximum rate of increase of the function f.

The question whether a vector is to be viewed as a row vector or a column can be avoided merely by writing a vector as an explicit linear combination of basis vectors in the vector space $\mathbb{R}^m$. If $f(\mathbf{x}) = x_j$ then

$$f'(\mathbf{x}) = \big(0, \quad \cdots, \quad 0, \quad 1, \quad 0, \quad \cdots, \quad 0\big),$$

where the entry 1 is in column j; thus $f'(\mathbf{x})$ is independent of $\mathbf{x}$ and actually is one of the standard basis vectors for the vector space $\mathbb{R}^m$. It is tempting to avoid introducing a separate notation and to denote this function just by x_j. The standard notation for its derivative then would be x_j', which is somewhat confusing since x_j' commonly is used to denote another set of variables in $\mathbb{R}^n$; it is clearer and rather more customary to denote the derivative of the function x_j by dx_j, so that

dffq1

$$dx_j = \big(0, \quad \cdots, \quad 0, \quad 1, \quad 0, \quad \cdots, \quad 0\big), \tag{2.18}$$

with the entry 1 in column j and all other entries 0. With this notation the vector $f'(\mathbf{x})$ for an arbitrary differentiable function f can be denoted correspondingly by $df(\mathbf{x})$ and written in terms of the basis (2.18) as

dffq2 (2.19)
$$df(\mathbf{x}) = \sum_{j=1}^{m} \partial_j f(\mathbf{x})\, dx_j.$$

This form for the derivative $f'(\mathbf{x})$ is called a **differential form**, or to be more explicit, a differential form of degree 1. More generally a mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^m$ when viewed as associating to a point $\mathbf{x} \in \mathbb{R}^m$ a vector also of dimension m is called a **vector field** on the subset $U \subset \mathbb{R}^m$; this is familiar from physics, for the electric, magnetic and gravitational fields. A vector field $\mathbf{f}$ in an open subset $U \subset \mathbb{R}^m$ with the coordinate functions $f_j(\mathbf{x})$ also can be written as a linear combination of the standard basis vectors in $\mathbb{R}^n$ and consequently as a differential form of degree 1, explicitly

dffq3 (2.20)
$$\omega_{\mathbf{f}}(\mathbf{x}) = \sum_{j=1}^{m} f_j(\mathbf{x})\, dx_j.$$

The discussion of differential forms will be taken up again in Chapter 5.

The condition that a function f is differentiable at a point $\mathbf{a}$ can be viewed as the condition that f can be approximated near the point $\mathbf{a}$ by an affine function, a polynomial of degree 1 in the variables in $\mathbb{R}^m$, and that the error in this approximation is fairly small; this interpretation is often quite useful in practice. It is often expressed as the condition that for small changes $\Delta\mathbf{x}$ in the coordinates of a point in $\mathbb{R}^m$ the change Δf in the value of the function is approximately a linear function of $\Delta\mathbf{x}$; for (2.1) can be written

difffc4 (2.21)
$$\Delta f(\mathbf{x}) = f(\mathbf{x} + \Delta\mathbf{x}) - f(\mathbf{x}) = f'(\mathbf{x})\Delta\mathbf{x} + \epsilon(\Delta\mathbf{x}),$$

so the change $\Delta f(\mathbf{x})$ in the value of the function $f(\mathbf{x})$ is approximately equal to the linear function $f'(\mathbf{x})\Delta\mathbf{x}$ of the change $\Delta\mathbf{x}$ in the variable $\mathbf{x}$, and the error $\epsilon(\Delta\mathbf{x})$ is much smaller than the change $\Delta\mathbf{x}$ in the variable since

$$\lim_{\Delta\mathbf{x}\to 0} \epsilon(\Delta\mathbf{x})/\|\Delta\mathbf{x}\| = 0.$$

If the function $f(\mathbf{x})$ is $\mathcal{C}^2$ in an open neighborhood U of the point $\mathbf{x}$ and if $M = \sup_{\mathbf{t}\in U} \|\partial_{j_1 j_2} f(\mathbf{t})\|_\infty$ is a bound on the Hessian of the function $f(\mathbf{t})$ for $\mathbf{t} \in U$ then

$$\left| \sum_{j_1, j_2 = 1}^{m} \partial_{j_1 j_2} f(\mathbf{t}) h_{j_1} h_{j_2} \right| \le M m^2 \|h\|_\infty^2$$

for all points $\mathbf{t} \in U$, and it follows from (2.15) that

difffc5 (2.22)
$$\Big|\epsilon(\Delta\mathbf{x})\Big| \le M m^2 \|\Delta\mathbf{x}\|_\infty^2;$$

this is a rough but sometimes useful estimate for the size of the error term in(diff1 2.1). In practice it is often convenient to have a quick approximation for the values of a function f in a neighborhood of a point when the value of the function at that point is known. Another of the standard approximations in one variable, the mean value theorem, does extend directly to functions of several variables, but extends only as an inequality to mappings from $\mathbb{R}^m$ to $\mathbb{R}^n$.

mvth **Theorem 2.7 (The Mean Value Theorem)** *If $f : U \longrightarrow \mathbb{R}$ is a differentiable function in an open set $U \subset \mathbb{R}^m$ and if two points $\mathbf{a}, \mathbf{b}$ and the line λ joining them lie in U then*

mvthq1
$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a}) \tag{2.23}$$

for some point $\mathbf{c}$ between $\mathbf{a}$ and $\mathbf{b}$ on the line λ joining these two points.

Proof: The function $g(t) = f\big(\mathbf{a} + t(\mathbf{b} - \mathbf{a})\big))$ is a differentiable function of the variable t in an open neighborhood of the interval $[0, 1]$, and by the mean value theorem for functions of one variable $g(1) - g(0) = g'(\tau)$ for a point $\tau \in (0, 1)$. Since $g(1) = f(\mathbf{b})$ and $g(0) = f(\mathbf{a})$, while by the chain rule

$$g'(t) = \sum_{j=1}^{m} \partial_j f\big(\mathbf{a} + t(\mathbf{b} - \mathbf{a})\big)(b_j - a_j),$$

it follows that

$$\begin{aligned} f(\mathbf{b}) - f(\mathbf{a}) = g(1) - g(0) = g'(\tau) &= \sum_{j=1}^{m} \partial_j f\big(\mathbf{a} + \tau(\mathbf{b} - \mathbf{a})\big)(b_j - a_j) \\ &= \sum_{j=1}^{m} \partial_j f(\mathbf{c})(b_j - a_j) = \nabla f(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a}) \end{aligned}$$

where $\mathbf{c} = \mathbf{a} + \tau(\mathbf{a} - \mathbf{b})$, and that suffices for the proof.

Although the preceding theorem can be applied to each coordinate function of the mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ for $n > 1$, the points at which the derivatives of the different coordinate functions are evaluated may be different. However for some purposes an estimate is useful enough and that problem can be avoided.

mvthcr **Theorem 2.8 (The Mean Value Inequality)** *If $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ is a differentiable mapping in an open set $U \subset \mathbb{R}^m$ and if two points $\mathbf{a}, \mathbf{b}$ and the line λ joining them lie in U then*

mvthcrq1
$$\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_\infty \le m\sqrt{n}\|\mathbf{f}'(\mathbf{c})\|_\infty \|\mathbf{b} - \mathbf{a}\|_\infty \tag{2.24}$$

for some point $\mathbf{c}$ between $\mathbf{a}$ and $\mathbf{b}$ on the line λ.

Proof: For any vector $\mathbf{u} \in \mathbb{R}^n$ for which $\|\mathbf{u}\|_2 = 1$ the dot product $f_{\mathbf{u}}(\mathbf{x}) = \mathbf{u} \cdot \mathbf{f}(\mathbf{x})$ is a real-valued function to which the Mean Value Theorem can be applied, so

$$\begin{aligned}
f_{\mathbf{u}}(\mathbf{b}) - f_{\mathbf{u}}(\mathbf{a}) &= \nabla f_{\mathbf{u}}(\mathbf{c}) \cdot (\mathbf{b} - \mathbf{a}) = \sum_{j=1}^{m} \partial_j(\mathbf{u} \cdot \mathbf{f})(\mathbf{c})\ (b_j - a_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} u_i \partial_j f_i(\mathbf{c})(b_j - a_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} u_i \cdot f'(\mathbf{c})_{ij}(b_j - a_j) \\
&= \mathbf{u} \cdot \Big(\mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a})\Big)
\end{aligned}$$

for some point $\mathbf{c}$ between $\mathbf{a}$ and $\mathbf{b}$ on the line λ joining these two points. If the unit vector $\mathbf{u}$ is chosen to lie in the direction of the vector $\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})$ then

$$\begin{aligned}
\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_2 = |f_{\mathbf{u}}(\mathbf{b}) - f_{\mathbf{u}}(\mathbf{a})| &= \Big|\mathbf{u} \cdot \Big(\mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a})\Big)\Big| \\
&\le \|\mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a})\|_2
\end{aligned}$$

normcmp natnorm2
by the Cauchy-Schwarz inequality, so by (1.7) and (1.10)

$$\begin{aligned}
\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_\infty &\le \|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_2 \le \|\mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a})\|_2 \\
&\le \sqrt{n}\|\mathbf{f}'(\mathbf{c})(\mathbf{b} - \mathbf{a})\|_\infty \le m\sqrt{n}\|\mathbf{f}'(\mathbf{c})\|_\infty \|\mathbf{b} - \mathbf{a}\|_\infty,
\end{aligned}$$

which suffices for the proof.

mvthcrx **Theorem 2.9** *If $\mathbf{f} : \Delta \longrightarrow \mathbb{R}^n$ is a differentiable mapping in a cell $\Delta \subset \mathbb{R}^m$ such that $\|\mathbf{f}'(\mathbf{x})\|_\infty \le M$ for all $\mathbf{x} \in \Delta$ then the mapping $\mathbf{f}$ is uniformly continuous in Δ.*

Proof: If $\|\mathbf{f}'(\mathbf{x})\|_\infty \le M$ for all $\mathbf{x} \in \Delta$ then for any two points $\mathbf{x}, \mathbf{y} \in \Delta$ the line joining them is contained in Δ so it follows from the Mean Value Inequality that

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\|_\infty \le M\sqrt{n}\|\mathbf{x} - \mathbf{y}\|_\infty,$$

which shows that the mapping $\mathbf{f}$ is uniformly continuous in Δ and thereby concludes the proof.

PROBLEMS: GROUP 1

(1) Find the derivatives of the following mappings:

(i) $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ where $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_1 + x_2 \\ x_2^3 + x_3^2 \\ x_1 x_3 \end{pmatrix}$

(ii) $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^1$ where $f(\mathbf{x}) = \sin(x_1^2 + x_2)$

(iii) $f : \mathbb{R}^1 \longrightarrow \mathbb{R}^2$ where $f(x) = \begin{pmatrix} \cos(\sin x) \\ e^{\sin x} \end{pmatrix}$

(2) Find the following iterated partial derivatives:

(i) $\partial_1 \partial_2 \sin(x_1 x_2)$, (ii) $\partial_1^2 \sin(x_1 x_2)$, (iii) $\partial_1 \partial_2 \partial_3\, e^{x_1+x_2+x_3}$.

(3) Find $\dfrac{\partial^2}{\partial r^2} f(r\cos\theta,\ r\sin\theta)$ where $f(x_1, x_2)$ is a twice continuously differentiable function of two variables.

(4) Find the local maxima and minima of the function

$$f(\mathbf{x}) = \log(x_1^2 + x_2^2 + 1)$$

in the plane $\mathbb{R}^2$.

(5) Show that the function

$$f(x_1, x_2) = (x_1^2 + 2x_2^2)e^{-x_1^2 - x_2^2}$$

has a minimal and a maximal value in $\mathbb{R}^2$; find the points at which these extrema are attained, and the values of the function f at those points.

(6) Find the directional derivative

$$\partial_u\big(x_1^2 x_2 + \sin(\pi x_1 x_3)\big)$$

at the point $(0, 1, 2)$ in the direction of the unit vector $u = \frac{1}{\sqrt{14}}(1, 2, 3)$.

(7) Find the direction in which the function

$$f(\mathbf{x}) = x_1 x_2 + x_2 x_3 + x_3 x_1$$

in $\mathbb{R}^3$ is increasing and decreasing most rapidly at the point $(1, 1, 2)$

PROBLEMS: GROUP 2

(8) Show that the function $f(\mathbf{x}) = \sqrt{|x_1 x_2|}$ (the non-negative square root) is not differentiable at the origin. Show that a function $f(\mathbf{x})$ such $|f(\mathbf{x})| \le \|\mathbf{x}\|_2^2$ in an open neighborhood of the origin is differentiable at the origin, and find its derivative at the origin. If a function $f(\mathbf{x})$ satisfies $|f(\mathbf{x})| \le \|\mathbf{x}\|_2$ in an open neighborhood of the origin is it necessarily differentiable at the origin? (Either show that it is differentiable or give a counterexample.)

(9) If f_1, f_2 are $\mathcal{C}^1$ functions defined in an open neighborhood of the point $(2,1)$ in $\mathbb{R}^2$ such that $f_1(2,1) = 4$ and $f_2(2,1) = 3$, and that

$$f_1^2 + f_2^2 + 4 = 29 \quad \text{and} \quad \frac{f_1^2}{x_1^2} + \frac{f_2^2}{x_2^2} + 4 = 17,$$

find $\partial_1 f_2(2,1)$ and $\partial_2 f_1(2,1)$.

(10) If f is a $\mathcal{C}^1$ real-valued function in $\mathbb{R}^2$ such that

$$f(1,2) = f(1,4) = f(2,1) = f(4,1) = 1$$

and

$$\begin{array}{lllll} \partial_1 f(1,1) = -1, & \partial_1 f(1,2) = 1, & \partial_1 f(1,4) = 3, & \partial_1 f(2,1) = 4, & \partial_1 f(4,1) = 2, \\ \partial_2 f(1,1) = -1, & \partial_2 f(1,2) = 5, & \partial_2 f(1,4) = -1, & \partial_2 f(2,1) = 2, & \partial_2 f(4,1) = -3, \end{array}$$

find $\partial_1 g(1,2)$ where

$$g(x_1, x_2) = f\big(f(x_1, x_2),\ f(x_2^2, x_1^3)\big).$$

(11) If g is a $\mathcal{C}^1$ function in $\mathbb{R}^2$ and

$$f(x_1, x_2, x_3) = g\big(x_1, g(x_1, g(x_2, x_3))\big),$$

find the derivative of the function f in terms of the partial derivatives of the function g.

(12) Show that the set $\mathcal{M}$ of real 2×2 matrices of the form

$$M(a,b) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

form a real algebra, in the sense that the sum and product of two matrices of this form have the same form. Show that the mapping that associates to a complex number $z = a + i\,b \in \mathbb{C}$ the matrix $M(z) = M(a,b) \in \mathcal{M}$ has the property that $M(z_1 + z_2) = M(z_1) + M(z_2)$ and $M(z_1 z_2) = M(z_1)M(z_2)$. (The algebra of such matrices thus is isomorphic to the algebra of complex numbers). Show that the set of continuously differentiable mappings $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ satisfying the partial differential equation $\mathbf{f}'(\mathbf{x}) \in \mathcal{M}$ is closed under composition, in the sense that if $\mathbf{f}$

and $\mathbf{g}$ are solutions of this partial differential equation then so is the composition $\mathbf{f} \circ \mathbf{g}$. Show that the identity mapping is a solution of this partial differential equation, but that the mapping $(x_1, x_2) \longrightarrow (x_1, -x_2)$ is not. (This partial differential equation is called the Cauchy-Riemann equation; the solutions are those functions that are limits of complex polynomials in the complex variable $z = x + i\,y$.) Describe the set of mappings : $\mathbb{R}^{\mathbf{2}} \longrightarrow \mathbb{R}^{\mathbf{2}}$ satisfying the partial differential equation $\mathbf{f}'(\mathbf{x}) \in \mathcal{N}$ for the algebra $\mathcal{N}$ of matrices of the form

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}.$$

(13) Show that the function

$$f(\mathbf{x}) = \begin{cases} \dfrac{x_1|x_2|}{\sqrt{x_1^2 + x_2^2}} & \text{if} \quad (x_1, x_2) \neq (0,0), \quad \text{positive square root} \\ 0 & \text{if} \quad (x_1, x_2) = (0,0), \end{cases}$$

has a well defined directional derivative in each direction at the origin but is not differentiable at the origin.

(14) Find real numbers a, b such that the line $y = ax + b$ in the plane is the best fit to n points $(x_i, y_i) \in \mathbb{R}^2$, in the sense that $\sum_{i=1}^{n}(y_i - ax_i - b)^2$ is minimized among all possible values of a and b.

(15) Show that a strictly subharmonic function in the plane, a $\mathcal{C}^2$ function $f(\mathbf{x})$ such that $\partial_1^2 f(\mathbf{x}) + \partial_2^2 f(\mathbf{x}) > 0$ for all points $\mathbf{x} \in \mathbb{R}^2$, cannot have a maximum at any point in the plane. Can such a function have a minimum in the plane? Why?

(16) Show that a twice continuously differentiable function $f(x_1, x_2)$ in $\mathbb{R}^2$ satisfies $\partial_1 \partial_2 f(x_1, x_2) = 0$ for all points $\mathbf{x} \in \mathbb{R}^2$ if and only if there are twice continuously differentiable functions g_1, g_2 in $\mathbb{R}$ such that $f(x_1, x_2) = g_1(x_1) + g_2(x_2)$. To what extent are the functions $g_1(x)$ and $g_2(x)$ uniquely determined?

(17) Show that if $f(x_1, x_2)$ is three times continuously differentiable in an open neighborhood of the origin, if the origin is a critical point at which the Hessian is also zero, and if at least one third partial derivative of f at the origin is nonzero, then the origin is not a local maximum or local minimum of f.

Chapter 3

The Rank Theorem

chrankth

3.1 The Inverse Mapping Theorem

secrankth

A continuous one-to-one mapping $\boldsymbol{\phi} : U \longrightarrow V$ between two open subsets $U, V \subset \mathbb{R}^n$ with a continuous inverse $\boldsymbol{\phi}^{-1} : V \longrightarrow U$ is called a **homeomorphism** between these two sets. A homeomorphism really identifies the two sets as far as topological properties are concerned. Indeed the mapping $\boldsymbol{\phi}$ takes open subsets of U to open subsets of V, since the image $\boldsymbol{\phi}(E) \subset V$ of any open subset $E \subset U$ is the inverse image $(\boldsymbol{\phi}^{-1})^{-1}(E)$ of the set E under the continuous mapping $\boldsymbol{\phi}^{-1} : V \longrightarrow U$, and correspondingly the inverse mapping $\boldsymbol{\phi}^{-1}$ takes open subsets of V to open subsets of U; consequently the mapping $\boldsymbol{\phi}$ identifies continuous functions $f : U \longrightarrow \mathbb{R}$ on U with continuous functions $f \circ \boldsymbol{\phi}^{-1} : V \longrightarrow \mathbb{R}$ on V, and correspondingly for the inverse mapping $\boldsymbol{\phi}^{-1}$. If the mapping and its inverse are both continuously differentiable, in which case the mapping is called a $\mathcal{C}^1$ **homeomorphism** or sometimes a $\mathcal{C}^1$ **diffeomorphism**, this mapping also identifies $\mathcal{C}^1$ functions on the two sets; and if the mapping $\boldsymbol{\phi}$ and its inverse is a $\mathcal{C}^r$ mappings it identifies $\mathcal{C}^r$ functions on U with those on V. It is a familiar result from calculus in one variable that if $f : I \longrightarrow \mathbb{R}^1$ is a continuously differentiable function in an open interval $I \subset \mathbb{R}^1$ such that $f'(x) \neq 0$ for all $x \in I$ then the mapping f is a $\mathcal{C}^1$ homeomorphism from I to its image $f(I)$. The proof of this result is often skipped or treated casually; but it is a model for one proof of the corresponding result in higher dimensions, so it may be helpful to review that special case first. Suppose that $f : [a, b] \longrightarrow [c, d]$ is a continuous mapping between two closed intervals in $\mathbb{R}^1$ such that f is $\mathcal{C}^1$ in (a, b) and $f'(x) \neq 0$ at each point $x \in (a, b)$; since the image $f([a, b])$ of the compact connected set $[a, b]$ is necessarily also compact and connected it must be a closed interval, so it can be assumed that $f([a, b]) = [c, d]$. The set of points where $f'(x) > 0$ and the set of points at which $f'(x) < 0$ are open subsets of (a, b), the union of which is the entire interval (a, b); and since an interval is connected it must be the case that one of these two sets is empty, so either $f'(x) > 0$ at all points $x \in (a, b)$ or $f'(x) < 0$ at all points $x \in (a, b)$. Suppose first that $f'(x) > 0$ at all points $x \in (a, b)$. It follows from the mean value theorem that whenever $a \le x_1 < x_2 \le b$ there is a point ξ such that $x_1 < \xi < x_2$ and $f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1) > 0$; thus f is a strictly increasing function in $[a, b]$, so it is a one-to-one mapping from the closed interval $[a, b]$ onto the closed interval $[c, d]$ and has a well defined inverse mapping $g : [c, d] \longrightarrow [a, b]$. For any $\delta > 0$ the continuous positive function $f'(x)$ attains its minimum value at a point in the compact subset $[a + \delta, b - \delta]$, and that minimum must be a positive number $m > 0$ since $f'(x) > 0$ everywhere; so $f'(x) \ge m > 0$ for all $x \in [a + \delta, b - \delta]$. Therefore whenever $a + \delta \le x_1 < x_2 \le b - \delta$ it follows from the mean value theorem that there is a point ξ such that $x_1 < \xi < x_2$ and $f(x_2) - f(x_1) = f'(\xi)(x_2 - x_1) \ge m(x_2 - x_1)$. If $y_i = f(x_i)$ then $x_i = g(y_i)$ and the preceding inequality can be written $y_2 - y_1 \ge m\big(g(y_2) - g(y_1)\big)$. That is the case for any points y_i for which $f(a + \delta) \le y_1 < y_2 \le f(b - \delta)$, which shows that the inverse function g is continuous in the interval $[f(a + \delta), f(b - \delta)]$; and since that is the case for any $\delta > 0$ it follows that the function g is continuous in (c, d). Furthermore in the interval $c \le y_1 < y_2 \le d$ if $x_i = g(y_i)$ then

$$\frac{g(y_2) - g(y_1)}{y_2 - y_1} = \frac{x_2 - x_1}{f(x_2) - f(x_1)};$$

since g is continuous, x_2 tends to x_1 as y_2 tends to y_1, so in the limit the preceding equation reduces to $g'(y_1) = 1/f'(x_1)$, showing that the function g is differentiable and that its derivative is also continuous. The corresponding result of course holds if $f'(x) < 0$ in (a, b). This argument critically used the hypothesis that the derivative $f'(x)$ is a continuous function, and the result does not hold without that hypothesis. Indeed the function

$$h(x) = \frac{1}{2}x + f(x) = \frac{1}{2}x + x^2 \sin\frac{1}{x},$$

where $f(x)$ is the function (diffexnc 2.4) considered earlier, is differentiable at all points $x \in \mathbb{R}^1$ and $h'(0) = \frac{1}{2}$; but the derivative of $h(x)$ alternates in sign in any open neighborhood of 0 so it cannot be a one-to-one mapping in any open neighborhood of 0.

If $\mathbf{f} : U \longrightarrow V$ is a $\mathcal{C}^1$ homeomorphism between two open subsets $U, V \subset \mathbb{R}^n$ and if $\mathbf{g} : V \longrightarrow U$ is its inverse then $\mathbf{g}\big(\mathbf{f}(\mathbf{x})\big) = \mathbf{x}$, and an application of the chain rule as on page comp3 24 shows that $\mathbf{g}'\big(\mathbf{f}(\mathbf{x})\big)\mathbf{f}'(\mathbf{x}) = \mathrm{I}$, the identity matrix, and consequently that $\det \mathbf{f}'(\mathbf{x}) \ne 0$ at each point $\mathbf{x} \in U$; thus a necessary condition for the mapping $\mathbf{f}$ to be a $\mathcal{C}^1$ homeomorphism is that $\det \mathbf{f}'(\mathbf{x}) \ne 0$ for all $\mathbf{x} \in U$. The matrix $\mathbf{f}'$ is called the **Jacobian matrix** of the mapping $\mathbf{f}$, and its determinant is called the **Jacobian** of the mapping $\mathbf{f}$. This necessary condition is sufficient locally. It is convenient first to establish a special case of that result, from which the general case follows easily.

invth **Theorem 3.1** *If* $\mathbf{f} : W \longrightarrow \mathbb{R}^n$ *is a continuously differentiable mapping defined in an open neighborhood* $W \subset \mathbb{R}^n$ *of the origin such that* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$ *and* $\mathbf{f}'(\mathbf{0}) = \mathrm{I}$, *the identity matrix, then for any sufficiently small open subneighborhood* $U \subset W$

of the origin the restriction of the mapping $\mathbf{f}$ *to* U *is a* $\mathcal{C}^1$ *homeomorphism* $\mathbf{f} : U \longrightarrow V$ *between* U *and an open neighborhood* $V \subset \mathbb{R}^n$ *of the origin.*

Proof: The mapping $\mathbf{r}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) - \mathbf{x}$ is a continuously differentiable mapping such that $\mathbf{r}(\mathbf{0}) = \mathbf{0}$ and $\mathbf{r}'(\mathbf{0}) = \mathbf{0}$; therefore there is a closed cell $\overline{\Delta} \subset W \subset \mathbb{R}^n$ centered at the origin $\mathbf{0}$ sufficiently small that the following conditions are satisfied:
(i) $\|\mathbf{r}'(\mathbf{x})\|_\infty \le \frac{1}{n^2}$ for all $\mathbf{x} \in \overline{\Delta}$,
(ii) $\det \mathbf{f}'(\mathbf{x}) \ne 0$ for all $\mathbf{x} \in \overline{\Delta}$.
The first step in the proof is to establish a basic inequality on which the remainder of the proof rests. From the Mean Value Inequality (2.24) and (i) it follows that for any points $\mathbf{a}, \mathbf{b} \in \overline{\Delta}$ there is a point $\mathbf{c}$ on the line connecting those two points for which

$$\|\mathbf{r}(\mathbf{b}) - \mathbf{r}(\mathbf{a})\|_\infty \le n\sqrt{n}\|\mathbf{r}'(\mathbf{c})\|_\infty \|\mathbf{b} - \mathbf{a}\|_\infty \le \frac{1}{2\sqrt{n}}\|\mathbf{b} - \mathbf{a}\|_\infty;$$

so by (1.7)

$$\|\mathbf{r}(\mathbf{b}) - \mathbf{r}(\mathbf{a})\|_2 \le \sqrt{n}\|\mathbf{r}(\mathbf{b}) - \mathbf{r}(\mathbf{a})\|_\infty \le \frac{1}{2}\|\mathbf{b} - \mathbf{a}\|_\infty \le \frac{1}{2}\|\mathbf{b} - \mathbf{a}\|_2.$$

Then from this inequality and the triangle inequality it follows that

$$\|\mathbf{b}-\mathbf{a}\|_2 \le \underbrace{\|\big(\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\big) - (\mathbf{b} - \mathbf{a})\|_2}_{=\mathbf{r}(\mathbf{b})-\mathbf{r}(\mathbf{a})} + \|\big(\mathbf{f}(\mathbf{b})-\mathbf{f}(\mathbf{a})\big)\|_2 \le \frac{1}{2}\|\mathbf{b}-\mathbf{a}\|_2 + \|\big(\mathbf{f}(\mathbf{b})-\mathbf{f}(\mathbf{a})\big)\|_2$$

hence that

invq1a (3.1)
$$\|\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a})\|_2 \ge \frac{1}{2}\|\mathbf{b} - \mathbf{a}\|_2,$$

which is the basic inequality.

It follows from (3.1) that the mapping $\mathbf{f} : \overline{\Delta} \longrightarrow \mathbb{R}^n$ is an injective mapping; for if $\mathbf{f}(\mathbf{b}) = \mathbf{f}(\mathbf{a})$ that inequality shows that $\|\mathbf{b} - \mathbf{a}\|_2 = 0$. In particular $\mathbf{f}(\mathbf{x}) \ne \mathbf{f}(\mathbf{0}) = \mathbf{0}$ for any point $\mathbf{x} \in \partial\Delta$, the boundary of the cell $\overline{\Delta}$; so since $\partial\Delta$ is compact there is a positive number $d > 0$ such that

invqleqq1 (3.2)
$$\|\mathbf{f}(\mathbf{x})\|_2 \ge d > 0 \quad \text{for all} \quad \mathbf{x} \in \partial\Delta.$$

The next step is to show that the open neighborhood V of the origin defined by

invqleqq2 (3.3)
$$V = \left\{ \mathbf{y} \in \mathbb{R}^n \;\middle|\; \|\mathbf{y}\|_2 < \frac{d}{2} \right\}$$

is contained in the image $\mathbf{f}(\overline{\Delta})$. To demonstrate that, for any point $\mathbf{b} \in V$ the function $\psi(\mathbf{x}) = \|\mathbf{f}(\mathbf{x}) - \mathbf{b}\|_2$ is a continuous function on $\overline{\Delta}$, and it is just necessary to show that $\psi(\mathbf{a}) = 0$ for some point $\mathbf{a} \in \Delta$. Since $\mathbf{b} \in V$ then

$\psi(\mathbf{0}) = \|\mathbf{b}\|_2 < \frac{d}{2}$ by (3.3), while if $\mathbf{x} \in \partial\Delta$ then in view of (3.2) and the triangle inequality

$$d \le \|\mathbf{f}(\mathbf{x})\|_2 \le \|\mathbf{f}(\mathbf{x}) - \mathbf{b}\|_2 + \|\mathbf{b}\|_2 = \psi(\mathbf{x}) + \|\mathbf{b}\|_2 < \psi(\mathbf{x}) + \frac{d}{2}$$

so $\psi(\mathbf{x}) > \frac{d}{2}$. Thus $\psi(\mathbf{0}) < \psi(\mathbf{x})$ for any point $\mathbf{x} \in \partial\Delta$, so the function $\psi(\mathbf{x})$ must take its minimum value at some interior point $\mathbf{a} \in \Delta$; and that point is the obvious candidate to be a zero of the function $\psi(\mathbf{x})$. The square of this function, the function $\psi(\mathbf{x})^2 = \sum_{i=1}^n \big(f_i(\mathbf{x}) - b_i\big)^2$, then also takes its minimum value at the point $\mathbf{a} \in \Delta$, hence that point is a critical point of $\psi(\mathbf{x})^2$, a zero of all of its partial derivatives $\partial_j\psi(\mathbf{x})^2 = \sum_{i=1}^n 2\big(f_i(\mathbf{x}) - b_i\big)\partial_j f_i(\mathbf{x})$; thus $0 = \sum_{i=1}^n 2\big(f_i(\mathbf{a}) - b_i\big)\partial_j f_i(\mathbf{a})$ for $1 \le j \le n$, which can be written equivalently in terms of the matrix $\mathbf{f}'(\mathbf{a})$ as

$$2\mathbf{f}'(\mathbf{a})\big(\mathbf{f}(\mathbf{a}) - \mathbf{b}\big) = 0.$$

However the matrix $\mathbf{f}'(\mathbf{a})$ is nonsingular by (ii), hence $\mathbf{f}(\mathbf{a}) - \mathbf{b} = 0$ as desired.

What has been shown so far is that the mapping $\mathbf{f} : \overline{\Delta} \longrightarrow \mathbb{R}^n$ is an injective mapping, and that $V \subset \mathbf{f}(\Delta)$. Consequently the subset $U = \mathbf{f}^{-1}(V) \subset \Delta$ is an open neighborhood of the point $\mathbf{a}$ for which $\mathbf{f}$ is a one-to-one mapping $\mathbf{f} : U \longrightarrow V$ from U onto V; so there is a well defined inverse mapping $\mathbf{g} = \mathbf{f}^{-1} : V \longrightarrow U$ from V onto U. To show that this mapping $\mathbf{g}$ is continuously differentiable, for any points $\mathbf{y}, \mathbf{y}+\mathbf{k} \in V$ let $\mathbf{x}, \mathbf{x}+\mathbf{h}$ be their images under the mapping $\mathbf{g} : V \longrightarrow U$, so that

invbq1 (3.4)
$$\begin{aligned} \mathbf{x} &= \mathbf{g}(\mathbf{y}) \quad \text{and} \quad \mathbf{x}+\mathbf{h} = \mathbf{g}(\mathbf{y}+\mathbf{k}) \\ \mathbf{y} &= \mathbf{f}(\mathbf{x}) \quad \text{and} \quad \mathbf{y}+\mathbf{k} = \mathbf{f}(\mathbf{x}+\mathbf{h}). \end{aligned}$$

By the basic inequality (3.1)

$$\|\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y})\|_2 = \|(\mathbf{x}+\mathbf{h}) - \mathbf{x}\|_2 \le 2\|\mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x})\|_2 = 2\|(\mathbf{y}+\mathbf{k}) - \mathbf{x}\|_2 = 2\|\mathbf{k}\|_2,$$

hence $\mathbf{g}$ is continuous. Since $\mathbf{f}$ is differentiable

$$\mathbf{k} = \mathbf{f}(\mathbf{x}+\mathbf{h}) - \mathbf{f}(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\mathbf{h} + \boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h}) \quad \text{where } \lim_{\mathbf{k}\to 0} \frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2} = 0,$$

from which it follows that

invthq1 (3.5)
$$\mathbf{g}(\mathbf{y}+\mathbf{k}) - \mathbf{g}(\mathbf{y}) = \mathbf{h} = \mathbf{f}'(\mathbf{x})^{-1}\mathbf{k} + \boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})$$

where $\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k}) = -\mathbf{f}'(\mathbf{x})^{-1}\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})$. It follows from (1.10) that $\|\mathbf{f}'(\mathbf{x})^{-1}\mathbf{k}\|_\infty \le A\|\mathbf{k}\|_\infty$ where $A = n\|\mathbf{f}'(\mathbf{x})^{-1}\|_\infty$, so by (1.7) equivalently $\|\mathbf{f}'(\mathbf{x})^{-1}\mathbf{k}\|_2 \le B\|\mathbf{k}\|_2$ where $B = \sqrt{n}A$, and similarly $\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_2 = \|-\mathbf{f}'(\mathbf{x})^{-1}\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2 \le B\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2$; altogether then

invthq1a (3.6)
$$\frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_2}{\|\mathbf{k}\|_2} \le B\frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2} \cdot \frac{\|\mathbf{h}\|_2}{\|\mathbf{k}\|_2} \le 2B\frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2}$$

since by the fundamental inequality (3.1) yet again

$$\frac{\|\mathbf{h}\|_2}{\|\mathbf{k}\|_2} = \frac{\|(\mathbf{x}+\mathbf{h})-\mathbf{x}\|_2}{\|\mathbf{f}(\mathbf{x}+\mathbf{h})-\mathbf{f}(\mathbf{x})\|_2} \le 2.$$

Since $\lim_{\mathbf{h}\to 0}\frac{\|\boldsymbol{\epsilon}_{\mathbf{f}}(\mathbf{h})\|_2}{\|\mathbf{h}\|_2} = 0$ while $\lim_{\mathbf{k}\to 0}\mathbf{h} = 0$ as a consequence of the continuity of the mapping $\mathbf{g}$, it follows from (3.6) that $\lim_{\mathbf{k}\to 0}\frac{\|\boldsymbol{\epsilon}_{\mathbf{g}}(\mathbf{k})\|_2}{\|\mathbf{k}\|_2} = 0$ and consequently from (3.5) that the mapping $\mathbf{g}$ is differentiable, and moreover that $\mathbf{g}'(\mathbf{y}) = \mathbf{f}'(\mathbf{x})^{-1} = \mathbf{f}'\big(\mathbf{g}(\mathbf{y})\big)^{-1}$ so the derivative $\mathbf{g}'(\mathbf{y})$ is also continuous; that suffices for the proof.

Theorem 3.2 (Inverse Mapping Theorem) *If $\mathbf{f}: W \longrightarrow \mathbb{R}^n$ is a continuously differentiable mapping in an open neighborhood $W \subset \mathbb{R}^n$ of a point $\mathbf{a} \in \mathbb{R}^n$ such that $\det \mathbf{f}'(\mathbf{a}) \neq 0$ then for any sufficiently small open subneighborhood $U \subset W$ of the point $\mathbf{a}$ the restriction of the mapping $\mathbf{f}$ to U is a $\mathcal{C}^1$ homeomorphism $\mathbf{f}: U \longrightarrow V$ between U and an open neighborhood V of the image point $\mathbf{b} = \mathbf{f}(\mathbf{a})$. If $\mathbf{g}: V \longrightarrow U$ is the inverse mapping then $\mathbf{g}'(\mathbf{y}) = \mathbf{f}'\big(\mathbf{g}(\mathbf{y})\big)^{-1}$ for each point $\mathbf{y} \in V$.*

Proof: Introduce the invertible affine mappings $\phi, \psi: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ defined by

$$\phi(\mathbf{x}) = \mathbf{f}'(\mathbf{a})^{-1}\mathbf{x} + \mathbf{a}, \quad \psi(\mathbf{x}) = \mathbf{x} - \mathbf{b}.$$

The composition $\mathbf{F} = \psi \circ \mathbf{f} \circ \phi$ is then a continuously differentiable mapping from an open neighborhood of the origin $\mathbf{0} \in \mathbb{R}^n$ into $\mathbb{R}^n$ such that $\mathbf{F}(\mathbf{0}) = \mathbf{0}$, and by the chain rule $F'(\mathbf{0}) = \psi'(\mathbf{b})\mathbf{f}'(\mathbf{a})\phi'(\mathbf{0}) = I \cdot \mathbf{f}'(\mathbf{a})\mathbf{f}'(\mathbf{a})^{-1} = I$. The preceding theorem shows that the mapping F is invertible, and that its inverse $\mathbf{G}$ is a continuously differentiable mapping from an open neighborhood of the origin into $\mathbb{R}^n$. The composition $\mathbf{g} = \phi \circ \mathbf{G} \circ \psi$ is a continuously differentiable mapping from an open neighborhood of $\mathbf{b}$ into $\mathbb{R}^n$, and since $\mathbf{f} = \psi^{-1} \circ \mathbf{F} \circ \phi^{-1}$ it follows that $\mathbf{g} \circ \mathbf{f} = \phi \circ \mathbf{G} \circ \psi \circ \psi^{-1} \circ \mathbf{F} \circ \phi^{-1}$ is the identity mapping so that $\mathbf{g}$ is the inverse mapping to $\mathbf{f}$. Finally since $\mathbf{y} = \mathbf{f}\big(\mathbf{g}(\mathbf{y})\big)$ for any point $\mathbf{y} \in V$ it follows from the chain rule that $I = \mathbf{f}'\big(\mathbf{g}(\mathbf{y})\big)\mathbf{g}'(\mathbf{y})$ where I is the identity matrix, so that $\mathbf{g}'(\mathbf{y}) = \mathbf{f}'\big(\mathbf{g}(\mathbf{y})\big)^{-1}$, and that concludes the proof.

One consequence of the Inverse Mapping Theorem is that if $\mathbf{f}: U \longrightarrow \mathbb{R}^n$ is a continuously differentiable mapping defined in an open subset $U \subset \mathbb{R}^n$ such that $\det \mathbf{f}'(\mathbf{x}) \neq 0$ for all points $\mathbf{x} \in U$ then $\mathbf{f}$ is an **open mapping**, in the sense that the image of any open subset $U_0 \subset U$ is an open subset $\mathbf{f}(U_0) \subset \mathbb{R}^n$. The inverse image of any open set under a continuous mapping is always open; but the image of an open subset under a general continuous mapping is not necessarily open, so open continuous mappings are a special subclass of continuous mappings. If $\mathbf{f}: U \longrightarrow \mathbb{R}^n$ is an invertible continuously differentiable mapping from an open subset $U \subset \mathbb{R}^n$ with coordinates $(t_1, \ldots, t_n)$ onto an open subset $V \subset \mathbb{R}^n$ with coordinates $(x_1, \ldots, x_n)$, points in V can be described either in terms of the coordinates $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$ or alternatively in terms of the coordinates

$(t_1, \ldots, t_n) \in U \subset \mathbb{R}^n$; for this reasons the parameters $(t_1, \ldots, t_n)$ are often described as being another **local coordinate system** in V, and the mapping $\mathbf{f}$ is called a $\mathcal{C}^1$ **local change of coordinates** in V. Of course it is also possible just to consider continuous changes of coordinates through a homeomorphism that is not necessarily given by a $\mathcal{C}^1$ mapping. The usefulness of a change of coordinates is that it may be possible to describe the local geometry much more simply in terms of the coordinates $(t_1, \ldots, t_n)$ than in terms of the initial coordinates $(x_1, \ldots, x_n)$. For example, an arc of a circle of radius 1 centered at the origin in the plane with a coordinate system (x_1, x_2) can be described locally as a straight line segment $r = 1$ in terms of polar coordinates (r, θ). It should be emphasized though that the Inverse Mapping Theorem is only a local result; a mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^m$ for which $\det \mathbf{f}'(\mathbf{x}) \neq 0$ at all points $\mathbf{x} \in U \subset \mathbb{R}^n$ need not be invertible if $n > 1$. For example the derivative of the mapping $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by $\mathbf{f}(r, \theta) = (r \cos\theta, r \sin\theta)$ is the matrix

plcrd1

$$\mathbf{f}'(r,\theta) = \begin{pmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \tag{3.7}$$

for which $\det \mathbf{f}'(r, \theta) = r$. The inverse Mapping Theorem asserts that the mapping $\mathbf{f}$ is locally invertible near any point $(r, \theta) \in \mathbb{R}^2$ for which $r \neq 0$; but $\mathbf{f}$ is clearly not a one-to-one mapping in the open subset $U = \{\ (r, \theta) \mid r > 0\ \}$ since $\mathbf{f}(r, \theta) = \mathbf{f}(r, \theta + 2n\pi)$ for any integer n. It is very difficult to find general conditions ensuring that a mapping is globally invertible[1]. Points at which the Jacobian of a differentiable mapping $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ vanishes are known as **singular points** or **singularities** of the mapping. Mappings can be very complicated indeed near their singular points. For a simple example, consider again the mapping $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ given by $\mathbf{f}(r, \theta) = (r \cos\theta, r \sin\theta)$; the origin $r = \theta = 0$ is a singular point since $\mathbf{f}'(0, 0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. The mapping is sketched

[1]The "Jacobian Problem" is whether a mapping $\mathbf{f} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ from the n-dimensional vector space to itself for $n > 1$ defined by polynomial equations for which $\det \mathbf{f}'(\mathbf{x}) = 1$ at all points $\mathbf{x} \in \mathbb{R}^n$ is necessarily a homeomorphism from $\mathbb{R}^n$ onto $\mathbb{R}^n$. This is a long standing open problem in mathematics, having been raised by O. H. Keller in 1939 for polynomial mappings $\mathbf{f} : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ over the complex numbers. It is known to be true for polynomial mappings $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ of degree at most 100, but not in general, at least at the time of this writing. It is known to be false for polynomial mappings over finite fields. It is also false for mappings involving functions more general than polynomials; for instance the mapping $\mathbf{f} : \mathbb{C}^2 \longrightarrow \mathbb{C}^2$ defined by $\mathbf{f}(x, y) = (e^x, e^{-x}y)$ has the Jacobian $\det \mathbf{f}'(x, y) = 1$ at all points (x, y) but its image excludes the axis $(0, y)$ and the mapping is not one-to-one. A famous example due to Fatou and Bieberbach, and discussed for example in the book *Several Complex Variables* by S. Bochner and W. T. Martin, (Princeton University Press, 1948), is an injective mapping $\mathbf{f} : \mathbb{C}^2 \longrightarrow \mathbb{C}^2$ given by functions that are everywhere analytic and for which the Jacobian $\det \mathbf{f}'(\mathbf{x}) = 1$ at all points $\mathbf{x} \in \mathbb{C}^2$, but the image omits an open subset of $\mathbb{C}^2$. A survey of the problem and some approaches, and some :saexamples of some false proofs, can be found in the paper by H. Bass, E.H. Connell, D. Wright in the *Bulletin of the American Mathematical Society*, volume 7, 1982. If it is just assumed that $\det \mathbf{f}'(\mathbf{x}) \neq 0$ for all points $\mathbf{x} \in \mathbb{C}^2$, the "strong Jacobian problem", there is a counterexample for mappings $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ by S Pinchuk in *Mathematische Zeitschrift*, volume 217, 1994.

in the accompanying Figure singmp 3.1. A cell centered at the origin is mapped to the bow-tie region, and the mapping is locally one-to-one except along the axis $r = 0$; that entire axis is mapped to a single point, the origin. The general study of the geometry of singularities of mappings is quite extensive, including what is known as the theory of catastrophes[2].

Figure 3.1: The mapping $\mathbf{f} : (r, \theta) \longrightarrow (r\cos\theta, r\sin\theta)$ in an open neighborhood of the origin. singmp

An important and sometimes somewhat confusing concept is that of **orientation**. If $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ is a continuously differentiable mapping from an open subset $U \subset \mathbb{R}^n$ into $\mathbb{R}^n$ and if $\det \mathbf{f}'(\mathbf{a}) \neq 0$ at a point $\mathbf{a} \in U$ then either $\det \mathbf{f}'(\mathbf{a}) > 0$, in which case the mapping $\mathbf{f}$ is said to be **orientation preserving** at the point $\mathbf{a}$, or $\det \mathbf{f}'(\mathbf{a}) < 0$, in which case the mapping $\mathbf{f}$ is said to be **orientation reversing** near the point $\mathbf{a}$. If U is a connected open set and the Jacobian of the mapping $\mathbf{f}$ is nowhere vanishing then the mapping $\mathbf{f}$ is either orientation preserving at all points of U or orientation reversing at all points of U. In the case $n = 1$ an orientation preserving mapping is a monotonically increasing mapping, and an orientation reversing mapping is a monotonically decreasing mapping; the orientation of $\mathbb{R}^1$, identified with the set of real numbers, is the direction of increasing numbers. If $n > 1$ each coordinate axis in $\mathbb{R}^n$ is naturally oriented in the direction in which the variable increases, and the orientation of $\mathbb{R}^n$ itself is defined by the choice of an order of the coordinate axes. For example in the case $n = 2$ with the variables x_1, x_2 it is customary to specify an orientation as either clockwise or counterclockwise, where clockwise determines the order x_2, x_1 of the axes while counterclockwise determines the reversed order x_1, x_2 of the axes; and in the case $n = 3$ the right-hand rule familiar from physics specifies the order of the axes as that determined by the order of the thumb and first two fingers on the right hand, when these fingers are viewed as lying along three coordinate axes. When the variables are denoted by $x_1, x_2, \ldots, x_n$ the natural order is the corresponding order of the axes of these variables, first the x_1 axis, then the x_2 axis, and so on. However for variables such as r, θ in $\mathbb{R}^2$ or $\alpha, a, \aleph$ in $\mathbb{R}^3$ there is no natural order so it is necessary to specify an order of the axes. Two different orders though may lead to the same orientation. A permutation of the variables is a linear transformation M described by a permutation matrix M, a matrix that has a single entry of 1 in each row and column but all other entries 0; the linear transformation described by a permutation matrix, a permutation of the coordinates, is orientation preserving if $\det M = +1$ but orientation reversing if $\det M = -1$. Thus for $n = 3$ for

[2]See for example the books *Catastrophe Theory and Its Applications* by Tim Poston and Ian Stewart, New York: Dover, 1998, or *Catastrophe Theory* by Vladimir Arnold, Berlin: Springer-Verlag, 1992.

example the orders x_1, x_2, x_3 and x_2, x_3, x_1 and x_3, x_1, x_2 determine the same orientation since these orders arise from the first by repeated applications of the permutation matrix $M_+ = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ for which $\det M_+ = +1$; on the other hand the orders x_2, x_1, x_3 and x_3, x_2, x_1 and x_1, x_3, x_2 arise from the three previous orders by by applications of the permutation matrix $M_- = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ for which $\det M_- = -1$, the permutation that switches the first two coordinates. A permutation that interchanges any two consecutive vectors is orientation reversing, since $\det\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = -1$, and any permutation can be written as a successive application of such interchanges; so beginning with any chosen order of the coordinates, those orders that arise by an even number of such interchanges beginning with the initially chosen order have the same orientation as the initially chosen order, while those that arise by an odd number of such interchanges have the opposite orientation. Thus there are aways just two possible orientations of the vector space $\mathbb{R}^n$ for any $n > 0$. For example, the vector space $\mathbb{R}^2$ with coordinates r, θ can be given the orientation r, θ and the vector space $\mathbb{R}^2$ with coordinates x, y can be given the orientation x, y. The mapping $\mathbf{f} : \mathbb{R}^2(r, \theta) \longrightarrow \mathbb{R}^2(x, y)$ that takes a point (r, θ) to the point $\mathbf{f}(r, \theta) = (r\cos\theta, r\sin\theta)$ has the derivative $\mathbf{f}'(r, \theta) = r$, and consequently

if $r > 0$ then $\mathbf{f}'(r, \theta) > 0$ and $\mathbf{f}$ is orientation preserving,
if $r < 0$ then $\mathbf{f}'(r, \theta) < 0$ and $\mathbf{f}$ is orientation reversing,
if $r = 0$ then $\mathbf{f}'(r, \theta) = 0$ and $\mathbf{f}$ is singular,

as indicated in the accompanying Figure orient 3.2. Orientation can be described

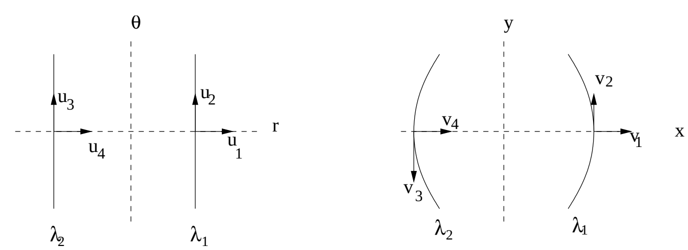

Figure 3.2: The mapping $\mathbf{f} : (r, \theta) \longrightarrow (r\cos\theta, r\sin\theta)$ preserves the orientation on the set where $r > 0$ but reverses the orientation on the set where $r < 0$. orient

alternatively in terms of the unit vectors dx_i along the coordinate axes rather than in terms of the coordinate axes themselves; and the equivalence of different choices of the order of the unit vectors can be handled through a special case of the exterior algebra generated by these vectors as discussed in the Appendix. Associate to any order of the basis vectors dx_i such as $dx_{i_1}, dx_{i_2}, \ldots, dx_{i_n}$ a formal vector denoted by

orien1 (3.8)
$$dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_n},$$

with the convention that when the order of two adjacent indices is reversed this vector changes sign; thus

orien2 (3.9) $dx_{i_1}\wedge\cdots\wedge dx_{i_k}\wedge dx_{i_{k+1}}\wedge\cdots\wedge dx_{i_n} = -dx_{i_1}\wedge\cdots\wedge dx_{i_{k+1}}\wedge dx_{i_k}\wedge\cdots\wedge dx_{i_n}.$

This vector is not to be interpreted as a vector in the space $\mathbb{R}^n$, but as a basis vector in the abstract one-dimensional vector space consisting of the products $t dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$ for all $t \in \mathbb{R}$. All the orders of the axes, or equivalently all the orders of the basis vectors dx_i, that determine the same vector (3.8) describe the same orientation of $\mathbb{R}^n$, and those that determine the negative of the vector (3.8) describe the opposite orientation; thus an orientation of $\mathbb{R}^n$ can be specified by the choice of a vector $dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$. For example, in $\mathbb{R}^2$ with the coordinates r, θ one orientation is specified by the vector $dr \wedge d\theta$ and the other possible orientation is specified by the vector $d\theta \wedge dr = -dr \wedge d\theta$.

Locally any $\mathcal{C}^1$ homeomorphism can be written as a composition of simpler homeomorphisms, which permits the local study of $\mathcal{C}^1$ homeomorphisms to be reduced to the examination of simpler mappings when that is helpful. A particularly simple homeomorphism is one that changes only a single variable, a mapping $\mathbf{f}: \mathbb{R}^n \longrightarrow \mathbb{R}^n$ for which the coordinate functions $f_i(\mathbf{x})$ have the form $f_i(\mathbf{x}) = x_i$ for $i \ne j$ while $f_j(\mathbf{x})$ is a function of all the variables.

mpcmp **Theorem 3.3** *A $\mathcal{C}^1$ mapping $\mathbf{f}: U \longrightarrow \mathbb{R}^n$ defined in an open neighborhood U of the origin in $\mathbb{R}^n$, that takes the origin to the origin and has a nonzero Jacobian at the origin, can be written in any sufficiently small open subneighborhood $V \subset U$ of the origin as the composition of $\mathcal{C}^1$ homeomorphisms that are either permutations of the variables in $\mathbb{R}^n$ or mappings that change only a single variable.*

Proof: Suppose that the coordinate functions f_i of the mapping $\mathbf{f}$ have the property that $f_i(\mathbf{x}) = x_i$ whenever $i < r$ for some index r; for convenience such a mapping will be called a mapping of type r, but just for the course of the present proof. Any mapping is of type 1, so it can be assumed that the mapping $\mathbf{f}$ is of type r for some r. It is convenient to simplify the notation in the proof by writing points $\mathbf{x} \in \mathbb{R}^n$ in the form

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_I \\ x_r \\ \mathbf{x}_{II} \end{pmatrix} \text{ where } \begin{cases} \mathbf{x}_I \in \mathbb{R}^{r-1} & I = (1, 2, \dots, r-1) \\ x_r \in \mathbb{R}^1 & \\ \mathbf{x}_{II} \in \mathbb{R}^{n-r-1} & II = (r+1, r+2, \dots, n) \end{cases}$$

and to write the mapping $\mathbf{f}$ correspondingly in the form

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} \mathbf{x}_I \\ f_r(\mathbf{x}) \\ \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix} \text{ where } \begin{cases} \mathbf{x}_I \in \mathbb{R}^{r-1} & I = (1, 2, \dots, r-1) \\ f_r(\mathbf{x}) \in \mathbb{R}^1 & \\ \mathbf{f}_{II}(\mathbf{x}) \in \mathbb{R}^{n-r-1} & II = (r+1, r+2, \dots, n). \end{cases}$$

In these terms the derivative of the mapping $\mathbf{f}$ is the $n \times n$ matrix

invnf (3.10)
$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} I_{r-1} & 0 & 0 \\ \partial_I f_r(\mathbf{x}) & \partial_r f_r(\mathbf{x}) & \partial_{II} f_r(\mathbf{x}) \\ \partial_I \mathbf{f}_{II}(\mathbf{x}) & \partial_r \mathbf{f}_{II}(\mathbf{x}) & \partial_{II} \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix}$$

in which I_{r-1} is the $(r-1)\times(r-1)$ identity matrix while

$$\partial_I f_r(\mathbf{x}) = \big(\partial_1 f_r(\mathbf{x}) \quad \cdots \quad \partial_{r-1} f_r(\mathbf{x})\big)\,, \quad \text{a } 1\times(r-1) \text{ matrix,}$$
$$\partial_I f_{II}(\mathbf{x}) = \Big\{\partial_j f_i(\mathbf{x})\Big| r{+}1 \le i \le n, 1 \le j \le r{-}1\Big\}, \text{ an } (n{-}r{-}1)\times(r{-}1) \text{ matrix,}$$

and similarly for the other components of the matrix (invnf 3.10). Since $\det \mathbf{f}'(\mathbf{0}) \ne 0$ not all the entries in column r of this matrix are zero; so after a suitable permutation of the coordinate functions of the mapping $\mathbf{f}$, that is, after replacing the mapping $\mathbf{f}$ by $P_r \circ \mathbf{f}$ for a suitable permutation matrix P_r, it can be assumed that $\partial_r\big(P_r \circ \mathbf{f}_r\big)(\mathbf{0}) \ne 0$. To simplify the notation assume this change has been made, that is, replace $P_r \circ \mathbf{f}$ by $\mathbf{f}$, so it can assumed now that $\partial_r f_r(\mathbf{0}) \ne 0$.

Let $\boldsymbol{\phi}_r : U \longrightarrow \mathbb{R}^n$ be the mapping defined by

$$\boldsymbol{\phi}_r \begin{pmatrix} \mathbf{x}_I \\ x_r \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_I \\ f_r(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix},$$

for which

invnf2 (3.11)
$$\boldsymbol{\phi}_r'(\mathbf{x}) = \begin{pmatrix} I_{r-1} & 0 & 0 \\ \partial_I f_r(\mathbf{x}) & \partial_r f_r(\mathbf{x}) & \partial_{II} f_r(\mathbf{x}) \\ 0 & 0 & I_{n-r-1} \end{pmatrix}.$$

Since $\det \boldsymbol{\phi}_r'(\mathbf{0}) = \partial_r f_r(\mathbf{0}) \ne 0$ it follows from the Inverse Mapping Theorem that the mapping $\boldsymbol{\phi}_r$ is invertible in a sufficiently small open neighborhood $U_r \subset U$ of the origin. The inverse mapping of course has the corresponding form

$$\boldsymbol{\phi}_r^{-1} \begin{pmatrix} \mathbf{x}_I \\ x_r \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_I \\ g_r(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix}$$

for some function $g_r(\mathbf{x})$; and since

$$\mathbf{x} = \boldsymbol{\phi}_r\big(\boldsymbol{\phi}_r^{-1}(\mathbf{x})\big) = \boldsymbol{\phi}_r \begin{pmatrix} \mathbf{x}_I \\ g_r(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_I \\ f_r\big(\boldsymbol{\phi}_r^{-1}(\mathbf{x})\big) \\ \mathbf{x}_{II} \end{pmatrix}$$

it follows that $f_r\big(\boldsymbol{\phi}_r^{-1}(\mathbf{x})\big) = x_r$ and therefore

$$\mathbf{f}\big(\boldsymbol{\phi}_r^{-1}(\mathbf{x})\big) = f \begin{pmatrix} \mathbf{x}_I \\ g_r(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_I \\ f_r\big(\boldsymbol{\phi}_r^{-1}(\mathbf{x})\big) \\ \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{x}_I \\ x_r \\ \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix},$$

so after further replacing the mapping $\mathbf{f}$ by the composition $\mathbf{f} \circ \boldsymbol{\phi}_r^{-1}$ it can be assumed that the mapping $\mathbf{f}$ is of type $r+1$. Altogether what has been demonstrated so far is that for any mapping $\mathbf{f}$ of type r there are a permutation matrix P_r and a $\mathcal{C}^1$ biholomorphic mapping $\boldsymbol{\phi}_r$ from an open neighborhood of the origin to an open neighborhood of the origin in $\mathbb{R}^n$ such that $P_r \circ \mathbf{f} \circ \boldsymbol{\phi}_r^{-1}$

is a mapping of type $r+1$. Repeating this observation shows that there are permutation matrices P_i and $\mathcal{C}^1$ local homeomorphisms $\boldsymbol{\phi}_i$ such that

$$P_n \circ P_{n-1} \circ \cdots \circ P_r \circ \mathbf{f} \circ \boldsymbol{\phi}_r^{-1} \circ \cdots \circ \boldsymbol{\phi}_{n-1}^{-1} \circ \boldsymbol{\phi}_n^{-1}$$

is a mapping of type $n+1$, hence is the identity mapping; consequently it follows that the mapping $\mathbf{f}$ can be written as the product of the permutation mapping $P_r^{-1} \circ \cdots P_{n-1}^{-1} \circ P_n^{-1}$ and the $\mathcal{C}^1$ local homeomorphism $\boldsymbol{\phi}_n \circ \boldsymbol{\phi}_{n-1} \circ \cdots \circ \boldsymbol{\phi}_r$, which suffices for the proof.

mpcmpcr **Corollary 3.1 (Decomposition Theorem)** *A $\mathcal{C}^1$ mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ defined in an open neighborhood U of the origin in $\mathbb{R}^n$ with a nonzero Jacobian at the origin can be written in any sufficiently small open subneighborhood $V \subset U$ of the origin as the composition of $\mathcal{C}^1$ homeomorphisms that are either translations or permutations of the variables in $\mathbb{R}^n$ or mappings that change only a single variable.*

Proof: If $\mathbf{f}(\mathbf{0}) = \mathbf{a}$ then the composition $\mathbf{g}(\mathbf{x}) = (T \circ \mathbf{f})(\mathbf{x})$ for the translation $T(\mathbf{x}) = \mathbf{x} - \mathbf{a}$ is a $\mathcal{C}^1$ mapping defined in the open neighborhood U of the origin that takes the origin to the origin and has a nonzero Jacobian at the origin; by the preceding theorem the homeomorphism $\mathbf{g}$ can be written in some open subneighborhood $V \subset U$ as a composition of $\mathcal{C}^1$ homeomorphisms that are either permutations of the variables or mappings that change only a single variable, so $\mathbf{f} = T^{-1} \circ \mathbf{g}$ is a composition of $\mathcal{C}^1$ homeomorphisms that are either translations or permutations of the variables in $\mathbb{R}^n$ or mappings that change only a single variable, and that suffices for the proof.

3.2 The Implicit Function Theorem

secimp

The Inverse Mapping Theorem can be applied to derive a simple local normal form at a point $\mathbf{a} \in \mathbb{R}^m$ for mappings $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ when $m > n$ and rank $\mathbf{f}'(\mathbf{a}) = n$, the maximal possible rank, showing that in suitable local coordinates near the point $\mathbf{a}$ the mapping is the simplest linear mapping between spaces of different dimensions; the general result will be discussed in Section 3.3, secrnk but a special case of particular importance and usefulness will be discussed in this section.

thimp1 **Theorem 3.4** *If $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ is a $\mathcal{C}^1$ mapping defined in an open neighborhood U of the origin $\mathbf{0} \in \mathbb{R}^m$ where $m > n$, and if $\mathbf{f}(\mathbf{0}) = \mathbf{0}$ and the $n \times n$ matrix consisting of the first n columns of the $n \times m$ matrix $\mathbf{f}'(\mathbf{0})$ is of rank n, then after shrinking the neighborhood U if necessary there is a $\mathcal{C}^1$ homeomorphism $\boldsymbol{\phi} : V \longrightarrow U$ between an open neighborhood V of the origin $\mathbf{0} \in \mathbb{R}^m$ and the open neighborhood U, where*

thimplq1 (3.12)
$$\boldsymbol{\phi}(t_1, \ldots, t_m) = \{\underbrace{g_1(\mathbf{t}), \ldots, g_n(\mathbf{t})}_{\mathbf{g}(\mathbf{t})}, t_{n+1}, \ldots, t_m\}$$

for some $\mathcal{C}^1$ functions $g_i(\mathbf{t})$ of the variables $\mathbf{t} \in V$, such that $g_i(\mathbf{0}) = 0$ and

thimplq2
$$(\mathbf{f} \circ \boldsymbol{\phi})(t_1 \ldots, t_m) = \{t_1, \ldots, t_n\}. \tag{3.13}$$

Proof: The assertion of the theorem can be summarized as the existence of the mapping $\boldsymbol{\phi}$ in the diagram

$$\begin{array}{ccccc} \mathbb{R}^m & & \mathbb{R}^m & & \mathbb{R}^n \\ \bigcup & & \bigcup & & \bigcup \\ V & \xrightarrow{\boldsymbol{\phi}} & U & \xrightarrow{\mathbf{f}} & \mathbb{R}^n \end{array}$$

with the properties (3.12) and (3.13). To simplify the notation write points $\mathbf{x} \in \mathbb{R}^m$ in the form

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_I \\ \mathbf{x}_{II} \end{pmatrix} \text{ where } \begin{cases} \mathbf{x}_I \in \mathbb{R}^n & I = (1, 2, \ldots, n) \\ \mathbf{x}_{II} \in \mathbb{R}^{m-n} & II = (n+1, n+2, \ldots, m). \end{cases}$$

The $n \times m$ matrix $\mathbf{f}'(\mathbf{x})$ can be written

$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} \partial_I \mathbf{f}(\mathbf{x}) & \partial_{II} \mathbf{f}(\mathbf{x}) \end{pmatrix}$$

in terms of the $n \times n$ matrix $\partial_I \mathbf{f}(\mathbf{x})$ having the entries $\Big(\partial_I\big(\mathbf{f}(\mathbf{x})\big)\Big)_{ij} = \partial_j f_i(\mathbf{x})$ for $1 \le\ i, j \le\ n$ and the $n \times (m-n)$ matrix $\partial_{II} \mathbf{f}(\mathbf{x})$ with entries $\Big(\partial_{II}\big(\mathbf{f}(\mathbf{x})\big)\Big)_{ij} = \partial_j f_i(\mathbf{x})$ for $1 \le i \le n$ and $n+1 \le j \le m$. Introduce the $\mathcal{C}^1$ mapping $\boldsymbol{\psi} : U \longrightarrow \mathbb{R}^m$ defined by

$$\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\psi}\begin{pmatrix} \mathbf{x}_I \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix}$$

and note that

$$\boldsymbol{\psi}'(\mathbf{x}) = \begin{pmatrix} \partial_I \mathbf{f}(\mathbf{x}) & \partial_{II} \mathbf{f}(\mathbf{x}) \\ \\ \partial_I \mathbf{x}_{II} & \partial_{II} \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \partial_I \mathbf{f}(\mathbf{x}) & \partial_{II} \mathbf{f}(\mathbf{x}) \\ \\ 0 & I_{m-n} \end{pmatrix}$$

where I_{m-n} is the identity $(m-n) \times (m-n)$ matrix; consequently $\det \boldsymbol{\psi}'(\mathbf{0}) = \det \partial_I \mathbf{f}(\mathbf{0}) \ne 0$, so it follows from the Inverse Mapping Theorem that after shrinking the neighborhood U if necessary the mapping $\boldsymbol{\psi}$ is a one-to-one mapping from U onto an open neighborhood V of the image $\boldsymbol{\psi}(\mathbf{0}) = \mathbf{0} \in \mathbb{R}^m$ with a $\mathcal{C}^1$ inverse $\boldsymbol{\phi} = \boldsymbol{\psi}^{-1}$, which of course has the form

$$\boldsymbol{\phi}(\mathbf{t}) = \begin{pmatrix} \mathbf{g}(\mathbf{t}) \\ \mathbf{t}_{II} \end{pmatrix}$$

for a mapping $\mathbf{g} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ for which $\mathbf{g}(\mathbf{0}) = \mathbf{0}$ since $\boldsymbol{\phi}(\mathbf{0}) = \mathbf{0}$; so $\boldsymbol{\phi}$ is a mapping of the form (3.12). Then

$$\begin{pmatrix} \mathbf{t}_I \\ \mathbf{t}_{II} \end{pmatrix} = (\boldsymbol{\psi} \circ \boldsymbol{\phi}) \begin{pmatrix} \mathbf{t}_I \\ \mathbf{t}_{II} \end{pmatrix} = \boldsymbol{\psi} \begin{pmatrix} \mathbf{g}(\mathbf{t}) \\ \mathbf{t}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{f}(\boldsymbol{\phi}(\mathbf{t})) \\ \mathbf{t}_{II} \end{pmatrix}$$

so $\mathbf{f}\big(\boldsymbol{\phi}(\mathbf{t})\big) = \mathbf{t}_I$, which demonstrates (3.13) and thereby concludes the proof.

thimpcr1 **Corollary 3.2** *If* $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ *is a* $\mathcal{C}^1$ *mapping defined in an open subset* $U \subset \mathbb{R}^m$ *where* $m > n$*, and if* $\operatorname{rank} \mathbf{f}'(\mathbf{a}) = n$ *at a point* $\mathbf{a} \in U$*, then there is a local coordinate system in an open neighborhood of the point* $\mathbf{a}$ *such that the mapping* $\mathbf{f} - \mathbf{f}(\mathbf{a})$ *is a linear mapping described by a matrix of rank* n *in terms of these coordinates.*

Proof: The mapping $\tilde{\mathbf{f}}(\mathbf{x}) = \mathbf{f}(\mathbf{x} + \mathbf{a}) - \mathbf{f}(\mathbf{a})$ is $\mathcal{C}^1$ in an open neighborhood of the origin, and rank $\tilde{\mathbf{f}}'(\mathbf{0}) = \operatorname{rank} \mathbf{f}'(\mathbf{a}) = n$ while $\tilde{\mathbf{f}}(\mathbf{0}) = \mathbf{0}$. By a permutation of the coordinates it can be supposed that the first n columns of the $n \times m$ matrix $\tilde{\mathbf{f}}'(\mathbf{0})$ are linearly independent. The mapping $\tilde{\mathbf{f}}$ thus satisfies the hypotheses of the preceding theorem, so after a further $\mathcal{C}^1$ change of coordinates near the origin it is a linear linear mapping of the form (3.13), a linear mapping of rank n; and since $\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{a}) = \tilde{\mathbf{f}}(\mathbf{x} - \mathbf{a})$ that suffices for the proof.

thimpcr2 **Corollary 3.3** *If* $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ *is a continuously differentiable mapping defined in an open subset* $U \subset \mathbb{R}^m$ *where* $m > n$*, and if* $\operatorname{rank} \mathbf{f}'(\mathbf{a}) = n$ *at a point* $\mathbf{a} \in U$*, then there is a local coordinate system in an open neighborhood* $U_\mathbf{a}$ *of the point* $\mathbf{a}$ *in terms of which the set* $\{\, \mathbf{x} \in U_\mathbf{a} \,\big|\, \mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) \,\}$ *is a linear subspace of dimension* $m - n$*.*

Proof: It follows from the preceding Corollary 3.2 that there is a local coordinate system in an open neighborhood $U_\mathbf{a}$ of the point $\mathbf{a}$ such that the mapping $\mathbf{f} - \mathbf{f}(\mathbf{a})$ is a linear mapping of rank n in terms of these coordinates; consequently the set of points in $U_\mathbf{a}$ at which the mapping $\mathbf{f} - \mathbf{f}(\mathbf{a})$ vanishes is a linear subspace of dimension $m - n$, which suffices for the proof.

The preceding corollaries involved only the assumption that $\det \mathbf{f}'(\mathbf{0}) \neq 0$; but one of the most important consequences of Theorem 3.4 involves the more precise assumption that a particular submatrix of $\mathbf{f}'(\mathbf{0})$ is nonsingular. The first $n \times n$ submatrix of the derivative $\mathbf{f}'(\mathbf{0})$ consists of the n columns $\partial_i \mathbf{f}(\mathbf{0})$ for $1 \le i \le n$; so the property of the matrix $\mathbf{f}'(\mathbf{0})$ in the hypothesis of Theorem 3.4 can be restated as the assumption that the matrix $\big(\partial_1 \mathbf{f}(\mathbf{0}) \quad \cdots \quad \partial_n \mathbf{f}(\mathbf{0})\big)$ is of its maximal rank n.

thimpcr3 **Theorem 3.5 (Implicit Function Theorem)** *If* $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ *is a* $\mathcal{C}^1$ *mapping defined in an open neighborhood* U *of the origin* $\mathbf{0} \in \mathbb{R}^m$ *where* $m > n$*, and if* $\mathbf{f}(\mathbf{0}) = \mathbf{0}$ *and* $\operatorname{rank} \big(\partial_1 \mathbf{f}(\mathbf{0}) \quad \cdots \quad \partial_n \mathbf{f}(\mathbf{0})\big) = n$*, then in a sufficiently small open cell* $\Delta = \Delta_I \times \Delta_{II} \subset U$ *containing the origin, written as the product of a cell* $\Delta_I \subset \mathbb{R}^n$ *in the space of the first* n *variables* $\mathbf{x}_I = \{x_1, \ldots x_n\}$ *and a cell* $\Delta_{II} \subset \mathbb{R}^{m-n}$ *in the space of the last* $m - n$ *variables* $\mathbf{x}_{II} = \{x_{n+1} \ldots, x_m\}$*, there is a mapping* $\mathbf{h} : \Delta_I \longrightarrow \Delta_{II}$ *such that* $\mathbf{f}(\mathbf{x}) = 0$ *for a point* $\mathbf{x} = \{\mathbf{x}_I, \mathbf{x}_{II}\}$ *in* $\Delta_I \times \Delta_{II}$ *if and only if* $\mathbf{x}_I = \mathbf{h}(\mathbf{x}_{II})$*.*

Proof: It follows from Theorem 3.4 that if the neighborhood U is sufficiently small there is a $\mathcal{C}^1$ homeomorphism $\boldsymbol{\phi} : V \longrightarrow U$ from an open neighborhood V of the origin in the space $\mathbb{R}^m$ of the variables $t_1, \ldots, t_m$ onto U such that (3.12) and (3.13) are satisfied. Any point $\mathbf{x} \in U$ is the image $\mathbf{x} = \boldsymbol{\phi}(\mathbf{t})$ of a unique point $\mathbf{t} \in V$, and it follows from (3.13) that $\mathbf{f}(\mathbf{x}) = 0$ if and only if $0 = (\mathbf{f} \circ \boldsymbol{\phi})(t) = \{t_1, \ldots, t_n\} = 0$, or equivalently if and only if $\mathbf{t}_I = 0$ when the point $\mathbf{t} \in \mathbb{R}^m$ is written $\mathbf{t} = \{\mathbf{t}_I, \mathbf{t}_{II}\}$ where $\mathbf{t}_I \in \mathbb{R}^n$ and $\mathbf{t}_{II} \in \mathbb{R}^{m-n}$; and in that case it follows from (3.12) that $\{\mathbf{x}_I, \mathbf{x}_{II}\} = \mathbf{x} = \boldsymbol{\phi}(\mathbf{t}) = \{\mathbf{g}(\mathbf{0}, \mathbf{t}_{II}), \mathbf{t}_{II}\}$, so that actually $\mathbf{t}_{II} = \mathbf{x}_{II}$ and consequently $\mathbf{x} = \{\mathbf{g}(\mathbf{0}, \mathbf{x}_{II}), \mathbf{x}_{II}\}$ or equivalently $\mathbf{x}_I = \mathbf{g}(\mathbf{0}, \mathbf{x}_{II})$, which suffices for the proof.

The argument in the proof of the preceding theorem is illustrated in the accompanying Figure 3.3. The zeros of the function $\mathbf{f}$ form the image of the linear subspace defined by the equations $t_1 = \cdots = t_n = 0$ under the $\mathcal{C}^1$ mapping $\boldsymbol{\phi} : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ that describes the change of coordinates in an open neighborhood of the point $\mathbf{a}$. The mapping $\boldsymbol{\phi}$ does not change the last $m - n$ coordinates but only the first n coordinates of the points; so the mapping preserves the "height" of points above the coordinate axis of the first n coordinates, but moves the points by changing their first n coordinates. Points on the linear subspace are described by the parameters $t_i = x_i$ for $n + 1 \le i \le m$ while the coordinates t_i are held constant for $1 \le i \le n$; the zeros of the function $\mathbf{f}$ are also described by the parameters $t_i = x_i$ for $n + 1 \le i \le m$.

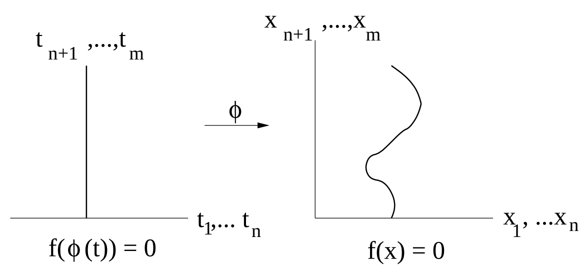

Figure 3.3: Illustration of the proof of the Implicit Mapping Theorem. impmap

In the Implicit Function Theorem, the equation $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ can be viewed as a description of relations among the coordinates $x_1, \ldots, x_m$ in $\mathbb{R}^m$, which implicitly describe some of the coordinates as functions of the remaining coordinates; the theorem provides conditions determining which coordinates really can be viewed as explicit functions of the remaining coordinates, at least locally. For example, if $f : U \longrightarrow \mathbb{R}^1$ is a continuously differentiable function in an open subset $U \subset \mathbb{R}^m$ and if $\partial_1 f(\mathbf{a}) \ne 0$ at some point $\mathbf{a} \in U$ then by the Implicit Function Theorem in an open cell $\Delta \subset U$ containing the point $\mathbf{a}$ there is a function $g(x_2, \ldots, x_n)$ such that $f(x_1, \ldots, x_m) = 0$ for a point $\{x_1, \ldots, x_m\} \in \Delta$ if and only if $x_1 = g(x_2, \ldots, x_m)$. The equation $f(x_1, x_2, \ldots x_m) = 0$ defines x_1 implicitly as a function of the remaining variables, and in an open neighborhood of any point $\mathbf{a}$ at which $\partial_1 f(\mathbf{a}) \ne 0$ this equation also defines x_1 explicitly as a function $x_1 = g(x_2, \ldots, x_m)$ of the remaining variables. It is obvious that some condition on the function f is necessary for this to be true; for instance without any condition the function $f(x_1, \ldots, x_m)$ might be independent of the variable x_1, in which case it can say nothing at all about that variable. The example in which $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - 1$ illustrates another way in which some condition on the function f is necessary. The equation $f(\mathbf{x}) = 0$ can be viewed as defining x_1 as the function $x_1 = \pm\sqrt{1 - x_2^2 - x_3^2}$, where the square root has two well defined locally continuous values so long as $x_2^2 + x_3^2 < 1$; but x_1 is not a well defined function in an open neighborhood of any point at which $x_2^2 + x_3^2 = 1$, or equivalently at which $\partial_1 f(x_1, x_2, x_3) = 0$. If the variable x_1 is defined explicitly by $x_1 = g(x_2, x_3)$ then $f\big(g(x_2, x_3), x_2, x_3\big) = 0$ identically in the variables x_2, x_3; and the chain rule can be applied to calculate the derivatives of the function $g(x_2, x_3)$, since

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_2} f\big(g(x_2, x_3), x_2, x_3\big) \\ &= \partial_1 f\big(g(x_2, x_3), x_2, x_3\big)\partial_1 g(x_2, x_3) + \partial_2 f\big(g(x_2, x_3), x_2, x_3\big). \end{aligned}$$

Thus the partial derivatives of g can be expressed in terms of the partial derivatives of f, and similarly for higher derivatives. Of course the same argument can be applied to another variable x_i, expressing it as a function of the remaining variables provided that $\partial_i f(\mathbf{x}) \ne 0$. If f_1, f_2 are continuously differentiable functions in in an open subset $U \subset \mathbb{R}^4$ and if

impex2 (3.14)
$$\det \begin{pmatrix} \partial_1 f_1(\mathbf{a}) & \partial_2 f_1(\mathbf{a}) \\ \partial_1 f_2(\mathbf{a}) & \partial_2 f_2(\mathbf{a}) \end{pmatrix} \ne 0$$

then the set of points at which $f(\mathbf{x}) = 0$ near $\mathbf{a}$ can be described explicitly as the set of points at which x_1 and x_2 are explicitly given functions of the remaining variables, say $x_1 = g_1(x_3, x_4)$ and $x_2 = g_2(x_3, x_4)$. Thus there are the identities $f_1\Big(g_1(x_3, x_4), g_2(x_3, x_4), x_3, x_4\Big) = f_2\Big(g_1(x_3, x_4), g_2(x_3, x_4), x_3, x_4\Big) = 0$; and the partial derivatives of these identities yield two linear equations relating the partial derivatives of the functions $g_1(x_3, x_4)$ and $g_2(x_3, x_4)$, which can be solved explicitly in view of (impex2 3.14).

The set $V = \{\, \mathbf{x} \in U \bigm| \mathbf{f}(x) = 0 \,\}$ of zeros of a mapping $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ defined in an open subset $U \subset \mathbb{R}^m$ is called the **zero locus** of the mapping $\mathbf{f}$. If the mapping is continuous its zero locus is a relatively closed subset of U, since it is the inverse image of the closed set $0 \in \mathbb{R}^n$ under a continuous mapping. A relatively closed subset $V \subset U$ of an open set $U \subset \mathbb{R}^n$ is called a **k-dimensional submanifold** of U if for any point $\mathbf{a} \in V$ there are an open neighborhood $U_\mathbf{a} \subset U$ and a continuous local change of coordinates in $U_\mathbf{a}$ such that $V \cap U_\mathbf{a}$ is a linear subspace of dimension k in terms of these coordinates. If the change of coordinates is described by a continuously differentiable mapping the subset V is said somewhat more precisely to be a $\mathcal{C}^1$ submanifold; similarly it is said to be a $\mathcal{C}^2$ submanifold if the change of coordinates is described by a $\mathcal{C}^2$ mapping, and so on. Corollary thimpcr2 3.3 can be interpreted as the assertion that if $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ is a $\mathcal{C}^1$ mapping from an open subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ for $m > n$ with the zero locus $V \subset U$, and if rank $\mathbf{f}'(\mathbf{x}) = n$ at each point $\mathbf{x} \in V$, then V is a $\mathcal{C}^1$ $(m - n)$-dimensional submanifold of U. It is quite common to rephrase this result just in terms of the coordinate functions $f_1, \ldots, f_n$ of the mapping **f** and to speak of the set of common zeros of these functions. In these terms, Corollary 3.3 asserts that if the gradient vectors $\nabla f_1, \ldots, \nabla f_n$ are linearly independent vectors at each point of the set V of the common zeros of these functions then V is a $\mathcal{C}^1$ submanifold of dimension $m - n$ in U. For example the circle

manexcr

$$V = \left\{ \, (x_1, x_2) \in \mathbb{R}^2 \; \middle| \; x_1^2 + x_2^2 = 1 \, \right\} \subset \mathbb{R}^2 \tag{3.15}$$

can be defined as the zero locus of the function $f(\mathbf{x}) = x_1^2 + x_2^2 - 1$ in the plane, and since $f'(\mathbf{x}) = \begin{pmatrix} 2x_1 & 2x_2 \end{pmatrix}$ is a matrix of rank 1 at each point $\mathbf{x} \in V$, at which not both $x_1 = 0$ and $x_2 = 0$, the zero locus V is a one-dimensional submanifold of the plane. Indeed it can be described locally as the linear space $r = 1$ in terms of polar coordinates (r, θ) in the plane in a neighborhood of each point of V; but of course there is no single coordinate neighborhood in which the entire circle can be described by the equation $r = 1$, since the other coordinate θ is not a single-valued function on the full circle. For some purposes an alternative characterization of submanifolds is quite convenient.

chsummadcr **Theorem 3.6** *A relatively closed subset $V \subset U$ of an open set $U \subset \mathbb{R}^m$ is an $(m-n)$-dimensional $\mathcal{C}^1$ submanifold of U if and only if for any point $\mathbf{a} \in V$, after a permutation of the coordinates in $\mathbb{R}^m$ if necessary, there is an open cell $\Delta_{\mathbf{a}} = \Delta_{I,\mathbf{a}} \times \Delta_{II,\mathbf{a}}$ centered at the point $\mathbf{a} \in V$, a product of cells $\Delta_{I,\mathbf{a}} \in \mathbb{R}^n$ and $\Delta_{II,\mathbf{a}} \in \mathbb{R}^{m-n}$, such that the intersection $V \cap \Delta_{\mathbf{a}}$ is described by the equations*

impthq2

$$V \cap \Delta_{\mathbf{a}} = \left\{ \, \mathbf{x} \in U_{\mathbf{a}} \; \middle| \; x_i = g_i(x_{n+1}, \ldots x_m), \;\; 1 \le i \le n \, \right\} \tag{3.16}$$

for some $\mathcal{C}^1$ functions g_i in the cell $\Delta_{II,\mathbf{a}} \in \mathbb{R}^k$.

Proof: If V is an $(m-n)$-dimensional $\mathcal{C}^1$ submanifold of U it follows from the Implicit Function Theorem that V can be described in an open neighborhood of any point $\mathbf{a} \in V$ by the equations (3.16). Conversely if V is defined by the equations (3.16) these equations can be rewritten $f_i(\mathbf{x}) = x_i - g_i(x_{n+1}, \ldots x_m) = 0$ for $1 \le i \le n$, and since $\nabla f_i(\mathbf{x})$ are clearly linearly independent vectors it follows that V is an $(m-n)$-dimensional $\mathcal{C}^1$ submanifold. That suffices for the proof.

The condition that a subset $V \subset U$ of an open set $U \subset \mathbb{R}^n$ be a submanifold is a local condition; so if a relatively closed subset V is a submanifold in a neighborhood of each of its points it is a submanifold in U. homsubman Two $\mathcal{C}^1$ submanifolds $V_1 \subset U_1, V_2 \subset U_2$ of open sets $U_1, U_2 \subset \mathbb{R}^m$ are called **homeomorphic $\mathcal{C}^1$ submanifolds** if there is a $\mathcal{C}^1$ homeomorphism $\boldsymbol{\phi} : U_1 \longrightarrow U_2$ such that $\boldsymbol{\phi}(V_1) = V_2$; this is clearly an equivalence relation that can be viewed as identifying two submanifolds that are essentially the same. The corresponding notion holds for topological submanifolds, for which the equivalence mapping is merely a homeomorphism, or for $\mathcal{C}^r$ submanifolds for $r > 1$. This equivalence relation really involves not just the point set V but the particular way in which that set is imbedded in the open set $U \subset \mathbb{R}^m$. For example a circle in $\mathbb{R}^3$ and a knotted curve in $\mathbb{R}^3$ are both submanifolds of $\mathbb{R}^3$; but they are not homeomorphic topological submanifolds, since there is no homeomorphism of $\mathbb{R}^3$ to itself that transforms a circle into a knotted curve[3].

The set of common zeros of a collection of functions $f_1, \ldots, f_n$ for which the gradients $\nabla f_1(\mathbf{x}), \ldots, \nabla f_n(\mathbf{x})$ are not linearly independent at some points is not necessarily a submanifold at these exceptional points but may have **singularities** of one sort or another. Of course if the gradients of these functions are linearly independent the squares $f_i(\mathbf{x})^2$ of these functions have the same set of common zeros as do the original functions, forming a submanifold; but the gradients of the squares of these functions vanish at all their common zeros. The study of singularities of differentiable mappings has been and still is a very active area of mathematical research[4].

3.3 The Rank Theorem

secrnk

The Inverse Mapping Theorem can be applied to describe mappings between spaces in situations more general than those to which the Implicit Function applies, in particular to describe mappings $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ where $m < n$.

[3]The study of knots is an old subject in mathematics, going back to work of P. G. Tait in the 1870's, but it is still a very active subject, with recent ties to mathematical physics and representation theory among other topics. Classical knot theory focuses on the classification of knots and the algebraic invariants that distinguish different knots. Further information can be found in the book by R. H. Crowell and R. H. Fox, *Introduction to Knot Theory*, Springer, 1963.

[4]Some classical examples of singularities are isolated singularities of algebraic curves, the topological properties of which were investigated by K. Brauner (Abh. Math. Sem. Hamburg, vol. 6 (1928), pages 8 - 54). A polynomial function $p(z_1, z_2)$ of two complex variables amounts to a pair of real functions of four real variables, the real and imaginary parts of the variables and of the function; so the set V of points at which $p(z_1, z_2) = 0$ is a candidate to be a submanifold of dimension 2 in $\mathbb{R}^4$. For a linear polynomial $p(z_1, z_2) = a_1 z_1 + a_2 z_2$ the set V is a linear submanifold, and the intersection $V \cap S^3_\epsilon$ of the set V and the sphere $S^3_\epsilon = \{\, (z_1, z_2) \,\big|\, |z_1|^2 + |z_2|^2 = \epsilon \,\}$ is a one-dimensional submanifold of the sphere, a linear slice of the sphere so just an ordinary circle. If the polynomial $p(z_1, z_2)$ has no constant term but a nontrivial linear term then again the set V is a 2-dimensional submanifold near the origin, and the intersection $V \cap S^3_\epsilon$ is a submanifold of the sphere that is $\mathcal{C}^1$ homeomorphic to a circle in the sphere. The set V of zeros of the polynomial $p(z_1, z_2) = z_1^2 + z_2^2$ is a 2-dimensional submanifold near the origin except at the origin itself, which is a singularity; actually this polynomial splits as the product $p(z_1, z_2) = (z_1 + i\, z_2)(z_1 - i\, z_2)$ of two linear functions both of which vanish at the origin, so near the origin the set V is the union of two linear spaces of dimension 2, the linear spaces $z_1 + i\, z_2 = 0$ and $z_1 - i\, z_2 = 0$, which meet only at a single point, the origin. In this case the intersection $V \cap S^3_\epsilon$ is a submanifold consisting of two disjoint circles which are not linked in the three-dimensional sphere. For the polynomial $p(z_1, z_2) = z_1^2 + z_2^3$ the origin is again a singularity, and in this case the intersection $V \cap S^3_\epsilon$ as a submanifold of the sphere is a nontrivial knot in the sphere, which is fairly easy to describe. In general the intersection $V \cap S^3_\epsilon$ is a submanifold of the sphere that is a knot of a special type, and its topological structure determines a good deal of the structure of the set V at its singular point. There are analogous topological sets that arise in the examination of isolated singularities of the sets of zeros of polynomials in several variables, a fascinating subject in its own right; these examples can be used to introduce the exotic differentiable structures on higher dimensional spheres, as discovered by J. W. Milnor. A more detailed discussion can be found in Milnor's book *Singular Points of Complex Hypersurfaces*, Annals of Mathematics Studies, number 61, Princeton University Press, 1968.

rankth **Theorem 3.7 (Rank Theorem)** *If* $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ *is a continuously differentiable mapping defined in an open subset* $U \subset \mathbb{R}^m$ *and if* $\operatorname{rank} \mathbf{f}'(\mathbf{x}) = r$ *for all points in an open neighborhood of a point* $\mathbf{a} \in U$ *then in suitable local coordinates near the points* $\mathbf{a} \in \mathbb{R}^m$ *and* $\mathbf{f}(\mathbf{a}) \in \mathbb{R}^n$ *the mapping* $\mathbf{f}$ *is a linear mapping of rank* r*.*

Proof: This theorem involves a change of local coordinates in both the domain and range of the mapping $\mathbf{f}$, so requires the existence at least locally of $\mathcal{C}^1$ homeomorphisms $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ in the diagram

$$\mathbb{R}^m \xrightarrow{\boldsymbol{\psi}} \mathbb{R}^m \xrightarrow{\mathbf{f}} \mathbb{R}^n \xrightarrow{\boldsymbol{\theta}} \mathbb{R}^n$$

so that the composition $\boldsymbol{\theta} \circ \mathbf{f} \circ \boldsymbol{\psi} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is a linear mapping of rank r. If $\operatorname{rank} \mathbf{f}'(\mathbf{a}) = r$ then by suitably relabeling the variables in the domain $\mathbb{R}^m$ and range $\mathbb{R}^n$ of the mapping $\mathbf{f}$ it can be assumed that the leading $r \times r$ submatrix of $\mathbf{f}'(\mathbf{a})$ is a nonsingular $r \times r$ matrix; and by continuity that is also the case for the matrix $\mathbf{f}'(\mathbf{x})$ for all points $\mathbf{x} \in U_{\mathbf{a}} \subset \mathbb{R}^m$ for a sufficiently small open neighborhood $U_{\mathbf{a}}$ of the point $\mathbf{a}$. As before it is convenient to write points $\mathbf{x} \in \mathbb{R}^m$ in the form $\mathbf{x} = \left(\begin{smallmatrix} \mathbf{x}_I \\ \mathbf{x}_{II} \end{smallmatrix}\right)$ where $\mathbf{x}_I \in \mathbb{R}^r$ and $\mathbf{x}_{II} \in \mathbb{R}^{m-r}$, and correspondingly to write points $\mathbf{y} \in \mathbb{R}^n$ in the form $\mathbf{y} = \left(\begin{smallmatrix} \mathbf{y}_I \\ \mathbf{y}_{II} \end{smallmatrix}\right)$ where $\mathbf{y}_I \in \mathbb{R}^r$ and $\mathbf{y}_{II} \in \mathbb{R}^{n-r}$. The mapping $\mathbf{f}$ can then be written

$$\mathbf{f}\begin{pmatrix} \mathbf{x}_I \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_I(\mathbf{x}) \\ \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix}$$

where $\mathbf{x}_I, \mathbf{f}_I(\mathbf{x}) \in \mathbb{R}^r$ while $\mathbf{x}_{II} \in \mathbb{R}^{m-r}$ and $\mathbf{f}_{II}(\mathbf{x}) \in \mathbb{R}^{n-r}$; and the matrix $\mathbf{f}'(\mathbf{x})$ then can be decomposed into the submatrices

$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} \partial_I \mathbf{f}_I(\mathbf{x}) & \partial_{II} \mathbf{f}_I(\mathbf{x}) \\ \partial_I \mathbf{f}_{II}(\mathbf{x}) & \partial_{II} \mathbf{f}_{II}(\mathbf{x}) \end{pmatrix}$$

where $\partial_I \mathbf{f}_I(\mathbf{x})$ is an $r \times r$ matrix which is nonsingular for all points $\mathbf{x} \in U_{\mathbf{a}}$. In these terms, introduce the mapping $\boldsymbol{\phi} : U_{\mathbf{a}} \longrightarrow \mathbb{R}^m$ defined by

$$\boldsymbol{\phi}\begin{pmatrix} \mathbf{x}_I \\ \mathbf{x}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_I(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix},$$

so that

$$\boldsymbol{\phi}'(\mathbf{x}) = \begin{pmatrix} \partial_I \mathbf{f}_I(\mathbf{x}) & \partial_{II} \mathbf{f}_I(\mathbf{x}) \\ 0 & I_{m-r} \end{pmatrix}$$

where I_{m-r} is the $(m-r) \times (m-r)$ identity matrix; since the matrix $\partial_I \mathbf{f}_I(\mathbf{x})$ is nonsingular for all points $\mathbf{x} \in U_{\mathbf{a}}$ it follows that the matrix $\boldsymbol{\phi}'(\mathbf{x})$ also is nonsingular for all points $\mathbf{x} \in U_{\mathbf{a}}$. If the neighborhood $U_{\mathbf{a}}$ is sufficiently small it follows from the Inverse Mapping Theorem that the restriction of the mapping $\boldsymbol{\phi}$ to U_a is a $\mathcal{C}^1$ homeomorphism $\boldsymbol{\phi} : U_{\mathbf{a}} \longrightarrow V$ between the open sets $U_{\mathbf{a}}$ and $V \subset \mathbb{R}^m$. Clearly the inverse mapping $\boldsymbol{\psi} : V \longrightarrow U_{\mathbf{a}}$ has the form

$$\boldsymbol{\psi}\begin{pmatrix} \mathbf{t}_I \\ \mathbf{t}_{II} \end{pmatrix} = \begin{pmatrix} \mathbf{g}(\mathbf{t}) \\ \mathbf{t}_{II} \end{pmatrix} \quad \text{for a } \mathcal{C}^1 \text{ mapping } \mathbf{g}_I : V \longrightarrow \mathbb{R}^r,$$

and then

$$\begin{pmatrix}\mathbf{t}_I\\ \mathbf{t}_{II}\end{pmatrix} = \mathbf{t} = \boldsymbol{\phi}\big(\boldsymbol{\psi}(\mathbf{t})\big) = \begin{pmatrix}\mathbf{f}_I\big(\boldsymbol{\psi}(\mathbf{t})\big)\\ \mathbf{t}_{II}\end{pmatrix} \quad \text{hence } \mathbf{f}_I\big(\boldsymbol{\psi}(\mathbf{t})\big) = \mathbf{t}_I;$$

consequently the composite mapping $\mathbf{h}(\mathbf{t}) = \big(\mathbf{f}\circ\boldsymbol{\psi}\big)(\mathbf{t})$ has the form

rkq00 (3.17)

$$\mathbf{h}(\mathbf{t}) = \mathbf{f}\big(\boldsymbol{\psi}(\mathbf{t})\big) = \begin{pmatrix}\mathbf{f}_I\big(\boldsymbol{\psi}(\mathbf{t})\big)\\ \mathbf{f}_{II}\big(\boldsymbol{\psi}(\mathbf{t})\big)\end{pmatrix} = \begin{pmatrix}\mathbf{t}_I\\ \mathbf{h}_{II}(\mathbf{t})\end{pmatrix} \quad \text{where } \mathbf{h}_{II}(\mathbf{t}) = \mathbf{f}_{II}\big(\boldsymbol{\psi}(\mathbf{t})\big) \in \mathbb{R}^{n-r},$$

hence

rkq1 (3.18)

$$\mathbf{h}'(\mathbf{t}) = \begin{pmatrix} I_r & 0\\ \partial_I\mathbf{h}_{II}(\mathbf{t}) & \partial_{II}\mathbf{h}_{II}(\mathbf{t})\end{pmatrix}$$

where I_r is the $r\times r$ identity matrix. On the other hand since $\mathbf{h} = \mathbf{f}\circ\boldsymbol{\psi}$ then by the chain rule $\mathbf{h}'(\mathbf{t}) = \mathbf{f}'\big(\boldsymbol{\psi}(\mathbf{t})\big)\cdot\boldsymbol{\psi}'(\mathbf{t})$; the $m\times m$ matrix $\boldsymbol{\psi}'(\mathbf{t})$ is nonsingular for all $\mathbf{t}\in V_{\mathbf{a}}$, while by hypothesis $\operatorname{rank}\mathbf{f}'\big(\boldsymbol{\psi}(\mathbf{t})\big) = r$ for all $\mathbf{t}\in V_{\mathbf{a}}$, and consequently $\operatorname{rank}\mathbf{h}'(\mathbf{t}) = r$ for all points $\mathbf{t}\in V_{\mathbf{a}}$ as well. However it is evident from (3.18) that $\operatorname{rank}\mathbf{h}'(\mathbf{t}) = r$ if and only if $\partial_{II}\mathbf{h}_{II}(\mathbf{t}) = 0$ for all points $\mathbf{t}\in U_{\mathbf{a}}$; so the mapping $\mathbf{h}_{II}$ is a function just of the variables $\mathbf{t}_I\in\mathbb{R}^r$. The mapping $\boldsymbol{\theta}:\mathbb{R}^n\longrightarrow\mathbb{R}^n$ defined by

$$\boldsymbol{\theta}\begin{pmatrix}\mathbf{y}_I\\ \mathbf{y}_{II}\end{pmatrix} = \begin{pmatrix}\mathbf{y}_I\\ \mathbf{y}_{II} - \mathbf{h}_{II}(\mathbf{y}_I)\end{pmatrix}$$

has the derivative

$$\boldsymbol{\theta}'(\mathbf{y}) = \begin{pmatrix} I_r & 0\\ -\partial_I\mathbf{h}_{II}(\mathbf{y}_I) & I_{n-r}\end{pmatrix},$$

so after restricting the neighborhood $U_{\mathbf{a}}$ further if necessary this mapping can be viewed as a local change of coordinates in that neighborhood by the Inverse Mapping Theorem; and it is evident from the definition of this mapping and the explicit form (3.17) of the mapping $\mathbf{h}$, in which $\mathbf{h}_{II}(t)$ is a function of the variables $\mathbf{t}_I$ alone, that

$$(\boldsymbol{\theta}\circ\mathbf{f}\circ\boldsymbol{\psi})(\mathbf{t}) = \boldsymbol{\theta}\big(\mathbf{h}(\mathbf{t})\big) = \boldsymbol{\theta}\begin{pmatrix}\mathbf{t}_I\\ \mathbf{h}_{II}(\mathbf{t})\end{pmatrix} = \begin{pmatrix}\mathbf{t}_I\\ \mathbf{h}_{II}(\mathbf{t}) - \mathbf{h}_{II}(\mathbf{t})\end{pmatrix} = \begin{pmatrix}\mathbf{t}_I\\ 0\end{pmatrix},$$

which is a linear mapping of rank r. That suffices for the proof.

rankthcra **Theorem 3.8** *If $\mathbf{f}: U\longrightarrow\mathbb{R}^n$ is a continuously differentiable mapping defined in an open subset $U\subset\mathbb{R}^m$ and if $\operatorname{rank}\mathbf{f}'(\mathbf{x}) = r$ for all points $\mathbf{x}\in U$ then*
(i) *the image $\mathbf{f}(U_{\mathbf{a}})$ of any sufficiently small open neighborhood $U_{\mathbf{a}}\subset U$ of a point $\mathbf{a}\in U$ is a $\mathcal{C}^1$ submanifold of dimension r of an open neighborhood of the point $\mathbf{f}(\mathbf{a})\in\mathbb{R}^n$;*
(ii) *for any point $\mathbf{b}\in\mathbf{f}(U)\subset\mathbb{R}^n$ the inverse image $\mathbf{f}^{-1}(\mathbf{b})$ is a $\mathcal{C}^1$ submanifold of dimension $m-r$ in U.*

Proof: It follows from the preceding theorem that after suitable changes of coordinates in open neighborhoods $U_{\mathbf{a}}$ of the point $\mathbf{a} \in U \subset \mathbb{R}^m$ and $V_{\mathbf{b}}$ of the point $\mathbf{b} = \mathbf{f}(\mathbf{a}) \in \mathbb{R}^n$ the mapping $\mathbf{f}$ is a linear mapping of rank r; consequently the image $\mathbf{f}(U_{\mathbf{a}})$ is a linear subspace of dimension r in terms of the new coordinates in $\mathbb{R}^n$ and the inverse image $\mathbf{f}^{-1}(\mathbf{y})$ of any point $\mathbf{y} \in \mathbf{f}(U_{\mathbf{a}}) \subset \mathbb{R}^m$ is a linear subspace of dimension $m - r$ in terms of the new coordinates in $\mathbb{R}^m$. Consequently the image $\mathbf{f}(U_{\mathbf{a}})$ is an r-dimensional submanifold of $V_{\mathbf{b}}$ in terms of the original coordinates in $\mathbb{R}^n$, and the inverse image $\mathbf{f}^{-1}(\mathbf{y}) \cap U_{\mathbf{a}}$ of any point $\mathbf{y} \in \mathbf{f}(U_{\mathbf{a}})$ is an $m-r$-dimensional $\mathcal{C}^1$ submanifold of $U_{\mathbf{a}}$ in terms of the original coordinates in $\mathbb{R}^m$. The latter statement holds for any point in $\mathbf{f}^{-1}(\mathbf{y})$; so since the inverse image $\mathbf{f}^{-1}(\mathbf{y}) \subset U$ is a relatively closed subset of U which is locally a submanifold of U it must be a submanifold of U, and that suffices for the proof.

Part (i) of the preceding theorem can be expected to hold only locally. For instance it is quite possible that the image of an open interval $I \subset \mathbb{R}^1$ under a continuously differentiable mapping $\boldsymbol{\phi} : I \longrightarrow \mathbb{R}^n$ is a curve that intersects itself at some points, so that although the image is locally a one-dimensional submanifold the full image has singularities; and the image of the open interval need not even be a closed subset of $\mathbb{R}^n$. The situation in higher dimensions of course is even more complicated; but the local result is of considerable use. For the Inverse Mapping Theorem and the Implicit Function Theorem the hypothesis only required that the mapping be of maximal rank at a point $\mathbf{a}$, or equivalently that there is a square submatrix of maximal size in the Jacobian matrix of the mapping so that the determinant of that matrix at the point $\mathbf{a}$ is nonzero; and since the Jacobian of a $\mathcal{C}^1$ mapping is a continuous function it is necessarily nonzero at all points near enough to $\mathbf{a}$. However if the rank of the Jacobian matrix is not maximal at a point $\mathbf{a}$ its rank at nearby points may well be greater at points arbitrarily close to $\mathbf{a}$. That is the reason that the hypothesis in the rank theorem requires that the rank be locally constant. In some applications of the Rank Theorem it is still the case that the rank is maximal; so if the rank is maximal at a point is maximal at all nearby points.

rankthcr0 **Theorem 3.9** *If $\boldsymbol{\phi} : U \longrightarrow \mathbb{R}^n$ is a continuously differentiable mapping from an open set $U \subset \mathbb{R}^r$ into $\mathbb{R}^n$ where $r < n$ and if $\operatorname{rank} \boldsymbol{\phi}'(\mathbf{x}) = r$ at some point $x \in U$ then the image under $\boldsymbol{\phi}$ of any sufficiently small open neighborhood of the point $\mathbf{a} \in U$ is a $\mathcal{C}^1$ submanifold of dimension r in an open neighborhood of the point $\boldsymbol{\phi}(\mathbf{a}) \in \mathbb{R}^n$.*

Proof: If rank $\boldsymbol{\phi}'(\mathbf{a}) = r$ at a point $\mathbf{a} \in U$ for a mapping $\boldsymbol{\phi} : \mathbb{R}^r \longrightarrow \mathbb{R}^n$ where $r < n$ then $\operatorname{rank} \boldsymbol{\phi}'(\mathbf{x}) =$ for all points $\mathbf{x}$ near $\mathbf{a}$, so the desired result follows immediately from the preceding theorem, and that suffices for the proof.

For example a mapping $\mathbf{f} : [0, 1] \longrightarrow \mathbb{R}^m$ can be viewed as a **parametrized curve** in the vector space $\mathbb{R}^m$, beginning at the point $\mathbf{f}(0)$ and ending at the point $\mathbf{f}(1)$; and if $\mathbf{f}'(t) \neq 0$ at each point $t \in [0, 1]$ then by Corollary 3.9 the image of an open neighborhood of any point $t_0 \in (0, 1)$ is a 1-dimensional submanifold of an open neighborhood of the point $\mathbf{f}(t_0) \in \mathbb{R}^m$. That the mapping $\mathbf{f}$ is differentiable at the point t_0 means that

$$\mathbf{f}(t) = \mathbf{f}(t_0) + \mathbf{f}'(t_0)(t - t_0) + \boldsymbol{\epsilon}(t) \quad \text{where} \quad \lim_{t \to t_0} \frac{\|\boldsymbol{\epsilon}(t)\|}{\|t - t_0\|} = 0$$

for all points t sufficiently near t_0. The parametrized curve

$$\mathbf{f}_0(t) = \mathbf{f}(t_0) + \mathbf{f}'(t_0)(t - t_0)$$

is a line in $\mathbb{R}^m$ passing through the point $\mathbf{f}(t_0)$, called the **tangent line** to the parametrized curve γ at the point $\mathbf{f}(t_0) = \mathbf{a}_0$, and its image is of course another submanifold of $\mathbb{R}^m$; the vector $\mathbf{f}'(t_0)$ is called the **tangent vector** to the parametrized curve $\mathbf{f}$ at the point $\mathbf{f}(t_0) = \mathbf{a}_0$. It follows from (2.1) and the uniqueness of the derivative that the tangent line to the parametrized curve $\mathbf{f}$ at the point $\mathbf{a}_0$ is the parametrized line through that point that is the best linear approximation to the $\mathbf{f}$ near that point. It is common to change the parametrization of the tangent line so that the point $\mathbf{a}$ corresponds to the parameter value $t = 0$, so to describe the tangent line as the parametrized curve $\mathbf{f}_0(t) = \mathbf{f}(t_0) + \mathbf{f}'(t_0)t$. A continuously differentiable mapping $\boldsymbol{\phi} : U \longrightarrow \mathbb{R}^n$ from an open subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ takes a parametrized curve $\mathbf{f}$ to the parametrized curve $\boldsymbol{\phi} \circ \mathbf{f}$, and by the chain rule the tangent vector to the image $\boldsymbol{\phi} \circ \mathbf{f}$ at the point $\boldsymbol{\phi}(\mathbf{a}_0)$ is the vector $(\boldsymbol{\phi} \circ \mathbf{f})'(t_0) = \boldsymbol{\phi}'(\mathbf{a}_0)\mathbf{f}'(t_0)$. Thus the derivative $\boldsymbol{\phi}'(\mathbf{a}_0)$ of the mapping $\boldsymbol{\phi}$ at the point $\mathbf{a}_0$ can be interpreted as the linear mapping that takes tangent vectors to parametrized curves through the point $\mathbf{a}_0$ to tangent vectors to the image parametrized curves through the point $\boldsymbol{\phi}(a_0)$.

If V is a $\mathcal{C}^1$ submanifold of an open subset $U \subset \mathbb{R}^m$ then through each point $\mathbf{a} \in V$ it is possible to pass a continuously differentiable parametrized curve contained entirely in the set V near that point; for in some local coordinates near the point $\mathbf{a}$ the subset V is a linear subspace, which contains straight lines through that point. Moreover if $\dim V = r$ there are parametrized curves through that point, the tangent vectors to which span a linear subspace of dimension r; for in some local coordinates near the point $\mathbf{a}$ the subset V is a linear subspace of dimension r, and there are r linearly independent straight lines through any point of a linear space of dimension r. For a submanifold $V \subset U$ of an open set $U \subset \mathbb{R}^m$ defined by equations

rnkdfv (3.19) $$V = \{\, \mathbf{x} \in U \bigm| f_1(\mathbf{x}) = \cdots = f_{m-r}(\mathbf{x}) = 0 \,\}$$

where the vectors $\nabla f_i(\mathbf{x})$ are linearly independent at each point $\mathbf{x} \in V$, the linear subspace of dimension r consisting of the tangent vectors at the point $\mathbf{a}$ of all continuously differentiable parametrized curves contained in V and passing through that point is called the **tangent space** to the submanifold V at the point $\mathbf{a}$ and is denoted by $\mathcal{T}_a(V)$. If $\boldsymbol{\phi} : I \longrightarrow V$ is a $\mathcal{C}^1$ parametrized curve for which $\boldsymbol{\phi}(0) = \mathbf{a}$, then since $f_i\big(\boldsymbol{\phi}(t)\big) = 0$ identically in t it follows from the chain rule that $f_i'(\mathbf{a})\boldsymbol{\phi}'(0) = 0$, or equivalently that $\nabla f_i(\mathbf{a}) \cdot \boldsymbol{\phi}'(0) = 0$; thus the tangent vector to the parametrized curve $\boldsymbol{\phi}$ is perpendicular to all of the gradient vectors

$\nabla f_i(\mathbf{a})$, hence the tangent space to V at the point $\mathbf{a} \in V$ is the linear subspace consisting of all points $\mathbf{x} \in \mathbb{R}^m$ such that the vector $\mathbf{x} - \mathbf{a}$ is perpendicular to the vectors $\nabla f_i(\mathbf{a})$. The equation of the tangent space therefore is

tp1
$$\mathcal{T}_a(V) = \Big\{ \mathbf{x} \in \mathbb{R}^m \ \Big|\ (\mathbf{x} - \mathbf{a}) \cdot \nabla f_i(\mathbf{a}) = 0 \ \text{ for } \ 1 \le i \le m - r \Big\}. \tag{3.20}$$

The tangent space is the linear subspace of $\mathbb{R}^n$ that is the best local approximation to the submanifold V near that point, just as is the tangent vector to a parametrized curve. For example, if $V \subset \mathbb{R}^3$ is the surface defined by

$$V = \Big\{ \mathbf{x} \in \mathbb{R}^3 \ \Big|\ f(\mathbf{x}) = x_1^2 + x_2^3 + x_3^4 - 3 = 0 \Big\}$$

then $f'(1,1,1) = \begin{pmatrix} 2 & 3 & 4 \end{pmatrix}$ so the equation of the tangent plane to V at the point $\{1,1,1\}$ is $(2,3,4) \cdot (x_1 - 1, x_2 - 1, x_3 - 1) = 0$ or equivalently $2x_1 + 3x_2 + 4x_3 = 9$.

To determine the extrema of a $\mathcal{C}^1$ function g defined on an r-dimensional submanifold $V \subset \mathbb{R}^m$ it is always possible theoretically to introduce local coordinates in an open neighborhood U of any point $\mathbf{a} \in V$ so that $V \cap U$ is a k-dimensional linear subspace in terms of those coordinates; the restriction of the function g to V then is locally just a function on a k-dimensional linear subspace of $\mathbb{R}^m$, so can be viewed as a function on an open subset of $\mathbb{R}^k$ and its critical points are candidates for points at which the function has an extremal value. However that can be a rather arduous calculation in all but the simplest situations, since determining the local coordinates in terms of which the submanifold is linear can be quite difficult to do explicitly. However the Implicit Function Theorem provides a much simpler and widely used method for finding local extrema on a submanifold. For the submanifold V defined by (3.19) where the vectors $\nabla f_i(\mathbf{x})$ are linearly independent at all points $\mathbf{x} \in V$, a point $\mathbf{a} \in V$ will be a local extremum of the restriction of the continuously differentiable function $g(\mathbf{x})$ to the submanifold V if and only if it is a local extrema of the restriction of the function $g(\mathbf{x})$ to any continuously differentiable parametrized curve lying in V and passing through the point $\mathbf{a}$. Then for any $\mathcal{C}^1$ parametrized curve $\phi : U \longrightarrow \mathbb{R}^m$ in an open neighborhood $U \subset \mathbb{R}^1$ of the origin $t = 0$ passing through the point $\mathbf{a}$, so that $\phi(0) = \mathbf{a}$, and contained in V, so that $f_i\big(\phi(t)\big) = 0$ for all $t \in U$, the origin $t = 0$ is a local extrema for the composite function $g\big(\phi(t)\big)$ and consequently

$$0 = \frac{d}{dt} g(\phi(t))\Big|_{t=0} = g'(\mathbf{a})\boldsymbol{\phi}'(0) = \nabla g(\mathbf{a}) \cdot \boldsymbol{\phi}'(0); .$$

thus $\nabla g(\mathbf{a})$ must be perpendicular to the tangent line to the parametrized curve, and since that is the case for all parametrized curves in V the vector $\nabla g(\mathbf{a})$ must be perpendicular to the tangent space $\mathcal{T}_a(V)$ to the submanifold V at the point $\mathbf{a}$. However the tangent space is characterized in (3.20) as the plane perpendicular to the gradients $\nabla f_i(\mathbf{a})$, so the vector $\nabla g(\mathbf{a})$ must be contained in the span of the vectors $\nabla f_i(\mathbf{a})$, consequently must be expressible as a linear combination

lagmq1

$$\nabla g(\mathbf{a}) = \sum_{i=1}^{m-r} \lambda_i \nabla f_i(\mathbf{a}) \tag{3.21}$$

for some parameters $\lambda_i \in \mathbb{R}$. This set of equations, combined with the equations $f_i(\mathbf{a}) = 0$ for the submanifold V, are an explicit set of equations that can determine the local extrema of the function $g(\mathbf{x})$, a technique called the method of **Lagrange multipliers**.

To apply this method to examine the extrema of a function g on a subset $V \subset \mathbb{R}^n$ defined by the vanishing of finitely many $\mathcal{C}^1$ functions f_i, the first step is to see whether the set V is a $\mathcal{C}^1$ submanifold; the method only provides candidates for extrema at points at which V is a submanifold, so if the set V has any singularities those points must be examined separately. The next step is to examine the equations (3.21) together with the equations $f_i(\mathbf{x}) = 0$, to find candidates for points at which the function g has local maxima or minima. It is often clear from the problem or can be determined fairly readily whether any of these points actually is a local maxima or minima. For instance, if M is a compact submanifold and there are only 2 candidates for local maxima and minima, one of these must be a global maximum and the other a global minimum, and which is determined simply by the values of the function g at these points. If it is not clear whether a point $\mathbf{a}$ is a local extrema or not, the manifold V can be parametrized in an open neighborhood of that point as in (3.16) in Theorem 3.6, and the restriction of the function g to that neighborhood is then a function of the variables $x_{m-r}, \ldots, x_m$ to which the usual test in terms of the Hessian of the function g in these coordinates can be tried, as discussed in Section 2.4; it is not necessary to determine the local parametrization (3.16) explicitly, of course, since the chain rule can be used to determine the second derivatives of the function g in terms of these variables.

As an example of the use of this method, find the minimum of $x + y + z$ subject to the restrictions that $\frac{a}{x} + \frac{b}{y} + \frac{c}{z} = 1$ where $x, y, z > 0$. The submanifold V is defined by the equation $f(x, y, z) = \frac{a}{x} + \frac{b}{y} + \frac{c}{z} - 1 = 0$, and the function of interest is $g(x, y, z) = x + y + z$, where

$$\nabla g = (1, 1, 1) \qquad \nabla f = \left(\frac{-a}{x^2}, \frac{-b}{y^2}, \frac{-c}{z^2} \right)$$

The Lagrange formula is the equation $\nabla g = \lambda \cdot \nabla f$; the first component of this vector equation is $\lambda = -x^2/a$, and upon substituting this into the other two components there result the equations

$$y = x\sqrt{\frac{b}{a}} \quad \text{and} \quad z = x\sqrt{\frac{c}{a}}$$

Since a, b and c are fixed there results a one-parameter set of solutions, parametrized by x. When restricted to the set $S = \{(x, y, z) : f(x, y, z) = 0, x > 0, y > 0, z >$

$0\}$ this is

$$\frac{a}{x} + \frac{b}{x\sqrt{b/a}} + \frac{c}{x\sqrt{c/a}} = \frac{\sqrt{a}}{x}\left(\sqrt{a}+\sqrt{b}+\sqrt{c}\right) = 1$$
$$\Rightarrow (x,y,z) = \left(\sqrt{a}+\sqrt{b}+\sqrt{c}\right)\left(\sqrt{a},\sqrt{b},\sqrt{c}\right).$$

Thus the extremum associated to this point is $f(x,y,z) = \left(\sqrt{a}+\sqrt{b}+\sqrt{c}\right)^2$. To see that this is the minimum, the set S is noncompact but the function g must have a minimum on any compact subset of S. The function g increases whenever one of the coordinates x, y, or z becomes very large. Since S is contained entirely in the first octant, it is only necessary to minimize f on a reasonable compact subset near the origin, which is the solution found.

When finding the extrema of a continuously differentiable function $g(\mathbf{x})$ on a set such as the closed semidisc

$$D = \left\{ \mathbf{x} \in \mathbb{R}^2 \;\middle|\; x_1^2 + x_2^2 \le 1, x_2 \ge 0 \right\}$$

in the plane, the procedure is first to examine critical points in the interior of the semidisc, then to examine the two pieces of the boundary of the semidisc that are local submanifolds, the semicircle $x_1^2 + x_2^2 = 1,\ x_2 \ge 0$ and the axis $-1 \le x_1 \le 1,\ x_2 = 0$, and then the two remaining points $(-1,0)$ and $(1,0)$ that are points where the boundary fails to be a submanifold. In general, the procedure is to examine critical points in the interior of the region, and then to examine the pieces of the boundary that are submanifolds and apply the method of Lagrange multipliers in all these cases.

If $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ is a continuously mapping defined in an open subset $U \subset \mathbb{R}^m$ where $m > n$ then $\operatorname{rank} \mathbf{f}'(\mathbf{a}) \le n$ at any point $\mathbf{a} \in U$; and if $\operatorname{rank} \mathbf{f}'(\mathbf{x}) = n$ at a point $\mathbf{x} = \mathbf{a} \in U$ then it is also the case that $\operatorname{rank} \mathbf{f}'(\mathbf{x}) = n$ for all points $\mathbf{x} \in U_{\mathbf{a}}$ for an open neighborhood V of the point $\mathbf{a}$, since if some $n \times n$ submatrix of $\mathbf{f}'(\mathbf{a})$ has a nonzero determinant then by continuity that submatrix of $\mathbf{f}'(\mathbf{x})$ will also have a nonzero determinant for all points $\mathbf{x}$ near enough $\mathbf{a}$. The Rank Theorem asserts that the mapping $\mathbf{f}$ is equivalent to a linear mapping of rank n in an open neighborhood of each point $\mathbf{a} \in V$, and consequently the image of the mapping $\mathbf{f}$ is a full open neighborhood of the point $\mathbf{f}(\mathbf{a})$. In particular the image of the mapping $\mathbf{f}$ is not contained within any proper linear subspace of the image space $\mathbb{R}^n$, so the coordinate functions f_i of the mapping $\mathbf{f}$ are linearly independent functions at each point $\mathbf{a} \in U$. On the other hand if $\operatorname{rank} \mathbf{f}'(\mathbf{x}) = r$ at all points $\mathbf{x} \in U$ where $r < n$, then by the Rank Theorem the mapping $\mathbf{f}$ will be a linear mapping of rank r in suitable local coordinates in open neighborhoods of each point $\mathbf{a} \in U \subset \mathbb{R}^m$ and its image $\mathbf{f}(\mathbf{a}) \in \mathbb{R}^n$; that means that the image of the restriction of the mapping $\mathbf{f}$ to an open neighborhood $V_{\mathbf{a}}$ of any point $\mathbf{a} \in U$ is contained in a submanifold of dimension r in an open neighborhood $V_{\mathbf{f}(\mathbf{a})} \subset \mathbb{R}^n$ of the image point $\mathbf{f}(\mathbf{a})$. This submanifold is described as the set of common zeros of $n - r$ linear functions in some local coordinates near the point $\mathbf{f}(\mathbf{a})$, and these functions will at least be some continuously differentiable functions

h_i near the point $\mathbf{f}(\mathbf{a})$ in the initial coordinates in $\mathbb{R}^n$. Therefore the component functions f_i of the mapping $\mathbf{f}$ satisfy a collection of $n-r$ nontrivial equations $h_\nu\big(f_1(\mathbf{x}), f_2(\mathbf{x})\ldots, f_n(\mathbf{x})\big) = 0$ for $1 \le \nu \le n-r$; that is usually expressed by saying that the functions f_i are **functionally dependent**.

The preceding discussion really focused only on local properties of submanifolds of open subsets $U \subset \mathbb{R}^n$; the definition and characterization of submanifolds are both local conditions. There is also an extensive global theory of submanifolds, and of a weaker equivalence class of submanifolds known as manifolds. Although these topics will not be pursued here, it may be of some interest at least to see the basic concepts involved. If V is a topological submanifold of an open subset $U \subset \mathbb{R}^m$ it has the relative topology induced by the topology of U, in which the open subsets of V are the intersections of V with the open subsets of U. Continuous functions on V can be defined as usual in terms of the open subsets of V. If $V_1 \subset U_1$ and $V_2 \subset U_2$ are topological submanifolds of open subsets $U_1 \subset \mathbb{R}^{n_1}$ and $U_2 \subset \mathbb{R}^{n_2}$ a mapping $\boldsymbol{\phi} : V_1 \longrightarrow V_2$ is continuous in the relative topologies if and only if $\boldsymbol{\phi}^{-1}(W) \subset V_1$ is an open subset of V_1 whenever $W \subset V_2$ is an open subset of V_2; and the submanifolds V_1 and V_2 are **weakly homeomorphic** if and only if there is a continuous one-to-one mapping $\boldsymbol{\phi} : V_1 \longrightarrow V_2$ with a continuous inverse. This must be distinguished from the notion that the two submanifolds V_1 and V_2 are homeomorphic, as defined on page 56; that refers to the way in which the submanifolds are imbedded in the open subsets U_i, so the submanifolds are equivalent if and only if there is a homeomorphism $\boldsymbol{\psi} : U_1 \longrightarrow U_2$ such that $\boldsymbol{\psi}(V_1) = V_2$. Two submanifolds that are weakly homeomorphic need not be submanifolds of subsets U_i in spaces $\mathbb{R}^{m_i}$ of the same dimension. For instance, a circle in the plane and a knotted curve in space are generally weakly homeomorphic, when the knotted curve is described as the image of a one-to-one $\mathcal{C}^1$ mapping of the circle into $\mathbb{R}^3$ with a nontrivial derivative; but they are certainly not homeomorphic as submanifolds, since they are imbedded in spaces of different dimensions. Weak homeomorphism of submanifolds is an equivalence relation in the customary sense; and a weak equivalence class is called a **manifold**, in this case a topological manifold. A given submanifold $V \subset U$ in an open subset $U \subset \mathbb{R}^m$ thus determines a manifold, an equivalence class of submanifolds under weak equivalence that includes the submanifold V. It is perhaps more common to speak of submanifolds being **equivalent as manifolds**, rather than as weakly equivalent submanifolds. [5]

[5] As might be expected, it is also possible to define manifolds even more intrinsically, without considering their embeddings. If $V \subset U$ is a k-dimensional topological submanifold of an open set $U \subset \mathbb{R}^m$ then by definition each point $\mathbf{a} \in V$ is contained in an open subset $U_\mathbf{a} \subset \mathbb{R}^m$ for which there is a homeomorphism $\boldsymbol{\phi}_\mathbf{a} : U_\mathbf{a} \longrightarrow W_\mathbf{a}$ between the open subset $U_\mathbf{a} \subset \mathbb{R}^n$ and another open subset $W_\mathbf{a} \subset \mathbb{R}^m$ such that $\boldsymbol{\phi}_\mathbf{a}(V \cap U_\mathbf{a})$ is a linear subspace $L_\mathbf{a} \subset W_\mathbf{a}$, which can be viewed as an open subset $L_\mathbf{a} \subset \mathbb{R}^k$; the restriction $\boldsymbol{\phi}_\mathbf{a} : V \cap U_\mathbf{a} \longrightarrow L_\mathbf{a}$ is a homeomorphism between the relatively open subset $V_\mathbf{a} = V \cap U_\mathbf{a} \subset V$ and the open subset $L_\mathbf{a} \subset \mathbb{R}^k$, which can be considered as describing a system of local coordinates in the relatively open neighborhood $V_\mathbf{a}$ of the point $\mathbf{a} \in V$, the local coordinates in the open subset $L_\mathbf{a} \subset \mathbb{R}^k$. The corresponding construction can be carried out for any other topological submanifold $V' \subset U'$ that is homeomorphic to V through a homeomorphism $\boldsymbol{\psi} : U \longrightarrow U'$ such that $\boldsymbol{\psi}(V) = V'$; if $\mathbf{a}' = \boldsymbol{\psi}(\mathbf{a})$ there is an open subset $U'_{\mathbf{a}'}$ with a homeomorphism $\boldsymbol{\phi}'_\mathbf{a} : U'_{\mathbf{a}'} \longrightarrow W'_{\mathbf{a}'}$ such that

PROBLEMS: GROUP 1

(1) Find the Jacobian of the mapping $\mathbf{f} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ defined by

$$\mathbf{f}(\mathbf{x}) = (e^{x_1}\cos x_2, e^{x_1}\sin x_2).$$

---

the restriction $\boldsymbol{\phi}'_{\mathbf{a}'} : V' \cap U'_{\mathbf{a}'} \longrightarrow L'_{\mathbf{a}'}$ is a homeomorphism between the relatively open subset $V'_{\mathbf{a}'} = V' \cap U'_{\mathbf{a}'} \subset V'$ and the open subset $L'_{\mathbf{a}'} \subset \mathbb{R}^k$, so the composition $\boldsymbol{\phi}'_{\mathbf{a}'} \circ \boldsymbol{\psi} : V_{\mathbf{a}} \longrightarrow L'_{\mathbf{a}'}$ can be considered as describing another system of local coordinate in the relatively open neighborhood $V_{\mathbf{a}}$ of the point $\mathbf{a} \in V$. These two sets of local coordinates near $\mathbf{a}$ are related by the homeomorphism $\boldsymbol{\phi}'_{\mathbf{a}'} \circ \boldsymbol{\psi} \circ (\boldsymbol{\phi}_{\mathbf{a}} \big| V_{\mathbf{a}}) : L_{\mathbf{a}} \longrightarrow L'_{\mathbf{a}'}$. Thus the system of local coordinates in neighborhoods of a point $\mathbf{a} \in V$ provided by equivalent topological k-dimensional submanifolds of $\mathbb{R}^m$ are all homeomorphic to one another. In particular if $V_{\mathbf{a}}$ and $V_{\mathbf{b}}$ are any two overlapping systems of local coordinates in V then they determine homeomorphic systems of local coordinates in an open neighborhood of any point $\mathbf{c} \in V_{\mathbf{a}} \cap V_{\mathbf{b}}$; thus the manifold V can be described by a collection of relatively open coordinate neighborhoods $V_{\mathbf{a}} \subset V$ which determine equivalent coordinate neighborhoods in any intersection $U_{\mathbf{a}} \cap U_{\mathbf{b}}$. Weak homeomorphisms between two submanifolds representing the same manifold can be described entirely in terms of these local coordinate neighborhoods. The same construction can be carried out for $\mathcal{C}^1$ submanifolds, providing systems of local coordinates that are related by $\mathcal{C}^1$ homeomorphisms in the intersections of coordinate neighborhoods, and of course also for $\mathcal{C}^r$ submanifolds for any r. The notion of a manifold can be extended to those defined more abstractly just by coverings of a topological space V by coordinate neighborhoods. Traditionally a k-dimensional **manifold** is defined as a topological space M such that each point $p \in M$ has an open neighborhood $U_p \subset M$ with a one-to-one continuous mapping $\mathbf{f}_p : U_p \longrightarrow V_p$ onto an open subset $V_p \subset \mathbb{R}^k$; the sets U_p are called **coordinate neighborhoods** on the manifold M, and the parameters in the image set $V_p \subset \mathbb{R}^k$ are called **local coordinates** on the manifold M in U_p. Two k-dimensional manifolds M and M' are **homeomorphic** if there is a continuous invertible one-to-one mapping $\boldsymbol{\phi} : M \longrightarrow M'$; this is an equivalence relation in the traditional sense, identifying manifolds that are topologically the same. It is customary to impose at least some restrictions on the topology of the set M to avoid odd and irregular cases.For example the real axis together with one additional point $0'$, where an open neighborhood of the zero 0 is defined as the set of real numbers $\{\, x \in \mathbb{R} \mid |x| \le \epsilon \,\}$ and an open neighborhood of the additional point $0'$ is defined to be that point together with the set of real numbers $\{\, x \in \mathbb{R} \mid 0 < |x| \le \epsilon \,\}$ is a one-dimensional manifold; but the points 0 and $0'$ do not have disjoint open neighborhoods. The usual regularity condition is that the topology is a **Hausdorff** topology, defined as a topology such that any two points of M have disjoint open neighborhoods. It is possible to speak of continuous functions on a manifold, since the manifold has a topology, but it is not possible to speak of differentiable functions; local coordinate neighborhoods look like pieces of $\mathbb{R}^k$, so that the notion of a function being differentiable in each of those coordinate neighborhoods makes sense, but the notions of differentiability in the two coordinate neighborhoods may be quite different, since it is only required that the mappings $\mathbf{f}_p : U_p \longrightarrow V_p$ be continuous. However if the manifold M has the additional property that in any intersection $U_p \cap U_q$ of coordinate neighborhoods of M the composition

$$\mathbf{f}_p^{-1} \circ \mathbf{f}_q : \mathbf{f}_q(U_p \cap U_q) \longrightarrow \mathbf{f}_p(U_p \cap U_q)$$

is a $\mathcal{C}^r$ mapping between the open subsets $\mathbf{f}_q(U_p \cap U_q) \subset \mathbb{R}^k$ and $\mathbf{f}_q(U_p \cap U_q) \subset \mathbb{R}^k$, then the notion of a function being $\mathcal{C}^r$ is well defined; a manifold M with this property is called a **differentiable manifold** of class $\mathcal{C}^r$, or just a $\mathcal{C}^r$ manifold for short. The Whitney Embedding Theorem was proved by H. Whitney, *Annals of Mathematics*, vol. 37 (1936), pages 645-680 and is discussed in a number of articles and books, for example in the article by Howard Jacobowitz on "The global isometric embedding problem" in Contemp. Math.,vol. 332, Amer. Math. Soc., Providence, RI, 2003, pages 131-137. That theorem shows that any abstractly defined manifold can be realized by a submanifold of some space $\mathbb{R}^n$, so that it is possible to define manifolds either as equivalence classes of submanifolds or as abstractly defined manifolds

Find formulas for the local inverse of this mapping at those points at which it exists. Does this mapping have a global inverse $\mathbf{g} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$?

(2) Consider the plane curve defined by the equation

$$f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2 = 0.$$

(i) Show that the portion of this curve in the half-plane $x_2 > 0$ is a submanifold.
(ii) Find the tangent vector to this curve at the point $(\frac{3}{2}, \frac{3}{2})$ and the equation of the tangent line to the curve at this point.

(3) Find the tangent vector to the curve $C \subset \mathbb{R}^3$ defined parametrically by

$$\mathbf{f}(t) = \{e^{(t-1)(t-2)}, t^2 - 4, t^2 - 1\}$$

at the point on the curve corresponding to the parameter value $t = 1$, and the equation of the tangent line to the curve C at that point.

(4) Find a normal vector and the equation of the tangent plane to the surface $x_1^2x_2 + x_2x_3 + x_4 = 3$ at the point $(1, 1, 1, 1)$.

(5) Find all points on the paraboloid $x_1^2 + x_2^2 - x_3^2 = 1$ where the tangent plane is parallel to the plane $x_1 + x_2 + x_3 = 1$, and find the equations of the tangent planes at these points.

(6) Find the maximal and minimal values of the function

$$f(\mathbf{x}) = 3x_1^2 - 2x_2^2 + 2x_2$$

on the closed unit disc $\{\ \mathbf{x} \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \le 1\ \}$, by the method of Lagrange multipliers, and determine the points at which these extreme values are taken.

(7) If $\mathbf{f} : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ for $m < n$ is a $\mathcal{C}^1$ mapping in an open neighborhood of a point $\mathbf{a} \in \mathbb{R}^m$ for which rank $\mathbf{f}'(\mathbf{a}) = m$, show that the image of an open neighborhood of the point $\mathbf{a} \in \mathbb{R}^m$ is a submanifold V of dimension m in an open neighborhood of the image $\mathbf{b} = \mathbf{f}(a) \in \mathbb{R}^n$. Show further that the tangent space to V at the point $\mathbf{b}$ is the image of the linear mapping $\mathbf{g} : \mathbb{R}^m \longrightarrow R^n$ defined by $\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{a}) + \mathbf{f}'(\mathbf{b})\ (\mathbf{x} - \mathbf{a})$.

PROBLEMS: GROUP 2

(8) Show that any 2×2 matrix A sufficiently close to the identity matrix has a unique square root near the identity matrix, a solution X of the equation $X^2 = A$ that is also near the identity matrix. This solution can be viewed as a function $\sqrt{A}$ of the 2×2 matrices A near the identity matrix I for which $\sqrt{I} = I$. What is the derivative of this function $\sqrt{A}$ in a neighborhood of the identity matrix I?

(9) Consider the function

$$f(x_1, x_2, x_3) = x_1^2 + x_1x_2 + x_2^2 + x_2x_3 + x_3^2.$$

(i) Find by using the method of Lagrange multipliers the maximum and minimum values of this function on the spherical surface S defined by $x_1^2+x_2^2+x_3^2 = 1$, and the points at which these extreme values are attained.
(ii) Find by using the method of Lagrange multipliers the maximum and minimum values of the function f on the intersection $S \cap T$ of the spherical surface S and the plane T defined by $c_1x_1 + c_2x_2 + c_3x_3 = 0$, where $(c_1, c_2, c_3) \in S$ is a point at which the function f takes its maximum value on S.

(10) Find by using the method of Lagrange multipliers all local maxima and minima of the function $f(x_1, x_2, x_3) = x_1^2+x_2^2+x_3^2$ on the surface $x_1^2+4x_2^2-2x_3^2 = 1$, and the points at which these local extemal values are taken.

(11) Find sufficient conditions on a $\mathcal{C}^1$ function $f(x_1, x_2)$ defined in an open neighborhood of the origin in $\mathbb{R}^2$ and vanishing at the origin so that there exists a $\mathcal{C}^1$ function $g(x)$ in an open neighborhood of the origin $0 \in \mathbb{R}^1$ such that $g(0) = 0$, $g'(0) \neq 0$ and $f\big(f(x, g(x)), g(x)\big) = 0$ for all values x near 0.

(12) Consider the surface S in $\mathbb{R}^3$ defined by the equation

$$x_1^2 - x_2^3 + 3x_1x_2x_3 = -1.$$

(i) Find the equation of the tangent plane to S at the point $(2, -1, 1)$.
(ii) Find the points on S at which the normal to S is a vector parallel to the vector $(1, 1, 0)$.
(iii) Show that S is a submanifold of $\mathbb{R}^3$.

(13) Show that the two equations

$$\sin(x_1 + 2x_2) + \cos(3x_3 + 4x_4) = 1, \quad e^{x_1+x_3} + e^{x_2+x_4} = 2$$

define x_1 and x_3 as functions of x_2 and x_4 near the origin $\mathbf{0} = (0, 0, 0, 0)$, and determine the value of $\dfrac{\partial x_1}{\partial x_2}$ at that point. Show that these two equations also define x_2 and x_4 as functions of x_1 and x_3 near the origin, and determine the value of $\dfrac{\partial x_2}{\partial x_1}$.

(14) Consider the functions in $\mathbb{R}^3$ defined by the equations

$$\begin{aligned} f_1(\mathbf{x}) &= e^{x_1} + e^{x_2} + e^{x_3} + x_1x_2x_3 - 3 \\ f_2(\mathbf{x}) &= x_1 + x_2^2 + x_3^3 + \sin\left(e^{x_1x_2x_3} - 1\right). \end{aligned}$$

(i) Show that the equations $f_1(\mathbf{x}) = f_2(\mathbf{x}) = 0$ define a one-dimensional submanifold V in an open neighborhood of the origin.
(ii) Show that the submanifold V can be described explicitly in an open neighborhood of the origin by equations of the form

$$x_1 = g_1(x_3) \quad \text{and} \quad x_2 = g_2(x_3),$$

for some $\mathcal{C}^1$ functions $g_1(x_3), g_2(x_3)$ in a neighborhood of the origin in $\mathbb{R}^1$ such that $g_1(0) = 0$ and $g_2(0) = 0$, and find the derivatives $g_1'(0)$ and $g_2'(0)$.
(iii) Show that the equation $f_2(\mathbf{x}) = 0$ defines a two-dimensional submanifold W in an open neighborhood of the origin, and find the tangent space to that submanifold at the origin.
(iv) In which directions are the functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ increasing most rapidly at the origin?
(v) If $f_3(\mathbf{x}) = \sin(x_1 + 2x_2 + 3x_3)$, show that the mapping $\mathbf{f} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ defined by $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})\}$ is a homeomorphism from an open neighborhood of the origin to another open neighborhood of the origin, and find the Jacobian at the origin of its inverse mapping.

(15) Show that the union of the two positive coordinate axes in $\mathbb{R}^2$, the set $\{\, (x_1, x_2) \in \mathbb{R}^2 \mid x_1x_2 = 0,\ x_1 \ge 0,\ x_2 \ge 0 \,\}$, is a continuous submanifold of $\mathbb{R}^2$ but is not a $\mathcal{C}^1$ submanifold.

(16) Suppose that S_1 and S_2 are two smooth surfaces in $\mathbb{R}^3$ (2-dimensional submanifolds) and that λ is a line segment from a point $\mathbf{a}_1 \in S^1$ to $\mathbf{a}_2 \in S_2$ that has the minimal length among all line segments joining points of these two surfaces. Show that the line λ is perpendicular to the tangent planes to the surfaces S_1 and S_2 at the points $\mathbf{a}_1$ and $\mathbf{a}_2$.

(17) Suppose that M is a compact two-dimensional submanifold in $\mathbb{R}^3$, such as the surface of a sphere or doughnut for example. Show that for every 2-dimensional linear subspace $L \subset \mathbb{R}^3$ there is a point $a \in M$ such that the tangent space to M at the point a is the plane $a + L$, the translate of the linear subspace L to the point $a \in M$.

Chapter 4

Integration

math218notes2

4.1 The Riemann Integral

riem

The definition of the integral of a function over a cell $\overline{\Delta} \subset \mathbb{R}^n$ is a straightforward extension of the definition of the integral of a function over an interval $I \subset \mathbb{R}^1$. The **content** of a closed cell

$$\overline{\Delta} = \left\{ \mathbf{x} \in \mathbb{R}^n \,\middle|\, a_j \le x_j \le b_j \ \text{ for } \ 1 \le j \le n \right\}$$

is defined by

intclct

$$|\overline{\Delta}| = \prod_{j=1}^{n} (b_j - a_j), \tag{4.1}$$

and is also considered as the content $|\Delta|$ of the corresponding open cell Δ. The content of a cell or interval in $\mathbb{R}^1$ is its length, the content of a cell in $\mathbb{R}^2$ is its area, and the content of a cell in $\mathbb{R}^3$ is its volume; so the content of a cell is just a generalization of the area or volume to arbitrary dimensions. The cell can be decomposed as a union of smaller closed cells by adjoining finitely many points of subdivision

$$a_j = c_{j,0} < c_{j,1} < \cdots < c_{j,n_j} = b_j$$

of each of its sides and writing $\overline{\Delta}$ as the union of the closed subcells

$$\left\{ \mathbf{x} \in \mathbb{R}^n \,\middle|\, c_{j,k-1} \le x_j \le c_{j,k} \ \text{ for } \ 1 \le j \le n, 1 \le k \le n_j \right\};$$

such a decomposition $\overline{\Delta} = \bigcup_i \overline{\Delta}_i$ of the cell is called a **partition** $\mathcal{P}$ of the cell. The subcells of the partition are not disjoint, but only have portions of their boundaries in common. For the case $n = 1$ this is the decomposition of an interval, a cell in $\mathbb{R}^1$, as a union of subintervals; and for $n = 2$ it is illustrated in the accompanying Figure partition 4.1. It is clear that $|\overline{\Delta}| = \sum_i |\overline{\Delta}_i|$ for any partition of a cell $\overline{\Delta}$. If f is a bounded function on a cell $\overline{\Delta} \subset \mathbb{R}^n$, its least upper bound

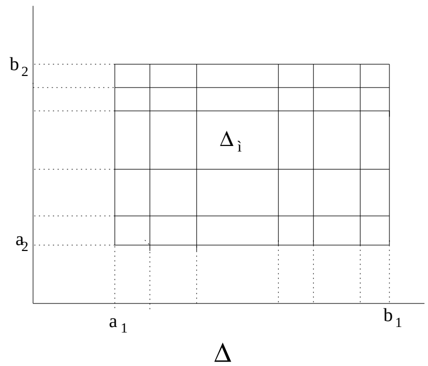

Figure 4.1: partition of cell $\Delta \subset \mathbb{R}^2$ partition

and greatest lower bound in the set $\overline{\Delta}$ are well defined finite real numbers; the **upper sum** of the function f for the partition $\mathcal{P}$ is defined by

intdfq1 (4.2)
$$S^*(f,\mathcal{P}) = \sum_i \sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\overline{\Delta}_i|,$$

and similarly the **lower sum** is defined by

intdfq2 (4.3)
$$S_*(f,\mathcal{P}) = \sum_i \inf_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\overline{\Delta}_i|.$$

It is clear that

intdfq2a (4.4)
$$S_*(f,\mathcal{P}) \le S^*(f,\mathcal{P})$$

for any partition $\mathcal{P}$ of the cell $\overline{\Delta}$. Adding further points of subdivision to the sides of the cell leads to a finer decomposition of the cell as a union of subcells, called a **refinement** of the partition. If $\mathcal{P}_2$ is a refinement of a partition $\mathcal{P}_1$ of a cell Δ then

intdfq3 (4.5)
$$S^*(f,\mathcal{P}_2) \le S^*(f,\mathcal{P}_1) \quad \text{and} \quad S_*(f,\mathcal{P}_2) \ge S_*(f,\mathcal{P}_1);$$

for if a cell $\overline{\Delta}_i$ in the partition $\mathcal{P}_1$ is decomposed into a union $\overline{\Delta}_i = \bigcup_j \overline{\Delta}_{ij}$ of cells $\overline{\Delta}_{ij}$ in the partition $\mathcal{P}_2$ then

$$\sup_{\mathbf{x}\in\overline{\Delta}_{ij}} f(\mathbf{x}) \le \sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x}) \quad \text{and} \quad \sum_j |\overline{\Delta}_{ij}| = |\overline{\Delta}_i|$$

and consequently

$$\begin{aligned} S^*(f,\mathcal{P}_2) &= \sum_{i,j} \sup_{\mathbf{x}\in\overline{\Delta}_{ij}} f(\mathbf{x})|\overline{\Delta}_{ij}| \\ &\le \sum_{i,j} \sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\overline{\Delta}_{ij}| = \sum_i \sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\Delta_i| = S^*(f,\mathcal{P}_1), \end{aligned}$$

and correspondingly for the lower sums with the reversed inequalities. It follows that if $\mathcal{P}_1$ and $\mathcal{P}_2$ are any two partitions of a cell Δ then

intdfq4
$$S_*(f,\mathcal{P}_1) \le S^*(f,\mathcal{P}_2); \tag{4.6}$$

for by considering all the points of subdivision of the sides of the cell $\overline{\Delta}$ from both partitions as defining a further partition $\mathcal{P}'$, a common refinement of both of these partitions, then by (4.4) and (4.5)

intdfq5
$$S_*(f,\mathcal{P}_1) \le S_*(f,\mathcal{P}') \le S^*(f,\mathcal{P}') \le S^*(f,\mathcal{P}_2). \tag{4.7}$$

In view of (4.7) the upper sums $S^*(f,\mathcal{P})$ for all of the partitions $\mathcal{P}$ of a cell $\overline{\Delta}$ are bounded below by any lower sum $S_*(f,\mathcal{P}_1)$, and the lower sums $S_*(f,\mathcal{P})$ for all of the partitions $\mathcal{P}$ of a cell $\overline{\Delta}$ are bounded above by any upper sum $S^*(f,\mathcal{P}_2)$; so it is possible to define the **upper integral** and **lower integral** of the function f over the cell $\overline{\Delta}$ by

intdfq6
$$\int_{\overline{\Delta}}^{*} f = \inf_{\mathcal{P}} S^*(f,\mathcal{P}) \quad \text{and} \quad \int_{*\,\overline{\Delta}} f = \sup_{\mathcal{P}} S_*(f,\mathcal{P}), \tag{4.8}$$

where the infimum and supremum are extended over all partitions $\mathcal{P}$ of the cell $\overline{\Delta}$, and clearly

intdfq7
$$\int_{*\,\overline{\Delta}} f \le \int_{\overline{\Delta}}^{*} f. \tag{4.9}$$

It may be the case though that the upper integral is strictly greater than the lower integral. For example, if f is the function on the real line $\mathbb{R}$ defined by

intex1
$$f(x) = \begin{cases} 0 & \text{if } x \text{ is rational,} \\ \\ 1 & \text{if } x \text{ is irrational,} \end{cases} \tag{4.10}$$

then $\sup_{x\in\overline{\Delta}_i} f(x) = 1$ and $\inf_{x\in\overline{\Delta}_i} f(x) = 0$ for any partition $\mathcal{P}$ of a cell $\overline{\Delta}$ into a union $\overline{\Delta} = \bigcup_i \overline{\Delta}_i$, and consequently $S^*(f,\mathcal{P}) = |\overline{\Delta}|$ and $S_*(f,\mathcal{P}) = 0$; and since that is the case for any partition $\mathcal{P}$ it follows that $\int_{\overline{\Delta}}^{*} f = |\overline{\Delta}|$ and $\int_{*\,\overline{\Delta}} f = 0$, so $\int_{\overline{\Delta}}^{*} f > \int_{*\,\overline{\Delta}} f$ in this case. On the other hand, it is clear that if $f(\mathbf{x}) = c$ at all points $\mathbf{x} \in \overline{\Delta}$ then $\int_{\overline{\Delta}}^{*} f = \int_{*\,\overline{\Delta}} f = c|\overline{\Delta}|$. More generally if f is a continuous function in a cell $\overline{\Delta}$ then $\int_{\overline{\Delta}}^{*} f = \int_{*\,\overline{\Delta}} f$. Indeed if f is continuous in $\overline{\Delta}$ then it is uniformly continuous since $\overline{\Delta}$ is compact, so for any $\epsilon > 0$ there is a partition $\mathcal{P}$ of the cell as a union $\overline{\Delta} = \cup_i \overline{\Delta}_i$ where the cells $\overline{\Delta}_i$ are sufficiently small that $\sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x}) - \inf_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x}) \le \epsilon$ for each cell $\overline{\Delta}_i$. Then

intoap
$$\begin{aligned} S^*(f,\mathcal{P}) - S_*(f,\mathcal{P}) &= \sum_i \Big(\sup_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\overline{\Delta}_i| - \inf_{\mathbf{x}\in\overline{\Delta}_i} f(\mathbf{x})|\overline{\Delta}_i| \Big) \\ &\le \sum_i \epsilon|\overline{\Delta}_i| = \epsilon|\overline{\Delta}|; \end{aligned} \tag{4.11}$$

that is the case for all partitions hence $\int_{\overline{\Delta}}^{*} f - \int_{*\overline{\Delta}} f \le \epsilon|\overline{\Delta}|$ for any $\epsilon > 0$, and consequently $\int_{\overline{\Delta}}^{*} f = \int_{*\overline{\Delta}} f$. This proof can be tweaked to yield a more general result that is a complete characterization of those functions for which the upper and lower integrals are actually equal; but that requires a digression to discuss some further properties of functions and point sets.

The **oscillation** of a bounded function f in a subset $S \subset \mathbb{R}^n$ is defined by

intos1
$$o_f(S) = \sup_{\mathbf{x}\in S} f(\mathbf{x}) - \inf_{\mathbf{x}\in S} f(\mathbf{x}); \tag{4.12}$$

this is well defined, since the supremum and infimum of any bounded collection of real numbers $\{f(\mathbf{x})\}$ are well defined. In view of (4.11) [intoap] it is clear that

intoapx
$$S^*(f,\mathcal{P}) - S_*(f,\mathcal{P}) = \sum_i o_f(\overline{\Delta}_i)|\overline{\Delta}_i| \tag{4.13}$$

for any partition $\mathcal{P}$, so the notion of oscillation is closely related to the question whether the upper and lower integrals of a function are equal. It is evident that if $S_1 \subset S_2$ then $o_f(S_1) \le o_f(S_2)$. In particular if a function f is defined and bounded in an open neighborhood $\mathcal{N}_r(\mathbf{a})$ of a point $\mathbf{a}$ then $o_f\big(\mathcal{N}_{r_1}(\mathbf{a})\big) \le o_f\big(\mathcal{N}_{r_2}(\mathbf{a})\big)$ whenever $r_1 \le r_2$ and r_2 is sufficiently small that these neighborhoods are contained in U, so the limit

intos2
$$o_f(\mathbf{a}) = \lim_{r\to 0} o_f\big(\mathcal{N}_r(\mathbf{a})\big) \tag{4.14}$$

is well defined; it is called the **oscillation** of the function $f(\mathbf{x})$ at the point $\mathbf{a}$.

intoslm **Lemma 4.1** *A bounded function f in an open neighborhood of a point $\mathbf{a} \in \mathbb{R}^n$ is continuous at the point $\mathbf{a}$ if and only if $o_f(\mathbf{a}) = 0$.*

Proof: If $o_f(\mathbf{a}) = 0$ then for any $\epsilon > 0$ there is a $\delta > 0$ such that $o_f\big(\mathcal{N}_\delta(\mathbf{a})\big) < \epsilon$, so in particular $|f(\mathbf{x}) - f(\mathbf{a})| < \epsilon$ whenever $\mathbf{x} \in \mathcal{N}_\delta(\mathbf{a})$ and consequently f is continuous at the point $\mathbf{a}$. Conversely if f is continuous at $\mathbf{a}$ then for any $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(\mathbf{x}) - f(\mathbf{a})| < \frac{1}{2}\epsilon$ whenever $\mathbf{x} \in \mathcal{N}_\delta(\mathbf{a})$; therefore $|f(\mathbf{a})| - \frac{1}{2}\epsilon \le |f(\mathbf{x})| \le |f(\mathbf{a})| + \frac{1}{2}\epsilon$ for any $\mathbf{x} \in \mathcal{N}_\delta(\mathbf{a})$, hence $o_f\big(\mathcal{N}_\delta(\mathbf{a})\big) < \epsilon$ and consequently $o_f(\mathbf{a}) = 0$. That suffices for the proof.

intoslm2 **Lemma 4.2** *If $f(\mathbf{x})$ is a bounded function in a subset $S \subset \mathbb{R}^n$ then*

$$E = \Big\{ \mathbf{x} \in S \;\Big|\; o_f(\mathbf{x}) < \epsilon \Big\}$$

is a relatively open subset of S for any $\epsilon > 0$.

Proof: If $\mathbf{a} \in E$ then $o_f(\mathbf{a}) < \epsilon$ so $o_f\big(\mathcal{N}_\delta(\mathbf{a}) \cap S\big) < \epsilon$ if δ is sufficiently small. If $\mathbf{b} \in \mathcal{N}_\delta(\mathbf{a}) \cap S$ and δ_0 is chosen sufficiently small that $\mathcal{N}_{\delta_0}(\mathbf{b}) \subset \mathcal{N}_\delta(\mathbf{a})$ then $o_f\big(\mathcal{N}_{\delta_0}(\mathbf{b}) \cap S\big) \le o_f\big(\mathcal{N}_\delta(\mathbf{a}) \cap S\big) < \epsilon$ and therefore $\Big(\mathcal{N}_{\delta_o}(\mathbf{b}) \cap S\Big) \subset E$. That shows that E is a relatively open subset of S and thereby concludes the proof.

A subset $S \subset \mathbb{R}^n$ has **content zero** if for any $\epsilon > 0$ there are *finitely* many open cells Δ_i such that $S \subset \bigcup_i \Delta_i$ and $\sum_i |\Delta_i| < \epsilon$; the set S has **measure zero** if for any $\epsilon > 0$ there are *countably* many open cells Δ_i such that $S \subset \bigcup_i \Delta_i$ and $\sum_i |\Delta_i| < \epsilon$. Clearly any set of content zero is also a set of measure zero; and a compact set of measure zero is also of content zero, since if S is compact and $S \subset \bigcup_i \Delta_i$ where $\sum_i |\Delta_i| < \epsilon$ then S is already contained in the union of finitely many of these cells Δ_i.

intoslm3 **Lemma 4.3** *If countably many subsets $E_i \subset \mathbb{R}^n$ each have measure zero then their union $E = \bigcup_{i=1}^\infty E_i$ also has measure zero.*

Proof: Since E_i has measure zero then for any $\epsilon > 0$ there are countably many open cells Δ_{ij} for $j = 1, 2, \ldots$ such that $E_i \subset \bigcup_{j=1}^\infty \Delta_{ij}$ and $\sum_{j=1}^\infty \Delta_{ij} < \epsilon/2^i$; then $E = \bigcup_{i=1}^\infty E_i \subset \bigcup_{i,j=1}^\infty \Delta_{ij}$, which is also a countable union of open cells, and $\sum_{i,j=1}^\infty |\Delta_{ij}| = \sum_{i=1}^\infty \sum_{j=1}^\infty |\Delta_{ij}| \le \sum_{i=1}^\infty \frac{1}{2^i}\epsilon = \epsilon$, which suffices for the proof.

For example, the set $\mathbb{Q}$ of rational numbers does not have content zero, since no finite number of finite open intervals can cover all the rational numbers; but it does have measure zero as a consequence of the preceding lemma, since each individual point of course has measure zero. There are uncountable sets of measure zero as well, for any proper submanifold of an open subset $U \subset \mathbb{R}^n$ is a set of measure zero and is a noncountable set if its dimension is greater than 0. Indeed if $V \subset U$ is a submanifold of dimension k where $1 \le k < n$ then by Corollary chsummadcr 3.6 for any point $\mathbf{a} \in V$ after relabeling the coordinates if necessary the intersection $V \cap \Delta$ of the submanifold V with a sufficiently small cell $\Delta_{\mathbf{a}} = \Delta'_{\mathbf{a}} \times \Delta''_{\mathbf{a}}$ centered at the point $\mathbf{a}$, where $\Delta'_{\mathbf{a}}$ is in the subspace $\mathbb{R}^{n-k}$ consisting of the first $n-k$ coordinates $x_1, \ldots, x_{n-k}$ and $\Delta''_{\mathbf{a}}$ is in the subspace $\mathbb{R}^k$ consisting of the last k coordinates $x_{n-k+1}, \ldots, x_n$, is described by the equations

$$V \cap \Delta_{\mathbf{a}} = \left\{ \mathbf{x} \in U_{\mathbf{a}} \;\middle|\; x_i = g_i(x_{n-k+1}, \ldots, x_n) \text{ for } 1 \le i \le n-k \right\}$$

for some continuously differentiable functions $g_i(\mathbf{x}')$ of the variables $\mathbf{x}' \in \Delta'_{\mathbf{a}}$, as in the accompanying Figure mano 4.2. Each point $\mathbf{x} \in V$ can be enclosed in a small cell $\Delta_i = \Delta'_i \times \Delta''_i$ where $|\Delta'_i| < \epsilon$, so that $\sum_i |\Delta_i| < \epsilon |\Delta''|$, from which it is evident that the set $V \cap \Delta_{\mathbf{a}}$ is of measure 0. The entire submanifold V can be covered by countably many such cells $\Delta_{\mathbf{a}}$, so by Lemma intoslm3 4.3 the set V is also of measure 0.

A function f is **continuous almost everywhere** in a set $S \subset \mathbb{R}^n$ if the subset $E \subset S$ of points where f fails to be continuous is a set of measure zero. For example, the function f on $\mathbb{R}$ defined by

cnex
$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = \frac{p}{q} \text{ where } p, q \text{ are relatively integers,} \\ 0 & \text{if } x \text{ is irrational,} \end{cases} \tag{4.15}$$

is continuous at the irrational numbers but is discontinuous at any rational number, so it is continuous almost everywhere. It is an amusing point, but a bit

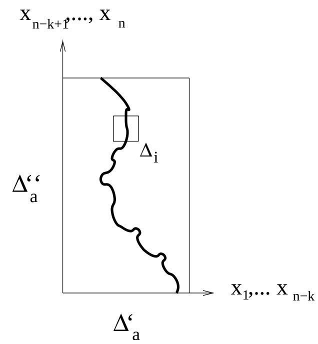

Figure 4.2: submanifold of a cell, a set of measure 0 mano

harder to prove, that there are no functions that are continuous at the rational numbers but discontinuous at any irrational number[1].

thleb **Theorem 4.1 (Riemann - Lebesgue Theorem)** *If f is a bounded function in a cell $\overline{\Delta} \subset \mathbb{R}^n$ then $\int^*_{\overline{\Delta}} f = \int_{*\overline{\Delta}} f$ if and only if f is continuous almost everywhere in $\overline{\Delta}$.*

Proof: First note that for any bounded function f in a cell $\overline{\Delta}$ and for any $\epsilon > 0$ let

$$E_\epsilon = \{\, \mathbf{x} \in \overline{\Delta} \mid o_f(\mathbf{x}) < \epsilon \,\} \quad \text{and} \quad F_\epsilon = \{\, \mathbf{x} \in \overline{\Delta} \mid o_f(\mathbf{x}) \ge \epsilon \,\},$$

so $F_\epsilon = \overline{\Delta} \sim E_\epsilon$. By Lemma 4.2 the set E_ϵ is relatively open in $\overline{\Delta}$, so its complement F_ϵ is a relatively closed subset of $\overline{\Delta}$; and since $\overline{\Delta}$ is compact it follows that the closed subset $F_\epsilon \subset \overline{\Delta}$ is compact. By Lemma 4.1 the function f is continuous at a point $\mathbf{a} \in \overline{\Delta}$ if and only if $o_f(\mathbf{a}) = 0$; so the set of points in $\overline{\Delta}$ at which f is not continuous is the set $F = \{\, \mathbf{x} \in \overline{\Delta} \mid o_f(\mathbf{x}) > o \,\}$, and $F_\epsilon \subset F$ for any $\epsilon > 0$ while $F = \bigcup_{n=1}^\infty F_{\frac{1}{n}}$.

To turn to the proof itself, first suppose that the function f is bounded and continuous almost everywhere in $\overline{\Delta}$. The set F then is of measure zero, as are the subsets F_ϵ; and since the sets F_ϵ are compact, they even are of content zero. Therefore there are finitely many open cells $\Delta'_j \subset \mathbb{R}^n$ such that $F_\epsilon \subset \bigcup_j \Delta'_j$ and $\sum_j |\Delta'_j| < \epsilon$. The complement $K = \overline{\Delta} \sim \bigcup_j \Delta'_j$ is a closed and consequently compact subset of $\overline{\Delta}$, and if $\mathbf{a} \in K$ then $\mathbf{a} \notin F_\epsilon$ so $o_f(\mathbf{a}) < \epsilon$; hence for any point $\mathbf{a} \in K$ there is an open cell $\Delta''_{\mathbf{a}}$ containing that point such that $o_f(\overline{\Delta}''_{\mathbf{a}}) < \epsilon$. Since K is compact finitely many of the open cells $\Delta''_{\mathbf{a}_k}$ cover the compact set K. Let $\mathcal{P}$ be the partition $\overline{\Delta} = \bigcup_i \overline{\Delta}_i$ of the cell $\overline{\Delta}$ determined by the points of subdivision of the finitely many cells Δ'_j and $\Delta''_{\mathbf{a}_k}$. Let I' be the set of those indices i such that the cell $\overline{\Delta}_i$ is contained within some of the cells $\overline{\Delta}'_j$, and let

[1] This result follows from an application of the Baire Category Theorem, which is discussed in most of the standard text books on real analysis.

I'' be the remaining set of indices. Thus $F_\epsilon \subset \bigcup_{i\in I'} \overline{\Delta}_i$ and $\sum_{i\in I'} |\overline{\Delta}_i| < \epsilon$, while if $i \in I''$ then $\overline{\Delta}_i \subset \overline{\Delta}''_{\mathbf{a}_k}$ for some of the cells $\overline{\Delta}''_{\mathbf{a}_k}$ so $o_f(\overline{\Delta}_i) < \epsilon$. If $|f(\mathbf{x})| \le M$ for all points $\mathbf{x} \in \overline{\Delta}$ then $o_f(D) \le 2M$ for any subset $D \subset \overline{\Delta}$ and consequently

$$\begin{aligned} S^*(f,\mathcal{P}) - S_*(f,\mathcal{P}) &= \sum_i o_f(\overline{\Delta}_i)|\overline{\Delta}_i| = \sum_{i\in I'} o_f(\overline{\Delta}_i)|\overline{\Delta}_i| + \sum_{i\in I''} o_f(\overline{\Delta}_i)|\overline{\Delta}_i| \\ &\le 2M \sum_{i\in I'} |\overline{\Delta}_i| + \epsilon \sum_{i\in I''} |\overline{\Delta}_i| \le 2M\epsilon + |\overline{\Delta}|\epsilon. \end{aligned}$$

Since that is the case for any ϵ it follows that

$$\int^*_{\overline{\Delta}} f = \int_{*\,\overline{\Delta}} f.$$

Next suppose that f is a bounded function in $\overline{\Delta}$ such that $\int^*_{\overline{\Delta}} f = \int_{*\,\overline{\Delta}}$. For any $\eta > 0$ there is a partition $\mathcal{P}$ of the cell $\overline{\Delta}$ as a union $\overline{\Delta} = \bigcup_i \overline{\Delta}_i$ such that

$$\sum_i o_f(\overline{\Delta}_i)|\overline{\Delta}_i| = S^*(f,\mathcal{P}) - S_*(f,\mathcal{P}) < \eta.$$

Let I' be the set of those indices i such that $\overline{\Delta}_i \cap F_\epsilon \ne \emptyset$ and let I'' be the remaining indices; thus $F_\epsilon \subset \bigcup_{i\in I'} \overline{\Delta}_i$, and if $i \in I'$ there is a point $\mathbf{a}_i \in \overline{\Delta}_i$ such that $o_f(\mathbf{a}_i) \ge \epsilon$, and consequently $o_f(\overline{\Delta}_i) \ge \epsilon$. Then

$$\eta > \sum_{i\in I'} o_f(\overline{\Delta}_i)|\overline{\Delta}_i| + \sum_{i\in I''} o_f(\overline{\Delta}_i)|\overline{\Delta}_i| \ge \sum_{i\in I'} o_f(\overline{\Delta}_i)|\overline{\Delta}_i| \ge \sum_{i\in I'} \epsilon|\overline{\Delta}_i|$$

so that $\sum_{i\in I'} |\overline{\Delta}_i| < \eta/\epsilon$; and since $F_\epsilon \subset \bigcup_{i\in I'} \overline{\Delta}_i$ for all $\eta > 0$ it follows that F_ϵ is a set of measure zero. Finally since $F \subset \bigcup_{n=1}^\infty F_{1/n}$ and all the sets $F_{1/n}$ are of measure zero the set F also is of measure zero, which suffices for the proof.

It is convenient to say that a function f is **Riemann integrable** over a subset $S \subset \mathbb{R}^n$ if it is bounded and continuous almost everywhere in S. The preceding theorem shows that the Riemann integrable functions in a cell $\overline{\Delta}$ are precisely those functions for which the upper and lower integrals over $\overline{\Delta}$ exist and are equal[2]. The common value of the upper and lower integrals of a Riemann integrable function f is called the **Riemann integral** of the function f over the cell $\overline{\Delta}$ and is denoted by $\int_{\overline{\Delta}} f$, so that

thlebap1

$$\int_{\overline{\Delta}} f = \int^*_{\overline{\Delta}} f = \int_{*\,\overline{\Delta}} f; \tag{4.16}$$

[2] It is perhaps more common to say that a function is Riemann integrable over a cell if and only if the upper and lower integrals exist and are equal; but that property rests on the results of rather nontrivial constructions involving the function rather on somewhat more intrinsic properties of the function itself. Riemann integrability as defined here makes equal sense in any subset of $\mathbb{R}^n$ without the need of defining the integral over more general sets.

for convenience this is often called just the integral of the function f, although some caution is necessary since there are other common notions of integrals, such as the Lebesgue or Denjoy integrals.

Although only the integral of the function f over a cell has been defined so far, it is possible to extend the notion of the integral of a Riemann integrable function to more general sets than just cells. A bounded subset $D \subset \mathbb{R}^n$ is called a **Jordan domain** if its boundary ∂D is a set of measure zero; since ∂D is a closed bounded set, and hence a compact set, then actually ∂D is a set of content zero for any Jordan domain D. It is easier to show that a domain is a Jordan domain by using the characterization that the boundary is a set of measure zero; but for many results about Jordan domains it is more convenient to use the equivalent characterization that the boundary is a set of content zero. The equivalence of these two characterizations should be kept in mind in the subsequent discussion. The definition of a Jordan domain does not impose any further restrictions on the set D; so a Jordan domain may be open or closed or even a more general point set[3]. An open or closed cell is a Jordan domain, since as noted earlier $\partial\Delta$ is a set of measure zero for any cell Δ. The intersection and union of two Jordan domains are again Jordan domains. To verify that first for the union, if D_1 and D_2 are Jordan domains then $\overline{D_1 \cup D_2} \subset \overline{D_1} \cup \overline{D_2}$ and $\overline{\sim (D_1 \cup D_2)} = \overline{\sim D_1 \cap \sim D_2} \subset \overline{\sim D_1} \cap \overline{\sim D_2}$ and consequently

$$\begin{aligned}\partial(D_1 \cup D_2) &= \left(\overline{D_1 \cup D_2}\right) \bigcap \overline{\sim (D_1 \cup D_2)} \\ &\subset \left(\overline{D_1} \cap (\overline{\sim D_1} \cap \overline{\sim D_2})\right) \bigcup \left(\overline{D_2} \cap (\overline{\sim D_1} \cap \overline{\sim D_2})\right),\end{aligned}$$

which is a set of measure zero since $\overline{D_1} \cap \overline{\sim D_1}$ and $\overline{D_2} \cap \overline{\sim D_2}$ are sets of measure zero. If $D \subset \overline{\Delta}$ is a Jordan domain then its complement $\overline{\Delta} \sim D$ is also a Jordan domain since $\partial(\overline{\Delta} \sim D) = \partial\overline{\Delta} \cup \partial D$; consequently the intersection of Jordan domains also is a Jordan domain.

A Jordan domain $D \subset \mathbb{R}^n$ is a bounded subset of $\mathbb{R}^n$ by definition, so $D \subset \Delta$ for some cell $\Delta \subset \mathbb{R}^n$. A Riemann integrable function f in the Jordan domain D can be extended to a function $\widetilde{f}$ in the cell $\overline{\Delta}$ by setting

ext (4.17)
$$\widetilde{f}(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \ \mathbf{x} \in D \\ \\ 0 & \text{if } \ \mathbf{x} \in \overline{\Delta} \sim D. \end{cases}$$

The extended function $\widetilde{f}$ of course is bounded in $\overline{D}$. It is continuous almost everywhere in the interior of D, since the function f is continuous almost everywhere in D by assumption, and it is also continuous in the complement $\overline{\Delta} \sim \overline{D}$ of the closure of D, where it vanishes identically by definition; and since $\overline{\Delta} = D^o \cup (\overline{\Delta} \sim D)^o \cup \partial D$ where ∂D is a set of measure 0 it follows that the

[3] It is more common to call such sets Jordan measurable sets; but since the interest here is not in Jordan measure as such the modified terminology seems more appropriate. The term domain is sometimes used just to denote an open set, but it is convenient to use that term somewhat more generally.

function $\widetilde{f}$ is also continuous almost everywhere in $\overline{\Delta}$. Consequently the function $\widetilde{f}$ is Riemann integrable in the cell Δ, so it is possible to define the integral of the initial function f over the Jordan domain D by

intjrd (4.18)
$$\int_D f = \int_{\overline{\Delta}} \widetilde{f}.$$

It is evident from the definition of the Riemann integral that the value of the integral $\int_D f$ is independent of the choice of the cell $\overline{\Delta}$; for if $\overline{\Delta} \subset \overline{\Delta'}$ then in calculating the integral $\int_{\overline{\Delta'}} \tilde{f}$ it is always possible to take partitions such that each cell $\overline{\Delta'}_i$ in the partition is either contained in $\overline{\Delta}$ or is disjoint from $\overline{\Delta}$ or is one of a finite number of cells containing $\partial\Delta$ and of total content less than any given ϵ. As a special case, if f is a Riemann integrable function in a closed cell $\overline{\Delta}$ its restriction $f|D$ to a Jordan domain $D \subset \overline{\Delta}$ is a Riemann integrable function in D and the extension $\widetilde{f|D}$ can be identified with the product $\chi_D f$, where χ_D is the **characteristic function** of the set D, the function defined by

dfch (4.19)
$$\chi_D(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad x \in D, \\ \\ 0 & \text{if} \quad x \notin D. \end{cases}$$

Thus if f is a Riemann integrable function in a closed cell $\overline{\Delta}$ then its integral over any Jordan domain $D \subset \overline{\Delta}$ can be defined equivalently by

dfintd (4.20)
$$\int_D f = \int_\Delta \chi_D\, f.$$

In particular the **content** of a Jordan domain $D \subset \Delta$ can be defined as the integral of the constant function 1 over the set D, so by the integral

intcnt (4.21)
$$|D| = \int_{\overline{\Delta}} \chi_D.$$

When $D = \overline{\Delta}$ this definition coincides with the previous definition of the content of a cell, since for any partition $\mathcal{P}$ of a cell $\overline{\Delta}$ it follows immediately from the definition of the upper sum that $S^*(\chi_{\overline{\Delta}}, \mathcal{P}) = \sum_i |\overline{\Delta}_i| = |\overline{\Delta}|$ and consequently that $\int_{\overline{\Delta}} \chi_{\overline{\Delta}} = |\overline{\Delta}|$.

rintp1 **Theorem 4.2** *If f_1 and f_2 are Riemann integrable functions in a Jordan domain D then so is any linear combination $f = c_1 f_1 + c_2 f_2$ for real constants c_1, c_2 and*

rintp1qq1 (4.22)
$$\int_D f = c_1 \int_D f_1 + c_2 \int_D f_2.$$

Proof: It is clear that any linear combination of functions that are bounded and continuous almost everywhere in a Jordan domain D is also bounded and continuous almost everywhere in D, so any linear combination of Riemann integrable functions over D is Riemann integrable over D. The integrals over

D of Riemann integrable functions are defined as the integrals over any cell Δ containing D of the extended functions (4.17); so it is sufficient just to prove the theorem for linear combinations of Riemann integrable functions over a cell $\overline{\Delta}$. If f is Riemann integrable in $\overline{\Delta}$ then for any partition $\mathcal{P}$ of that cell and any constant $c > 0$

$$S^*(cf, \mathcal{P}) = \sum_i \left(\sup_{x \in \overline{\Delta}_i} cf(x) \right) |\overline{\Delta}_i| = c \sum_i \left(\sup_{x \in \overline{\Delta}_i} f(x) \right) |\overline{\Delta}_i| = cS^*(f, \mathcal{P})$$

and consequently

$$\int_{\overline{\Delta}} cf = \int_{\overline{\Delta}}^* cf = c \int_{\overline{\Delta}}^* f = c \int_{\overline{\Delta}} f.$$

On the other hand

$$S^*(-f, \mathcal{P}) = \sum_i \left(\sup_{x \in \overline{\Delta}_i} -f(x) \right) |\overline{\Delta}_i| = -\sum_i \Big(\inf_{x \in \overline{\Delta}_i} f(x) \Big) |\overline{\Delta}_i| = -S_*(f, \mathcal{P})$$

and consequently

$$\begin{aligned} \int_{\overline{\Delta}} -f = \int_{\overline{\Delta}}^* -f &= \inf_{\mathcal{P}} S^*(-f, \mathcal{P}) = \inf_{\mathcal{P}} -S_*(f, \mathcal{P}) \\ &= -\sup_{\mathcal{P}} S_*(f, \mathcal{P}) = -\int_{*\overline{\Delta}} f = -\int_{\overline{\Delta}} f. \end{aligned}$$

If f_1 and f_2 are two Riemann integrable functions in $\overline{\Delta}$ then

$$\sup_{\mathbf{x} \in \Delta_i} (f_1(\mathbf{x}) + f_2(\mathbf{x})) \le \sup_{\mathbf{x} \in \Delta_i} f_1(\mathbf{x}) + \sup_{\mathbf{x} \in \Delta_i} f_2(\mathbf{x})$$

hence

$$\int_{\overline{\Delta}} (f_1 + f_2) = \int_{\overline{\Delta}}^* (f_1 + f_2) \le S^*(f_1 + f_2, \mathcal{P}) \le S^*(f_1, \mathcal{P}) + S^*(f_2, \mathcal{P});$$

the same argument shows that the corresponding inequality holds for the lower sums, so

$$S_*(f_1, \mathcal{P}) + S_*(f_2, \mathcal{P}) \le \int_{\overline{\Delta}} (f_1 + f_2) \le S^*(f_1, \mathcal{P}) + S^*(f_2, \mathcal{P}). \tag{4.23}$$

For any $\epsilon > 0$ there exist partitions $\mathcal{P}_i$ such that

$$S_*(f_i, \mathcal{P}_i) \le \int_{\overline{\Delta}} f_i \le S^*(f_i, \mathcal{P}_i) \quad \text{and} \quad S^*(f_i, \mathcal{P}_i) - S_*(f_i, \mathcal{P}_i) \le \epsilon;$$

these inequalities hold for any refinements of the partitions $\mathcal{P}_i$, so in particular hold for a common refinement $\mathcal{P}$ of these two partitions, and substituting them into (4.23) for this common refinement $\mathcal{P}$ shows that

$$\Big| \int_{\overline{\Delta}} (f_1 + f_2) - \int_{\overline{\Delta}} f_1 - \int_{\overline{\Delta}} f_2 \Big| < 2\epsilon.$$

Since that is the case for any $\epsilon > 0$ it follows that $\int_{\overline{\Delta}}(f_1 + f_2) = \int_{\overline{\Delta}} f_1 + \int_{\overline{\Delta}} f_2$, which suffices for the proof.

Integrability is really required for the preceding result to hold. For the function (4.10) for example $\int^*_{\overline{\Delta}} f = |\overline{\Delta}|$ and $\int_{*\overline{\Delta}} f = 0$, and it is clear that the same argument applied to the function $g(\mathbf{x}) = 1 - f(\mathbf{x})$ shows that $\int^*_{\overline{\Delta}} g = |\overline{\Delta}|$ and $\int_{*\overline{\Delta}} g = 0$; consequently $\int^*_{\overline{\Delta}} f + \int^*_{\overline{\Delta}} g = 2|\Delta|$ while $\int_{*\overline{\Delta}} f + \int_{*\overline{\Delta}} g = 0$. On the other hand $f+g = 1$ so $\int^*_{\overline{\Delta}}(f+g) = \int_{*\overline{\Delta}}(f+g) = \int_{\overline{\Delta}}(f+g) = |\Delta|$. Thus it is neither true that $\int^*_{\overline{\Delta}}(f+g) = \int^*_{\overline{\Delta}} f + \int^*_{\overline{\Delta}} g$ nor that $\int_{*\overline{\Delta}}(f+g) = \int_{*\overline{\Delta}} f + \int_{*\overline{\Delta}} g$.

rintp2 **Theorem 4.3** *Let f and g be Riemann integrable functions over a Jordan domain D.*
(i) *If $f(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in D$ then $\int_D f \ge 0$.*
(ii) *If $f(\mathbf{x}) \ge g(\mathbf{x})$ for all $\mathbf{x} \in D$ then $\int_D f \ge \int_D g$.*
(iii) $\left| \int_D f \right| \le \int_D |f|$.

Proof: Again it is sufficient just to prove these assertions for Riemann integrable functions over a cell $\overline{\Delta}$. If $f(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in \overline{\Delta}$ then $S^*(f, \mathcal{P}) \ge 0$ for any partition $\mathcal{P}$ of the cell $\overline{\Delta}$ so (i) holds. Then (ii) follows by applying (i) to the difference $f - g$ and using the preceding Theorem 4.2, while (iii) follows by applying (ii) to the functions $|f|$ and $\pm f$ since $|f(\mathbf{x})| \ge f(\mathbf{x})$ and $|f(\mathbf{x})| \ge -f(\mathbf{x})$ for all $\mathbf{x} \in \overline{\Delta}$; and that suffices for the proof.

A mapping $\mathbf{f} : D \longrightarrow \mathbb{R}^n$ from a Jordan domain $D \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is said to be **Riemann integrable** if each of its coordinate functions f_i is Riemann integrable over D, and the integral $\int_D \mathbf{f}$ is defined to be the vector having the coordinates $\int_D f_i$. The two preceding theorems extend quite directly from functions to mappings, with the appropriate modification for Theorem 4.3.

rintp3 **Theorem 4.4** (i) *Any linear combination $\mathbf{f} = c_1\mathbf{f}_1 + c_2\mathbf{f}_2$ of Riemann integrable mappings $\mathbf{f}_1$ and $\mathbf{f}_2$ from a Jordan domain $D \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is Riemann integrable and*

rintp3qq1 (4.24)
$$\int_D (c_1\mathbf{f}_1 + c_2\mathbf{f}_2) = c_1 \int_D \mathbf{f}_1 + c_2 \int_D \mathbf{f}_2.$$

(ii) *If $\mathbf{f}$ is a Riemann integrable mapping from a Jordan domain $D \subset \mathbb{R}^m$ into $\mathbb{R}^n$ then*

rintp3qq2 (4.25)
$$\Big\| \int_D \mathbf{f} \Big\|_2 \le \int_D \|\mathbf{f}\|_2.$$

Proof: (i) This is an immediate consequence of Theorem 4.2, since it just involves the individual coordinate functions of the mapping $\mathbf{f}$.
(ii) If $\mathbf{u} \in \mathbb{R}^n$ is a unit vector in the direction of the vector $\int_{\overline{\Delta}} \mathbf{f} \in \mathbb{R}^n$ then it follows from the Cauchy-Schwarz inequality and Theorems 4.2 and 4.3 for the function $\mathbf{u} \cdot \mathbf{f}$ that $\| \int_D \mathbf{f}\|_2 = |\mathbf{u} \cdot \int_D \mathbf{f}| = | \int_D \mathbf{u} \cdot \mathbf{f}| \le \int_D |\mathbf{u} \cdot \mathbf{f}| \le \int_D \|\mathbf{f}\|_2$, which suffices for the proof.

intexth **Theorem 4.5** *If f is a bounded function in a Jordan domain $D \subset \mathbb{R}^n$ such that $f(\mathbf{x}) = 0$ for all points $\mathbf{x} \in (D \sim E)$, where $E \subset D$ is a set of content 0, then f is Riemann integrable over D and $\int_D f = 0$.*

Proof: The Jordan domain D is bounded, so $\overline{D} \subset \overline{\Delta}$ for some cell $\overline{\Delta}$. The extended function $\widetilde{f}$ in $\overline{\Delta}$ also is zero outside the subset $E \subset \overline{\Delta}$ of content zero, so it is continuous there and hence it is Riemann integrable in $\overline{\Delta}$. Since E has content 0 then for any $\epsilon > 0$ it is possible to find a partition $\mathcal{P}$ of the cell $\overline{\Delta}$ as a union $\overline{\Delta} = \bigcup_{i \in I} \overline{\Delta}_i$ so that $E \subset \bigcup_{i \in I'} \overline{\Delta}_i$ for a subset $I' \subset I$ of the indices, where $\sum_{i \in I'} |\overline{\Delta}_i| < \epsilon$, while $\overline{\Delta}_i \cap E = \emptyset$ if $i \in I''$ for the complementary set of indices I''. Then $|\widetilde{f}(\mathbf{x})| < M$ for all points $\mathbf{x} \in \overline{\Delta}_i$ if $i \in I'$, where M is a bound of the function f, while $\widetilde{f}(\mathbf{x}) = 0$ for all points $\mathbf{x} \in \overline{\Delta}_i$ if $i \in I''$; consequently

$$\begin{aligned} S^*(\widetilde{f}, \mathcal{P}) &= \sum_{i \in I'} \left(\sup_{\mathbf{x} \in \overline{\Delta}_i} \widetilde{f}(\mathbf{x}) \right) |\overline{\Delta}_i| + \sum_{i \in I''} \left(\sup_{\mathbf{x} \in \overline{\Delta}_i} \widetilde{f}(\mathbf{x}) \right) |\overline{\Delta}_i| \\ &\le \sum_{i \in I'} M |\overline{\Delta}_i| + 0 \le M\epsilon, \end{aligned}$$

and since that is the case for any $\epsilon > 0$ it follows that $\int_D f = \int^*_{\overline{\Delta}} \widetilde{f} \le 0$. The corresponding argument shows that $\int_D f = \int_{*\overline{\Delta}} \widetilde{f} \ge 0$, and consequently $\int_D f = 0$, which suffices for the proof.

The preceding result fails if the exceptional set is only of measure 0 rather than of content 0; for a function f such that $f(\mathbf{x}) = 1$ at a countable dense set of points of a Jordan domain D and $f(\mathbf{x}) = -1$ at another countable dense set of points of D while $f(\mathbf{x}) = 0$ otherwise does have the property that it vanishes outside of a set of measure 0, but it is not Riemann integrable over D.

intext **Corollary 4.1** *If f is a Riemann integrable function over a Jordan domain D, and if g is a bounded function in D such that $g(\mathbf{x}) = f(\mathbf{x})$ at all points $\mathbf{x} \in (D \sim E)$ for a subset $E \subset D$ of content zero, then g is also Riemann integrable over D and $\int_D g = \int_D f$.*

Proof: The function $f - g$ is bounded in D and vanishes on the set $D \sim E$ where $E \subset D$ is a subset of measure zero, so by the preceding theorem this function is integrable in D and $\int_d (f - g) = 0$. The difference $g = f - (f - g)$ of these two integrable functions is then also integrable, and it follows from Theorem rintp1 4.2 that $\int_D g = \int_D f - \int_D (f - g) = \int_D f$, which suffices for the proof.

A consequence of the preceding corollary is that a Riemann integrable function f in a Jordan domain D can be modified on any subset of content 0 in D in any arbitrary way, so long as it remains bounded, and the modified function remains Riemann integrable with the same integral as the function f. The integral of a function f thus is not so much a property of the individual function f, but rather is a property of an equivalence class of functions, where two Riemann integrable functions in D are equivalent if and only if they are equal outside a set of content 0 in D. In particular when considering integrals or integrability of functions over cells or Jordan domains, the values of a function at the boundary of the set are immaterial, provided of course that the function remains bounded; so as far as integration is concerned, it is immaterial whether closed or open Jordan domains are considered.

intsm **Theorem 4.6** *If two Jordan domains D_1, D_2 are disjoint except possibly for a subset*

intsmqa1
$$E = D_1 \cap D_2 \subset \partial D_1 \cup \partial D_2 \tag{4.26}$$

and if f is a Riemann integrable function in the union $D_1 \cup D_2$ then

intsmq1
$$\int_{D_1 \cup D_2} f = \int_{D_1} f + \int_{D_2} f; \tag{4.27}$$

in particular

intsmq2
$$|D_1 \cup D_2| = |D_1| + |D_2|. \tag{4.28}$$

Proof: The union $D_1 \cup D_2$ of two Jordan domains D_1, D_2 is also a Jordan domain, as noted on page 78, so it is contained in a cell $\overline{\Delta}$; and then $\int_{D_1 \cup D_2} f = \int_{\overline{\Delta}} \chi_{D_1 \cup D_2} f$. The characteristic function of the union of the two Jordan domains is $\chi_{D_1 \cup D_2}(\mathbf{x}) = \chi_{D_1}(\mathbf{x}) + \chi_{D_2}(\mathbf{x}) - \chi_E(\mathbf{x})$, where the intersection $E = D_1 \cap D_2$ is also a Jordan domain; consequently

$$\begin{aligned}
\int_{D_1 \cup D_2} f &= \int_{\Delta} (\chi_{D_1} + \chi_{D_2} - \chi_E) f \\
&= \int_{\Delta} \chi_{D_1} f + \int_{\Delta} \chi_{D_2} f - \int_{\Delta} \chi_E f \\
&= \int_{D_1} f + \int_{D_2} f
\end{aligned}$$

since $E \subset (\partial D_1 \cup \partial D_2)$ and the sets ∂D_1 and ∂D_2 are sets of content 0 for the Jordan domains D_1 and D_2 so E is also a set of content 0 and therefore $\int_{\Delta} \chi_E f = 0$ by Theorem 4.5. That demonstrates (4.27), which holds in particular for $f = 1$ and therefore implies (4.28), thereby concluding the proof.

A subset $D \subset \overline{\Delta}$ has content zero if and only if $\int^*_{\overline{\Delta}} \chi_D = 0$, since for any partition $\mathcal{P}$ of the cell $\overline{\Delta}$ as the union $\overline{\Delta} = \bigcup_{i \in I} \overline{\Delta}_i$ it is clear that

$$S^*(\chi_D, \mathcal{P}) = \sum_{\{\, i \in I \mid D \cap \overline{\Delta}_i \neq \emptyset \,\}} |\overline{\Delta}_i|.$$

That suggests that the notion of the content of a Jordan domain might be extended to the notion of the content of an arbitrary subset $D \subset \overline{\Delta}$ by setting $|D| = \int^*_{\overline{\Delta}} \chi_D$; but for more general disjoint sets[4] it is not necessarily the case that $|D_1 \cup D_2| = |D_1| + |D_2|$. The Lebesgue integral does provide an extension of the notion of the content to some more general sets than Jordan domains for which this is true though.

[4]The Banach-Tarski paradox is the result that the ordinary solid ball of radius 1 in $\mathbb{R}^3$

fub

4.2 Fubini's Theorem

The explicit calculation of integrals in several variables can be reduced to a succession of calculations of integrals in a single variable. A cell $\overline{\Delta} \in \mathbb{R}^n = \mathbb{R}^{n'} \times \mathbb{R}^{n''}$ where $n = n' + n''$ can be written as a product $\overline{\Delta} = \overline{\Delta'} \times \overline{\Delta''}$ of cells $\overline{\Delta'} \in \mathbb{R}^{n'}$ and $\overline{\Delta''} \in \mathbb{R}^{n''}$, so a function f in the cell $\overline{\Delta} \in \mathbb{R}^n$ can be written as a function $f(\mathbf{x}', \mathbf{x}'')$ of variables $\mathbf{x}' \in \overline{\Delta'} \in \mathbb{R}^n$ and $\mathbf{x}'' \in \overline{\Delta''} \in \mathbb{R}^n$. For any fixed point $\mathbf{x}'' \in \overline{\Delta''}$ the function $f(\mathbf{x}', \mathbf{x}'')$ can be considered as a function of the variable $\mathbf{x}' \in \mathbb{R}^{n'}$ alone; so if f is bounded it is possible to consider the upper integral of this function of the variable $\mathbf{x}'$, denoted by $\int^*_{\mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'')$; this upper integral then is a well defined bounded function of the variable $\mathbf{x}'' \in \overline{\Delta''}$, so it is possible to consider the upper integral of this function of the variable $\mathbf{x}''$, denoted by $\int^*_{\mathbf{x}'' \in \overline{\Delta''}} \Big(\int^*_{\mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'') \Big)$ and called an **iterated integral**. Of course it is possible to use the lower rather than upper integral for one or another or both of these integrals, leading to iterated integrals such as $\int^*_{\mathbf{x}'' \in \overline{\Delta''}} \Big(\int_{*\, \mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'') \Big)$; and it is possible to proceed in the other order, integrating first with respect to the variable $\mathbf{x}''$ and then the variable $\mathbf{x}'$, leading to iterated integrals such as $\int^*_{\mathbf{x}' \in \overline{\Delta'}} \Big(\int^*_{\mathbf{x}'' \in \overline{\Delta''}} f(\mathbf{x}', \mathbf{x}'') \Big)$.

intfub **Theorem 4.7 (Fubini's Theorem)** *If f is a Riemann integrable function in a product cell $\overline{\Delta} = \overline{\Delta'} \times \overline{\Delta''}$ the integrals $\int^*_{\mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'')$ and $\int_{*\, \mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'')$ are Riemann integrable functions of the variable $\mathbf{x}'' \in \overline{\Delta''}$ and*

intfubqa
$$\int_{\overline{\Delta}} f = \int_{\mathbf{x}'' \in \overline{\Delta''}} \Big(\int^*_{\mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'') \Big) = \int_{\mathbf{x}'' \in \overline{\Delta''}} \Big(\int_{*\, \mathbf{x}' \in \overline{\Delta'}} f(\mathbf{x}', \mathbf{x}'') \Big). \tag{4.29}$$

Proof: If $\mathcal{P}'$ is a partition $\overline{\Delta'} = \bigcup_i \overline{\Delta'_i}$ of the cell $\overline{\Delta'}$ and $\mathcal{P}''$ is a partition $\overline{\Delta''} = \bigcup_i \overline{\Delta''_i}$ of the cell $\overline{\Delta''}$ these two separate partitions determine a partition $\mathcal{P} = \mathcal{P}' \times \mathcal{P}''$ of the cell $\overline{\Delta}$ and

$$\begin{aligned} S^*(f, \mathcal{P}) &= \sum_{ij} \sup_{\substack{\mathbf{x}' \in \overline{\Delta'_i} \\ \mathbf{x}'' \in \overline{\Delta''_j}}} f(\mathbf{x}', \mathbf{x}'') \left| \overline{\Delta'_i} \times \overline{\Delta''_j} \right| \\ &= \sum_j \left(\sum_i \sup_{\substack{\mathbf{x}' \in \overline{\Delta'_i} \\ \mathbf{x}'' \in \overline{\Delta''_j}}} f(\mathbf{x}', \mathbf{x}'') \left| \overline{\Delta'_i} \right| \right) \left| \overline{\Delta''_j} \right|. \end{aligned}$$

For any fixed point $\mathbf{x}''_0 \in \overline{\Delta''_j}$

$$\sum_i \sup_{\substack{\mathbf{x}' \in \overline{\Delta'_i} \\ \mathbf{x}'' \in \overline{\Delta''_j}}} f(\mathbf{x}', \mathbf{x}'') \left| \overline{\Delta'_i} \right| \ge \sum_i \sup_{\mathbf{x}' \in \overline{\Delta'_i}} f(\mathbf{x}', \mathbf{x}''_0) \left| \overline{\Delta'_i} \right| = S^*\Big(f(\mathbf{x}', \mathbf{x}''_0), \mathcal{P}' \Big) \ge F(\mathbf{x}''_0)$$

can be decomposed into a union of finitely many disjoint sets such that when these sets are suitably rotated and translated in $\mathbb{R}^3$ and put back together again the union is the solid ball of radius 2; this rather surprising result was demonstrated by Banach and Tarski in a joint paper in *Fundamenta Mathematicae* volume 6, (1924), pages 244 277, and is discussed in the book *The Banach-Tarski Paradox* by Stan Wagon, Cambridge University Press, 1994.

where $F(\mathbf{x}_0'') = \int^*_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}_0'')$. That is the case for any fixed point $\mathbf{x}_0'' \in \overline{\Delta_j''}$ so the same results hold for the supremum over all such points; hence

$$S^*(f,\mathcal{P}) \ge \sum_j \sup_{\mathbf{x}_0''\in\overline{\Delta_j''}} F(\mathbf{x}_0'')\Big|\overline{\Delta_j''}\Big| = S^*(F,\mathcal{P}'') \ge \int^*_{\mathbf{x}''\in\overline{\Delta''}} F(\mathbf{x}'')$$

so that

$$\int^*_{\overline{\Delta}} f \ge \int^*_{\mathbf{x}''\in\overline{\Delta''}} \Big(\int^*_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')\Big).$$

The corresponding argument for the lower integral leads to the reversed inequality, so altogether there is the chain of inequalities

$$\begin{aligned}\int^*_{\overline{\Delta}} f \ge \int^*_{\mathbf{x}''\in\overline{\Delta''}} \Big(\int^*_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}'\mathbf{x}'')\Big) &\ge \int_{*\mathbf{x}''\in\overline{\Delta''}} \Big(\int^*_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')\Big)\\ &\ge \int_{*\mathbf{x}''\in\overline{\Delta''}} \Big(\int_{*\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')\Big) \ge \int_{*\overline{\Delta}} f.\end{aligned}$$

Since the function f is Riemann integrable $\int^*_{\overline{\Delta}} f = \int_{*\overline{\Delta}} f$ and consequently all of the preceding inequalities must be equalities, which concludes the proof.

fubthcor **Corollary 4.2** *If f is a continuous function in a product cell $\overline{\Delta} = \overline{\Delta'} \times \overline{\Delta''}$ then*

fubthcorq1
$$\int_{\overline{\Delta}} f(\mathbf{x}) = \int_{\mathbf{x}''\in\overline{\Delta''}} \Big(\int_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')\Big) = \int_{\mathbf{x}'\in\overline{\Delta'}} \Big(\int_{\mathbf{x}''\in\overline{\Delta''}} f(\mathbf{x}',\mathbf{x}'')\Big). \tag{4.30}$$

Proof: If f is continuous in $\overline{\Delta}$ then for each fixed point $\mathbf{x}'' \in \overline{\Delta''}$ the function $f(\mathbf{x}',\mathbf{x}'')$ is a continuous function of $\mathbf{x}' \in \overline{\Delta'}$ so the upper integral or lower integral in (4.29) [intfubqa] can be replaced by the ordinary integral; and the same holds for iterated integrals in the reverse order, so that suffices for the proof.

It is customary also to set

$$\int_{\mathbf{x}''\in\overline{\Delta''}} \Big(\int_{\mathbf{x}'\in\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')\Big) = \int_{\overline{\Delta''}}\int_{\overline{\Delta'}} f(\mathbf{x}',\mathbf{x}'')d\mathbf{x}'d\mathbf{x}'',$$

another commonly used notation for an iterated integral, where the inner integral, that with respect to the inner variable $\mathbf{x}'$, is calculated first, and the result is then integrated over the outer integral, that with respect to the outer variable $\mathbf{x}''$. Integrals over Jordan domains $D \subset \overline{\Delta}$ are calculated similarly, by reducing them to integrals over cells; but some care often is required to keep the regions of integration straight. The process can be repeated, with more iterations; so an integral over a Jordan domain in $\mathbb{R}^n$ can be reduced to the calculation of n separate integrals of functions of a single variable.

For examples of the application of Fubini's Theorem, first let $D \subset \mathbb{R}^2$ be the region bounded by the lines $x_1 + x_2 = 0$, $x_1 - 2x_2 = 0$, and $x_2 = 1$ as in the

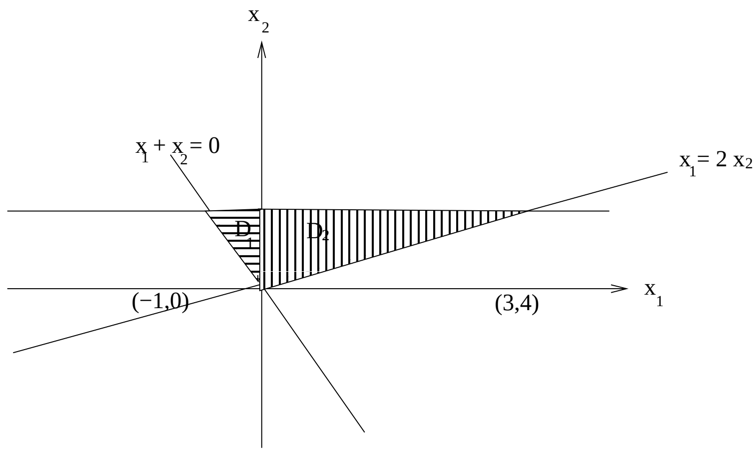

Figure 4.3: an example of Fubini's theorem in $\mathbb{R}^2$ fubex2

accompanying Figure 4.3. An application of Fubini's Theorem shows that for any continuous function $f(\mathbf{x})$ in D

$$\int_D f(\mathbf{x}) = \int_0^1 \int_{-x_2}^{2x_2} f(x_1, x_2) dx_1 dx_2,$$

thus reducing the integral to an iteration of integrals of functions of a single variable. The region D can be split into region D_1 consisting of those points $\mathbf{x} \in D$ for which $-1 \le x_1 \le 0$ and the region D_2 consisting of those points $\mathbf{x} \in D$ for which $0 \le x_1 \le 2$, and then an application of Fubini's Theorem shows that for any continuous function $f(\mathbf{x})$ in D

$$\int_D f(\mathbf{x}) = \int_{-1}^0 \int_{-x_1}^1 f(x_1, x_2) dx_2 dx_1 + \int_0^2 \int_{x_1/2}^1 f(x_1, x_2) dx_2 dx_1.$$

The order in which Fubini's Theorem is applied can make a difference in the calculation. The iterated integral

$$\int_0^1 \int_{\sqrt{x_1}}^1 \cos(x_2^3 + 1) dx_2 dx_1$$

involves a rather complicated trigonometric integral with respect to the variable x_2, but the equal iterated integral involves only simple integrations, since with the change of variables $t = x_2^3$

$$\begin{aligned}\int_0^1 \int_{\sqrt{x_1}}^1 \cos(x_2^3 + 1) dx_2 dx_1 &= \int_0^1 \int_0^{x_2^2} \cos(x_2^3 + 1) dx_1 dx_2 \\ &= \int_0^1 x_2^2 \cos(x_2^3 + 1) dx_2 = \int_0^1 \cos(t + 1) \frac{1}{3} dt = \frac{1}{3}\Big(\sin 2 - \sin 1\Big).\end{aligned}$$

There are some cases in which a Riemann integrable function $f(\mathbf{x}', \mathbf{x}'')$ in a product $\Delta = \Delta' \times \Delta''$ is not a Riemann integrable function of one variable when the other variable is held fixed. For instance the function $f(x_1, x_2)$ of variables $x_1, x_2 \in \mathbb{R}^1$ defined in a cell $\Delta = \Delta' \times \Delta'' \subset \mathbb{R}^2$ by

$$f(x', x'') = \begin{cases} 0 & \text{if either } x' \text{ or } x'' \text{ is irrational,} \\ \\ \frac{1}{q} & \text{if both } x' \text{ and } x'' \text{ are rational and } x'' = \frac{p}{q} \end{cases}$$

for coprime integers p and q actually is independent of the variable x', and as a function of the variable x'' it is continuous whenever x'' is irrational but discontinuous whenever x'' is rational. As a function of two variables its points of discontinuity are the products $\overline{\Delta'} \times \frac{p}{q}$ for all rational points $\frac{p}{q}$, a countable union of intervals; and since intervals are sets of measure zero in $\mathbb{R}^2$ the points of discontinuity form a set of measure zero, so the function f is Riemann integrable. Of course it is also Riemann integrable as a function of the variable x'' for each fixed point $x' \in \overline{\Delta'}$; but for any fixed point x'' the function takes the values 0 or $\frac{1}{q}$ on dense subsets of the interval $\overline{\Delta'}$ so the function is not Riemann integrable as a function of the variable x'. Fortunately in this case the integration can be taken first with respect to the variable x'', and the simpler version of Fubini's Theorem holds.

In Fubini's Theorem it is necessary to assume that a function f on the product cell $\overline{\Delta} = \overline{\Delta'} \times \overline{\Delta''}$ is Riemann integrable; there are functions that are Riemann integrable in the variable $\mathbf{x}' \in \overline{\Delta'}$ for each fixed point $\mathbf{x}'' \in \overline{\Delta''}$, and conversely are Riemann integrable in the variable $\mathbf{x}'' \in \overline{\Delta''}$ for each fixed point $\mathbf{x}' \in \overline{\Delta'}$, but are not Riemann integrable in the product cell $\overline{\Delta}$. For example let $f(x', x'')$ be a function on the cell $\overline{\Delta'} \times \overline{\Delta''} \in \mathbb{R}^2$ that takes the value 1 at the center of the square, that takes the value 1 at each of four points in the subsquares when the square is split in half on each line, but where these points are chosen so that no two lie on the same horizontal or vertical line, and so on; and takes the value 0 otherwise. Thus this function takes each of the values 1 and 0 on dense subsets of the square, so it not Riemann integrable; but on each horizontal or vertical line the function takes the value 1 at a single point and is 0 otherwise, so is Riemann integrable.

4.3 Limits and Improper Integrals

math218notes3

It is familiar that pointwise limits of continuous functions need no longer be continuous; for example the functions $f_\nu(x) = x^\nu$ in the closed interval $[0, 1]$ are continuous and

$$f(x) = \lim_{\nu \to \infty} f_\nu(x) = \begin{cases} 0 & \text{if} \quad 0 \le x < 1, \\ \\ 1 & \text{if} \quad x = 1, \end{cases}$$

which fails to be continuous at the point $x = 1$. It is even the case that limits of integrable functions need no longer be integrable; for example it is a straightforward matter to verify that

$$\lim_{m\to\infty}\lim_{n\to\infty}\big(\cos \pi m! x\big)^{2n} = \begin{cases} 0 & \text{if } \; x \text{ is irrational,} \\[2ex] 1 & \text{if } \; x \text{ is rational,} \end{cases}$$

which is not integrable[5]. However a stronger notion of convergence does preserve continuity. A sequence of mappings $\mathbf{f}_\nu : U \longrightarrow \mathbb{R}^n$ from a subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is **uniformly convergent** to a mapping $\mathbf{f}$ in U if for any $\epsilon > 0$ there is an integer N such that $\|\mathbf{f}_\nu(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| < \epsilon$ for all $\mathbf{x} \in U$ whenever $\nu > N$. It is evident from the inequalities (1.7) that the notion of uniform convergence is independent of the norm $\|\mathbf{x}\|$ is used in the definition.

serunif **Theorem 4.8** *The limit of a uniformly convergent sequence of continuous mappings from a subset $U \subset \mathbb{R}^m$ into $\mathbb{R}^n$ is continuous in U.*

Proof: If $\mathbf{f}_\nu$ converges uniformly to f in U then for any $\epsilon > 0$ there is an integer N sufficiently large that $\|\mathbf{f}_N(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| < \frac{1}{3}\epsilon$ for all $\mathbf{x} \in U$. For any point $\mathbf{a}$ there is a $\delta > 0$ such that $\|\mathbf{f}_N(\mathbf{a}) - \mathbf{f}_N(\mathbf{x})\| < \frac{1}{3}\epsilon$ whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$, since the mapping $\mathbf{f}_N$ is continuous at the point $\mathbf{a}$. Then

$$\begin{aligned} \|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{x})\| &\le \|\mathbf{f}(\mathbf{a}) - \mathbf{f}_N(\mathbf{a}\| + \|\mathbf{f}_N(\mathbf{a}) - \mathbf{f}_N(\mathbf{x})\| + \|\mathbf{f}_N(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \\ &< \frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon \end{aligned}$$

whenever $\|\mathbf{x} - \mathbf{a}\| < \delta$, so the mapping $\mathbf{f}$ is continuous at the point $\mathbf{a}$; and since that is the case at any point $\mathbf{a} \in U$ the mapping $\mathbf{f}$ is continuous in U, which concludes the proof.

Even if a sequence of integrable functions converges pointwise to an integrable function, the integral of the limit is not necessarily equal to the limit of the integral. For instance consider the sequence of continuous functions $f_\nu(x)$ on the unit interval $[0,1]$ where $f_\nu(0) = 0, f_\nu(\frac{1}{2\nu}) = 2\nu, f_u(\frac{1}{\nu}) = 0, f_\nu(1) = 0$, and the function f_ν is extended to be linear in the intervals between these points, as in the accompanying Figure 4.4. It easy to see that $\lim_{n\to\infty} f_n(x) = 0$ for all points $x \in [0,1]$ but that $\int_0^1 f_n = 1$.

convit **Theorem 4.9** *If f_ν is a uniformly convergent sequence of integrable functions in a cell $\overline{\Delta} \subset \mathbb{R}^n$ then its limit f is an integrable function in $\overline{\Delta}$ and*

$$\lim_{\nu\to\infty}\int_{\overline{\Delta}} f_\nu = \int_{\overline{\Delta}} f.$$

[5] It is amusing, although somewhat more difficult to demonstrate, that it does require two consecutive limiting processes to yield this function, which cannot be realized as a simple limit of continuous functions. R. Baire introduced a classification of functions, under which continuous functions are of class 0, functions that are not continuous but are pointwise limits of continuous functions are of class 1, functions that are not of class 1 but are pointwise limits of functions of class 1 are of class 2, and so on. He showed that there are functions of all these "Baire classes". This is discussed in some detail in C. Carathéodory's book *Vorlesungen über Reelle Funktionen*, 1935 (English translation *Real Functions*, Chelsea 1946) and F. Hausdorff's book *Grundzüge der Mengelehre*, 1937 (English translation *Set Theory*, AMS Chelsea 1957.) The subject is often treated in books on real analysis through a set of problems.

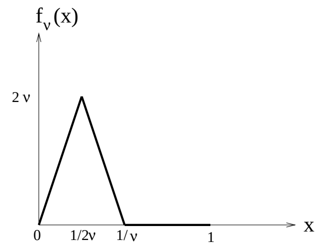

Figure 4.4: an example exlm

Proof: Since the functions f_ν converge uniformly to f there is an integer N such that $|f(\mathbf{x}) - f_N(\mathbf{x})| < 1$ for all $\mathbf{x} \in \overline{\Delta}$; the function f_N is integrable by hypothesis so it is bounded by some constant M in $\overline{\Delta}$, and then

$$|f(\mathbf{x})| \le |f(\mathbf{x}) - f_N(\mathbf{x})| + |f_N(\mathbf{x})| \le 1 + M$$

for all $\mathbf{x} \in \overline{\Delta}$, so the limit function is bounded in $\overline{\Delta}$. The set $E_\nu \subset \overline{\Delta}$ of points at which the function f_ν fails to be continuous is a set of measure zero, since the functions f_ν are integrable by hypothesis; so the union $E = \bigcup_{\nu=1}^\infty E_\nu$ is also a set of measure zero, and all the functions f_ν are continuous in $\overline{\Delta} \sim E$. The functions f_ν converge uniformly to f in $\overline{\Delta} \sim E$ so it follows from the preceding Theorem 4.8 that the limit function f is also continuous in $\overline{\Delta} \sim E$, hence that f is integrable. For any $\epsilon > 0$ there is an N such that $|f(\mathbf{x}) - f_\nu(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in \overline{\Delta}$ whenever $\nu > N$; since the functions f and f_ν are integrable it then follows from Theorem 4.3 that for any $\nu > N$

$$\Big| \int_{\overline{\Delta}} f - \int_{\overline{\Delta}} f_\nu \Big| = \Big| \int_{\overline{\Delta}} (f - f_\nu) \Big| \le \int_{\overline{\Delta}} |f - f_\nu| \le \int_{\overline{\Delta}} \epsilon = \epsilon|\overline{\Delta}|$$

and consequently that $\lim_{\nu\to\infty} \int_{\overline{\Delta}} f_\nu = \int_{\overline{\Delta}} f$, which concludes the proof.

Even when $\mathcal{C}^1$ functions converge uniformly their derivatives to not necessarily converge; for example the functions $f_n(x) = \frac{1}{n}\sin n^2 x$ converge uniformly to zero, but $f_n'(0) = n \longrightarrow \infty$.

lmdf **Theorem 4.10** *If f_ν are differentiable functions in a cell $\Delta \subset \mathbb{R}^n$ which converge at a point $\mathbf{a} \in \Delta$, and if the derivatives f_ν' converge uniformly to a function g in Δ, then the functions f_ν actually converge uniformly in Δ to a differentiable function f in Δ for which $f' = g$.*

Proof: For any $\epsilon > 0$ there is an N sufficiently large that $\|f_\mu'(\mathbf{x}) - f_\nu'(\mathbf{x})\|_\infty < \epsilon$ for all $\mathbf{x} \in \Delta$ and $\|f_\mu(\mathbf{a}) - f_\nu(\mathbf{a})\|_\infty < \epsilon$ whenever $\mu, \nu > N$. It then follows from the Mean Value Inequality, Corollary 2.8, that for any point $\mathbf{x} \in \Delta$ there is a point $\mathbf{y} \in \Delta$ such that whenever $\mu, \nu > N$

$$\|\big(f_\mu(\mathbf{x}) - f_\nu(\mathbf{x})\big) - \big(f_\mu(\mathbf{a}) - f_\nu(\mathbf{a})\big)\|_\infty \le n\|f_\mu'(\mathbf{y}) - f_\nu'(\mathbf{y})\|_\infty \|\mathbf{x} - \mathbf{a}\|_\infty < n\epsilon\|\mathbf{x} - \mathbf{a}\|_\infty$$

hence that

$$\|f_\mu(\mathbf{x}) - f_\nu(\mathbf{x})\|_\infty \le \|f_\mu(\mathbf{a}) - f_\nu(\mathbf{a})\|_\infty + n\epsilon\|\mathbf{x}-\mathbf{a}\|_\infty < (1+nd)\epsilon$$

where $d = \sup_{\mathbf{x}\in\Delta}\|\mathbf{x}-\mathbf{a}\|_\infty$; consequently the functions f_n converge uniformly to a function f in Δ. For any point $\mathbf{x}\in\Delta$ and any $\mathbf{h}\in\mathbb{R}^n$ the function

lmq1

$$\eta_\nu(\mathbf{h}) = \begin{cases} \frac{1}{\|\mathbf{h}\|_\infty}\Big(f_\nu(\mathbf{x}+\mathbf{h}) - f_\nu(\mathbf{x}) - f'_\nu(\mathbf{x})\mathbf{h}\Big) & \text{if} \quad \mathbf{h}\ne\mathbf{0}, \\ 0 & \text{if} \quad \mathbf{h}=\mathbf{0}, \end{cases} \tag{4.31}$$

is a continuous function of the variable $\mathbf{h}$ in an open neighborhood of the origin, since f_ν is differentiable. If $\mathbf{h}\ne\mathbf{0}$ then

$$\begin{aligned}\|\eta_\mu(\mathbf{h}) - \eta_\nu(\mathbf{h})\|_\infty \le \frac{1}{\|\mathbf{h}\|_\infty}\|\big(f_\mu(\mathbf{x}+\mathbf{h}) - f_\nu(\mathbf{x}+\mathbf{h})\big) - \big(f_\mu(\mathbf{x}) - f_\nu(\mathbf{x})\big)\|_\infty \\ + \|f'_\mu(\mathbf{x}) - f'_\nu(\mathbf{x})\|_\infty,\end{aligned}$$

and then from the Mean Value Inequality again for some point $\mathbf{y}$ between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$

$$\|\eta_\mu(\mathbf{h}) - \eta_\nu(\mathbf{h})\|_\infty \le n\|f'_\mu(\mathbf{y}) - f'_\nu(\mathbf{y})\|_\infty + \|f'_\mu(\mathbf{x}) - f'_\nu(\mathbf{x})\|_\infty,$$

and consequently $\|\eta_\mu(\mathbf{h}) - \eta_\nu(\mathbf{h})\|_\infty < (n+1)\epsilon$ whenever $\mu,\nu > N$, which also holds trivially for $\mathbf{h}=0$. Therefore the functions $\eta_\nu(\mathbf{h})$ are uniformly convergent in an open neighborhood of the origin to a continuous limit function $\eta(\mathbf{h})$, for which $\lim_{\mathbf{h}\to 0}\eta(\mathbf{h}) = \eta(\mathbf{0}) = 0$. For any fixed $\mathbf{h}\ne\mathbf{0}$ it follows from (4.31) that

$$\eta(\mathbf{h}) = \frac{1}{\|\mathbf{h}\|_\infty}\Big(f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - g(\mathbf{x})\mathbf{h}\Big)$$

since the functions f_ν converge to f and the functions f'_ν converge to g; but for $\mathbf{h}=\mathbf{0}$ it follows from (4.31) that $\eta(\mathbf{0}) = 0$. Therefore

$$\lim_{\mathbf{h}\to\mathbf{0}}\eta(\mathbf{h}) = \lim_{\mathbf{h}\to\mathbf{0}}\frac{1}{\|\mathbf{h}\|_\infty}\Big(f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - g(\mathbf{x})\mathbf{h}\Big) = 0,$$

which shows that $f'(\mathbf{x}) = g(\mathbf{x})$ and thereby concludes the proof.

Since the derivatives f'_ν determine the functions f_ν only up to additive constants it is apparent that it is not enough in the preceding theorem just to assume something about the convergence of the derivatives f'_ν and nothing about the convergence of the functions f_ν themselves; but by that theorem it is enough just to assume in addition to the uniform convergence of the derivatives that the functions f_ν converge at just a single point. The preceding theorem does not require any regularity properties of the derivatives f'_ν; but if the derivatives are assumed to be continuous there is a considerably simpler proof of the preceding theorem. For functions of a single variable $f_\nu(x) - f_\nu(a) = \int_a^x f'_\nu$ by the fundamental theorem of the calculus, so the convergence of the functions

f_ν follows directly from the convergence of the values $f_\nu(a)$ and of the uniform convergence of the functions f'_ν and the limit is $f(x) = \lim_{\nu\to\infty}\int_a^x f'_\nu = \int_a^x g$ by Theorem convit 4.9 so $f' = g$. The same argument can be applied in each variable for functions of several variables.

Another limiting process is the integral transformation defined by a kernel function $k(\mathbf{x},\mathbf{y})$ of the variables $\mathbf{x} \in D \subset \mathbb{R}^m$ and $\mathbf{y} \in E \subset \mathbb{R}^n$ for compact Jordan domains D and E. An integrable function $f(\mathbf{y})$ of the variable $\mathbf{y} \in E$ can be transformed to the function

kernq1 (4.32)
$$F(\mathbf{x}) = \int_{\mathbf{y}\in E} k(\mathbf{x},\mathbf{y})f(\mathbf{y})$$

of the variable $\mathbf{x} \in D$, provided that the kernel $k(\mathbf{x},\mathbf{y})$ is an integrable function of the variable $\mathbf{y} \in E$ for each point $\mathbf{x} \in D$. Of course the properties of the transform $F(\mathbf{x})$ depend on the properties of the kernel $k(\mathbf{x},\mathbf{y})$ as a function of the variable $\mathbf{x} \in D$; a sufficiently regular kernel $k(\mathbf{x},\mathbf{y})$ can transform a merely integrable function $f(\mathbf{y})$ of the variable $\mathbf{y} \in E$ to a continuous or differentiable function $F(\mathbf{x})$ of the variable $\mathbf{x} \in E$. The regularity conditions to be considered here are first that the kernel $k(\mathbf{x},\mathbf{y})$ is a continuous function of both variables $(\mathbf{x},\mathbf{y}) \in D \times E$, and second that in addition the kernel $k(\mathbf{x},\mathbf{y})$ is a differentiable function of the variable $\mathbf{x}$ and the derivative is continuous in both variables; more precisely the second condition is that for each fixed point $\mathbf{y} \in E$ the function $k(\mathbf{x},\mathbf{y})$ is a differentiable function of the variable x in an open neighborhood of the set $\overline{D}$ and the gradient of $k(\mathbf{x},\mathbf{y})$ when that function is viewed as a function of the variable $\mathbf{x}$ alone for a fixed point $\mathbf{y}$, which is a well defined vector-valued function $\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{y})$ of the variables $(\mathbf{x},\mathbf{y})$, is a continuous function in that neighborhood.

kern **Theorem 4.11** (i) *If $D \subset \mathbb{R}^m$ and $E \subset \mathbb{R}^n$ are compact Jordan domains, $k(x,y)$ is a continuous function of the variables $x \in D$, $\mathbf{y} \in E$, and $f(\mathbf{y})$ is a Riemann integrable function of the variable $\mathbf{y} \in E$, then the transform $F(\mathbf{x}) = \int_{\mathbf{y}\in E} k(\mathbf{x},\mathbf{y})f(\mathbf{y})$ is a continuous function of the variable $\mathbf{x} \in D$.*
(ii) *If in addition $k(\mathbf{x},\mathbf{y})$ is continuous in both variables and differentiable in the variable $\mathbf{x}$ in an open neighborhood of $D \times E$ with derivatives that are continuous in both variables, then the transform $F(\mathbf{x}) = \int_{\mathbf{y}\in E} k(\mathbf{x},\mathbf{y})f(\mathbf{y})$ is a continuously differentiable function of the variable $\mathbf{x} \in D$ and*

kernq2 (4.33)
$$\nabla F(\mathbf{x}) = \int_{\mathbf{y}\in E} \nabla_x k(\mathbf{x},\mathbf{y})f(\mathbf{y}).$$

Proof: (i) The integrable function $f(\mathbf{y})$ is bounded, so there is a constant M such that $|f(\mathbf{y})| \le M$ for all $\mathbf{y} \in E$. The continuous function $k(\mathbf{x},\mathbf{y})$ in the compact set $D \times E$ is uniformly continuous, so for any $\epsilon > 0$ there is a $\delta > 0$ such that $|k(x_1,y_1) - k(x_2,y_2)| < \epsilon$ whenever $\|(x_1,y_1) - (x_2,y_2)\|_2 < \delta$. Therefore if

$\mathbf{x} \in D$ and $\mathbf{h} \in \mathbb{R}^m$ satisfies $\|\mathbf{h}\|_2 < \delta$ then

$$\begin{aligned}|F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x})| &= \left|\int_{\mathbf{y}\in E}\Big(k(\mathbf{x}+\mathbf{h},\mathbf{y}) - k(\mathbf{x},\mathbf{y})\Big)f(\mathbf{y})\right| \\ &\le \int_{\mathbf{y}\in E} |k(\mathbf{x}+\mathbf{h},\mathbf{y}) - k(\mathbf{x},\mathbf{y})|\,|f(\mathbf{y})| \\ &\le \int_{\mathbf{y}\in E} \epsilon M = \epsilon M|E|,\end{aligned}$$

so $F(\mathbf{x})$ is a continuous function of the variable $\mathbf{x} \in D$.
(ii) By hypothesis the kernel $k(\mathbf{x},\mathbf{y})$ is continuous in the closure of an open neighborhood U of the set $D \times E$ and is a differentiable function of the variable $\mathbf{x}$ in U; so by the mean value theorem, Theorem 2.7, for any point $\mathbf{x} \in D$ and for any $\mathbf{h} \in \mathbb{R}^m$ sufficiently small

$$k(\mathbf{x}+\mathbf{h},\mathbf{y}) - k(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{x}} k(\mathbf{z},\mathbf{y}) \cdot \mathbf{h}$$

for some point $\mathbf{z} \in U$ between $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ on the line joining these two points. Then

$$\begin{aligned}&F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) - \int_{\mathbf{y}\in E} \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y}) \cdot \mathbf{h} f(\mathbf{y}) \\ &\quad = \int_{\mathbf{y}\in E}\Big(k(\mathbf{x}+\mathbf{h},\mathbf{y}) - k(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y}) \cdot \mathbf{h}\Big)f(\mathbf{y}) \\ &\quad = \int_{\mathbf{y}\in E}\Big(\nabla_{\mathbf{x}} k(\mathbf{z},\mathbf{y}) - \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\Big) \cdot \mathbf{h} f(\mathbf{y}).\end{aligned}$$

Since the function $\nabla_{\mathbf{x}} k(\mathbf{z},\mathbf{y})$ is assumed to be continuous on the compact set $\overline{U}$ it is uniformly continuous, and consequently for any $\epsilon > 0$ there is a $\delta > 0$ sufficiently small that whenever $\|\mathbf{h}\|_2 < \delta$

$$\|\nabla_{\mathbf{x}} k(\mathbf{z},\mathbf{y}) - \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\|_2 < \epsilon$$

and hence by the Cauchy-Schwartz inequality

$$\left|\Big(\nabla_{\mathbf{x}} k(\mathbf{z},\mathbf{y}) - \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\Big) \cdot \mathbf{h}\right| < \epsilon\|\mathbf{h}\|_2.$$

Thus if $|f(\mathbf{y})| \le M$ for all $\mathbf{y} \in E$ then

$$\left|F(\mathbf{x}+\mathbf{h}) - F(\mathbf{x}) - \int_{\mathbf{y}\in E} \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y}) \cdot \mathbf{h} f(\mathbf{y})\right| \le \int_{\mathbf{y}\in E} \epsilon M\|\mathbf{h}\|_2 \le \epsilon M|E|\|\mathbf{h}\|_2,$$

which shows that $F(\mathbf{x})$ is differentiable and that its derivative is given by (4.33), which suffices for the proof.

The integral has been defined so far just for integrable functions in a Jordan domain in $\mathbb{R}^n$. It is possible to extend the integral as an **improper integral** to a wider class of functions over a wider class of regions of integration, in particular to integrals over arbitrary open subsets $U \subset \mathbb{R}^n$ of functions that are continuous almost everywhere in U but are not necessarily bounded. The extension is called an improper integral since it does not necessarily satisfy all the conditions that the proper integals of Riemann integrable functions over Jordan domains satisfies; but with a little bit of care the extension is nonetheless very useful. The extension is a nontrivial extension, not just in that it provides integrals for functions that are not necessarily bounded but that it allows integrals over arbitrary open sets, bounded or unbounded; and not even all bounded open subsets of $\mathbb{R}^m$ are Jordan domains. For an example in $\mathbb{R}^1$, let r_j be a list of the rational numbers in the open unit interval $(0,1)$ where $j \ge 1$, let Δ_j be an open interval of length at most $|\Delta_j| = \frac{\epsilon}{2^j}$ contained in $(0,1)$ and containing the point r_j where $\epsilon < \frac{1}{2}$, and let $U = \bigcup \Delta_j$. Clearly $\overline{U} = [0,1]$, and it follows that $\partial U = [0,1] \sim U$. If U were a Jordan domain its boundary would be a compact set of content 0, which could be covered by finitely many open intervals of total length less than ϵ; thus there would be finitely many open intervals $I_j \subset \mathbb{R}^1$ such that $([0,1] \sim U) \subset \bigcup_{j=1}^N I_j$ and $\sum_{j=1}^N |I_j| < \epsilon$, and consequently

$$[0,1] \subset \Big(\bigcup_{j=1}^N I_j \cup \bigcup_{i=1}^\infty \Delta_i \Big).$$

However the compact set $[0,1]$ would be covered by finitely many of these open intervals, so that actually

$$[0,1] \subset \Big(\bigcup_{j=1}^N I_j \cup \bigcup_{i=1}^M \Delta_i \Big),$$

which is impossible since $\sum_{j=1}^n |I_j| + \sum_{i=1}^M |\Delta_i| < 2\epsilon < 1$; consequently U cannot be a Jordan domain.

gjop **Theorem 4.12** *Any open subset $U \subset \mathbb{R}^n$ can be written as a union $U = \bigcup_{i=1}^\infty D_i$ where D_i are open Jordan domains such that $\overline{D_i} \subset D_{i+1}$ for all indices i. Moreover these Jordan domains can be assumed to be unions of cells in a partition of $\mathbb{R}^n$.*

Proof: Let $\mathcal{P}_r$ be a partition of the entire space $\mathbb{R}^n$ into cells having as their sides intervals of length 2^{-r}. For a sufficiently large index r_1 there will be some closed cells in this partition that are contained in the open set U; the union of their interiors D_1 then is an open Jordan domain contained in U. For a sufficiently large index $r_2 > r_1$ each point on the compact boundary ∂D_1 will be contained in some open cells of the partition $\mathcal{P}_{r_2}$, and the closures of these cells also will be contained within U; the union D_2 of all of the open cells of this partition that are contained within U is another Jordan domain such that $\overline{D_1} \subset D_2$. The process continues, yielding a sequence of open Jordan domains D_i such that $\overline{D_i} \subset D_{i+1}$; and since the sizes of the cells in this partition tend to 0 any point in U will eventually lie in one of the Jordan domains D_i, which suffices to conclude the proof.

Consider then an open subset $U \subset \mathbb{R}^n$ that is a union $U = \bigcup_{i=1}^{\infty} D_i$ of open Jordan domains D_i such that $\overline{D_i} \subset D_{i+1}$ for all indices i. If f is a function that is continuous almost everywhere in U set

psng
$$f_+(\mathbf{x}) = \max(f(\mathbf{x}), 0) \quad \text{and} \quad f_-(\mathbf{x}) = \max(-f(\mathbf{x}), 0) \tag{4.34}$$

for any point $\mathbf{x} \in U$. Clearly $f_+(\mathbf{x}) \ge 0$ and $f_-(\mathbf{x}) \ge 0$ at all points $\mathbf{x} \in U$, and $f(\mathbf{x}) = f_+(\mathbf{x}) - f_-(\mathbf{x})$ is a unique decomposition of the function f as a difference of two non-negative functions; and both functions f_+ and f_- are also continuous almost everywhere in U. For any integer $i > 0$ the functions $\min(f_+(\mathbf{x}), i)$ and $\min(f_-(\mathbf{x}), i)$ also are bounded and continuous almost everywhere in U, and consequently they are integrable over the Jordan domains D_j. Moreover since these are positive functions and $\min(f_+(\mathbf{x}), i) \le \min(f_+(\mathbf{x}), i+1)$ while $D_i \subset D_{i+1}$ it follows that

$$\int_{D_i} \min(f_+(\mathbf{x}), i) \le \int_{D_{i+1}} \min(f_+(\mathbf{x}), i+1);$$

consequently the sequence of real numbers $\int_{D_i} \min(f_+(\mathbf{x}), i)$ for $i = 1, 2 \ldots$ has a well defined limit as $i \to \infty$, either a real number or $+\infty$, so it is possible to define

dfimpintp
$$\int_D f_+ = \lim_{i\to\infty} \int_{D_i} \min(f_+(\mathbf{x}), i) \tag{4.35}$$

and correspondingly of course

dfimpintn
$$\int_D f_- = \lim_{i\to\infty} \int_{D_i} \min(f_-(\mathbf{x}), i). \tag{4.36}$$

If at least one of these two integrals is finite the improper integral of the function f is defined to be

dfimpin
$$\int_D f = \int_D f_+ - \int_D f_- \tag{4.37}$$

which is either a real number or $+\infty$ or $-\infty$. This appears to depend on the particular choice of the sequence of Jordan domains D_i; but the value of the improper integral is actually independent of that choice. Indeed if U also is the union $U = \bigcup_{j=1}^{\infty} E_j$ for some Jordan domains $E_j \subset \mathbb{R}^n$ such that $\overline{E_j} \subset E_{j+1}$ then the same construction leads to other integrals, which for convenience will be denoted by $\int_E f_+$ and $\int_E f_-$; of course $E = U$ as sets, but that common set is represented as an increasing union of Jordan domains in two ways. Since each set $\overline{D_i}$ is compact and $D_i \subset \bigcup_{j=1}^{\infty} E_j$ for the open sets E_j then D_i is actually contained in a finite union of the sets E_j, and consequently in the largest of the sets E_j in this finite union, say the set E_{j_i}; and since D_i is contained in E_j whenever $j > j_i$ it is possible to assume that $j_i \ge i$. Therefore

$$\int_{D_i} \min(f_+, i) \le \int_{E_{j_i}} \min(f_+, j_i) \le \int_E f_+,$$

and since that is the case for any i it follows that $\int_D f_+ \le \int_E f_+$; and of course the same holds for the integrals of the function f_-. On the other hand the argument can be reversed, so that the opposite inequality $\int_E f_+ \le \int_D f_+$ also holds, and consequently $\int_E f_+ = \int_D f_+$; and again the same holds for integrals of the function f_-, so altogether $\int_E f = \int_D f$.

Some care must be taken with these improper integrals, as already mentionned. For instance if f is an unbounded function that is continuous almost everywhere in an open subset Jordan domain U but $\int_U f = \infty$ then $\int_U (f+g) = 0$ where $g = -f$ while $\int_U g = -\infty$ so the equation $\int_U (f+g) = \int_U f + \int_u g$ does not really have any sense. On the other hand if f and g are functions that are continuous almost everywhere in an open subset and have finite integrals then since $(f+g)_+(\mathbf{x}) \le f_+(\mathbf{x}) + g_+(\mathbf{x})$ and $(f+g)_-(\mathbf{x}) \ge f_-(\mathbf{x}) + g_-(\mathbf{x})$ at all points $\mathbf{x} \in U$ it follows fairly readily that $\int_U (f+g) = \int_U f + \int_U g$. It is possible to define the content of an arbitrary open subset U as the improper integral $|U| = \int_U 1$, which is always well defined although possibly $|U| = +\infty$. There are other ways of extending the class of functions and domains for integration that arise in various applications. The advantage of the procedure discussed here is that it is intrinsic: it attaches to any function f that is continuous almost everywhere in an open subset U a unique value (unless that value would be the indeterminate form $\infty - \infty$) for the integral $\int_U f$. There are situations in which it is possible to attach to an integral that would otherwise be of the indeterminate form $\infty - \infty$ a well defined value by specifying a particular way of taking the limit. The classical example is that of the Cauchy principal value of the integral of a function that is unbounded in a neighborhood of a point, as for example

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \left(\int_{-1}^{-\epsilon} \frac{1}{x} + \int_{\epsilon}^{1} \frac{1}{x} \right) = 0,$$

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \left(\int_{-1}^{-\epsilon} \frac{1}{x} + \int_{2\epsilon}^{1} \frac{1}{x} \right) = -\log 2.$$

In this way it is possible to assign a definite value to the integral $\int_{-1}^{1} \frac{1}{x}$ that would otherwise be of the indeterminate form $\infty - \infty$; but the actual value depends upon the particular choice of the way in which the limit is taken.

4.4 Change of Variables

math218notes4

In discussing the effects of a change of variables in integration it is convenient to invoke the notion of an oriented integral. That notion is familiar in the study of integrals of functions of a single variable; but it is customarily merely mentioned in passing just as a matter of notation if it is noted explicitly at all. A cell in $\mathbb{R}^1$ is an interval, customarily denoted by $[a, b]$ where $a \le b$, and oriented in the direction from a to b, the customary orientation of the real line by increasing values of the real numbers. The interval bounded by two real numbers a and b also can be denoted by $\Delta(a, b)$, when it is viewed as an unoriented cell and consequently as a cell that is defined independently of which of the two boundary points a or b is the larger number so that $\Delta(a, b) = \Delta(b, a)$. The content of this cell is $|\Delta(a, b)| = |b - a|$, and the Riemann integral $\int_{\Delta(a,b)} f$ is defined as in (4.8). The **oriented integral** is defined in terms of the Riemann integral by

intor
$$\int_a^b f = \begin{cases} \int_{\Delta(a,b)} f & \text{if } a \le b, \\ -\int_{\Delta(a,b)} f & \text{if } a > b\ . \end{cases} \tag{4.38}$$

This is the integral over the cell $\Delta(a, b) = \Delta(b, a)$ when that cell is oriented in the direction of the change of the variable x as it moves from the boundary point a to the boundary point b, so in the positive direction of the natural orientation of the real line of the variable x if $a < b$ but in the negative direction, the inverse of the natural orientation of the real line of the variable x, if $a > b$. It is clear from (4.38) that

intorv
$$\int_a^b f = -\int_b^a f \tag{4.39}$$

for any order of the real numbers a and b. For example, the Riemann integral of a constant function $f(x) = k$ is $\int_{\Delta(a,b)} k = k\,|b - a|$, a result that is independent of which of the two real numbers a or b is the larger; on the other hand it follows readily from (4.38) that the oriented integral is $\int_a^b k = k(b - a)$, also a result that is independent of which of the two real numbers is larger. In particular the integrals over $\Delta(a, x)$ for any fixed point a can be viewed as defining functions of the real variable $x \in \mathbb{R}^1$, and

intorq1
$$\int_{\Delta(a,x)} k = k|x - a| \quad \text{and} \quad \int_a^x k = k(x - a); \tag{4.40}$$

both of these integrals are continuous functions of the variable $x \in \mathbb{R}^1$, but the oriented integral is differentiable at the point a while the Riemann integral is not, one way in which the oriented integral is more convenient. In general, if $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a continuous function on the real line then for the Riemann integral it follows readily that

intorq2a
$$\int_{\Delta(a,b)} f + \int_{\Delta(b,c)} f = \begin{cases} \int_{\Delta(a,c)} f & \text{if} \quad a < b < c, \\ \int_{\Delta(a,c)} f + 2\int_{\Delta(b,c)} f & \text{if} \quad a < c < b, \end{cases} \tag{4.41}$$

while on the other hand for the oriented integral it follows equally readily that

intorq2b (4.42)
$$\int_a^b f + \int_b^c f = \int_a^c f \quad \text{for any order of } a, b, c.$$

fundthc **Theorem 4.13 (Fundamental Theorem of Calculus)** *If f is a continuous function on an interval $[a,b] \subset \mathbb{R}^1$ then $F_0(x) = \int_a^x f$ is a continuously differentiable function of the variable $x \in (a,b)$ and $F_0'(x) = f(x)$; and conversely if $F(x)$ is a continuously differentiable function in (a,b) such that $F'(x) = f(x)$ then $\int_a^x f = F(x) - F(a)$ for $x \in (a,b)$.*

Proof: If $F_0(x) = \int_a^x f$ it follows from (4.42) that $F_0(x+h) - F_0(x) = \int_x^{x+h} f$ whenever $x, x+h \in [a,b]$, where h can be either positive or negative; but for convenience assume that $h > 0$, since the case $h < 0$ can be treated in a similar manner. Since f is continuous then for any fixed point $x \in [a,b]$ and any $\epsilon > 0$ if h is sufficiently small

$$f(x) - \epsilon \le f(t) \le f(x) + \epsilon \quad \text{for all } t \in \Delta(x, x+h);$$

the oriented integral of this inequality over the interval $\Delta(x, x+h)$ yields the inequality

$$\big(f(x) - \epsilon\big)h \le F_0(x+h) - F_0(x) \le (f(x) + \epsilon)h,$$

and dividing by h and taking the limit as h tends to 0 shows that

$$f(x) - \epsilon \le F_0'(x) \le f(x) + \epsilon;$$

since that is the case for all $\epsilon > 0$ it follows that $F_0'(x) = f(x)$. If $F(x)$ is a differentiable function in $[a,b]$ such that $F'(x) = f(x)$ then $F(x) = F_0(x) + c$ for some constant c, since $\frac{d}{dx}\Big(F(x) - F_0(x)\Big) = 0$, and consequently $\int_a^x f = F_0(x) - F_0(a) = F(x) - F(a)$, which suffices for the proof.

A $\mathcal{C}^1$ homeomorphism $\phi : [a,b] \longrightarrow \Delta\big(\phi(a), \phi(b)\big)$ either is orientation preserving, so that $\phi(a) \le \phi(b)$ and $\phi'(x) > 0$ for all points $x \in (a,b)$, or is orientation reversing, so that $\phi(a) \ge \phi(b)$ and $\phi'(x) < 0$ for all points $x \in (a,b)$. In either case, if f is a continuous function in $[c,d] = \Delta\big(\phi(a), \phi(b)\big)$ then for any differentiable function F in $[a,b]$ such that $F'(x) = f(x)$ it follows from Theorem 4.13 that

$$\int_{\phi(a)}^{\phi(b)} f = F\big(\phi(b)\big) - F\big(\phi(a)\big).$$

By the chain rule $(F \circ \phi)' = (f \circ \phi)\,\phi'$, so by Theorem 4.13 again

$$\int_a^b \big(f \circ \phi\big)\,\phi' = (F \circ \phi)(b) - (F \circ \phi)(a) = F\big(\phi(b)\big) - F\big(\phi(a)\big).$$

Comparing the two preceding observations shows that

intorq5 (4.43)
$$\int_{\phi(a)}^{\phi(b)} f = \int_a^b \big(f \circ \phi\big)\,\phi'.$$

This is the familiar formula for the change of variables in an oriented integral; actually it is customary to set $y = \phi(x)$, $\phi(a) = c$ and $\phi(b) = d$ and to rewrite the preceding equation in the form

$$\int_c^d f(y)dy = \int_a^b f\big(y(x)\big)\,\frac{dy}{dx}\,dx. \tag{4.44}$$

On the other hand for the Riemann integral it follows from (4.38) that

$$\int_{\Delta(a,b)} f = \int_a^b f \quad \text{since } a \le b \tag{4.45}$$

and

$$\int_{\Delta(\phi(a),\phi(b))} f = \begin{cases} \int_{\phi(a)}^{\phi(b)} f & \text{if} \quad \phi(a) \le \phi(b), \\ \\ -\int_{\phi(a)}^{\phi(b)} f & \text{if} \quad \phi(a) \ge \phi(b); \end{cases} \tag{4.46}$$

consequently in terms of the unoriented integral (4.43) takes the form

$$i\int_{\Delta\big(\phi(a),\phi(b)\big)} f = \begin{cases} \int_{\Delta(a,b)} (f\circ\phi)\phi' & \text{if} \quad \phi(a) \le \phi(b), \\ \\ -\int_{\Delta(a,b)} (f\circ\phi)\phi' & \text{if} \quad \phi(a) \ge \phi(b). \end{cases} \tag{4.47}$$

In the first case in (4.47) the mapping ϕ is orientation-preserving and $\phi'(x) \ge 0$, while in the second case in (4.47) the mapping ϕ is orientation-reversing and $\phi'(x) \le 0$, so both cases can be subsumed under the general formula

$$\int_{\phi(\Delta)} f = \int_\Delta (f\circ\phi)|\phi'| \tag{4.48}$$

for any cell $\Delta \subset \mathbb{R}^1$. This is the form in which the result extends to mappings in higher dimensions.

Theorem 4.14 *If $\boldsymbol{\phi} : D \longrightarrow E$ is a $\mathcal{C}^1$ homeomorphism between two open subsets in $\mathbb{R}^n$ and if f is a function that is Riemann integrable in E then the function $(f\circ\boldsymbol{\phi})\,|\det\boldsymbol{\phi}'|$ is Riemann integrable in D and*

$$\int_E f = \int_D (f\circ\boldsymbol{\phi})\,|\det\boldsymbol{\phi}'|. \tag{4.49}$$

Proof: As a preliminary observation, if $\boldsymbol{\phi} : D \longrightarrow E$ is a $\mathcal{C}^1$ homeomorphism between two improper Jordan domains $D, E \subset \mathbb{R}^n$ then for any cell $\Delta \subset \overline{\Delta} \subset D$ the boundary of the image of Δ is the image of the boundary of Δ, that is, $\partial\boldsymbol{\phi}(\Delta) = \boldsymbol{\phi}(\partial\Delta)$. Since each face of Δ is locally a submanifold of $\mathbb{R}^n$ of dimension $n-1$ it follows from the Rank Theorem, Theorem 3.7, that the boundary of the image $\boldsymbol{\phi}(\Delta)$ consists of a collection of pieces that are also locally submanifolds of $\mathbb{R}^n$ of dimension $n-1$, so are subsets of measure zero as noted in the earlier discussion; consequently $\partial\boldsymbol{\phi}(\Delta)$ is a set of measure zero so the image $\boldsymbol{\phi}(\Delta)$ is a Jordan domain in $\mathbb{R}^n$. A set that is the image of a cell Δ under a $\mathcal{C}^1$ homeomorphism of an open neighborhood of the closure $\overline{\Delta}$ will be called a **special Jordan domain**, although only for the duration of this proof. It is evident from this definition that the image of a special Jordan domain under a $\mathcal{C}^1$ homeomorphism defined in an open neighborhood of its closure is also a special Jordan domain.

To turn to the proof of the theorem itself, first it will be demonstrated that for any cell $\Delta \subset \overline{\Delta} \subset E$ and for any constant c

varcht1q2

$$\int_\Delta c = \int_{\boldsymbol{\phi}^{-1}(\Delta)} c\,|\det\boldsymbol{\phi}'| \tag{4.50}$$

for three special cases of homeomorphisms $\boldsymbol{\phi}$.

(i) If the homeomorphism $\boldsymbol{\phi}: D \longrightarrow E$ is a translation then $|\det\boldsymbol{\phi}'| = 1$, and it obvious that any partition of the cell Δ translates to a corresponding partition of the translated cell $\boldsymbol{\phi}^{-1}(\Delta)$ for which the upper sums and consequently the upper integrals of the constant c over the partitions of Δ and $\boldsymbol{\phi}^{-1}(\Delta)$ are equal; and since c is Riemann integrable it follows that $\int_\Delta c = \int_{\boldsymbol{\phi}^{-1}(\Delta)} c$, which is (4.50) in this case.

(ii) If the homeomorphism $\boldsymbol{\phi}: D \longrightarrow E$ is a permutation of the variables then $|\det\boldsymbol{\phi}'| = 1$ again, and since the upper sums and hence the upper integrals of the constant function c for any partition of the cell Δ are unchanged by a relabeling of the variables it follows again that $\int_\Delta c = \int_{\boldsymbol{\phi}^{-1}(\Delta)} c$, which is (4.50) in this case.

(iii) For homeomorphisms $\boldsymbol{\phi}: D \longrightarrow E$ that change only a single variable the demonstration is really the same for any particular variable which is changed; but the notation is slightly simpler for a homeomorphism that changes only the variable x_1, so the details will be discussed just in that case. When points $\mathbf{x} \in D$ and $\mathbf{y} \in E$ are written

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \mathbf{x}_{II} \end{pmatrix} \text{ where } \begin{cases} x_1 \in \mathbb{R}^1 \\ \mathbf{x}_{II} \in \mathbb{R}^{n-1} \end{cases} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \mathbf{y}_{II} \end{pmatrix} \text{ where } \begin{cases} y_1 \in \mathbb{R}^1 \\ \mathbf{y}_{II} \in \mathbb{R}^{n-1} \end{cases}$$

the homeomorphism $\boldsymbol{\phi}$ has the form

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{pmatrix} \phi_1(\mathbf{x}) \\ \mathbf{x}_{II} \end{pmatrix} \quad \text{where} \quad \begin{cases} \phi_1(\mathbf{x}) \in \mathbb{R}^1 \\ \mathbf{x}_{II} \in \mathbb{R}^{n-1} \end{cases};$$

therefore

$$\boldsymbol{\phi}'(\mathbf{x}) = \begin{pmatrix} \partial_1\phi_1(\mathbf{x}) & \partial_{II}\phi_1(\mathbf{x}) \\ 0 & I_{n-1} \end{pmatrix}$$

where I_{n-1} is the $(n-1)\times(n-1)$ identity matrix, so $\det\boldsymbol{\phi}'(\mathbf{x}) = \partial_1\phi_1(\mathbf{x})$. The cell $\Delta \subset E$ is the product of an interval Δ_1 in the x_1-axis and a cell Δ_{II} in the space $\mathbb{R}^{n-1}$ of the variables $\{\mathbf{y}_{II}\}$, so it is the union of the one-dimenesional slices $\Delta(\mathbf{y}_{II}) = \Delta_1 \times \mathbf{y}_{II}$ for points $\mathbf{y}_{II} \in \Delta_{II}$. The image of the cell Δ

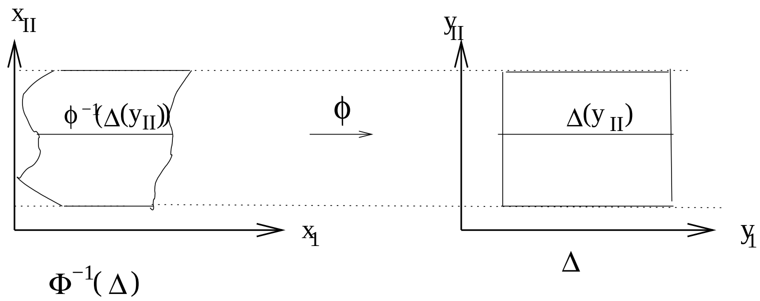

Figure 4.5: Mappings that change a single variable varch1

under the inverse homeomorphism $\boldsymbol{\phi}^{-1} : E \longrightarrow D$ is the special Jordan domain $\boldsymbol{\phi}^{-1}(\Delta) \subset D$, which correspondingly is the union of the one-dimensional slices

$$\boldsymbol{\phi}^{-1}\big(\Delta(\mathbf{y}_{II})\big) = \Big\{ \mathbf{x} = \{x_1, \mathbf{x}_{II}\} \in \boldsymbol{\phi}^{-1}(\Delta) \ \Big| \ \mathbf{x}_{II} = \mathbf{y}_{II} \Big\}$$

as sketched in the accompanying Figure 4.5. For the constant function $f = c$ in the cell Δ it follows from Fubini's Theorem and the change of variables formula (4.43) for the restriction $\boldsymbol{\phi} : \boldsymbol{\phi}^{-1}\big(\Delta(\mathbf{y}_{II})\big) \longrightarrow \Delta(\mathbf{y}_{II})$ of the homeomorphism $\boldsymbol{\phi}$ to the one-dimensional slices that

$$\begin{aligned} \int_\Delta c &= \int_{\mathbf{y}_{II}\in\Delta_{II}} \int_{y_1 \in \Delta(y_{II})} c = \int_{\mathbf{x}_{II}\in\Delta_{II}} \int_{x_1 \in \boldsymbol{\phi}^{-1}\left(\Delta(\mathbf{x}_{II})\right)} c\ |\boldsymbol{\phi}_1'| \\ &= \int_{\mathbf{x}_{II}\in\Delta_{II}} \int_{x_1 \in \boldsymbol{\phi}^{-1}\left(\Delta(\mathbf{x}_{II})\right)} c\ |\det \boldsymbol{\phi}'| = \int_{\boldsymbol{\phi}^{-1}(\Delta)} c\ |\det \boldsymbol{\phi}'|, \end{aligned}$$

which is equation (4.50) in this case.

Next it will be demonstrated that for any cell $\Delta \subset \overline{\Delta} \subset E$ and for any function f that is Riemann integrable in Δ the function $(f \circ \boldsymbol{\phi})|\det \boldsymbol{\phi}'|$ is Riemann integrable in $\boldsymbol{\phi}^{-1}(\Delta)$ and

varcht1q3 (4.51)
$$\int_\Delta f = \int_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\ |\det \boldsymbol{\phi}'|,$$

provided again that the homeomorphism $\boldsymbol{\phi} : D \longrightarrow E$ is any one of the three special cases just discussed. Of course it is clear that $(f \circ \boldsymbol{\phi})|\det \boldsymbol{\phi}'|$ is bounded in $\boldsymbol{\phi}^{-1}(\Delta)$, but the remainder is to be proved. For any partitition $\mathcal{P}$ of the cell $\overline{\Delta}$ as the union $\overline{\Delta} = \cup_i \overline{\Delta}_i$, introduce the function f^* in $\overline{\Delta}$ defined by

varcht1q4 (4.52)
$$f^*(\mathbf{y}) = \begin{cases} \sup_{\mathbf{z}\in\overline{\Delta}_i} f(\mathbf{z}) & \text{if} \quad \mathbf{y} \in \Delta_i, \\ \\ f(\mathbf{y}) & \text{if} \quad \mathbf{y} \in \bigcup_j \partial\Delta_j. \end{cases}$$

The function f^* is constant in each open cell Δ_i of the partition $\mathcal{P}$, so for any homeomorphism $\boldsymbol{\phi} : D \longrightarrow E$ of one of the three special cases considered earlier it follows from what has been demonstrated already that

varcht1q5 (4.53)
$$\int_{\Delta_i} f^* = \int_{\boldsymbol{\phi}^{-1}(\Delta_i)} (f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|.$$

The function f^* is continuous in Δ except for the union of the boundaries of the cells Δ_i, and that union is a set of measure 0 so f^* is Riemann integrable over Δ, as is $(f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|$. The special Jordan domains Δ_i are disjoint, as are the special Jordan domains $\boldsymbol{\phi}^{-1}(\Delta_i)$, so it follows from (4.53) and Theorem 4.6 that

varcht1q5a (4.54)
$$\int_{\Delta} f^* = \sum_i \int_{\Delta_i} f^* = \sum_i \int_{\boldsymbol{\phi}^{-1}(\Delta_i)} (f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'| = \int_{\boldsymbol{\phi}^{-1}(\Delta)} (f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|.$$

From the definition (4.52) it further follows that $\int_{\Delta_i} f^* = \left(\sup_{\mathbf{y}\in\Delta_i} f(\mathbf{y})\right)\,|\Delta_i|$, hence

varcht1q6 (4.55)
$$\int_{\Delta} f^* = \sum_i \int_{\Delta_i} f^* = \sum_i \left(\sup_{\mathbf{y}\in\Delta_i} f(\mathbf{y})\right)\,|\Delta_i| = S^*(f,\mathcal{P}).$$

It is also clear from (4.52) that $(f^* \circ \boldsymbol{\phi})(\mathbf{x})\,|\det \boldsymbol{\phi}'(\mathbf{x})| \ge (f \circ \boldsymbol{\phi})(\mathbf{x})\,|\det \boldsymbol{\phi}'(\mathbf{x})|$ at all points $\mathbf{x} \in \Delta$, so the upper sums of these two functions for any partition of a cell containing the special Jordan domain $\boldsymbol{\phi}^{-1}(\Delta)$ satisfy the corresponding inequality and consequently so do the upper integrals, hence

varcht1q7 (4.56)
$$\int_{\boldsymbol{\phi}^{-1}(\Delta)} (f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'| = \int^*_{\boldsymbol{\phi}^{-1}(\Delta)} (f^* \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'| \ge \int^*_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|.$$

It then follows from (4.54), (4.55) and (4.56) that

varcht1q8 (4.57)
$$S^*(f,\mathcal{P}) \ge \int^*_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|;$$

and since that is the case for any partition $\mathcal{P}$ of the cell Δ it follows that

varcht1q8x (4.58)
$$\int^*_{\Delta} f \ge \int^*_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|.$$

The corresponding argument for the lower integrals, for which the inequalities are reversed, shows that

varcht1q8y (4.59)
$$\int_{*\Delta} f \le \int_{*\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|,$$

and from (4.58) and (4.59) it follows that

varacht1q9 (4.60)
$$\int^*_{\Delta} f \ge \int^*_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'| \ge \int_{*\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'| \ge \int_{*\Delta} f.$$

Since the functiom f is Riemann integrable by hypothesis $\int_{\Delta}^{*} f = \int_{*\Delta} f = \int_{\Delta} f$ so the inequalities in (4.60) are equalities; hence the function $(f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|$ is Riemann integrable over $\boldsymbol{\phi}^{-1}(\Delta)$ and (4.51) holds as desired.

Next it will be demonsrated that if $\boldsymbol{\phi} : D \longrightarrow E$ is an arbitrary $\mathcal{C}^1$ homeomorphism and if f is a function that is Riemann integrable in E then for any point $\mathbf{b}_0 \in E$ and for any sufficiently small cell Δ such that $\mathbf{b}_0 \in \Delta \subset \overline{\Delta} \subset E$

varcht1qx1

$$\int_{\Delta} f = \int_{\boldsymbol{\phi}^{-1}(\Delta)} (f \circ \boldsymbol{\phi})\,|\det \boldsymbol{\phi}'|. \tag{4.61}$$

Indeed since $\det \boldsymbol{\phi}'\big(\boldsymbol{\phi}^{-1}(\mathbf{b}_0)\big) \ne 0$ for a $\mathcal{C}^1$ homeomorphism $\boldsymbol{\phi} : D \longrightarrow E$ it follows from Corollary 3.1 that over a sufficiently small open cell $\Delta \subset \overline{\Delta} \subset E$ containing $\mathbf{b}_0$ the homeomorphism $\boldsymbol{\phi}$ can be written as the composition

varcht1qy1

$$\boldsymbol{\phi} = \boldsymbol{\phi}_r \circ \boldsymbol{\phi}_{r-1} \circ \cdots \circ \boldsymbol{\phi}_2 \circ \boldsymbol{\phi}_1 \tag{4.62}$$

of $\mathcal{C}^1$ homeomorphisms that are either translations, permutations of the variables, or mappings that change only a single variable. By what has been demonstrated so far, each of these separate mappings satisfy the basic formula (4.49). Although that was only demonstrated for mappings to a cell, when the mapping is to a special Jordan domain the image can be viewed as a subset of a cell and the integral formula for the special cases of mappings to a cell can be applied. To demonstrate (4.61) it suffices then merely to show that if the basic formula (4.49) holds for two $\mathcal{C}^1$ homeomorphisms $\boldsymbol{\phi} : D \longrightarrow E$ and $\boldsymbol{\psi} : E \longrightarrow F$ then it also holds for the composition $\boldsymbol{\psi} \circ \boldsymbol{\phi} : D \longrightarrow F$. For that purpose, if f is a Riemann integrable function in F then by (4.61) applied to the two mappings $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ separately

$$\begin{aligned} \int_F f = \int_E (f \circ \boldsymbol{\psi})|\det \boldsymbol{\psi}'| &= \int_D \big((f \circ \boldsymbol{\psi}) \circ \boldsymbol{\phi}\big)|(\det \boldsymbol{\psi}') \circ (\boldsymbol{\phi})||\det \boldsymbol{\phi}'| \\ &= \int_D \big(f \circ (\boldsymbol{\psi} \circ \boldsymbol{\phi})\big)|\det(\boldsymbol{\psi} \circ \boldsymbol{\phi})'| \end{aligned}$$

since $(\boldsymbol{\psi} \circ \boldsymbol{\phi})'(\mathbf{x}) = \boldsymbol{\psi}'\big(\boldsymbol{\phi}(\mathbf{x})\big)\boldsymbol{\phi}'(\mathbf{x})$ by the chain rule for differentiation.

Finally to complete the proof of the theorem itself, by Theorem 4.12 any open subset $E \in \mathbb{R}^n$ can be written as the union of an increasing sequence of open Jordan domains E_i, each of which is a union $E_i = \bigcup_j \Delta_{ij}$ of cells Δ_{ij} that are disjoint aside from their boundaries; and D is correspondingly the union of the increasing sequence of open subsets $D_i = \boldsymbol{\phi}^{-1}(E_i)$ each of which is a union $D_i = \bigcup \boldsymbol{\phi}^{-1}(\Delta_{ij})$ of special Jordan domains that are disjoint aside from their boundaries. By passing to refinements of these cells it can be assumed by what already has been demonstrated that $\int_{\Delta_{ij}} f = \int_{\boldsymbol{\phi}^{-}(\Delta_{ij})} (f \circ \boldsymbol{\phi})|\det \boldsymbol{\phi}'|$ for any Riemann integrable function in E; and it then follows from Theorerm 4.6 that

$$\int_{E_i} f = \int_{D_i} (f \circ \boldsymbol{\phi})|\det \boldsymbol{\phi}'|$$

for each set E_i, so (4.49) holds in the limit for the integrals as in (4.35) and (4.36) and that suffices for the proof.

For an explicit example, consider the integral $\int_E \log(x_1^2 + x_2^2)$ where

$$E = \left\{ (x_1, x_2) \in \mathbb{R}^2 \;\middle|\; a \leq x_1^2 + x_2^2 \leq b, \;\; x_1 \geq 0, \;\; x_2 \geq 0 \right\}.$$

This is most convenently handled by introducing polar coordinates through the change of coordinates

$$\boldsymbol{\phi} : \mathbb{R}^2(x_1, x_2) \longrightarrow \mathbb{R}^2(r, \theta)$$

where $x_1 = r\cos\theta,\; x_2 = r\sin\theta$. The mapping $\boldsymbol{\phi}$ establishes a one-to-one mapping $\boldsymbol{\phi} : D \longrightarrow E$ where

$$D = \left\{ (r, \theta) \in \mathbb{R}^2 \;\middle|\; a \leq r \leq b, \;\; 0 \leq \theta \leq \tfrac{\pi}{2} \right\};$$

and

$$\boldsymbol{\phi}'(r, \theta) = \frac{\partial(x_1, x_2)}{\partial(r, \theta)} = \begin{pmatrix} \dfrac{\partial x_1}{\partial r} & \dfrac{\partial x_1}{\partial \theta} \\[2ex] \dfrac{\partial x_2}{\partial r} & \dfrac{\partial x_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

so $|\det \boldsymbol{\phi}'(r, \theta)| = r$. The formula for the change of variables and an application of Fubini's Theorem show that

$$\begin{aligned} \int_E \log(x_1^2 + x_2^2) &= \int_D \log(r^2)\, r \\ &= \int_a^b \int_0^{\pi/2} 2r \log r \; d\theta \, dr \\ &= \frac{\pi}{2} \int_a^b 2r \log r = \frac{\pi}{2} r^2 (\log r - \frac{1}{2})\Big|_a^b \\ &= \frac{\pi}{2} \pi \left(b^2 \log b - a^2 \log a \right) - \frac{\pi}{4}(b^2 - a^2). \end{aligned}$$

PROBLEMS: GROUP 1

(1) Evaluate the following integrals:

(i) $\int_E x_1^2$ where E is the set $x_1 \geq 0,\ 1 \leq x_1^2 + x_2^2 \leq 2$.

(ii) $\int_E x_1 e^{x_2}$ where $E \subset \mathbb{R}^2$ is the triangle with vertices $(0,0),(0,2),(2,2)$.

(iii) $\int_E x_2 e^{x_1 x_2}$ where $E \subset \mathbb{R}^2$ is the set defined by the conditions

$$1 \leq x_1 \leq 2,\ \ x_2 \leq 1,\ \ x_1 x_2 \geq 1.$$

(2) Show that the integral $\int_{R^2} x_1 x_2 e^{-x_1^2 - x_2^2}$ converges, and evaluate it.

(3) Consider the integral $\int_E x_3$ where $E \subset \mathbb{R}^3$ is the set defined by the conditions

$$x_1 \geq 0,\ \ x_2 \geq 0,\ \ 0 \leq x_3 \leq 1,\ \ x_1^2 + x_2^2 \leq 1.$$

(i) Write this integral as an interated integral in 3 different orders.

(ii) Evaluate this integral, by an iterated integration in any order you choose.

(4) Evaluate the following integrals, using a change of variables to simplify the calculation:

(i) $\int_E e^{-x_1^2 - x_2^2}$ where E is the set of points (x_1, x_2) in the unit disc in $\mathbb{R}^2$ for which $x_2 \geq 0$.

(ii) $\int_E \frac{1}{\sqrt{x_1^2 + x_2^2}}$ where E is the set of points $\mathbf{x} = (x_1, x_2)$ in the plane $\mathbb{R}^2$ for which $1 \leq \|\mathbf{x}\| \leq 2$.

(5) Spherical coordinates in $\mathbb{R}^3$ are defined by

$$x_1 = r \sin\phi \cos\theta,\ x_2 = r \sin\phi \sin\theta,\ x_3 = r \cos\phi.$$

Find explicitly the formula for the change of variables expressing integration in $\mathbb{R}^3$ in terms of spherical coordinates. Use this to evaluate the integral $\int_E r$ of the function $r = \sqrt{x_1^2 + x_2^2 + x_3^2} \geq 0$ over the solid ball E of radius 1.

PROBLEMS: GROUP 2

(6) Show that if $E \subset \mathbb{R}^n$ has content 0 then its boundary ∂E also has content 0, but that if $E \subset \mathbb{R}^n$ has measure 0 its boundary ∂E does not necessarily have measure 0.

(7) Show that if f is a non-negative integrable function in a cell $\overline{\Delta}$ such that $\int_{\overline{\Delta}} f = 0$ then $\{\, \mathbf{x} \in \overline{\Delta} \mid f(\mathbf{x}) \neq 0 \,\}$ is a set of measure 0.

(8) Show that if f is a monotonically increasing real-valued function in an interval $[a, b] \subset \mathbb{R}^1$ then f is necessarily continuous almost everywhere in $[a, b]$, and consequently is an integrable function.

(9) Let $E \subset \overline{\Delta}$ be a set of content 0 contained in the cell

$$\overline{\Delta} = \left\{\, (x_1, x_2) \in \mathbb{R}^2 \;\middle|\; a_1 \leq x_1 \leq b_1, \;\; a_2 \leq x_2 \leq b_2 \,\right\},$$

and let $F \subset [a_1, b_1]$ be the set of those points $x_1 \in [a_1, b_1]$ such that the set

$$F_{x_1} = \left\{\, x_2 \in [a_2, b_2] \;\middle|\; (x_1, x_2) \in E \,\right\} \subset [a_2, b_2]$$

as a subset of the one-dimensional cell $[a_2, b_2]$ does not have content 0. Show that F is a set of measure 0 in the interval $[a_1, b_1]$. [Suggestion: the characteristic function χ_E is integrable, and Fubini's Theorem can be applied to it.]

(10) Show that if $E \subset \Delta$ is a Jordan domain then for any $\epsilon > 0$ there exists a compact Jordan domain $F \subset E$ such that $|E \sim F| < \epsilon$.

(11) Use polar coordinates to evaluate the integral $\int_E (1 + x_1^2 + x_2^2)^{-2}$ over the triangular region bounded by the triangle with the vertices $(0, 0)$, $(2, 0)$, $(1, \sqrt{3})$. (There are other ways of evaluating this integral of course, but for this problem please use polar coordinates.)

Chapter 5

Differential Forms

math218notes5

5.1 Line Integrals

lineint

A particularly interesting special case of integration over proper subsets of $\mathbb{R}^n$ is that of integrals over curves in $\mathbb{R}^n$. As on page rankthcr0 60, a **parametrized curve** in an open subset $U \subset \mathbb{R}^n$ is a continuous mapping $\boldsymbol{\phi} : [a, b] \longrightarrow U$ from a closed interval into $\mathbb{R}^n$; the image of this mapping is a **curve** in U, a rather special type of subset $\boldsymbol{\phi}([a, b]) \subset U$. It is useful to distinguish between a mapping and its image, so to distinguish between parametrized curves and curves; a parametrized curve has as its image a unique curve, but a curve can be the image of a great many different parametrized curves. Much of the subsequent discussion will focus on $\mathcal{C}^1$ parametrized curves $\boldsymbol{\phi}$, those for which the mapping extends to a $\mathcal{C}^1$ mapping $\boldsymbol{\phi} : (a-\epsilon, b+\epsilon) \longrightarrow U$ for some $\epsilon > 0$; but it is common also to consider $\mathcal{C}^k$ parametrized curves for any integer $k > 0$, or even $\mathcal{C}^\infty$ parametrized curves, those which are $\mathcal{C}^k$ parametrized curves for all $k > 0$. Unless explicitly assumed otherwise, however, parametrized curves are merely continuous mappings. Also unless explicitly assumed otherwise, parametrized curves are not required to be one-to-one mappings, nor are $\mathcal{C}^1$ parametrized curves required to have a nonvanishing derivative; thus the images of $\mathcal{C}^1$ parametrized curves are considerably more general than one-dimensional $\mathcal{C}^1$ submanifolds. Two parametrized curves $\boldsymbol{\phi} : [a, b] \longrightarrow U$ and $\boldsymbol{\psi} : [c, d] \longrightarrow U$ are **equivalent** parametrized curves if there is a homeomorphism $h : [a, b] \longrightarrow [c, d]$ such that $\boldsymbol{\phi}(t) = \boldsymbol{\psi}\big(h(t)\big)$ for all $t \in [a, b]$; this is a natural equivalence relation among parametrized curves. It is clear that if $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are equivalent parametrized curves then $\boldsymbol{\phi}([a, b]) = \boldsymbol{\psi}([c, d])$; thus equivalent parametrized curves have the same image curves, but of course parametrized curves having the same image curves are not necessarily equivalent parametrized curves. In a natural extension of this definition, two $\mathcal{C}^1$ parametrized curves $\boldsymbol{\phi} : [a, b] \longrightarrow U$ and $\boldsymbol{\psi} : [c, d] \longrightarrow U$ are $\mathcal{C}^1$**-equivalent** parametrized curves if the mapping $h : [a, b] \longrightarrow [c, d]$ is a $\mathcal{C}^1$ homeomorphism, and of course correspondingly for $\mathcal{C}^k$ parametrized curves. It is often assumed as a standard convention that the parameter interval for a curve is the unit interval $[0,1]$, although other normalizations are sometimes more appropriate.

If $\boldsymbol{\phi} : [0,1] \longrightarrow U$ is a parametrized curve in an open subset $U \subset \mathbb{R}^n$ and if $f : U \longrightarrow \mathbb{R}$ is a function in U, the composition $f \circ \boldsymbol{\phi}$, also denoted by $\boldsymbol{\phi}^*(f)$, is a well defined function on the parameter interval $[0,1]$ called the **induced function** or the **pullback** of the function f under the mapping $\boldsymbol{\phi}$; and if the function f is continuous then the induced function $\boldsymbol{\phi}^*(f)$ is a continuous function in $[0,1]$. The **integral** of a continuous function f in U on the parametrized curve $\boldsymbol{\phi}$ is defined by

indintdf
$$\int_{\boldsymbol{\phi}} f = \int_{[0,1]} \boldsymbol{\phi}^*(f), \tag{5.1}$$

and has a well value for any parametrized curve. The integrals of a continuous function f along two equivalent parametrized curves in U need not be the same though. Indeed if $\boldsymbol{\phi} : [0,1] \longrightarrow \mathbb{R}^m$ and $\boldsymbol{\psi} : [0,1] \longrightarrow \mathbb{R}^n$ are $\mathcal{C}^1$ equivalent parametrized curves then $\boldsymbol{\phi} = \boldsymbol{\psi} \circ h$ for a continuously differentiable mapping $h : [0,1] \longrightarrow [0,1]$ and it follows from the change of variables theorem, Theorem varcht1 4.14, that $\int_{\boldsymbol{\psi}} f = \int_{[0,1]} f \circ \boldsymbol{\psi} = \int_{[0,1]} (f \circ \boldsymbol{\psi}) \circ h \ |h'| = \int_{[0,1]} f \circ \boldsymbol{\phi} \ |h'|$, which is not necessarily equal to $\int_{[0,1]} f \circ \boldsymbol{\phi} = \int_{\boldsymbol{\phi}} f$. Hence something further must be done in order to obtain an intrinsic integral of a function along a parametrized curve.

One way to obtain a more intrinsic integral is to use an intrinsic parametrization of the curve. If $\boldsymbol{\phi} : [0,1] \longrightarrow U$ is a parametrized curve with image curve $\gamma = \boldsymbol{\phi}([0,1]) \subset U$, then for any partition $\mathcal{P}$ of the interval $[0,1]$ by points of subdivision $0 = t_0 < t_1 < \cdots < t_m = 1$ the parametrized curve $\boldsymbol{\phi}$ can be approximated by the piecewise linear parametrized curve $\boldsymbol{\phi}_{\mathcal{P}}$ for which $\boldsymbol{\phi}_{\mathcal{P}}(t_i) = \boldsymbol{\phi}(t_i)$ for each point of subdivision while the mapping $\boldsymbol{\phi}_{\mathcal{P}}$ is extended to be a linear mapping between the points $\boldsymbol{\phi}(t_{i-1})$ and $\boldsymbol{\phi}(t_i)$, so that if $t_{i-1} \le t \le t_i$ then

dlgaqq
$$\boldsymbol{\phi}_{\mathcal{P}}(t) = \frac{t - t_i}{t_{i-1} - t_i} \boldsymbol{\phi}(t_{i-1}) + \frac{t - t_{i-1}}{t_i - t_{i-1}} \boldsymbol{\phi}(t_i). \tag{5.2}$$

This is rather like approximating the curve γ by the polygon joining the successive points $\boldsymbol{\phi}(t_i)$; but the mapping $\boldsymbol{\phi}$ may not be a one-to-one mapping, for instance it may go back and forth over various segments of the curve γ, and consequently the approximation of the mapping $\boldsymbol{\phi}$ by the piecewise linear mapping $\boldsymbol{\phi}_{\mathcal{P}}$ may differ significantly from a polygonal approximation of the point set γ. The length of the piecewise linear approximation $\boldsymbol{\phi}_{\mathcal{P}}$ to the parametrized curve $\boldsymbol{\phi}$ is

dlgapp
$$L(\boldsymbol{\phi}, \mathcal{P}) = \sum_{i=1}^{m} \|\boldsymbol{\phi}(t_i) - \boldsymbol{\phi}(t_{i-1})\|_2, \tag{5.3}$$

which can be considered as an approximation to the length of the parametrized curve $\boldsymbol{\phi}$. If $\mathcal{P}'$ is a refinement of the partition $\mathcal{P}$, obtained by adding a point of subdivision t' between t_{i-1} and t_i, the length $\|\boldsymbol{\phi}(t_i) - \boldsymbol{\phi}(t_{i-1})\|_2$ of the linear segment of the piecewise linear approximation $\boldsymbol{\phi}_{\mathcal{P}}$ between the points $\boldsymbol{\phi}(t_i)$ and

$\phi(t_{i-1})$ is replaced by the larger sum $\|\phi(t_i) - \phi(t')\|_2 + \|\phi(t') - \phi(t_{i-1})\|_2$ in the piecewise linear approximation $\phi_{\mathcal{P}'}$, so that $L(\phi, \mathcal{P}') \geq L(\phi, \mathcal{P})$; iterating this observation shows that the same inequality holds for any refinement $\mathcal{P}'$ of the partition $\mathcal{P}$. The **arc length** of the parametrized curve ϕ is defined by

dflg

$$L(\phi) = \sup_{\mathcal{P}} L(\phi, \mathcal{P}), \tag{5.4}$$

where the supremum is extended over all partitions $\mathcal{P}$ of the parameter interval $[0, 1]$. Of course it may be the case that this supremum is infinite; those parametrized curves for which the supremum is finite are called **rectifiable** curves, and for such curves the arc length has a uniquely determined finite value. There are parametrized curves that are not rectifiable, such as the parametrized curve in $\mathbb{R}^2$ given by

dflgnotrcx

$$\phi(t) = \begin{cases} \{t, t\sin\frac{\pi}{t}\} & \text{for} \quad 0 < t \leq 1, \\ \\ \{0, 0\} & \text{for} \quad t = 0. \end{cases} \tag{5.5}$$

If $\mathcal{P}_n$ is the partition given by $0 < \frac{1}{n} < \frac{1}{n-1} < \cdots < \frac{1}{2} < 1$ then since

$$\left\|\phi\left(\tfrac{1}{i}\right) - \phi\left(\tfrac{1}{i+1}\right)\right\|_2 \geq \left|\tfrac{1}{i}\sin\tfrac{\pi i}{2} - \tfrac{1}{i+1}\sin\tfrac{\pi(i+1)}{2}\right| = \begin{cases} \frac{1}{i+1} & \text{if } i \text{ is even} \\ \\ \frac{1}{i} & \text{if } i \text{ is odd,} \end{cases}$$

it is apparent that $\lim_{n\to\infty} L(\phi, \mathcal{P}_n) = \infty$, hence that this parametrized curve is not rectifiable.

dflglmrc

Lemma 5.1 *If $\phi : [a, b] \longrightarrow U$ and $\psi : [c, d] \longrightarrow U$ are two equivalent parametrized curves and if ϕ is rectifiable then ψ also is rectifiable and $L(\phi) = L(\psi)$.*

Proof: Since the parametrized curves ϕ and ψ are equivalent there is a homeomorphism $h : [a, b] \longrightarrow [c, d]$ for which $\phi = \psi \circ h$. Assume that ϕ is monotonically increasing; the proof in the case that ϕ is monotonically decreasing follows the same pattern. Any partition

$$\mathcal{P} : a = t_0 < t_1 < \cdots < t_m = b$$

of the interval $[a, b]$ determines a partition

$$h(\mathcal{P}) : c = h(a) = h(t_0) < h(t_1) < \cdots < h(t_m) = h(b) = d$$

of the interval $[c, d]$, and conversely any partition of the interval $[c, d]$ arises in this way from a partition of the interval $[a, b]$. Since $\psi\big(h(t_i)\big) = \phi(t_i)$ it is clear from the definition (5.3) [dlgapp] that $L\big(\psi, h(\mathcal{P})\big) = L(\phi, \mathcal{P})$, hence that $L(\psi) = \sup_{\mathcal{P}} L\big(\psi, h(\mathcal{P})\big) = \sup_{\mathcal{P}} L\big(\phi, \mathcal{P}\big) = L(\phi)$ and consequently that ψ also is rectifiable. That suffices for the proof.

dflglmrc1 **Lemma 5.2** *If* $\boldsymbol{\phi} : [0,1] \longrightarrow \mathbb{R}^n$ *is a rectifiable parametrized curve and* $L(\boldsymbol{\phi}, [a,b])$ *is the arc length of the restriction* $\boldsymbol{\phi} : [a,b] \longrightarrow \mathbb{R}^n$ *of this curve to a subinterval* $[a,b] \subset [0,1]$ *then*
(i) $L(\boldsymbol{\phi}, [a,b]) \ge \|\boldsymbol{\phi}(b) - \boldsymbol{\phi}(a)\|_2 \quad$ *if* $\quad 0 \le a \le b \le 1$*; and*
(ii) $L(\boldsymbol{\phi}, [a,b]) = L(\boldsymbol{\phi}, [a,c]) + L(\boldsymbol{\phi}, [c,b]) \quad$ *if* $\quad 0 \le a \le c \le b \le 1.$

Proof: (i) If $0 \le a \le b \le 1$ then $a \le b$ is a particular partition $\mathcal{P}$ of the interval $[a,b]$ so it follows from (5.3) and (5.4) that $\|\boldsymbol{\phi}(b) - \boldsymbol{\phi}(a)\|_2 = L(\boldsymbol{\phi}, \mathcal{P}) \le L(\boldsymbol{\phi}, [a,b])$.
(ii) For any partition $\mathcal{P}$ of the interval $[a,b]$ the length $L(\boldsymbol{\phi}, \mathcal{P})$ is increased by adjoining the point c to the other points of subdivision; so when examining the supremum of the lengths $L(\boldsymbol{\phi}, \mathcal{P})$ for partitions $\mathcal{P}$ of the interval $[0,1]$ it always can be assumed that c is one of the points of subdivision of the partition $\mathcal{P}$, and consequently that the partition $\mathcal{P}$ consists of a partition $\mathcal{P}_1$ of the interval $[a,c]$ and a partition $\mathcal{P}_2$ of the interval $[c,b]$. Therefore $L(\boldsymbol{\phi}, \mathcal{P}) = L(\boldsymbol{\phi}, \mathcal{P}_1) + L(\boldsymbol{\phi}, \mathcal{P}_2)$ for all partitions, so $L(\boldsymbol{\phi}, [a,b]) = L(\boldsymbol{\phi}, [a,c]) + L(\boldsymbol{\phi}, [c,b])$, which suffices for the proof.

dflglmrc1aa **Lemma 5.3** *If* $\boldsymbol{\phi} : [0,1] \longrightarrow \mathbb{R}^n$ *is a rectifiable parametrized curve and* $s(t) = L(\boldsymbol{\phi}, [0,t])$ *for any* $t \in [0,1]$ *then*
(i) $s(t)$ *is a monotonically increasing continuous function for which* $s(0) = 0$ *and* $s(1) = l = L(\boldsymbol{\phi})$*, and*
(ii) $s(t_1) = s(t_2)$ *for some parameter values* $0 \le t_1 < t_2 \le 1$ *if and only if the mapping* $\boldsymbol{\phi}$ *is constant in the interval* $[t_1, t_2]$*.*

Proof: (i) It follows immediately from Lemma 5.2 (ii) that if $0 \le t' \le t'' \le 1$ then $s(t'') = s(t') + L(\boldsymbol{\phi}, [t', t'']) \ge s(t')$, so the function $s(t)$ is monotonically increasing. For any $\epsilon > 0$ there is a $\delta > 0$ such that

dflglmrclaaq1a
$$\|\boldsymbol{\phi}(t') - \boldsymbol{\phi}(t'')\|_2 < \epsilon \quad \text{whenever } |t' - t''| < \delta \tag{5.6}$$

since the mapping $\boldsymbol{\phi}$ is uniformly continuous on the compact interval $[0,1]]$; moreover there is a partition $\mathcal{P}$ of the interval $[0,1]$ such that

dflglmrclaaq1
$$0 \le L(\boldsymbol{\phi}) - L(\boldsymbol{\phi}, \mathcal{P}) < \epsilon \tag{5.7}$$

since the parametrized curve $\boldsymbol{\phi}$ is rectifiable, and of course the same inequality holds for any refinement of the partition $\mathcal{P}$. Hence if $t_0 \in [0,1]$ then after a refinement of the partition $\mathcal{P}$ if necessary it can be assumed that $t_0 = t_k$ is one of the points of subdivision of the partition $\mathcal{P}$ and that $|t_k - t_{k-1}| < \delta$ and $|t_{k+1} - t_k| < \delta$. For any point $t \in [t_{k-1}, t_k]$ adding t to the points of subdivision of $\mathcal{P}$ yields a refinement $\mathcal{P}'$ of $\mathcal{P}$ such that $t = t'_{l-1}$ and $t_0 = t'_l$ for some index l; and it follows from (5.7), which holds also for the refinement $\mathcal{P}'$, and from (5.6) and Lemma 5.2 (i) and (ii) that

$$\begin{aligned}
\epsilon > L(\boldsymbol{\phi}) - L(\boldsymbol{\phi}, \mathcal{P}') &= \sum_i \Big(L(\boldsymbol{\phi}, [t'_{i-1}, t'_i]) - \|\boldsymbol{\phi}(t'_i) - \boldsymbol{\phi}(t'_{i-1})\|_2 \Big) \\
&\ge L(\boldsymbol{\phi}, [t, t_0]) - \|\boldsymbol{\phi}(t_0) - \boldsymbol{\phi}(t)\|_2 \ge s(t_0) - s(t) - \epsilon
\end{aligned}$$

and consequently that $0 \le s(t_0) - s(t) \le 2\epsilon$. The corresponding argument shows that $0 \le s(t) - s(t_0) \le 2\epsilon$ for points $t \in [t_k, t_{k+1}]$, so the function $s(t)$ is continuous.
(ii) If $\boldsymbol{\phi}(t) = \boldsymbol{\phi}(t_1)$ for all parameter values $t \in [t_1, t_2]$ it is clear from the definition that $L(\boldsymbol{\phi}, \mathcal{P}) = 0$ for any partition $\mathcal{P}$ of the interval $[t_1, t_2]$ and consequently that $0 = L(\boldsymbol{\phi}, [t_1, t_2]) = L(\boldsymbol{\phi}, [0, t_2]) - L(\boldsymbol{\phi}, [0, t_1]) = s(t_2) - s(t_1)$. On the other hand if $s(t_1) = s(t_2)$ then since $s(t)$ is monotonically increasing $s(t_1) = s(t)$ whenever $t_1 \le t \le t_2$, hence Lemma 5.2 (i) shows that that $0 = s(t) - s(t_1) = L(\boldsymbol{\phi}, [t_1, t]) \ge \|\boldsymbol{\phi}(t) - \boldsymbol{\phi}(t_1)\|_2$ and consequently that $\boldsymbol{\phi}(t) = \boldsymbol{\phi}(t_1)$. That suffices for the proof.

If $\boldsymbol{\phi} : [0, 1] \longrightarrow \mathbb{R}^n$ is a parametrized curve for which $s(t) = L(\boldsymbol{\phi}, [0, t])$ is strictly increasing then the mapping $s : [0, 1] \longrightarrow [0, l]$ is a homeomorphism, since it is a continuous one-to-one mapping between two compact sets; and if $t : [0, l] \longrightarrow [0, 1]$ is the inverse homeomorphism then the composition $\boldsymbol{\phi}_0(s) = \boldsymbol{\phi}(t(s))$ is a parametrized curve that is equivalent to the curve $\boldsymbol{\phi}$. From Lemma 5.1 it follows that

$$L(\boldsymbol{\phi}_0, [0, s]) = L\big(\boldsymbol{\phi}, [0, t(s)]\big) = s, \tag{5.8}$$

so the parametrized curve $\boldsymbol{\phi}_0$ is parametrized by arc length. On the other hand if the function $s(t)$ fails to be strictly increasing then by Lemma 5.3 (ii) that is because the mapping $\boldsymbol{\phi}$ is constant on some subintervals of $[0, 1]$; and the mapping $\boldsymbol{\phi}$ can be modified, essentially by shrinking such subintervals to points, so that $s(t)$ becomes strictly increasing. In more detail, $s : [0, 1] \longrightarrow [0, l]$ is a surjective mapping, since the image of the connected set $[0, 1]$ under the continuous mapping s is also connected and contains the points $s(0) = 0$ and $s(1) = l$. Therefore for any point $s_0 \in [0, l]$ there is some point $t_0 \in [0, l]$ for which $s(t_0) = s_0$; and if there is another point t_0' for which $s(t_0') = s_0$ it follows from Lemma 5.3 (ii) that $\boldsymbol{\phi}(t_0) = \boldsymbol{\phi}(t_0')$, so setting $\boldsymbol{\psi}(s_0) = \boldsymbol{\phi}\big(s(t_0)\big)$ yields a well-defined mapping $\boldsymbol{\psi} : [0, l] \longrightarrow \mathbb{R}^n$, independent of the choice of the particular point t_0, such that $\boldsymbol{\psi}\big(s(t)\big) = \boldsymbol{\phi}(t)$. This mapping $\boldsymbol{\psi}$ is a continuous mapping, since if $0 \le s_1 \le s_2 \le l$ and if $t_1, t_2 \in [0, 1]$ are any parameter values s for which $s(t_1) = s_1$ and $s(t_2) = s_2$ then by Lemma 5.2 (i)

$$\begin{aligned} \|\boldsymbol{\psi}(s_2) - \boldsymbol{\psi}(s_1)\|_2 &= \|\boldsymbol{\phi}(t_2) - \boldsymbol{\phi}(t_1)\|_2 \\ &\le L(\boldsymbol{\phi}, [t_1, t_2]) = s(t_2) - s(t_1) = s_2 - s_1. \end{aligned}$$

Further $L(\boldsymbol{\psi}, [0, s])$ is a strictly increasing function of s. Indeed if $L(\boldsymbol{\psi}, [0, s_1]) = L(\boldsymbol{\psi}, [0, s_2])$ where $s_1 < s_2$ then by Lemma 5.3 (ii) the mapping $\boldsymbol{\psi}$ must be constant in the interval $[s_1, s_2]$ so in particular $\boldsymbol{\psi}(s_1) = \boldsymbol{\psi}(s_2)$; and if $s_1 = s(t_1)$ and $s_2 = s(t_2)$ then $\boldsymbol{\phi}(t_1) = \boldsymbol{\psi}\big(s(t_1)\big) = \boldsymbol{\psi}\big(s(t_2)\big) = \boldsymbol{\phi}(t_2)$ and consequently $s_1 = s(t_1) = s(t_2) = s_2$ by Lemma 5.3 (ii) again, a contradiction. Consequently $\boldsymbol{\psi}$ also is equivalent to a parametrized curve $\boldsymbol{\phi}_0$ that is parametrized by arc length. Thus to any rectifiable parametrized curve $\boldsymbol{\phi}$ there is associated by the appropriate one of these two constructions a unique rectifiable parametrized curve $\boldsymbol{\phi}_0$ that is parametrized by arc length, called the **parametrization of**

the curve $\boldsymbol{\phi}$ by arc length; and it is evident from the construction that equivalent rectifiable parametrized curves have the same parametrization by arc length. It is then possible to associate to any continuous function f in an open neighborhood of the image $\boldsymbol{\phi}([0,1])$ of a rectifiable curve $\boldsymbol{\phi}$ a unique integral defined by

dfrectintdfa
$$\int_{\boldsymbol{\phi}} f = \int_{[0,l]} f\big(\boldsymbol{\phi}_0(s)\big) \tag{5.9}$$

where $\boldsymbol{\phi}_0$ is the parametrization of the curve $\boldsymbol{\phi}$ by arc length; this integral often is written just

dfrectintdfaa
$$\int_{\boldsymbol{\phi}} f\,ds, \tag{5.10}$$

with the convention that ds denotes the integral with respect to arc length.

dfrect **Theorem 5.1** *A $\mathcal{C}^1$ parametrized curve $\boldsymbol{\phi} : [0,1] \longrightarrow U$ in an open subset $U \subset \mathbb{R}^n$ is rectifiable, and its arc length is*

dfrectq1
$$L(\boldsymbol{\phi}) = \int_{[0,1]} \|\boldsymbol{\phi}'\|_2. \tag{5.11}$$

Proof: For any partition $\mathcal{P}$ of the interval $[0,1]$ it follows from the fundamental theorem of calculus in one variable applied to each coordinate function of the mapping $\boldsymbol{\phi}$ and the inequality (rintp3qq2 4.25) that

drectq2a
$$\begin{aligned} L(\boldsymbol{\phi},\mathcal{P}) &= \sum_{i=1}^{m} \|\boldsymbol{\phi}(t_i) - \boldsymbol{\phi}(t_{i-1})\|_2 = \sum_{i=1}^{m} \Big\| \int_{[t_{i-1},t_i]} \boldsymbol{\phi}' \Big\|_2 \\ &\le \sum_{i=1}^{m} \int_{[t_{i-1},t_i]} \|\boldsymbol{\phi}'\|_2 = \int_{[0,1]} \|\boldsymbol{\phi}'\|_2; \end{aligned} \tag{5.12}$$

and since $\|\boldsymbol{\phi}'\|_2$ is continuous on the full interval $[0,1]$ this integral is finite, so the curve γ is rectifiable and

dflgq01
$$L(\boldsymbol{\phi}) \le \int_{[0,1]} \|\boldsymbol{\phi}'\|_2. \tag{5.13}$$

For any point $t \in [0,1)$ and any $h > 0$ sufficiently small that $t+h \le 1$ it follows from (dflgq01 5.13) that

dflgq1
$$L(\boldsymbol{\phi},[t,t+h]) \le \int_{[t,t+h]} \|\boldsymbol{\phi}'\|_2. \tag{5.14}$$

On the other hand for any unit vector $\mathbf{u} \in \mathbb{R}^n$ it follows from Lemma dflglmrc1 5.2 (1), the Cauchy-Schwarz inequality and the mean value theorem applied to the real-valued function $\mathbf{u}\cdot\boldsymbol{\phi}(t)$ that

dflgq2
$$L(\boldsymbol{\phi},[t,t+h]) \ge \|\boldsymbol{\phi}(t+h)-\boldsymbol{\phi}(t)\|_2 \ge |\mathbf{u}\cdot\big(\boldsymbol{\phi}(t+h)-\boldsymbol{\phi}(t)\big)| = |\mathbf{u}\cdot\boldsymbol{\phi}'(\tau)|\ h \tag{5.15}$$

for some point τ in the interval $t < \tau < t+h$. Since

$$L(\boldsymbol{\phi},[t,t+h]) = L(\boldsymbol{\phi},[0,t+h]) - L(\boldsymbol{\phi},[0,t]) = s(t+h) - s(t)$$

by Lemma 5.2 (ii) it follows from the inequalities (5.14) and (5.15) that

dflgqa
$$|\mathbf{u}\cdot\boldsymbol{\phi}'(\tau)| \le \frac{1}{h}\big(s(t+h)-s(t)\big) \le \frac{1}{h}\int_{[t,t+h]}\|\boldsymbol{\phi}'\|_2; \tag{5.16}$$

by taking the limit as h tends to 0, and consequently as τ tends to t, it follows from the fundamental theorem of calculus, Theorem 4.13, that (5.16) reduces to the inequality

dflgqb
$$|\mathbf{u}\cdot\boldsymbol{\phi}'(t)| \le \lim_{h\to 0}\frac{1}{h}\big(s(t+h)-s(t)\big) \le \|\boldsymbol{\phi}'(t)\|_2. \tag{5.17}$$

In particular since $\mathbf{u}\cdot\boldsymbol{\phi}'(t) = \|\boldsymbol{\phi}'(t)\|_2$ when $\mathbf{u}$ is the unit vector in the direction of the vector $\boldsymbol{\phi}'(t)$, the inequality (5.17) shows that $\lim_{h\to 0}\frac{1}{h}\big(s(t+h)-s(t)\big) = \|\boldsymbol{\phi}'(t)\|_2$. The corresponding argument for $h < 0$ yields the same result, hence the function $s(t)$ is differentiable and

dflgq2xx
$$\frac{d}{dt}s(t) = \|\boldsymbol{\phi}'(t)\|_2. \tag{5.18}$$

It finally follows from the fundamental theorem of calculus again that

$$L(\boldsymbol{\phi}) = s(1) = \int_{[0,1]}\frac{d}{dt}s(t) = \int_{[0,1]}\|\boldsymbol{\phi}'\|_2,$$

and that concludes the proof.

For any $\mathcal{C}^1$ parametrized curve $\boldsymbol{\phi}$ it follows from (5.18) that its arc length $s(t) = L(\boldsymbol{\phi},[0,t])$ is a $\mathcal{C}^1$ mapping $s : [0,1] \longrightarrow [0,l]$ where $s(1) = l = L(\boldsymbol{\phi})$ is the arc length of $\boldsymbol{\phi}$; and if it is also assumed that $\boldsymbol{\phi}'(t) \ne 0$ at all points $t \in (0,1)$ then the function $s(t)$ is strictly increasing, and by the inverse mapping theorem it is a $\mathcal{C}^1$ homeomorphism so the curve $\boldsymbol{\phi}$ is $\mathcal{C}^1$ equivalent to a curve that is parametrized by arc length.

dfrectcr1
Corollary 5.1 *A $\mathcal{C}^1$ mapping $\boldsymbol{\phi} : [0,l] \longrightarrow \gamma \subset \mathbb{R}^n$ is the parametrization of a curve by arc length if and only if $\|\boldsymbol{\phi}'(s)\|_2 = 1$ for all points $s \in [0,l]$.*

Proof: If $\boldsymbol{\phi} : [0,l] \longrightarrow \mathbb{R}^n$ is a $\mathcal{C}^1$ parametrization by arc length then $s(t) = L(\boldsymbol{\phi},[0,t]) = t$ and it follows from (5.18) that $\|\boldsymbol{\phi}'(t)\|_2 = s'(t) = 1$. Conversely if $\boldsymbol{\phi} : [0,l] \longrightarrow \mathbb{R}^n$ is a parametrized curve for which $\|\boldsymbol{\phi}'(s)\|_2 = 1$ then it follows from (5.11) that $L(\boldsymbol{\phi},[0,t]) = \int_{[0,t]}\|\boldsymbol{\phi}'\|_2 = \int_{[0,t]} 1 = t$, so the parameter t is arc length, which suffices for the proof.

If $\boldsymbol{\phi} : [0,l] \longrightarrow \gamma \subset \mathbb{R}^n$ is a $\mathcal{C}^1$ curve parametrized by arc length then $\boldsymbol{\phi}'(t) \ne 0$ at all points $t \in [0,l]$ so it follows from Corollary 3.8 to the Rank Theorem that the image of an open neighborhood of any parameter value $s_0 \in [0,l]$ is a

$\mathcal{C}^1$ submanifold of an open neighborhood of the point $\boldsymbol{\phi}(s_0) \in U$. The vector $\boldsymbol{\phi}'(s_0)$ is a tangent vector to that submanifold, as in the discussion on page 62, and has length $\|\boldsymbol{\phi}'(s_0)\|_2 = 1$, by the preceding Corollary 5.1; thus it is the **unit tangent vector** to the local piece of the curve γ at the point $\boldsymbol{\phi}(s_0)$, and is denoted by $\boldsymbol{\tau}(s_0) = \boldsymbol{\phi}'(s_0)$. The curve γ may intersect itself at some points, at which there are unit tangent vectors to the various local components of the curve γ. If $\mathbf{f}$ is a continuous vector field in an open neighborhood of γ, the dot product $\mathbf{f}\big(\boldsymbol{\phi}(s)\big) \cdot \boldsymbol{\tau}(s) = \mathbf{f}\big(\boldsymbol{\phi}(s)\big) \cdot \boldsymbol{\phi}'(s)$ is a well defined continuous function of the variable $s \in [0, l]$ so it has a well defined integral that also is denoted by

lintq1 (5.19)
$$\int_{\boldsymbol{\phi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds = \int_{[0,l]} (\mathbf{f} \circ \boldsymbol{\phi}) \cdot \boldsymbol{\phi}',$$

where ds indicates integration with respect to arc length. This is called the **line integral** of the vector field $\mathbf{f}$ along the curve γ. Although this integral was defined only in terms of the parametrization of the curve by arc length, it actually has the same value for any equivalent parametrization of the curve, so in that sense it is intrinsically defined on equivalence classes of $\mathcal{C}^1$ parametrized curves.

lintind **Theorem 5.2** *If $\boldsymbol{\psi} : [a, b] \longrightarrow \mathbb{R}^n$ is a $\mathcal{C}^1$ parametrized curve and $\boldsymbol{\psi}'(t) \neq 0$ for $a \leq t \leq b$ then*

lintindq1 (5.20)
$$\int_{\boldsymbol{\psi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds = \int_{[a,b]} (\mathbf{f} \circ \boldsymbol{\psi}) \cdot \boldsymbol{\psi}'$$

for any continuous vector field $\mathbf{f}$ in an open neighborhood of $\boldsymbol{\psi}([a, b])$.

Proof: The parametrized curve $\boldsymbol{\psi}$ is equivalent to a curve $\boldsymbol{\phi}$ parametrized by arc length, since $\boldsymbol{\psi}'(t) \neq 0$ at all points $t \in [a, b]$; and the line integral of the vector field $\mathbf{f}$ along the parametrized curve $\boldsymbol{\psi}$ is defined in terms of the parametrized curve $\boldsymbol{\phi}$. If $h : [a, b] \longrightarrow [0, l]$ is a $\mathcal{C}^1$ homeomorphism such that $\boldsymbol{\psi}(t) = \boldsymbol{\phi}\big(h(t)\big)$ then by the chain rule $\boldsymbol{\psi}'(t) = \boldsymbol{\phi}'\big(h(t)\big)h'(t)$, so by the formula for the change of variables in integration for functions of a single variable it follows as in (5.19) that

$$\begin{aligned}\int_{\boldsymbol{\psi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds &= \int_{[0,l]} \mathbf{f}\big(\boldsymbol{\phi}(s)\big) \cdot \boldsymbol{\phi}'(s) = \int_{[a,b]} \mathbf{f}\big(\boldsymbol{\phi}(h(t))\big) \cdot \boldsymbol{\phi}'\big(h(t)\big)h'(t) \\ &= \int_{[a,b]} \mathbf{f}\big(\boldsymbol{\psi}(t)\big) \cdot \boldsymbol{\psi}'(t),\end{aligned}$$

which suffices for the proof.

For example, consider the integral of the vector field $\mathbf{f}(\mathbf{x}) = \{x_1^2, x_1x_2\}$ in terms of the variables x_1, x_2 in $\mathbb{R}^2$ over the parabola described parametrically by the mapping $\boldsymbol{\phi} : [0, 1] \longrightarrow \mathbb{R}^2$ for which $\boldsymbol{\phi}(t) = \{t, t^2\}$. Since $\boldsymbol{\phi}'(t) = \{1, 2t\}$

then $\mathbf{f}\big(\boldsymbol{\phi}(t)\big) = \{t^2,\, t^3\}$ so $\mathbf{f}\big(\boldsymbol{\phi}(t)\big)\cdot\boldsymbol{\phi}'(t) = t^2 + 2t^4$ and consequently

$$\begin{aligned}\int_\gamma \mathbf{f}\cdot d\boldsymbol{\tau}\, ds &= \int_{[0,1]} \mathbf{f}\big(\boldsymbol{\phi}(t)\big)\cdot\boldsymbol{\phi}'(t)\\ &= \int_{[0,1]} (t^2+2t^4) = \left(\frac{1}{3}t^3 + \frac{2}{5}t^5\right)\Big|_0^1 = \frac{11}{15}.\end{aligned}$$

A more interesting example is the line integral of a vector field that is the gradient $\mathbf{f}(\mathbf{x}) = \nabla h(\mathbf{x})$ of a function h in an open set $U \subset \mathbb{R}^n$, a vector field called a **conservative** vector field with the **potential** h.

lininepot **Theorem 5.3** *If $\mathbf{f}(\mathbf{x}) = \nabla h(\mathbf{x})$ is a conservative vector field in a connected open subset $U \subset \mathbb{R}^n$ with a $\mathcal{C}^1$ potential h, and if $\boldsymbol{\psi} : [0,1] \longrightarrow U$ is a $\mathcal{C}^1$ curve in U from a point $\boldsymbol{\psi}(0) = \mathbf{a} \in U$ to a point $\boldsymbol{\psi}(1) = \mathbf{b} \in U$ then*

lininepotq1 (5.21)
$$\int_{\boldsymbol{\psi}} \mathbf{f}\cdot\boldsymbol{\tau}\, ds = h(\mathbf{b}) - h(\mathbf{a}).$$

Proof: In view of the preceding theorem it is sufficient to demonstrate this result just for a curve parametrized by arc length; so let $\boldsymbol{\phi} : [0,l] \longrightarrow U$ be the parametrization of a curve by arc length, where $\boldsymbol{\phi}(s) = \{\phi_j(s)\}$ and $\boldsymbol{\phi}(0) = \mathbf{a}$, $\boldsymbol{\phi}(l) = \mathbf{b}$. By the chain rule

$$\begin{aligned}\frac{d}{ds}h\big(\boldsymbol{\phi}(s)\big) &= h'\big(\boldsymbol{\phi}(s)\big)\,\boldsymbol{\phi}'(s) = \sum_{j=1}^n \partial_j h\big(\boldsymbol{\phi}(s)\big)\phi_j'(s) = \sum_{j=1}^n f_j\big(\boldsymbol{\phi}(s)\big)\phi_j'(s)\\ &= \mathbf{f}\big(\boldsymbol{\phi}(s)\big)\cdot\boldsymbol{\phi}'(s),\end{aligned}$$

and consequently by the fundamental theorem of calculus

$$\begin{aligned}\int_{\boldsymbol{\psi}} \mathbf{f}\cdot\boldsymbol{\tau}\, ds &= \int_{[0,l]} \mathbf{f}\big(\boldsymbol{\phi}(s)\big)\cdot\boldsymbol{\phi}'(s) = \int_{[0,l]} h'\big(\boldsymbol{\phi}(s)\big)\\ &= h\big(\boldsymbol{\phi}(l)\big) - h\big(\boldsymbol{\phi}(0)\big) = h(\mathbf{b}) - h(\mathbf{a}),\end{aligned}$$

which suffices for the proof.

Thus the integral of a conservative vector field along a curve really depends only on the beginning and end points of the curve, and is quite independent of the particular path taken between these two points. There are vector fields that are not conservative, and for which the line integral between two points depends on the path chosen between those two points. For example for the vector field $\mathbf{f}(\mathbf{x}) = \{x_2, -x_1\}$ in $\mathbb{R}^2$, and the parametrized curves $\boldsymbol{\psi}_1(t) = \{t, -t\}$ and $\boldsymbol{\psi}_2(t) = \{t, -t^2\}$ for $t \in [0,1]$

$$\int_{\boldsymbol{\psi}_1} f\cdot\boldsymbol{\tau}\, ds = \int_{[0,1]} \{-t,\, -t\}\cdot\{1,\, -1\}\, dt = \int_{[0,1]} 0\ dt = 0,$$

$$\int_{\boldsymbol{\psi}_2} f\cdot\boldsymbol{\tau}\, ds = \int_{[0,1]} \{-t^2,\, -t\}\cdot\{1,\, -2t\} dt = \int_{[0,1]} t^2 dt = \frac{1}{3}.$$

Actually the condition that the line integral of a vector field is independent of the path characterizes conservative vector fields.

linth **Theorem 5.4** *A $\mathcal{C}^1$ vector field $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ in an open subset $U \subset \mathbb{R}^n$ is conservative if and only if the line integrals $\int_{\boldsymbol{\psi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds$ depend only on the end points of $\mathcal{C}^1$ parametrized curves $\boldsymbol{\psi}$ in U.*

Proof: The preceding theorem demonstrated that if $\mathbf{f}$ is a conservative vector field and $\boldsymbol{\psi} : [0,1] \longrightarrow U$ is a $\mathcal{C}^1$ parametrized curve in U then the line integral $\int_{\boldsymbol{\psi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds$ depends just on the beginning and end points of $\boldsymbol{\psi}$. Conversely suppose that $\mathbf{f}$ is a $\mathcal{C}^1$ vector field in a connected open subset $U \subset \mathbb{R}^n$ such that the line integrals $\int_{\boldsymbol{\psi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds$ depend only on the beginning and end points of all $\mathcal{C}^1$ parametrized curves $\boldsymbol{\psi} : [0,1] \longrightarrow U$. For a fixed base point $\mathbf{a} \in U$ the integral $h(x) = \int_{\boldsymbol{\psi}_{\mathbf{x}}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds$ along any $\mathcal{C}^1$ parametrized curve $\boldsymbol{\psi}_{\mathbf{x}}$ from the point $\mathbf{a}$ to the point $\mathbf{x}$ is a well defined function of the point $\mathbf{x} = \{x_1, \ldots, x_n\} \in U$, independent of the particular choice of the parametrized curve $\boldsymbol{\psi}_{\mathbf{x}}$, provided of course that $\mathbf{x}$ is in the same connected component of U as the point $\mathbf{a}$. In particular for a parametrized curve $\boldsymbol{\psi}_{\mathbf{x}}$ that describes a path from $\mathbf{a}$ to the straight line λ_i through the point $\mathbf{x}$ parallel to the coordinate axis of the variable x_i and then along λ_i to the point $\mathbf{x}$, near the point $\mathbf{x}$ the parametrization has the form

$$\boldsymbol{\psi}_{\mathbf{x}}(t) = \{x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n\} \quad \text{for } b_i \le t \le x_i$$

so $\boldsymbol{\psi}'_{\mathbf{x}} = \delta_i$ is the unit vector parallel to λ_i. The function $h(\mathbf{x})$ then can be written as the sum of the value A of the integral up to a point $\boldsymbol{\psi}_{\mathbf{x}}(b_i)$ on the line λ_i and the integral

$$h(\mathbf{x}) - A = \int_{t \in [b_i, x_i]} \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau}\, ds = \int_{t \in [b_i, x_i]} f_i(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)$$

along λ_i; and it then follows from the fundamental theorem of calculus in one variable that $\partial_i h(\mathbf{x}) = f_i(\mathbf{x})$. The corresponding argument for paths ending parallel to other coordinate axes in $\mathbb{R}^n$ shows that $\nabla h(\mathbf{x}) = \mathbf{f}(\mathbf{x})$, hence that the vector field $\mathbf{f}$ is conservative, thus concluding the proof.

5.2 Differential forms

math218notes5a

The differentials dx_j of the coordinate functions x_j in $\mathbb{R}^n$ are a basis for the vector space $\mathbb{R}^n$, as in Section 2.4. Vector fields over an open subset $U \subset \mathbb{R}^n$, mappings $\mathbf{f} : U \longrightarrow \mathbb{R}^n$ with the coordinate functions $\mathbf{f}(\mathbf{x}) = \{f_j(\mathbf{x})\}$, when written as the sums

DFq1 (5.22)
$$\omega(\mathbf{x}) = \sum_{j=1}^{n} f_j(\mathbf{x})\; dx_j$$

of multiples of the basis vectors dx_j, are called **differential forms** of degree 1 in U, or for short just differential 1-forms in U. Mappings $\mathbf{f} : U \longrightarrow \mathbb{R}^{\binom{n}{r}}$ for indices $2 \le r \le n$ with the coordinate functions $\mathbf{f} = \{f_{j_1\cdots j_r}(\mathbf{x})\}$ for $1 \le j_1 < j_2 < \cdots < j_r \le n$, when written as the sums

Vdf
$$\omega(\mathbf{x}) = \sum_{j_1<\cdots<j_r} f_{j_1\cdots j_r}(\mathbf{x}) dx_{j_1} \wedge \cdots \wedge dx_{j_r} \tag{5.23}$$

of multiples of the basis vectors $dx_{j_1} \wedge \cdots \wedge dx_{j_r}$ that span the exterior algebra Λ^r of r-vectors over the vector space $\mathbb{R}^n$ as discussed in Appendix A, are called **differential forms**[1] **of degree** r in U, or for short just differential r-forms in U. These are continuous differential forms if the coefficients $f_j(\mathbf{x})$ or $f_{j_1\cdots j_r}(\mathbf{x})$ are continuous functions in U, and are $\mathcal{C}^k$ differential forms if these coefficients are $\mathcal{C}^k$ functions in U. For differential forms of degree $r > 1$ it is sometimes convenient to introduce the additional vectors $dx_{i_1} \wedge \cdots \wedge dx_{i_r}$ for any order of the indices $i_1, \ldots, i_r$, where these satisfy the skew-symmetry conditions (A.2); thus each of these additional vectors is equal to either one of the basis vectors of Λ^r or the negative of one of these basis vectors. Differential r-forms in U can be written alternatively in terms of these vectors as sums

Vdf2
$$\omega(\mathbf{x}) = \sum_{i_1,\ldots,i_r=1}^{n} g_{i_1\cdots i_r}(\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_r}. \tag{5.24}$$

The coefficients $f_{j_1\cdots j_r}(\mathbf{x})$ in (5.23) are of course uniquely determined, while the coefficients $g_{i_1\cdots i_r}(\mathbf{x})$ in(5.24) are not uniquely determined. For example a differential form of degree 2 in $\mathbb{R}^3$ can be written

Vdf3
$$\begin{aligned}\omega(\mathbf{x}) = g_{12}(\mathbf{x})dx_1 \wedge dx_2 + g_{13}(\mathbf{x})dx_1 \wedge dx_3 \\ + g_{21}(\mathbf{x})dx_2 \wedge dx_1 + g_{23}(\mathbf{x})dx_2 \wedge dx_3 \\ + g_{31}(\mathbf{x})dx_3 \wedge dx_1 + g_{32}(\mathbf{x})dx_3 \wedge dx_2,\end{aligned} \tag{5.25}$$

or since $dx_{i_1} \wedge dx_{i_2} = -dx_{i_2} \wedge dx_{i_1}$ it can be written alternatively

Vdf4
$$\begin{aligned}\omega(\mathbf{x}) = \Big(g_{12}(\mathbf{x}) - g_{21}(\mathbf{x})\Big)dx_1 \wedge dx_2 + \Big(g_{13}(\mathbf{x}) - g_{31}(\mathbf{x})\Big)dx_1 \wedge dx_3 \\ + \Big(g_{23}(\mathbf{x}) - g_{32}(\mathbf{x})\Big)dx_2 \wedge dx_3\end{aligned} \tag{5.26}$$

[1] There is often some mystery made about what a differential form really is. Differential forms in open subsets of $\mathbb{R}^n$ are merely special vector-valued functions, described explicitly in terms of basis vectors associated to the basis vectors in $\mathbb{R}^n$; the detailed definitions are chosen to simplify the numerous calculations in linear algebra arising from the change of variables formula in various contexts, and to make the results of these calculations easier to remember and to use. The properties of differential forms under changes of coordinates in $\mathbb{R}^n$ show that they are really rather intrinsically defined, which of course raises the temptation to define them intrinsically. Unfortunately the traditional and perhaps easiest intrinsic approach is first through the dual vector space, and then through a more general discussion of the relations between the exterior algebras over dual vector spaces; for a first approach in multivariable calculus that seems rather unnecessarily complicated and distracting, which is the motivation for the alternative though nonintrinsic approach adopted in these notes.

thus as the differential form

Vdf5 (5.27)
$$\omega(\mathbf{x}) = f_{12}(\mathbf{x})dx_1 \wedge dx_2 + f_{13}(\mathbf{x})dx_1 \wedge dx_3 + f_{23}(\mathbf{x})dx_2 \wedge dx_3$$

where

Vdf6 (5.28)
$$f_{12}(\mathbf{x}) = g_{12}(\mathbf{x}) - g_{21}(\mathbf{x}),\ f_{13}(\mathbf{x}) = g_{13}(\mathbf{x}) - g_{31}(\mathbf{x}),\ f_{23}(\mathbf{x}) = g_{23}(\mathbf{x}) - g_{32}(\mathbf{x}).$$

The functions $f_{j_1 j_2}$ are uniquely determined; but it is possible for instance to replace $g_{12}(\mathbf{x})$ and $g_{21}(\mathbf{x})$ by $g_{12}(\mathbf{x}) + g(\mathbf{x})$ and $g_{21}(\mathbf{x}) + g(\mathbf{x})$ for any function $g(\mathbf{x})$ and still have the same differential 2-form. Of course since $dx_1 \wedge dx_1 = dx_2 \wedge dx_2 = dx_3 \wedge dx_3 = 0$ the coefficients $g_{ii}(\mathbf{x})$ are totally irrelevant. A differential form written as in (5.23) is said to be in the **reduced** form, while a differential form written as in (5.24) is said to be in the **unreduced** form. It is often convenient to use **multi-index notation**, writing the unreduced form (5.24) as

Vmlt (5.29)
$$\omega(\mathbf{x}) = \textstyle\sum_I g_I(\mathbf{x}) d\mathbf{x}_I$$

where $I = (i_1, i_2, \ldots, i_r)$ and $g_I(\mathbf{x})$ is an abbreviation for $g_{i_1 \cdots i_r}(\mathbf{x})$ while $d\mathbf{x}_I$ is an abbreviation for $dx_{i_1} \wedge \cdots \wedge dx_{i_r}$, and writing the reduced form (5.23) as

Vmltx (5.30)
$$\omega(\mathbf{x}) = \textstyle\sum_I' f_I(\mathbf{x}) d\mathbf{x}_I$$

with the corresponding meaning of the terms; thus $\sum_I$ denotes a sum over all indices $i_1, \ldots, i_r$ in the range $1 \le i_1, \ldots, i_r \le n$ while $\sum_I'$ denotes a sum over all ordered sets of indices $1 \le i_1 < \cdots < i_r \le n$. The set of $\mathcal{C}^k$ differential forms of degree r in the set U is denoted by $\Lambda^r_{(k)}(U)$; however it is quite common to drop the subscript k when the regularity of the differential forms is either clear from the context or irrelevant. By definition the set $\Lambda^0_{(k)}(U)$ of differential forms of degree 0 in U is just the set of $\mathcal{C}^k$ functions in U, while the set $\Lambda^n_{(k)}(U)$ consists of differential forms of degree n in U, each of which can be written uniquely in the reduced form as

Vdfnf (5.31)
$$\omega(\mathbf{x}) = f_{12\cdots n}(\mathbf{x}) dx_1 \wedge \cdots \wedge dx_n;$$

this set of differential forms also can be identified with the set of $\mathcal{C}^k$ functions in U by identifying the differential form (5.31) with the function $f_{12\cdots n}(\mathbf{x})$. Of course $\Lambda^r_{(k)}(U) = 0$ whenever $r > n$, as a consequence of (A.2).

Differential forms can be combined algebraically, either by linear combinations or by exterior multiplication. The **linear combination** $a\omega(\mathbf{x}) + b\sigma(\mathbf{x})$ of two differential r-forms $\omega(\mathbf{x}) = \sum_I f_I(\mathbf{x}) d\mathbf{x}_I$ and $\sigma(\mathbf{x}) = \sum_I g_I(\mathbf{x}) d\mathbf{x}_I$ for $a, b \in \mathbb{R}$ is defined to be the differential r-form

Vsmm (5.32)
$$\big(a\omega + b\sigma\big)(\mathbf{x}) = \sum_I \big(a\, f_I(\mathbf{x}) + b\, g_I(\mathbf{x})\big)\ d\mathbf{x}_I;$$

under this operation the set $\Lambda^r_{(k)}(U)$ has the natural structure of a real vector space, although an infinite-dimensional vector space. The **exterior product** $\omega(\mathbf{x}) \wedge \sigma(\mathbf{x})$ of a differential r-form $\omega(\mathbf{x}) = \sum_I f_I(\mathbf{x}) d\mathbf{x}_I \in \Lambda^r_{(k)}(U)$ and a differential s-form $\sigma(\mathbf{x}) = \sum_I g_I(\mathbf{x}) d\mathbf{x}_I \in \Lambda^s_{(k)}(U)$ is defined to be the differential $r+s$-form

prd
$$(\omega \wedge \sigma)(\mathbf{x}) = \sum_{I,J} f_I(\mathbf{x}) g_J(\mathbf{x}) d\mathbf{x}_I \wedge d\mathbf{x}_J \in \Lambda^{r+s}_{(k)}(U). \tag{5.33}$$

This operation is clearly associative, in the sense that

prdasc
$$\omega_1 \wedge \big(\omega_2 \wedge \omega_3\big) = \big(\omega_1 \wedge \omega_2\big) \wedge \omega_3 \tag{5.34}$$

for any differential forms $\omega_i \in \Lambda^{r_i}_{(k)}$. However it is not commutative but rather satisfies the condition that

ncom
$$(\omega \wedge \sigma)(\mathbf{x}) = (-1)^{rs} (\sigma \wedge \omega)(\mathbf{x}) \quad \text{if } \omega \in \Lambda^r_{(k)}(U) \text{ and } \sigma \in \Lambda^s_{(k)}(U); \tag{5.35}$$

for $d\mathbf{x}_I \wedge d\mathbf{x}_J = (-1)^{rs} d\mathbf{x}_J \wedge d\mathbf{x}_I$ since to rearrange the left-hand side each differential dx_j must be permuted with each of the r differentials dx_i, a change of sign $(-1)^r$ for each of the s differentials dx_j. In particular the exterior product of a 0-form, a function $f(\mathbf{x})$, and an r-form $\omega(\mathbf{x}) = \sum_J g_J(\mathbf{x}) d\mathbf{x}_J$ is the r-form $(f\omega)(\mathbf{x}) = \sum_J f(\mathbf{x}) g_J(\mathbf{x}) d\mathbf{x}_J$. Exterior products and linear operations are related by the distributive condition that

prdst
$$(a_1\omega_1 + a_2\omega_2) \wedge \sigma = a_1\omega_1 \wedge \sigma + a_2\omega_2 \wedge \sigma \tag{5.36}$$

for differential forms $\omega_i \in \Lambda^r_{(k)}(U)$ and $\sigma \in \Lambda^s_{(k)}(U)$, as is evident from the definitions.

These algebraic operations on differential forms are quite familiar in the special case of three dimensional vector spaces. A vector field in an open set $U \subset \mathbb{R}^3$ is naturally identified with a differential 1-form, as already discussed; but since $\dim \Lambda^2(\mathbb{R}^3) = 3$ a vector field also can be identified with a differential 2-form. Thus to a vector field $\mathbf{f}(\mathbf{x}) = \{f_j(\mathbf{x})\}$ in $U \subset \mathbb{R}^3$ there are naturally associated the two differential forms

ldfm
$$\begin{aligned} &\omega_{\mathbf{f}}(\mathbf{x}) = f_1(\mathbf{x}) dx_1 + f_2(\mathbf{x}) dx_2 + f_3(\mathbf{x}) dx_3 \in \Lambda^1(U) \\ &\text{and} \\ &\Omega_{\mathbf{f}}(\mathbf{x}) = f_1(\mathbf{x}) dx_2 \wedge dx_3 + f_2(\mathbf{x}) dx_3 \wedge dx_1 + f_3(\mathbf{x}) dx_1 \wedge dx_2 \in \Lambda^2(U), \end{aligned} \tag{5.37}$$

where the notation for the 2-form $\Omega_{\mathbf{f}}(\mathbf{x})$, involving cyclic permutations of the indices, is chosen to fit into the classical form for the algebraic operations. The exterior product of the 1-form associated to a vector field $\mathbf{f} : U \longrightarrow \mathbb{R}^3$ and the 2-form associated to a vector field $\mathbf{g} : U \longrightarrow \mathbb{R}^3$ is easily seen to be the 3-form

Vdot
$$\omega_{\mathbf{f}}(\mathbf{x}) \wedge \Omega_{\mathbf{g}}(\mathbf{x}) = \Big(f_1(\mathbf{x}) g_1(\mathbf{x}) + f_2(\mathbf{x}) g_2(\mathbf{x}) + f_3(\mathbf{x}) g_3(\mathbf{x}) \Big) dx_1 \wedge dx_2 \wedge dx_3; \tag{5.38}$$

this is the 3-form associated to the dot product or inner product $\mathbf{f}(\mathbf{x})\cdot\mathbf{g}(\mathbf{x})$ of the two vector fields $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$, so that

Vdotx (5.39)
$$\omega_{\mathbf{f}}(\mathbf{x})\wedge\Omega_{\mathbf{g}}(\mathbf{x}) = \mathbf{f}(x)\cdot\mathbf{g}(x)\ dx_1\wedge dx_2\wedge dx_3,$$

and in this case the exterior product is commutative. On the other hand a straightforward calculation shows that the exterior product of the 1-forms associated to two vector fields $\mathbf{f}: U \longrightarrow \mathbb{R}^3$ and $\mathbf{g}: U \longrightarrow \mathbb{R}^3$ is the 2 form

crs (5.40)
$$\omega_{\mathbf{f}}(\mathbf{x})\wedge\omega_{\mathbf{g}}(\mathbf{x}) = \det\begin{pmatrix} f_2(\mathbf{x}) & f_3(\mathbf{x}) \\ g_2(\mathbf{x}) & g_3(\mathbf{x})\end{pmatrix} dx_2\wedge dx_3 + \det\begin{pmatrix} f_3(\mathbf{x}) & f_1(\mathbf{x}) \\ g_3(\mathbf{x}) & g_1(\mathbf{x})\end{pmatrix} dx_3\wedge dx_1 + \det\begin{pmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) \\ g_1(\mathbf{x}) & g_2(\mathbf{x})\end{pmatrix} dx_1\wedge dx_2,$$

which is the 2 form associated to the vector field $\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})$, traditionally called **cross-product** of the two vector fields $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$, defined by

crsvf (5.41)
$$\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x}) = \left\{\det\begin{pmatrix} f_2(\mathbf{x}) & f_3(\mathbf{x}) \\ g_2(\mathbf{x}) & g_3(\mathbf{x})\end{pmatrix},\ \det\begin{pmatrix} f_3(\mathbf{x}) & f_1(\mathbf{x}) \\ g_3(\mathbf{x}) & g_1(\mathbf{x})\end{pmatrix},\ \det\begin{pmatrix} f_1(\mathbf{x}) & f_2(\mathbf{x}) \\ g_1(\mathbf{x}) & g_2(\mathbf{x})\end{pmatrix}\right\};$$

when this vector is expressed as a linear combination of some basis vectors $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$, $\boldsymbol{\epsilon}_3$ it can be written symbolically as the determinant

crsvfa (5.42)
$$\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x}) = \det\begin{pmatrix} \boldsymbol{\epsilon}_1 & \boldsymbol{\epsilon}_2 & \boldsymbol{\epsilon}_3 \\ f_1(\mathbf{x}) & f_2(\mathbf{x}) & f_3(\mathbf{x}) \\ g_1(\mathbf{x}) & g_2(\mathbf{x}) & g_3(\mathbf{x})\end{pmatrix},$$

which is the easiest form in which to remember the defining equation (5.41). In these terms then

drsvf1 (5.43)
$$\omega_{\mathbf{f}}(\mathbf{x})\wedge\omega_{\mathbf{g}}(\mathbf{x}) = \Omega_{\mathbf{f}\times\mathbf{g}}(\mathbf{x});$$

the exterior product of 1-forms is skew-symmetric, as is the cross-product, so

crsprdac (5.44)
$$\omega_{\mathbf{g}}(\mathbf{x})\wedge\omega_{\mathbf{f}}(\mathbf{x}) = -\omega_{\mathbf{f}}(\mathbf{x})\wedge\omega_{\mathbf{g}}(\mathbf{x}) \quad\text{and}\quad \mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x}) = -\mathbf{g}(\mathbf{x})\times\mathbf{f}(\mathbf{x}).$$

It is clear from (5.41) that the cross-product $\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})$ is trivial at a point $\mathbf{x}$ precisely when the two vectors $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ are linearly dependent. It is also clear from (5.41) that for any other vector field $\mathbf{h}(\mathbf{x})$ in U

crsq1 (5.45)
$$\mathbf{h}(\mathbf{x})\cdot\big(\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})\big) = \det\begin{pmatrix} h_1(\mathbf{x}) & h_2(\mathbf{x}) & h_3(\mathbf{x}) \\ f_1(\mathbf{x}) & f_2(\mathbf{x}) & f_3(\mathbf{x}) \\ g_1(\mathbf{x}) & g_2(\mathbf{x}) & g_3(\mathbf{x})\end{pmatrix},$$

which provides an alternative characterization of the cross-product. An evident consequence of (5.45) is that $\mathbf{f}(\mathbf{x})\cdot\big(\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})\big) = \mathbf{g}(\mathbf{x})\cdot\big(\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})\big) = 0$; hence the cross-product $\mathbf{f}(\mathbf{x})\times\mathbf{g}(\mathbf{x})$ is a vector in $\mathbb{R}^3$ that is perpendicular to both $\mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ if these vectors are linearly independent, so it is determined uniquely up to sign. As for the sign, it follows from (crsvf 5.41) then that if $\mathbf{f}(\mathbf{x}_0) = \{1, \quad 0, \quad 0\}$ and $\mathbf{g}(\mathbf{x}_0) = \{0, \quad 1, \quad 0\}$ at a point $\mathbf{x}_0 \in U$ then $\mathbf{f}(\mathbf{x}_0) \times \mathbf{g}(\mathbf{x}_0) = \{0, \quad 0, \quad 1\}$; consequently the three vectors $\mathbf{f}(\mathbf{x}_0), \mathbf{g}(\mathbf{x}_0), \mathbf{f}(\mathbf{x}_0) \times \mathbf{g}(\mathbf{x}_0)$ have the same orientation as the three basis vectors $dx_1,\ dx_2,\ dx_3 \in \mathbb{R}^3$, the natural orientation associated of the vector space $\mathbb{R}^3$ with this choice of coordinates.

In $\mathbb{R}^n$ the vector space of differential r-forms and the vector space of differential $(n-r)$-forms have bases of the same dimension; in particular in $\mathbb{R}^3$ both the 1-forms and the 2-forms can be represented by vector fields, while the 0-forms and 3-forms can be represented by ordinary functions. This property can be realized in general by Hodge **duality**, a linear mapping

hdgq1 (5.46)
$$* : \Lambda^r(U) \longrightarrow \Lambda^{n-r}(U) \quad \text{for open sets } U \subset \mathbb{R}^n$$

that associates to a differential form $\omega = \sum_I f_I\, d\mathbf{x}_I \in \Lambda^r(U)$ the differential form

hdgq2 (5.47)
$$\begin{aligned} *\omega &= \frac{1}{r!(n-r)!} \textstyle\sum_{I,J} \epsilon(I,J) f_I\, d\mathbf{x}_J \\ &= \textstyle\sum'_{I,J} \epsilon(IJ) f_I\, d\mathbf{x}_J \in \Lambda^{n-r}(U) \end{aligned}$$

where $\epsilon(I,J)$ is the sign of the permutation I,J of the integers $1,2,\ldots,n$. Thus for example in $\mathbb{R}^n$

hdgqzz (5.48)
$$*f = f dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$$

while in particular in $\mathbb{R}^2$

hdgq3 (5.49)
$$*(f_1 dx_1 + f_2 dx_2) = \sum_{1 \le i,j \le 2} \epsilon(ij) f_i\, dx_j = f_1\, dx_2 - f_2\, dx_1$$

and in $\mathbb{R}^3$

hdgq4 (5.50)
$$\begin{aligned} *(f_1\, dx_1 + f_2\, dx_2 + f_3\, dx_3) &= \frac{1}{2} \sum_{1 \le i, j_1, j_2 \le 3} \epsilon(i j_1 j_2) f_i\, dx_{j_1} \wedge dx_{j_2} \\ &= \sum_{1 \le i \le 3} \sum_{1 \le j_1 < j_2 \le 3} \epsilon(i j_1 j_2) f_i\, dx_{j_1} \wedge dx_{j_2} \\ &= f_1\, dx_2 \wedge dx_3 - f_2 dx_1 \wedge dx_3 + f_3 dx_1 \wedge dx_2 \end{aligned}$$

or equivalently $*\omega_{\mathbf{f}} = \Omega_{\mathbf{f}}$ for the differential forms associated to any vector field $\mathbf{f}$ in an open subset $U \subset \mathbb{R}^3$. This operation has the property that

hdgq4a (5.51)
$$* * \omega = (-1)^{r(n-r)} \omega \quad \text{for any differential } r\text{-form } \omega \text{ in } \mathbb{R}^n.$$

To demonstrate this, if $\omega = \sum_I f_I\, d\mathbf{x}_I$ is an r-form in $\mathbb{R}^n$ then by definition

hdgq4ax (5.52)
$$\begin{aligned} *(*\omega) &= *\Big(\textstyle\sum'_{I,J} \epsilon(IJ) f_I\, d\mathbf{x}_J \Big) \\ &= \textstyle\sum'_{I,J,K} \epsilon(JK) \epsilon(IJ) f_I\, d\mathbf{x}_K. \end{aligned}$$

For any ordered set of r indices $1 \le k_1 < \cdots < k_r \le n$ there is a unique complementary ordered set of $n-r$ indices $1 \le j_1 < \cdots < j_{n-r} \le n$ such that (J, K) is a permutation of the integers $(1, 2, \ldots, n)$, and then there is a further unique complementary ordered set of r indices $1 \le i_1 < \cdots < i_r \le n$ such that (I, J) also is is a permutation of the integers $(1, 2, \ldots, n)$, for which clearly $I = K$; thus the only nontrivial terms in the sum (5.51) are those for which the indices J and I are uniquely determined by the indices K, indeed for which $I = K$, and since $\epsilon(IJ)\epsilon(JK) = \epsilon(KJ)\epsilon(JK) = (-1)^{r(n-r)}$ just as in the earlier examination (5.35) of exterior products it follows that (5.52) reduces to (5.51) as desired. For instance since it was already noted in (5.50) that $*\omega_{\mathbf{f}} = \Omega_{\mathbf{f}}$ for any vector field $\mathbf{f}$ in $\mathbb{R}^3$, it follows from (5.51) that $*\Omega_{\mathbf{f}} = **\omega_{\mathbf{f}} = \omega_{\mathbf{f}}$.

The most characteristic property of Hodge duality[2] though is that if $\omega = \sum_I' f_I(\mathbf{x})d\mathbf{x}_I$ and $\sigma = \sum_J' g_J(\mathbf{x})d\mathbf{x}_J$ are two r-forms in $\mathbb{R}^n$ then

hdgq5 (5.53)
$$\omega \wedge *\sigma = \sum_I' f_I(\mathbf{x})g_I(\mathbf{x})dx_1 \wedge \cdots \wedge dx_n.$$

Indeed by definition

hdgq6 (5.54)
$$\begin{aligned}\omega \wedge *\sigma &= \sum_{I,J,K}' f_I(\mathbf{x})d\mathbf{x}_I \wedge \epsilon(JK)g_J(\mathbf{x})d\mathbf{x}_K\\ &= \sum_{I,J,K}' \epsilon(JK)f_I(\mathbf{x})g_J(\mathbf{x})d\mathbf{x}_I \wedge d\mathbf{x}_K.\end{aligned}$$

For any ordered set of r indices $1 \le i_1 < \cdots < i_r \le n$ there is a unique ordered set of $n-r$ indices $1 \le k_1 < \cdots < k_{n-r} \le n$ such that $d\mathbf{x}_I \wedge d\mathbf{x}_K$ is nontrivial, and for that ordered set of indices $d\mathbf{x}_I \wedge d\mathbf{x}_K = \epsilon(IK)dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$; then there is a unique ordered set of r indices $1 \le j_1 < \cdots < j_r \le n$ such that $\epsilon(JK) \ne 0$, and clearly $J = I$. Thus the choice of an ordered set of indices I determines unique ordered sets of indices J and K for which the associated term in the sum (5.54) is nontrivial, and since $\epsilon(IK)d\mathbf{x}_I \wedge d\mathbf{x}_K = dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n$ that sum reduces to (5.53).

If $\boldsymbol{\phi} : D \longrightarrow E$ is a mapping from an open subset $D \subset \mathbb{R}^m$ into an open subset $E \subset \mathbb{R}^n$ then to any function f in the set E there can be associated the **induced function** $\boldsymbol{\phi}^*(f)$ in the set D, the function defined by

Vindq1 (5.55)
$$\boldsymbol{\phi}^*(f)(\mathbf{x}) = f\big(\boldsymbol{\phi}(\mathbf{x})\big) \quad \text{for all points } \mathbf{x} \in D.$$

It is clear that $\boldsymbol{\phi}^*(c_1f_1 + c_2f_2) = c_1\boldsymbol{\phi}^*(f_1) + c_2\boldsymbol{\phi}^*(f_2)$ for any functions f_1, f_2 in D and any constants $c_1, c_2 \in \mathbb{R}$. To emphasize the main point, and avoid the complications involved in keeping track of how differentiable the differential forms, it will be assumed now that the changes of coordinates and the differential

[2]The version of Hodge duality discussed here is a special case of the general Hodge duality, for which the characteristic property is that
$$\omega \wedge *\sigma = \sum_{I,J}' h_{IJ}f_I(\mathbf{x})g_J(\mathbf{x})dx_1 \wedge \cdots \wedge dx_n$$
where $h_{IJ} = h_{i_1j_1}h_{i_2j_2}\cdots h_{i_nj_n}$ for some positive definite symmetric matrix h_{ij}; actually for some applications in physics the matrix h_{ij} is not even necessarily positive definite, but is a Lorentz metric, having one negative and $n-1$ positive eigenvalues.

forms have derivatives of all orders. Associating to a function $f \in \Lambda^0(E)$ the induced function $\boldsymbol{\phi}^*(f) \in \Lambda^0(D)$ then is a well defined linear mapping

$$\boldsymbol{\phi}^* : \Lambda^0(E) \longrightarrow \Lambda^0(D). \tag{5.56}$$

More generally to any mapping $\mathbf{f} : E \longrightarrow \mathbb{R}^N$ there can be associated the **induced mapping** $\boldsymbol{\phi}^*(\mathbf{f})$ in the set D, the mapping defined by

Vindq1mp
$$\boldsymbol{\phi}^*(\mathbf{f})(\mathbf{x}) = \mathbf{f}\big(\boldsymbol{\phi}(\mathbf{x})\big) \quad \text{for all points } \mathbf{x} \in D. \tag{5.57}$$

Differential forms $\omega \in \Lambda^r(E)$ in an open set $E \subset \mathbb{R}^n$ are just mappings from E into a vector space $\mathbb{R}^N$, expressed in terms of a basis for $\mathbb{R}^N$ that is determined by the basis $dy_1, \dots, dy_n$ for the vector space $\mathbb{R}^n$ if $y_1, \dots, y_n$ are the coordinates in $\mathbb{R}^n$; on the other hand differential forms $\omega \in \Lambda^r(D)$ in an open set $D \subset \mathbb{R}^m$ are mappings from D into a vector space R^m, expressed in terms of a basis for $\mathbb{R}^m$ that is determined by the basis $dx_1, \dots, dx_m$ for the vector space $\mathbb{R}^m$ if $x_1, \dots, x_m$ are the coordinates in $\mathbb{R}^m$. The induced differential form under the mapping $\boldsymbol{\phi} : D \longrightarrow E$ again is just the induced mappings as in (5.57), but must also reflect the change in the bases for the two vector spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ defined by the mapping $\boldsymbol{\phi}$; and by the chain rule the derivatives of these functions are the row vectors

$$y_i'(\mathbf{x}) = \left(\frac{\partial y_i}{\partial x_1}, \quad \frac{\partial y_i}{\partial x_2}, \quad \cdots, \quad \frac{\partial y_i}{\partial x_m} \right),$$

which can be written equivalently in terms of the bases dx_i of the space $\mathbb{R}^m$ and dy_i of the space $\mathbb{R}^n$ as the linear combination

Vinddfrl
$$dy_i = \sum_{j=1}^{m} \frac{\partial y_i}{\partial x_j} dx_j \tag{5.58}$$

of the basis vectors dx_j. The **induced differential form**, $\boldsymbol{\phi}^*(\omega) \in \Lambda^r(D)$ on D consequently is defined by

Vindq3
$$\begin{aligned} \boldsymbol{\phi}^*\Big(\sum_I f_I(\mathbf{y}) dy_I\Big) &= \sum_{I,J} f_I\big(\boldsymbol{\phi}(\mathbf{x})\big) \frac{\partial y_{i_1}}{\partial x_{j_1}} dx_{j_1} \wedge \cdots \wedge \frac{\partial y_{i_r}}{\partial x_{j_r}} dx_{j_r} \\ &= \sum_{I,J} f_I\big(\boldsymbol{\phi}(\mathbf{x})\big) \frac{\partial y_{i_1}}{\partial x_{j_1}} \cdots \frac{\partial y_{i_r}}{\partial x_{j_r}} \, dx_{j_1} \wedge \cdots \wedge dx_{j_r}. \end{aligned} \tag{5.59}$$

This is sometimes also referred to as the **pullback** of the differential form ω from E to D. It is clear that $\boldsymbol{\phi}^*\left(\Lambda^r(E)\right) \subset \Lambda^r(D)$ and that

Vindqqdfx1
$$\boldsymbol{\phi}^*(c_1\omega_1 + c_2\omega_2) = c_1\boldsymbol{\phi}^*(\omega_1) + c_2\boldsymbol{\phi}^*(\omega_2) \quad \text{for any } \omega_1, \omega_2 \in \Lambda^r(E) \text{ and } c_1, c_2 \in \mathbb{R}, \tag{5.60}$$

so the mapping $\boldsymbol{\phi}^*$ is a linear mapping

Vindqqdf1
$$\boldsymbol{\phi}^* : \Lambda^r(E) \longrightarrow \Lambda^r(D); \tag{5.61}$$

and it is also clear that

Vindzz1 (5.62) $$\boldsymbol{\phi}^*(\omega \wedge \sigma) = \boldsymbol{\phi}^*(\omega) \wedge \boldsymbol{\phi}^*(\sigma) \quad \text{for any } \omega \in \Lambda^r(E) \text{ and } \sigma \in \Lambda^s(E).$$

If $D, E \subset \mathbb{R}^n$ are both subsets of vector spaces of the same dimension and $\omega \in \Lambda^n_{(}E)$ is a differential n-form in E then $\omega(\mathbf{y}) = f(\mathbf{y})dy_1 \wedge \cdots \wedge dy_n$ for some function f in E and (Vindq3 5.59) reduces to

Vindexn (5.63) $$\boldsymbol{\phi}^*\Big(f(\mathbf{y})dy_1 \wedge \cdots \wedge dy_n\Big) = \sum_J f\big(\boldsymbol{\phi}(\mathbf{x})\big)\frac{\partial y_1}{\partial x_{j_1}} \cdots \frac{\partial y_n}{\partial x_{j_n}} dx_{j_1} \wedge \cdots \wedge dx_{j_n}$$
$$= f\big(\boldsymbol{\phi}(\mathbf{x})\big) \det \boldsymbol{\phi}'(\mathbf{x}) dx_1 \wedge \cdots \wedge dx_n.$$

This illustrates the role of exterior algebras as convenient tools in handling determinants, and in a sense underlies all the applications of differential forms in analysis and geometry.

Vcompin **Theorem 5.5** *If $\boldsymbol{\phi} : D \longrightarrow E$ is a $\mathcal{C}^\infty$ mapping between open subsets $D \subset \mathbb{R}^m$ and $E \subset \mathbb{R}^n$ while $\boldsymbol{\psi} : E \longrightarrow F$ is a $\mathcal{C}^\infty$ mapping between open subsets $E \subset \mathbb{R}^n$ and $F \subset \mathbb{R}^p$ then $(\boldsymbol{\psi} \circ \boldsymbol{\phi})^* : \Lambda^r(F) \longrightarrow \Lambda^r(D)$ is the composition $(\boldsymbol{\psi} \circ \boldsymbol{\phi})^* = \boldsymbol{\phi}^* \circ \boldsymbol{\psi}^*$ of the mapping $\boldsymbol{\phi}^* : \Lambda^r(E) \longrightarrow \Lambda^r(D)$ and the mapping $\boldsymbol{\psi}* : \Lambda^r(F) \longrightarrow \Lambda^r(E)$.*

Proof: If the coordinates in D are $t_1, \ldots, t_m$, the coordinates in E are $x_1, \ldots, x_n$, and the coordinates in F are $y_1, \ldots, y_l$, then for any differential form $\omega(\mathbf{y}) = \sum_I f_I(\mathbf{y})d\mathbf{y}_I$ it follows that

$$\boldsymbol{\psi}^*(\omega)(\mathbf{x}) = \sum_{I,J} f_I\big(\boldsymbol{\psi}(\mathbf{x})\big)\frac{\partial \mathbf{y}_I}{\partial \mathbf{x}_J}(\mathbf{x})d\mathbf{x}_J$$

and

$$\boldsymbol{\phi}^*\big(\boldsymbol{\psi}^*(\omega)\big)(\mathbf{t}) = \sum_{I,J,K} f_I\big(\boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{t}))\big)\frac{\partial \mathbf{y}_I}{\partial \mathbf{x}_J}\big(\boldsymbol{\phi}(\mathbf{t})\big)\frac{\partial \mathbf{x}_J}{\partial \mathbf{t}_K}(\mathbf{t})d\mathbf{t}_K.$$

On the other hand for the composite mapping $\boldsymbol{\psi} \circ \boldsymbol{\phi}$

$$(\boldsymbol{\psi} \circ \boldsymbol{\phi})^*(\omega)(\mathbf{t}) = \sum_{I,K} f_I\big(\boldsymbol{\psi} \circ \boldsymbol{\phi})(\mathbf{t})\frac{\partial \mathbf{y}_I}{\partial \mathbf{t}_K}(\mathbf{t})d\mathbf{t}_K,$$

and it follows from the chain rule that

$$\frac{\partial \mathbf{y}_I}{\partial \mathbf{t}_K} = \sum_J \frac{\partial \mathbf{y}_I}{\partial \mathbf{x}_J}(\boldsymbol{\phi}(\mathbf{t}))\frac{\partial \mathbf{x}_J}{\partial \mathbf{t}_K}(\mathbf{t}).$$

Consequently $(\boldsymbol{\psi} \circ \boldsymbol{\phi})^*(\omega) = \boldsymbol{\phi}^*\big(\boldsymbol{\psi}^*(\omega)\big)$, which suffices for the proof.

In addition to the algebraic operations and change of variables formula for differential forms discussed so far, there are analytic operations of differentiation and integration of differential forms. The **exterior derivative** of a differential form $\omega(\mathbf{x}) \in \Lambda^r_{(k)}(U)$ for $k > 0$ of degree r is defined to be a differential form $d\omega(\mathbf{x}) \in \Lambda^{r+1}_{(k-1)}(U)$ of degree $r+1$. Explicitly, if $f(\mathbf{x}) \in \Lambda^0_{(k)}$ is a differential form of degree 0, a function defined in the subset $U \subset \mathbb{R}^n$, its exterior derivative is defined to be the derivative of that function expressed as a differential form, so

Vstq1 (5.64)
$$df(\mathbf{x}) = \sum_{j=1}^{n} \partial_j f(\mathbf{x}) dx_j \quad \text{if } f \in \Lambda^0_k(U).$$

If $\omega(\mathbf{x}) = \sum_I f_I(\mathbf{x}) d\mathbf{x}_I \in \Lambda^r_{(k)}$ is a differential form of degree $r > 0$ its exterior derivative is defined by

Vstdfq1 (5.65)
$$d\omega(\mathbf{x}) = \sum_I df_I(\mathbf{x}) \wedge d\mathbf{x}_I,$$

so that

Vstdfq2 (5.66)
$$d\Big(\sum_{i_1,\ldots i_r=1}^{n} f_{i_1\cdots i_r}(\mathbf{x})\, dx_{i_1}\wedge\cdots\wedge dx_{i_r}\Big) = \sum_{j,i_1,\ldots i_r=1}^{n} \partial_j f_{i_1\cdots i_r}(\mathbf{x})\, dx_j\wedge dx_{i_1}\wedge\cdots\wedge dx_{i_r}.$$

A clear consequence of this is that $d\omega = 0$ for any n-form ω in an open subset $U \subset \mathbb{R}^n$.

Vexdfbs **Theorem 5.6** *Exterior differentiation is a linear mapping*

Vexdfbsq1 (5.67)
$$d : \Lambda^r_{(k)}(U) \longrightarrow \Lambda^{r+1}_{(k-1)}(U)$$

of $\mathcal{C}^\infty$ differential r-forms in an open subset $U \subset \mathbb{R}^n$ for $k \ge 1$, such that
(i) $d\big(\omega(\mathbf{x}) \wedge \sigma(\mathbf{x})\big) = d\omega(\mathbf{x}) \wedge \sigma(\mathbf{x}) + (-1)^r \omega(\mathbf{x}) \wedge d\sigma(\mathbf{x})$
if $\omega(\mathbf{x}) \in \Lambda^r_{(k)}(U)$ and $\sigma(\mathbf{x}) \in \Lambda^s_{(k)}(U)$ for $k \ge 1$, while
(ii) $d\,d\,\omega(\mathbf{x}) = 0$ *for any differential form $\omega \in \Lambda^r_{(2)}(U)$.*

Proof: That exterior differentiation is a linear mapping on r-forms is an immediate consequence of the definitions (5.64) and (5.65).
(i) For functions it follows from the usual rules for differentiation of a product that

Vexdfbsq2 (5.68)
$$d\big(f(\mathbf{x})g(\mathbf{x})\big) = f(\mathbf{x})dg(\mathbf{x}) + g(\mathbf{x})df(\mathbf{x}).$$

In general if $\omega = \sum_I f_I(\mathbf{x}) d\mathbf{x}_I \in \Lambda^r(D)$ and $\sigma = \sum_J g_J(\mathbf{x}) d\mathbf{x}_J \in \Lambda^s(D)$ it follows from (5.68) that

$$\begin{aligned}
d\big(\omega(\mathbf{x}) \wedge \sigma(\mathbf{x})\big) &= d\Bigg(\sum_{I,J} f_I(\mathbf{x}) g_J(\mathbf{x}) d\mathbf{x}_I \wedge d\mathbf{x}_J\Bigg) = \sum_{I,J} d\Bigg(f_I(\mathbf{x}) g_J(\mathbf{x})\Bigg) \wedge d\mathbf{x}_I \wedge d\mathbf{x}_J \\
&= \sum_{I,J} \Bigg(g_J(\mathbf{x}) df_I(\mathbf{x}) + f_I(\mathbf{x}) dg_J(\mathbf{x})\Bigg) \wedge d\mathbf{x}_I \wedge d\mathbf{x}_J \\
&= \sum_{I,J} df_I(\mathbf{x})) \wedge d\mathbf{x}_I \wedge g_J(\mathbf{x}) d\mathbf{x}_J + (-1)^r \sum_{I,J} f_I(\mathbf{x}) d\mathbf{x}_I \wedge dg_J(\mathbf{x}) \wedge d\mathbf{x}_J,
\end{aligned}$$

where the sign $(-1)^r$ arises since the 1-form $dg_J(\mathbf{x})$ must be moved across the r differentials in $d\mathbf{x}_I$.
(ii) If $\omega(\mathbf{x}) = \sum_I f_I(\mathbf{x})d\mathbf{x}_I \in \Lambda^r$ then

$$
\begin{aligned}
d\big(d\omega(\mathbf{x})\big) &= d\Big(\sum_I df_I(\mathbf{x}) \wedge d\mathbf{x}_I\Big) = d\Big(\sum_{I,j} \frac{\partial f_I(\mathbf{x})}{\partial x_j} \wedge dx_j \wedge d\mathbf{x}_I\Big) \\
&= \sum_{I,j,k} \frac{\partial^2 f_I(\mathbf{x})}{\partial x_j \partial x_k} dx_k \wedge dx_j \wedge d\mathbf{x}_I,
\end{aligned}
$$

which is zero since $\partial_{jk} f_I(\mathbf{x})$ is symmetric in the indices j,k by Theorem multord 2.4 while $dx_k \wedge dx_j$ is skew-symmetric in these indices. That suffices for the proof.

A differential form $\omega(\mathbf{x})$ is said to be **closed** if $d\omega(\mathbf{x}) = 0$, and it is said to be **exact** if $\omega(\mathbf{x}) = d\sigma(\mathbf{x})$ for some differential form $\sigma(\mathbf{x})$. A consequence of Theorem Vexdfbs 5.6 (ii) is that any exact differential form is necessarily a closed differential form; but the converse is not true in general. For example, in the complement of the origin in the plane $\mathbb{R}^2$ of the variables x_1, x_2 the function $\theta(x_1, x_2) = \tan^{-1}(x_2/x_1)$, the angle in polar coordinates, is well defined in any half-plane bounded by a line through the origin, such as the half-planes $x_1 > 0$ and $x_1 < 0$, but it cannot be defined in the entire complement of the origin as a single-valued function. However

$$
\omega(\mathbf{x}) = d\theta(x_1, x_2) = -\frac{x_2}{x_1^2 + x_2^2} dx_1 + \frac{x_1}{x_1^2 + x_2^2} dx_2
$$

is a well defined closed differential 1-form in the complement of the origin in $\mathbb{R}^2$. If $\omega(\mathbf{x})$ were an exact differential form in the full complement of the origin it would be the exterior derivative $\omega(\mathbf{x}) = df(\mathbf{x})$ of a well defined function $f(\mathbf{x})$ in the complement of the origin; but then $df(\mathbf{x}) = d\theta(\mathbf{x})$ so that $d\big(f(\mathbf{x}) - \theta(\mathbf{x})\big) = 0$ hence $f(\mathbf{x}) - \theta(\mathbf{x}) = c$ is a constant in the complement of the origin, which would imply that $\theta(\mathbf{x}) = f(\mathbf{x}) - c$ is a well defined function in the complement of the origin, a contradiction. However it is at least true locally that any closed form is exact.

Vpoin **Theorem 5.7 (Poincaré's Lemma)** *If $\omega(\mathbf{x}) \in \Lambda^r(\overline{\Delta})$ is a closed $\mathcal{C}^\infty$ differential form of degree $r > 0$ in the standard cell $\overline{\Delta} \subset \mathbb{R}^n$ then there is a $\mathcal{C}^\infty$ differential form $\sigma(\mathbf{x}) \in \Lambda^{r-1}(\overline{\Delta})$ of degree $r-1$ in $\overline{\Delta}$ such that $\omega(\mathbf{x}) = d\sigma(\mathbf{x})$.*

Proof: Consider differential r-forms $\omega(\mathbf{x}) = \sum_I' f(\mathbf{x})d\mathbf{x}_I$ that involve only the differentials $dx_1, dx_2, \ldots, dx_k$ for some index $1 \le k \le n$, so that

$$
\omega(\mathbf{x}) = \textstyle\sum_{1 \le i_1 < \cdots < i_r \le k} f_{i_1,\cdots,i_r}(\mathbf{x}) dx_{i_1} \wedge \cdots \wedge dx_{i_r}.
$$

The theorem will be demonstrated by induction on the index k. If $k = 1$ then $\omega(\mathbf{x}) = f(\mathbf{x})dx_1$ and $0 = d\omega(\mathbf{x}) = \sum_{i=2}^n \partial_i f(\mathbf{x}) dx_i \wedge dx_1$ so $\partial_i f(\mathbf{x}) = 0$ for $i > 1$ and consequently $f(\mathbf{x}) = f(x_1)$ is a function of the variable x_1 alone, where $x_1 \in [0,1]$. The integral $g(x_1) = \int_0^{x_1} f$ is a function of the variable $x_1 \in I$ for which $dg(\mathbf{x}) = dg(x_1) = f(x_1)dx_1 = \omega(\mathbf{x})$, which establishes the desired result in this special case. Assume then that the theorem has been demonstrated for an index k, and consider a differential r-form $\omega(\mathbf{x})$ that involves only the differentials $dx_1, \ldots, dx_k, dx_{k+1}$. This differential form can be written

$$\omega(\mathbf{x}) = \omega_1(\mathbf{x}) \wedge dx_{k+1} + \omega_2(\mathbf{x})$$

where the differential $(r-1)$-form $\omega_1(\mathbf{x})$ and the differential r-form $\omega_2(\mathbf{x})$ are in reduced form and involve only the differentials $dx_1, \ldots, dx_k$, so that

$$\omega_1(\mathbf{x}) = \textstyle\sum_I' f_I(\mathbf{x}) d\mathbf{x}_I \quad \text{and} \quad \omega_2(\mathbf{x}) = \textstyle\sum_J' g_J(\mathbf{x}) d\mathbf{x}_J$$

for indices

$$1 \le i_1 < i_2 < \ldots < i_{r-1} \le k \quad \text{and} \quad 1 \le j_1 < j_2 < \cdots < j_r \le k.$$

By assumption

$$\begin{aligned} 0 = d\omega(\mathbf{x}) &= d\omega_1(\mathbf{x}) \wedge dx_{k+1} + d\omega_2(\mathbf{x}) \\ &= \textstyle\sum_I' \sum_{i=1}^n \partial_i f_I(\mathbf{x}) dx_i \wedge d\mathbf{x}_I \wedge dx_{k+1} + \sum_J' \sum_{j=1}^n \partial_j g_J(\mathbf{x}) dx_j \wedge d\mathbf{x}_J. \end{aligned}$$

The only terms in this expression that involve dx_{k+1} and dx_l for $l > k+1$ are just in the first sum and just for the index $i = l$; and the basis vectors $dx_l \wedge d\mathbf{x}_I \wedge dx_{k+1}$ are linearly independent since $1 \le i_1 < i_1 < \cdots < i_{r-1} \le k$ and $k < k+1 < l$. Consequently $\partial_l f_I(\mathbf{x}) = 0$ for $l > k+1$, so $f_I(\mathbf{x}) = f_I(x_1, \cdots, x_{k+1})$ is a function just of the variables $x_1, \ldots, x_{k+1}$ for all $l > k+1$, where $0 \le x_i \le 1$. The integrals

$$h_I(x_1, \ldots, x_{k+1}) = \int_0^{x_{k+1}} f_I(x_1, \ldots, x_k, t) dt$$

then are functions in the cell Δ for which $\partial_{k+1} h_I(\mathbf{x}) = f_I(\mathbf{x})$, by the fundamental theorem of calculus for functions of a single variable; hence the differential $(r-1)$-form

$$\omega_3(\mathbf{x}) = \textstyle\sum_I' h_I(\mathbf{x}) d\mathbf{x}_I,$$

which involves just the differentials $dx_1, \ldots, dx_k$, satisfies

$$\begin{aligned} d\omega_3(\mathbf{x}) &= \textstyle\sum_I' \Big(\sum_{i=1}^k \partial_i h_I(\mathbf{x}) dx_i \wedge d\mathbf{x}_I + \partial_{k+1} h_I(\mathbf{x}) dx_{k+1} \wedge d\mathbf{x}_I \Big) \\ &= \textstyle\sum_I' \Big(\sum_{i=1}^k \partial_i h_I(\mathbf{x}) dx_i \wedge d\mathbf{x}_I + f_I(\mathbf{x}) dx_{k+1} \wedge d\mathbf{x}_I \Big) \\ &= \omega_4(\mathbf{x}) + (-1)^{r-1} \omega_1(\mathbf{x}) \wedge dx_{k+1} \end{aligned}$$

for a differential form $\omega_4(\mathbf{x})$ that involves only the differentials $dx_1, \ldots, dx_k$. Then

$$\begin{aligned} \omega(\mathbf{x}) - (-1)^{r-1} d\omega_3(\mathbf{x}) &= \omega_1(\mathbf{x}) \wedge dx_{k+1} + \omega_2(\mathbf{x}) - (-1)^{r-1}\omega_4(\mathbf{x}) - \omega_1(\mathbf{x}) \wedge dx_{k+1} \\ &= \omega_2(\mathbf{x}) - (-1)^{r-1}\omega_4(\mathbf{x}) \end{aligned}$$

involves only the differentials $dx_1, \ldots, dx_k$ and

$$d\big(\omega(\mathbf{x}) - (-1)^{r-1} d\omega_3(\mathbf{x})\big) = 0;$$

by the induction hypothesis then there is a differential form $\omega_5(\mathbf{x})$ in the cell Δ such that

$$\omega(\mathbf{x}) - (-1)^{r-1} d\omega_3(\mathbf{x}) = d\omega_5(\mathbf{x}),$$

hence such that $\omega(\mathbf{x}) = d\sigma(\mathbf{x})$ where $\sigma(\mathbf{x}) = (-1)^{r-1}\omega_3(\mathbf{x}) + \omega_5(\mathbf{x})$. That establishes the induction step and thereby concludes the proof.

The preceding theorem is an example of an integrability theorem for a linear system of partial differential equations. For example, the existence of a function h such that $dh = \sigma$ for a 1-form $\sigma = \sum_{j=1}^{n} f_j(\mathbf{x}) dx_j$ in $\mathbb{R}^n$ amounts to the existence of a solution of the system of partial differential equations $\partial_i h(\mathbf{x}) = f_i(\mathbf{x})$ for $1 \le i \le n$. It is clear that if there is a solution $h(\mathbf{x})$ then $\partial_j f_i(\mathbf{x}) = \partial_j \partial_i h(\mathbf{x}) = \partial_i \partial_j h(\mathbf{x}) = \partial_i f_j(\mathbf{x})$; that is just the condition that $d\sigma(\mathbf{x}) = 0$, which by Poincaré's Lemma is also a sufficient condition for the existence of the solution $h(\mathbf{x})$. There are corresponding interpretations for Poincaré's Lemma for differential forms of higher degree, although they are a bit more complicated to state without using the machinery of differential forms. Another important property of exterior differentiation is its invariance under the pullback of differential forms.

Vindth **Theorem 5.8** *If $\boldsymbol{\phi} : D \longrightarrow E$ is a $\mathcal{C}^\infty$ mapping from an open subset $D \subset \mathbb{R}^m$ into an open subset $E \subset \mathbb{R}^n$ then for any $\mathcal{C}^\infty$ differential form $\omega \in \Lambda^r(E)$*

Vindthqq1 (5.69)
$$d\boldsymbol{\phi}^*(\omega) = \boldsymbol{\phi}^*(d\omega) \in \Lambda^{r+1}_{(k-1)}.$$

Proof: For a differential form $\omega \in \Lambda^r(E)$ written explicitly as $\omega = \sum_I f_I(\mathbf{y}) dy_I$ it follows from the definition (5.59) that

$$\boldsymbol{\phi}^*(\omega) = \sum_{I,J} f_I\big(\boldsymbol{\phi}(\mathbf{x})\big) \frac{\partial y_{i_1}}{\partial x_{j_1}} \cdots \frac{\partial y_{i_r}}{\partial x_{j_r}} dx_{j_1} \wedge \cdots dx_{j_r}.$$

The exterior derivative of ω is

$$d\omega = \sum_{I,k} \frac{\partial f_I}{\partial y_k} dy_k \wedge dy_I$$

and it follows from the definition (5.59) that

$$\boldsymbol{\phi}^*(d\omega) = \sum_{I,J,k,l} \frac{\partial f_I}{\partial y_k}\big(\boldsymbol{\phi}(\mathbf{x})\big) \frac{\partial y_k}{\partial x_l} \frac{\partial y_{i_1}}{\partial x_{j_1}} \cdots \frac{\partial y_{i_r}}{\partial x_{j_r}}\, dx_l \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r}$$

On the other hand

$$\begin{aligned}
d\boldsymbol{\phi}^*(\omega) &= \sum_{I,J,l} \frac{\partial}{\partial x_l}\left(f_I\big(\boldsymbol{\phi}(\mathbf{x})\big)\frac{\partial y_{i_1}}{\partial x_{j_1}}\cdots\frac{\partial y_{i_r}}{\partial x_{j_r}}\right) dx_l\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_r}\\
&= \sum_{I,J,k,l} \frac{\partial f_I}{\partial y_k}\big(\boldsymbol{\phi}(\mathbf{x})\big)\,\frac{\partial y_k}{\partial x_l}\,\frac{\partial y_{i_1}}{\partial x_{j_1}}\cdots\frac{\partial y_{i_r}}{\partial x_{j_r}}\,dx_l\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_r}\\
&\quad+\sum_{I,J,k,l} f_I\big(\boldsymbol{\phi}(\mathbf{x})\big)\frac{\partial^2 y_{i_1}}{\partial x_l\partial x_{j_1}}\,\frac{\partial y_{i_2}}{\partial x_{j_2}}\cdots\frac{\partial y_{i_r}}{\partial x_{j_r}} dx_l\wedge dx_{j_1}\wedge\cdots\wedge dx_{j_r}\\
&\quad+\cdots
\end{aligned}$$

where the remaining terms involve successively the second derivatives of the the partial derivatives $\dfrac{\partial y_{i_2}}{\partial x_l\partial x_{j_2}}$, $\dfrac{\partial y_{i_3}}{\partial x_l\partial x_{j_3}}$, and so on. These terms that involve the second derivatives all vanish, since for example $\dfrac{\partial^2 y_{i_1}}{\partial x_l\partial x_{j_1}}$ is symmetric in the indices l and j_1 while $dx_l\wedge dx_{j_1}$ is skew-symmetric in these indices. That suffices for the proof.

The exterior derivatives of differential forms in $\mathbb{R}^3$ have classical forms, which are possibly quite familiar from physics. For the 2-form $\Omega_{\mathbf{f}}(\mathbf{x})$ associated to a vector field $\mathbf{f}(\mathbf{x})$ in an open set $U\subset\mathbb{R}^3$ as in (5.37) it follows quite directly from the definition of exterior derivative that

cldf1 (5.70)
$$d\Omega_{\mathbf{f}}(\mathbf{x}) = \left(\frac{\partial f_1(\mathbf{x})}{\partial x_1}+\frac{\partial f_2(\mathbf{x})}{\partial x_2}+\frac{\partial f_3(\mathbf{x})}{\partial x_3}\right)dx_1\wedge dx_2\wedge dx_3.$$

The coefficient of this differential 3-form is known as the **divergence** of the vector field $\mathbf{f}$ and is often denoted by $\mathbf{div}\,\mathbf{f}(\mathbf{x})$, so by definition it is the function

cldf2 (5.71)
$$\mathbf{div}\ \mathbf{f}(\mathbf{x}) = \frac{\partial f_1(\mathbf{x})}{\partial x_1}+\frac{\partial f_2(\mathbf{x})}{\partial x_2}+\frac{\partial f_3(\mathbf{x})}{\partial x_3}.$$

A traditional alternative notation is expressed in terms of the vector differential operator

cldf3 (5.72)
$$\nabla = \left\{\frac{\partial}{\partial x_1},\quad \frac{\partial}{\partial x_2},\quad \frac{\partial}{\partial x_3}\right\}.$$

When applied to functions $\nabla f(\mathbf{x})$ is the gradient of the function $f(\mathbf{x})$, as discussed on page 34; and the dot product of this operator with a vector field is the function

cldf4 (5.73)
$$\nabla\cdot\mathbf{f}(\mathbf{x}) = \frac{\partial}{\partial x_1}f_1(\mathbf{x})+\frac{\partial}{\partial x_2}f_2(\mathbf{x})+\frac{\partial}{\partial x_3{}_3}f_3(\mathbf{x}) = \mathbf{div}\,\mathbf{f}(\mathbf{x}).$$

In terms of this operator (5.70) can be rewritten

cldf5 (5.74)
$$d\Omega_{\mathbf{f}}(\mathbf{x}) = \begin{cases} \mathbf{div}\,\mathbf{f}(\mathbf{x})\ dx_1\wedge dx_2\wedge dx_3\\[2ex] \text{or}\\[2ex] \nabla\cdot\mathbf{f}(\mathbf{x})\ dx_1\wedge dx_2\wedge dx_3.\end{cases}$$

For the 1-form $\omega_{\mathbf{f}}(\mathbf{x})$ associated to a vector field $\mathbf{f}(\mathbf{x})$ in an open set $U \subset \mathbb{R}^3$ as in (5.37) it follows by a straightforward calculation that

cldf6 (5.75)
$$d\omega_{\mathbf{f}}(\mathbf{x}) = \big(\partial_2 f_3(\mathbf{x}) - \partial_3 f_2(\mathbf{x})\big)dx_2 \wedge dx_3 + \big(\partial_3 f_1(\mathbf{x}) - \partial_1 f_3(\mathbf{x})\big)dx_3 \wedge dx_1 + \big(\partial_1 f_2(\mathbf{x}) - \partial_2 f_1(\mathbf{x})\big)dx_1 \wedge dx_2;$$

this is the 2-form associated to the vector field called the **curl** of the vector field $\mathbf{f}(\mathbf{x})$, so by definition it is the vector field

cldf6a (5.76)
$$\textbf{curl } \mathbf{f}(\mathbf{x}) = \Big\{\partial_2 f_3(\mathbf{x}) - \partial_3 f_2(\mathbf{x}), \quad \partial_3 f_1(\mathbf{x}) - \partial_1 f_3(\mathbf{x}), \quad \partial_1 f_2(\mathbf{x}) - \partial_2 f_1(\mathbf{x})\Big\}.$$

It follows from (5.41) that the curl of a vector field can be written in terms of the operator ∇ as

cldf6xx (5.77)
$$\textbf{curl } \mathbf{f}(\mathbf{x}) = \nabla \times \mathbf{f}(\mathbf{x}),$$

so that

cldf77 (5.78)
$$d\omega_{\mathbf{f}}(\mathbf{x}) = \begin{cases} \Omega_{\textbf{curl } \mathbf{f}(\mathbf{x})} \\ \text{or} \\ \Omega_{\nabla\times\mathbf{f}(\mathbf{x})}. \end{cases}$$

The condition that $dd\omega = 0$ for any differential form ω, as in Theorem 5.6 (ii), has classical interpretations as well. For any $\mathcal{C}^\infty$ function $f(\mathbf{x})$ in $\mathbb{R}^3$ since $df(\mathbf{x}) = \omega_{\nabla f(\mathbf{x})}$, another interpretation of the first line of (5.37), it follows from (5.78) that $0 = d\,df(\mathbf{x}) = d\omega_{\nabla f(\mathbf{x})} = \Omega_{\nabla\times\nabla f(\mathbf{x})}$ and consequently that

Vclasc1q1 (5.79)
$$\nabla \times \nabla f(\mathbf{x}) = 0 \quad \text{or alternatively} \quad \textbf{curl}\,\textbf{grad} f(\mathbf{x}) = 0.$$

On the other hand for any $\mathcal{C}^\infty$ vector field $\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})\}$ since $d\omega_{\mathbf{f}(\mathbf{x})} = \Omega_{\nabla\times\mathbf{f}(\mathbf{x})}$ by (5.78) it follows from (5.70) that $0 = d\,d\omega_{\mathbf{f}(\mathbf{x})} = d\Omega_{\nabla\times\mathbf{f}(\mathbf{x})} = \nabla\cdot\big(\nabla\times\mathbf{f}(\mathbf{x})\big)\,dx_1\wedge dx_2\wedge dx_3$ and consequently that

Vclasc1q2 (5.80)
$$\nabla\cdot\big(\nabla\times\mathbf{f}(\mathbf{x})\big) = 0 \quad \text{or alternatively} \quad \textbf{div}\,\textbf{curl }\mathbf{f}(\mathbf{x}) = 0.$$

Both (5.79) and (5.80) of course can be demonstrated by direct calculation from the definition of the differential operator ∇. However Poincaré's Lemma is a far less trivial result, so its classical consequences cannot be proved merely by a direct calculation, as should be expected. If $\mathbf{f}(\mathbf{x})$ is a $\mathcal{C}^\infty$ vector field in $\mathbb{R}^3$ the condition that $\mathbf{f}(\mathbf{x}) = \nabla g(\mathbf{x})$ for a function $g(\mathbf{x})$ in Δ means in terms of differential forms that $\omega_{\mathbf{f}}(\mathbf{x}) = dg(\mathbf{x})$; and by Poincaré's Lemma that is the case for a vector field in a cell $\Delta \subset \mathbb{R}^3$ if $0 = d\omega_{\mathbf{f}(\mathbf{x})} = \Omega_{\nabla\times\mathbf{f}(\mathbf{x})}$, consequently

Vclasc2q1 (5.81)
$$\mathbf{f}(\mathbf{x}) = \nabla g(\mathbf{x}) \quad \text{for a function } g(\mathbf{x}) \text{ in } \Delta \text{ if } \nabla\times\mathbf{f}(\mathbf{x}) = 0,$$

or alternatively in the even more classical notation

Vclasc2q1a

$$\mathbf{f}(\mathbf{x}) = \mathbf{grad}\ g(\mathbf{x}) \quad \text{for a function } g(\mathbf{x}) \text{ in } \Delta \text{ if } \mathbf{curl}\ \mathbf{f}(\mathbf{x}) = 0. \tag{5.82}$$

On the other hand that $\mathbf{f}(\mathbf{x}) = \nabla \times \mathbf{g}(\mathbf{x})$ for a vector field $g(\mathbf{x})$ in Δ means in terms of differential forms that $\Omega_{\mathbf{f}(\mathbf{x})} = d\omega_{\mathbf{g}(\mathbf{x})}$; and by Poincaré's Lemma that is the case for differential forms in a cell $\Delta \subset \mathbb{R}^3$ if $0 = d\Omega_{\mathbf{f}(\mathbf{x})} = \nabla \cdot \mathbf{f}(\mathbf{x})\, dx_1 \wedge dx_2 \wedge dx_3$, and consequently

Vclasc2q2

$$\mathbf{f}(\mathbf{x}) = \nabla \times \mathbf{g}(\mathbf{x}) \quad \text{for a vector field } \mathbf{g}(\mathbf{x}) \text{ in } \Delta \text{ if } \nabla \cdot \mathbf{f}(\mathbf{x}) = 0 \tag{5.83}$$

or alternatively

Vclasc2q2a

$$\mathbf{f}(\mathbf{x}) = \mathbf{curl}\ \mathbf{g}(\mathbf{x}) \quad \text{for a vector field } \mathbf{g}(\mathbf{x}) \text{ in } \Delta \text{ if } \mathbf{div}\ \mathbf{f}(\mathbf{x}) = 0. \tag{5.84}$$

PROBLEMS: GROUP I

(1) Find the arc length of the curve $\gamma \in \mathbb{R}^2$ described parametrically by $\mathbf{f}(t) = (\frac{1}{3}t^3 - t, t^2)$ where $0 \leq t \leq 2$.

(2) Evaluate $\int_\gamma \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau}\, ds$ for the vector field $\mathbf{f}(\mathbf{x}) = \{x_1 + x_2, 1\}$ along the curve in $\mathbb{R}^2$ given parametrically by $(x_1, x_2) = (t, t^2)$ for $0 \leq t \leq 1$.

(3) In $\mathbb{R}^4$ consider the differential forms

$$\begin{aligned}
\omega_1 &= x_1 dx_2 + x_3 dx_4, \\
\omega_2 &= (x_1^2 + x_3^2) dx_1 \wedge dx_2 + (x_2^2 + x_4^2) dx_3 \wedge dx_4, \\
\omega_3 &= dx_1 \wedge dx_3 + x_2 dx_1 \wedge dx_4 + x_1 dx_2 \wedge dx_4.
\end{aligned}$$

(i) Find the exterior products $\omega_1 \wedge \omega_1$, $\omega_1 \wedge \omega_2$ and $\omega_2 \wedge \omega_2$ in reduced form.

(ii) Find the exterior derivatives $d\omega_1$, $d\omega_2$ and $d\omega_3$ in reduced form.

(iii) Find the Hodge duals $*\omega_1$ and $*\omega_3$ in reduced form.

(iv) Which of these three differentials forms are closed, that is, satisfy $d\omega = 0$?

(v) For each of the differentials ω that is closed find a differential form σ such that $\omega = d\sigma$. Having found one such form σ, what is the most general such form? (Answer this part of the question without giving explicit formulas for all possible solutions, but just by characterizing all possible solutions using your general knowledge of the properties of differential forms).

(4) Evaluate $\int_\gamma \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau}\, ds$ for the vector field $\mathbf{f}(\mathbf{x}) = \{x_1, x_1^2, x_2\}$ along the curve $x_1 = x_2 = x_3$ from $(0,0,0)$ to $(1,1,1)$ in $\mathbb{R}^3$.
(5) Show that the differential form

$$\omega(\mathbf{x}) = 2x_1 x_2 dx_1 + (x_1^2 + x_3^2) dx_2 + 2x_2 x_3 dx_3$$

in $\mathbb{R}^3$ is closed, and find a function $f(\mathbf{x})$ such that $df(\mathbf{x}) = \omega(\mathbf{x})$. What is the most general function $f(\mathbf{x})$ with this property?

PROBLEMS: GROUP II

(6) Find a curve γ in the plane that has length 1 and has the property that $\int_\gamma \mathbf{f} \cdot \boldsymbol{\tau}\, ds = 0$ for the vector field $\mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

(7) Let γ be a $\mathcal{C}^\infty$ curve in $\mathbb{R}^3$ parametrized by arc length by a $\mathcal{C}^\infty$ mapping $\mathbf{f} : [0, L] \longrightarrow \mathbb{R}^3$, and let $\boldsymbol{\tau}(t) = \mathbf{f}'(t)$ be the unit tangent vector to γ at the point $\mathbf{f}(t)$.
(i) Show that if $\boldsymbol{\tau}'(t) \ne 0$ then that vector is perpendicular to $\boldsymbol{\tau}(t)$.

The length $\|\boldsymbol{\tau}'(t)\|_2 = \kappa(t)$ of this vector is called the **curvature** of γ at the point $\mathbf{f}(t)$, and its inverse $1/\kappa(t)$ is called the **radius of curvature** of γ at the point $\mathbf{f}(t)$.
(ii) Show that $\kappa(t) = 0$ for all $t \in [0, L]$ if and only if γ is a straight line in $\mathbb{R}^3$.
(iii) Show that the radius of curvature of a circle in a plane in $\mathbb{R}^3$ is equal to the radius of the circle.

If $\boldsymbol{\tau}'(t) \ne 0$ the unit vector $\boldsymbol{\nu}(t) = \frac{1}{\kappa(t)}\boldsymbol{\tau}'(t)$ is called the **principal normal vector** to the curve γ. The vector $\boldsymbol{\beta}(t) = \boldsymbol{\tau}(t) \times \boldsymbol{\nu}(t)$, called the **binormal vector** to the curve γ at the point $\mathbf{f}(t)$, is a unit vector that is perpendicular both to the tangent vector $\boldsymbol{\tau}(t)$ and to the principal normal vector $\boldsymbol{\nu}(t)$.
(iv) Show that $\boldsymbol{\beta}'(t) \cdot \boldsymbol{\beta}(t) = \boldsymbol{\beta}'(t) \cdot \mathbf{f}'(t) = 0$ and consequently that $\boldsymbol{\beta}'(t) = -\tau(t)\boldsymbol{\nu}(t)$ for a real-valued function $\tau(t)$ and that $\boldsymbol{\nu}'(t) = -\kappa(t)\boldsymbol{\tau}(t) + \tau(t)\boldsymbol{\beta}(t)$.

The function $\tau(t)$ is called the **torsion** of the curve γ. The set of differential equations in (iv), called the **Serret-Frenet equations**, can be written equivalently

sefrq
$$\begin{pmatrix} \boldsymbol{\tau}'(t) & \boldsymbol{\nu}'(t) & \boldsymbol{\beta}'(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\tau}(t) & \boldsymbol{\nu}(t) & \boldsymbol{\beta}(t) \end{pmatrix} \begin{pmatrix} 0 & -\kappa(t) & 0 \\ \kappa(t) & 0 & -\tau(t) \\ 0 & \tau(t) & 0 \end{pmatrix}. \tag{5.85}$$

(v) Show that $\tau(t) = 0$ if and only if the curve γ is contained in a 2-dimensional linear subspace of $\mathbb{R}^3$.
(vi) Assuming the general result from the elementary theory of ordinary differential equations that all parametrized curves for which the parametrization $\mathbf{f}(t)$ satisfies $f_i'(t) = \sum_{j=1}^3 m_{ij}(t) f_j(t)$ for some given functions $m_{ij}(t)$ differ by a linear transformation of $\mathbb{R}^3$, show that any two curves with the same curvature and torsion can be transformed into one another by a linear transformation in $\mathbb{R}^3$. For this reason curvature and torsion are sometimes called the intrinsic equations of space curves. [Suggestion: The differential equations ($\overset{\text{sefrq}}{5.85}$) can be viewed as a set of linear equations for the vector-valued function in $\mathbb{R}^9$ with components the entries in the 3×3 matrix $\begin{pmatrix} \boldsymbol{\tau}(t) & \boldsymbol{\nu}(t) & \boldsymbol{\beta}(t) \end{pmatrix}$.]
For more information about the geometry of curves see for example volume II of Spivak's *Differential Geometry*.}

(8) To a vector field $\mathbf{f} = \{f_i\}$ in $\mathbb{R}^4$ there can be associated the differential

forms

$$\begin{aligned}\omega_{\mathbf{f}} &= f_1 dx_1 + f_2 dx_2 + f_3 dx_3 + f_4 dx_4,\\ \Omega_{\mathbf{f}} &= f_1 dx_2 \wedge dx_3 \wedge dx_4 + f_2 dx_3 \wedge dx_4 \wedge dx_1 + \cdots .\end{aligned}$$

(i) For any 3 vector fields $\mathbf{f}_i$ show that the vector field $\mathbf{F}$ for which $\Omega_{\mathbf{F}} = \omega_{\mathbf{f}_1} \wedge \omega_{\mathbf{f}_2} \wedge \omega_{\mathbf{f}_3}$ can be interpreted as an analogue $\mathbf{F} = \mathbf{f}_1 \times \mathbf{f}_2 \times \mathbf{f}_3$ of the cross-product in $\mathbb{R}^3$, so as a vector field perpendicular to the vector fields $\mathbf{f}_i$ with the appropriate orientation.
(ii) If σ and τ are 2-forms in $\mathbb{R}^4$ show that $\sigma \wedge \tau$ can be interepreted as the dot product of two vectors in $\mathbb{R}^6$.
(iii) If $\mathbf{f}$ is a vector field and σ is a 2-form in $\mathbb{R}^4$ show that $\omega_{\mathbf{f}} \wedge \sigma$ can be interpreted as a collection of dot products of vectors in R^3.

(9) For the differential 1-form $\omega = \omega_{\mathbf{f}}$ associated to a $\mathcal{C}^\infty$ vector field $\mathbf{f}$ in $\mathbb{R}^3$ calculate and interpret (in terms of the vector field $\mathbf{f}$) the results of the differential operators

$$\mathfrak{d}\ \omega = *d*\omega, \quad \Box\ \omega = d\ \mathfrak{d}\ \omega(\mathbf{x}) - \mathfrak{d}\ d\ \omega(\mathbf{x}).$$

(The operator $\Box$ is Hodge's harmonic operator.)

Chapter 6

Integration of Differential Forms

6.1 Integrals of Differential Forms

dfint

Integrals of differential n-forms in open subsets $E \subset \mathbb{R}^n$ can be defined, but only in terms of the orientation of the vector space $\mathbb{R}^n$; these integrals are **oriented integrals**, extending that notion from spaces of dimension 1 as defined in (intor 4.38) to spaces of dimensions $n > 1$. The integral of a continuous n-form $\omega(\mathbf{y}) = f(\mathbf{y})dy_1 \wedge \cdots \wedge dy_n$ over an open subset $E \subset \mathbb{R}^n$ of an oriented vector space $\mathbb{R}^n$ with the orientation $dy_1 \wedge \cdots \wedge dy_n$ in terms of coordinates $y_1, \ldots, y_n$ is defined by

Vorindfa
$$\int_E f(\mathbf{y})dy_1 \wedge \cdots \wedge dy_n = \int_E f, \tag{6.1}$$

so that

Vintfq2
$$\int_E f(\mathbf{y})dy_{i_1} \wedge \cdots \wedge dy_{i_n} = \epsilon \begin{pmatrix} i_2 & i_2 & \cdots & i_n \\ 1 & 2 & \cdots & n \end{pmatrix} \int_E f(\mathbf{y}) \tag{6.2}$$

where $\epsilon \begin{pmatrix} i_1 & i_2 & \cdots & i_n \\ 1 & 2 & \cdots & n \end{pmatrix}$ is the sign of the permutation $\begin{pmatrix} i_1 & i_2 & \cdots & i_n \end{pmatrix}$ of the integers $\begin{pmatrix} 1 & 2 & \cdots & n \end{pmatrix}$; thus under a permutation of the variables the integral of a differential form changes by the sign of the permutation.

varchdfcr

Theorem 6.1 *If $\boldsymbol{\phi} : D \longrightarrow E$ is a $\mathcal{C}^1$ homeomorphism between two connected open subsets $D, E \subset \mathbb{R}^n$ then for any continuous differential n-form ω in E*

varchdfcrq1
$$\int_E \omega = \pm \int_D \boldsymbol{\phi}^*(\omega), \tag{6.3}$$

where the sign is $+$ if the mapping $\boldsymbol{\phi}$ prerserves the orientation and $-$ if the mapping $\boldsymbol{\phi}$ reverses the orientation.

Proof: Suppose that the vector space $\mathbb{R}^n$ containing D is oriented by the differential form $dx_1 \wedge \cdots \wedge dx_n$ in terms of coordinates $x_1, \ldots, x_n$, and that the vector space $\mathbb{R}^n$ containing E is oriented by the differential form $dy_1 \wedge \cdots \wedge dy_n$ in terms of coordinates $y_1, \ldots, y_n$. The integral of a continuous differential n-form $\omega(\mathbf{y}) = f(\mathbf{y}) dy_1 \wedge \cdots \wedge dy_n$ in E is

$$\int_E \omega = \int_E f, \tag{6.4}$$

as in (6.1); and the integral of the induced differential form $\boldsymbol{\phi}^*(\omega)$ in D, which by (5.63) has the explicit form $\boldsymbol{\phi}^*(\omega)(\mathbf{x}) = f\big(\boldsymbol{\phi}(\mathbf{x})\big) \det \boldsymbol{\phi}'(\mathbf{x}) dx_1 \wedge \cdots \wedge dx_n$, is

$$\int_D \boldsymbol{\phi}^*(\omega) = \int_D (f \circ \boldsymbol{\phi}) \det \boldsymbol{\phi}'. \tag{6.5}$$

If the mapping $\boldsymbol{\phi}$ preserves orientation then $\boldsymbol{\phi}'(\mathbf{x}) > 0$ at all points $\mathbf{x} \in D$ and consequently $\det \boldsymbol{\phi}'(\mathbf{x}) = |\det \boldsymbol{\phi}'(\mathbf{x}|$, while if the mapping $\boldsymbol{\phi}$ reverses orientation then $\boldsymbol{\phi}'(\mathbf{x}) < 0$ at all points $\mathbf{x} \in D$ and consequently $\det \boldsymbol{\phi}'(\mathbf{x}) = -|\det \boldsymbol{\phi}'(\mathbf{x})|$; therefore (6.5) can be rewritten

$$\int_D \boldsymbol{\phi}^*(\omega) = \pm \int_D (f \circ \boldsymbol{\phi}) |\det \boldsymbol{\phi}'| \tag{6.6}$$

where the sign is $+$ if the mapping $\boldsymbol{\phi}$ preserves orientation and $-$ if the mapping $\boldsymbol{\phi}$ reverses orientation. By the change of variables formula (4.49) in Theorem 4.14

$$\int_E f = \int_D (f \circ \boldsymbol{\phi}) |\det \boldsymbol{\phi}'|, \tag{6.7}$$

and (6.3) follows immediately by substituting (6.4) and (6.6) into (6.7), to conclude the proof.

There also are oriented integrals of m-forms over suitable lower-dimensional subsets of the spaces $\mathbb{R}^n$. An m-dimensional **singular cell** in an open subset $U \subset \mathbb{R}^n$ is a continuous mapping $\boldsymbol{\phi} : \overline{\Delta} \longrightarrow U$ from an oriented closed cell $\overline{\Delta} \subset \mathbb{R}^m$ into the open subset $U \subset \mathbb{R}^n$. It is often assumed as a standard convention that the parameter cells are unit cells

$$\overline{\Delta}_m = \Big\{ \{x_1, \ldots, x_m\} \in \mathbb{R}^m \Bigm| 0 \le x_j \le 1 \quad \text{for} \quad 1 \le j \le m \Big\}, \tag{6.8}$$

although other normalizations are sometimes more appropriate. Much of the subsequent discussion will focus on $\mathcal{C}^1$ singular cells, those for which the mappings $\boldsymbol{\phi}$ extend to $\mathcal{C}^1$ mappings from an open neighborhood of the closed cell $\overline{\Delta}$ into U. It is also common to consider $\mathcal{C}^k$ singular cells for any integer $k > 0$ or even $\mathcal{C}^\infty$ singular cells, those for which the mappings $\boldsymbol{\phi}$ are of class $\mathcal{C}^k$ for all integers $k > 0$. Unless explicitly assumed otherwise, though, singular cells are just continuous mappings. A 1-dimensional singular cell is a parametrized curve, as defined on page 60, since an interval $[0, 1]$ is a 1-dimensional cell that is naturally oriented. For singular cells of dimension $m > 1$ however the orientation always must be specified explicitly. As in the case of curves, the mappings describing singular cells are not required to be one-to-one mappings, nor for $\mathcal{C}^k$ singular cells is it required that the derivative $\boldsymbol{\phi}'$ of the mapping $\boldsymbol{\phi}$ have maximal rank; so the images of singular cells are not necessarily submanifolds of U even locally, and that is the reason for the adjective singular. It is even possible for the image of a singular m-cell to be a submanifold of $\mathbb{R}^n$ of a dimension lower than m. However for some purposes it is necessary or convenient to consider more regular singular cells. A **simple singular cell** is a continuous mapping $\boldsymbol{\phi} : \overline{\Delta} \longrightarrow \mathbb{R}^n$ that is a homeomorphism between the cell $\overline{\Delta} \subset \mathbb{R}^m$ and its image $\Gamma = \boldsymbol{\phi}(\overline{\Delta}) \subset \mathbb{R}^n$, where Γ has the induced topology as a subset of $\mathbb{R}^n$. Two m-dimensional singular cells $\boldsymbol{\phi}_1 : \overline{\Delta}_1 \longrightarrow \mathbb{R}^{n_1}$ and $\boldsymbol{\phi}_2 : \overline{\Delta}_2 \longrightarrow \mathbb{R}^{n_2}$ are **equivalent** singular cells if there is an orientation-preserving homeomorphism $\mathbf{h} : \overline{\Delta}_1 \longrightarrow \overline{\Delta}_2$ such that $\boldsymbol{\phi}_1(\mathbf{t}) = \boldsymbol{\phi}_2\big(\mathbf{h}(\mathbf{t})\big)$; this is a natural equivalence relation between singular cells. Equivalent singular cells of course are of the same dimension and have the same image; but in general singular cells having the same image are not necessarily equivalent singular cells and are not necessarily of the same dimension. However two simple singular cells are equivalent if and only if they have the same image; for if $\boldsymbol{\phi}_1 : \overline{\Delta}_1 \longrightarrow \mathbb{R}^n$ and $\boldsymbol{\phi}_2 : \overline{\Delta}_2 \longrightarrow \mathbb{R}^n$ are singular cells having the same image $\boldsymbol{\phi}_1(\overline{\Delta}_1) = \boldsymbol{\phi}_2(\overline{\Delta}_2)$ then $\boldsymbol{\phi}_1(\mathbf{t}) = \boldsymbol{\phi}_2\big((\boldsymbol{\phi}_2^{-1} \circ \boldsymbol{\phi}_1)(\mathbf{t})\big)$, and $\boldsymbol{\phi}_2^{-1} \circ \boldsymbol{\phi}_1$ is a homeomorphism so the singular cells $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are equivalent singular cells. Two $\mathcal{C}^k$ singular cells $\boldsymbol{\phi}_1 : \overline{\Delta}_1 \longrightarrow \mathbb{R}^{n_1}$ and $\boldsymbol{\phi}_2 : \overline{\Delta}_2 \longrightarrow \mathbb{R}^{n_2}$ are **$\mathcal{C}^k$-equivalent** singular cells if there is an orientation preserving $\mathcal{C}^k$ homeomorphism $\mathbf{h} : V_1 \longrightarrow V_2$ between open neighborhoods V_1 of the cell $\overline{\Delta}_1$ and V_2 of the cell $\overline{\Delta}_2$ such that $\boldsymbol{\phi}_1(\mathbf{t}) = \boldsymbol{\phi}_2\big(\mathbf{h}(\mathbf{t})\big)$ for all $\mathbf{t} \in V_1$, and of course the corresponding holds for $\mathcal{C}^\infty$ singular cells. For simplicity the subsequent discussion will generally deal just with $\mathcal{C}^\infty$ singular cells.

If $\boldsymbol{\phi} : \overline{\Delta} \longrightarrow U$ is an m-dimensional $\mathcal{C}^\infty$ singular cell in an open subset $U \subset \mathbb{R}^n$, and if $\omega \in \Lambda^m(U)$ is a continuous differential m-form in the open set U, the integral of the differential form ω over the singular cell $\boldsymbol{\phi}$ is defined by

VintdfA (6.9)
$$\int_{\boldsymbol{\phi}} \omega = \int_{\overline{\Delta}} \boldsymbol{\phi}^*(\omega).$$

If $\boldsymbol{\phi}_1 : \overline{\Delta_1} \longrightarrow U$ and $\boldsymbol{\phi}_2 : \overline{\Delta_2} \longrightarrow U$ are $\mathcal{C}^\infty$ equivalent singular cells, so that $\boldsymbol{\phi}_1(\mathbf{t}) = \boldsymbol{\phi}_2\big(\mathbf{h}(\mathbf{t})\big)$ where $\mathbf{h} : V_1 \longrightarrow V_2$ is an orientation preserving $\mathcal{C}^\infty$ homeomorphism between open neighborhoods V_i of the cells $\overline{\Delta}_i$, then by Theorems Vcompin 5.5 and varchdfcr 6.1

dfinq1a (6.10)
$$\int_{\overline{\Delta}_1} \boldsymbol{\phi}_1^*(\omega) = \int_{\overline{\Delta}_1} \mathbf{h}^*\big(\boldsymbol{\phi}_2^*(\omega)\big) = \int_{\overline{\Delta}_2} \boldsymbol{\phi}_2^*(\omega);$$

thus the integrals of a differential form over $\mathcal{C}^\infty$ equivalent singular m-cells are equal. Since simple singular cells $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are equivalent if and only if they have the same image $\Gamma = \boldsymbol{\phi}_1(\overline{\Delta}_1) = \boldsymbol{\phi}_2(\overline{\Delta}_1)$, the integral of a differential form ω along a simple singular cell is determined uniquely by the image $\Gamma = \boldsymbol{\phi}(\overline{\Delta})$;

consequently for simple singular cells it is customary to set

dfinq2b (6.11)
$$\int_{\boldsymbol{\phi}} \omega = \int_{\Gamma} \omega \quad \text{where} \quad \Gamma = \boldsymbol{\phi}(\overline{\Delta}).$$

This notation of course is limited to the case of simple singular cells.

Line integrals in $\mathbb{R}^n$ are special cases of integrals of differential forms over singular cells, since a parametrized curve is a special case of a singular cell. If $\boldsymbol{\phi} : [0,1] \longrightarrow U$ is a $\mathcal{C}^\infty$ parametrized 1-cell in an open subset $U \subset \mathbb{R}^n$, where the parameter in $[0,1]$ is t and the parameters in $\mathbb{R}^n$ are $x_1, \ldots, x_n$, and if $\omega(\mathbf{x}) = \sum_{j=1}^n f_j(\mathbf{x})dx_j$ is a continuous differential form in U, the induced differential form is $\boldsymbol{\phi}^*(\omega)(t) = \sum_{j=1}^n f_j\big(\boldsymbol{\phi}(t)\big)\dfrac{dx_j}{dt}dt$; and if $\mathbf{f}(\mathbf{x}) = \{f_j(\mathbf{x})\}$ is the vector field with component functions $f_j(\mathbf{x})$ this induced differential form can be written as the dot product $\boldsymbol{\phi}^*(\omega)(t) = \mathbf{f}\big(\boldsymbol{\phi}(t)\big) \cdot \boldsymbol{\phi}'(t)dt$. The integral of the differential 1-form $\omega(\mathbf{x})$ over the singular 1-cell $\boldsymbol{\phi}$ hence is just the familiar line integral

Vintdfd1q2 (6.12)
$$\int_{\boldsymbol{\phi}} \omega = \int_{\boldsymbol{\phi}} \mathbf{f} \cdot \boldsymbol{\tau}\, ds$$

as in (lintq1 5.19); and for simple parametrized curves the integral is determined by the image curve $\gamma = \boldsymbol{\phi}([0,1])$ so it is often denoted just by $\int_\gamma \omega = \int_\gamma \mathbf{f} \cdot \boldsymbol{\tau}\, ds$. For example if $\gamma \subset \mathbb{R}^3$ is the curve from the origin $\mathbf{a} = \{0,0,0\}$ to the point $\mathbf{b} = \{1,1,1\}$ that is the image of the mapping $\boldsymbol{\phi} : [0,1] \longrightarrow \mathbb{R}^3$ given by $\boldsymbol{\phi}(t) = \{t, t^2, t^3\}$, and if $\mathbf{f}$ is the vector field $\mathbf{f}(\mathbf{x}) = \{x_1^2, x_2, x_1x_2x_3\}$, then

$$\begin{aligned}
\int_\gamma \mathbf{f} \cdot \boldsymbol{\tau}\; ds &= \int_\gamma (x_1^2 dx_1 + x_2 dx_2 + x_1x_2x_3 dx_3) \\
&= \int_{[0,1]} \big(t^2 \cdot dt + t^2 \cdot 2tdt + t^6 \cdot 3t^2 dt\big) \\
&= \int_{[0,1]} \big(3t^2 + 3t^8\big)dt = \frac{4}{3}.
\end{aligned}$$

As this example illustrates, the explicit calculation of line integrals is often simpler when these integrals are expressed in terms of differential 1-forms.

Another commonly considered special case is that of the integration of differential 2-forms over 2-dimensional singular cells in $\mathbb{R}^3$. Let $\boldsymbol{\phi} : \overline{\Delta} \longrightarrow U$ be a $\mathcal{C}^\infty$ singular 2-cell in an open subset $U \subset \mathbb{R}^3$, where the parameters in $\mathbb{R}^2$ are t_1, t_2 with the natural orientation $dt_1 \wedge dt_2$ and the parameters in $\mathbb{R}^3$ are x_1, x_2, x_3 with the natural orientation $dx_1 \wedge dx_2 \wedge dx_3$. A continuous 2-form in $\mathbb{R}^3$ can be viewed as the differential form

$$\Omega_{\mathbf{f}}(\mathbf{x}) = f_1(\mathbf{x})dx_2 \wedge dx_3 + f_2(\mathbf{x})dx_3 \wedge dx_1 + f_3(\mathbf{x})dx_1 \wedge dx_2$$

associated to the vector field $\mathbf{f}(\mathbf{x})$, with the usual cyclic notation as in (5.37). The induced differential form under the mapping $\boldsymbol{\phi}$ is the differential 2-form

$$\begin{aligned}\boldsymbol{\phi}^*(\Omega_{\mathbf{f}})(\mathbf{t}) &= \sum_{j_1 j_2=1}^{2} \Big(f_1\big(\boldsymbol{\phi}(\mathbf{t})\big)\frac{\partial x_2}{\partial t_{j_1}}\frac{\partial x_3}{\partial t_{j_2}} + f_2\big(\boldsymbol{\phi}(\mathbf{t})\big)\frac{\partial x_3}{\partial t_{j_1}}\frac{\partial x_1}{\partial t_{j_2}} \\ &\qquad\qquad + f_3\big(\boldsymbol{\phi}(\mathbf{t})\big)\frac{\partial x_1}{\partial t_{j_1}}\frac{\partial x_2}{\partial t_{j_2}}\Big) dt_{j_1}\wedge dt_{j_2} \\ &= \Bigg(f_1\big(\boldsymbol{\phi}(\mathbf{t})\big)\Big(\frac{\partial x_2}{\partial t_1}\frac{\partial x_3}{\partial t_2} - \frac{\partial x_2}{\partial t_2}\frac{\partial x_3}{\partial t_1}\Big) + f_2\big(\boldsymbol{\phi}(\mathbf{t})\big)\Big(\frac{\partial x_3}{\partial t_1}\frac{\partial x_1}{\partial t_2} - \frac{\partial x_3}{\partial t_2}\frac{\partial x_1}{\partial t_1}\Big) \\ &\qquad\qquad + f_3\big(\boldsymbol{\phi}(\mathbf{t})\big)\Big(\frac{\partial x_1}{\partial t_1}\frac{\partial x_2}{\partial t_2} - \frac{\partial x_1}{\partial t_2}\frac{\partial x_2}{\partial t_1}\Big)\Bigg) dt_1\wedge dt_2,\end{aligned}$$

which can be written more succinctly as

Vcxdt (6.13)
$$\boldsymbol{\phi}^*(\Omega_{\mathbf{f}})(\mathbf{t}) = \det\begin{pmatrix} f_1\big(\boldsymbol{\phi}(\mathbf{t})\big) & f_2\big(\boldsymbol{\phi}(\mathbf{t})\big) & f_3\big(\boldsymbol{\phi}(\mathbf{t})\big) \\ \dfrac{\partial x_1}{\partial t_1} & \dfrac{\partial x_2}{\partial t_1} & \dfrac{\partial x_3}{\partial t_1} \\ \dfrac{\partial x_1}{\partial t_2} & \dfrac{\partial x_2}{\partial t_2} & \dfrac{\partial x_3}{\partial t_2}\end{pmatrix} dt_1\wedge dt_2,$$

or using the cross-product introduced in (5.41) as

Vcxcx (6.14)
$$\boldsymbol{\phi}^*(\Omega_{\mathbf{f}})(\mathbf{t}) = \mathbf{f}\big(\boldsymbol{\phi}(\mathbf{t})\big)\cdot\Big(\frac{\partial\mathbf{x}}{\partial t_1}\times\frac{\partial\mathbf{x}}{\partial t_2}\Big)dt_1\wedge dt_2$$

in terms of the vector fields

$$\frac{\partial\mathbf{x}}{\partial t_j} = \left\{\frac{\partial x_1}{\partial t_j}, \frac{\partial x_2}{\partial t_j}, \frac{\partial x_3}{\partial t_j}\right\} \quad \text{for } j=1,2.$$

Thus the integral of the differential 2-form $\Omega_{\mathbf{f}}$ on the singular 2-cell $\boldsymbol{\phi}$ is given by

Vintcx (6.15)
$$\int_{\boldsymbol{\phi}}\Omega_{\mathbf{f}} = \int_{\Delta}\mathbf{f}\big(\boldsymbol{\phi}(t)\big)\cdot\Big(\frac{\partial\mathbf{x}}{\partial t_1}\times\frac{\partial\mathbf{x}}{\partial t_2}\Big)dt_1\wedge dt_2;$$

and if $\boldsymbol{\phi}$ is a simple 2-cell with image $\boldsymbol{\phi}(\overline{\Delta}) = \Gamma\subset U$ this can be viewed as the integral $\int_\Gamma\Omega_{\mathbf{f}}$. In either case the value of this integral depends on both the orientation of the space $\mathbb{R}^3$, which determines the order of the columns in (6.13), and the orientation of the space $\mathbb{R}^2$, which determines the order of the rows in (6.13). If the two vectors $\partial\mathbf{x}/\partial t_1$ and $\partial\mathbf{x}/\partial t_2$ are linearly independent at a point $\mathbf{t}\in\Delta$ then rank $\boldsymbol{\phi}'(\mathbf{t}) = 2$ and it follows from the rank theorem that the image of an open neighborhood of the point $\mathbf{t}\in\Delta$ is a 2-dimensional submanifold of U; but of course this may not be all of the image Γ near the point $\boldsymbol{\phi}(\mathbf{t}) \in U$ if the 2-cell is not a simple 2-cell. The three vectors

$$\frac{\partial \mathbf{x}}{\partial t_1}, \quad \frac{\partial \mathbf{x}}{\partial t_2}, \quad \frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2}$$

have the orientation compatible with the chosen orientation of $\mathbb{R}^3$, described by the order $dx_1, \ dx_2, \ dx_3$ of the basis vectors along the coordinate axes. The unit vector in the direction of the cross-product $\dfrac{\partial \mathbf{x}}{\partial t_1} \times \dfrac{\partial \mathbf{x}}{\partial t_2}$ is the **unit normal vector** to that oriented submanifold, which is denoted by $\boldsymbol{\nu}(\mathbf{t})$; and in these terms

Vintcky (6.16)
$$\frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} = \Big\| \frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} \Big\|_2 \boldsymbol{\nu}(\mathbf{t}).$$

In parallel to the interpretation of line integrals in (6.12), the integral (6.15) can be written

Vintckz (6.17)
$$\int_\Gamma \Omega_{\mathbf{f}} = \int_\Gamma \mathbf{f} \cdot \boldsymbol{\nu} \ dS$$

and can be viewed as the integral over the surface Γ of the dot product of the vector field $\mathbf{f}$ and the unit normal vector $\boldsymbol{\nu}$ to Γ, where dS indicates integration with respect to the **surface area** of Γ where the surface area $|\Gamma|$ is defined by[1]

Vintcksa (6.18)
$$\Big|\Gamma\Big| = \int_\Delta \Big\| \frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} \Big\|_2.$$

The calculation of the surface area can be a rather complicated one, involving explicit integrals of the form

Varcl (6.19)
$$|\Gamma| = \int_\Delta \sqrt{\det \begin{pmatrix} \frac{\partial x_2}{\partial t_1} & \frac{\partial x_3}{\partial t_1} \\ \frac{\partial x_2}{\partial t_2} & \frac{\partial x_3}{\partial t_2} \end{pmatrix}^2 + \det \begin{pmatrix} \frac{\partial x_3}{\partial t_1} & \frac{\partial x_1}{\partial t_1} \\ \frac{\partial x_3}{\partial t_2} & \frac{\partial x_1}{\partial t_2} \end{pmatrix}^2 + \det \begin{pmatrix} \frac{\partial x_1}{\partial t_1} & \frac{\partial x_2}{\partial t_1} \\ \frac{\partial x_1}{\partial t_2} & \frac{\partial x_2}{\partial t_2} \end{pmatrix}^2}.$$

As in the case of line integrals it is generally much much simpler to calculate the integral (6.15) as the integral of the differential form $\Omega_{\mathbf{f}}$ in terms of any parametrization of the singular cell Γ, without calculating the surface area.

[1] The direct definition of surface area is an extremely complicated matter, as discussed in detail for instance in the classical book by T. Radoó *Length and Area*, American Mathematical Society Colloquium Publications, New York, 1948. Those difficulties are finessed by using the pattern of the calculation of arc length through line integrals to define surface area, reversing the order of the discussion in Section 5.19. It is customary to attempt to justify this definition of surface area by observing that the length $\Big\| \frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} \Big\|_2$ of the cross product of the two vectors $\frac{\partial \mathbf{x}}{\partial t_1}$ and $\frac{\partial \mathbf{x}}{\partial t_2}$ is the area of the parallelogram spanned by these two vectors, which can be viewed as a planar approximation to the surface since the vectors are tangent vectors approximating the lengths of the two curves that are the image of the t_1 and t_2 coordinate axes under the mapping $\boldsymbol{\phi} : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$. What is perhaps more convincing though is that the integral $\int_\Gamma dS$ for some standard surfaces yields the usual values for surface areas.

For example the cylindrical surface $\Gamma \subset \mathbb{R}^3$ defined by $x_1^2 + x_2^2 = 1,\ 0 \le x_3 \le 1$, can be described as the image of the singular 2-cell $\boldsymbol{\phi} : \overline{\Delta}_2 \longrightarrow \Gamma$ given by $\boldsymbol{\phi}(\mathbf{t}) = \{\cos t_1,\ \sin t_1,\ t_2\}$, where the orientation of the cell is described by the differential form $dt_1 \wedge dt_2$. At the point $\mathbf{0} = \{0,0,1\} = \boldsymbol{\phi}\{0,0\}$ on the boundary of the surface Γ

$$\begin{aligned}\frac{\partial \mathbf{x}}{\partial t_1} &= \{-\sin t_1,\ \cos t_1,\ 0\}\Big|_{t_1=t_2=0} = \{0 \quad 1 \quad 0\},\\ \frac{\partial \mathbf{x}}{\partial t_2} &= \{0 \quad 0 \quad 1\}\end{aligned}$$

hence

$$\frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} = \{1,\ 0,\ 0\},$$

which points outward from the interior of the cylinder. (For the opposite orientation of the cylinder the orientation of the unit cell $\overline{\Delta}_2$ would be described by the differential form $dt_2 \wedge dt_1$.) The differential 2-form

$$\Omega_{\mathbf{f}} = x_2\ dx_2 \wedge dx_3 + x_1 dx_3 \wedge dx_1 + x_3 dx_1 \wedge dx_2,$$

is that described by the vector field $\mathbf{f}$ where $\mathbf{f}(\mathbf{x}) = \{x_1, x_2, x_3\}$, and its integral is

$$\begin{aligned}\int_\Gamma \mathbf{f} \cdot \boldsymbol{\nu}\ dS = \int_\Gamma \Omega_{\mathbf{f}} &= \int_\Delta x_2\ dx_2 \wedge dx_3 + x_1 dx_3 \wedge dx_1 + x_3 dx_1 \wedge dx_2\\ &= \det\begin{pmatrix} \cos t_1 & \sin t_1 & t_2 \\ -\sin t_1 & \cos t_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} dt_1 \wedge dt_2\\ &= \int_{t_1 \in [0,2\pi]} \int_{t_2 \in [0,1]} dt_2 dt_1 = 2\pi.\end{aligned}$$

Furthermore since

$$\frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} = \cos t_1 dx_1 + \sin t_1 dx_2 + 0 dx_3$$

when expressed in terms of the basis vectors $dx_1,\ dx_2,\ dx_3$ it follows that

$$\left\| \frac{\partial \mathbf{x}}{\partial t_1} \times \frac{\partial \mathbf{x}}{\partial t_2} \right\|_2 = 1$$

so the area of the cylinder Γ is

$$|\Gamma| = \int_\Gamma 1 dS = \int_\Delta 1 = 2\pi.$$

That is of course a familiar result, and it should be somewhat comforting that the general formula does work in this case.

6.2 Stokes's Theorem

stokes

The boundary $\partial\Delta$ of an ordinary n-cell Δ consists of a collection of singular $(n-1)$-cells; to discuss this boundary in more detail it is convenient to introduce some further notation. Denote the unit cell in the space $\mathbb{R}^n$ by

nwq1 (6.20)
$$\Delta(x_1, x_2, \ldots, x_n) = \Big\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \Big| \ 0 < x_i < 1 \quad \text{for } 1 \le i \le n\Big\},$$

with the orientation described by the differential form $dx_1 \wedge dx_2 \wedge \ldots \wedge dx_n$; and correspondingly denote the closed unit cell by

nwq1c (6.21)
$$\overline{\Delta}(x_1, x_2, \ldots, x_n) = \Big\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \Big| \ 0 \le x_i \le 1 \quad \text{for } 1 \le i \le n\Big\},$$

also with the orientation described by the differential form $dx_1 \wedge dx_2 \wedge \ldots \wedge dx_n$. The boundary of either cell consists of those points of the closed cell for which at least one of the coordinates takes the value 0 or 1; for instance when $n = 2$ the boundary of the cell $\Delta(x_1, x_2)$ consists of the 4 segments

stkxq2 (6.22)
$$\begin{aligned}
\overline{\Delta}(0, x_2) \in \mathbb{R}^2 &= \Big\{ \ (0, x_2) \ \Big| \ 0 \le x_2 \le 1 \ \Big\}\\
\overline{\Delta}(1, x_2) \in \mathbb{R}^2 &= \Big\{ \ (1, x_2) \ \Big| \ 0 \le x_2 \le 1 \ \Big\}\\
\overline{\Delta}(x_1, 0) \in \mathbb{R}^2 &= \Big\{ \ (x_1, 0) \ \Big| \ 0 \le x_1 \le 1 \ \Big\}\\
\overline{\Delta}(x_1, 1) \in \mathbb{R}^2 &= \Big\{ \ (x_1, 1) \ \Big| \ 0 \le x_1 \le 1 \ \Big\}.
\end{aligned}$$

The segment $\overline{\Delta}(0, x_2)$ can be viewed as a singular 1-cell, the image of the closed unit 1-cell $\overline{\Delta}(x_2) \subset \mathbb{R}^1$ under the mapping that takes a point $x_2 \in \overline{\Delta}(x_2)$ to the point $(0, x_2) \in \overline{\Delta}(x_1, x_2)$, and correspondingly for the other three segments. With the orientations of these singular cells defined by the natural orientation of the 1-cells $\Delta(x_i)$, the collection of these four singular cells is sketched in the accompanying Figure bndnew 6.1. However the more natural and more traditional expression for the boundary of the cell $\Delta(x_1, x_2)$ consists of these 4 singular 1-cells but with the orientations changed so that the path described by the collection of these singular cells traverses the entire boundary of the cell $\Delta(x_1, x_2)$ in the counterclockwise direction; that involves reversing the orientations of the singular cells $\Delta(x_1, 1)$ and $\Delta(0, x_2)$. It is convenient to denote these singular cells with the orientation reversed by $-\Delta(x_1, 1)$ and $-\Delta(x_1, 1)$, and to view the collection of all four of these cells, with the chosen orientations, as a formal sum

newbnd1 (6.23)
$$\partial\Delta(x_1, x_2) = \Delta(x_1, 0) + \Delta(1, x_2) - \Delta(x_1, 1) - \Delta(0, x_2)$$

of singular 1-cells in $\mathbb{R}^2$. In general, a **singular n-chain** in an open subset $U \subset \mathbb{R}^m$ is defined to be a finite formal sum

dfinq2 (6.24)
$$\mathcal{C} = \sum_{j=1}^{r} \nu_j \cdot \boldsymbol{\phi}_j \quad \text{for integers } \nu_j \in \mathbb{Z} \text{ and singular } n\text{-cells } \boldsymbol{\phi}_j \text{ in } U;$$

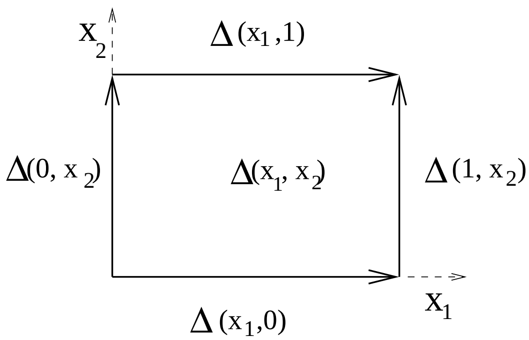

Figure 6.1: The singular cells forming the boundary of a cell $\Delta(x_1, x_2) \subset \mathbb{R}^2$. bndnew

and the boundary of the cell $\Delta(x_1, x_2, \ldots, x_n)$ is defined to be the singular $(n-1)$-chain

newbnd2 (6.25)
$$\partial\Delta(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \sum_{\epsilon=0}^{1} (-1)^{i+\epsilon} \overline{\Delta}(x_1, \ldots, x_{i-1}, \epsilon, x_{i+2}, \ldots, x_n).$$

The notation $\partial\Delta$ is the same as that used for the point set boundary of Δ; generally it should be clear from the context whether $\partial\Delta$ is to be viewed as a point set or as a singular chain, but if necessary for clarity the set will be called either the point set boundary or the boundary chain. In particular the boundary of the 1-cell $\Delta(x_1)$ is the singular 0-cell

newbnd1a (6.26)
$$\partial\Delta(x_1) = \Delta(1) - \Delta(0),$$

in which the point or 0-cell $\Delta(1)$ at the head of the 1-cell has the positive sign and the point or 0-cell $\Delta(0)$ at the tail of the 1-cell has the negative sign; the boundary of the 2-cell $\Delta(x_1, x_2)$ is the singular 1-chain (6.23); and the boundary of the 3-cell $\Delta(x_1, x_2, x_3)$ is the singular 2-chain

stkxq9 (6.27)
$$\begin{aligned}\partial\Delta(x_1, x_2, x_3) = {} & -\overline{\Delta}(0, x_2, x_3) + \overline{\Delta}(1, x_2, x_3) + \overline{\Delta}(x_1, 0, x_3) - \overline{\Delta}(x_1, 1, x_3) \\ & - \overline{\Delta}(x_1, x_2, 0) + \overline{\Delta}(x_1, x_2, 1).\end{aligned}$$

The geometric significance of the orientation of the boundary $\partial\overline{\Delta}(x_1, x_2, x_3)$ as just defined may not be all that clear, and perhaps merits a bit more discussion, for which Figure 6.2 may be helpful. The boundary segment $\overline{\Delta}(x_1, x_2, 1)$ is the top side while the boundary segment $\overline{\Delta}(x_1, x_2, 0)$ is the bottom side of the cell $\Delta(x_1, x_2, x_3)$. By (6.27) the orientation of the top side is described by the differential form $dx_1 \wedge dx_2$ while the orientation of the bottom side is described by the differential form $-dx_1 \wedge dx_2 = dx_2 \wedge dx_1$. If $\mathbf{e}_i$ denotes the unit vector in the direction of the x_i-axis in $\mathbb{R}^3$, the orientation of the top side also can be

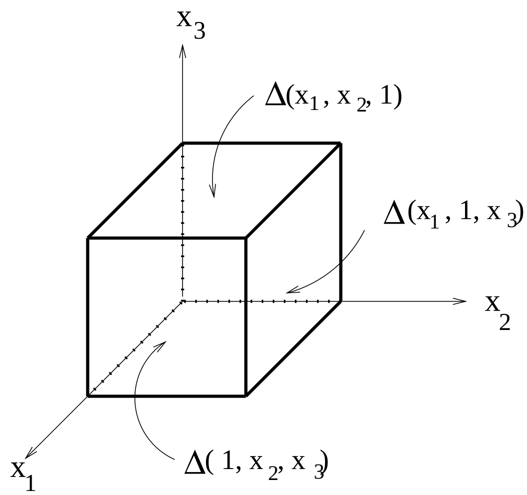

Figure 6.2: The boundary of a cell $\Delta(x_1, x_2, x_3) \subset \mathbb{R}^3$. 3cellnew

described by the ordered pair of vectors $\mathbf{e}_1, \mathbf{e}_2$, or equivalently by their cross-product $\mathbf{e}_1 \wedge \mathbf{e}_2 = \mathbf{e}_3$, a vector perpendicular to the top boundary segment and in the direction pointing outward from the 3-cell $\Delta(x_1, x_2, x_3)$; and the orientation of the bottom side is described by the ordered pair of vectors $\mathbf{e}_2, \mathbf{e}_1$, or equivalently by their cross-product $\mathbf{e}_2 \wedge \mathbf{e}_1 = -\mathbf{e}_3$, a vector perpendicular to the bottom boundary segment and in the direction pointing outward from the 3-cell $\Delta(x_1, x_2, x_3)$. The boundary segment $\overline{\Delta}(1, x_2, x_3)$ is the front side while the boundary segment $\overline{\Delta}(0, x_2, x_3)$ is the back side of the cell $\Delta(x_1, x_2, x_3)$. The orientation of the front side is described by the differential form $dx_2 \wedge dx_3$ or the cross-product $\mathbf{e}_2 \times \mathbf{e}_3 = \mathbf{e}_1$, a vector perpendicular to that side and pointing outward, while the orientation of the back side is described by the differential form $-dx_2 \wedge dx_3$ or the cross-product $\mathbf{e}_3 \times \mathbf{e}_2 = -\mathbf{e}_1$, again a vector perpendicular to that side and pointing outward. The boundary segment $\overline{\Delta}(x_1, 1, x_3)$ is the right-hand side while the boundary segment $\overline{\Delta}(x_1, 0, x_3)$ is the left-hand side of the cell $\Delta(x_1, x_2, x_3)$. The orientation of the right-hand side is described by the differential form $-dx_1 \wedge dx_3$ or the cross-product $\mathbf{e}_3 \times \mathbf{e}_1 = \mathbf{e}_2$, a vector perpendicular to that side and pointing outward, while the orientation of the left-hand side is described by the differential form $dx_1 \wedge dx_3$ or the cross-product $\mathbf{e}_1 \times \mathbf{e}_3 = -\mathbf{e}_2$, yet again a vector perpendicular to that side and pointing outward. Thus the orientations of the 6 faces of the 3-cell $\Delta(x_1, x_2, x_3)$ are all described by the outward pointing normal vectors to these sides, an easy way to remember the orientation and to use it in explicit calculations.

The notion of the boundary chain of a standard n-cell $\Delta \subset \mathbb{R}^n$ can be extended quite simply to singular n-cells. A singular n-cell $\boldsymbol{\phi}$ is a continuous mapping $\boldsymbol{\phi} : \Delta \longrightarrow \mathbb{R}^n$, and its boundary is defined by

newbndsg (6.28)
$$\partial\boldsymbol{\phi} = \boldsymbol{\phi}\big(\partial\Delta(x_1, x_2, \ldots, x_n)\big) = \sum_{i=1}^{n}\sum_{\epsilon=0}^{1}(-1)^{i+\epsilon}\boldsymbol{\phi}\big(\overline{\Delta}(x_1, \ldots, x_{i-1}, \epsilon, x_{i+1}, \ldots, x_n)\big).$$

In this definition the singular cell $\overline{\Delta}(x_1, \ldots, x_{n-1}, 0)$ is the mapping from the standard $(n-1)$-cell $\overline{\Delta}(x_1, \ldots, x_{n-1}) \subset \mathbb{R}^{n-1}$ into $\mathbb{R}^n$ that takes a point $(x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}$ to the point $(x_1, \ldots, x_{n-1}, 0) \in \mathbb{R}^n$, so $\boldsymbol{\phi}\big(\overline{\Delta}(x_1, \ldots, x_{n-1}, 0)\big)$ is just the singular $(n-1)$-cell resulting from the composition of these two mappings; the other terms in (6.28) can be interpreted correspondingly, as is illustrated in Figure 6.3. A singular n-chain $\mathcal{C}$ is said to be a **cycle** or an n**-cycle** if $\partial\mathcal{C} = 0$,

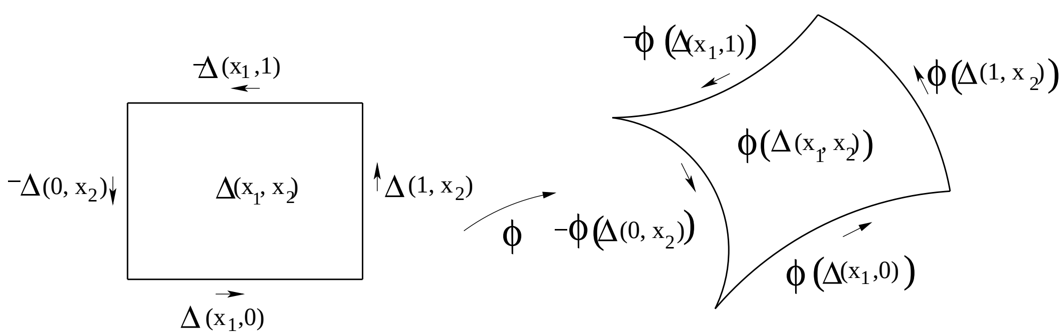

Figure 6.3: The boundary of a singular cell $\Phi\Big(\Delta(x_1, x_2)\Big) \subset \mathbb{R}^3$, where the mapping $\boldsymbol{\phi}$ is orientation preserving. newsgbnd

and it is said to be a **boundary** if $\mathcal{C} = \partial\mathcal{C}_0$ for some singular $(n-1)$ chain $\mathcal{C}_0$. For example, in Figure 6.3 the boundary chain of the cell $\Delta(x_1, x_2)$ is the boundary 1-chain $\partial\Delta(x_1, x_2) = \Delta(x_1, 0) + \Delta(1, x_2) - \Delta(x_1, 1) - \Delta(0, x_2)$. Each vertex $(\epsilon_1, \epsilon_2) \in \mathbb{R}^2$ of the cell $\Delta(x_1, x_2)$ is part of the boundary of two of these singular cells in the boundary chain $\partial\Delta(x_1, x_2)$, at the head of one and the tail of the other; so altogether the boundary of the boundary 1-chain $\partial\Delta(x_1, x_2)$ is the trivial 0-chain

$$\begin{aligned}\partial\partial\Delta(x_1, x_2) &= \partial\Delta(x_1, 0) + \partial\Delta(1, x_2) - \partial\Delta(x_1, 1) - \partial\Delta(0, x_2)\\ &= \big((1,0) - (0,0)\big) + \big((1,1) - (1,0)\big) + \big((0,1) - (1,1)\big) + \big((0,0) - (0,1)\big)\\ &= 0,\end{aligned}$$

hence the boundary chain $\partial\Delta(x_1, x_2)$ is a cycle. That is indeed the general situation.

dfinq9 **Theorem 6.2** *The boundary chain of a singular chain is a cycle, or equivalently* $\partial(\partial\mathcal{C}) = 0$ *for any singular chain* $\mathcal{C}$*.*

Proof: The boundary of the standard cell $\overline{\Delta}(x_1, x_2, \ldots, x_n)$ was defined in (6.25) and the boundary of a chain was defined in (6.28), and from these definitions it follows that

newbnditq1

$$(6.29)\quad \partial\partial\overline{\Delta}(x_1,\ldots,x_n) = \partial\left(\sum_{i=1}^{n}\sum_{\epsilon=0}^{1}(-1)^{i+\epsilon}\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n)\right)$$
$$= \sum_{i=1}^{n}\sum_{\epsilon=0}^{1}(-1)^{i+\epsilon}\partial\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n).$$

Here $\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n)$ is a singular $(n-1)$-cell, the image of the standard $(n-1)$-cell $\overline{\Delta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) \in \mathbb{R}^{n-1}$ by the mapping

$$\boldsymbol{\phi} : \overline{\Delta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) \longrightarrow \overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n)$$

for which

$$\boldsymbol{\phi}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = (x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n) \in \mathbb{R}^n;$$

it then follows from the definition (6.25) of the boundary of a cell and the definition (6.28) of the boundary of a singular cell that

$$\partial\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n) = \partial\boldsymbol{\phi}\overline{\Delta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$$
$$= \boldsymbol{\phi}\partial\overline{\Delta}(x_1,\ldots,x_{i-1}x_{i+1},\ldots,x_n)$$
$$= \boldsymbol{\phi}\Bigg(\sum_{j=1}^{i-1}\sum_{\delta=0}^{1}(-1)^{j+\delta}\overline{\Delta}(x_1,\ldots,x_{j-1},\delta,x_{j+1},\ldots x_{i-1},x_{i+1},\ldots,x_n)$$
$$+\sum_{j=i+1}^{n}\sum_{\delta=0}^{1}(-1)^{j-1+\delta}\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots x_{j-1},\delta,x_{j+1},\ldots,x_n)\Bigg)$$
$$= \sum_{j=1}^{i-1}\sum_{\delta=0}^{1}(-1)^{j+\delta}\overline{\Delta}(x_1,\ldots,x_{j-1},\delta,x_{j+1},\ldots x_{i-1},\epsilon,x_{i+1},\ldots,x_n)$$
$$+\sum_{j=i+1}^{n}\sum_{\delta=0}^{1}(-1)^{j-1+\delta}\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots x_{j-1},\delta,x_{j+1},\ldots,x_n).$$

The sign change in the preceding equation is a consequence of the indexing of the cell $\overline{\Delta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ since x_j is the j-th variable in $\overline{\Delta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ if $j < i$ but is the $(j-1)$-th variable if $j > i$; and $\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots x_{j-1},\delta,x_{j+1},\ldots,x_n)$ is the obvious singular $(n-2)$-cell with the natural extension of the definition of the singular $(n-1)$-cell $\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots,x_n)$. Substituting the preceding observation into (6.29) yields

$$\begin{aligned}
&\partial\partial\overline{\Delta}(x_1,\ldots,x_n)\\
&\quad=\sum_{\substack{i,j=1\\ j<i}}^{n}\sum_{\epsilon,\delta=0}^{1}(-1)^{i+j+\epsilon+\delta}\overline{\Delta}(x_1,\ldots,x_{j-1},\delta,x_{j+1},\ldots x_{i-1},\epsilon,x_{i+1},\ldots,x_n)\\
&\quad+\sum_{\substack{i,j=1\\ j>i}}^{n}\sum_{\epsilon,\delta=0}^{1}(-1)^{i+j-1+\epsilon+\delta}\overline{\Delta}(x_1,\ldots,x_{i-1},\epsilon,x_{i+1},\ldots x_{j-1},\delta,x_{j+1},\ldots,x_n).
\end{aligned}$$

If the indices i, j are interchanged and the indices ϵ, δ are simultaneously interchanged in the first line on the right-hand side of the preceding equation the result is evidently equal to the second line in that equation, except for the sign, so that $\partial\partial\Delta(x_1,\ldots,x_n) = 0$, which suffices for the proof.

ssch

The boundary chain of the standard n-cell is a particularly simple singular $(n-1)$-chain; indeed it is an example of a **simple singular chain**, defined to be a singular n-chain $\mathcal{C} = \sum_{j=1}^{r} \nu_j \cdot \boldsymbol{\phi}_j$ in $\mathbb{R}^n$ where all the singular cells $\boldsymbol{\phi}_j$ are simple singular cells for which the mappings are orientation preserving and their images are Jordan domains that are disjoint aside for their boundaries. Simple singular chains are useful in describing and working with more general geometric objects. For example, the annular region in Figure 6.4, the region between two concentric circles, can be described as the image of the simple singular 2-chain $\mathcal{C} = \boldsymbol{\phi}_1 + \boldsymbol{\phi}_2$, where the images of the 2-cells Δ_1 and Δ_2 are the sets Γ_1 and Γ_2 that cover the annular region and that intersect only in the boundary segments at the top and bottom of the annular region. When the cells Δ_i have the standard orientation, so that their boundaries are traversed in the counter-clockwise order, and the mappings $\boldsymbol{\phi}_i$ are orientation preserving, the boundary of the singular 2-chain reduces to the outer circle of the annular region, traversed in the counter-clockwise direction, and the inner circle of the annular region, traversed in the clockwise direction, since the two boundary segments at the top and bottom of the annular region are traversed in opposite directions so cancel in the full boundary of the annular region. Of course there are useful singular chains that are not simple, such as the boundary chains of singular cells that are not simple singular cells.

The set of singular n-chains in an open subset $U \subset \mathbb{R}^m$ form an abelian group $C_n(U,\mathbb{Z})$, a set with an addition operation for which the sum of two singular chains $\mathcal{C}' = \sum_i \nu_i' \boldsymbol{\phi}_i'$ and $\mathcal{C}'' = \sum_j \nu_j'' \boldsymbol{\phi}_j''$ is defined to be the singular chain $\mathcal{C}' + \mathcal{C}'' = \sum_i \nu_i' \boldsymbol{\phi}_i' + \sum_j \nu_j'' \boldsymbol{\phi}_j''$. It is clear from the definition of the boundary chain that $\partial\big(\mathcal{C}'+\mathcal{C}''\big) = \partial\mathcal{C}'+\partial\mathcal{C}''$, so the set of singular n-cycles form a subgroup $Z_n(U,\mathbb{Z}) \subset C_n(U,\mathbb{Z})$, a subset closed under the operation of addition of singular chains; and it follows from Theorem 6.2 that the set of singular n-boundaries form a further subgroup $B_n(U,\mathbb{Z}) \subset Z_n(U,\mathbb{Z})$. The quotient group, defined in analogy with the quotient of two vector spaces, is a group

dfhom (6.30)
$$H_n(U,\mathbb{Z}) = \frac{Z_n(U,\mathbb{Z})}{B_n(U,\mathbb{Z})}$$

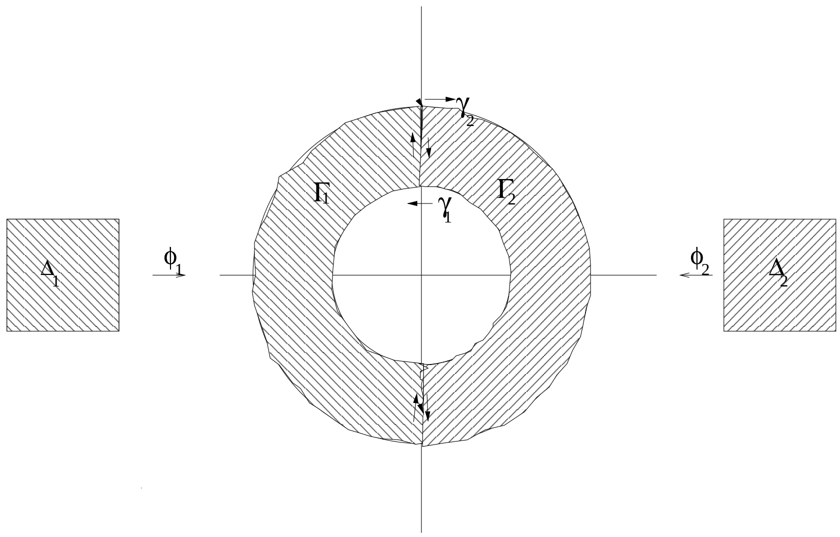

Figure 6.4: An annular region viewed as the image of a singular cell $\mathcal{C} = \boldsymbol{\phi}_1 + \boldsymbol{\phi}_2$ grnex

called the n-dimensional **singular homology group** of the set U. These groups can be defined for more general topological spaces than just subsets of $\mathbb{R}^m$; they are algebraic structures that reflect the geometrical properties of topological spaces, and some examples of their usefulness will be included in the subsequent discussion here[2].

The integral of a differential n-form ω over a singular n-chain is defined by

dfinq2a (6.31)
$$\int_{\mathcal{C}} \omega = \sum_{j=1}^{r} \nu_j \int_{\boldsymbol{\phi}_j} \omega \quad \text{for the chain } \mathcal{C} = \textstyle\sum_{j=1}^{r} \nu_j \cdot \boldsymbol{\phi}_j.$$

For example, twice the integral of a differential form ω over a singular cell $\boldsymbol{\phi}$ can be viewed alternatively as the integral of that differential form over the singular chain $\mathcal{C} = 2 \cdot \boldsymbol{\phi}$; and the negative of the integral of ω over $\boldsymbol{\phi}$, or what is the same thing the integral of ω over $\boldsymbol{\phi}$ with the opposite orientation, can be viewed alternatively as the integral of that differential form over the singular chain $\mathcal{C} = -1 \cdot \boldsymbol{\phi} = -\boldsymbol{\phi}$. The fundamental theorem of calculus for functions of a single variable extends to functions of several variables in the form of Stokes's Theorem, which is expressed most naturally in terms of the integrals of differential forms over singular chains. The proof really is just a straightforward application of the classical fundamental theorem of calculus in a single variable, which can be established readily in the following special case.

[2]Homology groups are discussed in detail in most books on algebraic topology, as for instance in the book by Allen Hatcher *Algebraic Topology*, (Cambridge University Press, 2002).

sto1 **Theorem 6.3** *If ω is a $\mathcal{C}^\infty$ differential form of degree $n-1$ defined in an open neighborhood of the cell $\overline{\Delta}(x_1, x_2, \ldots, x_n) \subset \mathbb{R}^m$ then*

stoqa (6.32)
$$\int_{\partial\Delta(x_1,x_2,\ldots,x_n)} \omega = \int_{\Delta(x_1,x_2,\ldots,x_n)} d\omega.$$

Proof: First for a fixed index r introduce for convenience but just for the course of this proof the simplifying notation

$$\begin{aligned}
&\Delta = \Delta(x_1, x_2, \ldots, x_n) \subset \mathbb{R}^n,\\
&\Delta' = \Delta(x_1, \ldots, x_{r-1}, x_{r+1}, \ldots, x_n) \subset \mathbb{R}^{n-1}\\
&\Delta_\epsilon = \Delta(x_1, \ldots, x_{r-1}, \epsilon, x_{r+1}, \ldots, x_n) \subset \mathbb{R}^n\\
&dX_r = dx_1 \wedge \cdots \wedge dx_{r-1} \wedge dx_{r+1} \wedge \cdots \wedge dx_n,\\
&dX = dx_1 \wedge dx_2 \wedge \cdots \wedge dx_n.
\end{aligned}$$

Consider then a differential $(n-1)$-form given by $\omega_r = f(\mathbf{x})\ dX_r$ for some function $f(\mathbf{x})$. For this differential form

sto1q1 (6.33)
$$d\omega_r = \partial_r f(\mathbf{x})\ dx_r \wedge dX_r = (-1)^{r-1}\partial_r f(\mathbf{x})\ dX.$$

Furthermore on a segment $\overline{\Delta}(x_1, \ldots, x_{i-1}, \epsilon, x_{i+1}, \ldots, x_n)$ of the boundary $\partial\Delta$ the variable x_i is constant, so if $i \neq r$ the differential form $\omega_r = f(\mathbf{x})\ dX_r$ vanishes identically; therefore the integral of the differential form ω_r over the boundary $\partial\Delta$ reduces to the integral just over the singular $(n-1)$-cells $(-1)^{r+1}\Delta_1$ and $(-1)^r\Delta_0$ of the boundary. It then follows from the definition of the integral of a differential form over a singular cell, the fundamental theorem of calculus in one variable and Fubini's Theorem that

$$\begin{aligned}
\int_{\partial\Delta} \omega_r &= (-1)^{r+1}\int_{\Delta_1} \omega_r + (-1)^r \int_{\Delta_0} \omega_r\\
&= (-1)^{r+1}\int_{\Delta'} \Big(f(x_1, \ldots, x_{r-1}, 1, x_{r+1}, \ldots, x_n)\\
&\qquad\qquad - f(x_1, \ldots, x_{r-1}, 0, x_{r+1}, \ldots, x_n) \Big) dX_r\\
&= (-1)^{r+1}\int_{\Delta'}\int_{\Delta(x_r)} \partial_r f(\mathbf{x})\ dx_r \wedge dX_r\\
&= \int_\Delta d\omega_r,
\end{aligned}$$

which is (6.32) for the differential form ω_r. Since any differential form ω can be written as a sum of differential forms of the special type ω_r for $r = 1, 2, \cdots, n$ and (6.32) is linear in the differential form ω it follows that (6.32) holds for arbitrary differential forms ω, which suffices for the proof.

sto2 **Corollary 6.1 (Stokes's Theorem)** *If $\mathcal{C} \subset U$ is a $\mathcal{C}^\infty$ n-dimensional singular chain in an open subset $U \subset \mathbb{R}^m$ and ω is a $\mathcal{C}^\infty$ differential $(n-1)$-form in U then*

sto2q1 (6.34)
$$\int_{\partial\mathcal{C}} \omega = \int_{\mathcal{C}} d\omega.$$

Proof: First for the special case of a singular n-cell $\boldsymbol{\phi} : \Delta \longrightarrow U$ in an open subset $U \subset \mathbb{R}^m$, if ω is an arbitrary $\mathcal{C}^1$ differential $(n-1)$-form ω in U then since $\partial\boldsymbol{\phi} = \boldsymbol{\phi}\partial\Delta$ by (6.28) it follows from the definition (6.31) of the integral of a differential form over a singular chain, the definition (6.9) of the integral of a differential form over a singular cell, and the preceding Theorem 6.3 that

$$\int_{\partial\boldsymbol{\phi}} \omega = \int_{\partial\overline{\Delta}} \boldsymbol{\phi}^*(\omega) = \int_{\boldsymbol{\phi}(\Delta} \omega = \int_{\overline{\Delta}} d\boldsymbol{\phi}^*(\omega) = \int_{\overline{\Delta}} \boldsymbol{\phi}^*(d\omega) = \int_{\boldsymbol{\phi}} d\omega,$$

since $d\boldsymbol{\phi}^*(\omega) = \boldsymbol{\phi}^*(d\omega)$ by Theorem 5.8; and that is (6.34) for the case of a singular cell. Then if $\mathcal{C} = \sum_i \nu_i \cdot \boldsymbol{\phi}_i$ is an arbitrary singular chain where Γ_i are singular cells in U and if ω is a $\mathcal{C}^\infty$ differential $(n-1)$-form in U it follows from what has just been demonstrated and the definition (6.31) of the integral over a singular chain that

$$\int_{\partial\mathcal{C}} \omega = \sum_i \nu_i \int_{\partial\boldsymbol{\phi}_i} \omega = \sum_i \nu_i \int_{\boldsymbol{\phi}_i} d\omega = \int_{\sum_i \nu_i \cdot \boldsymbol{\phi}_i} d\omega = \int_{\mathcal{C}} d\omega,$$

and that suffices for the proof.

Examples of the application of Stokes's Theorem abound. It can be used for instance to complement the discussion of conservative vector fields in Section 5.1. It was noted there that any closed 1-form is locally the exterior derivative of a function, indeed is the exterior derivative of a function in any cell but not necessarily so in more general regions; but it is the exterior derivative of a function in a class of regions that can be described topologically.

stokcr1 **Corollary 6.2** [3] *If $U \subset \mathbb{R}^n$ is an open subset such that $H_1(U, \mathbb{Z}) = 0$ then any closed $\mathcal{C}^\infty$ differential 1-form in U is the exterior derivative of a function in U.*

Proof: It follows from Theorem 5.4 that it is sufficient to show that the integral $\int_\gamma \omega$ of any closed 1-form ω on curves γ between two points $\mathbf{a}, \mathbf{b} \in U$ depends only on the points $\mathbf{a}, \mathbf{b}$ and not on the choice of the curve γ from $\mathbf{a}$ to $\mathbf{b}$. If γ_1, γ_2 are two curves from $\mathbf{a}$ to $\mathbf{b}$, viewed as singular 1-chains, then $\gamma = \gamma_1 - \gamma_2$ is a closed singular 1-chain U so as a consequence of the hypothesis it is the boundary $\gamma = \partial\mathcal{C}$ of a singular 2-chain $\mathcal{C} \subset U$; then by Stokes's Theorem

$$\int_{\gamma_1} \omega - \int_{\gamma_2} \omega = \int_\gamma \omega = \int_{\partial\mathcal{C}} \omega = \int_{\mathcal{C}} d\omega = \int_{\mathcal{C}} 0 = 0,$$

[3] For extensions of this corollary see for instance the discussion of De Rham's theorem in "From Calculus to Cohomology: De Rham Cohomology and Characteristic Classes" by Ib H. Madsen and Jurgen Tornehave, (Cambridge University Press, 1997), among other places.

which suffices for the proof.

A number of applications of Stokes's Theorem can be restated in more classical terms in the special cases of the simple singular n-chains in open subsets $U \subset \mathbb{R}^n$, which were discussed on page 147. The **image** of a simple singular chain $\mathcal{C} = \sum_i \boldsymbol{\phi}_i$ is the subset $|\mathcal{C}| = \bigcup_i \boldsymbol{\phi}_i(\overline{\Delta}_n) \subset U$; it is a union of Jordan domains $\boldsymbol{\phi}_i(\overline{\Delta}_n)$ that are disjoint aside from their boundaries.

Theorem 6.4 *If $D = |\mathcal{C}| \subset \mathbb{R}^n$ is the image of a $\mathcal{C}^\infty$ simple singular n-chain $\mathcal{C}$ in an open subset $U \subset \mathbb{R}^n$ and if*

$$\omega = \sum_{i=1}^{n} f_i dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n \tag{6.35}$$

is a $\mathcal{C}^\infty$ differential $(n-1)$-form in U then

$$\int_{\partial\mathcal{C}} \omega = \int_D \sum_{i=1}^{n} (-1)^{i-1} \frac{\partial f_i}{\partial x_i}. \tag{6.36}$$

Proof: By Stokes's Theorem $\int_{\partial\mathcal{C}} \omega = \int_{\mathcal{C}} d\omega$ for the differential form (6.35) and the simple singular n-chain $\mathcal{C}$. The integral of any differential n-form $\sigma = h(\mathbf{x})dx_1 \wedge \cdots \wedge dx_n$ over the simple singular n-chain $\mathcal{C} = \sum_i \boldsymbol{\phi}_i$ is $\int_{\mathcal{C}} \sigma = \sum_i \int_{\boldsymbol{\phi}_i} \sigma = \sum_i \int_{\overline{\Delta}_n} \boldsymbol{\phi}_i^*(\sigma)$. Since the $\mathcal{C}^1$ mappings $\boldsymbol{\phi}_i : \overline{\Delta}_m \longrightarrow \Gamma_i$ are orientation-preserving homeomorphisms by definition, it follows from Theorem 6.1 that $\int_{\overline{\Delta}_n} \boldsymbol{\phi}_i^*(\sigma) = \int_{\Gamma_i} \sigma$; and $\int_{\Gamma_i} \sigma = \int_{\Gamma_i} h$ as in (6.1). The Jordan domains Γ_i are disjoint, except for their boundaries, so from Theorem 4.6 it follows that $\sum_i \int_{\Gamma_i} h = \int_{|\mathcal{C}|} h = \int_D h$; altogether therefore $\int_{\mathcal{C}} \sigma = \int_D h$. In particular for the differential form $\sigma = d\omega$ in which $h = \sum_{i=1}^{n} (-1)^{i-1} \partial f_i / \partial x_i$ it follows that (6.36) holds as desired, to conclude the proof.

Corollary 6.3 (Green's Theorem) *If $D \subset U$ is the image of a simple singular 2-chain $\mathcal{C}$ in an open subset $U \subset \mathbb{R}^2$ then*

$$\int_{\partial\mathcal{C}} \Big(f_1 dx_1 + f_2 dx_2 \Big) = \int_D \left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2} \right) \tag{6.37}$$

for any $\mathcal{C}^1$ vector field $\mathbf{f} = \{f_1, f_2\}$ in U.

Proof: This is just the special case of the preceding theorem for the differential form $\omega_{\mathbf{f}}$ in $\mathbb{R}^2$, so no further proof is required.

Green's Theorem can be applied for example to the annular region D in Figure 6.4, which is exhibited as the image of the simple singular 2-chain $\mathcal{C} = \boldsymbol{\phi}_1 + \boldsymbol{\phi}_2$; the boundary $\partial\mathcal{C}$ consists of the outer circle γ_1 oriented in the counterclockwise direction and the inner circle γ_2 oriented in the clockwise direction. If $\omega = f_1 dx_1 + f_2 dx_2$ is a $\mathcal{C}^1$ differential 1-form in an open neighborhood of D then Green's Theorem asserts that $\int_{\gamma_1}\omega - \int_{\gamma_2}\omega = \int_D\left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}\right)$. In particular it follows that

stokcrg1
$$\int_{\gamma_1} x_1dx_2 - \int_{\gamma_2} x_1dx_2 = \int_D 1 = |D|, \tag{6.38}$$

where $|D|$ is the area of the annular region D. Of course the same formula can be applied to any set that is the image of a simple singular chain, showing that the area of any such set can be calculated from its boundary alone[4]. As a more interesting example of this technique consider the hypocycloid curve in Figure hypc 6.5, defined by the equation

$$x_1^{2/3} + x_2^{2/3} = a^{2/3}$$

for some constant $a > 0$. This curve can be viewed as the boundary of a simple singular chain $\mathcal{C}$ consisting of four simple singular 2-cells, the images of which are the segments in the four quadrants in Figure hypc 6.5. The boundary chain reduces to the union of the four curved arcs, traversed in the counterclockwise direction, so is the image of the singular 1-cell

$$\boldsymbol{\phi}(t) = \{a\,\cos^3 t, a\,\sin^3 t\} \quad \text{for} \quad 0 \le t \le 2\pi;$$

consequently the area $|D|$ of the region enclosed by the hypocycloid is

$$\begin{aligned}
|D| &= \int_{[0,2\pi]} (x_1dx_2 - x_2dx_1) \\
&= \int_{[0,2\pi]} \big(a\cos^3 t\cdot 3a\sin^2 t\cos t - a\sin^3 t\cdot 3a\cos^2 t\big(-\sin t\big)\big) \\
&= \int_{[0,2\pi]} 3a^2\sin^2 t\cos^2 t) = \int_{[0,2\pi]} \frac{3}{4}a^2\sin^2 2t \\
&= \int_{[0,2\pi]} \frac{3}{8}a^2(1-\cos 4t) = \frac{3}{8}\pi a^2.
\end{aligned}$$

Green's Theorem can be written in terms of the differential 1-form $\omega_{\mathbf{f}} = f_1dx_1 + f_2dx_2$ associated to a vector field $\mathbf{f}$ as the line integral

stokcrg2
$$\int_\gamma \mathbf{f}\cdot\boldsymbol{\tau}\,ds = \int_D\left(\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}\right) \tag{6.39}$$

where $\boldsymbol{\tau}$ is the unit tangent vector to γ and ds indicates integration with respect to the arc length of γ. Green's Theorem also can be written in terms of the

[4]This method for calculating the area of a region from its boundary is embodied in mechanical and electronic machines, called planimeters, that have long been used in engineering and surveying; these machines are discussed for example in the book by John Bryant and Christopher J. Sangwin *How round is your circle? Where engineering and mathematics meet.* Princeton University Press, 2008.

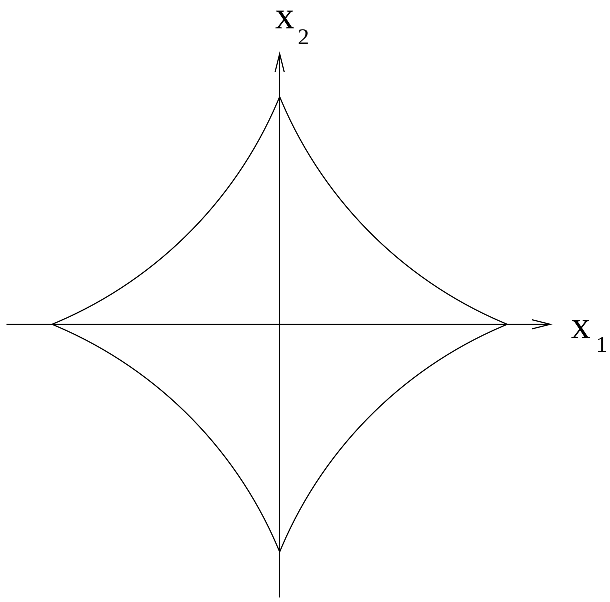

Figure 6.5: A hypocycloid hypc

Hodge dual $*\omega_{\mathbf{f}} = f_1 dx_2 - f_2 dx_1$ of the differential form $\omega_{\mathbf{f}}$ and since $d * \omega_{\mathbf{f}} = (\partial_1 f_1 + \partial_2 f_2) dx_1 \wedge dx_2$

stokcrg3ex (6.40)
$$\int_\gamma *\omega_{\mathbf{f}} = \int_D \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} \right).$$

If the curve γ is parametrized by a mapping $\boldsymbol{\phi} : [0,1] \longrightarrow U$ the integral of the differential form $\omega_{\mathbf{f}}$ is given by

stokcrg4 (6.41)
$$\int_\gamma *\omega_{\mathbf{f}} = \int_{t \in [0,1]} \left(f_1 \frac{dx_2}{dt} - f_2 \frac{dx_1}{dt} \right).$$

This can be viewed as the integral over the curve γ of the inner product $\mathbf{f} \cdot \mathbf{v}$ of the vector field $\mathbf{f}$ and the vector field

stokcrg4a (6.42)
$$\mathbf{v} = \left\{ \frac{dx_2}{dt}, \quad -\frac{dx_1}{dt} \right\};$$

and since $\|\mathbf{v}\|_2 = \sqrt{\left(\frac{dx_2}{dt}\right)^2 + \left(\frac{dx_1}{dt}\right)^2}$ then with the conventions adopted in the discussion of line integrals the integral (stokcrg4 6.41 can be rewritten in terms of the unit vector $\boldsymbol{\nu} = \mathbf{v}/\|\mathbf{v}\|_2$ as the line integral

stokcrg4ax (6.43)
$$\int_\gamma *\omega_{\mathbf{f}} = \int_\gamma \mathbf{f} \cdot \boldsymbol{\nu} \, ds$$

with respect to the arc length of the curve γ. The vector field $\boldsymbol{\nu}$ is orthogonal to the tangent vector field $\boldsymbol{\tau}$ along the curve since

$$\mathbf{v}\cdot\boldsymbol{\tau} = \left\{\frac{dx_2}{dt}, -\frac{dx_1}{dt}\right\}\cdot\left\{\frac{dx_1}{dt}, \frac{dx_2}{dt}\right\} = 0;$$

and when $\boldsymbol{\tau} = \{1, 0\}$ then $\boldsymbol{\nu} = (0, -1)$, so the pair of orthogonal vectors $\boldsymbol{\nu}, \boldsymbol{\tau}$ has the orientation of the vector space $\mathbb{R}^2$. In these terms the version (6.40) of Green's Theorem takes the form

stokcrg6

$$\int_\gamma \mathbf{f}\cdot\boldsymbol{\nu}\, ds = \int_D \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}\right). \tag{6.44}$$

While the line integral $\int_\gamma \mathbf{f}\cdot\boldsymbol{\tau}\, ds$ measures the component of the vector field $\mathbf{f}$ in the direction of the tangent vector to γ, the integral $\int_\gamma \mathbf{f}\cdot\boldsymbol{\nu}\, ds$ Maure's the component of the vector field $\mathbf{f}$ in the direction of the outward normal vector to γ. The first integral sometimes is interpreted as the work done by the vector field $\mathbf{f}$ in pushing a particle along the curve γ, while the second integral sometimes is interpreted as the total flow across the curve γ of a substance moving with the velocity at any point $\mathbf{x}$ described by the vector $\mathbf{f}(\mathbf{x})$. For a conservative vector field $\mathbf{f} = \nabla h$ with the potential function h Green's Theorem in these two cases takes the forms

stokcrg7a

$$\begin{aligned}\int_\gamma \nabla h\cdot\boldsymbol{\tau}\, ds &= \int_D \left(\frac{\partial^2 h}{\partial x_1\partial x_2} - \frac{\partial^2 h}{\partial x_2\partial x_1}\right) = 0,\\ \int_\gamma \nabla h\cdot\boldsymbol{\nu}\, ds &= \int_D \left(\frac{\partial^2 h}{\partial x_1^2} + \frac{\partial^2 h}{\partial x_2^2}\right) = \int_D \Delta h\end{aligned} \tag{6.45}$$

where Δh is the Laplacian of the function h, defined as $\Delta h(\mathbf{x}) = \frac{\partial^2 h(\mathbf{x})}{\partial x_1^2} + \frac{\partial^2 h(\mathbf{x})}{\partial x_2^2}$.

Another classical special case of Theorem 6.4 arises in three dimensions, and has a great number of applications in physics.

stokcrgss

Corollary 6.4 (Gauss's Theorem) *If $D \subset U$ is the image of a simple singular 3-chain $\mathcal{C}$ in $U \subset \mathbb{R}^3$ then*

stokth1q3x

$$\int_{\partial\mathcal{C}} \Big(f_1 dx_2\wedge dx_3 + f_2 dx_3\wedge dx_1 + f_3 dx_1\wedge dx_2\Big) = \int_D \left(\sum_{i=1}^{3}\frac{\partial f_i}{\partial x_i}\right) \tag{6.46}$$

for any $\mathcal{C}^\infty$ vector field $\mathbf{f} = \{f_1, f_2, f_3\}$ in U.

Proof: This is just the special case of the Theorem 6.4 for the differential form $\Omega_\mathbf{f}$ in $\mathbb{R}^3$, so no further proof is required.

For the differential form $\Omega_\mathbf{f}$ associated to a vector field $\mathbf{f}$ the surface integral in (6.46) can be written as in (6.17), while the integrand in the integral over the set D can be written as in (5.73), so Gauss's Theorem can be written in the alternative form

stokgssx1

$$\int_\Gamma \mathbf{f}\cdot\boldsymbol{\nu}\,dS = \int_D \nabla\cdot\mathbf{f} = \int_D \mathbf{div}\ \mathbf{f}(\mathbf{x}) \tag{6.47}$$

where Γ is the boundary chain of the simple singular 3-chain $\mathcal{C}$ and $D = |\mathcal{C}|$ is the image of the chain $\mathcal{C}$. As in the discussion of Green's Theorem, the surface integral can be interpreted as the total flow across the boundary Γ of a substance moving with the velocity at a point $\mathbf{x}\in\mathbb{R}^3$ described by the vector $\mathbf{f}(\mathbf{x})$. With this interpretation, $\nabla\cdot\mathbf{f}(\mathbf{x}) = \mathbf{div}\ \mathbf{f}(\mathbf{x})$ describes the amount of this mysterious substance created at the point $\mathbf{x}$, so the volume integral can be interpreted as the total amount of the substance being created within D. Gauss's Theorem asserts that the amount of substance being created in D is equal to the amount flowing outward across the boundary Γ of D, the amount of material diverging from the domain D and measured locally by $\mathbf{div}\ \mathbf{f}(\mathbf{x})$. For the particular case in which the vector field $\mathbf{f}$ is conservative, so is the gradient $\mathbf{f} = \nabla h$ of a potential function, Gauss's Theorem takes the form

stokgssx2

$$\int_\Gamma \nabla h(\mathbf{x})\cdot\boldsymbol{\nu}\,dS = \int_D \nabla\cdot\nabla h(\mathbf{x}) = \int_D \Delta h(\mathbf{x}) \tag{6.48}$$

in terms of the harmonic differential operator $\Delta h = \sum_{i=1}^3 \frac{\partial^2 h}{\partial x_i^2}$ in $\mathbb{R}^3$. For a simple example of Gauss's Theorem, to integrate the normal component of the vector field $\mathbf{f} = \{2x_1,\, x_2^2,\, x_3^2\}$ over the boundary Γ of the unit sphere $D = \{\ \mathbf{x}\in\mathbb{R}^3\ \big|\, \|\mathbf{x}\|_2^2 \le 1\ \}$, since $\nabla\cdot\mathbf{f} = 2 + 2x_2 + 2x_3$ it follows from (6.47) that $\int_\Gamma \mathbf{f}\cdot\boldsymbol{\nu}\,dS = 2\int_D(1 + x_2 + x_3)$. The unit ball is symmetric with respect to the variables x_2 and x_3 so the second two terms in the integral yield 0 and consequently $\int_\Gamma \mathbf{f}\cdot\boldsymbol{\nu}\,dS = 2\int_D 1 = \frac{8}{3}\pi$.

A final classical interpretation of the general Stokes's Theorem of Corollary 6.1, which was traditionally called Stokes's Theorem before that term was adopted for the general case, involves line integrals in $\mathbb{R}^3$ that are the boundaries of singular 2-chains.

stokcrstok

Corollary 6.5 *If $\mathcal{C}\subset U$ is a continuously differentiable singular 2-chain in an open subset $U\subset\mathbb{R}^3$ then*

stokcrstokq1

$$\int_{\partial\mathcal{C}} \mathbf{f}(\mathbf{x})\cdot\boldsymbol{\tau}\,ds = \int_{\mathcal{C}} \Big(\nabla\times\mathbf{f}(\mathbf{x})\Big)\cdot\boldsymbol{\nu}\,dS = \int_{\mathcal{C}} \mathbf{curl}\ \mathbf{f}(\mathbf{x})\cdot\boldsymbol{\nu}\,dS \tag{6.49}$$

for any $\mathcal{C}^\infty$ vector field $\mathbf{f}(\mathbf{x})$ in U.

Proof: This is just the special case of Corollary 6.1 for the differential form $\omega_{\mathbf{f}}$, so no further proof is necessary.

For a simple example of the classical Stokes's Theorem, let γ be the intersection of the unit sphere $S\subset\mathbb{R}^3$, the set $S = \{\,(x_1,x_2,x_3)\,|\,x_1^2+x_2^2+x_3^2 = 1\,\} = 0$, with the plane $x_3 = 0$ and $\mathbf{f}(\mathbf{x})$ be the vector field $\mathbf{f}(\mathbf{x}) = \{2x_1,\, x_2^2,\, x_3^2\}$; since the curve γ is the boundary of the disc $D = \{(x_1, x_2, x_3) | x_1^2 + x_2^2 \le 1,\ x_3\} = 0$ when the orientation of the disc is described by the differential form $dx_2 \wedge dx_1$ and $\nabla \times \mathbf{f}(\mathbf{x}) = 0$ Stokes's Theorem asserts that $\int_\gamma \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau}\, ds = \int_D \big((\nabla \times \mathbf{f}(\mathbf{x})\big) \cdot \boldsymbol{\nu}\, dS = 0$. For a perhaps more interesting observation though, if Ω is a closed 2-form in $\mathbb{R}^3$ and S is the unit sphere in $\mathbb{R}^3$, so that S is the boundary $S = \partial D$ of the unit ball, then by Gauss's Theorem $\int_S \Omega = \int_D d\Omega = 0$; but that does require that the differential form Ω is defined and closed in the entire ball D, just as with the corresponding example in $\mathbb{R}^2$ discussed on page Vpoin 126. However if $\Omega = d\omega$ is an exact differential form defined just in an open neighborhood of the unit sphere S, and not necessarily in the entire ball D, it is actually the case that $\int_S \Omega = 0$. Indeed if γ is a circle on S, the intersection of S with a plane in $\mathbb{R}^3$, then γ splits the sphere S into two pieces, $S = S_1 \cup S_2$, where γ is the boundary of both S_1 and S_2; but if S_1 and S_2 have the same orientation as S and if γ is oriented so that $\partial S_1 = \gamma$ then clearly $\partial S_2 = -\gamma$. Then by the classical Stokes's Theorem $\int_S \Omega = \int_{S_1} \Omega + \int_{S_2} \Omega = \int_\gamma \omega - \int_\gamma \omega = 0$. The classical Stokes's Theorem was also used to give a geometrical interpretation of the curl of a vector field. If $\mathbf{f}(\mathbf{x})$ is a $\mathcal{C}^1$ vector field in an open neighborhood U of a point $\mathbf{a} \in \mathbb{R}^3$, if D_ϵ is a disc of radius ϵ centered at the point $\mathbf{a}$ in a plane $L \subset \mathbb{R}^3$ through the point $\mathbf{a}$, with boundary $\gamma_\epsilon = \partial D_\epsilon$ in L, and if the plane L is chosen so that the unit normal vector $\boldsymbol{\nu}$ to the plane L is parallel to the vector **curl** $\mathbf{f}(\mathbf{a})$,then by Stokes's Theorem $\int_{\gamma_\epsilon} \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau} ds = \int_{D_\epsilon} \mathbf{curl}\ \mathbf{f}(\mathbf{x}) \cdot \nu\, dS$; and since the area of the disc D_ϵ is $|D_\epsilon| = \int_{D_\epsilon} dS$ it follows that

$$\frac{1}{\pi\epsilon^2} \int_{\gamma_\epsilon} \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau} ds = \frac{\int_{D_\epsilon} \mathbf{curl}\ \mathbf{f}(\mathbf{x}) \cdot \nu\, dS}{\int_{D_\epsilon} dS.}$$

The function **curl** $\mathbf{f}(\mathbf{x}) \cdot \nu$ is a continuous function of the point $\mathbf{x}$, so its average over the disc D_ϵ approaches the value $\mathbf{curl}\ \mathbf{f}(\mathbf{a}) \cdot \boldsymbol{\nu} = \|\mathbf{curl}\ \mathbf{f}(\mathbf{a})\|_2$ in the limit as ϵ tends to zero; consequently

stokcrstokq2

$$\|\mathbf{curl}\ \mathbf{f}(\mathbf{a})\|_2 = \lim_{\epsilon \to 0} \frac{1}{\pi\epsilon^2} \int_{\gamma_\epsilon} \mathbf{f}(\mathbf{x}) \cdot \boldsymbol{\tau}\, ds, \tag{6.50}$$

so that the length of the vector $\mathbf{curlf}(\mathbf{a})$ is the limit of the line integral of the vector $\mathbf{f}(\mathbf{x})$ around the curve γ_ϵ as ϵ tends to zero, where this integral can be viewed as describing the curl of the vector field $\mathbf{f}(\mathbf{x})$ at the point $\mathbf{a}$.

PROBLEMS: GROUP I

(1) Show that the integral

$$\int_\gamma \sin x_2 \sin x_3 \ dx_1 + x_1 \cos x_2 \sin x_3 \ dx_2 + x_1 \sin x_2 \cos x_3 \ dx_3$$

along a path from the point $(0,0,0)$ to the point $(1,1,1)$ in $\mathbb{R}^3$ is independent of the choice of the path γ.

(2) Let S be the cylindrical surface in $\mathbb{R}^3$ defined by the equations $x_1^2 + x_2^2 = 1,\ 0 \le x_3 \le 1$, so described parametrically by $(x_1, x_2, x_3) = (\cos\theta, \sin\theta, x_3)$ for $0 \le \theta \le 2\pi$, $1 \le x_3 \le 1$, with the orientation described by the differential form $d\theta \wedge dx_3$; and let $\tilde{S}$ be the union of the cylindrical surface S and the discs closing up the top and bottom of the cylinder S, so that $\tilde{S}$ is boundary of a solid cylinder, oriented with the orientation of S extended to an orientation of $\tilde{S}$ as the full boundary of the solid cylinder. Consider also the differential form

$$\omega = x_3\, dx_1 \wedge dx_2 + x_1 x_3\, dx_2 \wedge dx_3 + x_3\, dx_1 \wedge dx_3.$$

(i) Evaluate $\int_S \omega$.
(ii) Evaluate $\int_{\tilde{S}} \omega$, using Stokes's Theorem.

(3) Compute the surface integral

$$\int_X (x_1^2 + x_2^2) dx_1 \wedge dx_2 + x_1 dx_2 \wedge dx_3$$

where X is the curved part of the boundary of the cylinder of height 1 with base the unit circle at the origin in the plane of the variables x_1, x_2.

PROBLEMS: GROUP II

(4) Calculate $\int_X (\nabla \times \mathbf{f}) \cdot \boldsymbol{\nu}\, dS$ for the vector field $\mathbf{f}(\mathbf{x}) = \{x_2,\ x_1 x_2 x_3,\ x_2)$ where X is the piece of the paraboloid $x_3 = 3 - x_1^2 - x_2^2$ for which $x_3 \ge 0$, and $\boldsymbol{\nu}$ is the unit outward normal vector to X.

(5) In $\mathbb{R}^3$ let S be the surface $x_1^2 + x_2^2 = 1 - x_3$, $0 \le x_3 \le 1$, let S_0 be the disc $x_1^2 + x_2^2 \le 1$, $x_3 = 0$ and let R be the solid region $x_1^2 + x_2^2 \le 1 - x_3$, $0 \le x_3$; assume the usual orientation in $\mathbb{R}^3$, so that $dx_1 \wedge dx_2 \wedge dx_3$ is the volume element, and orient the surfaces S and S_0 so that the boundary of R is $\partial R = S + S_0$.

(i) Find $\int_S x_3 dx_1 \wedge dx_2$.

(ii) Find $\int_{\partial R} x_3 dx_1 \wedge dx_2 + x_2 dx_2 \wedge dx_3$.

(iii) Find $\int_S F \cdot \nu \, dS$ for the vector field $F = (x_1, x_2, 2x_3 - x_1 - x_2)$ where ν is the outward normal to the surface S and dS is the element of area of S.

(iv) Find $\int_\gamma x_1 dx_1 + 2x_2 dx_2 + 3x_3 dx_3$ where $\gamma \subset S$ is the curve in R that begins at the point $(1,0,0)$, continues around the boundary ∂S_0 in the direction of the natural orientation of that boundary to the point $(0,1,0)$, continues along the intersection of the surface S and the plane $x_1 = 0$ to the point $(0, \frac{1}{\sqrt{2}}, \frac{1}{2})$, continues then in a straight line in the plane $x_3 = \frac{1}{2}$ to the central line $x_1 = x_2 = 0$ and then along that line to the vertex $(0,0,1)$; and do it the easy way.

(6) Express the integral $\int_S \mathbf{f} \cdot \nu \, dS$, the area integral of the dot product of the unit normal vector ν and the vector $\mathbf{f}(x_1, x_2, x_3) = (x_1^2, x_3, -x_2)$ along the surface S of the unit sphere $x_1^2 + x_2^2 + x_3^2 = 1$, as an integral of an appropriate differential form on S. Evaluate the integral by using Stokes's Theorem.

(7) If ω is a closed differential 1-form on the complement of the origin in $\mathbb{R}^2$ show that there is a unique constant c such that $\omega = c\, d\theta + df$ where θ is the angle (or argument) in polar coordinates and f is a differentiable function in the complement of the origin in $\mathbb{R}^2$.

(8) Let $\Lambda^1_*(\Delta)$ be the set of $\mathcal{C}^\infty$ differential 1-forms in the cell $\Delta \subset \mathbb{R}^2$ that are identically 0 in an open neighborhood of the boundary $\partial\Delta$. If $\omega_1, \omega_2 \in \Lambda^1_*(\Delta)$ show that

$$(\omega_1, \omega_2) = \int_\Delta \omega_1 \wedge *\omega_2$$

is an inner product, where * is the Hodge duality operator. Why is the integral $\int_\Delta \omega_1 \wedge \omega_2$ not a reasonable candidate for an inner product? Show however that for any $\mathcal{C}^\infty$ function f in $\overline{\Delta}$ and any 1-form $\omega \in \Lambda^1_*(\Delta)$

$$\int_\Delta df \wedge \omega = \int_\Delta -f d\omega.$$

(9) A function f in $\mathbb{R}^n$ is said to have *compact support* if there is a compact set $S \subset \mathbb{R}^n$ such that $f(\mathbf{x}) = 0$ whenever $x \notin S$.
(i) Show that if f_i for $1 \le i \le n$ are continuously differentiable functions in $\mathbb{R}^n$ and at least one of them has compact support then

$$\int_{\mathbb{R}^n} df_1 \wedge df_2 \wedge \cdots \wedge df_n = 0.$$

(ii) Find an explicit example of continuously differentiable functions f_i in $\mathbb{R}^n$, of course none of which has compact support, such that

$$\lim_{r\to\infty} \int_{B_r} df_1 \wedge df_2 \wedge \cdots \wedge df_n = \infty$$

where $B_r = \{\, \mathbf{x} \in \mathbb{R}^n \;\big|\; \|\mathbf{x}\|_2 \le r \,\}$.
(iii) Show that if f_1, f_2 are $\mathcal{C}^\infty$ functions in $\mathbb{R}^n$ and at least one of them has compact support then

$$\int_{\mathbb{R}^n} f_1(\mathbf{x})\, df_2(\mathbf{x}) \wedge dx_2 \cdots \wedge dx_n = -\int_{\mathbb{R}^n} f_2(\mathbf{x})\, df_1(\mathbf{x}) \wedge dx_2 \cdots \wedge dx_n.$$

(10) Let $\mathbf{f} = \{f_1, f_2, f_3\}$ be a $\mathcal{C}^\infty$ vector field and g be a $\mathcal{C}^\infty$ function in an open subset $U \subset \mathbb{R}^3$; and let $\overline{D}$ be a simple singular cell in U, such that its boundary ∂D is a $\mathcal{C}^\infty$ submanifold of U.
(i) Find $d * \omega_{\mathbf{f}}$, where $*$ indicates the Hodge dual form.
(ii) Find $* \omega_{\mathbf{f}} \wedge dg$.
(iii) Show that if $\mathbf{f}(\mathbf{x})$ is a tangent vector to the surface ∂D at each point $\mathbf{x} \in \partial D$ then

$$\int_D \Big(f_1 \partial_1 g + f_2 \partial_2 g + f_3 \partial_3 g \Big) = -\int_D \Big(\partial_1 f_1 + \partial_2 f_2 + \partial_3 f_3 \Big) g.$$

(iv) Show that if $\nabla g(\mathbf{x})$ is a tangent vector to ∂D at each point $\mathbf{x} \in \partial D$ and if $\Delta g(\mathbf{x}) = \partial_1^2 g(\mathbf{x}) + \partial_2^2 g(\mathbf{x}) + \partial_3^2 g(\mathbf{x}) = 0$ at each point $\mathbf{x} \in D$ then $g(\mathbf{x})$ is constant in D.

Appendix A

Exterior Algebras

math218notesa

Algebra is the study of a variety of structures: groups, rings, algebras, fields, and so on. A structure of considerable use in the calculus of several variables is exterior algebra, which is really a topic in linear algebra, or in multilinear algebra to be more precise, but which is not covered in most standard courses in linear algebra; so a short survey of the relevant material in an elementary form sufficient for purposes here is included in this appendix.

If V is a real vector space of dimension n with the basis $\xi_1, \ldots, \xi_n$, then to any integer k in the range $1 \le k \le n$ there is associated another vector space $\Lambda^k(V)$, the basis of which consists of a set of vectors called **basic k-vectors** and denoted by

apkv (A.1) $$\xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k} \quad \text{where } 1 \le i_1 < i_2 < \ldots < i_k \le n.$$

For $k = 1$ the basic 1-vectors thus are just the basis vectors ξ_i themselves, so that $\Lambda^1(V) = V$ and consequently $\dim \Lambda^1(V) = n$. For $k = n$ on the other hand there is just a single basic n-vector $\xi_1 \wedge \xi_2 \wedge \cdots \wedge \xi_n$, so $\dim \Lambda^n(V) = 1$. In general for $1 \le k \le n$ the basic k-vectors are described by the sets of k-tuples of integers from 1 to n in ascending order, so by the number of distinct subsets of k integers from the set of the first n integers, and consequently $\dim \Lambda^k(V) = \binom{n}{k}$. It is customary to extend this set of vector spaces by setting $\Lambda^0(V) = \mathbb{R}$. It is also customary, and very convenient, to introduce the **formal k-vectors** $\xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k}$ for arbitrary sets of k indices i_j in the range $1 \le i_j \le n$, with the understanding that

ap1 (A.2)
$$\begin{aligned} &\xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k} = 0 \quad \text{if the indices } i_1, i_2, \ldots, i_k \text{ are not distinct,} \\ &\xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k} = (\operatorname{sgn} \pi) \xi_{j_1} \wedge \cdots \wedge \xi_{j_k} \\ &\qquad\qquad\qquad\qquad \text{if } j_1, \ldots, j_k \text{ is a permutation } \pi \text{ of } i_1, \ldots, i_k, \end{aligned}$$

where $\operatorname{sgn} \pi = \pm 1$ is the sign of the permutation π. If the permutation is described by a permutation matrix M_π, with a single entry of 1 in each row and column and entries of 0 elsewhere, then $\operatorname{sgn}\pi = \det M_\pi$; equivalently $\operatorname{sgn}\pi$ is the number of transpositions of adjacent integers that transform the set $j_1, \ldots, j_k$ to the set $i_1, \ldots, i_k$. Any nonzero formal k-vector thus is equal to either a unique basic k-vector or the negative of a unique basic k-vector. A vector $\omega \in \Lambda^k(V)$ can be written as a unique linear combination of the basic k-vectors

ap3 (A.3)
$$\omega = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} a_{i_1 i_2 \cdots i_k} \xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k}$$

for some real numbers $a_{i_1 i_2 \cdots i_k}$; this is called the *reduced* form for the vector ω. Alternatively the vector ω can be written in various ways as a linear combination of the formal k-vectors

ap3a (A.4)
$$\omega = \sum_{i_1, i_2, \cdots, i_k = 1}^{n} b_{i_1 i_2 \cdots i_k} \xi_{i_1} \wedge \xi_{i_2} \wedge \cdots \wedge \xi_{i_k}$$

for some real numbers $a_{i_1 i_2 \cdots i_k}$; and this is called the **unreduced** form for that vector. By using (A.2) an unreduced form for a vector ω can be reduced to a unique reduced form for that vector. For example if $\dim V = 3$ a basis for the 3-dimensional vector space $\Lambda^2(V)$ consists of the basic 2-vectors $\xi_1 \wedge \xi_2$, $\xi_1 \wedge \xi_3$ and $\xi_2 \wedge \xi_3$; a vector ω in the unreduced form

$$\begin{aligned} \omega = b_{12}\xi_1 \wedge \xi_2 &+ b_{13}\xi_1 \wedge \xi_3 + b_{23}\xi_2 \wedge \xi_3 \\ &+ b_{21}\xi_2 \wedge \xi_1 + b_{31}\xi_3 \wedge \xi_1 + b_{32}\xi_3 \wedge \xi_2 \end{aligned}$$

is in the reduced form

$$\omega = (b_{12} - b_{21})\xi_1 \wedge \xi_2 + (b_{13} - b_{31})\xi_1 \wedge \xi_3 + (b_{23} - b_{32})\xi_2 \wedge \xi_3.$$

The addition of vectors can be carried out in either the reduced or unreduced form, whichever is more convenient.

There is an additional algebraic operation on vectors in the spaces $\Lambda^k(V)$ that associates to an ordered pair of vectors $\omega \in \Lambda^k(V)$ and $\sigma \in \Lambda^l(V)$ the **exterior product** or **wedge product** $\omega \wedge \sigma \in \Lambda^{k+l}(V)$, of these two vectors; explicitly if

apk (A.5)
$$\omega = \sum_{i_1, \ldots, i_k = 1}^{n} a_{i_1 \cdots i_k} \xi_{i_1} \wedge \cdots \wedge \xi_{i_k} \in \Lambda^k(V)$$

and

apl (A.6)
$$\sigma = \sum_{j_1, \ldots, j_1 = 1}^{n} b_{j_1 \cdots j_l} \xi_{j_1} \wedge \cdots \wedge \xi_{j_l} \in \Lambda^l(V)$$

then

appr (A.7)
$$\omega \wedge \sigma = \sum_{\substack{i_1, \ldots, i_k \\ j_1, \ldots, j_l}}^{n} a_{i_1 \cdots i_k} b_{j_1 \cdots j_l} \xi_{i_1} \wedge \cdots \wedge \xi_{i_k} \wedge \xi_{j_1} \wedge \cdots \wedge \xi_{j_l}.$$

The unreduced form of the vectors is clearly very convenient for handling wedge products. The order of the vectors ω and σ is important, since

apord (A.8) $$\omega \wedge \sigma = (-1)^{kl} \sigma \wedge \omega.$$

To demonstrate that, when the vectors ω and σ in (A.7) are reversed of course

$$b_{j_1 \cdots j_l} a_{i_1 \cdots i_k} = a_{i_1 \cdots i_k} b_{j_1 \cdots j_l}$$

since these are merely real numbers; but the permutation that exchanges

$$\xi_{j_1} \wedge \cdots \wedge \xi_{j_l} \wedge \xi_{i_1} \wedge \cdots \wedge \xi_{i_k} \quad \text{and} \quad \xi_{i_1} \wedge \cdots \wedge \xi_{i_k} \wedge \xi_{j_1} \wedge \cdots \wedge \xi_{j_l}$$

can be written as the composition first of the permutation that transfers ξ_{i_1} across $\xi_{j_1} \wedge \cdots \wedge \xi_{j_l}$, a permuation with sign $(-1)^l$, followed by the permutation that transfers ξ_{i_2} across $\xi_{j_1} \wedge \cdots \wedge \xi_{j_l}$, another permutation of sign $(-1)^l$, and so on, and this product of k permutations each of sign $(-1)^l$ is a permutation of sign $(-1)^{kl}$. In the special case in which $k = 0$ a vector $\omega \in \Lambda^0(V)$ is just a real number a, and the wedge product $\omega \wedge \sigma$ with the vector (A.6) is the vector $a\sigma$. Of course $\omega \wedge \sigma = 0$ if $k + l > n$. The formal union $\Lambda(V) = \bigcup_{k=0}^n \Lambda^k(V)$, or alternativaly the direct sum $\Lambda(V) = \bigoplus_{k=0}^n \Lambda^k(V)$, of the vector spaces $\Lambda^k(V)$ thus has two algebraic operations, the addition of vectors in each vector space $\Lambda^k(V)$ separately and the exterior product of vectors in any two of the vector spaces $\Lambda^k(V)$ and $\Lambda^l(V)$; the set $\Lambda(V)$ with this algebraic structure is called an **exterior algebra**, or more precisely the exterior algebra over a real vector space of dimension n.

The basic role of the exterior product is to handle a number of calculations involving determinants. For example if $\omega_1 = \sum_{j=1}^n a_{1j}\xi_j$ and $\omega_2 = \sum_{k=1}^n a_{2k}\xi_k$ then

$$\begin{aligned}
\omega_1 \wedge \omega_2 &= \sum_{j,k=1}^n a_{1j}a_{2k}\xi_j \wedge \xi_k = \sum_{1 \le j < k \le n} (a_{1j}a_{2k} - a_{1k}a_{2j})\xi_j \wedge \xi_k \\
&= \sum_{1 \le j < k \le n} \det \begin{pmatrix} a_{1j} & a_{1k} \\ a_{2j} & a_{2k} \end{pmatrix} \xi_j \wedge \xi_k,
\end{aligned}$$

so $\omega_1 \wedge \omega_2$ is the vector consisting of all the distinct determinants of 2×2 submatrices of the $2 \times n$ matrix $\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \end{pmatrix}$. For another example, if ω_i are any n vectors in V written explicitly in terms of the basis vectors $\xi_1, \ldots, \xi_n$ as

$$\omega_i = \sum_{j_i=1}^n a_{ij_i}\xi_{j_i} \quad \text{for } 1 \le i \le n,$$

then

$$\begin{aligned}
\omega_1 \wedge \cdots \wedge \omega_n &= \sum_{j_1,\ldots,j_n=1}^n a_{1j_1}a_{2j_2} \cdots a_{nj_n} \xi_{j_1} \wedge \xi_{j_2} \wedge \cdots \wedge \xi_{j_n} \\
&= \sum_{j_1,\ldots,j_n=1}^n a_{1j_1}a_{2j_2} \cdots a_{nj_n} \operatorname{sgn} \begin{pmatrix} j_1 & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix} \xi_1 \wedge \xi_2 \cdots \wedge \xi_n
\end{aligned}$$

where $\operatorname{sgn}\begin{pmatrix} j_1 & j_2 & \cdots & j_n \\ 1 & 2 & \cdots & n \end{pmatrix}$ is the sign of the permutation that takes the ordered set of integers $j_1, j_2, \cdots, j_n$ to the ordered set of integers $1, 2 \cdots, n$; and as is familiar from linear algebra the preceding equation can be rewritten

apdets
$$\omega_1 \wedge \cdots \wedge \omega_n = \det A \;\; \xi_1 \wedge \xi_2 \cdots \wedge \xi_n \tag{A.9}$$

for the matrix $A = \{a_{ij}\}$. To any $n \times n$ matrix $A = \{a_{ij}\}$ there is naturally associated the 2-form $\Omega_A = \sum_{i,j=1}^{n} a_{ij} \xi_i \wedge \xi_j$; and if $n = 2m$ is an even integer the m-th exterior product of the 2-form Ω_A with itself is an n-form

Apff
$$\Omega_A^m = P(A) \xi_1 \wedge \xi_2 \wedge \cdots \wedge \xi_n. \tag{A.10}$$

The coefficient $P(A)$ is a polynomial in the entries of the matrix A which is nontrivial if the matrix A is skew-symmetric; in that case it is called the **Pfaffian** of the matrix A, and is another invariant attached to that matrix, with the property that $P(A)^2 = \det A$.

The special cases of vector spaces V of dimensions 2 and 3 arise in a great many applications. If $\dim V = 2$ then $\dim \Lambda^0(V) = 1$ since $\Lambda^0(V) = \mathbb{R}$ while $\dim \Lambda^1(V) = 2$ with the basis ξ_1, ξ_2 and $\dim \Lambda^2(V) = 1$ with the basis $\xi_1 \wedge \xi_2$, which is the only nontrivial wedge product. For any two vectors $\omega = a_1 \xi_1 + a_2 \xi_2$ and $\sigma = b_1 \xi_1 + b_2 \xi_2$ then as above $\omega \wedge \sigma = \det \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}$. If $\dim V = 3$ then $\dim \Lambda^0(V) = 1$ since $\Lambda^0(V) = \mathbb{R}$ while $\dim \Lambda^1(V) = 3$ with the basis ξ_1, ξ_2, ξ_3 and $\dim \Lambda^2(V) = 3$ with the basis $\xi_1 \wedge \xi_2, \xi_1 \wedge \xi_3, \xi_2 \wedge \xi_3$ and $\dim \Lambda^3(V) = 1$ with the basis $\xi_1 \wedge \xi_2 \wedge \xi_3$. What is particularly interesting in this case is that both $\Lambda^1(V) = V$ and $\Lambda^2(V)$ are vector spaces of dimension 3, so an element of $\Lambda^2(V)$ also can be identified with a vector in the space V; it is traditional and natural to make this identification by associating to a vector $\mathbf{a} \in \mathbb{R}^3$ the vector $\omega_{\mathbf{a}} \in \Lambda^1(V)$ defined by

apx3a
$$\omega_{\mathbf{a}} = a_1 \xi_1 + a_2 \xi_2 + a_3 \xi_3 \tag{A.11}$$

and the vector $\Omega_{\mathbf{a}} \in \Lambda^2(V)$ defined by

apx3
$$\Omega_{\mathbf{a}} = a_1 \xi_2 \wedge \xi_3 + a_2 \xi_3 \wedge \xi_1 + a_3 \xi_1 \wedge \xi_2, \tag{A.12}$$

the obvious cyclic order. Various wedge products then introduce various algebraic operations between pairs of vectors in V. Thus

apx12
$$\omega_{\mathbf{a}} \wedge \Omega_{\mathbf{b}} = (a_1 b_1 + a_2 b_2 + a_3 b_3) \; \xi_1 \wedge \xi_2 \wedge \xi_3 \tag{A.13}$$

this exterior product when identified with a real number is just the dot product of the two vectors $\mathbf{a}$ and $\mathbf{b}$. On the other hand

$$\begin{aligned} \omega_{\mathbf{a}} \wedge \omega_{\mathbf{b}} &= \sum_{j,k=1}^{3} a_j b_k \, \xi_j \wedge \xi_k \\ &= (a_2 b_3 - a_3 b_2) \xi_2 \wedge \xi_3 + (a_3 b_1 - a_1 b_3) \xi_3 \wedge \xi_1 + (a_1 b_2 - a_2 b_1) \xi_1 \wedge \xi_2 \\ &= \omega_{\mathbf{c}}, \end{aligned}$$

for the vector $\mathbf{c} = \{c_1, c_2, c_3\} \in \mathbb{R}^3$ with the components

$$c_1 = \det\begin{pmatrix} a_2 & a_3 \\ b_2 & b_3 \end{pmatrix}, c_2 = \det\begin{pmatrix} a_3 & a_1 \\ b_3 & b_1 \end{pmatrix} c_3 = \det\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}; \tag{A.14}$$

this is just the cross-product $\mathbf{c} = \mathbf{a} \times \mathbf{b}$ of the two vectors $\mathbf{a}$ and $\mathbf{b}$, so $\omega_{\mathbf{a}} \wedge \omega_{\mathbf{b}} = \omega_{\mathbf{a}\times\mathbf{b}}$. It is an amusing exercise to develop the corresponding algebraic operations on 4-dimensional vector spaces.

Index