Calculus

Swapneel Mahajan

Contents

Contents iii
References 1
Remarks on the conditional 1
Pattern of mathematical writing 1
Pattern of mathematical learning 2

Part I. Functions of one real variable 3

Chapter 1. Sets and functions 4
1.1. Sets 4
1.1.1. Sets 4
1.1.2. Number systems 4
1.1.3. Set of real numbers 4
1.1.4. Properties of $\mathbb{R}$ 5
1.1.5. Intervals 5
1.2. Functions 6
1.2.1. Functions between sets 6
1.2.2. Graph of a function 7
1.2.3. Functions between real numbers 7
1.2.4. Absolute value function 7
1.2.5. Sine and cosine functions 7
1.2.6. Exponential and logarithm functions 8
1.2.7. Integer part function 8
1.2.8. Polynomial functions 9
1.2.9. Bounded and monotone functions 9
1.2.10. Convex functions 10

Chapter 2. Sequences 12
2.1. Sequences 12
2.1.1. Sequences 12
2.1.2. Visualizing a sequence 13
2.1.3. Bounded and monotone sequences 14
2.1.4. Convergence of sequences 14
2.1.5. Uniqueness of a limit 15
2.1.6. Convergent implies bounded 16
2.1.7. Algebra of sequences 16
2.1.8. Completeness property 17
2.1.9. Important limits 18
2.1.10. Convergence to infinity 18

Chapter 3. Continuity 19
3.1. Continuity 19
3.1.1. Continuous functions 19
3.1.2. Algebra of continuous functions 20
3.1.3. Characterization using sequences 20
3.1.4. Further properties of continuous functions 21
3.2. Limit of a function 23
3.2.1. Limit of a function 23
3.2.2. Algebra of limits of functions 23
3.2.3. Continuity and limit 24
3.2.4. Left and right limits 24
3.2.5. Types of discontinuities 24
3.2.6. Convergence to and at infinity of a function 25

Chapter 4. Differentiability 26
4.1. Differentiability 26
4.1.1. Differentiable functions 26
4.1.2. Left and right derivatives 27
4.1.3. Derivative function 27
4.1.4. Increment function 28
4.1.5. Algebra of differentiable functions 29
4.2. Maxima and minima 30
4.2.1. Global and local maxima/minima 31
4.2.2. Local maxima/minima: necessary condition 31
4.2.3. Rolle's theorem and mean value theorem 32
4.2.4. Mean value inequality 33
4.2.5. Increasing and decreasing functions 34
4.2.6. Convex functions 34
4.2.7. Critical points and global maxima/minima 35
4.2.8. Local maxima/minima: sufficient conditions 36
4.2.9. Points of inflection 37
4.2.10. Asymptotes 40

Chapter 5. Integration 42
5.1. Riemann integral 42
5.1.1. Riemann integrable functions 42
5.1.2. Riemann integral 43
5.1.3. Riemann sums 43
5.1.4. Domain additivity 43
5.1.5. Monotone functions 44
5.1.6. Continuous functions 44
5.1.7. Algebra of Riemann integrable functions 45
5.1.8. Further properties of the Riemann integral 45
5.1.9. Application: computing limits 46
5.2. Fundamental theorem of calculus 46
5.2.1. FTC. Part I 47
5.2.2. FTC. Part II 48
5.2.3. Integration by parts 48

5.2.4. Integration by substitution 48
5.3. Defining functions using the Riemann integral 49
5.3.1. Logarithmic function 49
5.3.2. Exponential function 50
5.3.3. Real powers of positive real numbers 50
5.3.4. Inverse trigonometric and trigonometric functions 51
5.4. Lengths, areas, volumes 51
5.4.1. Areas between curves 51
5.4.2. Volumes of solids 53
5.4.3. Arc length of a parametrized curve 55
5.4.4. Area of surface of revolution 57

Part II. Functions of several real variables 59

Chapter 6. Continuity 60
6.1. Real vector space 60
6.1.1. Real vector space 60
6.1.2. Dot product 60
6.1.3. Norm 61
6.1.4. Ball around a point 61
6.2. Functions of two real variables 62
6.2.1. Natural domain 62
6.2.2. Interior and boundary points 62
6.2.3. Bounded region 63
6.2.4. Graph of a function 63
6.2.5. Level curves and contour lines 64
6.3. Sequences 64
6.3.1. Sequences in $\mathbb{R}^2$ 64
6.3.2. Bounded sequences 65
6.3.3. Convergence of sequences 65
6.4. Continuity 66
6.4.1. Continuous functions 66
6.4.2. Algebra of continuous functions 66
6.4.3. Characterization using sequences 67
6.4.4. Further properties of continuous functions 68
6.5. Limit of a function 69
6.5.1. Limit of a function 69
6.5.2. Algebra of limits of functions 69
6.5.3. Continuity and limit 70

Chapter 7. Differentiability 71
7.1. Differentiability 71
7.1.1. Partial derivatives 71
7.1.2. Directional derivatives 73
7.1.3. Differentiability 73
7.1.4. Pair of increment functions 74
7.1.5. Algebra of differentiable functions 75
7.1.6. Geometric interpretation of the gradient 78

7.1.7. Higher partial derivatives 79
7.2. Tangent plane to a surface 79
7.2.1. Tangent line to a curve 80
7.2.2. Tangent plane to a surface 80
7.3. Maxima and minima 81
7.3.1. Global and local maxima/minima 81
7.3.2. Saddle points 82
7.3.3. Local maxima/minima: necessary condition 83
7.3.4. Local maxima/minima, saddle points: sufficient condition 84
7.3.5. Critical points and global maxima/minima 86
7.3.6. Contrained extrema 87

Chapter 8. Integration 88
8.1. Riemann integral on a rectangle 88
8.1.1. Riemann integrable functions 88
8.1.2. Riemann integral on a rectangle 89
8.1.3. Riemann sums 89
8.1.4. Monotone functions 90
8.1.5. Continuous functions 90
8.1.6. Fubini's theorem 90
8.1.7. Application: computing limits 92
8.2. Riemann integral in the plane 92
8.2.1. Riemann integral on a general region 92
8.2.2. Algebra of Riemann integrable functions 93
8.2.3. Elementary regions 93
8.2.4. Area of a general region 94
8.3. Change of variables 95
8.3.1. Jacobian matrix 95
8.3.2. Change of variables formula 96
8.3.3. Polar coordinates 97
8.4. Riemann integral in space 98
8.4.1. Riemann integral on a cuboid 98
8.4.2. Riemann integral on a general region 98
8.4.3. Change of variables 99
8.4.4. Cylindrical coordinates 99
8.4.5. Spherical coordinates 100

Chapter 9. Differential forms 102
9.1. Scalar and vector fields 102
9.1.1. Scalar fields 102
9.1.2. Vector fields 103
9.2. Gradient, curl, divergence 106
9.2.1. Gradient in three dimensions 106
9.2.2. Curl 106
9.2.3. Divergence in three dimensions 107
9.2.4. Gradient, curl, divergence 107
9.2.5. When is a vector field a gradient vector field? 107
9.3. Line integrals and FTC 108

9.3.1. Parametrized curve 108
9.3.2. Length of a parametrized curve 109
9.3.3. Line integral of a scalar field 109
9.3.4. Differential notation 110
9.3.5. Invariance under reparametrization 110
9.3.6. Arc length parametrization 110
9.3.7. Line integral of a vector field 111
9.3.8. Differential notation 111
9.3.9. Relating ds and $|ds|$ 112
9.3.10. Line integral of a gradient vector field 112
9.3.11. Path independence of line integrals 113
9.3.12. Invariance under reparametrization up to sign 114
9.3.13. Geometric curve 115
9.4. Green's theorem 116
9.4.1. Orienting the boundary curve 116
9.4.2. Green's theorem 116
9.4.3. Principle of deformation 117
9.4.4. Area calculation 119
9.5. Surface integrals 121
9.5.1. Parametrized surface 121
9.5.2. Fundamental vector product 121
9.5.3. Area of a parametrized surface 122
9.5.4. Surface integral of a scalar field 123
9.5.5. Invariance under reparametrization 125
9.5.6. Surface integral of a vector field 125
9.5.7. Differential notation 126
9.5.8. Invariance under reparametrization up to sign 126
9.5.9. Relating dS and $|dS|$ 127
9.5.10. Geometric surface 127
9.6. Gauss's divergence theorem 127
9.6.1. Orienting the boundary surface 128
9.6.2. Gauss's divergence theorem 128
9.6.3. Principle of deformation 129
9.6.4. Volume calculation 130
9.7. Stokes theorem 131
9.7.1. Orienting the boundary curve 131
9.7.2. Stokes theorem 131
9.7.3. Curl probe 133
9.7.4. Principle of deformation 133
9.7.5. Simply connected region 134
9.8. Differential forms 135
9.8.1. Orientations 135
9.8.2. Differential forms 136
9.8.3. Exterior derivative 136
9.8.4. Stokes theorem 137

Bibliography 139

References. Here is a list of useful general references, which is by no means exhaustive.

- For set theory: Halmos [14], Munkres [20, Chapter 1].
- For category theory: Mac Lane [17].
- For calculus: Apostol [1, 2], Ghorpade-Limaye [12], Marsden, Tromba, Weinstein [18].
- For analysis: Rudin [24], Pugh [22], Browder [8], Munkres [20]. Also useful are Apostol [3], Simmons [26], Tao [28, 29].
- For geometry on surfaces: Pressley [21], Thorpe [30], do Carmo [9].
- For manifolds and differential forms: do Carmo [10], Boothby [6], Morita [19], Lee [15], Spivak [27].
- Wikipedia is a good online source for getting a birds-eye-view of many concepts discussed in these notes. Blogs are also useful.

Pick a book that suits you. To understand the subject matter, it is not necessary to understand each and every sentence written in a particular book.

Remarks on the conditional. Consider the statements.

(1) If A, then B.
(2) If not B, then not A.
(3) If B, then A.
(4) If not A, then not B.

Statements (1) and (2) imply each other. Similarly, statements (3) and (4) imply each other.

Statements (1) and (3) are converses of each other. It is possible that one is true, while the other is false. Similarly, statements (2) and (4) are converses of each other.

Avoid/minimize usage of the symbol $\implies$. Note very carefully:

- The statement "A $\implies$ B." means "If A, then B.".
- The statement "A. $\implies$ B." means "A. Hence B.".

The two are different. If we write using $\implies$, then the two statements only differ in a fullstop which can be easily missed. So it is better not to use it.

Now appreciate the difference in the statements.

- A $\implies$ B. $\implies$ C. Better to say: A implies B. Hence C.
- A. $\implies$ B. $\implies$ C. Better to say: A. Hence B. Hence C.
- A. $\implies$ B $\implies$ C. Better to say: A. Hence B implies C.

The terms 'necessary condition' and 'sufficient condition' also appear often in mathematical writing. Their precise relation to a conditional is as follows. Let us go back to the statement 'If A, then B.' Here B is a necessary condition for A, while A is a sufficient condition for B.

Pattern of mathematical writing. While writing mathematics, one makes use of some technical constructs. They are as follows, and usually appear in the order given below.

- Definitions,
- Lemmas, Propositions,
- Theorems,
- Corollaries,

- Examples.

Many times, in mathematical discovery, it is the right definition that one is searching for to explain a bunch of phenomena that are known/believed to be true.

Pattern of mathematical learning. Many times, it is hard to immediately comprehend a definition. So one goes ahead, and reads the subsequent lemmas, theorems, examples. Then one again goes back to the definition followed by the lemmas and so on. This time round, things makes more sense. Then we repeat this process again, and again. Eventually everything makes sense. This process is called "rote learning" which is seeped in the indian tradition of learning.

Part I

Functions of one real variable

CHAPTER 1

Sets and functions

1.1. Sets

Sets are the building blocks of modern mathematics. We recall them briefly, focussing on number systems, particularly on the set of real numbers.

1.1.1. Sets. A set consists of elements. Let us begin with a couple of examples of sets.

- A = set of dogs in iitb campus,
- B = set of students in MA 105.

You can write down many similar examples.

1.1.2. Number systems. Now let us look at some standard sets related to number systems.

(1) $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ = set of natural numbers
(2) $\mathbb{N}_+ = \{1, 2, 3, \ldots\}$ = set of positive natural numbers
(3) $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ = set of integers
(4) $\mathbb{Q} = \{m/n : m, n \in \mathbb{Z},\ n \neq 0\}$ = set of rational numbers
(5) $\mathbb{R}$ = set of real numbers
(6) $\mathbb{R} \setminus \mathbb{Q}$ = set of irrational numbers

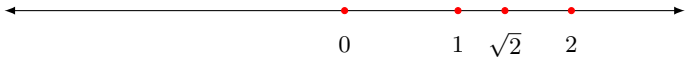
Lemma 1.1. *There is no rational number whose square is* 2.

PROOF. Suppose $(p/q)^2 = 2$, that is, $p^2 = 2q^2$ for some integers p, q such that $q \neq 0$, and p and q have no common factor. Now 2 divides p^2, and hence also divides p. So $p = 2r$ for an integer r. Then $2q^2 = p^2 = (2r)^2 = 4r^2$, and so $q^2 = 2r^2$. Now 2 divides q^2, and hence also divides q. Thus 2 is a common factor of p and q, which is a contradiction. □

The above result motivates the consideration of number systems which are larger than $\mathbb{Q}$ such as $\mathbb{R}$.

Remark 1.2 (Algebraic structures). There is no formal definition of a number system. However, the above considerations led to abstract concepts such as monoids, groups, rings, fields (in the later part of the nineteenth and early part of the twentieth century). For example, $\mathbb{Z}$ is an example of a ring, while $\mathbb{Q}$ and $\mathbb{R}$ are examples of fields. For more details, see Artin [4], Dummit-Foote [11].

1.1.3. Set of real numbers. It is customary to represent the set of real numbers $\mathbb{R}$ as a line as follows.



Elements of $\mathbb{R}$ are points on the line. Have we filled all the "holes" in the line? The set of rational numbers does not achieve this goal, but we believe that the set of real numbers does. There are two standard ways to pass from $\mathbb{Q}$ to $\mathbb{R}$, namely,

- Dedekind cuts,
- Cauchy sequences.

These two constructions were made in the nineteenth century around 1870. Note: $\sqrt{2} \in \mathbb{R}$.

1.1.4. Properties of $\mathbb{R}$. We mention that the set of real numbers satisfies the following properties.

- algebraic properties (related to addition and multiplication).
- order properties (related to greater than and less than).
- completeness property.
- archimedean property (implied by completeness property).

The archimedean property says that for any $x \in \mathbb{R}$, there is a natural number $n \in \mathbb{N}$ such that $n > x$.

Let us use this property to prove that between any two distinct real numbers, there is a rational number and an irrational number:

Lemma 1.3. *Let $a, b \in \mathbb{R}$ with $a < b$. Then there is $r \in \mathbb{Q}$ and $s \in \mathbb{R} \setminus \mathbb{Q}$ such that $a < r, s < b$.*

Proof. Let us do this in two steps.

(i) Let $[x]$ denote the integer part of x, that is, $x - 1 < [x] \leq x$. Pick $n > \frac{1}{b-a}$, and put $m = [na] + 1$. Then $a < \frac{m}{n} < b$. Now take $r := m/n$.

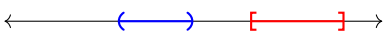
(ii) Using item (i), find $r \in \mathbb{Q}$ such that $a + \sqrt{2} < r < b + \sqrt{2}$. Then $a < r - \sqrt{2} < b$. Now take $s := r - \sqrt{2}$.

□

1.1.5. Intervals. We say $I \subseteq \mathbb{R}$ is an *interval* if $a, b \in I$ and $a < x < b$, then $x \in I$. Some standard examples of intervals are given below. For $a \leq b \in \mathbb{R}$, define

$$(a, b) := \{x \in \mathbb{R} \colon a < x < b\} \quad \text{and} \quad [a, b] := \{x \in \mathbb{R} \colon a \leq x \leq b\}.$$

These are the *open interval* and *closed interval*, respectively, from a to b. See illustrations below.



Similarly, define

$$(a, \infty) := \{x \in \mathbb{R} \colon a < x\} \quad \text{and} \quad [a, \infty) := \{x \in \mathbb{R} \colon a \leq x\},$$

and

$$(-\infty, b) := \{x \in \mathbb{R} \colon x < b\} \quad \text{and} \quad (-\infty, b] := \{x \in \mathbb{R} \colon x \leq b\}.$$

Note: The empty set $\emptyset$ and $\mathbb{R}$ are also intervals.

Observe:

$$\bigcap_{n=1}^{\infty} (a, b + \frac{1}{n}) = (a, b] \quad \text{and} \quad \bigcup_{n=1}^{\infty} [a, b - \frac{1}{n}] = [a, b).$$

Puzzle 1.4. A man has no money, but fortunately he has a silver bar which is 31 inches long. So he enters into the following agreement with his landlord for paying his March rent. He will pay one inch of his silver bar for each of the 31 days of March. The question is: What is the minimum of pieces he can cut his silver bar into in order to fulfil this requirement?

The silliest thing would be to cut the bar into 31 pieces and pay one piece each day. A better way to start would be to have 2 one inch pieces and a 3 inch piece, so that he can pay the first two days with the one inch pieces, and on the third day he can give the 3 inch piece and take back the 2 one inch pieces. He can use these to pay off the fourth and fifth days as well.

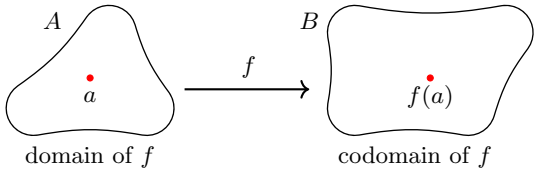
Puzzle 1.5. A shopkeeper has a single weight of 40 kilos. One day, his son mistakenly drops it on the floor, and it breaks into 4 pieces. The shopkeeper is very angry but his clever son shows him that with these 4 pieces, he can weigh on his balance any item whose weight is an integer between 1 to 40 (both inclusive). What are these 4 weights?

1.2. Functions

When we talk of sets, we also need to talk of ways to relate them. This is the notion of a function. We focus mainly on real-valued functions of a real variable. We discuss bounded, monotone, convex functions. We also informally recall many familiar examples; some of them are formalized later in Section 5.3.

For functions of more than one real variable, see Section 6.2.

1.2.1. Functions between sets. We specify a function as $f : A \to B$. Here A and B are sets. We say A is the domain of f, and B is the codomain of f. To every element $a \in A$, we have $f(a) = b \in B$.

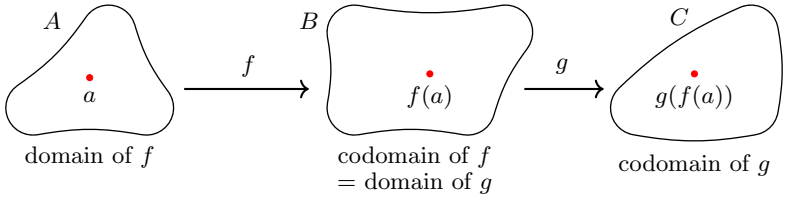


We write $f(A)$ for the image of f. It is the set of values taken by f. It is a subset of B.

For $f : A \to B$ and $g : B \to C$, define composite function $g \circ f : A \to C$ by

$$(g \circ f)(a) := g(f(a))$$

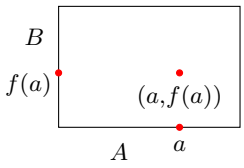
for $a \in A$.



1.2.2. Graph of a function. The *graph* of $f : A \to B$ is the subset of $A \times B$ defined by

$$\{(a, f(a)) \colon a \in A\}.$$

A schematic illustration is shown below.



1.2.3. Functions between real numbers. If the codomain of f is $\mathbb{R}$, that is, $f : A \to \mathbb{R}$, then we say f is real-valued. For example,

- for $A =$ set of dogs in iitb campus, consider $f(a) =$ weight of dog a,
- for $B =$ set of students in MA 105, consider $f(B) =$ IQ of student b.

We will mainly deal with functions f whose domain is $A \subseteq \mathbb{R}$. For functions on intervals, consider

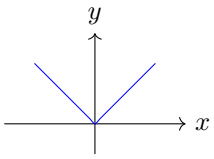
$$f : [0, 1] \to \mathbb{R}, \quad f(x) = x^2 + 5,$$
$$g : [0, 1] \to (3, 10), \quad g(x) = x^2 + 5.$$

Note very carefully: f and g are different functions because their codomains are different!

1.2.4. Absolute value function. An important real-valued function on $\mathbb{R}$ is the absolute value function. It is defined by

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = |x|,$$

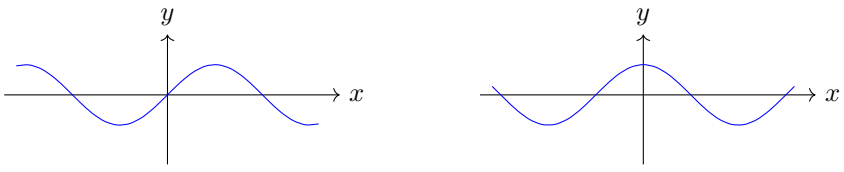
the absolute value of x. Its graph is shown below.



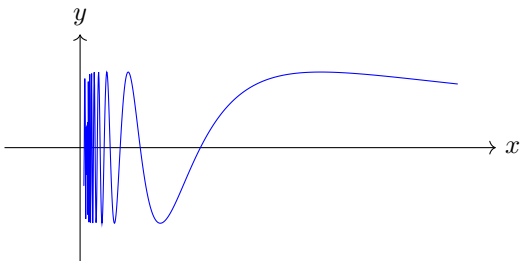
The absolute value function satisfies the following properties.

(i) $|x| \geq 0$ with equality iff $x = 0$. Thus, the image of f is $[0, \infty)$.
(ii) $|x| = |-x|$.
(iii) $|xy| = |x||y|$.
(iv) $-|x| \leq x \leq |x|$.
(v) $|x + y| \leq |x| + |y|$. This is known as the triangle inequality.

1.2.5. Sine and cosine functions. The graphs of the functions $f(x) = \sin x$ and $f(x) = \cos x$ are shown below.

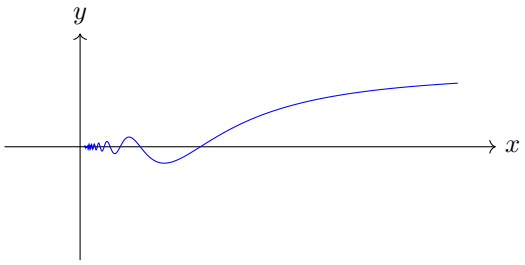


The graph of the function $f(x) = \sin(1/x)$, for $x > 0$, is shown below.



(For clarity, we have stretched the x-axis.) The graph oscillates rapidly as it approaches the y-axis.

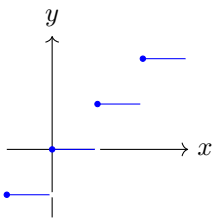
The graph of the function $f(x) = x\sin(1/x)$, for $x > 0$, is shown below.



The graph oscillates exactly as before, but now the amplitude of the oscillations goes to zero as it approaches the y-axis.

1.2.6. Exponential and logarithm functions. The graphs of the functions $f(x) = e^x$ and $f(x) = \log x$ are shown below.

1.2.7. Integer part function. The integer part $[x]$ of a real number x is the greatest integer which is less than or equal to x. For example, $[.5] = 0$, $[2] = 2$, $[2.1] = 2$. The graph of the integer part function $f(x) = [x]$ is shown below.



1.2.8. Polynomial functions. Polynomials in one variable are functions which are finite linear combinations of 1, x, x^2 and so on. Each polynomial has a degree. Polynomials of

- degree zero are constants $p(x) = c$,
- degree one are linear functions $p(x) = ax + b$ with $a \neq 0$,
- degree two are quadratic functions $p(x) = ax^2 + bx + c$ with $a \neq 0$,

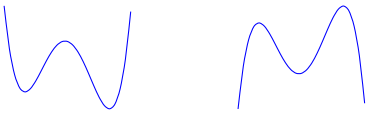
and so on.

The graph of a degree one polynomial (linear) looks as follows.

The graph of a degree two polynomial (quadratic) looks as follows.

The graph of a degree three polynomial (cubic) looks as follows.

The graph of a degree four polynomial (quartic) looks as follows.



For each degree, we have drawn two graphs depending on the sign of the leading coefficient. Also, the above pictures show the generic case. They may degenerate in specific cases. For example, compare the graph of $f(x) = x^3$ with the left picture shown above for a cubic.

1.2.9. Bounded and monotone functions. There are properties which a given function may or may not have. For example, for a function f, we can ask whether f is injective (into) or surjective (onto) or bijective (into and onto). Some other important properties are listed below.

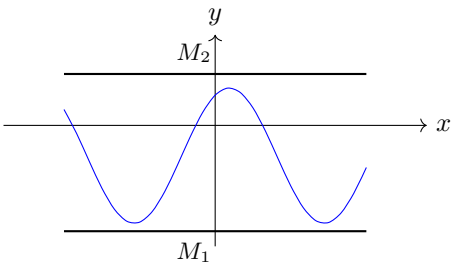
Definition 1.6. A function $f : A \to \mathbb{R}$ is

(i) *bounded above* if there is a real number M (upper bound) such that
$$f(x) \leq M$$
for $x \in A$,

(ii) *bounded below* if there is a real number M (lower bound) such that
$$M \leq f(x)$$
for $x \in A$,

(iii) *bounded* if it is bounded above and bounded below, that is, if there are real numbers M_1, M_2 such that
$$M_1 \leq f(x) \leq M_2$$
for $x \in A$.

A bounded function can be visualized as follows.



The *global maximum* of f is its least upper bound, and the *global minimum* of f is its greatest lower bound. By completeness property of $\mathbb{R}$, these necessarily exist, but they may not be attained at any point of A.

Definition 1.7. Let I be an interval, and let $f : I \to \mathbb{R}$. We say f is

(i) *(monotonically) increasing* on I if for $x_1, x_2 \in I$,

$$x_1 < x_2 \implies f(x_1) \leq f(x_2).$$

(ii) *(monotonically) decreasing* on I if for $x_1, x_2 \in I$,

$$x_1 < x_2 \implies f(x_1) \geq f(x_2).$$

(iii) *monotonic* on I if it increasing on I, or it is decreasing on I.

We use the terms *strictly increasing* and *strictly decreasing* if the inequalities $\leq$ and $\geq$ above can be replaced by $<$ and $>$.

Note: The constant function $f(x) = 3$ is both increasing and decreasing, but it is not strictly increasing or strictly decreasing.

1.2.10. Convex functions.

Definition 1.8. For I an interval, let $f : I \to \mathbb{R}$ be a function.

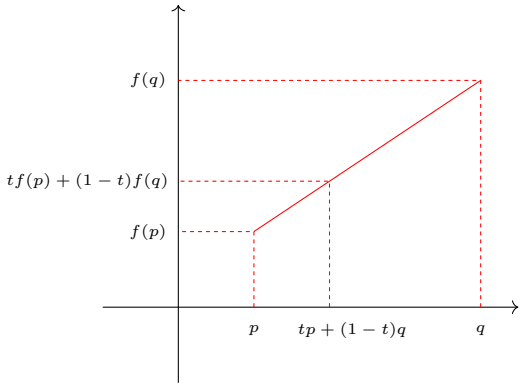
(i) f is *convex* if for $p < q$ in I and $t \in (0, 1)$,

$$f(tp + (1-t)q) \leq tf(p) + (1-t)f(q).$$

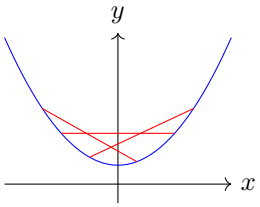
We use the term *strictly convex* if the above inequality is strict.

(ii) f is *(strictly) concave* if $-f$ is (strictly) convex. This can also be defined directly by reversing the inequality.

The different points involved in the definition of a convex function are illustrated below.



In geometric terms, a function f is convex if the chord joining any pair of points $(p, f(p))$ and $(q, f(q))$ on the graph of f lies on or above the graph of f. This is illustrated below.

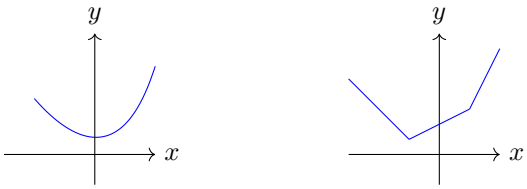


Exercise 1.9. Show: The convexity condition can be equivalently written as: For any $p < x < q$ in I,

$$f(x) \leq f(p) + \frac{f(q) - f(p)}{q - p}(x - p).$$

For strict convexity, we replace $\leq$ by $<$ above. For (strictly) concave, we use $\geq$ and $>$.

The graph of a typical convex function on $\mathbb{R}$ is shown below on the left. The graph of a function on $\mathbb{R}$ which is convex but not strictly convex is shown below on the right. A concrete example is the absolute value function.



Mention convex sets, and the fact that a convex set in $\mathbb{R}$ is the same as an interval.

CHAPTER 2

Sequences

2.1. Sequences

We introduce sequences of real numbers, and define the notion of convergence of such a sequence. We connect convergence to the property of being monotone and bounded. This is related to completeness property of $\mathbb{R}$.

2.1.1. Sequences.

Definition 2.1. A *sequence of real numbers* is a function $f : \mathbb{N}_+ \to \mathbb{R}$ from the set of positive integers to the set of real numbers.

Put $f(n) = a_n$. Thus specifying the function f is the same as specifying

$$a_1, a_2, a_3, \ldots.$$

We shall use the notation $\{a_n\}$ for short. We call a_n the n-th term of the sequence.

Example 2.2. Here are a few sample examples of sequences.

(1) $a_n = 1/n$.

$$1, 1/2, 1/3, 1/4, \ldots$$

(2) $a_n = n$.

$$1, 2, 3, 4, \ldots$$

(3) $a_n = (-1)^n$.

$$-1, 1, -1, 1, \ldots$$

(4) $a_n = n^2$.

$$1, 4, 9, 16, \ldots$$

(5) $a_n = \sqrt{2}$.

$$\sqrt{2}, \sqrt{2}, \sqrt{2}, \ldots$$

This is a *constant sequence*.

(6) $a_n = 2^n$.

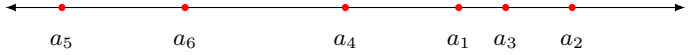
$$2, 4, 8, 16, \ldots$$

(7) $a_1 = 1, a_2 = 1$ and $a_n = a_{n-1} + a_{n-2}$ for $n \geq 3$.

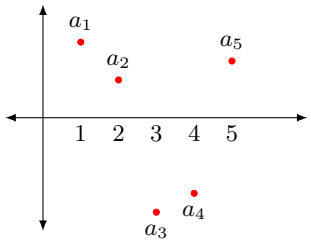
$$1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

This is the *Fibonacci sequence*.

2.1.2. Visualizing a sequence. A sequence may be visualized on the real line as follows by marking its terms $a_1, a_2, a_3, \ldots$.



It may also be visualized as the graph of the function $\mathbb{N}_+ \to \mathbb{R}$. In the picture below, we have marked the first 5 terms of the sequence.



Remark 2.3. We make some remarks related to the notion of a sequence.

(1) A sequence is always infinite. For example,
$$a_1, a_2, a_3, a_4$$
which is a tuple of four real numbers is not a sequence.

(2) A sequence need not be given by an algebraic formula. For example, we can define a sequence using the digits in the decimal expansion of $\sqrt{2}$. We can also do something like
$$-1, 3, 4, 5, 2, 2, 2, 2, \ldots,$$
that is, the sequence is constant barring the first few terms.

(3) ∞ is not a real number. Thus,
$$-1, 2, \infty, 1/5, \ldots$$
is not a sequence. Similarly, $\{\frac{1}{n-1}\}$ does not define a sequence since it is not defined at $n = 1$.

(4) The formula $a_n = \frac{1}{n-5}$ does not define a sequence (since it is not defined at $n = 5$).

(5) The following is not a sequence.
$$\ldots, a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2, a_3, \ldots.$$
It arises from a function $f : \mathbb{Z} \to \mathbb{R}$.

(6) An example of a sequence which contains each integer exactly once is
$$0, 1, -1, 2, -2, 3, -3, \ldots.$$

(7) If $\{a_n\}$ and $\{b_n\}$ are two sequences, then interleaving gives a third sequence
$$a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, \ldots$$
For example, the sequence $a_n = (-1)^n$ arises by interleaving the constant -1 sequence and constant 1 sequence.

Exercise 2.4. Construct a sequence which contains all rational numbers. (One way is to use Cantor's famous diagonalization argument.)

2.1.3. Bounded and monotone sequences. We now define some properties which a given sequence may or may not have.

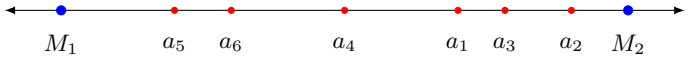
Definition 2.5. A sequence $\{a_n\}$ of real numbers is

(i) *bounded above* if there is a real number M such that
$$a_n \leq M$$
for $n \geq 1$,

(ii) *bounded below* if there is a real number M such that
$$M \leq a_n$$
for $n \geq 1$,

(iii) *bounded* if it is bounded above and bounded below, that is, if there are real numbers M_1, M_2 such that
$$M_1 \leq a_n \leq M_2$$
for $n \geq 1$.

A bounded sequence can be visualized on the real line as follows.



Definition 2.5 is the special case $A := \mathbb{N}_+$ of Definition 1.6.

Definition 2.6. A sequence $\{a_n\}$ of real numbers is

(i) *(monotonically) increasing* if
$$a_1 \leq a_2 \leq a_3 \leq \dots,$$

(ii) *(monotonically) decreasing* if
$$a_1 \geq a_2 \geq a_3 \geq \dots,$$

(iii) *monotonic* if it is either (monotonically) increasing or decreasing.

Exercise 2.7. For sequences in Example 2.2, which of the bounded and monotone properties hold?

2.1.4. Convergence of sequences. Where is a sequence heading?

Definition 2.8 (ϵ–n_0). Let $\{a_n\}$ be a sequence of real numbers. We say $\{a_n\}$ is *convergent* if there is $a \in \mathbb{R}$ such that the following condition holds.

For every $\epsilon > 0$, there is $n_0 \in \mathbb{N}_+$ such that
$$|a_n - a| < \epsilon$$
for $n \geq n_0$.

In this case, we say $\{a_n\}$ *converges* to a, or a is the *limit* of $\{a_n\}$, and write
$$\lim_{n\to\infty} a_n = a \qquad \text{or} \qquad a_n \to a \text{ (as } n \to \infty).$$

If a sequence does not converge, we say the sequence *diverges* or is *divergent*.

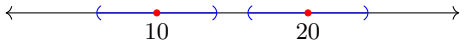
Example 2.9. Let us look at convergence in some of our examples.

(1) The sequence $a_n = 1/n$ converges to 0, or equivalently,
$$\lim_{n\to\infty} \tfrac{1}{n} = 0.$$
Why? Let $\epsilon > 0$. By archimedean property, there is $n_0 \in \mathbb{N}_+$ such that $\frac{1}{n_0} < \epsilon$. Therefore,
$$|a_n - a| = |\tfrac{1}{n}| \leq \tfrac{1}{n_0} < \epsilon$$
for $n \geq n_0$.
(2) The sequence $a_n = n$ does not converge, or equivalently, $\lim_{n\to\infty} n$ does not exist. (Use archimedean property.)
(3) The sequence $a_n = (-1)^n$ does not converge. (Take for example, $\epsilon = 1/2$.)

Remark 2.10. We make some remarks related to the notion of convergence.

(1) Suppose we know the first 100 terms of a sequence satisfy the formula $a_n = 1/n$. We cannot conclude from this that $\{a_n\}$ converges. It does not make sense to say $\{a_n\}$ converges at $n = 100$.
(2) Suppose we know $|a_n - 2| < 0.3$ for $n \geq 100$. This says that from a_{100} onwards, the sequence is confined to the interval $(1.7, 2.3)$. However, this does not imply that $\{a_n\}$ converges.
(3) Suppose we know $a_n = 1/n$ for $n \geq 100$. Then $a_n \to 0$ irrespective of the values $a_1, a_2, \ldots, a_{99}$. In general, the convergence of a sequence is unaltered if a finite number of its terms are replaced by some other terms.
(4) Consider $a_n = 1/n \to 0$. For $\epsilon = 1/10$, the smallest n_0 which works is 11. However, if $n_0 = 11$ works, then so does any number > 11. Note very carefully: The definition of convergence only requires us to find one n_0, not necessarily the smallest one. However, it is a good practice to specify the smallest n_0 for a given ϵ whenever possible.
(5) Many times, we will be dealing with two convergent sequences $a_n \to a$ and $b_n \to b$ at the same time. In such cases: For $\epsilon > 0$, the sequence $\{a_n\}$ will have its n_0, and $\{b_n\}$ will have its n_0. By taking the larger of the two, we will have an n_0 which works for both.

2.1.5. Uniqueness of a limit. Let $\{a_n\}$ be any sequence of real numbers. Parvati says that $\{a_n\}$ converges to 10, while Shankar says that $\{a_n\}$ converges to 20. Can both of them be right?



No. Give $\epsilon = 4$ to both of them, and ask them to provide n_0. Both cannot succeed since the open intervals $(6, 14)$ and $(16, 24)$ are disjoint as shown in the picture.

This argument generalizes to yield the following.

Lemma 2.11. *Limit of a sequence of real numbers is unique whenever it exists.*

Proof. Let $\{a_n\}$ be such a sequence. Suppose $a_n \to a$ and $a_n \to b$ with $a \neq b$. Take $\epsilon = |a - b|/2 > 0$. Let $n_0 \in \mathbb{N}_+$ be such that
$$|a_n - a| < \epsilon \quad \text{and} \quad |a_n - b| < \epsilon$$

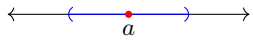
for $n \geq n_0$. Then

$$|a - b| \leq |a - a_{n_0}| + |a_{n_0} - b| < \epsilon + \epsilon = |a - b|,$$

which is a contradiction. Hence $a = b$. □

2.1.6. Convergent implies bounded. We now relate convergence of a sequence to its property of being bounded.

Proposition 2.12. *Let $\{a_n\}$ be a sequence of real numbers. If $\{a_n\}$ converges, then it is bounded. Equivalently, if $\{a_n\}$ is not bounded, then it does not converge.*



PROOF IDEA. A finite set of real numbers is always bounded. The problem is that a sequence contains infinitely many real numbers. But if $\{a_n\}$ converges, then some tail of this sequence lies in a finite neighborhood of the limit a. In the above picture, only finitely many terms of the sequence will be outside the blue interval. □

For example: The sequences $\{n\}$, $\{n^2\}$, $\{2^n\}$ are not bounded, and hence are divergent.

The converse of Proposition 2.12 is false. For example, take $a_n = (-1)^n$. This sequence is bounded but it does not converge.

2.1.7. Algebra of sequences. One can add two sequences, multiply two sequences, scalar multiply a sequence (by a real number). These operations are compatible with the notion of convergence in the following sense.

Lemma 2.13 (Limit theorems). *Suppose $a_n \to a$ and $b_n \to b$ are two convergent sequences of real numbers. Then*

(i) $a_n + b_n \to a + b$,
(ii) $ra_n \to ra$ *for* $r \in \mathbb{R}$,
(iii) $a_n b_n \to ab$,
(iv) $1/a_n \to 1/a$ *if* $a \neq 0$.

PROOF. For item (i): Let $\epsilon > 0$. Since $a_n \to a$ and $b_n \to b$, there is $n_0 \in \mathbb{N}_+$ such that

$$|a_n - a| < \epsilon/2 \quad \text{and} \quad |b_n - b| < \epsilon/2$$

for $n \geq n_0$. Now using triangle inequality,

$$|(a_n + b_n) - (a + b)| \leq |a_n - a| + |b_n - b| < \epsilon/2 + \epsilon/2 = \epsilon$$

for $n \geq n_0$. Thus, $a_n + b_n \to a + b$.

Proofs of items (ii), (iii), (iv) use similar ideas. □

Remark 2.14. For item (iv), strictly speaking, we must require $a_n \neq 0$ for $1/a_n$ to make sense. However, since $a_n \to a$ and $a \neq 0$, from some point on, the a_n are indeed nonzero (and convergence of a sequence is not affected if we change finitely many of its terms).

Lemma 2.15 (Sandwich lemma). *If $a_n \leq b_n \leq c_n$, and $a_n \to a$ and $c_n \to a$, then $b_n \to a$.*

PROOF. Let $\epsilon > 0$. Since $a_n \to a$ and $c_n \to a$, there is $n_0 \in \mathbb{N}_+$ such that

$$a - \epsilon < a_n < a + \epsilon \quad \text{and} \quad a - \epsilon < c_n < a + \epsilon$$

for $n \geq n_0$. Since $a_n \leq b_n \leq c_n$,

$$a - \epsilon < b_n < a + \epsilon$$

for $n \geq n_0$. □

Example 2.16. Let us illustrate the sandwich lemma.

(1) Let $a_n = \frac{n^3+3n^2+2}{n^4+7n^2+5}$. Then $a_n \to 0$ since

$$0 \leq a_n \leq \tfrac{1}{n} + \tfrac{3}{n^2} + \tfrac{2}{n^4} \to 0.$$

(2) Let $a_n = \frac{1}{n}\sin(\frac{1}{n})$. Then $a_n \to 0$ since

$$-\tfrac{1}{n} \leq a_n \leq \tfrac{1}{n} \quad \text{and} \quad \tfrac{1}{n} \to 0.$$

2.1.8. Completeness property. We now give two sufficient conditions for a sequence to converge. This is a partial converse to Proposition 2.12.

Proposition 2.17. *Let $\{a_n\}$ be a sequence of real numbers. Then:*

(i) *If $\{a_n\}$ is increasing and bounded above, then $\{a_n\}$ is convergent.*
(ii) *If $\{a_n\}$ is decreasing and bounded below, then $\{a_n\}$ is convergent.*

This result can be deduced using completeness property of $\mathbb{R}$. Since we have not discussed the latter, we take the above result for granted. Note: Items (i) and (ii) imply each other by replacing a sequence by its negative.

Example 2.18. Let us illustrate the completeness property.

(1) The sequence $a_n = 1/n$ is decreasing and bounded below by 0, hence it converges.

(2) Let $a_1 = 1$ and $a_n = \frac{3a_{n-1}+2}{6} = \frac{1}{2}a_{n-1} + \frac{1}{3}$ for $n \geq 2$. This sequence is bounded below by 0. Is it decreasing? The first few values are $a_1 = 1$, $a_2 = 5/6$, $a_3 = 3/4$. Now

$$a_n \leq a_{n-1} \iff \tfrac{1}{2}a_{n-1} + \tfrac{1}{3} \leq a_{n-1} \iff \tfrac{2}{3} \leq a_{n-1}$$

for $n \geq 2$.

Note: $a_1 \geq \frac{2}{3}$. If $a_{n-1} \geq \frac{2}{3}$ for some $n \geq 2$, then $a_n \geq \frac{1}{2}(\frac{2}{3}) + \frac{1}{3} = \frac{2}{3}$. So by induction, $a_n \geq \frac{2}{3}$ for $n \geq 1$. Hence $\{a_n\}$ is decreasing. By completeness property, $\{a_n\}$ converges (say to a).

To compute a, we may proceed as follows. In $a_n = \frac{1}{2}a_{n-1} + \frac{1}{3}$, lhs goes to a and rhs goes to $\frac{1}{2}a + \frac{1}{3}$. So $a = \frac{1}{2}a + \frac{1}{3}$, and hence $a = \frac{2}{3}$.

Exercise 2.19. Give an example of a sequence $\{a_n\}$ of real numbers which is strictly decreasing in absolute value, that is, $|a_n| > |a_{n+1}|$ for $n \geq 1$, but which does not converge.

Remark 2.20. Proposition 2.17 fails for $\mathbb{Q}$. For example, we can take the sequence of rational numbers $1, 1.4, 1.41, 1.414, \ldots$ arising from the decimal expansion of $\sqrt{2}$. This sequence is increasing and bounded above by say the rational number 1.5. But it does not converge in $\mathbb{Q}$. What we are seeing here is the fact that the set of rational numbers $\mathbb{Q}$ is not complete.

2.1.9. Important limits. We mention a couple of important limits.

Lemma 2.21. *Let $a \in \mathbb{R}$. Then:*

(i) *If $|a| < 1$, then $\lim_{n\to\infty} a^n = 0$.*
(ii) *If $a > 0$, then $\lim_{n\to\infty} a^{1/n} = 1$.*

PROOF. For item (i): The result is clear if $a = 0$. Let $0 < |a| < 1$. Then $\frac{1}{|a|} > 1$. Write $\frac{1}{|a|} = 1 + h$ for $h > 0$. Then

$$\frac{1}{|a|^n} = (1+h)^n = 1 + nh + \cdots + h^n \geq 1 + nh \geq nh.$$

Therefore,

$$0 \leq |a|^n \leq \frac{1}{nh} \to 0.$$

Result follows by sandwich lemma.

For item (ii): The result is clear if $a = 1$. Let $a > 1$. Then $a^{1/n} > 1$. Write $a^{1/n} = 1 + h_n$ for $h_n > 0$. Now $a = (1 + h_n)^n \geq nh_n$. Therefore, $0 \leq h_n \leq \frac{a}{n}$. So $h_n \to 0$, and $a^{1/n} \to 1$. Finally, let $0 < a < 1$. Then $\frac{1}{a} > 1$. So by previous case, $(\frac{1}{a})^{1/n} \to 1$. Therefore, $a^{1/n} \to 1$. □

2.1.10. Convergence to infinity. Suppose a sequence $\{a_n\}$ diverges. Then it makes sense to ask whether $\{a_n\}$ is converging to ∞ or $-\infty$ as explained below. We emphasize again that $\pm\infty$ are not real numbers.

Definition 2.22. Let $\{a_n\}$ be a sequence of real numbers.

(i) We say $\{a_n\}$ *converges to* ∞ or $\lim_{n\to\infty} a_n = \infty$ or $a_n \to \infty$ if the following condition holds. For every $\alpha \in \mathbb{R}$, there is $n_0 \in \mathbb{N}_+$ such that $a_n > \alpha$ for $n \geq n_0$.
(ii) We say $\{a_n\}$ *converges to* $-\infty$ or $\lim_{n\to\infty} a_n = -\infty$ or $a_n \to -\infty$ if the following condition holds. For every $\beta \in \mathbb{R}$, there is $n_0 \in \mathbb{N}_+$ such that $a_n < \beta$ for $n \geq n_0$.

For example: The sequence $a_n = n^2 \to \infty$ and $a_n = -n^3 \to -\infty$. The sequence $a_n = (-1)^n n$ is unbounded but does not converge either to ∞ or to $-\infty$.

Remark 2.23 (Metric spaces). We have focussed on sequences of real numbers. More generally, a sequence can take values in any set A. However, to define convergence, one needs a notion of distance in A. Such a set A is called a *metric space*. For $A = \mathbb{R}$, the distance is defined by $\text{dist}(x, y) := |x - y|$, and convergence as in Definition 2.8. This example generalizes to $A = \mathbb{R}^m$. The case $m = 2$ is explained in Definition 6.10.

CHAPTER 3

Continuity

3.1. Continuity

The intuitive idea of a continuous function f is that the graph of f has no "breaks". We now formalize this notion.

3.1.1. Continuous functions.

Definition 3.1 (ϵ–δ). Let $f : A \to \mathbb{R}$. We say f is *continuous* at $c \in A$ if the following condition holds.

For every $\epsilon > 0$, there is $\delta > 0$ such that

$$|x - c| < \delta \quad \Longrightarrow \quad |f(x) - f(c)| < \epsilon.$$

We say f is *continuous* on A if f is continuous at each point of A.

Example 3.2. Let us illustrate the notion of continuity.

(1) Let $f(x) = x$. Then f is continuous at all $c \in \mathbb{R}$.
Take $\delta = \epsilon$.

(2) Let $f(x) = 3x - 5$. Then f is continuous at all $c \in \mathbb{R}$.
Take $\delta = \epsilon/3$. Then $|x - c| < \delta$ implies

$$|(3x - 5) - (3c - 5)| = 3|x - c| < \epsilon.$$

(3) Let $f(x) = [x]$. Then f is continuous at non-integer points and discontinuous at integer points.
- c is a non-integer point. Pick $\delta > 0$ which avoids the adjacent integer points.
- c is an integer point. Give $\epsilon = 1/2$. No choice of δ works.

(4) Consider the *Dirichlet function*

$$f : [0, 1] \to \mathbb{R}, \qquad f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

It is discontinuous at all points. Give $\epsilon = 1/2$. No choice of δ works because in any open interval there is always a rational and an irrational by Lemma 1.3.

Exercise 3.3. Let $f : A \to \mathbb{R}$ be continuous at $c \in A$, and $f(c) > 0$. Then there is an open interval I containing c such that $f(x) > 0$ for all $c \in I$.

3.1.2. Algebra of continuous functions. One can add two functions, multiply two functions, scalar multiply a function (by a real number). These operations are compatible with the notion of continuity in the following sense.

Lemma 3.4. *Suppose $f, g : A \to \mathbb{R}$ are continuous at $c \in A$. Then so are*

(i) $f + g$,
(ii) rf *for* $r \in \mathbb{R}$,
(iii) fg,
(iv) $1/f$ *if* $f(c) \neq 0$.

PROOF. For item (i): Let $\epsilon > 0$. Since f and g are continuous at c, there is $\delta > 0$ such that

$$|x - c| < \delta \quad \Longrightarrow \quad |f(x) - f(c)| < \epsilon/2 \quad \text{and} \quad |g(x) - g(c)| < \epsilon/2.$$

Now using triangle inequality,

$$|(f + g)(x) - (f + g)(c)| \leq |f(x) - f(c)| + |g(x) - g(c)| < \epsilon/2 + \epsilon/2 = \epsilon.$$

Proofs of items (ii), (iii), (iv) use similar ideas. For item (iv): It suffices to prove that the function $1/x$ is continuous, and use Lemma 3.5 below. □

Lemma 3.5. *Let $f : A \to B$ and $g : B \to \mathbb{R}$. If f is continuous at $c \in A$ and g is continuous at $f(c) \in B$, then the composite $g \circ f$ is continuous at $c \in A$.*

PROOF IDEA. Given $\epsilon > 0$, pick $\delta' > 0$ using continuity of g at $f(c)$. Now taking $\delta' > 0$ as the ϵ, pick the required $\delta > 0$ using continuity of f at c. □

As a consequence:

- polynomials in x such as $p(x) = x^2$ and $p(x) = 2x^3 - 3x + 1$ are continuous,
- a rational function in x, that is $r(x) = p(x)/q(x)$, where p and q are polynomials, is continuous at c if $q(c) \neq 0$,
- a function such as $f(x) = x^3 \sin|x| + \cos x^2$ is continuous.

Example 3.6. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Then f is continuous at $c \neq 0$ since it is formed out of continuous functions. Let us see what happens at $c = 0$. Given $\epsilon > 0$, let $\delta = \epsilon$. Then $|x - 0| < \delta \implies |f(x) - f(0)| \leq |x| < \delta = \epsilon$. Hence f is continuous at 0.

Exercise 3.7. Define f as above but with $x \sin(1/x)$ replaced by $\sin(1/x)$. Show: f is not continuous at 0.

3.1.3. Characterization using sequences. We now characterize continuity of a function using sequences. This forges a connection between Definition 3.1 and Definition 2.8.

Proposition 3.8. *Let $f : A \to \mathbb{R}$. Then f is continuous at $c \in A$ iff the following condition holds.*

For any sequence $\{x_n\}$ in A with $x_n \to c$, we have $f(x_n) \to f(c)$.

PROOF. Suppose f is continuous at $c \in A$, and $x_n \to c$. We want to show $f(x_n) \to f(c)$. Let $\epsilon > 0$. Continuity of f at c yields a δ. Using this δ, we find a n_0 for $x_n \to c$. Thus for $n \geq n_0$, we have $|x_n - c| < \delta$, and hence $|f(x_n) - f(c)| < \epsilon$ as required.

Conversely, suppose the condition holds. We prove f is continuous at c by contradiction. So suppose f is not continuous at c. Then there is $\epsilon > 0$ for which no δ works. This gives a sequence $x_n \to c$ for which $|f(x_n) - f(c)| > \epsilon$ for $n \geq 1$. This is a contradiction. □

Example 3.9. Let us use Proposition 3.8 to show that certain functions are not continuous at a point.

(1) Consider the integer part function $f(x) = [x]$. At $c = 5$, $f(c) = 5$. Let $x_n = 5 - \frac{1}{n}$. Then $x_n \to 5$, but $[x_n] = 4$ and so $[x_n] \not\to 5$. Thus, f is not continuous at $c = 5$.

(2) Define

$$f(x) = \begin{cases} \sin(1/x) & \text{if } x \neq 0, \\ r & \text{if } x = 0. \end{cases}$$

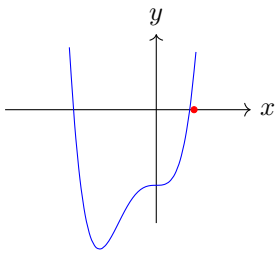
Then f is continuous at $c \neq 0$ since it is formed out of continuous functions. Let us see what happens at $c = 0$. Let $x_n = \frac{2}{(2n+1)\pi}$. Then $x_n \to 0$, but $f(x_n) = \sin(\frac{(2n+1)\pi}{2}) = (-1)^n$ does not converge. So f is not continuous at $c = 0$, no matter what r is.

3.1.4. Further properties of continuous functions.

Theorem 3.10 (Intermediate value property). *Let I be an interval, and $f : I \to \mathbb{R}$ be a continuous function. Let $r \in \mathbb{R}$ be such that $f(x_1) < r < f(x_2)$ for some $x_1 < x_2$ in I. Then there is $x \in (x_1, x_2)$ such that $f(x) = r$.*

The proof uses completeness property of $\mathbb{R}$, and is omitted.

Example 3.11. Let us show that the function $f(x) = x^4 + 2x^3 - 2$ has a root in $(0, 1)$. Its graph is shown below. The red point is $x = 1$.



Since f is a polynomial, it is continuous. Now $f(0) = -2$ and $f(1) = 1$. So by IVP, f attains every value between -2 and 1 in the interval $(0, 1)$, and in particular, the value 0.

Corollary 3.12. *Let $f : A \to \mathbb{R}$ be a continuous function, and $I \subseteq A$ be an interval. Then $f(I)$ is an interval.*

Exercise 3.13. Is there a continuous function from $[0, 1]$ onto $[2, 3]$? onto $[2, 3] \cup [4, 5]$? onto $(0, \infty)$? onto $[-1, 1]$?

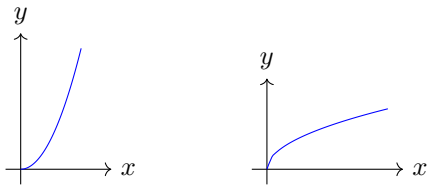
Corollary 3.14. *Let $f : I \to \mathbb{R}$ be continuous and injective. Then f is either increasing or decreasing. Also, $f^{-1} : f(I) \to \mathbb{R}$ is continuous.*

PROOF. Exercise. □

Let us use the above result to deduce the existence of the square root function

$$g : [0, \infty) \to [0, \infty), \qquad g(x) = \sqrt{x}.$$

Take $f : [0, \infty) \to [0, \infty)$ with $f(x) = x^2$. This function is continuous and injective. Also $f([0, \infty)) = [0, \infty)$. Put $g = f^{-1}$. The graph of f on $(0, 2)$ and of g on $(0, 4)$ are shown below.



Theorem 3.15. *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is bounded on $[a, b]$ and attains its global maximum and global minimum on $[a, b]$. Further, $f([a, b])$ is a closed and bounded interval.*

The proof is omitted.

Example 3.16. Let us see what can go wrong if the domain is an interval but not a closed interval.

(1) Take $f : (0, 1) \to \mathbb{R}$ with $f(x) = \frac{1}{x}$. Then f is continuous but not bounded.
(2) Take $f : [0, \infty) \to \mathbb{R}$ with $f(x) = x$. Then f is continuous but not bounded.
(3) Take $f : (0, 1) \to \mathbb{R}$ with $f(x) = x$. Then f is continuous and bounded, but does not attain its global maximum or global minimum.

Exercise 3.17. Construct a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that f takes every value exactly three times.

Exercise 3.18. Define the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 0 & \text{if } x \text{ is irrational,} \\ 1/q & \text{if } x = p/q \text{ in lowest terms.} \end{cases} \tag{3.1}$$

Show: f is continuous at all irrational points, but discontinuous at all rational points.

Puzzle 3.19. A pilgrim wants to go to a temple on the top of a mountain. He starts from the bottom at 8 in the morning, and reaches the top at 12. He stays there for a week. While coming down, he again starts at 8 in the morning, and reaches the bottom at 11. Show that there is a time between 8 and 11 when the pilgrim was at the same point on the mountain while ascending and descending.

3.2. Limit of a function

3.2.1. Limit of a function. Let $f : A \to \mathbb{R}$ and $c \in \mathbb{R}$ be such that there is $r > 0$ with $(c - r, c) \cup (c, c + r) \subseteq A$. In other words, A contains all points within distance r of c, except perhaps the point c.

Definition 3.20. We say $\lim_{x\to c} f(x)$ exists if there is $\ell \in \mathbb{R}$ such that for every sequence $\{x_n\}$ in A with $x_n \neq c$ and $x_n \to c$, we have $f(x_n) \to \ell$.

In this case, we write

$$\ell = \lim_{x\to c} f(x),$$

and say f has a limit at c.

Example 3.21. Let us illustrate the notion of limit.

(1) Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} 3x + 5 & \text{if } x \neq 0, \\ 1 & \text{if } x = 0. \end{cases}$$

Let $x_n \to 0$, $x_n \neq 0$ for $n \geq 1$. Then $f(x_n) = 3x_n + 5 \to 5$. Hence $\lim_{x\to 0} f(x) = 5$.

(2) Let $f(x) = [x]$.
- Let $x_n = 5 + (1/n)$, so $x_n \to 5$. Also $f(x_n) = 5$, so $f(x_n) \to 5$.
- Let $x_n = 5 - (1/n)$, so $x_n \to 5$. Also $f(x_n) = 4$, so $f(x_n) \to 4$.

Thus $\lim_{x\to 5} f(x)$ does not exist.

(3) Let $f(x) = \sin(1/x)$ for $x \in \mathbb{R} \setminus \{0\}$.

Let $x_n = \frac{2}{(2n+1)\pi}$, so $x_n \to 0$, but $f(x_n) = \sin(\frac{(2n+1)\pi}{2}) = (-1)^n$ does not converge.

Thus $\lim_{x\to 0} f(x)$ does not exist.

Remark 3.22 (ϵ–δ). Equivalently, similar to Definition 3.1 for continuity, we say:

$$\lim_{x\to c} f(x) = \ell$$

if the following condition holds.

For every $\epsilon > 0$, there is $\delta > 0$ such that

$$0 < |x - c| < \delta \implies |f(x) - \ell| < \epsilon.$$

It is possible to take this as a definition, and deduce Definition 3.20 as a consequence.

3.2.2. Algebra of limits of functions. The operations of addition, multiplication, scalar multiplication on functions are compatible with the notion of taking limits in the following sense.

Lemma 3.23 (Limit theorems). *Suppose* $\lim_{x\to c} f(x)$ *and* $\lim_{x\to c} g(x)$ *exist. Then*

(i)
$$\lim_{x\to c}(f + g)(x) = \lim_{x\to c} f(x) + \lim_{x\to c} g(x),$$

(ii)
$$\lim_{x\to c} rf(x) = r \lim_{x\to c} f(x) \text{ for } r \in \mathbb{R}.$$

(iii)

$$\lim_{x\to c}(fg)(x) = (\lim_{x\to c} f(x))(\lim_{x\to c} g(x)),$$

(iv)

$$\lim_{x\to c}\left(\frac{1}{f}\right)(x) = \frac{1}{\lim_{x\to c} f(x)} \quad (\textit{if denominator} \neq 0).$$

PROOF. Follows from Lemma 2.13 for sequences. □

Lemma 3.24 (Sandwich lemma). *If $f(x) \leq g(x) \leq h(x)$, and $\lim_{x\to c} f(x) = \ell$ and $\lim_{x\to c} h(x) = \ell$, then $\lim_{x\to c} g(x) = \ell$.*

PROOF. Follows from sandwich Lemma 2.15 for sequences. □

3.2.3. Continuity and limit. We say $c \in \mathbb{R}$ is an *interior point* of $A \subseteq \mathbb{R}$ if there is $r > 0$ such that $(c - r, c + r) \subseteq A$.

Proposition 3.25. *Let $f : A \to \mathbb{R}$, and c be an interior point of A. Then f is continuous at c iff $\lim_{x\to c} f(x)$ exists and is equal to $f(c)$.*

PROOF IDEA. We use characterization of continuity given by Proposition 3.8. Forward implication is straightforward. For backward implication: Let $x_n \to c$. Break $\{x_n\}$ into two subsequences: One contains terms not equal to c, and other contains terms equal to c. Both subsequences, after applying f, converge to $f(c)$. Hence, $f(x_n) \to f(c)$, as required. (Ignore either of the two subsequences if it is finite.) □

3.2.4. Left and right limits.

Definition 3.26. We build on Definition 3.20.

(i) We say $\lim_{x\to c^-} f(x)$ exists if there is $\ell \in \mathbb{R}$ such that for every sequence $\{x_n\}$ in A with $x_n < c$ and $x_n \to c$, we have $f(x_n) \to \ell$. In this case, we say f has a left limit at c.

(ii) We say $\lim_{x\to c^+} f(x)$ exists if there is $\ell \in \mathbb{R}$ such that for every sequence $\{x_n\}$ in A with $x_n > c$ and $x_n \to c$, we have $f(x_n) \to \ell$. In this case, we say f has a right limit at c.

Proposition 3.27. *We have: f has a limit at c iff f has a left limit and right limit at c, and they are equal.*

3.2.5. Types of discontinuities. Suppose $f : A \to \mathbb{R}$ is discontinuous at an interior point $c \in A$. Then one of the following happens.

- $\lim_{x\to c} f(x)$ does not exist.
  - Either left limit or right limit of $f(x)$ at c does not exist (essential discontinuity).
  - Left and right limits of $f(x)$ at c exist, but are not equal (jump discontinuity).
- $\lim_{x\to c} f(x)$ exists, but is not equal to $f(c)$ (removable discontinuity).

3.2.6. Convergence to and at infinity of a function. We mention that it is possible to make sense of the limits

$$\lim_{x\to\infty} f(x) = \ell, \qquad \lim_{x\to-\infty} f(x) = \ell,$$

and also of

$$\lim_{x\to c} f(x) = \infty, \qquad \lim_{x\to c} f(x) = -\infty.$$

The latter two can also be applied to left and right limits.

For example,

$$\lim_{x\to\infty} \frac{1}{x} = 0, \quad \lim_{x\to-\infty} \frac{1}{x} = 0, \quad \lim_{x\to 0^+} \frac{1}{x} = \infty, \quad \lim_{x\to 0^-} \frac{1}{x} = -\infty.$$

Remark 3.28 (Metric spaces). We build on Remark 2.23. Let X and Y be metric spaces. It makes sense to define a continuous function $f : X \to Y$ as in Definition 3.1, with $|x - c|$ replaced by $\operatorname{dist}(x, c)$ (distance in X), and $|f(x) - f(c)|$ replaced by $\operatorname{dist}(f(x), f(c))$ (distance in Y). For the example of $f : \mathbb{R}^2 \to \mathbb{R}$, see Definition 6.12.

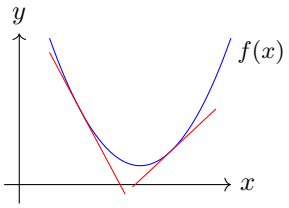
An even more general context for continuous functions is that of topological spaces (in which there is a qualitative rather than quantitative notion of what it means for two points to be close to each other). For more details, see Munkres [20, Chapter 2].

CHAPTER 4

Differentiability

4.1. Differentiability

The intuitive idea of a differentiable function f is that the graph of f has tangents which are not vertical (that is, of finite slope). See illustration below. We now formalize this notion.



4.1.1. Differentiable functions. Let $A \subseteq \mathbb{R}$, and c be an interior point of A.

Definition 4.1. A function $f : A \to \mathbb{R}$ is *differentiable* at c if the limit

$$\lim_{h \to 0} \frac{f(c+h) - f(c)}{h}$$

exists. We denote it by $f'(c)$, and call it the *derivative* of f at c.

Equivalently, a function $f : A \to \mathbb{R}$ is *differentiable* at c if there is a real number α such that

$$\lim_{h \to 0} \frac{f(c+h) - f(c) - \alpha h}{h} = 0. \tag{4.1}$$

In this case, we say α is the *derivative* of f at c. Note: One may also replace h by $|h|$ in the denominator in (4.1).

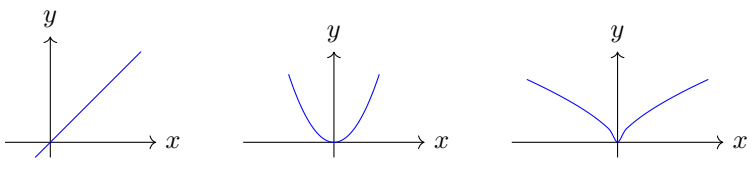
Example 4.2. Let us illustrate the notion of differentiability.

(1) Let $f : \mathbb{R} \to \mathbb{R}$ be a constant function. Then f is differentiable and $f'(c) = 0$ for all $c \in \mathbb{R}$.

(2) Let $f_1, f_2, f_3 : \mathbb{R} \to \mathbb{R}$ be

$$f_1(x) = x, \qquad f_2(x) = x^2, \qquad f_3(x) = x^{\frac{2}{3}}.$$

Their graphs are shown below.



We have:

- f_1 is differentiable and $f_1'(c) = 1$ for all $c \in \mathbb{R}$.
- f_2 is differentiable and $f_2'(c) = 2c$ for all $c \in \mathbb{R}$.
- f_3 is differentiable at $c \neq 0$, but, it is not differentiable at 0 since

$$\frac{f_3(0+h) - f_3(0)}{h} = \frac{1}{h^{1/3}}$$

whose limit does not exist as $h \to 0$.

(3) Let $f(0) = 0$ and $f(x) = x \sin(1/x)$ for $x \in \mathbb{R} \setminus \{0\}$. Then f is not differentiable at 0 since

$$\frac{f(0+h) - f(0)}{h} = \sin\big(\frac{1}{h}\big)$$

whose limit does not exist as $h \to 0$.

4.1.2. Left and right derivatives. Let $f : A \to \mathbb{R}$.

(i) Suppose $c \in A$ is such that $[c, c+r) \subseteq A$ for some $r > 0$. If the limit

$$\lim_{h \to 0^+} \frac{f(c+h) - f(c)}{h}$$

exists, then we call it the *right derivative* of f at c, and denote it by $f_+'(c)$.

(ii) Suppose $c \in A$ is such that $(c-r, c] \subseteq A$ for some $r > 0$. If the limit

$$\lim_{h \to 0^-} \frac{f(c+h) - f(c)}{h}$$

exists, then we call it the *left derivative* of f at c, and denote it by $f_-'(c)$.

Lemma 4.3. *If c is an interior point of A, then $f : A \to \mathbb{R}$ is differentiable at c iff $f_+'(c)$ and $f_-'(c)$ both exist and are equal.*

Example 4.4. Let $f(x) = |x|$. Then $f_-'(0) = -1$ and $f_+'(0) = 1$. Hence f is not differentiable at 0.

4.1.3. Derivative function. Let us now focus on the case when the domain of f is an interval I.

We say $f : (a, b) \to \mathbb{R}$ is differentiable on (a, b) if f is differentiable at every $c \in (a, b)$. In this case, define

$$f' : (a, b) \to \mathbb{R}, \qquad c \mapsto f'(c).$$

We call f' the *derivative* of f. We make a similar definition when the domain of f is (a, ∞), $(-\infty, b)$, $\mathbb{R}$.

We say $f : [a, b] \to \mathbb{R}$ is differentiable on $[a, b]$ if f is differentiable on (a, b), and $f_+'(a)$ and $f_-'(b)$ exist. In this case, define

$$f' : [a, b] \to \mathbb{R}, \qquad a \mapsto f_+'(a),\ c \mapsto f'(c),\ b \mapsto f_-'(b)$$

for $c \in (a, b)$. We make a similar definition when the domain of f is $[a, b)$, $(a, b]$, $[a, \infty)$, $(-\infty, b]$.

4.1.4. Increment function.

Lemma 4.5 (Caratheodory lemma). *A function $f : A \to \mathbb{R}$ is differentiable at an interior point c of A iff there is a function $f_1 : A \to \mathbb{R}$ which is continuous at c such that*

$$f(x) - f(c) = (x - c)f_1(x)$$

for $x \in A$. Moreover, $f'(c) = f_1(c)$.

We call $f_1 : A \to \mathbb{R}$ the *increment function*. Note very carefully: f_1 depends on the point c.

Proof. We make use of Proposition 3.25.

Forward implication. Let f be differentiable at c. Define

$$f_1(x) := \begin{cases} \frac{f(x)-f(c)}{x-c} & \text{if } x \in A \setminus \{c\}, \\ f'(c) & \text{if } x = c. \end{cases}$$

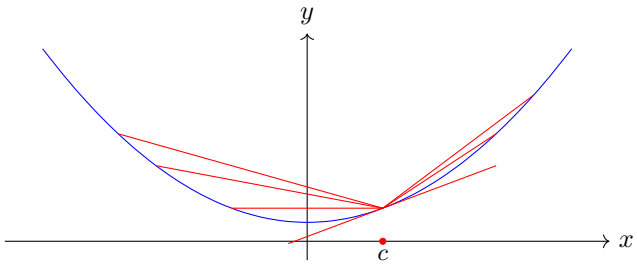
Then f_1 is continuous at c since $\lim_{x\to c} f_1(x) = f'(c) = f_1(c)$.

Backward implication. Let f_1 be as stated. Then

$$\lim_{h\to 0} \frac{f(c+h) - f(c)}{h} = \lim_{h\to 0} f_1(c+h) = \lim_{x\to c} f_1(x) = f_1(c)$$

since f_1 is continuous at c. Hence f is differentiable at c. □

In other words, the increment function f_1 keeps track of slopes of all secants drawn from $(c, f(c))$. More precisely, $f_1(x)$ is the slope of the line segment joining $(c, f(c))$ to $(x, f(x))$ for $x \neq c$, and $f_1(c)$ is the slope of the tangent line at $(c, f(c))$.



Corollary 4.6. *If f is differentiable at c, then f is continuous at c.*

Proof. Let f be differentiable at c. Using Caratheodory Lemma 4.5, write

$$f(x) = f(c) + (x - c)f_1(x).$$

Since f_1 is continuous, so is f by Lemma 3.4. Alternatively,

$$\lim_{x\to c} f(x) = \lim_{x\to c} f(c) + (x - c)f_1(x) = f(c)$$

by Lemma 3.23. Now use Proposition 3.25. □

Remark 4.7. If f is not continuous at c, then it is not differentiable at c. For example: The function $f(x) = [x]$ is not continuous at 5, hence it is not differentiable at 5.

The converse of Corollary 4.6 is false. For example: The function $f(x) = |x|$ is continuous at 0, but it is not differentiable at 0.

Remark 4.8. Here is an alternative way to phrase Caratheodory lemma. A function $f : A \to \mathbb{R}$ is differentiable at an interior point c of A iff there is a real number α such that

$$f(c+h) = f(c) + \alpha\, h + \epsilon(h)\, h$$

where $\epsilon(h)$ is defined for small h, and $\epsilon(h) \to 0$ as $h \to 0$. Moreover, $f'(c) = \alpha$.

4.1.5. Algebra of differentiable functions. The operations of addition, multiplication, scalar multiplication on functions are compatible with the notion of differentiability in the following sense.

Lemma 4.9. *Suppose $f, g : A \to \mathbb{R}$ are differentiable at $c \in A$. Then*

(i) *$f + g$ is differentiable at c, and*

$$(f+g)'(c) = f'(c) + g'(c),$$

(ii) *rf is differentiable at c, and*

$$(rf)'(c) = rf'(c)$$

for $r \in \mathbb{R}$,

(iii) *fg is differentiable at c, and*

$$(fg)'(c) = f'(c)g(c) + f(c)g'(c),$$

(iv) *$1/f$ is differentiable at c, and*

$$(1/f)'(c) = \frac{-f'(c)}{f(c)^2}$$

if $f(c) \neq 0$.

Proof. For item (i): Write

$$f(x) = f(c) + (x-c)f_1(x) \quad \text{and} \quad g(x) = g(c) + (x-c)g_1(x).$$

Then

$$f(x) + g(x) = f(c) + g(c) + (x-c)[f_1(x) + g_1(x)].$$

Thus,

$$(f+g)(x) = (f+g)(c) + (x-c)(f_1 + g_1)(x).$$

Since f_1 and g_1 are both continuous at c, so is $f_1 + g_1$. It serves as the increment function for $f + g$ at the point c. Thus by Caratheodory Lemma 4.5, $f + g$ is differentiable at c. Moreover,

$$(f+g)'(c) = (f+g)_1(c) = (f_1 + g_1)(c) = f_1(c) + g_1(c) = f'(c) + g'(c).$$

Proofs of items (ii), (iii), (iv) use similar ideas. □

Lemma 4.10 (Chain rule). *Let $f : A \to B$ and $g : B \to \mathbb{R}$. Let c be an interior point of A, and $f(c)$ be an interior point of B. If f is differentiable at c, and g is differentiable at $f(c)$, then the composite $g \circ f : A \to \mathbb{R}$ is differentiable at c, and*

$$(g \circ f)'(c) = g'(f(c))f'(c). \tag{4.2}$$

Proof. Exercise. □

Example 4.11. Let $\varphi(x) = (4x^3+3)^7 + 2$. Define $f(x) = 4x^3+3$ and $g(y) = y^7 + 2$. Then $\varphi = g \circ f$. Hence,

$$\varphi'(c) = g'(f(c))f'(c) = 7(4c^3+3)^6(12c^2).$$

Lemma 4.12. *Let $f : (a,b) \to (p,q)$ be continuous, and a bijection. Let $f^{-1} : (p,q) \to (a,b)$ be the inverse function. Let f be differentiable at $c \in (a,b)$, and $f'(c) \neq 0$. Then f^{-1} is differentiable at $f(c) \in (p,q)$, and*

$$(f^{-1})'(f(c)) = \frac{1}{f'(c)}.$$

PROOF. Put $g = f^{-1}$. Then $\varphi = g \circ f$ is the identity function on (a,b). By the chain rule, $1 = \varphi'(c) = g'(f(c))f'(c)$. Therefore, $g'(f(c)) = 1/f'(c)$. □

Draw a picture.

Example 4.13. Let us illustrate Lemma 4.12.

(1) Let

$$f : (-\frac{\pi}{2}, \frac{\pi}{2}) \to (-1,1), \qquad f(x) = \sin(x).$$

Then f is continuous and a bijection. Its inverse function is denoted $\sin^{-1}$. Put $f(c) = d$. Thus,

$$(\sin^{-1})'(d) = (f^{-1})'(d) = \frac{1}{f'(c)} = \frac{1}{\cos(c)} = \frac{1}{\sqrt{1-\sin^2 c}} = \frac{1}{\sqrt{1-d^2}}.$$

(2) Fix a positive integer $n \geq 1$. Let

$$f : (0,\infty) \to (0,\infty), \qquad f(x) = x^n.$$

Then f is continuous and a bijection. Put $f(c) = d$. Thus,

$$(f^{-1})'(d) = \frac{1}{f'(c)} = \frac{1}{nc^{n-1}} = \frac{1}{nd^{(n-1)/n}} = \frac{1}{n}d^{(1/n)-1}.$$

Remark 4.14. The derivative of a trigonometric function is again a trigonometric function. However, the derivative of an inverse trigonometric function is algebraic involving rational functions and square roots. This is because the relations among different trigonometric functions are algebraic, and usually quadratic. For instance, in the above calculation of the derivative of $\sin^{-1}$, we used the quadratic relation $\sin^2\theta + \cos^2\theta = 1$.

4.2. Maxima and minima

The derivative provide an effective tool to solve maxima and minima (optimization) problems. Conversely, one can use these ideas to prove results about the derivative such that the mean value theorem. This establishes a clear connection between sign of the first derivative and increasing/decreasing functions. Going one step further, there is a connection between sign of the second derivative and convex/concave functions.

4.2.1. Global and local maxima/minima. Let $f : A \to \mathbb{R}$ be a function.

Definition 4.15. We say:

(i) f has a *global maximum* at c if $f(x) \leq f(c)$ for $x \in A$. In this case, $f(c)$ is the least upper bound of f, and it is attained at c.
(ii) f has a *global minimum* at c if $f(x) \geq f(c)$ for $x \in A$. In this case, $f(c)$ is the greatest lower bound of f, and it is attained at c.

Definition 4.16. We say:

(i) f has a *local maximum* at c if there is $\delta > 0$ such that $|x-c| < \delta$ implies $f(x) \leq f(c)$.
(ii) f has a *local minimum* at c if there is $\delta > 0$ such that $|x-c| < \delta$ implies $f(x) \geq f(c)$.

Note: Global maximum (minimum) implies local maximum (minimum), but the converse is false.

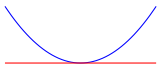
Note: A constant function has both a global maximum and a global minimum at all points.

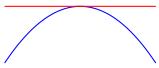
We say f has a global (local) *extremum* at c if it has either a global (local) maximum at c, or a global (local) minimum at c.

4.2.2. Local maxima/minima: necessary condition.

Lemma 4.17. *Let c be an interior point of A. If $f : A \to \mathbb{R}$ is differentiable at c, and has either a local maximum or a local minimum at c, then $f'(c) = 0$.*

See illustrations below.





PROOF. Suppose f has a local minimum at c. Thus, for small h, $f(c+h) - f(c) \geq 0$.

$$h > 0 : \quad \frac{f(c+h)-f(c)}{h} \geq 0. \qquad \text{Hence, } f'_+(c) \geq 0.$$

$$h < 0 : \quad \frac{f(c+h)-f(c)}{h} \leq 0. \qquad \text{Hence, } f'_-(c) \geq 0.$$

Now f differentiable at c implies $f'_+(c) = f'_-(c) = f'(c) = 0$.

Equivalently, we may argue using the increment function: $f_1(c+h) \geq 0$ for $h > 0$, and $f_1(c+h) \leq 0$ for $h < 0$. And $f_1(x)$ is continuous at c, so $f_1(c) = 0$.

We can make a similar argument when f has a local maximum at c. □

Remark 4.18. We make some remarks related to the above result.

(1) Let $f : [-1, 1] \to \mathbb{R}$ with $f(x) = x^2$. Then f has a local minimum at the interior point 0, and indeed $f'(0) = 0$ as claimed by Lemma 4.17.
(2) Let $f : [0, 1] \to \mathbb{R}$ with $f(x) = x$. Then f has a local minimum at 0 and local maximum at 1. But $f'_+(0) \neq 0$ and $f'_-(1) \neq 0$. This does not contradict Lemma 4.17 since 0 and 1 are not interior points.
(3) Let $f : [-1, 1] \to \mathbb{R}$ with $f(x) = x^3$. Then $f'(0) = 0$, but f does not have a local maximum or a local minimum at 0. Thus, the converse to Lemma 4.17 is false.

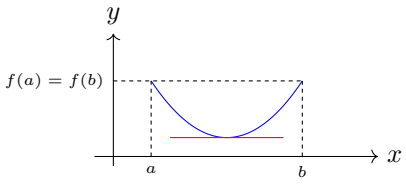
4.2.3. Rolle's theorem and mean value theorem. We now discuss Rolle's theorem and the mean value theorem. The former is a special case of the latter. The latter is attributed to Lagrange.

Theorem 4.19 (Rolle's theorem). *Let $f : [a, b] \to \mathbb{R}$ be such that*

(i) *f is continuous on $[a, b]$,*
(ii) *f is differentiable on (a, b),*
(iii) *$f(a) = f(b)$.*

Then there is $c \in (a, b)$ such that $f'(c) = 0$.

See illustration below.



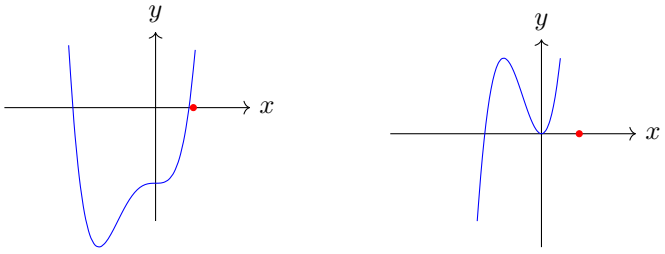
Proof. We consider two cases.

- f is constant. Then $f'(c) = 0$ for all $c \in (a, b)$.
- f is not a constant. Then the global minimum of f is strictly smaller than the global maximum of f. Since f is continuous, by Theorem 3.15, both are attained on $[a, b]$. Both cannot be at a and b since $f(a) = f(b)$. Hence, there is $c \in (a, b)$ such that f has either a global maximum or a global minimum at c. Global maximum/minimum implies local maximum/minimum, so by Lemma 4.17, $f'(c) = 0$.

□

Example 4.20. Let us return to Example 3.11. We saw by IVP that the function $f(x) = x^4 + 2x^3 - 2$ has a root in $(0, 1)$. Now let us show that $f(x) = x^4 + 2x^3 - 2$ has exactly one root in $(0, 1)$.

Suppose there are two roots in $(0, 1)$. Say $f(a) = 0 = f(b)$ for $0 < a < b < 1$. Then by Rolle's theorem, $f'(c) = 0$ for some $c \in (a, b)$. Now $f'(x) = 4x^3 + 6x^2 = 2x^2(2x + 3) \neq 0$ for $x \in (0, 1)$. This is a contradiction.



The graphs of f and f' are shown above. The red point is $x = 1$.

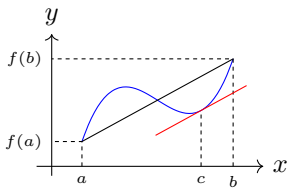
Theorem 4.21 (Mean value theorem). *Let $f : [a, b] \to \mathbb{R}$ be such that*

(i) *f is continuous on $[a, b]$,*
(ii) *f is differentiable on (a, b).*

Then there is $c \in (a,b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

See illustration below.



PROOF. For $x \in [a,b]$, define

$$F(x) := f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Then F is continuous on $[a,b]$, differentiable on (a,b) and $F(a) = f(a) = F(b)$. By Rolle's theorem, there is $c \in (a,b)$ such that $F'(c) = 0$, that is, $f'(c) = \frac{f(b)-f(a)}{b-a}$. □

Remark 4.22 (Physical interpretation). Let $f(t)$ denote the displacement of a particle at time t for $a \le t \le b$. Then the average speed is $\frac{f(b)-f(a)}{b-a}$, and speed at time c is $f'(c)$. Thus, MVT says that there is a time c such that the speed at time c equals the average speed.

Remark 4.23. Note very carefully: Rolle's theorem and the mean value theorem are results about the derivative, and make no direct reference to the notions of minima and maxima. Then why are they in this section, and not in Section 4.1? The reason is that the proof of Rolle's theorem uses a result about minima and maxima.

Rolle's theorem is a corollary of the mean value theorem obtained by imposing the additional hypothesis $f(a) = f(b)$. Then why is it stated earlier rather than later? The reason is that Rolle's theorem is used in the proof of the mean value theorem.

4.2.4. Mean value inequality.

Lemma 4.24. *Let $f : [a,b] \to \mathbb{R}$ be such that f is continuous on $[a,b]$, and differentiable on (a,b). If $m \le f'(x) \le M$ for all $x \in (a,b)$, then*

$$m(b - a) \le f(b) - f(a) \le M(b - a).$$

This is the *mean value inequality*.

PROOF. This follows from Theorem 4.21 (MVT). □

Example 4.25. Fix n. Define $f : [n, n+1] \to \mathbb{R}$ by $f(x) = \sqrt{x}$. Then $f'(x) = \frac{1}{2\sqrt{x}}$. Moreover,

$$\frac{1}{2\sqrt{n+1}} \le f'(x) \le \frac{1}{2\sqrt{n}}.$$

Therefore, by the mean value inequality,

$$\frac{1}{2\sqrt{n+1}}(n+1-n) \leq \sqrt{n+1}-\sqrt{n} \leq \frac{1}{2\sqrt{n}}(n+1-n).$$

For $n=1$, we get $\frac{1}{2\sqrt{2}} \leq \sqrt{2}-1 \leq \frac{1}{2}$. Therefore, $\sqrt{2} < \frac{3}{2}$. To get a lower bound, we use $\frac{1}{\sqrt{2}} > \frac{2}{3}$. So $\frac{1}{2}\frac{2}{3} < \sqrt{2}-1$ which yields $\frac{4}{3} < \sqrt{2}$. Thus,

$$\frac{4}{3} < \sqrt{2} < \frac{3}{2}.$$

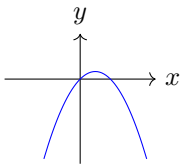
4.2.5. Increasing and decreasing functions.

Lemma 4.26. *Let $f : [a,b] \to \mathbb{R}$ be such that f is continuous on $[a,b]$, and differentiable on (a,b).*

(1) *If $f'(x) = 0$ for $x \in (a,b)$, then f is constant on $[a,b]$. (Converse true).*
(2) (i) *If $f'(x) \geq 0$ for $x \in (a,b)$, then f is increasing on $[a,b]$. (Converse true).*
(ii) *If $f'(x) \leq 0$ for $x \in (a,b)$, then f is decreasing on $[a,b]$. (Converse true).*
(iii) *If $f'(x) > 0$ for $x \in (a,b)$, then f is strictly increasing on $[a,b]$. (Converse false).*
(iv) *If $f'(x) < 0$ for $x \in (a,b)$, then f is strictly decreasing on $[a,b]$. (Converse false).*

PROOF. These can be deduced from Theorem 4.21 (MVT). □

Example 4.27. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x(1-x)$. Its graph is shown below.



Then $f'(x) = 1-2x$. Thus, $f'(x) > 0$ if $x < \frac{1}{2}$, and $f'(x) < 0$ if $x > \frac{1}{2}$. So, f is strictly increasing on $(-\infty, \frac{1}{2})$, and strictly decreasing on $(\frac{1}{2}, \infty)$.

4.2.6. Convex functions.

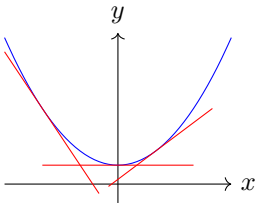
Recall convex functions from Section 1.2.10. We now relate them to differentiability.

Lemma 4.28. *Let I be an interval and $f : I \to \mathbb{R}$ be differentiable. Then*

(i) *f' is increasing on I iff f is convex on I.*
(ii) *f' is decreasing on I iff f is concave on I.*
(iii) *f' is strictly increasing on I iff f is strictly convex on I.*
(iv) *f' is strictly decreasing on I iff f is strictly concave on I.*

PROOF. See [12, Proposition 4.31]. Note: Items (i) and (ii) imply each other, while items (iii) and (iv) imply each other. □

An illustration of item (i) is shown below.



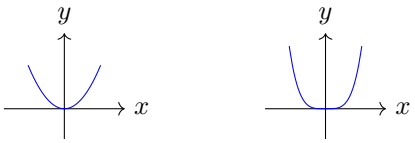
Note how the slopes of the tangents increase as we move from left to right.

Corollary 4.29. *Let I be an interval and $f : I \to \mathbb{R}$ be twice differentiable. Then*

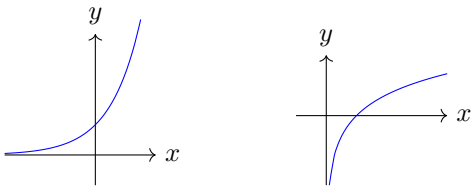
(i) *$f'' \geq 0$ on I iff f is convex on I.*
(ii) *$f'' \leq 0$ on I iff f is concave on I.*
(iii) *If $f'' > 0$ on I, then f is strictly convex on I.*
(iv) *If $f'' < 0$ on I, then f is strictly concave on I.*

Example 4.30. This result gives a test for convexity as illustrated below.

(1) The function $f(x) = x^2$ is strictly convex since $f''(x) = 2 > 0$ at all points. Its graph is shown below on the left. The function $f(x) = x^4$ is convex since $f''(x) = 12x^2 \geq 0$. Its graph is shown below on the right. In fact, it is strictly convex, even though the second derivative is not strictly positive at all points.



(2) The exponential function $f(x) = e^x$ is strictly convex since $f''(x) = e^x > 0$. Its graph is shown below on the left. The logarithm function $f(x) = \log x$ is strictly concave since $f''(x) = -1/x^2 < 0$. Its graph is shown below on the right.



4.2.7. Critical points and global maxima/minima. Let $f : A \to \mathbb{R}$. An interior point c of A is a *critical point* of f if either f is not differentiable at c, or if f is differentiable at c and $f'(c) = 0$.

Lemma 4.31. *Let $f : [a, b] \to \mathbb{R}$ be continuous. Then the global minimum and global maximum of f are attained at points which are either critical points of f or endpoints of $[a, b]$.*

PROOF. Suppose $f(c)$ is a global maximum. We consider two cases.

- $c = a$ or $c = b$. Then c is an endpoint of $[a, b]$.

- $c \in (a,b)$. We consider two subcases.
  - f is not differentiable at c. Then c is a critical point of f.
  - f is differentiable at c. Since f has a global maximum at c, it has a local maximum at c. Hence $f'(c) = 0$ by Lemma 4.17, and c is a critical point of f.

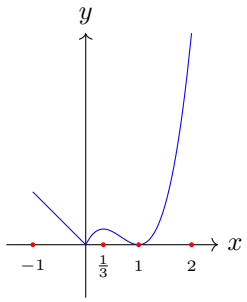
The argument for a global minimum is similar. □

Thus, to find the global maximum and global minimum of f, we first find the critical points of f. Then we evaluate f at these critical points, and at endpoints of $[a,b]$. Among these values, the largest is the global maximum, and smallest is the global minimum.

Example 4.32. Let $f : [-1,2] \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} -x & \text{if } -1 \le x \le 0, \\ 2x^3 - 4x^2 + 2x & \text{if } 0 \le x \le 2. \end{cases}$$

It is continuous everywhere. Its graph is shown below.



Observe

$$f'(x) = \begin{cases} -1 & \text{if } -1 \le x < 0, \\ 2(3x-1)(x-1) & \text{if } 0 < x \le 2. \end{cases}$$

Note: f is not differentiable at $x = 0$. Also, $f'(x) = 0$ iff $x = \frac{1}{3}$ or $x = 1$. So $x = 0, \frac{1}{3}, 1$ are the critical points. Let us tabulate f at the critical points, and at the endpoints:

x	-1	0	$\frac{1}{3}$	1	2
$f(x)$	1	0	$\frac{8}{27}$	0	4

Thus, we see f has global maximum 4 attained at $x = 2$, and f has global minimum 0 attained at $x = 0$ and $x = 1$.

4.2.8. Local maxima/minima: sufficient conditions. Let $f : A \to \mathbb{R}$. Let c be an interior point of A with $(c-\delta, c+\delta) \subseteq A$.

Lemma 4.33 (First derivative test). *Let f be differentiable on $(c-\delta, c)$ and $(c, c+\delta)$. Then:*

(i) *If $f' \ge 0$ on $(c-\delta, c)$ and $f' \le 0$ on $(c, c+\delta)$, and f is continuous at c, then f has a local maximum at c.*

(ii) *If $f' \le 0$ on $(c-\delta, c)$ and $f' \ge 0$ on $(c, c+\delta)$, and f is continuous at c, then f has a local minimum at c.*

PROOF. For item (i): Since $f' \geq 0$ on $(c-\delta, c)$, f is increasing on $(c-\delta, c)$. Similarly, since $f' \leq 0$ on $(c, c+\delta)$, f is decreasing on $(c, c+\delta)$. Finally, since f is continuous at c, we deduce that $f(c) \geq f(x)$ for $x \in (c-\delta, c+\delta)$.

Argument for item (ii) is similar. □

Lemma 4.34 (Second derivative test). *Let f be differentiable in an open interval containing c, and twice differentiable at c.*

(i) *If $f'(c) = 0$ and $f''(c) < 0$, then f has a local maximum at c.*
(ii) *If $f'(c) = 0$ and $f''(c) > 0$, then f has a local minimum at c.*

PROOF. For item (i):

$$f''(c) = \lim_{h \to 0} \frac{f'(c+h) - f'(c)}{h} = \lim_{h \to 0} \frac{f'(c+h)}{h} < 0.$$

Thus, there is $\delta > 0$ such that

$$f'(c+h) < 0 \quad \text{for} \quad 0 < h < \delta,$$

and

$$f'(c+h) > 0 \quad \text{for} \quad -\delta < h < 0.$$

So f has a local maximum at c by Lemma 4.33.

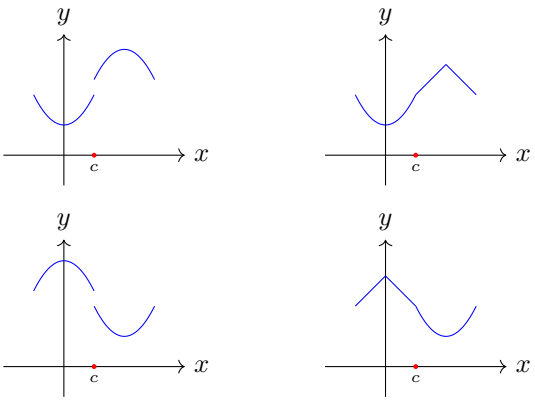
Argument for item (ii) is similar. □

4.2.9. Points of inflection. Let c be an interior point of an interval I, and let $f : I \to \mathbb{R}$.

Definition 4.35. We say c is a *point of inflection* for f if for some $\delta > 0$,

(i) either f is convex on $(c-\delta, c)$ and concave on $(c, c+\delta)$,
(ii) or f is concave on $(c-\delta, c)$ and convex on $(c, c+\delta)$.

See illustrations below. In the first two pictures, item (i) holds, while in the next two pictures, item (ii) holds.



Remark 4.36. We make some remarks related to the notion of an inflection point.

(1) There are two different kinds of inflection points, depending on whether item (i) holds or item (ii) holds. However, we use the same terminology for both. In the setting of maximum/minimum, we indeed use two

different terminologies. To refer to them together, we use the term extremum.

(2) The definition of inflection point c does not put any condition on $f(c)$. In particular, f is not required to be continuous or differentiable at c (or at any other point).

(3) A linear function is both convex and concave, so every point of a linear function is an inflection point.

Recall: We have related the notions of convexity and differentiability. This leads to the following.

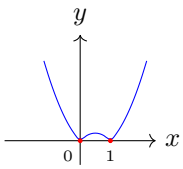
Lemma 4.37 (Derivative tests).

(i) *Suppose f is differentiable in an open interval containing c, except possibly at c. Then c is a point of inflection for f iff f' is increasing on $(c-\delta, c)$ and decreasing on $(c, c+\delta)$, or vice versa, for some $\delta > 0$.*

(ii) *Suppose f is twice differentiable in an open interval containing c, except possibly at c. Then c is a point of inflection for f iff $f'' \geq 0$ on $(c-\delta, c)$ and $f'' \leq 0$ on $(c, c+\delta)$, or vice versa, for some $\delta > 0$.*

PROOF. Item (i) follows from Lemma 4.28. Similarly, item (ii) follows from Corollary 4.29. □

Example 4.38. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = |x(1-x)|$. Its graph is shown below. Compare with the graph in Example 4.27.



In the interval $(0, 1)$, $f'(x) = 1 - 2x$ is strictly decreasing. In the intervals $(-\infty, 0)$ and $(1, \infty)$, $f'(x) = 2x - 1$ is strictly increasing. Thus, f is strictly concave in $(0, 1)$, and strictly convex in $(-\infty, 0)$ and $(1, \infty)$. It has inflection points at 0 and 1.

Exercise 4.39. Check: $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x|x|$ is strictly concave in $(-\infty, 0)$ and strictly convex in $(0, \infty)$, and has a point of inflection at $x = 0$.

Lemma 4.40 (Necessary condition). *Let f be twice differentiable at a point of inflection c for f. Then $f''(c) = 0$.*

PROOF. The hypothesis implies that f is differentiable in an open interval containing c. By Lemma 4.37, item (i), f' is increasing on $(c-\delta, c)$ and decreasing on $(c, c+\delta)$, or vice versa, for some $\delta > 0$. Let us assume the former, the latter is similar. Since f' is differentiable at c, it is continuous at c. So f' has a local maximum at c. Hence, by Lemma 4.17 applied to f', we get $f''(c) = 0$. □

The condition $f''(c) = 0$ is not sufficient to have an inflection point. For example: Take $f(x) = x^4$. Then $f''(0) = 0$, and f has a local minimum at $x = 0$ (not an inflection point).

Lemma 4.41 (Sufficient condition). *Suppose f is thrice differentiable at c. If $f''(c) = 0$ and $f'''(c) \neq 0$, then c is a point of inflection for f.*

PROOF. The hypothesis implies that f is twice differentiable in an open interval containing c. Suppose $f'''(c) < 0$.

$$f'''(c) = \lim_{h\to 0} \frac{f''(c+h) - f''(c)}{h} = \lim_{h\to 0} \frac{f''(c+h)}{h} < 0.$$

Thus, there is $\delta > 0$ such that

$$f''(c+h) < 0 \quad \text{for} \quad 0 < h < \delta,$$

and

$$f''(c+h) > 0 \quad \text{for} \quad -\delta < h < 0.$$

So f has a point of inflection at c by Lemma 4.37, item (ii).

The case $f'''(c) > 0$ is similar. □

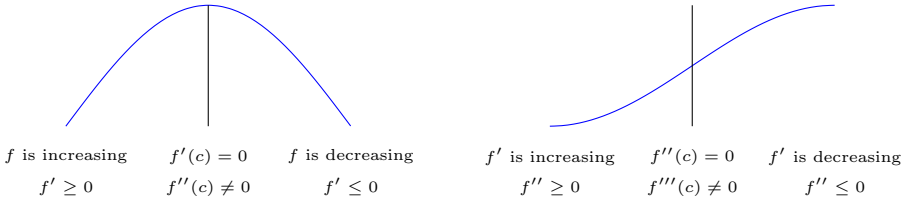
Example 4.42. Take $f(x) = x^3$. Then $f''(0) = 0$ and $f'''(0) = 6 \neq 0$, and hence f has an inflection point at $x = 0$.

However, the condition $f'''(c) \neq 0$ is not necessary to have an inflection point. For example: Take $f(x) = x^5$. Then $f'''(0) = 0$, and yet f has an inflection point at $x = 0$. We may also take $f(x) = x$. Then $f'''(c) = 0$ for all c, and yet each c is an inflection point for f.

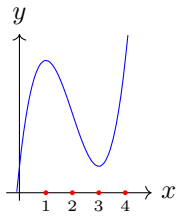
TABLE 4.1. Comparison of local extremum and inflection point.

	local extremum for f at c	inflection point for f at c
necesary condition	$f'(c) = 0$ (Lemma 4.17)	$f''(c) = 0$ (Lemma 4.40)
sufficient condition	$f'(c) = 0$ and $f''(c) \neq 0$ (Lemma 4.34)	$f''(c) = 0$ and $f'''(c) \neq 0$ (Lemma 4.41)

A comparison of local extremum and inflection point (under suitable differentiability hypothesis) is shown in Table 4.1. An illustration of the same is shown below.



Example 4.43. Let $f : [-.1, 4.1] \to \mathbb{R}$ with $f(x) = x^3 - 6x^2 + 9x + 1$. Its graph is shown below.



Note:

$$f'(x) = 3(x-3)(x-1), \quad f''(x) = 6(x-2), \quad f'''(x) = 6.$$

It is easy to determine the intervals where f', f'', f''' are positive, zero, negative. Also, $f(0) = 1$, $f(1) = 5$, $f(2) = 3$, $f(3) = 1$, $f(4) = 5$. We deduce that f has a local maximum at 1, a local minimum at 3, a point of inflection at 2. Further, it has a global minimum at the left endpoint $-.1$, and a global maximum at the right endpoint 4.1.

For any function f, it is easier to answer questions about its local maximum/minimum and inflection points if we have a general idea of how the graph of f looks. In this regard, note that Example 4.43 fits into a class of functions called polynomials. The graphs of polynomials are briefly discussed in Section 1.2.8.

4.2.10. Asymptotes. There are three types of asymptotes to a function $f : A \to \mathbb{R}$, namely, horizontal, vertical, oblique. Let us go over them one by one.

(i) The line $y = b$ is a *horizontal asymptote* to $y = f(x)$ if

$$\lim_{x\to-\infty} f(x) = b \qquad \text{or} \qquad \lim_{x\to\infty} f(x) = b.$$

These correspond, respectively, to the two pictures below.

(ii) The line $x = a$ is a *vertical asymptote* to $y = f(x)$ if

$$\lim_{x\to a^-} f(x) = \pm\infty \qquad \text{or} \qquad \lim_{x\to a^+} f(x) = \pm\infty.$$

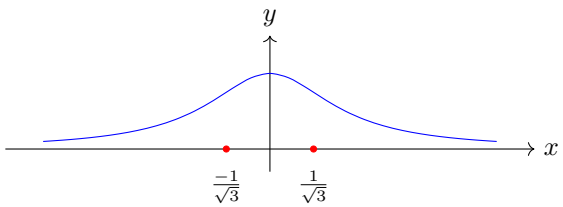
These correspond, respectively, to the two pictures below. In both cases, the limit is ∞. Similar pictures can be drawn when the limit is $-\infty$.

(iii) The line $y = ax + b$ is an *oblique asymptote* to $y = f(x)$ if

$$\lim_{x\to-\infty} [f(x) - (ax+b)] = 0 \qquad \text{or} \qquad \lim_{x\to\infty} [f(x) - (ax+b)] = 0.$$

These correspond, respectively, to the two pictures below.

Example 4.44. Let $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = \frac{1}{1+x^2}$. Its graph is shown below.

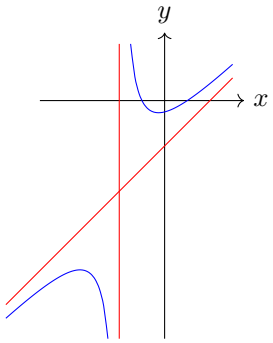


The line $y = 0$ is a horizontal asymptote. Note:

$$f'(x) = \frac{-2x}{(1+x^2)^2} \quad \text{and} \quad f''(x) = \frac{2(3x^2-1)}{(1+x^2)^3}.$$

The function f has a local maximum at 1. It is convex in the intervals $(-\infty, \frac{-1}{\sqrt{3}})$ and $(\frac{1}{\sqrt{3}}, \infty)$, and concave in the interval $(\frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$.

Example 4.45. Let $f : \mathbb{R} \setminus \{-2\} \to \mathbb{R}$ with $f(x) = \frac{x^2-1}{x+2}$. Its graph is shown below.



Note:

$$f'(x) = \frac{(x+2+\sqrt{3})(x+2-\sqrt{3})}{(x+2)^2} \quad \text{and} \quad f''(x) = \frac{6}{(x+2)^3}.$$

We deduce:

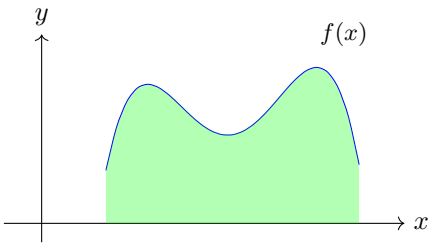
- f is increasing on $(-\infty, -2-\sqrt{3})$ and $(-2+\sqrt{3}, \infty)$, and decreasing on $(-2-\sqrt{3}, -2)$ and $(-2, -2+\sqrt{3})$.
- f has a local maximum at $-2-\sqrt{3}$ and local minimum at $-2+\sqrt{3}$.
- f is concave on $(-\infty, -2)$ and convex on $(-2, \infty)$.
- f has no point of inflection.
- The line $x = -2$ is a vertical asymptote, and $y = x - 2$ is an oblique asymptote.

CHAPTER 5

Integration

5.1. Riemann integral

The intuitive idea of the integral of a function f is the area under the graph of f. See illustration below. We now formalize this notion.



5.1.1. Riemann integrable functions. Let $a < b$ be real numbers, and $f : [a, b] \to \mathbb{R}$ be a bounded function. Let M be the global maximum of f, and m the global minimum of f. A partition P of $[a, b]$ is a sequence of points

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

We write

$$P = \{x_0 < x_1 < \cdots < x_{n-1} < x_n\}.$$

The norm of partition P is defined as

$$\|P\| := \max\{x_i - x_{i-1} \colon 1 \le i \le n\}.$$

For each $1 \le i \le n$, let M_i be the global maximum of f on $[x_{i-1}, x_i]$, and m_i the global minimum of f on $[x_{i-1}, x_i]$. Let

$$U(P, f) = \sum_{i=1}^{n} M_i(x_i - x_{i-1}) \quad \text{and} \quad L(P, f) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}).$$

We call $U(P, f)$ the *upper sum*, and $L(P, f)$ the *lower sum*. Then

$$m(b - a) \le L(P, f) \le U(P, f) \le M(b - a).$$

Draw a picture.

Definition 5.1. A bounded function $f : [a, b] \to \mathbb{R}$ is *Riemann integrable* if there is a sequence $\{P_n\}$ of partitions of $[a, b]$ such that

$$U(P_n, f) - L(P_n, f) \to 0 \quad \text{as} \quad n \to \infty.$$

5.1.2. Riemann integral. Let $f : [a, b] \to \mathbb{R}$ be Riemann integrable.

Proposition 5.2. *There is a real number A such that*

$$L(P, f) \leq A \leq U(P, f)$$

for every partition P of $[a, b]$, and

$$\lim_{n\to\infty} L(P_n, f) = A = \lim_{n\to\infty} U(P_n, f)$$

for every sequence $\{P_n\}$ of partitions of $[a, b]$ with $\|P_n\| \to 0$.

The proof uses the completeness property of $\mathbb{R}$, and is omitted. We write

$$\int_a^b f(x)\,dx = A,$$

and call it the *Riemann integral* of f.

5.1.3. Riemann sums. For P a partition of $[a, b]$, let

$$S(P, f) = \sum_{i=1}^{n} f(c_i)(x_i - x_{i-1})$$

where $c_i \in [x_{i-1}, x_i]$ for $1 \leq i \leq n$. We call $S(P, f)$ a *Riemann sum*. Note very carefully: Along with P and f, a Riemann sum $S(P, f)$ depends on the choice of points c_i.

Observe:

$$L(P, f) \leq S(P, f) \leq U(P, f).$$

In words, any Riemann sum lies between the lower and upper sums.

Proposition 5.3. *Suppose*

- *f is Riemann integrable on $[a, b]$,*
- *$\{P_n\}$ is a sequence of partitions of $[a, b]$ with $\|P_n\| \to 0$,*
- *$S(P_n, f)$ is any Riemann sum for P_n and f.*

Then

$$S(P_n, f) \to \int_a^b f(x)\,dx$$

as $n \to \infty$.

PROOF. By Proposition 5.2, sequences $L(P_n, f)$ and $U(P_n, f)$ have the same limit, namely, $\int_a^b f(x)\,dx$. Now apply sandwich Lemma 2.15 to $L(P_n, f) \leq S(P_n, f) \leq U(P_n, f)$. □

5.1.4. Domain additivity.

Lemma 5.4. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function, and let $c \in (a, b)$. Then f is Riemann integrable on $[a, b]$ iff f is Riemann integrable on $[a, c]$ and on $[c, b]$. In this case,*

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

PROOF. The proof is straightforward. For details, see [12, Proposition 6.7]. □

Convention:

- For $a = b$, we set $\int_a^b f(x)\,dx = 0$.
- For $b < a$, we set $\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx$.

5.1.5. Monotone functions. Recall monotonic functions from Definition 1.7.

Lemma 5.5. *If $f : [a, b] \to \mathbb{R}$ is monotonic, then f is Riemann integrable.*

Draw a picture.

PROOF. Let $P = \{x_0 < x_1 < \cdots < x_{n-1} < x_n\}$ be a partition of $[a, b]$. Suppose f is increasing on $[a, b]$. Then $M_i = f(x_i)$ and $m_i = f(x_{i-1})$ for $1 \le i \le n$. Hence

$$U(P, f) - L(P, f) = \sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))(x_i - x_{i-1}) \le \|P\|(f(b) - f(a)).$$

Now take any sequence $\{P_n\}$ with $\|P_n\| \to 0$. (For instance, P_n partitions $[a, b]$ into n equal parts.) Then

$$U(P_n, f) - L(P_n, f) \le \|P_n\|(f(b) - f(a)) \to 0 \quad \text{as} \quad n \to \infty.$$

So f is Riemann integrable.

The case when f is decreasing on $[a, b]$ is similar. □

Example 5.6. Let us illustrate Lemma 5.5.

(1) The function $f : [-1, 1] \to \mathbb{R}$ given by $f(x) = x^{\frac{2}{3}}$ is Riemann integrable. To see this,
   - f is decreasing in $[-1, 0]$, and hence Riemann integrable on $[-1, 0]$,
   - f is increasing in $[0, 1]$, and hence Riemann integrable on $[0, 1]$.

   By Lemma 5.4, f is Riemann integrable on $[-1, 1]$.
(2) The integer part function $f(x) = [x]$ on $[a, b]$ is increasing, and hence Riemann integrable.

5.1.6. Continuous functions.

Lemma 5.7. *If $f : [a, b] \to \mathbb{R}$ is bounded, and has at most finitely many discontinuities, then f is Riemann integrable.*

In particular, if $f : [a, b] \to \mathbb{R}$ is continuous, then f is Riemann integrable.

PROOF. We omit the proof. The idea is to use that any continuous function on $[a, b]$ is uniformly continuous. For details, see [1, Theorem 3.14] or [12, Proposition 6.9, item (ii)]. □

Example 5.8. Let us illustrate Lemma 5.7.

(1) Any polynomial function $p : [a, b] \to \mathbb{R}$ is continuous, and hence Riemann integrable.
(2) The functions $f : [a, b] \to \mathbb{R}$ with $f(x) = \sin x$ or $f(x) = \cos x$ are continuous, and hence Riemann integrable.
(3) The function $f : [-1, 1] \to \mathbb{R}$ given by $f(x) = x^{\frac{2}{3}}$ is continuous, and hence Riemann integrable.
(4) The integer part function $f(x) = [x]$ on $[a, b]$ has only finitely many discontinuities, and hence it is Riemann integrable.

5.1.7. Algebra of Riemann integrable functions. The operations of addition, multiplication, scalar multiplication on functions are compatible with the notion of Riemann integrability in the following sense.

Lemma 5.9. *Suppose $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable. Then*

(i) *$f + g$ is Riemann integrable, and*

$$\int_a^b (f+g)(x)\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx,$$

(ii) *rf is Riemann integrable, and*

$$\int_a^b (rf)(x)\,dx = r\int_a^b f(x)\,dx$$

for $r \in \mathbb{R}$,

(iii) *fg is Riemann integrable,*

(iv) *$1/f$ is Riemann integrable if there is $\delta > 0$ such that $|f(x)| \geq \delta$ for $x \in [a, b]$ (so that $1/f$ is bounded).*

Proof. We omit the proof. See [12, Proposition 6.15]. □

5.1.8. Further properties of the Riemann integral.

Lemma 5.10. *For $a < b$, suppose $f, g : [a, b] \to \mathbb{R}$ are Riemann integrable.*

- *If $f(x) \leq g(x)$, then $\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx$.*

In particular:

- *If $f(x) \geq 0$, then $\int_a^b f(x)\,dx \geq 0$.*

Proof. The particular case is clear. We get the general case by applying the particular case to $g - f$, and using Lemma 5.9. □

Lemma 5.11. *Suppose $f : [a, b] \to \mathbb{R}$ is Riemann integrable. Then so is $|f|$, and moreover*

$$\Big|\int_a^b f(x)\,dx\Big| \leq \int_a^b |f(x)|\,dx.$$

Proof. Observe: $0 \leq U(P, |f|) - L(P, |f|) \leq U(P, f) - L(P, f)$. Now apply sandwich Lemma 2.15. □

The converse of Lemma 5.11 is false. In other words, $|f|$ Riemann integrable does not imply f Riemann integrable. For example, take f to be the function

$$f : [0, 1] \to \mathbb{R}, \qquad f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q}, \\ -1 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q}. \end{cases} \tag{5.1}$$

It is not Riemann integrable because $L(P, f) = -1$ and $U(P, f) = 1$ for any partition P of $[0, 1]$. In contrast, $|f|$ is the constant function 1, so it is clearly Riemann integrable.

Exercise 5.12. Define the function $f : [0, 1] \to \mathbb{R}$ by

$$f(x) = \begin{cases} 0 & \text{if } x \text{ is irrational}, \\ 1/q & \text{if } x = p/q \text{ in lowest terms}. \end{cases} \tag{5.2}$$

Show: f is Riemann integrable and $\int_0^1 f(x)\,dx = 0$. This gives an example of a Riemann integrable function with infinitely many points of discontinuity (by Exercise 3.18).

5.1.9. Application: computing limits. It is possible to evaluate limits of certain sequences by interpreting their terms as Riemann sums for a suitable function over a suitable interval. Let us illustrate this principle.

Example 5.13. Suppose we want to find the limit of the sequence

$$a_n = \sum_{i=1}^{n} \frac{n}{n^2 + i^2}.$$

For that, rewrite a_n as

$$a_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + (\frac{i}{n})^2}.$$

Now define

$$f : [0,1] \to \mathbb{R}, \qquad f(x) = \frac{1}{1+x^2}.$$

Then f is decreasing, and f is continuous. So either by Lemma 5.5 or by Lemma 5.7, f is Riemann integrable.

Now let P_n be the partition of $[0,1]$ defined by $x_i = \frac{i}{n}$ for $0 \le i \le n$. Take $c_i = \frac{i}{n}$. The key observation is that $a_n = S(P_n, f)$, the Riemann sum for P_n and f. Since $\|P_n\| \to 0$,

$$a_n = S(P_n, f) \longrightarrow \int_0^1 \frac{1}{1+x^2}\,dx = \tan^{-1}(x)\Big|_0^1 = \frac{\pi}{4} - 0 = \frac{\pi}{4}.$$

Thus, $a_n \to \frac{\pi}{4}$.

Remark 5.14 (Lebesgue integration). In these notes, we give several applications of Riemann integration. But it also suffers from many drawbacks. To overcome these, it has now been replaced by Lebesgue integration. The main difference between the two is the following.

- To define Riemann integral of f, one divides the domain of f into small pieces.
- To define Lebesgue integral of f, one divides the codomain of f into small pieces. This necessitates the concept of a measure, and of measurable spaces.

For more details on Lebesgue integration, see Royden [23, Chapters 3,4,5], Rudin [25, Chapters 1,2,3]. For an exposition directed towards probability theory, see Billingsley [5, Chapters 2 and 3].

5.2. Fundamental theorem of calculus

Differentiation and integration are inverse processes. This contains two statements, namely, $\int F' = F$ and $(\int f)' = f$. The fundamental theorem of calculus makes these two statements precise.

5.2.1. FTC. Part I. This pertains to integration followed by differentiation.

Theorem 5.15. *Let f be Riemann integrable on $[a,b]$. For $x \in [a,b]$, define*

$$F(x) := \int_a^x f(t)\,dt. \tag{5.3}$$

Then F is continuous on $[a,b]$. Moreover, if f is continuous at $c \in [a,b]$, then F is differentiable at c, and $F'(c) = f(c)$.

PROOF. Let us prove the second part. So let f be continuous at c. By domain additivity (Lemma 5.4),

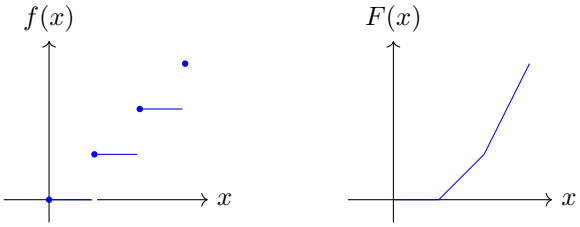
$$\frac{F(c+h) - F(c)}{h} = \frac{1}{h}\int_c^{c+h} f(t)\,dt.$$

Therefore,

$$\begin{aligned}\Big|\frac{F(c+h) - F(c)}{h} - f(c)\Big| &= \Big|\frac{1}{h}\int_c^{c+h} f(t) - f(c)\,dt\Big| \\ &\le \frac{1}{|h|}\,|h|\,\max\{|f(t) - f(c)| \colon |t-c| \le h\}.\end{aligned}$$

This goes to 0 as $h \to 0$ since f is continuous at c. □

Example 5.16. Let $f : [0,3] \to \mathbb{R}$ be the integer part function given by $f(x) = [x]$. Now define $F : [0,3] \to \mathbb{R}$ by formula (5.3) with $a = 0$. It is a piecewise linear function. The graphs of f and F are shown below.



Note very carefully: f is not continuous at 1 and 2. But F is continuous at all points. However, F is not differentiable at 1 and 2.

Example 5.17. Let $f : [-1,1] \to \mathbb{R}$ be the absolute value function given by $f(x) = |x|$. Now define $F : [-1,1] \to \mathbb{R}$ by formula (5.3) with $a = -1$. Explicitly,

$$F(x) = \begin{cases} \frac{1-x^2}{2} & \text{if } -1 \le x \le 0, \\ \frac{1+x^2}{2} & \text{if } 0 < x \le 1. \end{cases}$$

The graphs of f and F are shown below.

Note very carefully: f is not differentiable at 0. However, f is continuous at 0, and hence F is differentiable at 0.

5.2.2. FTC. Part II. This pertains to differentiation followed by integration.

Theorem 5.18. *Let f be Riemann integrable on $[a,b]$. Let F be continuous on $[a,b]$, differentiable on (a,b), and $F'=f$ on (a,b). Then*

$$\int_a^b f(x)\,dx = F(b)-F(a). \tag{5.4}$$

PROOF. For $P=\{x_0<x_1<\cdots<x_{n-1}<x_n\}$ a partition of $[a,b]$,

$$\begin{aligned} F(b)-F(a) &= \sum_{i=1}^{n} F(x_i)-F(x_{i-1}) \\ &= \sum_{i=1}^{n} F'(c_i)(x_i-x_{i-1}) \\ &= \sum_{i=1}^{n} f(c_i)(x_i-x_{i-1}) \\ &= S(P,f). \end{aligned}$$

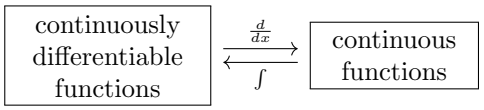
In the second step, we used mean value Theorem 4.21.

Thus, for P any partition, $F(b)-F(a)$ equals a Riemann sum associated to P and f. Now take any sequence $\{P_n\}$ with $\|P_n\|\to 0$. Then

$$F(b)-F(a)=S(P_n,f)\longrightarrow \int_a^b f(x)\,dx$$

by Proposition 5.3. □

In view of Theorems 5.15 and 5.18, we get the following correspondence.



5.2.3. Integration by parts.

Proposition 5.19. *Let f,g be differentiable on $[a,b]$, and f',g' be Riemann integrable on $[a,b]$. Then*

$$\int_a^b f(x)g'(x)\,dx = f(b)g(b)-f(a)g(a)-\int_a^b f'(x)g(x)\,dx. \tag{5.5}$$

PROOF. Put $h:=fg$. Then by chain rule, $h'=fg'+f'g$. Thus, h' is Riemann integrable by Lemma 5.9. Now integrate and use Theorem 5.18. □

5.2.4. Integration by substitution.

Proposition 5.20. *Consider functions $[\alpha,\beta]\xrightarrow{\varphi}[a,b]\xrightarrow{f}\mathbb{R}$. Suppose*

- *φ is differentiable on $[\alpha,\beta]$,*
- *φ' is Riemann integrable on $[\alpha,\beta]$,*
- *f is continuous on $[a,b]$.*

Then $(f \circ \varphi)\,\varphi'$ is Riemann integrable on $[\alpha, \beta]$, and

$$\int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,dx = \int_{\alpha}^{\beta} f(\varphi(t))\varphi'(t)\,dt. \tag{5.6}$$

PROOF. Since f and φ are continuous, their composite $f \circ \varphi$ is continuous, and hence Riemann integrable. By hypothesis, φ' is Riemann integrable. Hence, the product $(f \circ \varphi)\,\varphi'$ is Riemann integrable.

Define $F : [a, b] \to \mathbb{R}$ by

$$F(x) := \int_a^x f(u)\,du.$$

Define $H : [\alpha, \beta] \to \mathbb{R}$ as the composite $H = F \circ \varphi$. By chain rule and Theorem 5.15,

$$H'(t) = F'(\varphi(t))\,\varphi'(t) = f(\varphi(t))\varphi'(t).$$

Integrating,

$$\begin{aligned}\int_{\alpha}^{\beta} f(\varphi(t))\varphi'(t)\,dt = \int_{\alpha}^{\beta} H'(t)\,dt &= H(\beta) - H(\alpha) \\ &= F(\varphi(\beta)) - F(\varphi(\alpha)) = \int_{\varphi(\alpha)}^{\varphi(\beta)} f(x)\,dx.\end{aligned}$$

We used Theorem 5.18 in the second step and in the last step. □

5.3. Defining functions using the Riemann integral

We can use the Riemann integral to construct functions of a real variable such as $\log x$, $\sin x$, and so on, and prove their basic properties. Usually, these functions are introduced informally; their graphs are recalled in Section 1.2.

5.3.1. Logarithmic function. Define the *logarithmic function*

$$\log : (0, \infty) \to \mathbb{R}$$

by

$$\log x := \int_1^x \frac{1}{t}\,dt.$$

Note: $f(t) = \frac{1}{t}$ is continuous on $(0, \infty)$, and bounded on $[1, x]$ or on $[x, 1]$ (depending on whether $x \geq 1$ or $x \leq 1$). So the Riemann integral exists by Lemma 5.7.

Proposition 5.21. *The log function satisfies the following properties.*

(i) $\log x > 0$ *for* $x > 1$, $\quad \log 1 = 0$, $\quad \log x < 0$ *for* $0 < x < 1$.
(ii) log *is differentiable and* $(\log)'(x) = 1/x$.
(iii) log *is strictly increasing and strictly concave.*
(iv) $\log(xy) = \log(x) + \log(y)$ *for* $x, y > 0$.
(v) $\log x \to \infty$ *as* $x \to \infty$, *and* $\log x \to -\infty$ *as* $x \to 0^+$.
(vi) log *is bijective.*

PROOF.

- Item (i) is clear since $1/t > 0$.

- For item (ii): Use Theorem 5.15.
- For item (iii):
$$(\log)'(x) > 0 \quad \text{and} \quad (\log)''(x) = -1/x^2 < 0.$$
Now use Lemma 4.26 and Corollary 4.29.
- For item (iv): Let $f(x) = \log(xy) - \log(x)$, with y fixed. Then $f'(x) = \frac{y}{xy} - \frac{1}{x} = 0$. Therefore, f is a constant function with $f(1) = \log(y)$. Hence, $\log(xy) = \log(x) + \log(y)$. Alternatively, one can use substitution to show
$$\int_y^{xy} \frac{1}{t}\, dt = \int_1^x \frac{1}{t}\, dt.$$
- For item (v): $\log 2^n = n \log 2$ and $\log(1/x) = -\log x$.
- For item (vi): log is strictly increasing, and hence injective. By item (v) and IVP, we see that it is also surjective.

□

5.3.2. Exponential function. Define the *exponential function*
$$\exp : \mathbb{R} \to (0, \infty)$$
as the inverse of the logarithmic function $\log : (0, \infty) \to \mathbb{R}$. Thus,
$$\exp x = y \iff \log y = x.$$

Proposition 5.22. *The exp function satisfies the following properties.*

(i)
$$\exp x > 0 \text{ for all } x, \qquad \exp 0 = 1.$$
(ii) exp *is differentiable and* $(\exp)'(x) = \exp x$.
(iii) exp *is strictly increasing and strictly convex.*
(iv) $\exp(x + y) = (\exp x)(\exp y)$ *for all* x, y.
(v) $\exp x \to \infty$ *as* $x \to \infty$, *and* $\exp x \to 0$ *as* $x \to -\infty$.
(vi) $\exp : \mathbb{R} \to (0, \infty)$ *is bijective.*

Proof.

- Item (i) is clear.
- For item (ii): If $x = \log y$, then $(\exp)'(x) = 1/(\log)'(y) = y = \exp x$.
- For item (iii): $(\exp)'x > 0$ and $(\exp)''x > 0$.
- For item (iv): Applying *exp* to $\log(xy) = \log(x) + \log(y)$ yields $xy = \exp(\log(x) + \log(y))$. Now put $\log(x) = a$ and $\log(y) = b$, to get $\exp(a + b) = (\exp a)(\exp b)$.
- Items (v) and (vi) follow from corresponding properties of log.

□

Put $e := \exp(1)$. It is called the *Euler number*. Note: $\log e = 1$.

5.3.3. Real powers of positive real numbers. Let $a > 0$. Define
$$a^x := \exp(x \log a).$$
In particular,
$$e^x = \exp(x \log e) = \exp x.$$
Thus $a^x = e^{x \log a}$.

5.3.4. Inverse trigonometric and trigonometric functions. Define

$$\tan^{-1} : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$$

by

$$\tan^{-1} x := \int_0^x \frac{1}{1+t^2}\, dt.$$

Note: $f(t) = \frac{1}{1+t^2}$ is continuous on $\mathbb{R}$, and bounded on $[0, x]$ or on $[x, 0]$ (depending on whether $x \geq 0$ or $x \leq 0$). So the Riemann integral exists by Lemma 5.7. .

We can then establish properties of the function $\tan^{-1}$, namely,

- an expression for its derivative,
- that it is strictly increasing,
- that it is strictly concave on $(-\frac{\pi}{2}, 0)$ and strictly convex on $(0, \frac{\pi}{2})$.

This is similar to how we established properties of log.

Next we can define the function

$$\tan : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$$

as the inverse of $\tan^{-1}$, and establish its properties.

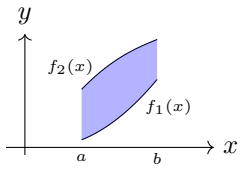
Once we have the tan function, we can use it to define the sin and cos functions, and so on. For more details, see [12, Chapter 7].

5.4. Lengths, areas, volumes

We have an intuitive understanding of lengths, areas and volumes. However, it is important that we give formal definitions of these concepts. We will do this in Chapter 8 when we study the Riemann integral in higher dimensions. However for the present, we proceed intuitively, and explain how the Riemann integral (in one dimension) can be used to compute lengths of curves, areas of plane regions or surfaces, volumes of solids.

5.4.1. Areas between curves. We discuss three cases of computing areas between curves depending on the integration variable.

(1) Integration variable is x. Suppose $f_1, f_2 : [a, b] \to \mathbb{R}$ are Riemann integrable, and $f_1 \leq f_2$. Let R be the region bounded by curves $y = f_1(x)$ and $y = f_2(x)$ and lines $x = a$ and $x = b$. See picture below.

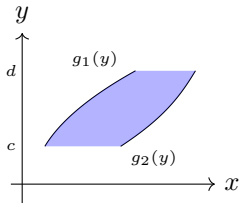


Then the area of R is

$$\text{Area}(R) = \int_a^b (f_2(x) - f_1(x))\, dx. \tag{5.7}$$

In the special case when $f_1 = 0$, we get the area below the curve $f_2(x)$ between $x = a$ and $x = b$.

(2) Integration variable is y. Suppose $g_1, g_2 : [c, d] \to \mathbb{R}$ are Riemann integrable, and $g_1 \leq g_2$. Let R be the region bounded by curves $x = g_1(y)$ and $x = g_2(y)$ and lines $y = c$ and $y = d$. See picture below.

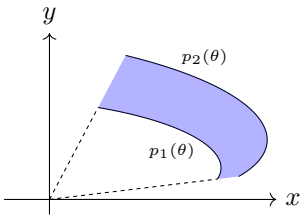


Then the area of R is

$$\text{Area}(R) = \int_c^d (g_2(y) - g_1(y))\, dy. \tag{5.8}$$

In the special case when $g_1 = 0$, we get the area below the curve $g_2(y)$ between $y = c$ and $y = d$.

(3) Integration variable is angle θ. Suppose $p_1, p_2 : [\alpha, \beta] \to \mathbb{R}$ are Riemann integrable, and $p_1 \leq p_2$. Let R be the region bounded by curves $r = p_1(\theta)$ and $r = p_2(\theta)$ and rays $\theta = \alpha$ and $\theta = \beta$. See picture below.



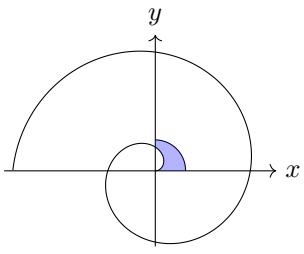
Then the area of R is

$$\text{Area}(R) = \frac{1}{2} \int_\alpha^\beta (p_2(\theta)^2 - p_1(\theta)^2)\, d\theta. \tag{5.9}$$

In the special case when $p_1 = 0$, we get the area below the curve $p_2(\theta)$ between $\theta = \alpha$ and $\theta = \beta$.

For some worked out examples of formulas (5.7) and (5.8), see [1, Section 2.3] or [12, Examples 8.1].

Example 5.23. Let R be the region between the circle $r = 2$ and spiral $r = \theta$ for $0 \leq \theta \leq \pi/2$. See picture below.



Then

$$\text{Area}(R) = \frac{1}{2}\int_0^{\pi/2} (2^2 - \theta^2)\, d\theta = \pi - \frac{\pi^3}{48}.$$

Note: $\theta < 2$ for $0 \le \theta \le \pi/2$, that is, the spiral lies inside the circle.

Let us now consider a more general setup. We elaborate using item (1). Suppose we want to find the area between arbitrary curves $y = f_1(x)$ and $y = f_2(x)$. In this case, we separate the calculation in two parts, namely,

$$\text{Area}(R) = \int\limits_{f_1 \le f_2} (f_2(x) - f_1(x))\, dx + \int\limits_{f_2 \le f_1} (f_1(x) - f_2(x))\, dx, \tag{5.10}$$

depending on which curve lies above the other.

5.4.2. Volumes of solids. Let D denote a solid. We discuss three methods to compute its volume.

Slice method. Let $A(x_0)$ be the area of the slice of the solid D by the plane perpendicular to the x-axis at x_0. Then the volume of D is given by

$$\text{Vol}(D) = \int_a^b A(x)\, dx, \tag{5.11}$$

assuming D lies between the planes $x = a$ and $x = b$.

Example 5.24. Let us find the volume of the solid D enclosed by the cylinders $x^2 + y^2 = a^2$ and $x^2 + z^2 = a^2$. Note:

$$-a \le x \le a \quad \text{and} \quad -\sqrt{a^2 - x^2} \le y, z \le \sqrt{a^2 - x^2}.$$

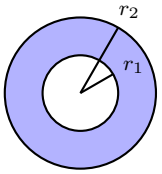
Thus, each slice is a square of side $2\sqrt{a^2 - x^2}$. Therefore,

$$A(x) = (2\sqrt{a^2 - x^2})(2\sqrt{a^2 - x^2}) = 4(a^2 - x^2).$$

Hence,

$$\text{Vol}(D) = 4\int_{-a}^{a} (a^2 - x^2)\, dx = 8\int_0^a (a^2 - x^2)\, dx = 8(a^3 - \frac{a^3}{3}) = \frac{16}{3}a^3.$$

Washer method. This is a special case of the slice method in which each slice of D is a washer with inner radius r_1, outer radius r_2, and area $\pi(r_2^2 - r_1^2)$. See picture below.



It is a disc if $r_1 = 0$, in which case, this method is called the disc method.

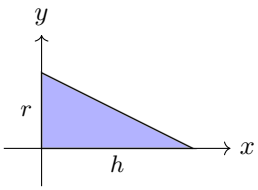
Let R be the region bounded by curves $y = f_1(x)$ and $y = f_2(x)$ (with $f_1 \le f_2$) and lines $x = a$ and $x = b$. Let D be the solid obtained by revolving R about the x-axis. Then the volume of D is given by

$$\text{Vol}(D) = \int_a^b \pi(f_2(x)^2 - f_1(x)^2)\, dx. \tag{5.12}$$

Remember in washer method: axis of revolution = variable of integration.

Example 5.25. Let D be the circular cone of radius r and height h. Place it so that the center of its base is at the origin, h is along the x-axis, and r is along the y-axis.

Let R be the triangular region bounded by lines $y = 0$, $y = r - \frac{rx}{h}$ and $x = 0$. See picture below.

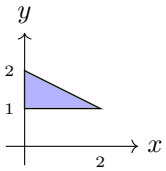


Then D is the solid obtained by revolving R about the x-axis. So

$$\text{Vol}(D) = \int_0^h \pi r^2 (1 - \frac{x}{h})^2 \, dx = \frac{1}{3}\pi r^2 h.$$

In this case, the slices are discs.

Example 5.26. As a variant of Example 5.25, let R be the triangular region bounded by lines $y = 1$, $y = 2 - \frac{x}{2}$ and $x = 0$. See picture below.



Let D be the solid obtained by revolving R about the x-axis. So

$$\text{Vol}(D) = \int_0^2 \pi[(2 - \frac{x}{2})^2 - 1^2] \, dx = \pi \int_0^2 (3 - 2x + \frac{x^2}{4}) \, dx$$
$$= \pi(3(2) - 2(2) + \frac{2}{3}) = \frac{8\pi}{3}.$$

In this case, the slices are no longer discs.

Shell method. Let R be the region bounded by curves $y = f_1(x)$ and $y = f_2(x)$ (with $f_1 \leq f_2$) and lines $x = a$ and $x = b$. Let D be the solid obtained by revolving R about the y-axis. Then the volume of D is given by

$$\text{Vol}(D) = \int_a^b 2\pi x(f_2(x) - f_1(x)) \, dx. \tag{5.13}$$

Remember in shell method: axis of revolution $\neq$ variable of integration.

Example 5.27. Let us go back to Example 5.25, and compute the volume of the cone by the shell method. We place the cone as before, and integrate along the y-axis. So we write the oblique line as $x = h(1 - \frac{y}{r})$. Thus,

$$\text{Vol}(D) = \int_0^r 2\pi y \, [h(1 - \frac{y}{r})] \, dy = 2\pi h \int_0^r y - \frac{y^2}{r} \, dy = \frac{1}{3}\pi r^2 h.$$

Origin of washers and shells. Consider a vertical line segment L in the positive quadrant of the xy-plane. In our setup, this is the portion between $f_1(x)$ and $f_2(x)$ (for a fixed x).

- If L is revolved about the x-axis, then it produces an annulus of inner radius $f_1(x)$ and outer radius $f_2(x)$.
- If L is revolved about the y-axis, then it produces a cylinder of radius x and height $f_2(x) - f_1(x)$.

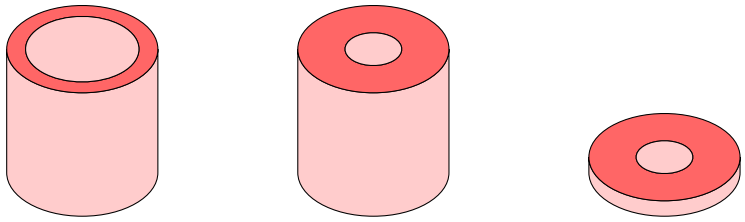
Now thicken L in the x-direction so that it becomes a thin rectangle R with a tiny breadth. In our setup, this is the portion between $f_1(x)$ and $f_2(x)$ multiplied with dx (for a fixed x).

- If R is revolved about the x-axis, then it produces a thickened annulus which is the same as a washer of thickness dx.
- If R is revolved about the y-axis, then it produces a thickened cylinder which is the same as a shell of thickness dx.

Integrating along the x-axis yields the respective volume formulas (5.12) and (5.13).

Shell vs washer, a picture perspective. Consider the thick cylinder shown in the middle picture below, whose thickness and height are of comparable size.

- Suppose we keep the height the same, and make the thickness very small, then we get the shell shown on the left.
- Suppose we keep the thickness the same, and make the height very small, then we get the washer shown on the right.



shell shell or washer? washer

5.4.3. Arc length of a parametrized curve. Let

$$C : [a, b] \to \mathbb{R}^2, \qquad C(t) = (x(t), y(t)),$$

where $x(t)$ and $y(t)$ are differentiable, and their derivatives are continuous on $[a, b]$. Then the length of the curve C is given by

$$\ell(C) = \int_a^b \sqrt{x'(t)^2 + y'(t)^2}\, dt. \tag{5.14}$$

ORIGIN OF THE FORMULA. Let $s, t \in [a, b]$ be 'close' to each other. Then

$$x(s) - x(t) \approx x'(t)(s - t) \quad \text{and} \quad y(s) - y(t) \approx y'(t)(s - t).$$

Length of the line segment joining $(x(s), y(s))$ and $(x(t), y(t))$ is

$$\sqrt{(x(s) - x(t))^2 + (y(s) - y(t))^2} \;\approx\; \sqrt{x'(t)^2 + y'(t)^2}\,(s - t)$$

which leads to (5.14). □

Special case. We highlight some special cases:

(1) Curve parameter t is the x variable.

$$C : [a, b] \to \mathbb{R}^2, \qquad C(x) = (x, f(x)), \qquad \ell(C) = \int_a^b \sqrt{1 + f'(x)^2}\, dx.$$

(2) Curve parameter t is the y variable.

$$C : [c, d] \to \mathbb{R}^2, \qquad C(y) = (g(y), y), \qquad \ell(C) = \int_c^d \sqrt{1 + g'(y)^2}\, dy.$$

(3) Curve parameter t is the angle θ variable. The curve $r = p(\theta)$ can be expressed as

$$C : [\alpha, \beta] \to \mathbb{R}^2, \qquad C(\theta) = (p(\theta)\cos\theta, p(\theta)\sin\theta).$$

So

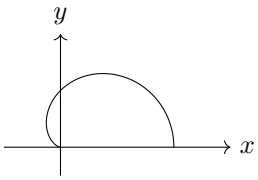
$$\begin{aligned}\ell(C) &= \int_\alpha^\beta \sqrt{(p'(\theta)\cos\theta - p(\theta)\sin\theta)^2 + (p'(\theta)\sin\theta + p(\theta)\cos\theta)^2}\, d\theta \\ &= \int_\alpha^\beta \sqrt{p(\theta)^2 + p'(\theta)^2}\, d\theta.\end{aligned}$$

Example 5.28. Consider the curve $C(x) = (x, x^2)$ for $0 \le x \le 1$. By our formula,

$$\ell(C) = \int_0^1 \sqrt{1 + 4x^2}\, dx = \frac{1}{2}\int_0^2 \sqrt{1 + u^2}\, du = \frac{1}{2}\sqrt{5} + \frac{1}{4}\log(2 + \sqrt{5}).$$

Recall: $\int_0^x \sqrt{1 + t^2}\, dt = \frac{1}{2}(x\sqrt{1 + x^2} + \log(x + \sqrt{1 + x^2}))$.

Example 5.29 (Cycloid). Consider the curve $r = p(\theta) = 1 + \cos\theta$ for $0 \le \theta \le \pi$. See picture below.



By our formula,

$$\begin{aligned}\ell(C) &= \int_0^\pi \sqrt{(1 + \cos\theta)^2 + (-\sin\theta)^2}\, d\theta \\ &= \int_0^\pi \sqrt{2(1 + \cos\theta)}\, d\theta \\ &= 2\int_0^\pi \cos\frac{\theta}{2}\, d\theta \\ &= 4.\end{aligned}$$

Fact: Formula (5.14) does not depend on the specific parametrization. Let us illustrate this fact by a simple example. Parametrize the semicircle of radius 1 between angles 0 and π in the standard way by

$$C : [0, \pi] \to \mathbb{R}, \qquad C(t) = (\cos t, \sin t).$$

Then

$$\ell(C) = \int_0^\pi \sqrt{(-\sin t)^2 + (\cos t)^2}\, dt = \int_0^\pi dt = \pi.$$

Now parametrize the same semicircle by

$$C' : [0, \pi/2] \to \mathbb{R}, \qquad C'(t) = (\cos 2u, \sin 2u).$$

Then

$$\ell(C') = \int_0^{\pi/2} \sqrt{(-2\sin 2u)^2 + (2\cos 2u)^2}\, du = 2\int_0^{\pi/2} du = \pi.$$

5.4.4. Area of surface of revolution. Consider a curve

$$C : [\alpha, \beta] \to \mathbb{R}^2, \qquad C(t) = (x(t), y(t)),$$

and a line L given by the equation $ax + by + c = 0$ with $a^2 + b^2 \neq 0$. Suppose C does not cross L. The distance of point $(x(t), y(t))$ from L is given by

$$\rho(t) = \frac{|a\, x(t) + b\, y(t) + c|}{\sqrt{a^2 + b^2}}.$$

Let S be the surface generated by revolving C about L. Then

$$\text{Area}(S) = 2\pi \int_\alpha^\beta \rho(t) \sqrt{x'(t)^2 + y'(t)^2}\, dt. \tag{5.15}$$

Special case. We highlight some special cases:

(1) Curve parameter t is the x variable. Let C be $y = f(x)$ with $x \in [a, b]$, and L be the line $y = 0$. Then

$$\text{Area}(S) = 2\pi \int_a^b |f(x)| \sqrt{1 + f'(x)^2}\, dx.$$

(2) Curve parameter t is the y variable. Let C be $x = g(y)$ with $y \in [c, d]$, and L be the line $x = 0$. Then

$$\text{Area}(S) = 2\pi \int_c^d |g(y)| \sqrt{1 + g'(y)^2}\, dy.$$

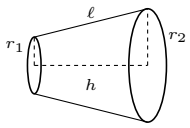
(3) Curve parameter t is the angle θ variable. Let C be $r = p(\theta)$ with $\theta \in [\alpha, \beta]$, and L be the line $\theta = \gamma$. Then

$$\text{Area}(S) = 2\pi \int_\alpha^\beta p(\theta) |\sin(\theta - \gamma)| \sqrt{p(\theta)^2 + p'(\theta)^2}\, d\theta.$$

Example 5.30 (Surface area of a cone and frustum). The surface area of a cone is $\pi r \ell$, where r is radius of the base, and ℓ is distance from the vertex to any point on the base.

More generally, the surface area of a frustum is $\pi(r_1 + r_2)\ell$, where r_1 and r_2 are radii of the two bases, and ℓ is distance between their boundary circles.

Let h be the height of the frustum. Thus, $\ell^2 = h^2 + (r_2 - r_1)^2$. See picture below.



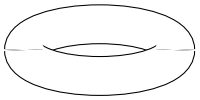
This area calculation fits into the special case of item (1) above. It is given below.

$$\begin{aligned}\text{Area}(S) &= 2\pi \int_0^h \left(r_1 + \frac{x}{h}(r_2 - r_1)\right) \sqrt{1 + \left(\frac{r_2 - r_1}{h}\right)^2}\, dx \\ &= 2\pi \frac{\ell}{h} \left(r_1 h + \frac{h^2}{2h}(r_2 - r_1)\right) \\ &= \pi(r_1 + r_2)\ell.\end{aligned}$$

Example 5.31 (Surface area of a torus). Let $0 < b < a$. Consider the circle $(x-a)^2 + y^2 = b^2$ of radius b with center at $(a, 0)$. We parametrize it as

$$C : [-\pi, \pi] \to \mathbb{R}^2, \qquad C(\theta) = (a + b\cos\theta, b\sin\theta).$$

Let S be the surface generated by revolving C about the y-axis. It is called a torus.



By formula (5.15),

$$\begin{aligned}\text{Area}(S) &= 2\pi \int_{-\pi}^{\pi} (a + b\cos\theta)\sqrt{(-b\sin\theta)^2 + (b\cos\theta)^2}\, d\theta \\ &= 2\pi b \int_{-\pi}^{\pi} (a + b\cos\theta)\, d\theta \\ &= (2\pi b)(2\pi a) \\ &= 4\pi^2 ab.\end{aligned}$$

We mention a theorem of Pappus. It says

$$\text{Area}(S) = \ell(C) \times d, \tag{5.16}$$

where d is the distance travelled by the centroid of C. In Example 5.31, $\ell(C) = 2\pi b$. The centroid of C is at $(a, 0)$, and it travels a distance of $2\pi a$. Thus, $\text{Area}(S) = (2\pi b)(2\pi a) = 4\pi^2 ab$.

Part II

Functions of several real variables

CHAPTER 6

Continuity

6.1. Real vector space

We take a very brief look at real vector spaces, and set up the required terminology and notations. For a more thorough introduction, see [1, Chapter 15], [2, Chapter 1].

6.1.1. Real vector space. Let $\mathbb{R}$ denote the set of real numbers. Any real number is called a *scalar*.

Now fix a positive integer m. Let $\mathbb{R}^m$ denote the set consisting of m-tuples

$$x = (x_1, \ldots, x_m),$$

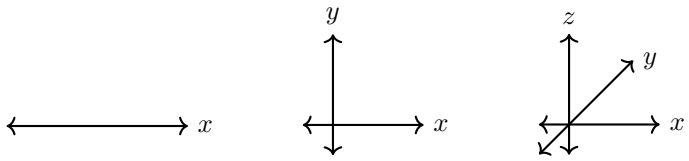
with each $x_i \in \mathbb{R}$, that is, each x_i is a scalar. Any element x of $\mathbb{R}^m$ is called a *vector*.

We can add vectors and scalar multiply them: For vectors $x, y \in \mathbb{R}^m$ and scalar $r \in \mathbb{R}$,

$$x + y = (x_1, \ldots, x_m) + (y_1, \ldots, y_m) = (x_1 + y_1, \ldots, x_m + y_m),$$
$$rx = r(x_1, \ldots, x_m) = (rx_1, \ldots, rx_m).$$

We refer to $\mathbb{R}^m$ as a *vector space*.

For $m = 1, 2, 3$, it is customary to visualize these vector spaces as follows.



6.1.2. Dot product. For vectors $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$, define

$$x \cdot y := \sum_{i=1}^{m} x_i y_i.$$

The scalar $x \cdot y$ is called the *dot product* of x and y. For example, in $\mathbb{R}^3$,

$$(1, 2, -1) \cdot (3, -1, -5) = 3 - 2 + 5 = 6,$$
$$(1, 2, -1) \cdot (3, -1, 5) = 3 - 2 - 5 = -4,$$
$$(1, -2, -1) \cdot (3, -1, 5) = 3 + 2 - 5 = 0.$$

Thus, the dot product can be positive, negative or zero. We say vectors x and y are *orthogonal* if $x \cdot y = 0$.

Lemma 6.1. *The dot product on $\mathbb{R}^m$ satisfies the following properties. It is*

- bilinear, *that is, it is linear in both coordinates:*
$$(rx + y) \cdot z = r(x \cdot z) + y \cdot z$$
$$x \cdot (ry + z) = r(x \cdot y) + x \cdot z.$$
- symmetric, *that is,* $x \cdot y = y \cdot x$,
- positive definite, *that is,*
$$x \cdot x = x_1^2 + \cdots + x_m^2 \geq 0,$$
and $x \cdot x = 0$ *iff* $x = 0$.

PROOF. Exercise. □

Due to the above properties, we say the dot product defines an *inner product* on $\mathbb{R}^m$.

6.1.3. Norm. Define the *norm* or *length* of $x = (x_1, \ldots, x_m)$ by
$$\|x\| := \sqrt{x \cdot x} = \sqrt{x_1^2 + \cdots + x_m^2}.$$
For example,
$$\|(1, 2, -1)\| = \sqrt{6}.$$

Lemma 6.2. *The norm on* $\mathbb{R}^m$ *satisfies the following properties.*
$$\|0\| = 0 \text{ and } \|x\| > 0 \text{ if } x \neq 0,$$
$$\|x + y\| \leq \|x\| + \|y\|,$$
$$\|rx\| = |r|\|x\|.$$

PROOF. Exercise. □

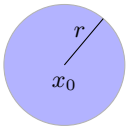
Due to the above properties, we say $\mathbb{R}^m$ is a *normed linear space*. For more details, see [16].

Note for carefully: For $m = 1$, norm is the same as absolute value, that is, $\|x\| = |x|$.

6.1.4. Ball around a point. For $x, y \in \mathbb{R}^m$, define the distance between x and y to be
$$\operatorname{dist}(x, y) := \|x - y\|. \tag{6.1}$$
Note how we used the linear structure of $\mathbb{R}^m$ to define distance. Properties of norm above translate to familiar properties of distance.

For $r > 0$ and $x_0 \in \mathbb{R}^m$, define the *open ball* of radius r around the point x_0 by
$$\begin{aligned} B(x_0, r) &:= \{x \in \mathbb{R}^m : \|x - x_0\| < r\} \\ &= \{x \in \mathbb{R}^m : \operatorname{dist}(x_0, x) < r\}. \end{aligned}$$
It consists of all points x whose distance from x_0 is strictly smaller than r.

This is illustrated below for $m = 2$.



You can imagine a similar picture for $m = 3$ consisting of points inside a sphere of radius r with center x_0. For $m = 1$, $B(x_0, r)$ is the same as the open interval $(x_0 - r, x_0 + r)$.

6.2. Functions of two real variables

We now focus on real-valued functions f of two real variables. Thus, we write $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^2$.

6.2.1. Natural domain. To define a function f, we need to first specify its domain. However, it is possible to start with a formula for f, and then figure out all points where that formula makes sense. This is called the natural domain of f.

For example, the natural domain of

$$f(x, y) = \sqrt{4 - x^2 - y^2}$$

is

$$\{(x, y) \in \mathbb{R}^2 \colon 4 - x^2 - y^2 \geq 0\} = \{(x, y) \in \mathbb{R}^2 \colon \|(x, y)\| \leq 2\}.$$

6.2.2. Interior and boundary points. A point (a, b) of $\mathbb{R}^2$ is an *interior point* of D if there is $r > 0$ such that $B((a, b), r) \subseteq D$. That is, if there is an open ball around (a, b) which lies inside D. In particular, an interior point of D lies in D. Let $\operatorname{Int} D$ denote the set of interior points of D. We say D is an *open set* if all points of D are interior points, that is, $\operatorname{Int} D = D$.

A point (a, b) of $\mathbb{R}^2$ is a *boundary point* of D if $B((a, b), r)$ intersects both D and $\mathbb{R}^2 \setminus D$ for every $r > 0$. Note very carefully: A boundary point of D may not lie in D. Let ∂D denote the set of boundary points of D. We say D is a *closed set* if it contains all its boundary points, that is, $\partial D \subseteq D$.

Example 6.3 (Open and closed unit discs). Let D_1 be the open unit disc, and D_2 be the closed unit disc. That is,

$$D_1 := \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 < 1\} \quad \text{and} \quad D_2 := \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\}.$$

See pictures below.

The interior points of both D_1 and D_2 are those strictly inside the unit circle. The boundary points of both D_1 and D_2 are those on the unit circle. That is,

$$\partial D_1 = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 = 1\} = \partial D_2.$$

Thus, D_1 is an open set, but not a closed set, while D_2 is a closed set, but not a open set.

One can also imagine a subset D in between D_1 and D_2 which contains only a part of the boundary such as a semicircle. Then interior and boundary points of D are the same as above, and it is neither open nor closed.

The analogues of D_1, D_2, D for $m = 1$ are

$$(a, b), \qquad [a, b], \qquad [a, b),$$

respectively. The interval (a, b) is an open set, $[a, b]$ is a closed set, while $[a, b)$ is neither open nor closed.

Example 6.4 (Open and closed rectangles). Let D_1 be the open rectangle $(a, b) \times (c, d)$. Let D_2 be the closed rectangle $[a, b] \times [c, d]$. See pictures below.

D_1 D_2

The interior points of D_1 and D_2 are precisely those contained in D_1. The boundary points of both D_1 and D_2 are those on the four sides of the rectangle. Thus, D_1 is an open set, but not a closed set, while D_2 is a closed set, but not a open set.

One may also consider more complicated shapes such as an annulus (with two boundary circles), or a rectangle with a disc removed from its interior, and so on.

6.2.3. Bounded region. We say D is *bounded* if there is a real number M such that

$$\|(x, y)\| \leq M$$

for $(x, y) \in D$. In other words, D is contained inside a closed disc of radius M centered at the origin.

Exercise 6.5. A nice example of a closed and bounded set in $\mathbb{R}^2$ is the closed unit disc. Another example is the closed rectangle $[a, b] \times [c, d]$.

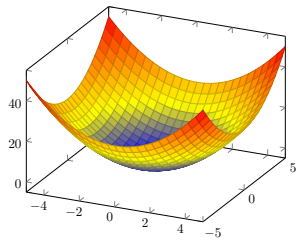
- Give an example of a closed set in $\mathbb{R}^2$ which is not bounded.
- Give an example of a bounded set in $\mathbb{R}^2$ which is not closed.

6.2.4. Graph of a function. The *graph* of $f : D \to \mathbb{R}$ is the subset of $\mathbb{R}^3$ defined by

$$\{(x, y, f(x, y)) \colon (x, y) \in D\}.$$

This is an example of a surface in $\mathbb{R}^3$.

There are many examples of functions of two real variables such as polynomial functions, rational functions, and so on. The graph of the polynomial function $f(x, y) = x^2 + y^2$ is shown below. It is called a *paraboloid*.

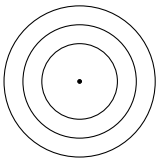


6.2.5. Level curves and contour lines. For $c \in \mathbb{R}$, the *level curve* of f corresponding to c is the subset of $\mathbb{R}^2$ defined by

$$\{(x, y) \in D \colon f(x, y) = c\}.$$

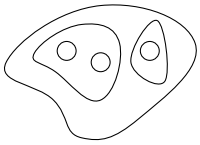
It consists of those points in the domain where f takes value c.

Example 6.6. Consider $f(x, y) = x^2 + y^2$ whose graph was drawn above. The level curve for $c > 0$ is the circle centered at the origin and of radius $\sqrt{c}$. Level curves become smaller and smaller as c decreases. The level curve for $c = 0$ consists of a single point, namely, the origin. See picture below.



Now consider $f(x, y) = 100 - x^2 - y^2$. Its graph is obtained by inverting the previous graph and shifting it up by 100. It looks like the surface of a mountain. Level curves are again circles. These are points on sea level where the mountain has the same height. Now level curves become smaller and smaller as c increases. The level curve for $c = 100$ consists of a single point, namely, the origin. This corresponds to the peak of the mountain.

A more generic picture of level curves is shown below.



For $c \in \mathbb{R}$, the *countour curve* of f corresponding to c is the subset of $\mathbb{R}^3$ defined by

$$\{(x, y, f(x, y)) \colon (x, y) \in D, f(x, y) = c\}.$$

It is the intersection of the graph of f and the horizontal plane $z = c$.

6.3. Sequences

We studied sequences of real numbers in Section 2.1. We now briefly consider sequences in $\mathbb{R}^2$ following the same general ideas.

6.3.1. Sequences in $\mathbb{R}^2$.

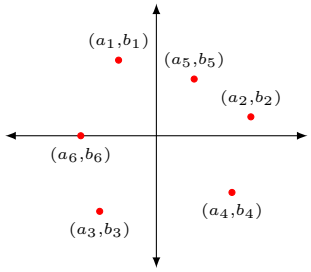
Definition 6.7. A *sequence in* $\mathbb{R}^2$ is a function $f : \mathbb{N}_+ \to \mathbb{R}^2$ from the set of positive integers to the set of pairs of real numbers.

Put $f(n) = (a_n, b_n)$. Thus specifying f is the same as specifying

$$(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots.$$

We shall use the notation $\{(a_n, b_n)\}$ for short. We call (a_n, b_n) the n-th term of the sequence.

A sequence in $\mathbb{R}^2$ may be visualized as follows by marking its terms $(a_1, b_1), (a_2, b_2), (a_3, b_3), \ldots$.



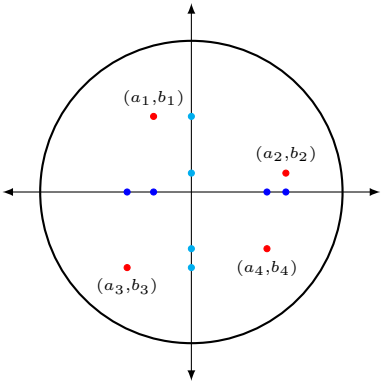
6.3.2. Bounded sequences.

Definition 6.8. A sequence $\{(a_n, b_n)\}$ in $\mathbb{R}^2$ is *bounded* if there is a real number M such that

$$\|(a_n, b_n)\| \leq M$$

for $n \geq 1$.

Thus, a bounded sequence lies entirely in some disc of radius M. See picture below.



The picture also shows the corresponding sequence $\{a_n\}$ on the x-axis and sequence $\{b_n\}$ on the y-axis. This ties with the result below.

Lemma 6.9. *The sequence $\{(a_n, b_n)\}$ is bounded iff sequences $\{a_n\}$ and $\{b_n\}$ are both bounded.*

Proof. Exercise. □

Note: We have not defined a monotone sequence in $\mathbb{R}^2$. In contrast to $\mathbb{R}$, there is no canonical linear order on $\mathbb{R}^2$.

6.3.3. Convergence of sequences.

Definition 6.10 (ϵ–n_0). A sequence $\{(a_n, b_n)\}$ in $\mathbb{R}^2$ is *convergent* if there is $(a, b) \in \mathbb{R}^2$ such that the following condition holds.

For every $\epsilon > 0$, there is $n_0 \in \mathbb{N}_+$ such that

$$\|(a_n, b_n) - (a, b)\| < \epsilon$$

for $n \geq n_0$.

Lemma 6.11. *The sequence $\{(a_n, b_n)\}$ converges iff sequences $\{a_n\}$ and $\{b_n\}$ both converge. In this case,*

$$\lim_{n\to\infty}(a_n, b_n) = (\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n).$$

PROOF. Use the inequalities

$$|u_1|, |u_2| \le \|(u_1, u_2)\| \le |u_1| + |u_2|.$$

We leave further details as an exercise. □

6.4. Continuity

We studied continuity of functions f of a real variable in Section 3.1. We now look at continuity of functions f of two real variables. Thus, we write $f : D \to \mathbb{R}$, where $D \subseteq \mathbb{R}^2$. The intuitive idea remains the same, that is, the graph of f has no "breaks".

6.4.1. Continuous functions.

Definition 6.12 (ϵ–δ). Let $f : D \to \mathbb{R}$. We say f is *continuous* at $(a, b) \in D$ if the following condition holds.

For every $\epsilon > 0$, there is $\delta > 0$ such that

$$\|(x, y) - (a, b)\| < \delta \quad \Longrightarrow \quad |f(x, y) - f(a, b)| < \epsilon.$$

We say f is *continuous* on D if f is continuous at each point of D.

Draw picture.

6.4.2. Algebra of continuous functions.

Lemma 6.13. *Suppose $f, g : D \to \mathbb{R}$ are continuous at $(a, b) \in D$. Then so are*

(i) $f + g$,
(ii) rf *for* $r \in \mathbb{R}$,
(iii) fg,
(iv) $1/f$ *if* $f(a, b) \neq 0$.

PROOF. The proof of item (i) is identical in form to the one that we gave for Lemma 3.4, item (i). The main notational change is to replace $|x - c| < \delta$ by $\|(x, y) - (a, b)\| < \delta$. □

Lemma 6.14. *Let $f : D \to E$ and $g : E \to \mathbb{R}$. If f is continuous at $(a, b) \in D$ and g is continuous at $f(a, b) \in E$, then the composite $g \circ f$ is continuous at $(a, b) \in D$.*

As a consequence:

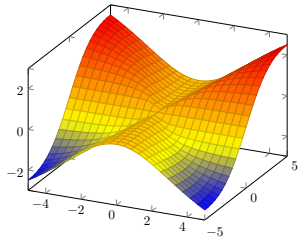
- polynomials in x and y such as $p(x, y) = x^2 + y^2$ and $p(x, y) = 2x^3y - 3x + y + 1$ are continuous,
- a rational function in x and y, that is $r(x, y) = p(x, y)/q(x, y)$, where p and q are polynomials, is continuous at (a, b) if $q(a, b) \neq 0$,
- functions such as $f(x, y) = x^3 \sin|y| + \cos(x^2 + y)$ and $f(x, y) = e^{x^2+xy}$ are continuous.

Example 6.15. Let us illustrate Definition 6.12.

(1) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

Its graph is shown below.



Then f is continuous at $(a,b) \neq (0,0)$ since it is formed out of continuous functions. It is also continuous at $(0,0)$. Why? We can estimate as follows.

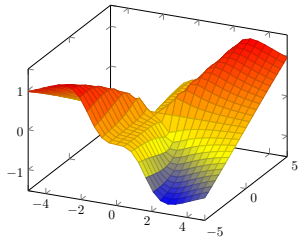
$$\left|\frac{x^2 y}{x^2+y^2}\right| = |x| \left|\frac{xy}{x^2+y^2}\right| \leq \frac{1}{2}|x| \leq \|(x,y)\|.$$

Given $\epsilon > 0$, we can take $\delta = \epsilon$, yielding continuity at $(0,0)$.

(2) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{x^3 y}{x^4+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

Its graph is shown below.



Then f is continuous at $(a,b) \neq (0,0)$ since it is formed out of continuous functions. It is also continuous at $(0,0)$. Why?

6.4.3. Characterization using sequences.

Proposition 6.16. *Let $f : D \to \mathbb{R}$. Then f is continuous at $(a,b) \in D$ iff the following condition holds.*

For any sequence $\{(x_n, y_n)\}$ in D with $(x_n, y_n) \to (a,b)$, we have $f(x_n, y_n) \to f(a,b)$.

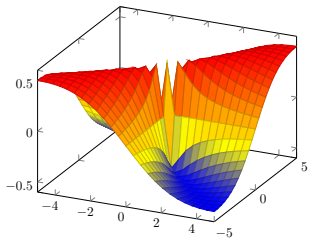
Proof. See [13, Proposition 2.22]. □

Example 6.17. Let us use Proposition 6.16 to show that certain functions are not continuous at a point.

(1) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

It is bounded by $\frac{1}{2}$. Its graph is shown below.

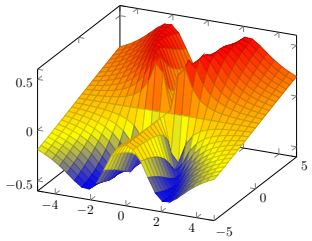


It is identically zero on the x-axis and on the y-axis. Consider the sequence $\{(\frac{1}{n}, \frac{1}{n})\}$. It converges to $(0,0)$ along the line $y = x$. However, the sequence $\{f(\frac{1}{n}, \frac{1}{n})\}$ converges to $\frac{1}{2}$ which is not equal to $f(0,0)$. Thus, f is not continuous at $(0,0)$. What happens if we approach along the line $y = mx$?

(2) Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} \frac{x^2y}{x^4+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0). \end{cases}$$

It is bounded by $\frac{1}{2}$. Its graph is shown below.



It is identically zero on the x-axis and on the y-axis. Consider the sequence $\{(\frac{1}{n}, \frac{1}{n^2})\}$. It converges to $(0,0)$ along the parabola $y = x^2$. However, the sequence $\{f(\frac{1}{n}, \frac{1}{n^2})\}$ converges to $\frac{1}{2}$ which is not equal to $f(0,0)$. Thus, f is not continuous at $(0,0)$. Note: If we approach along the line $y = mx$, then $f(x,y)$ indeed goes to 0.

6.4.4. Further properties of continuous functions. We say D is *path connected* if for any points $x, y \in D$, there is a curve in D joining x and y. We say D is *convex* if for any points $x, y \in D$, the line segment joining x and y lies entirely in D.

A convex set is path connected. The converse is false. For example, an annulus is path connected but not convex.

Proposition 6.18. *Let $f : D \to \mathbb{R}$ be a continuous function, and D be path connected. Then $f(D)$ is an interval.*

PROOF. Given points (x_1, y_1) and (x_2, y_2) in D, join them by a path. In other words, let

$$C : [\alpha, \beta] \to D, \qquad C(\alpha) = (x_1, y_1), \quad C(\beta) = (x_2, y_2).$$

Now apply IVP (Theorem 3.10) to the composite $f \circ C$. □

Theorem 6.19. *Let $f : D \to \mathbb{R}$ be continuous, and D be closed and bounded. Then f is bounded on D and attains its global maximum and global minimum on D.*

PROOF. See [13, Proposition 2.25]. □

6.5. Limit of a function

6.5.1. Limit of a function. Let $f : D \to \mathbb{R}$ and $(a, b) \in \mathbb{R}^2$ be such that there is $r > 0$ with $B((a, b), r) \setminus \{(a, b)\} \subseteq D$. In other words, D contains all points within distance r of (a, b), except perhaps the point (a, b).

Definition 6.20. We say $\lim_{(x,y)\to(a,b)} f(x, y)$ exists if there is $\ell \in \mathbb{R}$ such that for every sequence $\{(x_n, y_n)\}$ in D with $(x_n, y_n) \neq (a, b)$ and $(x_n, y_n) \to (a, b)$, we have $f(x_n, y_n) \to \ell$.

In this case, we write

$$\ell = \lim_{(x,y)\to(a,b)} f(x, y),$$

and say f has a limit at (a, b).

Remark 6.21 (ϵ–δ). Equivalently, similar to Definition 6.12 for continuity, we say:

$$\lim_{(x,y)\to(a,b)} f(x, y) = \ell$$

if the following condition holds.

For every $\epsilon > 0$, there is $\delta > 0$ such that

$$0 < \|(x, y) - (a, b)\| < \delta \implies |f(x, y) - \ell| < \epsilon.$$

It is possible to take this as a definition, and deduce Definition 6.20 as a consequence.

6.5.2. Algebra of limits of functions.

Lemma 6.22 (Limit theorems). *Suppose $\lim_{(x,y)\to(a,b)} f(x, y)$ and $\lim_{(x,y)\to(a,b)} g(x, y)$ exist. Then*

(i)
$$\lim_{(x,y)\to(a,b)} (f + g)(x, y) = \lim_{(x,y)\to(a,b)} f(x, y) + \lim_{(x,y)\to(a,b)} g(x, y),$$

(ii)
$$\lim_{(x,y)\to(a,b)} rf(x, y) = r \lim_{(x,y)\to(a,b)} f(x, y) \text{ for } r \in \mathbb{R}.$$

(iii)
$$\lim_{(x,y)\to(a,b)} (fg)(x, y) = (\lim_{(x,y)\to(a,b)} f(x, y))(\lim_{(x,y)\to(a,b)} g(x, y)),$$

(iv)

$$\lim_{(x,y)\to(a,b)} \left(\frac{1}{f}\right)(x,y) = \frac{1}{\lim_{(x,y)\to(a,b)} f(x,y)} \quad (\textit{if denominator} \neq 0).$$

PROOF. Follows from Lemma 2.13 for sequences. □

6.5.3. Continuity and limit. We say $(a,b) \in \mathbb{R}^2$ is an *interior point* of $D \subseteq \mathbb{R}^2$ if there is $r > 0$ such that $B((a,b),r) \subseteq D$.

Proposition 6.23. *Let $f : D \to \mathbb{R}$, and (a,b) be an interior point of D. Then f is continuous at (a,b) iff $\lim_{(x,y)\to(a,b)} f(x,y)$ exists and is equal to $f(a,b)$.*

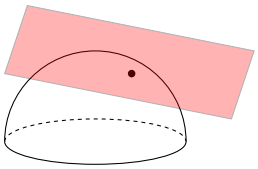
PROOF. See [13, Proposition 2.48]. □

CHAPTER 7

Differentiability

7.1. Differentiability

We studied differentiability of functions f of a real variable in Section 4.1. We now look at differentiability of functions f of two real variables. The intuitive idea is that the graph of f has tangent planes instead of tangent lines. See illustration below.



We also encounter the notions of partial derivatives, and the more general directional derivatives, by restricting f to a line inside its domain.

7.1.1. Partial derivatives. Let $D \subseteq \mathbb{R}^2$. Let $f : D \to \mathbb{R}$, and (a, b) be an interior point of D. The *partial derivative* of f wrt x at (a, b), denoted $f_x(a, b)$, is defined by

$$f_x(a,b) := \lim_{h\to 0} \frac{f(a+h,b) - f(a,b)}{h} \tag{7.1}$$

assuming this limit exists. It represents the rate of change in f in the x-direction near (a, b).

Similarly, the *partial derivative* of f wrt y at (a, b), denoted $f_y(a, b)$, is defined by

$$f_y(a,b) := \lim_{k\to 0} \frac{f(a,b+k) - f(a,b)}{k} \tag{7.2}$$

assuming this limit exists. It represents the rate of change in f in the y-direction near (a, b).

The partial derivatives f_x and f_y are also denoted by $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, respectively.

The *gradient* of f at (a, b), denoted $\nabla f(a, b)$, is defined by

$$\nabla f(a,b) := (f_x(a,b), f_y(a,b)). \tag{7.3}$$

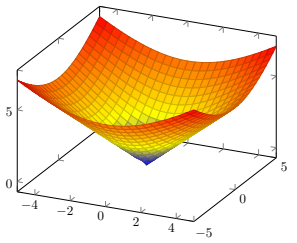
Example 7.1. Let us illustrate the notion of partial derivative.

(1) Let $f(x,y) = x^2 + y^2$. Then $f_x(a,b) = 2a$ and $f_y(a,b) = 2b$. For instance,

$$\begin{aligned} f_x(a,b) &= \lim_{h\to 0} \frac{(a+h)^2 + b^2 - a^2 - b^2}{h} \\ &= \lim_{h\to 0} \frac{2ah + h^2}{h} \\ &= 2a. \end{aligned}$$

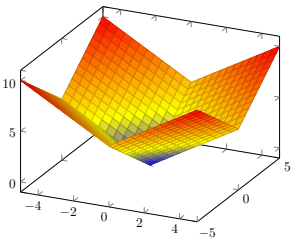
This is similar to the calculation one does to show that $f'(x^2) = 2x$ from first principles.

(2) Let $f(x,y) = \sqrt{x^2+y^2}$. Its graph is shown below. It is the surface of a *cone*.



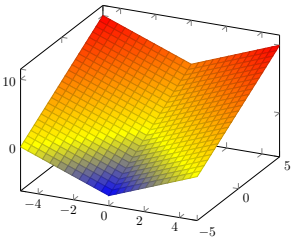
Then f is continuous at $(0,0)$, but neither partial derivative exists at $(0,0)$. This is because $f(x,0) = |x|$ and $f(0,y) = |y|$.

(3) Let $f(x,y) = |x| + |y|$. Its graph is shown below. It consists of four quarter-planes glued together.



Then f is continuous at $(0,0)$, but neither partial derivative exists at $(0,0)$. This is again because $f(x,0) = |x|$ and $f(0,y) = |y|$.

(4) Let $f(x,y) = |x| + y$. Its graph is shown below. It consists of two half-planes glued together.



Then f is continuous at $(0,0)$. Now f_x does not exist at $(0,0)$, but f_y does exist at $(0,0)$.

(5) Let $f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if } (x,y) \neq (0,0), \\ 0 & \text{if } (x,y) = (0,0) \end{cases}$ as in Example 6.17, item (1). Then f is not continuous at $(0,0)$, but both partial derivatives exist at $(0,0)$, with $f_x(0,0) = f_y(0,0) = 0$.

7.1.2. Directional derivatives. Let $f : D \to \mathbb{R}$, and (a,b) be an interior point of D. Let $u = (u_1, u_2)$ with $\|u\| = 1$. The *directional derivative* of f along u at (a,b), denoted $D_u f(a,b)$, is defined by

$$D_u f(a,b) := \lim_{t\to 0} \frac{f(a+tu_1, b+tu_2) - f(a,b)}{t}. \tag{7.4}$$

It represents the rate of change in f in the u-direction near (a,b).

Note:

$$D_{(1,0)} f = f_x, \qquad D_{(0,1)} f = f_y.$$

Also, $D_{-u} f = -D_u f$.

Example 7.2. Let $f(x,y) = x^2 + y^2$. Then $f_x(a,b) = 2a$ and $f_y(a,b) = 2b$. Let us compute the directional derivative of f at (a,b).

$$\begin{aligned} \lim_{t\to 0} \frac{(a+tu_1)^2 + (b+tu_2)^2 - a^2 - b^2}{t} &= \lim_{t\to 0} \frac{2atu_1 + 2btu_2 + t^2u_1^2 + t^2u_2^2}{t} \\ &= 2au_1 + 2bu_2 \\ &= f_x(a,b)u_1 + f_y(a,b)u_2 \\ &= \nabla f(a,b) \cdot (u_1, u_2). \end{aligned}$$

Thus, we see that $D_u f(a,b)$ is the dot product of $\nabla f(a,b)$ and (u_1, u_2). This is not true in general. A sufficient condition is given in Lemma 7.4 below.

Remark 7.3. For a function $f(x,y,z)$ of three variables, we have three partial derivatives, namely, f_x, f_y, f_z. Also, we have the directional derivative of f along u at (a,b,c), denoted $D_u f(a,b,c)$, for $u = (u_1, u_2, u_3)$ with $\|u\| = 1$. It is defined by

$$D_u f(a,b,c) := \lim_{t\to 0} \frac{f(a+tu_1, b+tu_2, c+tu_3) - f(a,b,c)}{t}. \tag{7.5}$$

7.1.3. Differentiability. Let $f : D \to \mathbb{R}$, and (a,b) be an interior point of D. We say f is *differentiable* at (a,b) if there is $(\alpha, \beta) \in \mathbb{R}^2$ such that

$$\lim_{(h,k)\to(0,0)} \frac{f(a+h, b+k) - f(a,b) - \alpha h - \beta k}{\sqrt{h^2+k^2}} = 0. \tag{7.6}$$

We call the pair (α, β) as the *total derivative* of f at (a,b). Compare with the one variable formulation in (4.1).

Lemma 7.4. *Suppose f is differentiable at (a,b) with (α, β) as its total derivative. Then*

(i) *$f_x(a,b)$ exists and equals α,*
(ii) *$f_y(a,b)$ exists and equals β,*
(iii) *$D_u f(a,b)$ exists and equals $\alpha u_1 + \beta u_2$. That is,*

$$D_u f(a,b) = \nabla f(a,b) \cdot u. \tag{7.7}$$

Note: Items (i) and (ii) are special cases of item (iii).

PROOF.

- For item (i), set $k = 0$ in (7.6).
- For item (ii), set $h = 0$ in (7.6).
- For item (iii), set $h = tu_1$ and $k = tu_2$ in (7.6).

□

Example 7.5. Let $f(x, y) = x^2 + y^2$. Then $f_x(a, b) = 2a$ and $f_y(a, b) = 2b$. Let us check that f is differentiable at (a, b).

$$\lim_{(h,k)\to(0,0)} \frac{(a+h)^2 + (b+k)^2 - a^2 - b^2 - 2ah - 2bk}{\sqrt{h^2+k^2}}$$
$$= \lim_{(h,k)\to(0,0)} \frac{h^2+k^2}{\sqrt{h^2+k^2}} = \lim_{(h,k)\to(0,0)} \sqrt{h^2+k^2} = 0.$$

As a consequence, (7.7) holds, and we recover the observation in Example 7.2.

In pictorial terms, it is the two-dimensional limit in (7.6) which guarantees the existence of the tangent plane to the graph of f at (a, b). It implies the existence of the one-dimensional limits in (7.1), (7.2), (7.3) as we saw in Lemma 7.4. Note very carefully: The converse is false, that is, the existence of the one-dimensional limits do not impy the existence of the two-dimensional limit. A sufficient condition for the latter is given in Proposition 7.10 below.

7.1.4. Pair of increment functions.

Lemma 7.6 (Caratheodory lemma). *A function $f : D \to \mathbb{R}$ is differentiable at an interior point (a, b) of D iff there are functions $f_1, f_2 : D \to \mathbb{R}$ which are continuous at (a, b) such that*

$$f(x, y) - f(a, b) = (x - a)f_1(x, y) + (y - b)f_2(x, y) \tag{7.8}$$

for $(x, y) \in D$. Moreover,

$$\nabla f(a, b) = (f_1(a, b), f_2(a, b)).$$

PROOF. See [13, Proposition 3.25]. □

We call (f_1, f_2) a pair of increment functions. Note very carefully: f_1 and f_2 depend on (a, b). Moreover, they are not necessarily unique. An illustration (in slightly different notation) is given in Example 7.9 below. This is in contrast to what happened in the one variable case in Caratheodory Lemma 4.5.

Corollary 7.7. *If f is differentiable at (a, b), then f is continuous at (a, b).*

PROOF. Let f be differentiable at (a, b). Using (7.8), write

$$f(x, y) = f(a, b) + (x - a)f_1(x, y) + (y - b)f_2(x, y)$$

Since f_1 and f_2 are continuous, so is f by Lemma 6.13. □

In pictorial terms, Corollary 7.7 says that if the graph of f has a break at (a, b), then there is no hope of drawing a tangent plane at (a, b).

Here is an alternative way to phrase Caratheodory lemma.

Lemma 7.8. *A function $f : D \to \mathbb{R}$ is differentiable at an interior point (a, b) of D iff there are real numbers α, β such that*

$$f(a+h, b+k) = f(a,b) + \alpha\, h + \beta\, k + \epsilon_1(h,k)\, h + \epsilon_2(h,k)\, k \tag{7.9}$$

where $\epsilon_1(h,k)$ and $\epsilon_2(h,k)$ are defined for small h, k, and $\epsilon_1(h,k) \to 0$ and $\epsilon_2(h,k) \to 0$ as $(h,k) \to (0,0)$. Moreover,

$$f_x(a,b) = \alpha \quad \text{and} \quad f_y(a,b) = \beta.$$

Proof. To link (7.8) and (7.9), put

$$x = a + h, \quad y = b + k, \quad h = x - a, \quad k = y - b,$$
$$f_1(x,y) = \alpha + \epsilon_1(h,k), \quad f_2(x,y) = \beta + \epsilon_2(h,k).$$

The continuity claims about f_1 and f_2 link to the continuity claims about ϵ_1 and ϵ_2. □

Example 7.9. Using the calculation in Example 7.5: For $f(x,y) = x^2 + y^2$,

$$\epsilon_1(h,k)\, h + \epsilon_2(h,k)\, k = h^2 + k^2.$$

So a possible choice is $\epsilon_1(h,k) = h$ and $\epsilon_2(h,k) = k$. Another choice is $\epsilon_1(h,k) = h - k$ and $\epsilon_2(h,k) = h + k$. Thus, we explicitly see that these two functions are not unique.

Proposition 7.10 (Sufficient condition for differentiability). *Let $f : D \to \mathbb{R}$ and (a,b) be an interior point of D. Suppose f_x and f_y exist in $B((a,b), r)$ for some $r > 0$, and are continuous at (a,b). Then f is differentiable at (a,b).*

Proof. By mean value Theorem 4.21,

$$f(a+h, b) - f(a,b) = f_x(c,b)h,$$
$$f(a+h, b+k) - f(a+h,b) = f_y(a+h,d)k$$

for some c and d. Adding,

$$\begin{aligned} &f(a+h,b+k) - f(a,b) \\ &\quad = f_x(c,b)h + f_y(a+h,d)k \\ &\quad = f_x(a,b)h + f_y(a,b)k + [f_x(c,b) - f_x(a,b)]\, h + [f_y(a+h,d) - f_y(a,b)]\, k \\ &\quad = f_x(a,b)h + f_y(a,b)k + \epsilon_1(h,k)\, h + \epsilon_2(h,k)\, k \end{aligned}$$

for some $\epsilon_1(h,k)$ and $\epsilon_2(h,k)$. These go to 0 as $(h,k) \to (0,0)$ since f_x and f_y are continuous at (a,b) by hypothesis. Now use Lemma 7.8 to deduce that f is differentiable at (a,b).

For the equivalent argument using Caratheodory Lemma 7.6, see [13, Proposition 3.33]. □

7.1.5. Algebra of differentiable functions.

Lemma 7.11. *Suppose $f, g : D \to \mathbb{R}$ are differentiable at $(a,b) \in D$. Then*

(i) *$f + g$ is differentiable at (a,b), and*

$$\nabla(f+g)(a,b) = \nabla f(a,b) + \nabla g(a,b),$$

(ii) *rf is differentiable at (a,b), and*

$$\nabla(rf)(a,b) = r\nabla f(a,b)$$

for $r \in \mathbb{R}$,

(iii) *fg is differentiable at (a,b), and*

$$\nabla(fg)(a,b) = \nabla f(a,b)g(a,b) + f(a,b)\nabla g(a,b),$$

(iv) *$1/f$ is differentiable at (a,b), and*

$$\nabla(1/f)(a,b) = \frac{-\nabla f(a,b)}{f(a,b)^2}$$

if $f(a,b) \neq 0$.

PROOF. We can employ Caratheodory Lemma 7.6, as we did in the proof of the one variable case in Lemma 4.9. For details, see [13, Proposition 3.30]. □

A function $g : \mathbb{R} \to \mathbb{R}^2$ is the same as a pair of functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$. Write $g = (g_1, g_2)$. Explicitly, $g(t) = (g_1(t), g_2(t))$. We say g is differentiable at c if both g_1 and g_2 are differentiable at c.

Recall: $g : \mathbb{R} \to \mathbb{R}^2$ is the same as a parametrized curve in $\mathbb{R}^2$, where the parameter varies over all of $\mathbb{R}$. A familiar example is $g(t) = (\cos t, \sin t)$.

Lemma 7.12 (Chain rule). *Let*

$$\mathbb{R} \xrightarrow{g} \mathbb{R}^2 \xrightarrow{f} \mathbb{R}.$$

Let g be differentiable at c, and f be differentiable at $g(c) =: (a,b)$. Then the composite $f \circ g$ is differentiable at c, and

$$(f \circ g)'(c) = f_x(a,b)g_1'(c) + f_y(a,b)g_2'(c). \tag{7.10}$$

PROOF. Exercise. □

The chain rule (7.10) can be visualized in matrix notation as

$$\begin{bmatrix}(f \circ g)'(c)\end{bmatrix} = \begin{bmatrix}f_x(a,b) & f_y(a,b)\end{bmatrix} \begin{bmatrix}g_1'(c) \\ g_2'(c)\end{bmatrix}.$$

This can be written in short hand as

$$\begin{bmatrix}(f \circ g)'\end{bmatrix} = \begin{bmatrix}f_x & f_y\end{bmatrix} \begin{bmatrix}g_1' \\ g_2'\end{bmatrix}.$$

Example 7.13. Let $f(x,y) = x^2 + y^2$ and $g(t) = (e^t, t)$. Then

$$(f \circ g)(t) = f(e^t, t) = e^{2t} + t^2.$$

By chain rule,

$$(f \circ g)'(t) = 2(e^t)e^t + 2t(1) = 2e^{2t} + 2t.$$

In matrix notation,

$$\begin{bmatrix}(f \circ g)'(t)\end{bmatrix} = \begin{bmatrix}2x & 2y\end{bmatrix} \begin{bmatrix}e^t \\ 1\end{bmatrix} = \begin{bmatrix}2e^t & 2t\end{bmatrix} \begin{bmatrix}e^t \\ 1\end{bmatrix} = 2e^{2t} + 2t.$$

As a variant, let $f(x,y) = x^2 + y^2$ and $g(t) = (e^t, \sin t)$. Then

$$(f \circ g)(t) = f(e^t, \sin t) = e^{2t} + \sin^2 t.$$

By chain rule,

$$(f \circ g)'(t) = 2(e^t)e^t + 2\sin t(\cos t) = 2e^{2t} + 2\sin t\cos t.$$

We leave it to you to write this in matrix notation.

Another way of writing or thinking about (7.10) is shown below.

$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt}. \tag{7.11}$$

Here t is the variable in the first $\mathbb{R}$, (x, y) are the variables in the middle $\mathbb{R}^2$, and w is the variable in the last $\mathbb{R}$.

A function $g : \mathbb{R}^2 \to \mathbb{R}^2$ is the same as a pair of functions $g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}$. Write $g = (g_1, g_2)$. Explicitly, $g(u, v) = (g_1(u, v), g_2(u, v))$. We say g is differentiable at (c, d) if both g_1 and g_2 are differentiable at (c, d).

Lemma 7.14 (Chain rule). *Let*

$$\mathbb{R}^2 \xrightarrow{g} \mathbb{R}^2 \xrightarrow{f} \mathbb{R}.$$

Let g be differentiable at (c, d), and let f be differentiable at $g(c, d) =: (a, b)$. Then the composite $f \circ g$ is differentiable at (c, d), and

$$\begin{aligned}(f \circ g)_u(c, d) &= f_x(a, b)(g_1)_u(c, d) + f_y(a, b)(g_2)_u(c, d),\\ (f \circ g)_v(c, d) &= f_x(a, b)(g_1)_v(c, d) + f_y(a, b)(g_2)_v(c, d).\end{aligned} \tag{7.12}$$

The chain rule (7.12) can be visualized in matrix notation as

$$\begin{bmatrix}(f \circ g)_u(c, d) & (f \circ g)_v(c, d)\end{bmatrix} = \begin{bmatrix}f_x(a, b) & f_y(a, b)\end{bmatrix} \begin{bmatrix}(g_1)_u(c, d) & (g_1)_v(c, d)\\ (g_2)_u(c, d) & (g_2)_v(c, d)\end{bmatrix}.$$

This can be written in short hand as

$$\begin{bmatrix}(f \circ g)_u & (f \circ g)_v\end{bmatrix} = \begin{bmatrix}f_x & f_y\end{bmatrix} \begin{bmatrix}(g_1)_u & (g_1)_v\\ (g_2)_u & (g_2)_v\end{bmatrix}.$$

Example 7.15. Let $f(x, y) = x^2 + y^2$ and $g(u, v) = (u^2 - v^2, 2uv)$. Then

$$(f \circ g)(u, v) = f(u^2 - v^2, 2uv) = (u^2 - v^2)^2 + (2uv)^2 = u^4 + 2u^2v^2 + v^4.$$

By chain rule,

$$\begin{aligned}(f \circ g)_u &= 2(u^2 - v^2)(2u) + 2(2uv)(2v) = 4u(u^2 + v^2),\\ (f \circ g)_v &= 2(u^2 - v^2)(-2v) + 2(2uv)(2u) = 4v(u^2 + v^2).\end{aligned}$$

In matrix notation,

$$\begin{aligned}\begin{bmatrix}(f \circ g)_u & (f \circ g)_v\end{bmatrix} &= \begin{bmatrix}2x & 2y\end{bmatrix} \begin{bmatrix}2u & -2v\\ 2v & 2u\end{bmatrix}\\ &= \begin{bmatrix}2(u^2 - v^2) & 4uv\end{bmatrix} \begin{bmatrix}2u & -2v\\ 2v & 2u\end{bmatrix}\\ &= \begin{bmatrix}4u(u^2 + v^2)\\ 4v(u^2 + v^2)\end{bmatrix}.\end{aligned}$$

Another way of writing or thinking about (7.12) is shown below.

$$\frac{\partial w}{\partial u} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u} \quad \text{and} \quad \frac{\partial w}{\partial v} = \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v}. \tag{7.13}$$

Here (u, v) are the variables in the first $\mathbb{R}^2$, (x, y) are the variables in the middle $\mathbb{R}^2$, and w is the variable in the last $\mathbb{R}$.

Remark 7.16 (Chain rule and matrix multiplication). In general, for a map $f : \mathbb{R}^n \to \mathbb{R}^m$, we write a $m \times n$ matrix, called the *jacobian matrix*, whose entries consist of all possible partial derivatives. More precisely, write $f = (f_1, \ldots, f_m)$. In the first row, we write the n partial derivatives of f_1, in the second row, we write the n partial derivatives of f_2, and so on.

The matrix for the composite map

$$f \circ g : \mathbb{R}^n \xrightarrow{g} \mathbb{R}^m \xrightarrow{f} \mathbb{R}^p$$

is of size $p \times n$. It is the product of the matrix for f of size $p \times m$ with the matrix for g of size $m \times n$. This is how the chain rule works in general.

The special cases that we have seen before are tabulated below.

dimensions	chain rule
$n = 1$, $m = 1$, $p = 1$	(4.2)
$n = 1$, $m = 2$, $p = 1$	(7.10)
$n = 2$, $m = 2$, $p = 1$	(7.12)

Remark 7.17 (Derivatives and linear maps). The general philosophy of differential calculus is that derivative of a function f at a point x is a linear map. (These linear maps vary with the point x.) The derivative of a composite function $f \circ g$ is the composite of the derivatives of f and g. This is the chain rule. It amounts to writing a linear map as a composite of two linear maps. Finally, a linear map can be represented by a matrix, and composite of two linear maps is multiplication of the two corresponding matrices.

7.1.6. Geometric interpretation of the gradient. Let $f : D \to \mathbb{R}$ be differentiable at (a, b). Suppose $\nabla f(a, b) \neq (0, 0)$. Then by (7.7),

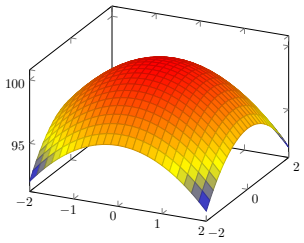
$$D_u f(a, b) = \nabla f(a, b) \cdot u = \|\nabla f(a, b)\| \cos\theta,$$

where $\theta \in [0, \pi]$ is the angle between the vectors $\nabla f(a, b)$ and u. Hence:

- $D_u f(a, b)$ is maximum when $\cos\theta = 1$, that is, $\theta = 0$. Therefore, near (a, b), f increases most rapidly in the direction $u = \frac{\nabla f(a,b)}{\|\nabla f(a,b)\|}$.
- $D_u f(a, b)$ is minimum when $\cos\theta = -1$, that is, $\theta = \pi$. Therefore, near (a, b), f decreases most rapidly in the direction $u = -\frac{\nabla f(a,b)}{\|\nabla f(a,b)\|}$.
- $D_u f(a, b) = 0$ when $\cos\theta = 0$, that is, $\theta = \pi/2$. Therefore, near (a, b), the directions of no rate of change in f are those perpendicular to $\nabla f(a, b)$.

Suppose $\nabla f(a, b) = (0, 0)$. Then $D_u f(a, b) = 0$ for all u. Hence, the rate of change in f near (a, b) is zero in all directions.

Example 7.18. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = 100 - x^2 - y^2$. Its graph is shown below.



Then $\nabla f(a, b) = (-2a, -2b)$. In particular, $\nabla f(1, 1) = (-2, -2)$. Thus, on the surface $z = f(x, y)$ at the point $(1, 1)$,

- the steepest ascent is in the direction $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$,
- the steepest descent is in the direction $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$,
- zero rate of change are in the directions $\pm(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$.

Recall from Example 6.6 that level curves of f are circles round the origin. The gradient is perpendicular to the level curves. Note that the peak of the mountain is at the origin. The steepest ascent at any point (a, b) points towards the origin. Now suppose you are standing on the mountain (that is, on the graph of f) at the point $(a, b, f(a, b))$. If you are feeling energetic, then you can take the steepest path and head straight towards the peak. If you are feeling tired, then you can take it easy and follow a contour curve for a while before starting your ascent.

7.1.7. Higher partial derivatives. Let $f : D \to \mathbb{R}$. Suppose f_x and f_y exist at all points in D. This defines functions $f_x, f_y : D \to \mathbb{R}$. So we can consider their partial derivatives. Put

$$f_{xx}(a, b) := (f_x)_x(a, b), \qquad f_{xy}(a, b) := (f_x)_y(a, b),$$
$$f_{yx}(a, b) := (f_y)_x(a, b), \qquad f_{yy}(a, b) := (f_y)_y(a, b).$$

Proposition 7.19 (Equality of mixed partials). *Let $f : D \to \mathbb{R}$, and (a, b) be an interior point of D. Suppose f_x and f_y exist on $B((a, b), r)$ for some $r > 0$. If either f_{xy} or f_{yx} exists on $B((a, b), r)$ and is continuous at (a, b), then the other exists at (a, b), and $f_{xy}(a, b) = f_{yx}(a, b)$.*

Proof. We omit the proof. See [13, Proposition 3.14]. □

Example 7.20. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = \sin(x^2 y)$. Then

$$f_x(x, y) = 2xy \cos(x^2 y) \quad \text{and} \quad f_y(x, y) = x^2 \cos(x^2 y).$$

Hence,

$$f_{xy}(x, y) = 2x \cos(x^2 y) - 2x^3 y \sin(x^2 y) = f_{yx}(x, y).$$

7.2. Tangent plane to a surface

We know that the notion of derivative of a function f (of one variable) allows us to write the equation of the tangent line to the graph of f at any point on it. More generally, if a curve is defined implicitly as the zero set of a function F (of two variables), then the partial derivatives of F, or equivalently,

the gradient of F, allow us to write the equation of the tangent line at any point on the curve.

Similarly, we can use partial derivatives of a function f (of two variables) to write the equation of the tangent plane to the graph of f at any point on it. More generally, if a surface is defined implicitly as the zero set of a function F (of three variables), then the partial derivatives of F, or equivalently, the gradient of F, allow us to write the equation of the tangent plane at any point on the surface.

7.2.1. Tangent line to a curve. Consider the curve $y = f(x)$. The tangent line to this curve at the point $(c, f(c))$ is given by

$$y - f(c) = f'(c)(x - c). \tag{7.14}$$

Suppose the curve is defined implicitly by $F(x, y) = 0$. Let (a, b) be a point on the curve, that is, $F(a, b) = 0$. Suppose $\nabla F(a, b) \neq (0, 0)$. Then the tangent line to the curve at (a, b) is given by

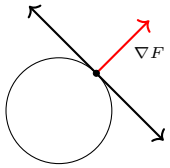
$$F_x(a, b)(x - a) + F_y(a, b)(y - b) = 0. \tag{7.15}$$

Setting $F(x, y) = y - f(x)$ recovers (7.14).

Example 7.21 (Circle). Consider the circle defined by $F(x, y) = x^2 + y^2 - 1 = 0$, and let $(a, b) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Then $\nabla F(a, b) = (2a, 2b) = (\sqrt{2}, \sqrt{2})$. The tangent line at (a, b) is

$$\sqrt{2}(x - \tfrac{1}{\sqrt{2}}) + \sqrt{2}(y - \tfrac{1}{\sqrt{2}}) = 0$$

which is the same as the line $x + y = \sqrt{2}$. See picture below.



The gradient vector $\nabla F(a, b)$ is orthogonal to the tangent line at (a, b).

Example 7.22. Consider the curve defined by $F(x, y) = y^3 - x^2 = 0$. It is the graph of the function $y = x^{\frac{2}{3}}$. See Example 4.2, item (2) for a picture. Let $(a, b) = (0, 0)$. Then $\nabla F(a, b) = (-2a, 3b) = (0, 0)$. So the above method fails.

7.2.2. Tangent plane to a surface. Consider the surface $z = f(x, y)$. The tangent plane to this surface at the point $(a, b, f(a, b))$ is given by

$$z - f(a, b) = f_x(a, b)(x - a) + f_y(a, b)(y - b). \tag{7.16}$$

Suppose the surface is defined implicitly by $F(x, y, z) = 0$. Let (a, b, c) be a point on the surface, that is, $F(a, b, c) = 0$. Suppose $\nabla F(a, b, c) \neq (0, 0, 0)$. Then the tangent plane to the surface at (a, b, c) is given by

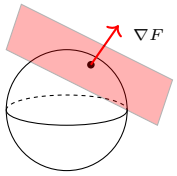
$$F_x(a, b, c)(x - a) + F_y(a, b, c)(y - b) + F_z(a, b, c)(z - c) = 0. \tag{7.17}$$

Setting $F(x, y, z) = z - f(x, y)$ recovers (7.16).

Example 7.23 (Sphere). Consider the sphere defined by $F(x,y,z) = x^2 + y^2 + z^2 - 1 = 0$, and let $(a,b,c) = (\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$. Then $\nabla F(a,b,c) = (2a, 2b, 2c) = (\frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}}, \frac{2}{\sqrt{3}})$. The tangent plane at (a,b,c) is

$$\tfrac{2}{\sqrt{3}}(x - \tfrac{1}{\sqrt{3}}) + \tfrac{2}{\sqrt{3}}(y - \tfrac{1}{\sqrt{3}}) + \tfrac{2}{\sqrt{3}}(z - \tfrac{1}{\sqrt{3}}) = 0$$

which is the same as the plane $x + y + z = \sqrt{3}$. See picture below.



The gradient vector $\nabla F(a,b,c)$ is orthogonal to the tangent plane at (a,b,c).

Example 7.24. Consider the surface defined by $F(x,y,z) = e^x + \sin y - \cos z = 0$. Then $\nabla F(0,0,0) = (1,1,0)$. Hence the tangent plane at $(0,0,0)$ is

$$1(x-0) + 1(y-0) + 0(z-0) = 0$$

which is the same as $x + y = 0$.

7.3. Maxima and minima

We studied maxima and minima for functions of a real variable in Section 4.2. We now extend these considerations to functions of two real variables.

7.3.1. Global and local maxima/minima.

Let $f : D \to \mathbb{R}$ be a function.

Definition 7.25. We say:

(i) f has a *global maximum* at (a,b) if $f(x,y) \leq f(a,b)$ for $(x,y) \in D$. In this case, $f(a,b)$ is the least upper bound of f, and it is attained at (a,b).

(ii) f has a *global minimum* at (a,b) if $f(x,y) \geq f(a,b)$ for $(x,y) \in D$. In this case, $f(a,b)$ is the greatest lower bound of f, and it is attained at (a,b).

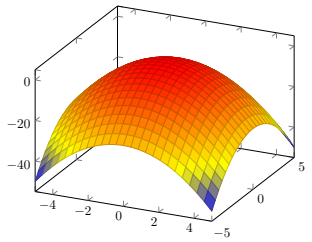
Definition 7.26. We say:

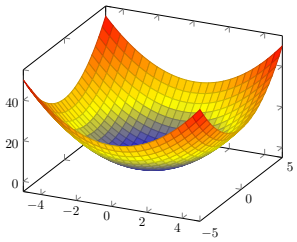
(i) f has a *local maximum* at (a,b) if there is $\delta > 0$ such that $\|(x,y) - (a,b)\| < \delta$ implies $f(x,y) \leq f(a,b)$.

(ii) f has a *local minimum* at (a,b) if there is $\delta > 0$ such that $\|(x,y) - (a,b)\| < \delta$ implies $f(x,y) \geq f(a,b)$.

Note: Global maximum (minimum) implies local maximum (minimum), but the converse is false.

Example 7.27. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = -x^2 - y^2$. It has a global maximum at $(0, 0)$. See left picture below.





Similarly, the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x^2 + y^2$ has a global minimum at $(0, 0)$. See right picture above.

7.3.2. Saddle points. Let $f : D \to \mathbb{R}$ be a function which is differentiable at (a, b).

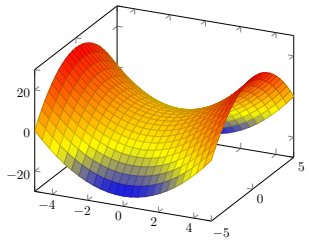
Definition 7.28. We say: f has a *saddle point* at (a, b) if the tangent plane to $z = f(x, y)$ at $(a, b, f(a, b))$ is horizontal (that is, z is constant), and for every $\delta > 0$ there are (x_1, y_1) and (x_2, y_2) with $\|(x_1, y_1) - (a, b)\| < \delta$ and $\|(x_2, y_2) - (a, b)\| < \delta$ such that

$$f(x_1, y_1) < f(a, b) < f(x_2, y_2).$$

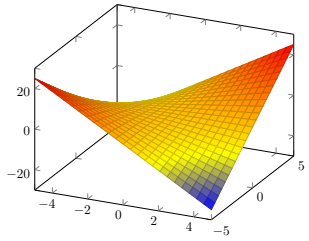
In words, a saddle point for f is a point with a horizontal tangent plane such that the surface in the neighborhood of that point does not lie entirely on one side of the tangent plane.

Example 7.29. In each of the examples below the function $f : \mathbb{R}^2 \to \mathbb{R}$ has a saddle point at $(0, 0)$.

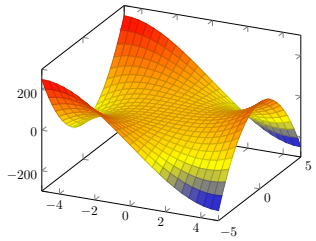
(1) $f(x, y) = x^2 - y^2$.



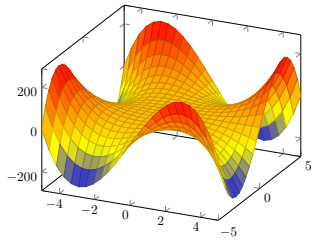
(2) $f(x, y) = xy$.



(3) $f(x,y) = x^3 - 3xy^2$.



(4) $f(x,y) = xy(x^2 - y^2)$.



Draw plus-minus diagrams on the right.

7.3.3. Local maxima/minima: necessary condition.

Lemma 7.30. *Let $f : D \to \mathbb{R}$, and (a,b) be an interior point of D. Then:*

(i) *If gradient $\nabla f(a,b)$ exists, and f has either a local maximum or a local minimum at (a,b), then $\nabla f(a,b) = 0$.*
(ii) *If directional derivative $D_u f(a,b)$ exists for some u, and f has either a local maximum or a local minimum at (a,b), then $D_u f(a,b) = 0$.*

Proof. Item (ii) implies item (i) since partial derivatives are special cases of the directional derivative. So let us prove item (ii). Put

$$g(t) := f(a + tu_1, b + tu_2) - f(a,b).$$

In particular, $g(0) = 0$. Then by hypothesis and formula (7.4),

$$g'(0) = \lim_{t \to 0} g(t)/t = D_u f(a,b)$$

exists. Also by hypothesis, g has either a local maximum or a local minimum at 0. So by Lemma 4.17, $g'(0) = 0$, that is, $D_u f(a,b) = 0$. □

The converse of Lemma 7.30 is false. For example, for $f(x,y) = x^2 - y^2$, we have $\nabla f(0,0) = 0$, but f does not have a local maximum or a local minimum at $(0,0)$. In fact, it has a saddle point at $(0,0)$. Note very carefully, we got an example in degree 2 itself. For higher degree examples, we can take $f(x,y) = x^3 + y^3$, and so on, similar to what we do in the one variable case, where the starting example is $f(x) = x^3$.

7.3.4. Local maxima/minima, saddle points: sufficient condition. Let (a,b) be an interior point of D. Suppose $f : D \to \mathbb{R}$ is such that second order partial derivatives of f exist and are continuous in a neighborhood of (a,b). Define the *discriminant* of f at (a,b) by

$$\Delta f(a,b) := f_{xx}(a,b)f_{yy}(a,b) - f_{xy}(a,b)^2. \tag{7.18}$$

Equivalently,

$$\Delta f(a,b) = \det \begin{bmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{yx}(a,b) & f_{yy}(a,b) \end{bmatrix} \tag{7.19}$$

The above matrix is called the *hessian matrix* of f at (a,b).

Lemma 7.31 (Discriminant test). *Suppose $\nabla f(a,b) = (0,0)$. Then:*

(i) *If $\Delta f(a,b) > 0$ and $f_{xx}(a,b) < 0$, then f has a local maximum at (a,b).*
(ii) *If $\Delta f(a,b) > 0$ and $f_{xx}(a,b) > 0$, then f has a local minimum at (a,b).*
(iii) *If $\Delta f(a,b) < 0$, then f has a saddle point at (a,b).*

PROOF. See [2, Theorem 9.7]. We provide a sketch. Write

$$f(x+h, y+k) = f(x,y) + hf_x + kf_y + \frac{1}{2}[h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy}] + \dots.$$

Since $\nabla f(a,b) = (0,0)$,

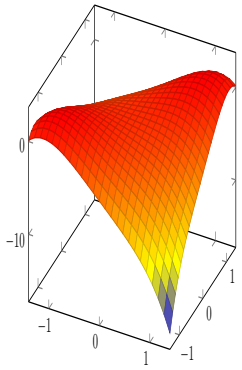
$$f(a+h, b+k) - f(a,b) \approx \frac{1}{2}[h^2 f_{xx}(a,b) + 2hkf_{xy}(a,b) + k^2 f_{yy}(a,b)].$$

Note: The discriminant of this quadratic in h and k is the negative of (7.19). □

The discriminant test is inconclusive if $\Delta f(a,b) = 0$.

Example 7.32. Let us illustrate Lemma 7.31.

(1) $f(x,y) = 4xy - x^4 - y^4$. Its graph is shown below.



We compute:

$$f_x(x,y) = 4y - 4x^3 = 4(y - x^3) \quad \text{and} \quad f_y(x,y) = 4x - 4y^3 = 4(x - y^3),$$
$$f_{xx}(x,y) = -12x^2, \quad f_{xy}(x,y) = f_{yx}(x,y) = 4, \quad f_{yy}(x,y) = -12y^2.$$

Hence the discriminant of f is

$$\Delta f(x,y) = 144x^2y^2 - 16 = 16(9x^2y^2 - 1).$$

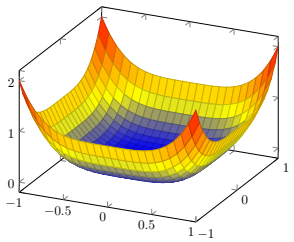
Let us first find the points where the gradient of f is zero.

$$\nabla f(x,y) = (0,0) \iff y = x^3 \text{ and } x = y^3 \iff (x,y) = (0,0), (1,1), (-1,-1).$$

Now let us evaluate the discriminant of f at each of these three points.

- $\Delta f(0,0) = -16 < 0$. Thus, f has a saddle point at $(0,0)$.
- $\Delta f(1,1) = 128 > 0$ and $f_{xx}(1,1) = -12 < 0$. Thus, f has a local maximum at $(1,1)$.
- $\Delta f(-1,-1) = 128 > 0$ and $f_{xx}(-1,-1) = -12 < 0$. Thus, f has a local maximum at $(-1,-1)$.

(2) $f(x,y) = x^4 + y^4$. Its graph is shown below.



We compute:

$$f_x(x,y) = 4x^3 \quad \text{and} \quad f_y(x,y) = 4y^3,$$

$$f_{xx}(x,y) = 12x^2, \quad f_{xy}(x,y) = f_{yx}(x,y) = 0, \quad f_{yy}(x,y) = 12y^2.$$

The gradient of f is zero at $(0,0)$. The discriminant of f is $\Delta f(x,y) = 144x^2y^2$. Thus, $\Delta f(0,0) = 0$, and the test is inconclusive. However, it is clear that f has a local minimum at $(0,0)$.

(3) $f(x,y) = x^2 - y^2$. This is the same as Example 7.29, item (1).
We compute:

$$f_x(x,y) = 2x \quad \text{and} \quad f_y(x,y) = -2y,$$

$$f_{xx}(x,y) = 2, \quad f_{xy}(x,y) = f_{yx}(x,y) = 0, \quad f_{yy}(x,y) = -2.$$

The gradient of f is zero at $(0,0)$. Further, $\Delta f(0,0) = -4 < 0$. Thus, f has a saddle point at $(0,0)$.

(4) $f(x,y) = xy(x^2 - y^2)$. This is the same as Example 7.29, item (4).
We compute:

$$f_x(x,y) = 3x^2y - y^3 \quad \text{and} \quad f_y(x,y) = x^3 - 3xy^2,$$

$$f_{xx}(x,y) = 6xy, \quad f_{xy}(x,y) = f_{yx}(x,y) = 3x^2 - 3y^2, \quad f_{yy}(x,y) = -6xy.$$

The gradient of f is zero at $(0,0)$. Further, $\Delta f(0,0) = 0$, and the test is inconclusive. However, f has a saddle point at $(0,0)$.

From the above examples, we observe that the maxima-minima-saddlepoint test is inconclusive for functions such as $f(x,y) = x^4 + y^4$ (involving high powers of x and y) since $\Delta f(0,0) = 0$. This is similar to the one variable case in which the maxima-minima test is inconclusive for functions such as $f(x) = x^4$ since $f''(0) = 0$.

7.3.5. Critical points and global maxima/minima. Let $f : D \to \mathbb{R}$. An interior point (a, b) of D is a *critical point* of f if either ∇f does not exist at (a, b), or if ∇f exists at (a, b) and is equal to 0.

Lemma 7.33. *Let D be closed and bounded, and $f : D \to \mathbb{R}$ be continuous. Then the global minimum and global maximum of f are attained at points which are either critical points of f or boundary points of D.*

PROOF. The same argument as for Lemma 4.31 works. Suppose $f(a, b)$ is a global maximum. We consider two cases.

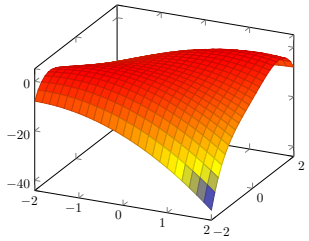
- (a, b) is a boundary point. Then we are fine.
- (a, b) is an interior point. We consider two subcases.
  - ∇f does not exist at (a, b). Then (a, b) is a critical point of f.
  - ∇f exists at (a, b). Since f has a global maximum at (a, b), it has a local maximum at (a, b). Hence $\nabla f(a, b) = 0$ by Lemma 7.30, and (a, b) is a critical point of f.

The argument for a global minimum is similar. □

Example 7.34. Consider the rectangle $D = [-2, 2] \times [-2, 2]$, and

$$f : D \to \mathbb{R}, \qquad f(x, y) = 4xy - 2x^2 - y^4.$$

Its graph is shown below.



Note: D is closed and bounded, and f is continuous on D. We compute:

$$f_x(x, y) = 4y - 4x \quad \text{and} \quad f_y(x, y) = 4x - 4y^3.$$

Thus,

$$\nabla f(x, y) = (0, 0) \iff (x, y) = (0, 0), (1, 1), (-1, -1).$$

These are the critical points of f.

Now let us analyze f on the boundary of D.

- Let $h(y) = f(2, y) = 8y - 8 - y^4$ for $y \in [-2, 2]$. Thus, $h'(y) = 8 - 4y^3 = 0$ iff $y = 2^{1/3}$. Hence, we need to consider $(2, 2^{1/3})$, and also endpoints $(2, -2)$ and $(2, 2)$.
- Let $g(x) = f(x, 2) = 8x - 2x^2 - 16$ for $x \in [-2, 2]$. Thus, $g'(x) = 8 - 4x = 0$ iff $x = 2$. Hence, we need to consider $(2, 2)$ and also the other endpoint $(-2, 2)$.

Due to the symmetry $f(-x, -y) = f(x, y)$, we do not need to analyze the other two sides separately.

(x, y)	$(0, 0)$	$(1, 1)$	$(2, -2)$	$(2, 2)$	$(2, 2^{1/3})$
$f(x, y)$	0	1	-40	-8	$6 \times 2^{1/3} - 8$

Thus, we see f has global maximum 1 attained at $(x, y) = (1, 1)$ and $(x, y) = (-1, -1)$, and f has global minimum -40 attained at $(x, y) = (2, -2)$ and $(x, y) = (-2, 2)$.

7.3.6. Contrained extrema. Let (a, b) be an interior point of D. Suppose $f, g : D \to \mathbb{R}$ is such that partial derivatives of f and g exist and are continuous in a neighborhood of (a, b).

Lemma 7.35 (Lagrange multiplier). *Let $C = \{(x, y) \in D \colon g(x, y) = 0\}$, the zero set of g. Suppose*

(1) $g(a, b) = 0$,
(2) $\nabla g(a, b) \neq (0, 0)$,
(3) *the function f restricted to the curve C has a local extremum at (a, b).*

Then $\nabla f(a, b) = \lambda \nabla g(a, b)$ for some real number λ.

Example 7.36. Let $f(x, y) = xy$ and $g(x, y) = x^2 + y^2 - 1$. Thus, we solve for

$$y = 2\lambda x, \quad x = 2\lambda y, \quad x^2 + y^2 - 1 = 0.$$

These imply $4\lambda^2 = 1$, thus $2\lambda = \pm 1$. Thus, the points which solve the above equations are

$$(\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}}).$$

Note: $\nabla g \neq (0, 0)$ on the unit circle. Also, the unit circle is closed and bounded, so f attains its global maximum and global minimum on it. By Lemma 7.35, these can only occur at the above four points. It is easy to check:

- f has a global maximum of $\frac{1}{2}$ at $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$,
- f has a global minimum of $-\frac{1}{2}$ at $(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$.

Puzzle 7.37. Consider a 10×10 array of soldiers. Assume that each soldier has a distinct height. From each column, locate the tallest soldier, and let P denote the shortest soldier among these 10 soldiers. Similarly, from each row, locate the shortest soldier, and let D denote the tallest soldier among these 10 soldiers. Who among P and D is taller, and who is shorter?

CHAPTER 8

Integration

8.1. Riemann integral on a rectangle

We studied integrability of functions f of a real variable in Section 5.1. We now look at integrability of functions f of two real variables. The intuitive idea of the integral of a function f is the volume under the graph of f. We also encounter the notion of iterated integrals.

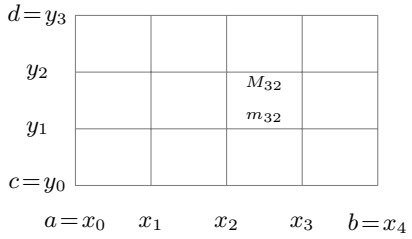
8.1.1. Riemann integrable functions. Let $a < b$ and $c < d$ be real numbers, and $f : [a, b] \times [c, d] \to \mathbb{R}$ be a bounded function. Let M be the global maximum of f, and m the global minimum of f. A partition P of $[a, b] \times [c, d]$ is defined by

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b,$$
$$c = y_0 < y_1 < \cdots < y_{k-1} < y_k = d.$$

The norm of partition P is defined as

$$\|P\| := \max\{x_i - x_{i-1}, y_j - y_{j-1} : 1 \le i \le n, 1 \le j \le k\}.$$

We think of P as a subdivision of the rectangle $[a, b] \times [c, d]$ into smaller rectangles $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. An illustration with $n = 4$ and $k = 3$ is shown below.



For each $1 \le i \le n$ and $1 \le j \le k$, let M_{ij} be the global maximum of f on $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$, and m_{ij} the global minimum of f on $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. Let

$$U(P, f) = \sum_{i=1}^{n} \sum_{j=1}^{k} M_{ij}(x_i - x_{i-1})(y_j - y_{j-1}),$$
$$L(P, f) = \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}(x_i - x_{i-1})(y_j - y_{j-1}).$$

We call $U(P, f)$ the *upper sum*, and $L(P, f)$ the *lower sum*. Then

$$m(b - a)(d - c) \le L(P, f) \le U(P, f) \le M(b - a)(d - c).$$

Definition 8.1. A bounded function $f : [a,b] \times [c,d] \to \mathbb{R}$ is *Riemann integrable* if there is a sequence $\{P_n\}$ of partitions of $[a,b] \times [c,d]$ such that

$$U(P_n, f) - L(P_n, f) \to 0 \quad \text{as} \quad n \to \infty.$$

8.1.2. Riemann integral on a rectangle. Let $f : [a,b] \times [c,d] \to \mathbb{R}$ be Riemann integrable.

Proposition 8.2. *There is a real number A such that*

$$L(P, f) \leq A \leq U(P, f)$$

for every partition P of $[a,b] \times [c,d]$, and

$$\lim_{n\to\infty} L(P_n, f) = A = \lim_{n\to\infty} U(P_n, f)$$

for every sequence $\{P_n\}$ of partitions of $[a,b] \times [c,d]$ with $\|P_n\| \to 0$.

We write

$$\iint\limits_{[a,b]\times[c,d]} f(x,y)\, d(x,y) = A,$$

and call it the *Riemann integral* of f.

8.1.3. Riemann sums. For P a partition of $[a,b] \times [c,d]$, let

$$S(P, f) = \sum_{i=1}^{n} \sum_{j=1}^{k} f(s_i, t_j)(x_i - x_{i-1})(y_j - y_{j-1})$$

where $s_i \in [x_{i-1}, x_i]$ for $1 \leq i \leq n$ and $t_j \in [y_{j-1}, y_j]$ for $1 \leq j \leq k$. We call $S(P, f)$ a *Riemann sum*.

Observe:

$$L(P, f) \leq S(P, f) \leq U(P, f).$$

In words, any Riemann sum lies between the lower and upper sums.

Proposition 8.3. *Suppose*

- *f is Riemann integrable on $[a,b] \times [c,d]$,*
- *$\{P_n\}$ is a sequence of partitions of $[a,b] \times [c,d]$ with $\|P_n\| \to 0$,*
- *$S(P_n, f)$ is any Riemann sum for P_n and f.*

Then

$$S(P_n, f) \to \iint\limits_{[a,b]\times[c,d]} f(x,y)\, d(x,y)$$

as $n \to \infty$.

Proof. By Proposition 8.2, sequences $L(P_n, f)$ and $U(P_n, f)$ have the same limit, namely, $\iint_{[a,b]\times[c,d]} f(x,y)\, d(x,y)$. Now apply sandwich Lemma 2.15 to $L(P_n, f) \leq S(P_n, f) \leq U(P_n, f)$. □

8.1.4. Monotone functions. Recall monotonic functions from Definition 1.7.

Lemma 8.4. *If $f : [a,b]\times[c,d] \to \mathbb{R}$ is monotonic in each of the two variables, then f is Riemann integrable.*

PROOF. The argument given for Lemma 5.5 generalizes. For details, see [13, Proposition 5.12, item (i)]. □

Example 8.5. Let us illustrate Lemma 8.4. The function $f(x) = [x] + [y]$ on $[a,b] \times [c,d]$ is monotonically increasing in each of the two variables. So it is Riemann integrable.

8.1.5. Continuous functions.

Lemma 8.6. *If $f : [a,b] \times [c,d] \to \mathbb{R}$ is bounded, and has at most finitely many discontinuities, then f is Riemann integrable.*

In particular, if $f : [a,b] \times [c,d] \to \mathbb{R}$ is continuous, then f is Riemann integrable.

PROOF. See [2, Theorems 11.16 and 11.17] or [13, Proposition 5.12, item (ii) and Lemma 5.41]. □

Example 8.7. Let us illustrate Lemma 8.6.

(1) Any polynomial function $p : [a,b] \times [c,d] \to \mathbb{R}$ such as $p(x,y) = x^2y + 3xy - 1$ is continuous, and hence Riemann integrable.
(2) The functions $f : [a,b] \times [c,d] \to \mathbb{R}$ such as $f(x,y) = \sin xy$ or $f(x,y) = \cos xy$ are continuous, and hence Riemann integrable.
(3) The functions with saddle points in Example 7.29 are continuous, and hence Riemann integrable.

8.1.6. Fubini's theorem. The next question is how does one evaluate a double integral. For a single integral, we use FTC, Part II, that is, formula (5.4). We will see this technique later when we develop FTC in higher dimensions in Chapter 9. (Evaluating limits of Riemann sums is a cumbersome task even in one variable, so we do not pursue that line of thought.)

However, there is another way to evaluate a double integral. The idea is to evaluate two single integrals one after the other as explained below.

Theorem 8.8. *Let $f : [a,b] \times [c,d] \to \mathbb{R}$ be Riemann integrable.*

(i) *If $\int_c^d f(x,y)\,dy$ exists for each $x \in [a,b]$, then $\int_a^b \left(\int_c^d f(x,y)\,dy\right) dx$ exists and equals $\iint_{[a,b]\times[c,d]} f(x,y)\,d(x,y)$.*
(ii) *If $\int_a^b f(x,y)\,dx$ exists for each $y \in [c,d]$, then $\int_c^d \left(\int_a^b f(x,y)\,dx\right) dy$ exists and equals $\iint_{[a,b]\times[c,d]} f(x,y)\,d(x,y)$.*

PROOF. We omit the proof. See [13, Proposition 5.28]. □

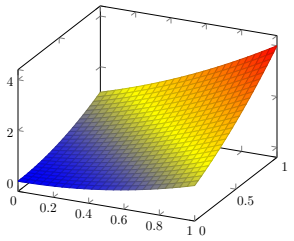
If the hypotheses in both (i) and (ii) above hold, then

$$\int_a^b \left(\int_c^d f(x,y)\,dy\right) dx = \iint\limits_{[a,b]\times[c,d]} f(x,y)\,d(x,y) = \int_c^d \left(\int_a^b f(x,y)\,dx\right) dy.$$

This holds, in particular, if f is continuous.

Example 8.9. Let us illustrate Theorem 8.8.

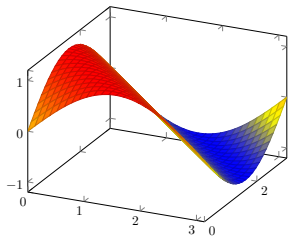
(1) Let $f : [0,1] \times [0,1] \to \mathbb{R}$ be given by $f(x,y) = (x+y)^2$. Its graph is shown below.



Let us now compute the volume below this surface.

$$\begin{aligned}\iint\limits_{[0,1]\times[0,1]} (x+y)^2\, d(x,y) &= \int_0^1 \left(\int_0^1 (x+y)^2 dx \right) dy \\ &= \int_0^1 \frac{(x+y)^3}{3}\Big|_0^1 dy \\ &= \frac{1}{3}\int_0^1 [(1+y)^3 - y^3]\, dy \\ &= \frac{1}{12}(2^4 - 1 - 1) \\ &= \frac{7}{6}.\end{aligned}$$

(2) Let $f : [0,\pi] \times [0,\pi] \to \mathbb{R}$ be given by $f(x,y) = \sin(x+y)$. Its graph is shown below.



Then

$$\begin{aligned}\iint\limits_{[0,\pi]\times[0,\pi]} \sin(x+y)\, d(x,y) &= \int_0^\pi \left(\int_0^\pi \sin(x+y)\, dy \right) dx \\ &= \int_0^\pi [-\cos(x+\pi) + \cos x]\, dx \\ &= 2\int_0^\pi \cos x\, dx \\ &= 0.\end{aligned}$$

This makes sense since part of the graph is above and part of it is below the xy-plane.

Special case. Suppose $f : [a,b] \times [c,d] \to \mathbb{R}$ is given by $f(x,y) := \varphi(x)\psi(y)$ for some Riemann integrable functions $\varphi : [a,b] \to \mathbb{R}$ and $\psi : [c,d] \to \mathbb{R}$. Then

$$\iint\limits_{[a,b]\times[c,d]} \varphi(x)\psi(y)\, d(x,y) = \left(\int_a^b \varphi(x)\, dx\right)\left(\int_c^d \psi(y)\, dy\right). \tag{8.1}$$

To prove (8.1),

$$\begin{aligned}\iint\limits_{[a,b]\times[c,d]} \varphi(x)\psi(y)\, d(x,y) &= \int_a^b \Big(\int_c^d \varphi(x)\psi(y)\, dy\Big)\, dx \\ &= \int_a^b \varphi(x)\Big(\int_c^d \psi(y)\, dy\Big)\, dx \\ &= \left(\int_a^b \varphi(x)\, dx\right)\left(\int_c^d \psi(y)\, dy\right).\end{aligned}$$

Example 8.10. For real numbers $r, s \geq 0$,

$$\iint\limits_{[a,b]\times[c,d]} x^r y^s\, d(x,y) = \left(\frac{b^{r+1} - a^{r+1}}{r+1}\right)\left(\frac{d^{s+1} - c^{s+1}}{s+1}\right)$$

assuming $0 < a < b$ and $0 < c < d$.

8.1.7. Application: computing limits. It is possible to evaluate limits of certain sequences by interpreting their terms as Riemann sums. We illustrate with an example:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n^4} \sum_{i=1}^n \sum_{j=1}^n (i+j)^2 &= \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \big(\frac{i}{n} + \frac{j}{n}\big)^2 \\ &= \iint\limits_{[0,1]\times[0,1]} (x+y)^2\, d(x,y) \\ &= \frac{7}{6}.\end{aligned}$$

8.2. Riemann integral in the plane

In Section 8.1, we looked at Riemann integral on the rectangle. Now we consider Riemann integral on more general regions in the plane such as the closed unit disc. This also allows us to formalize the notion of area of a region in the plane.

8.2.1. Riemann integral on a general region. Let D be a bounded subset of $\mathbb{R}^2$, and let $f : D \to \mathbb{R}$ be a bounded function. Pick a rectangle $[a,b] \times [c,d]$ containing D, and define

$$f^* : [a,b] \times [c,d] \to \mathbb{R}, \qquad f^*(x,y) := \begin{cases} f(x,y) & \text{if } (x,y) \in D, \\ 0 & \text{otherwise.}\end{cases}$$

We say f is *Riemann integrable* on D if f^* is Riemann integrable on $[a,b]\times[c,d]$, and in this case, we define

$$\iint_D f(x,y)\,d(x,y) := \iint_{[a,b]\times[c,d]} f^*(x,y)\,d(x,y). \tag{8.2}$$

This definition is independent of the choice of the rectangle containing D.

8.2.2. Algebra of Riemann integrable functions.

Lemma 8.11. *Suppose $f,g : D \to \mathbb{R}$ are Riemann integrable. Then*

(i) *$f+g$ is Riemann integrable, and*

$$\iint_D (f+g)(x,y)\,d(x,y) = \iint_D f(x,y)\,d(x,y) + \iint_D g(x,y)\,d(x,y),$$

(ii) *rf is Riemann integrable, and*

$$\iint_D (rf)(x,y)\,d(x,y) = r\iint_D f(x,y)\,d(x,y)$$

for $r\in\mathbb{R}$,

(iii) *fg is Riemann integrable,*

(iv) *$1/f$ is Riemann integrable if there is $\delta>0$ such that $|f(x,y)|\geq\delta$ for $(x,y)\in D$ (so that $1/f$ is bounded).*

Proof. We omit the proof. See [13, Proposition 5.34]. □

In continuation of Example 8.10: By items (i) and (ii), Riemann integral is linear, so we can now evaluate the Riemann integral of any polynomial in x and y.

8.2.3. Elementary regions.

Let $\varphi_1,\varphi_2 : [a,b]\to\mathbb{R}$ be continuous such that $\varphi_1\leq\varphi_2$, and define

$$D := \{(x,y)\in\mathbb{R}^2 : a\leq x\leq b,\ \varphi_1(x)\leq y\leq\varphi_2(x)\}. \tag{8.3}$$

Similarly, let $\psi_1,\psi_2 : [c,d]\to\mathbb{R}$ be continuous such that $\psi_1\leq\psi_2$, and define

$$D := \{(x,y)\in\mathbb{R}^2 : c\leq y\leq d,\ \psi_1(y)\leq x\leq\psi_2(y)\}. \tag{8.4}$$

In both cases, we say D is an *elementary region* in $\mathbb{R}^2$.

Theorem 8.12 (Fubini's theorem over elementary regions).

(i) *For elementary region D of the form (8.3),*

$$\iint_D f(x,y)\,d(x,y) = \int_a^b \Big(\int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,dy\Big)\,dx.$$

(ii) *For elementary region D of the form (8.4),*

$$\iint_D f(x,y)\,d(x,y) = \int_c^d \Big(\int_{\psi_1(y)}^{\psi_2(y)} f(x,y)\,dx\Big)\,dy.$$

PROOF. This follows from Fubini's Theorem 8.8 over a rectangle. For instance,

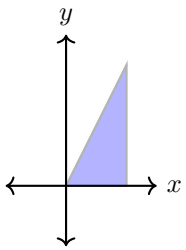
$$\begin{aligned}\iint_D f(x,y)\,d(x,y) &= \iint_{[a,b]\times[c,d]} f^*(x,y)\,d(x,y)\\ &= \int_a^b \Big(\int_c^d f^*(x,y)\,dy\Big)\,dx\\ &= \int_a^b \Big(\int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\,dy\Big)\,dx.\end{aligned}$$

Here c and d are chosen so that $c \le \varphi_1(x) \le \varphi_2(x) \le d$. In the last step, we improved the integration limits since f^* is zero when y is not between $\varphi_1(x)$ and $\varphi_2(x)$. □

Example 8.13. Consider $f : D \to \mathbb{R}$ with $f(x,y) = e^{x^2}$ and

$$D = \{(x,y) \in \mathbb{R}^2 : 0 \le x \le 1,\ 0 \le y \le 2x\}.$$

See picture below.



By Fubini's Theorem 8.12, item (i),

$$\iint_D f(x,y)\,d(x,y) = \int_0^1 \left(\int_0^{2x} e^{x^2}\,dy\right)\,dx = \int_0^1 2x\,e^{x^2}\,dx = e-1.$$

Also, note

$$D = \{(x,y) \in \mathbb{R}^2 : 0 \le y \le 2,\ y/2 \le x \le 1\}.$$

So, by Fubini's Theorem 8.12, item (ii),

$$\iint_D f(x,y)\,d(x,y) = \int_0^2 \left(\int_{y/2}^1 e^{x^2}\,dx\right)\,dy = ??.$$

It is not clear how to proceed.

8.2.4. Area of a general region. Let D be a bounded subset of $\mathbb{R}^2$. Let 1_D denote the function which is identically 1 on D. Suppose 1_D is Riemann integrable. Then define

$$\text{Area}(D) = \iint_D 1_D\,d(x,y), \tag{8.5}$$

with rhs defined by (8.2).

By Fubini's Theorem 8.12, this definition is consistent with the earlier definition of area under the curve given by $y = f(x)$, or by $x = g(y)$. See

formulas (5.7) and (5.8). Explicitly, for area between the curves $y = f_1(x)$ and $y = f_2(x)$ with $f_1(x) \le f_2(x)$ for $a \le x \le b$,

$$\text{Area}(D) = \iint_D 1_D \, d(x,y) = \int_a^b \Big(\int_{f_1(x)}^{f_2(x)} 1 \, dy \Big) dx = \int_a^b f_2(x) - f_1(x) \, dx.$$

This is the same as (5.7).

8.3. Change of variables

Just as in the one variable case, the calculation of a double integral can often be simplified by making a substitution, that is, a change of variables. This can either simplify the region over which we are integrating or simplify the function that we are integrating or both. A particular example that we consider is that of polar coordinates.

8.3.1. Jacobian matrix. Consider the linear map

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^2, \qquad (u,v) \mapsto (a_1 u + b_1 v, a_2 u + b_2 v). \tag{8.6}$$

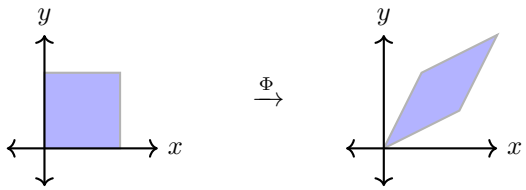
Define the *jacobian* $J(\Phi)$ of Φ by

$$J(\Phi) = \det \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} = a_1 b_2 - a_2 b_1. \tag{8.7}$$

We refer to the above 2×2 matrix as the *jacobian matrix* of Φ. Observe: Φ is a bijection iff $J(\Phi) = a_1 b_2 - a_2 b_1 \neq 0$. In this case, the unit square $E = [0,1] \times [0,1]$ maps to the parallelogram $D = \Phi(E)$ with vertices

$$\Phi(0,0) = (0,0), \quad \Phi(1,0) = (a_1, a_2),$$
$$\Phi(0,1) = (b_1, b_2), \quad \Phi(1,1) = (a_1 + b_1, a_2 + b_2).$$

See picture below.



Thus Φ scales area by $|J(\Phi)|$.

Definition 8.14 (Jacobian matrix). Let Ω be an open subset of $\mathbb{R}^2$ (that is, each point of Ω is an interior point of Ω). Let

$$\Phi : \Omega \to \mathbb{R}^2, \qquad \Phi(u,v) = (\varphi_1(u,v), \varphi_2(u,v)), \tag{8.8}$$

where partial derivatives of φ_1 and φ_2 exist, and are continuous. Define the *jacobian* $J(\Phi)$ of Φ by

$$J(\Phi)(a,b) = \det \begin{bmatrix} \frac{\partial \varphi_1}{\partial u}(a,b) & \frac{\partial \varphi_1}{\partial v}(a,b) \\ \frac{\partial \varphi_2}{\partial u}(a,b) & \frac{\partial \varphi_2}{\partial v}(a,b) \end{bmatrix}. \tag{8.9}$$

This depends on the point (a,b). We refer to the above 2×2 matrix as the *jacobian matrix* of Φ.

In the special case when Φ is a linear map, formula (8.9) reduces to formula (8.7) which is a constant (with no dependence on (a,b)). In the general case, the area scaling depends on (a,b).

8.3.2. Change of variables formula.

Theorem 8.15. *Let map Φ be as in (8.8), with Φ injective and $J(\Phi)(u,v) \neq 0$ for all $(u,v) \in \Omega$. Let D be an elementary region and let $f : D \to \mathbb{R}$ be continuous. Let $E \subseteq \Omega$ be such that $\Phi(E) = D$. Then*

$$f \circ \Phi : E \xrightarrow{\Phi} D \xrightarrow{f} \mathbb{R}$$

is Riemann integrable, and

$$\iint_D f(x,y)\, d(x,y) = \iint_E (f \circ \Phi)(u,v) |J(\Phi)(u,v)|\, d(u,v). \tag{8.10}$$

PROOF. We omit the proof. See [13, Proposition 5.61] for more detail. □

Note: The following weakening of the hypothesis is permitted. Instead of $J(\Phi)(u,v) \neq 0$ for all $(u,v) \in \Omega$, it suffices to assume either $J(\Phi)(u,v) \geq 0$ or $J(\Phi)(u,v) \leq 0$ for all $(u,v) \in \Omega$, and $J(\Phi)(u,v) = 0$ only on a 'thin' subset of Ω like a point or a line segment or a curve $v = \psi(u)$.

Example 8.16. Consider the trapezium

$$D = \{(x,y) \in \mathbb{R}^2 : x \geq 0,\ y \geq 0,\ \frac{1}{2} \leq x + y \leq 1\}.$$

and define

$$f : D \to \mathbb{R}, \qquad f(x,y) = \frac{y}{x+y}.$$

We want to find the Riemann integral of f. For that, we make a change of variables as follows. Let

$$u = x + y,\ v = \frac{y}{x+y},\ \text{that is, } x = u(1-v),\ y = uv.$$

We let $\Omega = \{(u,v) \in \mathbb{R}^2 : u > 0\}$, and define

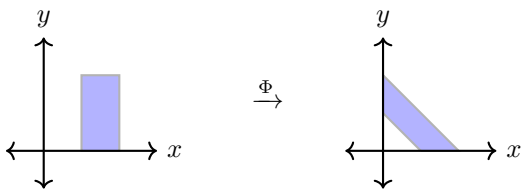
$$\Phi : \Omega \to \mathbb{R}^2, \qquad \Phi(u,v) = (u(1-v), uv).$$

Then the map Φ is injective and onto $\{(x,y) \in \mathbb{R}^2 : x + y > 0\}$. Also, by formula (8.9),

$$J(\Phi)(u,v) = \det \begin{bmatrix} 1-v & -u \\ v & u \end{bmatrix} = u \neq 0$$

for all $(u,v) \in \Omega$.

Further, if $E = [1/2, 1] \times [0,1]$, then $\Phi(E) = D$. See picture below.



By formula (8.10),

$$\begin{aligned}\iint_D f(x,y)\,d(x,y) &= \iint_E f(u(1-v),uv)|u|\,d(u,v)\\ &= \iint_E uv\,d(u,v)\\ &= \left(\int_{1/2}^1 u\,du\right)\left(\int_0^1 v\,dv\right)\\ &= \frac{3}{16}.\end{aligned}$$

8.3.3. Polar coordinates. We now discuss polar coordinates. These are useful to compute double integrals of functions or over regions which have a circular symmetry.

Define

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^2, \qquad \Phi(r,\theta) := (r\cos\theta, r\sin\theta). \tag{8.11}$$

Let us compute the jacobian $J(\Phi)$ of Φ. By formula (8.9),

$$J(\Phi)(r,\theta) = \det\begin{bmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{bmatrix} = r.$$

Under suitable hypothesis,

$$\iint_D f(x,y)\,d(x,y) = \iint_E f(r\cos\theta, r\sin\theta)\,r\,d(r,\theta). \tag{8.12}$$

As formulated, this does not follow from formula (8.10) since the map Φ above is not injective.

Example 8.17. Consider the closed unit disc

$$D = \{(x,y)\in\mathbb{R}^2 \colon x^2+y^2\le 1\}.$$

Define

$$\begin{aligned}E &= [0,1]\times[-\pi,\pi]\\ &= \{(r,\theta)\in\mathbb{R}^2 \colon r\ge 0, -\pi\le\theta\le\pi,\ (r\cos\theta, r\sin\theta)\in D\}.\end{aligned}$$

Thus $\Phi(E) = D$.

We now consider two examples of $f : D\to\mathbb{R}$.

(i) Let $f(x,y) = \sqrt{1-x^2-y^2}$. By formula (8.12),

$$\begin{aligned}\iint_D f(x,y)\,d(x,y) &= \iint_E f(r\cos\theta, r\sin\theta)\,r\,d(r,\theta)\\ &= \int_{-\pi}^{\pi}\left(\int_0^1 \sqrt{1-r^2}\,r\,dr\right)d\theta\\ &= \int_{-\pi}^{\pi}\frac{1}{2}\left(\int_0^1\sqrt{s}\,ds\right)d\theta\\ &= \frac{2\pi}{3}.\end{aligned}$$

(ii) Let $f(x,y) = e^{x^2+y^2}$. By formula (8.12),

$$\begin{aligned}\iint_D f(x,y)\,d(x,y) &= \iint_E f(r\cos\theta, r\sin\theta)\,r\,d(r,\theta)\\ &= \int_{-\pi}^{\pi}\left(\int_0^1 e^{r^2} r\ dr\right) d\theta\\ &= \int_{-\pi}^{\pi}\left(\frac{e-1}{2}\right) d\theta\\ &= \pi(e-1).\end{aligned}$$

8.4. Riemann integral in space

We now go from two to three dimensions by following the general procedure in the preceding sections. We begin with Riemann integral on a cuboid. Then we go to Riemann integral on more general regions in space such as the solid cylinder or unit solid sphere. This also allows us to formalize the notion of volume of a region in space. We also see how change of variables works in three dimensions. Particular examples that we consider are that of cylindrical coordinates and spherical coordinates.

8.4.1. Riemann integral on a cuboid. Let $a < b$, $c < d$, $p < q$ be real numbers, and $f : [a,b] \times [c,d] \times [p,q] \to \mathbb{R}$ be a bounded function. We then define a partition P of $[a,b] \times [c,d] \times [p,q]$, upper sum $U(P,f)$, lower sum $L(P,f)$, norm $\|P\|$, Riemann sums $S(P,f)$, Riemann integral

$$\iiint\limits_{[a,b]\times[c,d]\times[p,q]} f(x,y,z)\,d(x,y,z).$$

Fubini's theorem takes the form

$$\iiint\limits_{[a,b]\times[c,d]\times[p,q]} f(x,y,z)\,d(x,y,z) = \int_a^b \Big(\int_c^d \Big(\int_p^q f(x,y,z)\,dz\Big)\,dy\Big)\,dx,$$

and so on.

8.4.2. Riemann integral on a general region. Let D be a bounded subset of $\mathbb{R}^3$, and let $f : D \to \mathbb{R}$ be a bounded function. Pick a rectangle $[a,b] \times [c,d] \times [p,q]$ containing D, and define

$$f^* : [a,b] \times [c,d] \times [p,q] \to \mathbb{R}, \qquad f^*(x,y,z) := \begin{cases} f(x,y,z) & \text{if } (x,y,z) \in D,\\ 0 & \text{otherwise,}\end{cases}$$

and

$$\iiint\limits_D f(x,y,z)\,d(x,y,z) := \iiint\limits_{[a,b]\times[c,d]\times[p,q]} f^*(x,y,z)\,d(x,y,z). \tag{8.13}$$

Let D_0 be an elementary region in $\mathbb{R}^2$. Let $\varphi_1, \varphi_2 : D_0 \to \mathbb{R}$ be continuous such that $\varphi_1 \le \varphi_2$, and define

$$D := \{(x,y,z) \in \mathbb{R}^3 \colon (x,y) \in D_0,\ \varphi_1(x,y) \le z \le \varphi_2(x,y)\}. \tag{8.14}$$

It is the region between the two surfaces $z = \varphi_1(x, y)$ and $z = \varphi_2(x, y)$. We make similar definitions, viewing D_0 in the xz-plane or in the yz-plane. In all cases, we say D is an *elementary region* in $\mathbb{R}^3$.

Fubini's theorem for elementary region (8.14) says

$$\iiint_D f(x, y, z)\, d(x, y, z) = \iint_{D_0} \Big(\int_{\varphi_1(x,y)}^{\varphi_2(x,y)} f(x, y, z)\, dz \Big)\, d(x, y).$$

Let D be a bounded subset of $\mathbb{R}^3$. Let 1_D denote the function which is identically 1 on D. Suppose 1_D is Riemann integrable. Then define

$$\operatorname{Vol}(D) = \iiint_D 1_D\, d(x, y, z), \tag{8.15}$$

with rhs defined by (8.13).

8.4.3. Change of variables.

Definition 8.18 (Jacobian matrix). Let Ω be an open subset of $\mathbb{R}^3$ (that is, each point of Ω is an interior point of Ω). Let

$$\Phi : \Omega \to \mathbb{R}^3, \qquad \Phi(u, v, w) = (\varphi_1(u, v, w), \varphi_2(u, v, w), \varphi_3(u, v, w)), \tag{8.16}$$

where partial derivatives of φ_1, φ_2, φ_3 exist, and are continuous. Define the *jacobian* $J(\Phi)$ of Φ by

$$J(\Phi)(a, b, c) = \det \begin{bmatrix} \frac{\partial \varphi_1}{\partial u}(a, b, c) & \frac{\partial \varphi_1}{\partial v}(a, b, c) & \frac{\partial \varphi_1}{\partial w}(a, b, c) \\ \frac{\partial \varphi_2}{\partial u}(a, b, c) & \frac{\partial \varphi_2}{\partial v}(a, b, c) & \frac{\partial \varphi_2}{\partial w}(a, b, c) \\ \frac{\partial \varphi_3}{\partial u}(a, b, c) & \frac{\partial \varphi_3}{\partial v}(a, b, c) & \frac{\partial \varphi_3}{\partial w}(a, b, c) \end{bmatrix}. \tag{8.17}$$

We refer to the above 3×3 matrix as the *jacobian matrix* of Φ.

If Φ is a linear map, then $J(\Phi)$ is a constant. Its absolute value is the volume of the parallelopiped which is the image of the unit cube. Thus, $|J(\Phi)|$ is the factor by which volumes get scaled.

Theorem 8.19. *Let map Φ be as in (8.16), with Φ injective and $J(\Phi)(u, v, w) \neq 0$ for all $(u, v, w) \in \Omega$. Let D be an elementary region and let $f : D \to \mathbb{R}$ be continuous. Let $E \subseteq \Omega$ be such that $\Phi(E) = D$. Then*

$$f \circ \Phi : E \xrightarrow{\Phi} D \xrightarrow{f} \mathbb{R}.$$

is Riemann integrable, and

$$\iiint_D f(x, y, z)\, d(x, y, z) = \iiint_E (f \circ \Phi)(u, v, w) |J(\Phi)(u, v, w)|\, d(u, v, w). \tag{8.18}$$

Proof. We omit the proof. See [13, Proposition 5.71] for more detail. □

8.4.4. Cylindrical coordinates.

We now discuss cylindrical coordinates. These are useful to compute triple integrals of functions or over regions which have a circular symmetry in the first two coordinates.

Define $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$\Phi(r, \theta, z) := (r \cos\theta,\ r \sin\theta,\ z). \tag{8.19}$$

Let us compute the jacobian $J(\Phi)$ of Φ. By formula (8.17),

$$J(\Phi)(r,\theta,z) = \det \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = r.$$

The function

$$\Phi : \{(r,\theta,z) \in \mathbb{R}^3 \colon r > 0,\ -\pi < \theta \le \pi\} \to \{(x,y,z) \in \mathbb{R}^3 \colon (x,y) \neq (0,0)\}$$

is a bijection.

Therefore, under suitable hypothesis,

$$\iiint_D f(x,y,z)\, d(x,y,z) = \iiint_E f(r\cos\theta, r\sin\theta, z) r\, d(r,\theta,z). \tag{8.20}$$

This formula arises from formula (8.18), by substituting the jacobian that we computed above.

Example 8.20. Consider the solid cylinder

$$D = \{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 \le 1,\ 0 \le z \le 1\},$$

and the function

$$f : D \to \mathbb{R}, \qquad f(x,y,z) = z\sqrt{1-x^2-y^2}.$$

Define

$$\begin{aligned} E &= [0,1] \times [-\pi,\pi] \times [0,1] \\ &= \{(r,\theta,z) \in \mathbb{R}^3 \colon r \ge 0,\ -\pi \le \theta \le \pi,\ (r\cos\theta, r\sin\theta, z) \in D\}. \end{aligned}$$

Thus $\Phi(E) = D$. Hence, by formula (8.20),

$$\begin{aligned} \iiint_D f(x,y,z)\, d(x,y,z) &= \iiint_E f(r\cos\theta, r\sin\theta, z)\, r\, d(r,\theta,z) \\ &= \int_0^1 \left[\int_{-\pi}^{\pi} \left(\int_0^1 z\sqrt{1-r^2}\, r\, dz \right) d\theta \right] dr \\ &= \pi \int_0^1 \sqrt{1-r^2}\, r\, dr \\ &= \frac{\pi}{3}. \end{aligned}$$

8.4.5. Spherical coordinates. We now discuss spherical coordinates. These are useful to compute triple integrals of functions or over regions which have a spherical symmetry.

Define $\Phi : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$\Phi(\rho,\varphi,\theta) := (\rho\sin\varphi\cos\theta,\ \rho\sin\varphi\sin\theta,\ \rho\cos\varphi). \tag{8.21}$$

Let us compute the jacobian $J(\Phi)$ of Φ. By formula (8.17),

$$J(\Phi)(\rho,\varphi,\theta) = \det \begin{bmatrix} \sin\varphi\cos\theta & \rho\cos\varphi\cos\theta & -\rho\sin\varphi\sin\theta \\ \sin\varphi\sin\theta & \rho\cos\varphi\sin\theta & \rho\sin\varphi\cos\theta \\ \cos\varphi & -\rho\sin\varphi & 0 \end{bmatrix} = \rho^2\sin\varphi.$$

The function

$$\Phi : \{(\rho, \varphi, \theta) \in \mathbb{R}^3 : \rho > 0,\ 0 < \varphi < \pi,\ -\pi < \theta \leq \pi\} \longrightarrow \{(x, y, z) \in \mathbb{R}^3 : (x, y) \neq (0, 0)\}$$

is a bijection.

Therefore, under suitable hypothesis,

$$\iiint_D f(x, y, z)\, d(x, y, z) = \iiint_E f(\rho \sin\varphi \cos\theta, \rho \sin\varphi \sin\theta, \rho \cos\varphi)\rho^2 \sin\varphi\, d(\rho, \varphi, \theta). \tag{8.22}$$

This formula arises from formula (8.18), by substituting the jacobian that we computed above.

Example 8.21. Consider the unit solid sphere

$$D = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + z^2 \leq 1\}$$

and the function

$$f : D \to \mathbb{R}, \qquad f(x, y, z) = z^2.$$

Define

$$\begin{aligned} E &= [0, 1] \times [0, \pi] \times [-\pi, \pi] \\ &= \{(\rho, \varphi, \theta) \in \mathbb{R}^3 : \rho \geq 0,\ 0 \leq \varphi \leq \pi,\ -\pi \leq \theta \leq \pi, \\ &\qquad (\rho \sin\varphi \cos\theta, \rho \sin\varphi \sin\theta, \rho \cos\varphi) \in D\}. \end{aligned}$$

Thus $\Phi(E) = D$. Hence, by formula (8.22),

$$\begin{aligned} \iiint_D f(x, y, z)\, d(x, y, z) &= \iiint_E f(\rho \sin\varphi \cos\theta, \rho \sin\varphi \sin\theta, \rho \cos\varphi)\, \rho^2 \sin\varphi\, d(\rho, \varphi, \theta) \\ &= \int_0^1 \left[\int_0^\pi \left(\int_{-\pi}^\pi \left(\rho^2 \cos^2\varphi\right) \rho^2 \sin\varphi\, d\theta \right) d\varphi \right] d\rho \\ &= 2\pi \int_0^1 \rho^4 \left(\int_0^\pi \cos^2\varphi \sin\varphi\, d\varphi \right) d\rho \\ &= \frac{2\pi}{5} \cdot \frac{2}{3} \\ &= \frac{4\pi}{15}. \end{aligned}$$

CHAPTER 9

Differential forms

We studied FTC in one real variable in Section 5.2 which says that differentiation and integration are inverse processes. We would now like to extend this result to higher dimensions. We will mainly focus on three different special cases, namely, Green's theorem, Gauss's theorem, Stokes theorem. Their formulations involve line and surface integrals. A summary is given in Table 9.1.

TABLE 9.1. Various avatars of FTC.

	curve C, $\dim C = 1$	surface S, $\dim S = 2$	solid W, $\dim W = 3$
1	Theorem 5.18		
2	Theorem 9.24, $m = 2$	Green's theorem 9.30	
3	Theorem 9.24, $m = 3$	Stokes theorem 9.61	Gauss's theorem 9.54

The first row indicates dimension of the object under consideration, namely, a curve, a surface, a solid. The first column indicates dimension of the ambient space in which the object lies. For example, a curve may lie in dimension 1, or dimension 2, or dimension 3, and so on.

The special cases of FTC in Table 9.1 can be unified into one result by using the notion of differential forms which we will indicate towards the end of the chapter.

In this chapter, we assume that all functions involved are smooth. That is, they are continuous, differentiable, and so on.

9.1. Scalar and vector fields

We introduce scalar and vector fields. A scalar field on $\mathbb{R}^m$ is the same as a real-valued function on $\mathbb{R}^m$, while a vector field on $\mathbb{R}^m$ is the same as a $\mathbb{R}^m$-valued function on $\mathbb{R}^m$. These concepts may then look like an unnecessary renaming of known concepts, however the perspective is new. This becomes more apparent when we consider vector fields on spaces such as the circle, the sphere, and so on.

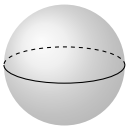
9.1.1. Scalar fields. Let $D \subseteq \mathbb{R}^m$. A function $f : D \to \mathbb{R}$ is called a *scalar field* on D. In other words, a scalar field on D is the same as a real-valued function on D.

We have seen many examples of real-valued functions, so each one of them yields an example of a scalar field.

Remark 9.1. It is customary to refer to a real number as a scalar. Thus, a scalar field on D specifies a scalar at each point of D (which explains the terminology). In many real-life examples, the scalar represent some physical quantity. For example, consider the unit sphere

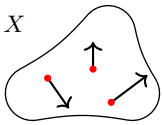
$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + z^2 = 1\}.$$

See picture below.



As an idealistic model, we may identify the surface of the earth with S^2. Then the temperature on earth's surface (at a fixed time) specifies a scalar field on S^2.

9.1.2. Vector fields. Informally: A vector field on a space X is a choice of a tangent vector at each point of X. The following is a way to visualize a vector field.



We have shown the tangent vectors only at three points, but one has to imagine a tangent vector at each point of X.

Remark 9.2 (Manifolds). One need to make precise what one means by space above. The technical term for space is differentiable manifold. See Section 9.8 in this regard. The technical setup for vector field is that of a tangent bundle of a manifold. For starting points, you may look at Boothby [6], do Carmo [10].

For $X = \mathbb{R}^m$, a vector field on X is the same as a function $F : \mathbb{R}^m \to \mathbb{R}^m$. We focus on the cases $m = 1$, $m = 2$, $m = 3$ below.

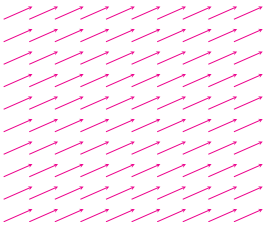
Example 9.3 (Vector fields on the line). For $X = \mathbb{R}$, a vector field on X is the same as a function $F : \mathbb{R} \to \mathbb{R}$. For example, let $F(x) = 1$. Then the vector at each point on the line points one unit to the right as illustrated below.

$$\longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow$$

Now let $F(x) = x$. Can you imagine this vector field? It is zero at the origin. At positive real numbers, vectors point to the right, while at negative real numbers, vectors point to the left. Their magnitudes increase linearly as we move away from the origin on either side.

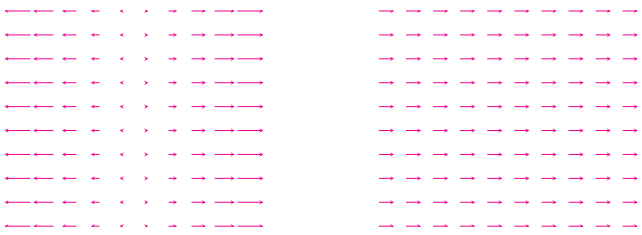
Example 9.4 (Vector fields on the plane). For $X = \mathbb{R}^2$, a vector field on X is the same as a function $F : \mathbb{R}^2 \to \mathbb{R}^2$. Many examples along with illustrations are shown below.

(1) Constant vector fields. Let $F(x, y) = (2, 1)$.



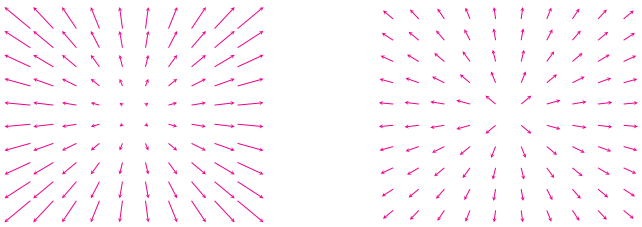
This is an example of a constant vector field.

(2) Horizontal vector fields. Let $F(x, y) = (x, 0)$. See left picture below. Vectors on y-axis are 0. For positive values of x, vectors point to the right. For negative values of x, vectors point to the left. Lengths of vectors increase as we move to the right or to the left.



Let $F(x, y) = (1, 0)$. This is a constant horizontal vector field where all vectors point one unit to the right. See right picture above.

(3) Radial vector fields. Let $F(x, y) = (x, y)$. See left picture below. Vectors point radially outward. Their lengths increase as we move outward.

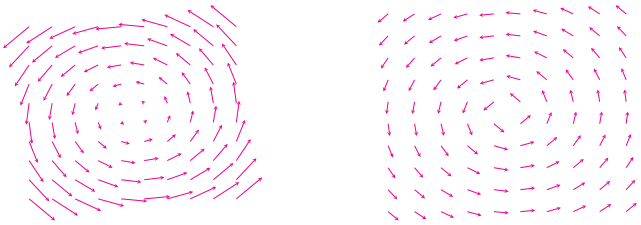


Let $F(x, y) = (\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}})$ on $\mathbb{R}^2 \setminus \{(0, 0)\}$. This is a variant of the previous example where vectors point radially outward and have length 1. See right picture above.

We may also consider $F(x, y) = (-x, -y)$, where vectors point radially inward, and its normalized version on $\mathbb{R}^2 \setminus \{(0, 0)\}$ where vectors also have length 1.

(4) Rotational vector fields. Let $F(x, y) = (-y, x)$. See left picture below. Vectors are orthogonal to the radial direction, and point in the

anticlockwise direction. Their lengths increase as we move outward.



Let $F(x,y) = (\frac{-y}{\sqrt{x^2+y^2}}, \frac{x}{\sqrt{x^2+y^2}})$ on $\mathbb{R}^2 \setminus \{(0,0)\}$. This is a variant of the previous example where vectors in addition have length 1. See right picture above.

Example 9.5 (Vector fields in space). For $X = \mathbb{R}^3$, a vector field on X is the same as a function $F : \mathbb{R}^3 \to \mathbb{R}^3$.

A vector field on $\mathbb{R}^2$ yields a vector field on $\mathbb{R}^3$ by setting the z-coordinate to zero. We can visualize it by translating the vector field in the xy-plane along the z-axis. For instance, for the rotational vector field in Example 9.4, item (4), we consider $F(x,y,z) = (y,-x,0)$. This can serve as a model for a fluid moving in $\mathbb{R}^3$ which is rotating about the z-axis.

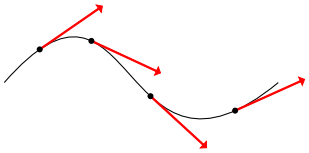
Example 9.6 (Vector fields on a curve). Consider the unit circle

$$S^1 = \{(x,y) \in \mathbb{R}^2 \colon x^2 + y^2 = 1\}.$$

A vector field on S^1 is shown below on the left. For simplicity, only four tangent vectors have been drawn.

The picture on the right cannot be part of a vector field on S^1 since there are vectors (two of them are shown) which are not tangential to the circle.

A way to visualize a vector field on a general curve is shown below.



Suppose a particle is moving along a curve X. Then the velocity vector of the particle yields a vector field on X.

Example 9.7 (Vector fields on a surface). Consider a surface X in $\mathbb{R}^3$. For example, X can be the unit sphere S^2, or X can be the graph $z = f(x,y)$ of a function f of two variables. For each point p on X, there is a plane tangent to X at p. To specify a vector field on X, we choose a vector from this tangent plane at p, and do this for each point p on X. Can you now imagine a vector field on a surface?

9.2. Gradient, curl, divergence

We now introduce three fundamental operations related to scalar and vector fields on $\mathbb{R}^3$. These are called the gradient, curl, divergence.

In this section, D denotes an open subset of $\mathbb{R}^3$.

9.2.1. Gradient in three dimensions. Let f be a scalar field on D. The *gradient vector field* of f is the vector field on D defined by

$$\operatorname{grad} f := \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) = (f_x, f_y, f_z). \tag{9.1}$$

In this case, we say $F = \operatorname{grad} f$ is a *conservative vector field* with *potential function* f.

Example 9.8. Let $D = \mathbb{R}^3$. Consider the vector field $F(x, y, z) = (yz, zx, xy)$ on D. It is the gradient vector field of $f(x, y, z) = xyz$, that is, $F = \operatorname{grad} f$.

Example 9.9 (Gravitational force field). Let $D = \mathbb{R}^3 \setminus \{(0, 0, 0)\}$, that is, $\mathbb{R}^3$ minus the origin. Consider the vector field on D defined by

$$F(x, y, z) = -\alpha \left(\frac{x}{(x^2 + y^2 + z^2)^{3/2}}, \frac{y}{(x^2 + y^2 + z^2)^{3/2}}, \frac{z}{(x^2 + y^2 + z^2)^{3/2}}\right).$$

In physics, this is called the *gravitational force field* (with $\alpha = mMG$). It is the gradient vector field of

$$f(x, y, z) = \frac{\alpha}{(x^2 + y^2 + z^2)^{1/2}}.$$

The scalar field f is the *potential*.

Definition (9.1) generalizes to open subsets of $\mathbb{R}^m$. In particular, for open subsets of $\mathbb{R}^2$, we define $\operatorname{grad} f = (f_x, f_y)$.

9.2.2. Curl. Let $F = (P, Q, R)$ be a vector field on D. The *curl vector field* of F is the vector field on D defined by

$$\operatorname{curl} F := \nabla \times F := \left(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}, \frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}, \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right). \tag{9.2}$$

The notation $\nabla \times F$ is to be interpreted as

$$\nabla \times F = \nabla \times (P, Q, R) = \det \begin{bmatrix} i & j & k \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ P & Q & R \end{bmatrix}.$$

Example 9.10. Let $F = \omega(-y, x, 0)$, where we interpret w as angular speed. Then $\operatorname{curl} F = (0, 0, 2\omega)$. This is a constant vector field pointing along the positive z-axis.

9.2.3. Divergence in three dimensions. Let $F = (P, Q, R)$ be a vector field on D. The *divergence field* of F is the scalar field on D defined by

$$\operatorname{div} F := \nabla \cdot F = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} + \frac{\partial R}{\partial z}. \tag{9.3}$$

For example,

$$\operatorname{div}(x, y, z) = 3 \quad \text{and} \quad \operatorname{div}(xy^2, z \sin y, xe^y) = y^2 + z \cos y.$$

Definition (9.3) generalizes to open subsets of $\mathbb{R}^m$. In particular, for open subsets of $\mathbb{R}^2$, we define $\operatorname{div}(P, Q) = P_x + Q_y$.

9.2.4. Gradient, curl, divergence. Gradient, curl, divergence can be assembled into a sequence of maps

$$\left\{\begin{matrix}\text{scalar}\\ \text{fields}\end{matrix}\right\} \xrightarrow{\text{grad}} \left\{\begin{matrix}\text{vector}\\ \text{fields}\end{matrix}\right\} \xrightarrow{\text{curl}} \left\{\begin{matrix}\text{vector}\\ \text{fields}\end{matrix}\right\} \xrightarrow{\text{div}} \left\{\begin{matrix}\text{scalar}\\ \text{fields}\end{matrix}\right\}. \tag{9.4}$$

Lemma 9.11. *We have*

$$\begin{aligned} \operatorname{curl}(\operatorname{grad} f) &= \nabla \times (\nabla f) = 0, \\ \operatorname{div}(\operatorname{curl} F) &= \nabla \cdot (\nabla \times F) = 0. \end{aligned} \tag{9.5}$$

Proof. We compute:

$$\operatorname{curl}(\operatorname{grad}(f)) = \det \begin{bmatrix} i & j & k \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ f_x & f_y & f_z \end{bmatrix} = 0.$$

Similarly,

$$\begin{aligned} \operatorname{div}(\operatorname{curl} F) &= (R_y - Q_z)_x + (P_z - R_x)_y + (Q_x - P_y)_z \\ &= R_{yx} - Q_{zx} + P_{zy} - R_{xy} + Q_{xz} - P_{yz} \\ &= 0. \end{aligned}$$

In either calculation, we used the mixed partials result in Proposition 7.19. □

9.2.5. When is a vector field a gradient vector field? Given a vector field F on D, we would like to know if F is the gradient vector field of some scalar field f on D?

Lemma 9.12 (Necessary condition for a gradient vector field). *Suppose $F = \operatorname{grad} f$, that is, $F = (P, Q, R)$ and $P = f_x$, $Q = f_y$, $R = f_z$. Then $\operatorname{curl} F = 0$, that is, $Q_x = P_y$, $R_y = Q_z$, $P_z = R_x$.*

Proof. We saw this in (9.5). □

Lemma 9.12 in two dimensions takes the following form. Suppose $F = \operatorname{grad} f$, that is, $F = (P, Q)$ and $P = f_x$, $Q = f_y$. Then by the mixed partials result, $Q_x = P_y$.

Example 9.13. Consider the inward radial vector field $F(x, y) = (-x, -y)$ in Example 9.4, item (3). It is the gradient vector field of $f(x, y) = -\frac{1}{2}(x^2 + y^2)$, that is, $F = \operatorname{grad} f$.

Example 9.14. Consider the rotational vector field $F(x,y) = (-y,x)$ in Example 9.4, item (4). This is not a gradient vector field since $F = (P,Q)$ and $Q_x = 1 = -P_y$ which violates the necessary condition above.

Now consider

$$F(x,y) = \Big(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\Big) \tag{9.6}$$

on $\mathbb{R}^2 \setminus \{(0,0)\}$. Now

$$F = (P,Q) \quad \text{and} \quad Q_x = -\frac{x^2-y^2}{(x^2+y^2)^2} = P_y.$$

So the necessary condition is satisfied, yet F is not a gradient vector field. We will see this later in Example 9.27.

We mention that the converse of Lemma 9.12 holds when the region D is simply connected. See Section 9.7.3 in this regard.

Remark 9.15 (Homological algebra). The sequence of maps in (9.4) is an example of a chain complex. It has the property that composite of any two consecutive maps in the sequence is zero. We saw this in (9.5). To any chain complex are associated its homology groups H_0, H_1, and so on, depending on the length of the sequence. In particular, (9.4) has four homology groups H_0, H_1, H_2, H_3. They depend on the space X whose scalar and vector fields we are considering. The question of the converse of Lemma 9.12 mentioned above is related to the first homology group H_1. For more details on this topic, see Bott and Tu [7]. For abstract homological algebra, see Weibel [31]. These ideas developed mainly in the first half of the twentieth century.

9.3. Line integrals and FTC

Now we look at line integrals of scalar fields and of vector fields along a parametrized curve. The length of a curve is the line integral of the scalar field which is identically 1. The line integral of a gradient vector field yields a generalization of FTC, Part II in one variable.

9.3.1. Parametrized curve. A *path* or a *parametrized curve* in $\mathbb{R}^m$ is a map

$$\gamma : [a,b] \to \mathbb{R}^m, \qquad t \mapsto (x_1(t), \ldots, x_m(t)). \tag{9.7}$$

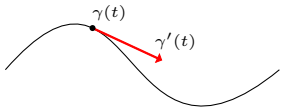
We say $\gamma(a)$ is the *initial point* and $\gamma(b)$ is the *final point* of the path γ. We say γ is *closed* if $\gamma(a) = \gamma(b)$.

Define

$$\gamma'(t) := \lim_{h\to 0} \frac{\gamma(t+h) - \gamma(t)}{h} = (x_1'(t), \ldots, x_m'(t)). \tag{9.8}$$

It is the tangent vector to γ at t.

The picture below shows a parametrized curve in $\mathbb{R}^2$, with the tangent vector marked at one point.



In physics terms, one may think of a path as a particle moving along a curve in $\mathbb{R}^m$ with t as the time variable. The particle is at the initial point at $t = a$, at the final point at $t = b$, and in general at point $\gamma(t)$ at time t. The vector $\gamma'(t)$ is then the velocity of the particle at time t.

Example 9.16 (Graph). Let $f : [a, b] \to \mathbb{R}$ be a function. Then the graph of f is a parametrized curve

$$\gamma : [a, b] \to \mathbb{R}^2, \qquad \gamma(t) = (t, f(t)), \tag{9.9}$$

and $\gamma'(t) = (1, f'(t))$.

We say γ is *regular* if $\gamma'(t) \neq 0$ for all $a \leq t \leq b$.

Example 9.17. Consider the graph of the function $y = x^{\frac{2}{3}}$, see Example 4.2, item (2). This curve can be parametrized by $\gamma(t) = (t^3, t^2)$ say for $t \in [-1, 1]$. Then $\gamma'(t) = (3t^3, 2t)$. Observe that $\gamma'(0) = (0, 0)$. Thus, γ is not regular at $t = 0$.

9.3.2. Length of a parametrized curve. For a parametrized curve γ, define the length of γ to be

$$\ell(\gamma) := \int_a^b \|\gamma'(t)\|\, dt = \int_a^b \sqrt{x_1'(t)^2 + \cdots + x_m'(t)^2}\, dt. \tag{9.10}$$

The special case when $m = 2$ was considered in (5.14).

Example 9.18 (Helix). Consider the curve $\gamma(t) = (\cos t, \sin t, t)$ for $a \leq t \leq b$. Then

$$\ell(\gamma) = \int_a^b \sqrt{(-\sin t)^2 + (\cos t)^2 + 1}\, dt = \sqrt{2}\,(b - a).$$

9.3.3. Line integral of a scalar field. Put $C = \gamma([a, b])$. Let $f : C \to \mathbb{R}$ be a scalar field. Define the *line integral* of f along a path γ by

$$\int_\gamma f\, |ds| := \int_a^b f(\gamma(t))\, \|\gamma'(t)\|\, dt. \tag{9.11}$$

If $f \equiv 1$, then formula (9.11) coincides with (9.10) and yields the length of γ.

In physics terms, one may think of f as the density function of a one-dimensional object such as a wire, and (9.11) then yields the mass of the wire. Alternatively, one may think of f as the charge density function, and (9.11) then yields the total charge present in the wire.

9.3.4. Differential notation. Put

$$|ds| = \|\gamma'(t)\|\,dt = \sqrt{dx_1^2 + \cdots + dx_m^2}.$$

Then

$$\int_\gamma f\,|ds| = \int_\gamma f\,\sqrt{dx_1^2 + \cdots + dx_m^2} = \int_a^b f(\gamma(t))\|\gamma'(t)\|\,dt.$$

9.3.5. Invariance under reparametrization.

Lemma 9.19. *Let*

$$[\alpha,\beta] \xrightarrow{h} [a,b] \xrightarrow{\gamma} \mathbb{R}^m,$$

with $h' \neq 0$ and h onto. Then

$$\int_{\gamma\circ h} f\,|ds| = \int_\gamma f\,|ds|. \tag{9.12}$$

Proof. We calculate:

$$\begin{aligned}\int_{\gamma\circ h} f\,|ds| &= \int_\alpha^\beta f((\gamma\circ h)(u))\,\|(\gamma\circ h)'(u)\|\,du\\ &= \int_\alpha^\beta f(\gamma(h(u)))\,\|\gamma'(h(u))\|\,|h'(u)|\,du\\ &= \int_a^b f(\gamma(t))\,\|\gamma'(t)\|\,dt\\ &= \int_\gamma f\,|ds|.\end{aligned}$$

We used substitution formula (5.6) in the third step. □

In particular, taking $f \equiv 1$, we see that the length of a curve does not depend on the chosen parametrization.

Exercise 9.20. Parametrize the unit circle S^1 by $(\cos t, \sin t)$ for $t \in [0, 2\pi]$, and also by $(\cos 2t, \sin 2t)$ for $t \in [0, \pi]$. Check directly that in both cases, we get $\ell(S^1)$ to be 2π.

9.3.6. Arc length parametrization. For $t \in [a, b]$, define

$$u(t) := \int_a^t \|\gamma'(t)\|\,dt. \tag{9.13}$$

This yields a bijective map from $[a, b]$ to $[0, \ell(\gamma)]$. Composing its inverse with γ yields

$$\tilde\gamma : [0, \ell(\gamma)] \to [a, b] \xrightarrow{\gamma} \mathbb{R}^m.$$

This parametrization of C is known as its *arc length parametrization*, and u is called the *arc length parameter*.

In physics terms, for a particle moving along C, the particle is at the initial point at $u = 0$, at the final point at $u = \ell(C)$, and in general at point $\tilde\gamma(u)$ at time u, where the length of the curve from the initial point $\tilde\gamma(0)$ to point $\tilde\gamma(u)$ is u.

Example 9.21. For the circle of radius a centered at the origin, an arc length parametrization is

$$\tilde{\gamma} : [0, 2\pi a] \to \mathbb{R}^2, \qquad u \mapsto (a\cos\big(\frac{u}{a}\big), a\sin\big(\frac{u}{a}\big)).$$

Starting with the usual parametrization $(a\cos t, a\sin t)$, formula (9.13) yields $u = at$ which leads to the above map.

9.3.7. Line integral of a vector field. Define the *line integral* of a vector field F on $\mathbb{R}^m$ along a path γ in $\mathbb{R}^m$ by

$$\int_\gamma F \cdot ds := \int_a^b F(\gamma(t)) \cdot \gamma'(t)\, dt. \tag{9.14}$$

The vector field F is defined everywhere, but for the above definition, only its values on points of the curve matter. The dot product says that we take the component of $F(\gamma(t))$ along the tangential component $\gamma'(t)$.

In physics terms, one may think of (9.14) as the work done by the force field F along the curve γ.

More generally, we can consider the situation where F is a vector field on an open subset D of $\mathbb{R}^m$, and γ is a path which lies in D, that is, $\gamma(t) \in D$ for all $t \in [a, b]$.

Example 9.22. Consider the radial vector field $F(x, y, z) = (x, y, z)$ on $\mathbb{R}^3$ and the path

$$\gamma : [0, 2\pi] \to \mathbb{R}^3, \qquad \gamma(t) = (\cos t, \sin t, t).$$

The latter is a parametrized helix. Then

$$\begin{aligned}\int_\gamma F \cdot ds &= \int_0^{2\pi} (\cos t, \sin t, t) \cdot (-\sin t, \cos t, 1)\, dt \\ &= \int_0^{2\pi} t\, dt \\ &= 2\pi^2.\end{aligned}$$

9.3.8. Differential notation. Let $F = (P, Q, R)$ and $\gamma(t) = (x(t), y(t), z(t))$. Put

$$ds = \gamma'(t)\, dt = (x'(t)\, dt, y'(t)\, dt, z'(t)\, dt) = (dx, dy, dz).$$

Then

$$F \cdot ds = Px'(t)\, dt + Qy'(t)\, dt + Rz'(t)\, dt = P\, dx + Q\, dy + R\, dz,$$

and

$$\int_\gamma F \cdot ds = \int_\gamma P\, dx + Q\, dy + R\, dz = \int_a^b Px'(t)\, dt + Qy'(t)\, dt + Rz'(t)\, dt.$$

Example 9.23. Consider the vector field $F(x, y, z) = (\cos z, e^x, e^y)$ on $\mathbb{R}^3$ and the path

$$\gamma : [0, 2] \to \mathbb{R}^3, \qquad \gamma(t) = (1, t, e^t).$$

Then

$$\begin{aligned}\int_\gamma F\cdot ds &= \int_\gamma \cos z\,dx + e^x\,dy + e^y\,dz\\ &= \int_0^2 [(\cos e^t)(0) + (e^1)(1) + (e^t)(e^t)]dt\\ &= 2e + \frac{1}{2}(e^4-1).\end{aligned}$$

9.3.9. Relating *ds* and *|ds|*. Suppose $\gamma'(t)\neq 0$. That is, the tangent vector to the curve at t under the parametrization γ is nonzero. Put

$$T(\gamma(t)) := \frac{\gamma'(t)}{\|\gamma'(t)\|}, \tag{9.15}$$

the *unit tangent vector* to the curve γ at t. Then

$$\int_\gamma F\cdot ds = \int_\gamma F\cdot T\,|ds|, \tag{9.16}$$

with lhs defined by (9.14) and rhs by (9.11).

Proof. This is a consequences of our definitions:

$$\begin{aligned}\int_\gamma F\cdot ds &= \int_a^b F(\gamma(t))\cdot\gamma'(t)\,dt\\ &= \int_a^b F(\gamma(t))\cdot T(\gamma(t))\,\|\gamma'(t)\|\,dt\\ &= \int_\gamma F\cdot T\,|ds|.\end{aligned}$$

In the second step, we used definition (9.15). □

9.3.10. Line integral of a gradient vector field. The following is a generalization of Theorem 5.18 which was FTC, Part II. We may refer to it as FTC on a parametrized curve; the earlier FTC is on a line segment (which is the simplest example of a parametrized curve with $\gamma:[a,b]\to\mathbb{R}$ defined by $\gamma(x)=x$).

Theorem 9.24. *Let $F=\operatorname{grad} f$ be a gradient vector field on an open subset D of $\mathbb{R}^m$. Let $\gamma:[a,b]\to\mathbb{R}^m$ be a path which lies in D. Then*

$$\int_\gamma F\cdot ds = f(\gamma(b)) - f(\gamma(a)). \tag{9.17}$$

PROOF. For simplicity, let us take $m = 3$. Also, put $\gamma(t) = (x(t), y(t), z(t))$. Then

$$\begin{aligned}\int_\gamma F \cdot ds &= \int_a^b \nabla f(\gamma(t)) \cdot \gamma'(t)\, dt \\ &= \int_a^b [f_x(\gamma(t))x'(t) + f_y(\gamma(t))y'(t) + f_z(\gamma(t))z'(t)]\, dt \\ &= \int_a^b (f \circ \gamma)'(t)\, dt \\ &= f(\gamma(b)) - f(\gamma(a)).\end{aligned}$$

The third equality is by the chain rule. In matrix form,

$$(f \circ \gamma)'(t) = \begin{bmatrix} f_x(\gamma(t)) & f_y(\gamma(t)) & f_z(\gamma(t)) \end{bmatrix} \begin{bmatrix} x'(t) \\ y'(t) \\ z'(t) \end{bmatrix}.$$

The last equality is by Theorem 5.18 (which was FTC, Part II on an interval $[a, b]$). □

Example 9.25. Recall from Example 9.8, the vector field on $\mathbb{R}^3$ given by $F(x, y, z) = (yz, zx, xy)$. It is the gradient vector field of $f(x, y, z) = xyz$. Let

$$\gamma : [0, \pi/4] \to \mathbb{R}^3, \qquad \gamma(t) = (\cos^4 t, \sin^4 t, \tan^4 t).$$

By formula (9.17),

$$\int_\gamma F \cdot ds = f(\gamma(\pi/4)) - f(\gamma(0)) = f\Big(\frac{1}{4}, \frac{1}{4}, 1\Big) - f(1, 0, 0) = \frac{1}{16}.$$

9.3.11. Path independence of line integrals. Let F be a gradient vector field on an open subset D of $\mathbb{R}^m$. We say line integrals of F are *path independent* in D if

$$\int_{\gamma_1} F \cdot ds = \int_{\gamma_2} F \cdot ds$$

for any paths γ_1 and γ_2 lying in D with the same initial point and same final point. The above condition is equivalent to requiring that

$$\int_\gamma F \cdot ds = 0$$

for any closed path γ lying in D.

Proposition 9.26. *Let F be a gradient vector field on an open subset D of $\mathbb{R}^m$. Then line integrals of F are path independent in D. In particular, the line integral of F along any closed path γ is zero.*

PROOF. This follows from formula (9.17). □

Example 9.27. Consider the vector field

$$F(x, y) = \Big(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2}\Big)$$

on $\mathbb{R}^2 \setminus \{(0, 0)\}$ from (9.6), and the closed path

$$\gamma : [0, 2\pi] \to \mathbb{R}^2, \qquad \gamma(t) = (\cos t, \sin t).$$

Then

$$\int_\gamma F \cdot ds = \int_0^{2\pi} (-\sin t, \cos t) \cdot (-\sin t, \cos t)\, dt = \int_0^{2\pi} dt = 2\pi.$$

Hence F is not a gradient vector field (since its line integral along the above closed path γ is not zero).

The converse of Proposition 9.26 is also true.

Proposition 9.28. *Let F be a vector field on an open path connected subset D of $\mathbb{R}^m$. Suppose the line integrals of F are path independent in D. Then F is a gradient vector field on D.*

Proof. We illustrate with $m = 3$. Let $F = (P, Q, R)$. Fix a point $(a_0, b_0, c_0) \in D$. Put

$$\int_{(a_0,b_0,c_0)}^{(x,y,z)} F \cdot ds := \int_\gamma F \cdot ds,$$

where γ is any path in D from (a_0, b_0, c_0) to (x, y, z). This is well-defined by the hypothesis that line integrals of F only depend on endpoints of γ. Define

$$f(x, y, z) := \int_{(a_0,b_0,c_0)}^{(x,y,z)} F \cdot ds.$$

We claim that f is the required potential function. That is, $\nabla f = F$, or equivalently, $f_x = P$, $f_y = Q$, $f_z = R$. Let us first deal with the claim $f_x = P$. We calculate:

$$\begin{aligned} f_x(a, b, c) &= \lim_{h\to 0} \frac{f(a + h, b, c) - f(a, b, c)}{h} \\ &= \lim_{h\to 0} \frac{\int_{(a,b,c)}^{(a+h,b,c)} F \cdot ds}{h} \\ &= \lim_{h\to 0} \frac{\int_0^h P(a + x, b, c)\, dx}{h} \\ &= P(a, b, c). \end{aligned}$$

In the third step, we took the straight line path in the x-direction, namely,

$$\gamma : [0, h] \to \mathbb{R}^3, \qquad \gamma(x) = (a + x, b, c),$$

and so $\gamma'(x) = (1, 0, 0)$. The last step can be deduced from Theorem 5.15, which is FTC, Part I. This shows $f_x = P$. Analogously, we can show $f_y = Q$, $f_z = R$. □

9.3.12. Invariance under reparametrization up to sign.

Lemma 9.29. *Let*

$$[\alpha, \beta] \xrightarrow{h} [a, b] \xrightarrow{\gamma} \mathbb{R}^m,$$

with $h' \neq 0$ and h onto. Then

$$\int_{\gamma\circ h} F \cdot ds = \pm \int_\gamma F \cdot ds. \tag{9.18}$$

PROOF. There are two cases.

- $h' > 0$ on $[\alpha, \beta]$.

$$\begin{aligned}\int_{\gamma\circ h} F \cdot ds &= \int_\alpha^\beta F((\gamma\circ h)(u)) \cdot (\gamma\circ h)'(u)\, du \\ &= \int_\alpha^\beta F(\gamma(h(u))) \cdot \gamma'(h(u))h'(u)\, du \\ &= \int_{h(\alpha)=a}^{h(\beta)=b} F(\gamma(t)) \cdot \gamma'(t)\, dt \\ &= \int_\gamma F \cdot ds.\end{aligned}$$

- $h' < 0$ on $[\alpha, \beta]$. We repeat the above calculation. We now get a minus sign since $h(\alpha) = b$ and $h(\beta) = a$.

□

Special case. Consider

$$[-b, -a] \xrightarrow{h} [a, b] \xrightarrow{\gamma} D,$$

with $h(u) = -u$. We denote $\gamma \circ h$ by $-\gamma$, and call it the negative of γ. Since $h' = -1 < 0$,

$$\int_{-\gamma} F \cdot ds = -\int_\gamma F \cdot ds. \tag{9.19}$$

9.3.13. Geometric curve. A *geometric curve* is the image of a parametrized curve. An *orientation* of a geometric curve C is an assignment of a unit tangent vector, that is, a specification of a direction at each point of C in a continuous manner.

This can be done in several ways.

(1) If C is not closed, then specify the 'start' and 'finish' points.
(2) If C is closed and is the boundary of a planar region S, then specify whether S lies to the left (or to the right) as C is traversed (with head in upright position).
(3) If $\gamma : [a, b] \to \mathbb{R}^2$ is a 'nice' parametrized curve, then for each t, $\frac{\gamma'(t)}{\|\gamma'(t)\|}$ is a unit tangent vector to C at $\gamma(t)$, and we obtain an orientation of C.

An *oriented geometric curve* is a geometric curve with a specified orientation. If C is the image of a parametrized curve γ, then we say γ is orientation preserving if $\frac{\gamma'(t)}{\|\gamma'(t)\|}$ equals the prescribed unit tangent vector. In this case, we define

$$\int_C F \cdot ds = \int_\gamma F \cdot ds.$$

If this is not the case, that is, γ is orientation reversing, then $-\gamma$ is orientation preserving, and we define

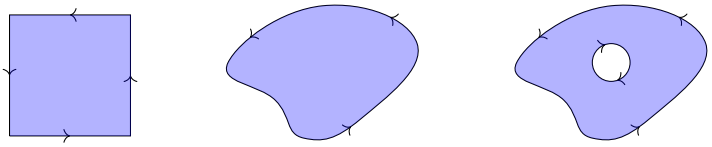
$$\int_C F \cdot ds = \int_{-\gamma} F \cdot ds = -\int_\gamma F \cdot ds.$$

9.4. Green's theorem

We now generalize FTC, Part II to two dimensions, where the two-dimensional region is a subset of $\mathbb{R}^2$.

All curves in this section are in $\mathbb{R}^2$.

9.4.1. Orienting the boundary curve. Let D be a bounded subset of $\mathbb{R}^2$ such that ∂D consists of a finite number of simple closed nonintersecting piecewise smooth geometric curves. Orient each such curve so that as one travels along that curve, the set D lies to the left (with head in upright position). Then we say ∂D is *positively oriented*. See illustrations below.



When viewed from above, the outer boundary of D is traversed anticlockwise, and each of the inner boundary curves is traversed clockwise.

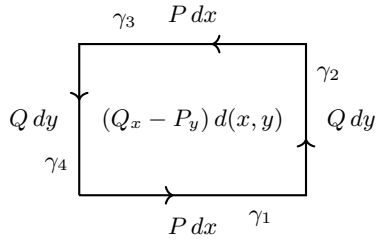
9.4.2. Green's theorem. This result relates an integral on a two-dimensional region D to a line integral on its boundary ∂D.

Theorem 9.30. *Let $D \subseteq \mathbb{R}^2$ be such that ∂D is positively oriented. If P and Q are scalar fields defined on an open set containing D, then*

$$\int_{\partial D} P\,dx + Q\,dy = \iint_D (Q_x - P_y)\,d(x,y). \tag{9.20}$$

Proof. We break the proof in three steps.

(i) D is a rectangle. Let $D = [a,b] \times [c,d]$. Let γ_1, γ_2, γ_3, γ_4 be oriented line segments corresponding to the four sides of the rectangle, as shown below.



Using formula (5.4) of FTC, Part II,

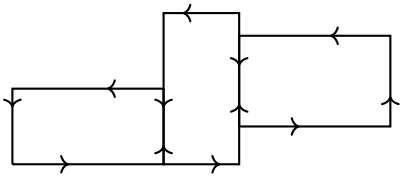
$$\begin{aligned}\iint_D Q_x\,d(x,y) &= \int_c^d \int_a^b Q_x\,d(x,y) \\ &= \int_c^d [Q(b,y) - Q(a,y)]\,dy \\ &= \int_{\gamma_2} Q\,dy + \int_{\gamma_4} Q\,dy.\end{aligned}$$

Similarly,

$$\begin{aligned}\iint_D -P_y\, d(x,y) &= \int_a^b \int_c^d -P_y\, d(x,y) \\ &= \int_a^b [-P(x,d) + P(x,c)]\, dx \\ &= \int_{\gamma_3} P\, dx + \int_{\gamma_1} P\, dx.\end{aligned}$$

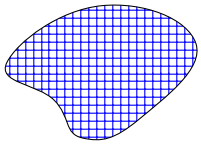
Adding the two proves formula (9.20) when D is a rectangle.

(ii) D is a "union" of rectangles. Now suppose D is as shown below.



Overlapping line-segments have opposite orientations, so line integrals over them cancel by (9.19). Thus, formula (9.20) holds when D is formed by adjoining rectangles.

(iii) D is a general region. Draw a grid inside D as shown below.



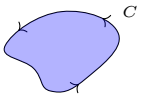
Let D' be the largest region which fits inside D which is of the form in item (ii) above and is made out of the sides of the grid. By item (ii), formula (9.20) holds for D'. Finally, we approach formula (9.20) for D by considering finer and finer grids.

□

9.4.3. Principle of deformation.

Corollary 9.31. *If C is a simple closed geometric curve and $Q_x = P_y$ inside C, then*

$$\int_C P\, dx + Q\, dy = 0. \tag{9.21}$$



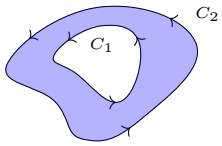
Proof. By (9.20),

$$\int_{C=\partial D} P\, dx + Q\, dy = \iint_D (Q_x - P_y)\, d(x,y) = 0.$$

□

Corollary 9.32. *Suppose C_1 and C_2 are simple closed geometric curves with C_1 lying inside C_2 and with both oriented anticlockwise. If $Q_x = P_y$ in the region D between C_1 and C_2, then*

$$\int_{C_1} P\,dx + Q\,dy = \int_{C_2} P\,dx + Q\,dy. \tag{9.22}$$



PROOF. Let ∂D be the positively oriented boundary of D. By (9.20),

$$\int_{\partial D} P\,dx + Q\,dy = \iint_D (Q_x - P_y)\,d(x,y) = 0.$$

Therefore,

$$\int_{C_2} P\,dx + Q\,dy + \int_{-C_1} P\,dx + Q\,dy = 0.$$

Using (9.19), the required formula follows. □

Note: In (9.22), P and Q must be well-behaved in the region between C_1 and C_2. They may have singularities at other points.

Example 9.33. Let

$$P(x,y) = \frac{-y}{x^2+y^2} \quad \text{and} \quad Q(x,y) = \frac{x}{x^2+y^2}$$

on $\mathbb{R}^2 \setminus \{(0,0)\}$. Then

$$Q_x = -\frac{x^2-y^2}{(x^2+y^2)^2} = P_y$$

as noted in Example 9.14.

Let C be any simple closed geometric curve oriented anticlockwise. Let us compute $\int_C P\,dx + Q\,dy$. There are two cases.

(i) $(0,0)$ lies outside C. Then $\int_C P\,dx + Q\,dy = 0$ by formula (9.21) in Corollary 9.31.

(ii) $(0,0)$ lies inside C. Then $\int_C P\,dx + Q\,dy = 2\pi$. Why? Let γ_ϵ be the circle of radius ϵ and center $(0,0)$ lying inside C. Then by formula (9.22) in Corollary 9.32,

$$\int_C P\,dx + Q\,dy = \int_{\gamma_\epsilon} P\,dx + Q\,dy = \int_{\gamma_1} P\,dx + Q\,dy = 2\pi.$$

The second equality is by the same reasoning. The last equality was calculated in Example 9.27.

Thus, the calculation of a line integral over a complicated curve C is simplified.

9.4.4. Area calculation.

Corollary 9.34. *Let C be a simple closed geometric curve oriented anticlockwise which encloses a region D. Then*

$$\text{Area}(D) = \frac{1}{2}\int_C x\,dy - y\,dx = \int_C x\,dy = -\int_C y\,dx. \tag{9.23}$$

Proof. Put $P(x,y) = \frac{-y}{2}$ and $Q(x,y) = \frac{x}{2}$. Then $Q_x - P_y = \frac{1}{2} + \frac{1}{2} = 1$. By formula (8.5),

$$\begin{aligned}\text{Area}(D) &= \iint_D d(x,y)\\ &= \iint_D (Q_x - P_y)\,d(x,y)\\ &= \int_{\partial D = C} P\,dx + Q\,dy\\ &= \frac{1}{2}\int_C x\,dy - y\,dx.\end{aligned}$$

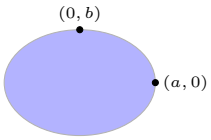
The third equality is by (9.20). □

Note very carefully: If in (9.23) we replace $x\,dy - y\,dx$ by $x\,dy + y\,dx$, then we get a line integral of a gradient vector field over a closed curve which is zero by formula (9.17).

Corollary 9.35. *For a simple closed curve C parametrized by $(x(t), y(t))$ for $a \le t \le b$ which is traversed anticlockwise as t goes from a to b,*

$$\begin{aligned}\text{Area}(D) &= \frac{1}{2}\int_a^b x(t)y'(t) - y(t)x'(t)\,dt\\ &= \frac{1}{2}\int_a^b \det\begin{bmatrix} x(t) & y(t)\\ x'(t) & y'(t)\end{bmatrix} dt.\end{aligned} \tag{9.24}$$

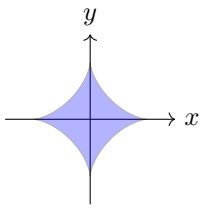
Example 9.36 (Ellipse). Let C be the ellipse defined by $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. See picture below.



Parametrize C by $x = a\cos t$ and $y = b\sin t$ for $0 \le t \le 2\pi$. Then, by formula (9.24), area of the elliptical region enclosed by C is

$$\frac{1}{2}\int_0^{2\pi} (a\cos t)(b\cos t) - (b\sin t)(-a\sin t)\,dt = \pi ab.$$

Example 9.37 (Hypocycloid). Let C be the hypocycloid given by $x = a\cos^3 t$ and $y = a\sin^3 t$ for $0 \le t \le 2\pi$. See picture below.



Then

$$\det \begin{bmatrix} x(t) & y(t) \\ x'(t) & y'(t) \end{bmatrix} = 3a^2 \cos^2 t \sin^2 t.$$

Thus, by formula (9.24), area enclosed by C is

$$\begin{aligned} \frac{1}{2}\int_0^{2\pi} 3a^2\cos^2 t\sin^2 t\,dt &= \frac{3a^2}{2}\int_0^{2\pi}\left(\frac{\sin 2t}{2}\right)^2 dt \\ &= \frac{3a^2}{8}\int_0^{2\pi}\frac{1-\cos 4t}{2}\,dt \\ &= \frac{3\pi a^2}{8}. \end{aligned}$$

Special case. Let the parameter be θ, and the curve C be given by $x(\theta) = p(\theta)\cos\theta$ and $y(\theta) = p(\theta)\sin\theta$ for $0 \le \theta \le 2\pi$. In this case, C is traversed anticlockwise. Now

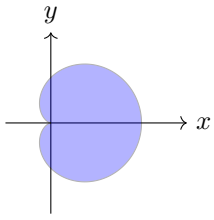
$$\begin{aligned} \det \begin{bmatrix} x(\theta) & y(\theta) \\ x'(\theta) & y'(\theta) \end{bmatrix} &= \det \begin{bmatrix} p(\theta)\cos\theta & p(\theta)\sin\theta \\ p'(\theta)\cos\theta - p(\theta)\sin\theta & p'(\theta)\sin\theta + p(\theta)\cos\theta \end{bmatrix} \\ &= p(\theta)^2. \end{aligned}$$

Thus, by formula (9.24), area of the region D enclosed by the curve C is

$$\text{Area}(D) = \frac{1}{2}\int_0^{2\pi} p(\theta)^2\,d\theta. \tag{9.25}$$

This is a special case of (5.9).

Example 9.38 (Cycloid). Consider the curve C defined by $r = p(\theta) = a(1+\cos\theta)$ for $0 \le \theta \le 2\pi$. See picture below. We had looked at half of this curve in Example 5.29.



By formula (9.25), area enclosed by C is

$$\frac{1}{2}\int_0^{2\pi} a^2(1+\cos\theta)^2\,d\theta = \frac{3\pi a^2}{2}.$$

9.5. Surface integrals

Now we look at surface integrals of scalar fields and of vector fields across a parametrized surface. The area of a surface is the surface integral of the scalar field which is identically 1.

We proceed in analogy with Section 9.3 on line integrals. A comparison is shown in Table 9.2.

TABLE 9.2. Line integrals vs surface integrals.

parametrized curve $\gamma(t)$	(9.7)	parametrized surface $\Phi(u,v)$	(9.26)
$\|\gamma'(t)\|$		$\|(\Phi_u \times \Phi_v)(u,v)\|$	
$\ell(\gamma)$	(9.10)	$\mathrm{Area}(\Phi)$	(9.32)
$\int_\gamma f\,\|ds\|$	(9.11)	$\iint_\Phi f\,\|dS\|$	(9.33)
reparametrization	(9.12)	reparametrization	(9.35)
$\gamma'(t)$	(9.8)	$\Phi_u \times \Phi_v$	(9.28)
$\int_\gamma F \cdot ds$	(9.14)	$\iint_\Phi F \cdot dS$	(9.36)
$(P,Q,R)\cdot ds = P\,dx + Q\,dy + R\,dz$		$(P,Q,R)\cdot dS = P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy$	
reparametrization	(9.18)	reparametrization	(9.37)
oriented geometric curve		oriented geometric surface	
ds and $\|ds\|$	(9.16)	dS and $\|dS\|$	(9.40)

9.5.1. Parametrized surface. A *parametrized surface* in $\mathbb{R}^3$ is a map

$$\Phi : D \to \mathbb{R}^3, \qquad (u,v) \mapsto (x(u,v), y(u,v), z(u,v)), \tag{9.26}$$

where $D \subseteq \mathbb{R}^2$. Define

$$\Phi_u := (x_u, y_u, z_u) \quad \text{and} \quad \Phi_v := (x_v, y_v, z_v), \tag{9.27}$$

where x_u is the partial derivative of x wrt u, and so on.

9.5.2. Fundamental vector product. The *fundamental vector product* of the parametrization Φ is defined by

$$\begin{aligned}(\Phi_u \times \Phi_v)(u,v) &= \det \begin{bmatrix} i & j & k \\ x_u & y_u & z_u \\ x_v & y_v & z_v \end{bmatrix} \\ &= \left(\det \begin{bmatrix} y_u & z_u \\ y_v & z_v \end{bmatrix}, \det \begin{bmatrix} z_u & x_u \\ z_v & x_v \end{bmatrix}, \det \begin{bmatrix} x_u & y_u \\ x_v & y_v \end{bmatrix} \right) \\ &=: \left(\frac{\partial(y,z)}{\partial(u,v)}, \frac{\partial(z,x)}{\partial(u,v)}, \frac{\partial(x,y)}{\partial(u,v)} \right).\end{aligned} \tag{9.28}$$

This is a function of (u,v).

Example 9.39 (Cylinder). Let $D = [0, 2\pi] \times [0,h]$. Fix $a > 0$. Define

$$\Phi : D \to \mathbb{R}^3, \qquad \Phi(\theta, v) := (a\cos\theta,\ a\sin\theta,\ v). \tag{9.29}$$

Contrast with cylindrical coordinates (8.19) on $\mathbb{R}^3$, or with the helix.

Now

$$\Phi_\theta = (-a\sin\theta,\, a\cos\theta,\, 0) \quad \text{and} \quad \Phi_v = (0,0,1),$$

and

$$\begin{aligned}(\Phi_\theta \times \Phi_v)(\theta, v) &= \det \begin{bmatrix} i & j & k \\ -a\sin\theta & a\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= (a\cos\theta, a\sin\theta, 0).\end{aligned}$$

Example 9.40 (Sphere). Let $D = [0,\pi] \times [0, 2\pi]$. Fix $a > 0$. Define

$$\Phi : D \to \mathbb{R}^3, \qquad \Phi(\varphi,\theta) := (a\sin\varphi\cos\theta,\, a\sin\varphi\sin\theta,\, a\cos\varphi). \tag{9.30}$$

Contrast with spherical coordinates (8.21) on $\mathbb{R}^3$.

Now

$$\begin{aligned}\Phi_\varphi &= (a\cos\varphi\cos\theta, a\cos\varphi\sin\theta, -a\sin\varphi), \\ \Phi_\theta &= (-a\sin\varphi\sin\theta, a\sin\varphi\cos\theta, 0)\end{aligned}$$

and

$$\begin{aligned}(\Phi_\varphi \times \Phi_\theta)(\varphi,\theta) &= \det \begin{bmatrix} i & j & k \\ a\cos\varphi\cos\theta & a\cos\varphi\sin\theta & -a\sin\varphi \\ -a\sin\varphi\sin\theta & a\sin\varphi\cos\theta & 0 \end{bmatrix} \\ &= a\sin\varphi\,\Phi(\varphi,\theta).\end{aligned}$$

Example 9.41 (Graph). Let $f : D \to \mathbb{R}$ be a function of two variables. Then the graph of f is a parametrized surface

$$\Phi : D \to \mathbb{R}^3, \qquad \Phi(u,v) = (u, v, f(u,v)). \tag{9.31}$$

Now

$$\Phi_u = (1, 0, f_u) \quad \text{and} \quad \Phi_v = (0, 1, f_v),$$

and

$$\begin{aligned}(\Phi_u \times \Phi_v)(u,v) &= \det \begin{bmatrix} i & j & k \\ 1 & 0 & f_u \\ 0 & 1 & f_v \end{bmatrix} \\ &= (-f_u, -f_v, 1).\end{aligned}$$

9.5.3. Area of a parametrized surface. Recall: For vectors $u, v \in \mathbb{R}^3$, area of the parallelogram spanned by u and v is given by $\|u \times v\|$.

Let $\Phi : D \to \mathbb{R}^3$ be a parametrized surface. Define

$$\operatorname{Area}(\Phi) := \iint_D \|(\Phi_u \times \Phi_v)(u,v)\|\, d(u,v), \tag{9.32}$$

where $\Phi_u \times \Phi_v$ is the fundamental vector product in (9.28).

Example 9.42 (Cylinder). Consider the cylinder parametrized by (9.29). By formula (9.32),

$$\begin{aligned}\operatorname{Area}(\Phi) &= \iint\limits_{[0,2\pi]\times[0,h]} \|(a\cos\theta, a\sin\theta, 0)\|\, d(\theta, v)\\ &= \iint\limits_{[0,2\pi]\times[0,h]} a\, d(\theta, v)\\ &= 2\pi a h.\end{aligned}$$

This is the familiar formula for the surface area of cylinder of radius a and height h.

Example 9.43 (Sphere). Consider the sphere parametrized by (9.30). By formula (9.32),

$$\begin{aligned}\operatorname{Area}(\Phi) &= \iint\limits_{[0,\pi]\times[0,2\pi]} \|a\sin\varphi\, \Phi(\varphi,\theta)\|\, d(\varphi,\theta)\\ &= 2\pi a^2 \int_0^\pi \sin\varphi\, d\varphi\\ &= 4\pi a^2.\end{aligned}$$

This is the familiar formula for the surface area of sphere of radius a.

Example 9.44 (Graph). Consider the surface parametrized by (9.31). By formula (9.32),

$$\begin{aligned}\operatorname{Area}(\Phi) &= \iint_D \|(-f_u, -f_v, 1)\|\, d(u,v)\\ &= \iint_D \sqrt{1+f_u^2+f_v^2}\, d(u,v).\end{aligned}$$

As an example, consider the paraboloid parametrized by $f(u,v) = u^2+v^2$ for $(u,v)\in D$, where D is the disc of radius a around $(0,0)$. Then

$$\begin{aligned}\operatorname{Area}(\Phi) &= \iint_D \sqrt{1+4u^2+4v^2}\, d(u,v)\\ &= \int_0^{2\pi} d\theta \int_0^a \sqrt{1+4r^2}\, r\, dr\\ &= \frac{\pi}{6}[(1+4a^2)^{3/2}-1].\end{aligned}$$

9.5.4. Surface integral of a scalar field. Put $S = \Phi(D)$. Let $f : S \to \mathbb{R}$ be a scalar field. Define the *surface integral* of f across Φ by

$$\iint_\Phi f\, |dS| := \iint_D f(\Phi(u,v))\, \|(\Phi_u \times \Phi_v)(u,v)\|\, d(u,v), \tag{9.33}$$

where $\Phi_u \times \Phi_v$ is the fundamental vector product in (9.28).

If $f \equiv 1$, then formula (9.33) coincides with (9.32) and yields the area of the surface parametrized by Φ.

Example 9.45. We do a minor variation on Example 9.43. Consider the hemisphere parametrized by (9.30) on $D = [0, \frac{\pi}{2}] \times [0, 2\pi]$. Let $f(x, y, z) = z$. Then

$$\begin{aligned}\iint_\Phi f\,|dS| &= \iint_{[0,\frac{\pi}{2}]\times[0,2\pi]} a\cos\varphi\,\|a\sin\varphi\,\Phi(\varphi,\theta)\|\,d(\varphi,\theta)\\ &= 2\pi a^3\int_0^{\frac{\pi}{2}}\cos\varphi\sin\varphi\,d\varphi\\ &= \pi a^3.\end{aligned}$$

Example 9.46. Let $f(x, y, z) = xy + z$. Let S be the portion of the cylinder $y^2 + z^2 = 9$ in the first octant with $0 \le y \le 2$ and $0 \le x \le 4$. We want to evaluate (9.33).

Observe: S is the graph of the function $z = g(x, y) = \sqrt{9 - y^2}$ with $0 \le y \le 2$ and $0 \le x \le 4$. Thus,

$$\begin{aligned}\iint_\Phi f\,|dS| &= \int_0^2\int_0^4 (xy + g(x,y))\sqrt{1 + g_x^2 + g_y^2}\,dxdy\\ &= \int_0^2\int_0^4 (xy + \sqrt{9-y^2})\frac{3}{\sqrt{9-y^2}}\,dxdy\\ &= \int_0^2\int_0^4 \frac{3xy}{\sqrt{9-y^2}}\,dxdy + \int_0^2\int_0^4 3\,dxdy\\ &= 24(3 - \sqrt{5}) + 24\\ &= 24(4 - \sqrt{5}).\end{aligned}$$

Alternatively, we can parametrize S using cylindrical coordinates:

$$\Phi(x, \theta) = (x, 3\cos\theta, 3\sin\theta)$$

with $0 \le x \le 4$ and $\tan^{-1}(\frac{\sqrt{5}}{2}) \le \theta \le \frac{\pi}{2}$. Thus,

$$\begin{aligned}\iint_\Phi f\,|dS| &= \int_{\tan^{-1}(\frac{\sqrt{5}}{2})}^{\frac{\pi}{2}}\int_0^4 (3x\cos\theta + 3\sin\theta)3\,dxd\theta\\ &= 9\int_{\tan^{-1}(\frac{\sqrt{5}}{2})}^{\frac{\pi}{2}}\int_0^4 (x\cos\theta + \sin\theta)\,dxd\theta\\ &= 9\int_{\tan^{-1}(\frac{\sqrt{5}}{2})}^{\frac{\pi}{2}} (8\cos\theta + 4\sin\theta)\,d\theta\\ &= 24(4 - \sqrt{5}).\end{aligned}$$

Example 9.47 (Helicoid). Let $D = [0, 1] \times [0, 2\pi]$. Define

$$\Phi : D \to \mathbb{R}^3, \qquad \Phi(u, v) = (u\cos v, u\sin v, v). \tag{9.34}$$

This looks like a staircase curling around a vertical pillar. Then

$$\begin{aligned}(\Phi_u \times \Phi_v)(u, v) &= \det\begin{bmatrix} i & j & k\\ \cos v & \sin v & 0\\ -u\sin v & u\cos v & 1\end{bmatrix}\\ &= (\sin v, -\cos v, u).\end{aligned}$$

Let $f(x,y,z) = \sqrt{1+x^2+y^2}$. Thus,

$$\begin{aligned}\iint_\Phi f\,|dS| &= \iint_{[0,1]\times[0,2\pi]} \sqrt{1+u^2\cos^2 v + u^2\sin^2 v}\sqrt{\sin^2 v+\cos^2 v+u^2}\,d(u,v)\\ &= 2\pi\int_0^1 1+u^2\,du\\ &= \frac{8\pi}{3}.\end{aligned}$$

9.5.5. Invariance under reparametrization. Recall from (8.9) the jacobian of a map.

Lemma 9.48. *Let*

$$E \xrightarrow{h} D \xrightarrow{\Phi} \mathbb{R}^3,$$

with $J(h) \neq 0$. *Then*

$$\iint_{\Phi\circ h} f\,|dS| = \iint_\Phi f\,|dS|. \tag{9.35}$$

9.5.6. Surface integral of a vector field. Define the *surface integral* of a vector field F on $\mathbb{R}^3$ across a parametrized surface Φ in $\mathbb{R}^3$ by

$$\iint_\Phi F\cdot dS := \iint_D F(\Phi(u,v))\cdot(\Phi_u\times\Phi_v)(u,v)\,d(u,v), \tag{9.36}$$

where $\Phi_u\times\Phi_v$ is the fundamental vector product in (9.28).

Example 9.49 (Cylinder). Consider the radial vector field $F(x,y,z) = (x,y,z)$ on $\mathbb{R}^3$ and the cylinder parametrized by (9.29). Then

$$\begin{aligned}\iint_\Phi F\cdot dS &= \iint_{[0,2\pi]\times[0,h]} (a\cos\theta, a\sin\theta, v)\cdot(a\cos\theta, a\sin\theta, 0)\,d(\theta,v)\\ &= a^2\int_0^{2\pi} d\theta\int_0^h dv\\ &= 2\pi a^2 h.\end{aligned}$$

Example 9.50 (Sphere). Consider the radial vector field $F(x,y,z) = (x,y,z)$ on $\mathbb{R}^3$ and the sphere parametrized by (9.30). Then

$$\begin{aligned}\iint_\Phi F\cdot dS &= \iint_{[0,\pi]\times[0,2\pi]} \Phi(\varphi,\theta)\cdot\Phi(\varphi,\theta)\,a\sin\varphi\,d(\varphi,\theta)\\ &= \int_0^\pi\int_0^{2\pi} a^3\sin\varphi\,d(\varphi,\theta)\\ &= 4\pi a^3.\end{aligned}$$

Example 9.51 (Graph). Consider the surface parametrized by (9.31). Let $F = (P,Q,R)$. Then

$$\begin{aligned}\iint_\Phi F\cdot dS &= \iint_D (P,Q,R)\cdot(-f_u,-f_v,1)\,d(u,v)\\ &= \iint_D (-P\,f_u - Q\,f_v + R)\,d(u,v).\end{aligned}$$

As an example, consider the paraboloid parametrized by $f(u,v) = u^2+v^2$ for $(u,v) \in D$, where $D = [0,1]\times[0,1]$ is the unit square. Then

$$\begin{aligned}\iint_\Phi F\cdot dS &= \iint_{[0,1]\times[0,1]} (-u(2u) - v(2v) + (u^2+v^2))\,d(u,v)\\ &= -\iint_{[0,1]\times[0,1]} (u^2+v^2)\,d(u,v)\\ &= -\frac{2}{3}.\end{aligned}$$

Note very carefully: $\iint_\Phi F\cdot dS$ can be negative in general.

9.5.7. Differential notation. Let $F = (P,Q,R)$. Put

$$\begin{aligned}dS &:= (\Phi_u\times\Phi_v)(u,v)\,d(u,v)\\ &= \left(\frac{\partial(y,z)}{\partial(u,v)}\,d(u,v), \frac{\partial(z,x)}{\partial(u,v)}\,d(u,v), \frac{\partial(x,y)}{\partial(u,v)}\,d(u,v)\right)\\ &= (dy\wedge dz, dz\wedge dx, dx\wedge dy).\end{aligned}$$

Then

$$\begin{aligned}F\cdot dS &= P\frac{\partial(y,z)}{\partial(u,v)}\,d(u,v) + Q\frac{\partial(z,x)}{\partial(u,v)}\,d(u,v) + R\frac{\partial(x,y)}{\partial(u,v)}\,d(u,v)\\ &= P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy,\end{aligned}$$

and

$$\begin{aligned}\iint_\Phi F\cdot dS &= \iint_D P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy\\ &= \iint_D P\frac{\partial(y,z)}{\partial(u,v)}\,d(u,v) + Q\frac{\partial(z,x)}{\partial(u,v)}\,d(u,v) + R\frac{\partial(x,y)}{\partial(u,v)}\,d(u,v).\end{aligned}$$

9.5.8. Invariance under reparametrization up to sign. Recall from (8.9) the jacobian of a map.

Lemma 9.52. *Let*

$$E \xrightarrow{h} D \xrightarrow{\Phi} \mathbb{R}^3,$$

with $J(h) \neq 0$. *Then*

$$\iint_{\Phi\circ h} F\cdot dS = \pm\iint_\Phi F\cdot dS. \tag{9.37}$$

Special case. For $D\subseteq\mathbb{R}^2$, let $E\subseteq\mathbb{R}^2$ be defined by $\{(p,q)\colon (q,p)\in D\}$. Consider

$$E \xrightarrow{h} D \xrightarrow{\Phi} \mathbb{R}^3,$$

with $h(p,q) = (q,p)$. We denote $\Phi\circ h$ by $-\Phi$, and call it the negative of Φ. Since $J(h) = -1 < 0$,

$$\iint_{-\Phi} F\cdot dS = -\iint_\Phi F\cdot dS. \tag{9.38}$$

9.5.9. Relating dS and $|dS|$. Suppose $(\Phi_u \times \Phi_v)(u,v) \neq 0$. That is, the normal vector to the surface at (u,v) under the parametrization Φ is nonzero. Put

$$n(\Phi(u,v)) := \frac{(\Phi_u \times \Phi_v)(u,v)}{\|(\Phi_u \times \Phi_v)(u,v)\|}, \tag{9.39}$$

the *unit normal vector* to the surface Φ at (u,v). Then

$$\iint_\Phi F \cdot dS = \iint_\Phi F \cdot n \, |dS|, \tag{9.40}$$

with lhs defined by (9.36) and rhs by (9.33).

PROOF. This is a consequences of our definitions:

$$\begin{aligned}\iint_\Phi F \cdot dS &= \iint_D F(\Phi(u,v)) \cdot (\Phi_u \times \Phi_v)(u,v)\, d(u,v) \\ &= \iint_D F(\Phi(u,v)) \cdot n(\Phi(u,v)) \, \|(\Phi_u \times \Phi_v)(u,v)\| \, d(u,v) \\ &= \iint_\Phi F \cdot n \, |dS|.\end{aligned}$$

In the second step, we used definition (9.39). □

9.5.10. Geometric surface. A *geometric surface* in $\mathbb{R}^3$ is the image of a parametrized surface. An *orientation* of a geometric surface S is an assignment of a unit normal vector at each point of S in a continuous manner.

This can be done in several ways.

(1) If S is the graph of a function of two variables, then specify whether the unit normal points upward with positive z-component (or downward with negative z-component).
(2) If S is a bounded surface without boundary, then specify whether the unit normal points outward (or inward).
(3) If $\Phi : D \to \mathbb{R}^3$ is a 'nice' parametrized surface, then for each (u,v), $\frac{\Phi_u \times \Phi_v}{\|\Phi_u \times \Phi_v\|}$ is a unit normal vector to S at $\Phi(u,v)$, and we obtain an orientation of S.

An *oriented geometric surface* is a geometric surface with a specified orientation. If S is the image of a parametrized surface Φ, then we say Φ is orientation preserving if $\frac{\Phi_u \times \Phi_v}{\|\Phi_u \times \Phi_v\|}$ equals the prescribed unit normal vector. In this case, we define

$$\iint_S F \cdot dS = \iint_\Phi F \cdot dS.$$

If this is not the case, that is, Φ is orientation reversing, then $-\Phi$ is orientation preserving, and we define

$$\iint_S F \cdot dS = \iint_{-\Phi} F \cdot dS = -\iint_\Phi F \cdot dS.$$

9.6. Gauss's divergence theorem

We now generalize FTC, Part II to three dimensions, where the three-dimensional region is a subset of $\mathbb{R}^3$.

9.6.1. Orienting the boundary surface. Let W be a bounded subset of $\mathbb{R}^3$ such that ∂W consists of a finite number of nonintersecting geometric surfaces. Orient each such surface such that the unit normal vector points out of W. Then we say ∂W is *positively oriented*.

Example 9.53. Let $W = \{(x, y, z) \in \mathbb{R}^3 \colon 1 \le x^2 + y^2 + z^2 \le 4\}$. Then ∂W consists of

$$S_1 = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + z^2 = 4\}$$

and

$$S_2 = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + z^2 = 1\}.$$

The positive orientation on ∂W is given by outward normals on S_1 and inward normals on S_2.

More generally, the normal to the outer boundary is outward, and to each of the inner boundaries is inward.

9.6.2. Gauss's divergence theorem. This result relates an integral on a three-dimensional region W to a surface integral on its boundary ∂W.

Theorem 9.54. *Let $W \subseteq \mathbb{R}^3$ be such that ∂W is positively oriented. If P, Q, R are scalar fields defined on an open set containing W, then*

$$\iint_{\partial W} P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy = \iiint_W (P_x + Q_y + R_z)\,d(x, y, z). \tag{9.41}$$

Equivalently, for $F = (P, Q, R)$,

$$\iint_{\partial W} F \cdot dS = \iiint_W \operatorname{div} F\,d(x, y, z). \tag{9.42}$$

Proof. We break the proof in three steps as in the proof of Theorem 9.30.

(i) W is a cuboid. Let $W = [a, b] \times [c, d] \times [p, q]$. Let S_1 be the face of W where $z = p$, and S_2 be the face of W where $z = q$. We orient these surfaces by the normal going out of W. In other words, S_1 is oriented by $n_1 = (0, 0, -1)$, and S_2 is oriented by $n_2 = (0, 0, 1)$. Using formula (5.4) of FTC, Part II,

$$\begin{aligned}\iiint_W R_z\,d(x, y, z) &= \iint_{[a,b]\times[c,d]} R(x, y, q)\,d(x, y) - \iint_{[a,b]\times[c,d]} R(x, y, p)\,d(x, y)\\ &= \iint_{S_2} F \cdot n_2\,|dS| + \iint_{S_1} F \cdot n_1\,|dS|.\end{aligned}$$

Similarly, rewrite $\iiint_W Q_y\,d(x, y, z)$ and $\iiint_W P_x\,d(x, y, z)$ to surface integrals over the remaining two pairs of faces. Adding these equalities proves formula (9.41) when W is a cuboid.

(ii) W is a "union" of cuboids. Now suppose W is formed by adjoining cuboids. Overlapping faces have opposite orientations, so surface integrals over them cancel by (9.38). Thus, formula (9.41) holds when W is formed by adjoining cuboids.

(iii) W is a general region. Draw a cuboidal grid inside W as shown below. Let W' be the largest region which fits inside W which is of the form in item (ii) above and is made out of the faces of the cuboidal grid. By item (ii), formula (9.41) holds for W'. Finally, we approach formula (9.41) for W by considering finer and finer cuboidal grids.

□

Theorem 9.54 gives us an interpretation of divergence: $\operatorname{div} F(a,b,c)$ is the flux of the vector field F through the boundary of a small box around (a,b,c), that is, the net flow of F out of the box. This explains the terminology "divergence".

9.6.3. Principle of deformation.

Corollary 9.55. *If S is a simple closed geometric surface and $P_x+Q_y+R_z = 0$ inside S, then*

$$\iint_S P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy = 0. \tag{9.43}$$

Proof. By (9.41),

$$\iint_{S=\partial W} P\,dy\wedge dz+Q\,dz\wedge dx+R\,dx\wedge dy = \iiint_W (P_x+Q_y+R_z)\,d(x,y,z) = 0.$$

□

Corollary 9.56. *Suppose S_1 and S_2 are simple closed geometric surfaces with S_1 lying inside S_2 and with both oriented by the outward normal. If $P_x + Q_y + R_z = 0$ in the region W between S_1 and S_2, then*

$$\iint_{S_1} P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy = \iint_{S_2} P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy. \tag{9.44}$$

Proof. Let ∂W be the positively oriented boundary of W. By (9.41),

$$\iint_{\partial W} P\,dy\wedge dz+Q\,dz\wedge dx+R\,dx\wedge dy = \iiint_W (P_x+Q_y+R_z)\,d(x,y,z) = 0.$$

Therefore,

$$\iint_{S_2} P\,dy\wedge dz+Q\,dz\wedge dx+R\,dx\wedge dy+\iint_{-S_1} P\,dy\wedge dz+Q\,dz\wedge dx+R\,dx\wedge dy = 0.$$

Using (9.38), the required formula follows. □

Note: In (9.44), P, Q, R must be well-behaved in the region between S_1 and S_2. They may have singularities at other points.

Example 9.57. Let

$$F(x,y,z) = (P,Q,R) = \left(\frac{x}{r^3}, \frac{y}{r^3}, \frac{z}{r^3}\right)$$

on $\mathbb{R}^3 \setminus \{(0,0,0)\}$. Then, one may check that $\operatorname{div} F = P_x + Q_y + R_z = 0$.

Let S be any simple closed geometric surface oriented by the outward normal. Let us compute $\iint_S F\cdot dS$. There are two cases.

(i) $(0,0,0)$ lies outside S. Then $\iint_S F \cdot dS = 0$ by formula (9.43) in Corollary 9.55.

(ii) $(0,0,0)$ lies inside S. Then $\iint_S F \cdot dS = 4\pi$. Why? Let T_ϵ be the sphere of radius ϵ and center $(0,0)$ lying inside S. Then by formula (9.44) in Corollary 9.56,

$$\iint_S F \cdot dS = \iint_{T_\epsilon} F \cdot dS = \iint_{T_1} F \cdot dS = 4\pi.$$

The second equality is by the same reasoning.

9.6.4. Volume calculation.

Corollary 9.58. *Let S be a simple closed geometric surface oriented by the outward normal which encloses a region W. Then*

$$\text{Vol}(W) = \frac{1}{3}\iint_S x\,dy\wedge dz + y\,dz\wedge dx + z\,dx\wedge dy \tag{9.45}$$
$$= \iint_S x\,dy\wedge dz = \iint_S y\,dz\wedge dx = \iint_S z\,dx\wedge dy.$$

Proof. Put $P(x,y,z) = \frac{x}{3}$, $Q(x,y,z) = \frac{y}{3}$, $R(x,y,z) = \frac{z}{3}$. Then $P_x + Q_y + R_z = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$. By formula (8.15),

$$\begin{aligned}\text{Vol}(W) &= \iiint_W d(x,y,z)\\ &= \iiint_W (P_x + Q_y + R_z)\,d(x,y,z)\\ &= \iint_{\partial W = S} P\,dy\wedge dz + Q\,dz\wedge dx + R\,dx\wedge dy\\ &= \frac{1}{3}\iint_S x\,dy\wedge dz + y\,dz\wedge dx + z\,dx\wedge dy.\end{aligned}$$

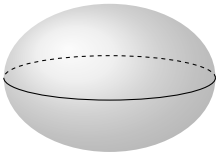
The third equality is by (9.41). □

Corollary 9.59. *For a simple closed surface S parametrized by $(x(u,v), y(u,v), z(u,v))$ which agrees with the outward normal,*

(9.46)

$$\text{Vol}(W) = \frac{1}{3}\iint_D [x(u,v)\frac{\partial(y,z)}{\partial(u,v)} + y(u,v)\frac{\partial(z,x)}{\partial(u,v)} + z(u,v)\frac{\partial(x,y)}{\partial(u,v)}]\,d(u,v)$$
$$= \frac{1}{3}\iint_D \det\begin{bmatrix} x(u,v) & y(u,v) & z(u,v)\\ x_u(u,v) & y_u(u,v) & z_u(u,v)\\ x_v(u,v) & y_v(u,v) & z_v(u,v)\end{bmatrix} d(u,v).$$

Example 9.60. Let S be the surface defined by $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$. See picture below.



Parametrize S by

$$x = a\sin\varphi\cos\theta, \quad y = b\sin\varphi\sin\theta, \quad z = c\cos\varphi$$

for $0 \le \varphi \le \pi$ and $0 \le \theta \le 2\pi$. Then

$$\begin{bmatrix} x(\varphi,\theta) & y(\varphi,\theta) & z(\varphi,\theta) \\ x_\varphi(\varphi,\theta) & y_\varphi(\varphi,\theta) & z_\varphi(\varphi,\theta) \\ x_\theta(\varphi,\theta) & y_\theta(\varphi,\theta) & z_\theta(\varphi,\theta) \end{bmatrix} = \begin{bmatrix} a\sin\varphi\cos\theta & b\sin\varphi\sin\theta & c\cos\varphi \\ a\cos\varphi\cos\theta & b\cos\varphi\sin\theta & -c\sin\varphi \\ -a\sin\varphi\sin\theta & b\sin\varphi\cos\theta & 0 \end{bmatrix}$$

whose determinant equals $abc\sin\varphi$. Thus, by formula (9.46), volume of the elliptical region enclosed by S is

$$\frac{1}{3}\iint\limits_{[0,\pi]\times[0,2\pi]} abc\sin\varphi\, d(\varphi,\theta) = \frac{4}{3}\pi abc.$$

In particular, the volume of a sphere of radius a is $\frac{4}{3}\pi a^3$.

9.7. Stokes theorem

We now generalize FTC, Part II to two dimensions, where the two-dimensional region is a surface in $\mathbb{R}^3$. In the special case when this surface lies in $\mathbb{R}^2$, we recover Green's theorem from Section 9.4.

9.7.1. Orienting the boundary curve. Let S be an oriented geometric surface in $\mathbb{R}^3$. It induces an orientation on ∂S as follows. Walk along ∂S with the prescribed unit normal vector as our upright direction so that S lies to our left. This is the *induced orientation* on ∂S.

9.7.2. Stokes theorem. This result relates a surface integral on an oriented surface S to a line integral on its boundary ∂S.

Theorem 9.61. *Let S be an oriented geometric surface in $\mathbb{R}^3$. Suppose ∂S consists of a finite number of nonintersecting simple closed curves. Let ∂S be given the induced orientation. If P, Q, R are scalar fields defined on a region containing S, then*

$$\begin{aligned} (9.47)\qquad &\int_{\partial S} P\,dx + Q\,dy + R\,dz \\ &= \iint_S (R_y - Q_z)\,dy\wedge dz + (P_z - R_x)\,dz\wedge dx + (Q_x - P_y)\,dx\wedge dy. \end{aligned}$$

Equivalently, for $F = (P, Q, R)$,

$$(9.48)\qquad \int_{\partial S} F\cdot ds = \iint_S \operatorname{curl} F\cdot dS.$$

Proof. We break the proof in three steps.

(i) S is a parallelogram. We assume that the sides of the parallelogram are parallel to the x-axis and y-axis. Let

$$\Phi : [a,b]\times[c,d] \to S, \qquad (x,y)\mapsto (x, y, \alpha x + \beta y).$$

Hence, $\Phi_x \times \Phi_y = (-\alpha, -\beta, 1)$.

$$\iint_S \operatorname{curl} F\cdot dS = \iint_D -\alpha(R_y - Q_z) - \beta(P_z - R_x) + (Q_x - P_y)\,dx\wedge dy.$$

$$\int_{\partial S} F \cdot ds = \int_{\partial D} (P + \alpha R)\, dx + (Q + \beta R)\, dy.$$

Why are the two results equal? The answer is Green's theorem on the rectangle! Let us check this.

$$\begin{aligned}(Q + \beta R)_x - (P + \alpha R)_y &= Q_x + \alpha Q_z + \beta(R_x + \alpha R_z) - (P_y + \beta P_z) - \alpha(R_y + \beta R_z) \\ &= -\alpha(R_y - Q_z) - \beta(P_z - R_x) + (Q_x - P_y).\end{aligned}$$

For the first equality, note very carefully that we are taking partial wrt x of $Q(x, y, \alpha x + \beta y)$, and so on. This proves formula (9.48) when S is a parallelogram.

(ii) S is a "union" of parallelograms. This follows from item (i). The contribution along common edges cancels.

(iii) S is a general surface. Divide S into small pieces so that each piece is roughly a parallelogram. Use item (ii), and then take limits.

□

Example 9.62. Let us verify Stokes theorem, that is, formula (9.48) for the upper hemisphere for the vector field $F = (x, y, z)$.

Observe: $F = \nabla f$ for $f(x, y, z) = \frac{x^2+y^2+z^2}{2}$. Thus, F is a gradient vector field. Hence, $\operatorname{curl} F = 0$ by (9.5). Thus, the surface integral in (9.48) is zero. Also,

$$\int_{\partial S} F \cdot ds = \int_0^{2\pi} (\cos t, \sin t, 0) \cdot (-\sin t, \cos t, 0)\, dt = 0.$$

Thus, the line integral in (9.48) is zero.

Example 9.63. Consider the surface defined by

$$S := \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 + (z - \sqrt{3})^2 = 4,\ z \geq 0\}.$$

It is the portion of the sphere of radius 2 centered at $(0, 0, \sqrt{3})$ which lies above the xy-plane. Orient it by the outward normal of the sphere. Observe:

$$\partial S = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 = 1,\ z = 0\}.$$

This is the unit circle in the xy-plane. The induced orientation on ∂S is anticlockwise. Hence the parametrization $(\cos\theta, \sin\theta, 0)$ for $0 \leq \theta \leq 2\pi$ is orientation preserving.

Consider the vector field $F = (y, -x, e^{xz})$. By (9.48),

$$\begin{aligned}\iint_S \operatorname{curl} F \cdot dS &= \int_{\partial S} F \cdot ds \\ &= \int_0^{2\pi} (\sin\theta, -\cos\theta, e^{(\cos\theta)(0)}) \cdot (-\sin\theta, \cos\theta, 0)\, d\theta \\ &= -2\pi.\end{aligned}$$

Note: Calculating the surface integral directly is not easy.

Example 9.64. Let C be the intersection of the cylinder $x^2 + y^2 = 1$ and plane $x + y + z = 1$. Orient it clockwise when viewed from top. We want to compute

$$\int_C -y^3\, dx + x^3\, dy - z^3\, dz.$$

Observe: C is the boundary of the surface

$$S = \{(x, y, z) \in \mathbb{R}^3 \colon x^2 + y^2 \le 1,\ x + y + z = 1\}.$$

This is the graph of the function $z = f(x, y) = 1 - x - y$ for $(x, y) \in D$, the closed unit disc. Orient S by the unit normal $n = \frac{1}{\sqrt{3}}(1, 1, 1)$. Then the induced orientation on C matches its given orientation.

Let $F(x, y, z) = (-y^3, x^3, -z^3)$. Then $\operatorname{curl} F = (0, 0, 3x^2 + 3y^2)$. Also $(-f_x, -f_y, 1) = (1, 1, 1)$. By (9.48),

$$\begin{aligned}\int_C F \cdot ds &= \iint_S \operatorname{curl} F \cdot dS \\ &= \iint_D (0, 0, 3x^2 + 3y^2) \cdot (1, 1, 1)\, d(x, y) \\ &= 3 \iint_D (3x^2 + 3y^2)\, d(x, y) \\ &= \frac{3\pi}{2}.\end{aligned}$$

Note: Calculating the line integral directly is not easy.

9.7.3. Curl probe. Theorem 9.61 gives us an interpretation of curl: $(\operatorname{curl} F)(a, b, c)$ is the flow of F around a small loop lying in the plane perpendicular to $\operatorname{curl} F$. This explains the terminology "curl".

Curl of a vector field F can be detected using the curl probe. The vector $(\operatorname{curl} F)(a, b, c)$ points in the direction such that if you insert the paddle of the curl probe with its axis in that direction, then it will spin the fastest. The speed at which it spins is proportional to the magnitude of the curl.

9.7.4. Principle of deformation.

Corollary 9.65. *Let C be a simple closed geometric curve in $\mathbb{R}^3$. Let S be an oriented geometric surface in $\mathbb{R}^3$ whose boundary is C. If $Q_x = P_y$, $P_z = R_x$, $R_y = Q_z$, that is, $\operatorname{curl}(P, Q, R) = 0$ in a region containing S, then*

$$\int_C P\, dx + Q\, dy + R\, dz = 0. \tag{9.49}$$

Proof. By (9.47),

$$\begin{aligned}&\int_{C=\partial S} P\, dx + Q\, dy + R\, dz \\ &= \iint_S (R_y - Q_z)\, dy \wedge dz + (P_z - R_x)\, dz \wedge dx + (Q_x - P_y)\, dx \wedge dy = 0.\end{aligned}$$

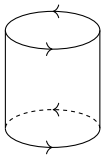
□

Corollary 9.31 is the special case when S lies in the xy-plane.

Corollary 9.66. *Suppose the boundary of S consists of two closed curves C_1 and C_2. If $Q_x = P_y$, $P_z = R_x$, $R_y = Q_z$, then*

$$\int_{C_1} P\, dx + Q\, dy + R\, dz = \int_{C_2} P\, dx + Q\, dy + R\, dz, \tag{9.50}$$

with appropriate orientations on C_1 and C_2 as illustrated in the figure below.



Corollary 9.32 is the special case when S lies in the xy-plane. The above result can be proved in the same way as this special case.

9.7.5. Simply connected region. Let us now look at the converse of Lemma 9.12. We know $\operatorname{curl}(\operatorname{grad} f) = 0$. Suppose we are given $\operatorname{curl} F = 0$. Then is $F = \operatorname{grad} f$ for some f? That is, is F a gradient vector field? That is, is $\int_C F \cdot ds = 0$ for every closed curve C?

The answer is yes when the domain of F is simply connected. In rough terms, it means that the domain has no holes. More precisely, it means that every closed curve in the domain can be shrunk to a point by staying within that domain.

Example 9.67 (Simply connected). We illustrate this concept.

(1) $\mathbb{R}^2$ is simply connected. But $\mathbb{R}^2$ with a point removed is not simply connected.
(2) A closed disc in $\mathbb{R}^2$ is simply connected. But a closed disc with an interior point removed is not simply connected. Similarly, an annulus is not simply connected.
(3) $\mathbb{R}^3$ is simply connected. $\mathbb{R}^3$ with a point removed is simply connected. But $\mathbb{R}^3$ with the entire z-axis removed is not simply connected.

Suppose the domain of F is simply connected, and $\operatorname{curl} F = 0$. We indicate an argument as to why F is a gradient vector field. Let C be any closed curve. Since the domain of F is simply connected, one can find an orientable surface S whose boundary is C. Now apply (9.48) to S.

Exercise 9.68. Consider the vector field F of Example 9.27 but only on the right half plane $x > 0$. This region is simply connected. So F is a gradient vector field on this region. Check that $f(x, y) = \tan^{-1}(\frac{y}{x})$ is a potential function for F on this region. Note very carefully that f is constant along radial lines. This makes sense since F is orthogonal to the radial direction.

Example 9.69. This is a three-dimensional version of Example 9.27. Consider the vector field

$$F(x, y, z) = \left(\frac{-y}{x^2 + y^2}, \frac{x}{x^2 + y^2}, 0\right)$$

on $\mathbb{R}^3 \setminus \{(0, 0, z)\}$, that is, on $\mathbb{R}^3$ minus z-axis. Note: $\operatorname{curl} F = 0$. For the path

$$\gamma : [0, 2\pi] \to \mathbb{R}^3, \qquad \gamma(t) = (\cos t, \sin t, 0),$$

$$\int_\gamma F \cdot ds = \int_0^{2\pi} (-\sin t, \cos t, 0) \cdot (-\sin t, \cos t, 0)\, dt = \int_0^{2\pi} dt = 2\pi.$$

Hence F is not a gradient vector field.

Note: γ goes round the z-axis. So there is no surface S in $\mathbb{R}^3 \setminus \{(0, 0, z)\}$ whose boundary is γ. So the argument given above does not work.

9.8. Differential forms

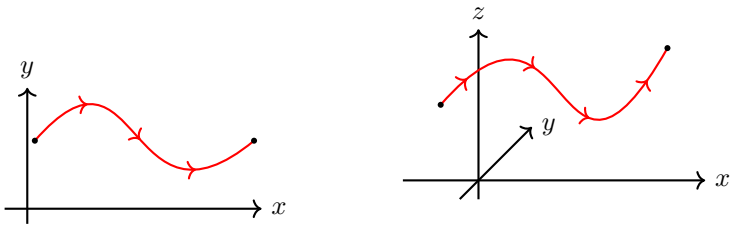
Differentiable manifolds, and more specifically differential forms, provide a general framework in which to express the content of the preceding sections. We provide a brief informal discussion on this point in this concluding section. Note: Lines, planes, curves, surfaces, three-space, and so on are examples of differentiable manifolds.

9.8.1. Orientations. Let us try to understand the notion of orientation of a manifold through examples.

- Dimension zero. Let us start with $\mathbb{R}^0$ which is a point. To orient a point, we specify a number for it, which is either $+1$ or -1. See below.

- Dimension one. For $\mathbb{R}$, to orient it, we specify a direction as shown below.

- Dimension one. Orientations of more complicated one-dimensional objects, that is, curves in $\mathbb{R}^2$ and $\mathbb{R}^3$ are shown below.



- Dimension two. What does it mean to orient $\mathbb{R}^2$? It means that we order the positive x-axis and positive y-axis:
  - x followed by y is the standard orientation, denoted $dx \wedge dy$, and
  - y followed by x, namely, $dy \wedge dx$ is the negative of the standard orientation.

  Thus,
$$dy \wedge dx = -dx \wedge dy.$$
- Dimension two. Orientations of more complicated two-dimensional objects, that is, surfaces in $\mathbb{R}^3$ are defined as follows. Parametrize the surface S by a region D in $\mathbb{R}^2$ whose coordinates we denote by u and v. To orient S, we either orient u before v (standard orientation), or v before u.

  This way of orienting is equivalent to the way we have discussed before, that is, by specifying a normal.
- Dimension three. What does it mean to orient $\mathbb{R}^3$? It means that we order the positive x-axis, positive y-axis, positive z-axis: x followed by y followed by z is the standard orientation, denoted $dx \wedge dy \wedge dz$. Changing the order of any two consecutive axes changes the orientation

to its negative as illustrated below.

$$\begin{array}{ccccc} dx\wedge dy\wedge dz & = & -dy\wedge dx\wedge dz & = & dy\wedge dz\wedge dx \\ \| & & & & \| \\ -dx\wedge dz\wedge dy & = & dz\wedge dx\wedge dy & = & -dz\wedge dy\wedge dx. \end{array}$$

In general: Let M be an oriented manifold of dimension n. Then its boundary ∂M is a manifold of dimension $n-1$. The orientation of M induces an orientation on ∂M.

Illustrate.

9.8.2. Differential forms. Differential forms on the manifold $\mathbb{R}^m$ can be understood in a formal manner. We illustrate them for $m = 0, 1, 2, 3$ in Table 9.3.

TABLE 9.3. Differential forms.

	0-form	1-form	2-form	3-form
$\mathbb{R}^0$	scalar	–	–	–
$\mathbb{R}^1$	$f(x)$	$g(x)dx$	–	–
$\mathbb{R}^2$	$f(x,y)$	$Pdx+Qdy$	$gdx\wedge dy$	–
$\mathbb{R}^3$	$f(x,y,z)$	$Pdx+Qdy+Rdz$	$Pdy\wedge dz+Qdz\wedge dx+Rdx\wedge dy$	$gdx\wedge dy\wedge dz$

9.8.3. Exterior derivative. The *exterior derivative* d is a map which takes k-forms to $(k+1)$-forms. We illustrate it for $k = 0, 1, 2$ below.

- 0 to 1.

$$d : 0\text{-forms} \longrightarrow 1\text{-forms}.$$

We illustrate on $m = 0, 1, 2, 3$.

$$d(\text{scalar}) = 0.$$

$$d\big(f(x)\big) = f'(x)\,dx.$$

$$d\big(f(x,y)\big) = f_x(x,y)\,dx + f_y(x,y)\,dy.$$

$$d\big(f(x,y,z)\big) = f_x(x,y,z)\,dx + f_y(x,y,z)\,dy + f_z(x,y,z)\,dz.$$

- 1 to 2.

$$d : 1\text{-forms} \longrightarrow 2\text{-forms}.$$

We illustrate on $m = 1, 2, 3$.

$$d\big(f(x)\,dx\big) = 0.$$

$$\begin{aligned} d\big(P(x,y)\,dx + Q(x,y)\,dy\big) &= P_y(x,y)\,dy\wedge dx + Q_x(x,y)\,dx\wedge dy \\ &= \big(Q_x(x,y) - P_y(x,y)\big)\,dx\wedge dy. \end{aligned}$$

$$\begin{aligned} d\big(P(x,y,z)\,dx + Q(x,y,z)\,dy + R(x,y,z)\,dz\big) \\ &= P_y(x,y,z)\,dy\wedge dx + P_z(x,y,z)\,dz\wedge dx \\ &\quad + Q_x(x,y,z)\,dx\wedge dy + Q_z(x,y,z)\,dz\wedge dy \\ &\quad + R_x(x,y,z)\,dx\wedge dz + R_y(x,y,z)\,dy\wedge dz \\ &= (R_y(x,y,z) - Q_z(x,y,z))\,dy\wedge dz \\ &\quad + (P_z(x,y,z) - R_x(x,y,z))\,dz\wedge dx \\ &\quad + (Q_x(x,y,z) - P_y(x,y,z))\,dx\wedge dy. \end{aligned}$$

- 2 to 3.

$$d : \text{2-forms} \longrightarrow \text{3-forms}.$$

We illustrate on $m = 2, 3$.

$$d\big(g(x,y)\,dx\wedge dy\big) = 0.$$

$$\begin{aligned} d\big(P(x,y,z)\,dy\wedge dz + Q(x,y,z)\,dz\wedge dx + R(x,y,z)\,dx\wedge dy\big) \\ &= P_x(x,y,z)\,dx\wedge dy\wedge dz + Q_y(x,y,z)\,dy\wedge dz\wedge dx \\ &\quad + R_z(x,y,z)\,dz\wedge dx\wedge dy \\ &= \big(P_x(x,y,z) + Q_y(x,y,z) + R_z(x,y,z)\big)\,dx\wedge dy\wedge dz. \end{aligned}$$

Note very carefully how formulas for gradient, curl, divergence in Section 9.2 emerge naturally in the consideration of the map d.

The map d has the important property that $d^2 = 0$. For $M = \mathbb{R}^3$, we may 'write'

$$0\text{-forms} \xrightarrow{\text{gradient}} 1\text{-forms} \xrightarrow{\text{curl}} 2\text{-forms} \xrightarrow{\text{divergence}} 3\text{-forms}.$$

The fact that curl(gradient) $= 0$ and divergence(curl) $= 0$ that we saw in (9.5) are instances of the property $d^2 = 0$.

9.8.4. Stokes theorem. This result links an integral on a manifold M to an integral on its boundary ∂M.

Theorem 9.70. *Let M be an oriented manifold of dimension m. Put the induced orientation on ∂M. Let ω be an $(m-1)$-form on M. Then*

$$\int_M d\omega = \int_{\partial M} \omega. \tag{9.51}$$

In the lhs, the m-form $d\omega$ on M is obtained by applying the exterior derivative d to the $(m-1)$-form ω on M.

Special case. Let us see how formula (9.51) specializes to formulas that we have seen earlier.

- $m = 1$. We rewrite formula (5.4), namely,

$$\int_a^b f'(x)\,dx = f(b) - f(a)$$

as

$$\int_{[a,b]} f'(x)\,dx = \int_{\partial[a,b]} f(x).$$

Here $M = [a, b]$ which is a one-dimensional manifold. The 0-form on M is $f(x)$, and the resulting 1-form on M is $f'(x)\,dx$.

- $\underline{m = 1}$. More generally: We rewrite formula (9.17), namely,

$$\int_\gamma F \cdot ds = f(\gamma(b)) - f(\gamma(a))$$

as

$$\int_\gamma f_x\,dx + f_y\,dy = \int_{\partial\gamma} f(x)$$

or as

$$\int_\gamma f_x\,dx + f_y\,dy + f_z\,dz = \int_{\partial\gamma} f(x),$$

depending on whether γ is in $\mathbb{R}^2$ or in $\mathbb{R}^3$.

- $\underline{m = 2}$. We rewrite formula (9.20), namely,

$$\iint_D (Q_x - P_y)\,d(x, y) = \int_{\partial D} P\,dx + Q\,dy$$

as

$$\int_D (Q_x - P_y)\,dx \wedge dy = \int_{\partial D} P\,dx + Q\,dy.$$

We now use only one integral sign even for a double integral.

- $\underline{m = 2}$. Formula (9.47), namely,

$$\int_S (R_y - Q_z)\,dy \wedge dz + (P_z - R_x)\,dz \wedge dx + (Q_x - P_y)\,dx \wedge dy = \int_{\partial S} P\,dx + Q\,dy + R\,dz$$

does not need any rewriting! The only minor difference is that we now use only one integral sign for the surface integral.

- $\underline{m = 3}$. We rewrite formula (9.41), namely,

$$\iiint_W (P_x + Q_y + R_z)\,d(x, y, z) = \iint_{\partial W} P\,dy \wedge dz + Q\,dz \wedge dx + R\,dx \wedge dy$$

as

$$\int_W (P_x + Q_y + R_z)\,dx \wedge dy \wedge dz = \int_{\partial W} P\,dy \wedge dz + Q\,dz \wedge dx + R\,dx \wedge dy.$$

Bibliography

[1] Tom M. Apostol, *Calculus. Vol. I: One-variable calculus, with an introduction to linear algebra*, Second edition, Blaisdell Publishing Co. Ginn and Co., Waltham, Mass.-Toronto, Ont.-London, 1967. 1, 44, 52, 60

[2] ———, *Calculus. Vol. II: Multi-variable calculus and linear algebra, with applications to differential equations and probability*, Second edition, Blaisdell Publishing Co. Ginn and Co., Waltham, Mass.-Toronto, Ont.-London, 1969. 1, 60, 84, 90

[3] ———, *Mathematical analysis*, second ed., Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1974. 1

[4] Michael Artin, *Algebra*, Prentice Hall Inc., Englewood Cliffs, NJ, 1991. 4

[5] Patrick Billingsley, *Probability and measure*, third ed., Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1995, A Wiley-Interscience Publication. 46

[6] William M. Boothby, *An introduction to differentiable manifolds and Riemannian geometry*, second ed., Pure and Applied Mathematics, vol. 120, Academic Press Inc., Orlando, FL, 1986. 1, 103

[7] Raoul Bott and Loring W. Tu, *Differential forms in algebraic topology*, Graduate Texts in Mathematics, vol. 82, Springer-Verlag, New York, 1982. 108

[8] Andrew Browder, *Mathematical analysis*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1996, An introduction. 1

[9] Manfredo Perdigão do Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall Inc., Englewood Cliffs, N.J., 1976, Translated from the Portuguese. 1

[10] ———, *Differential forms and applications*, Universitext, Springer-Verlag, Berlin, 1994, Translated from the 1971 Portuguese original. 1, 103

[11] David S. Dummit and Richard M. Foote, *Abstract algebra*, third ed., John Wiley & Sons Inc., Hoboken, NJ, 2004. 4

[12] Sudhir R. Ghorpade and Balmohan V. Limaye, *A course in calculus and real analysis*, Undergraduate Texts in Mathematics, Springer, New York, 2006. 1, 34, 43, 44, 45, 51, 52

[13] ———, *A course in multivariable calculus and analysis*, Undergraduate Texts in Mathematics, Springer, New York, 2010. 67, 69, 70, 74, 75, 76, 79, 90, 93, 96, 99

[14] Paul R. Halmos, *Naive set theory*, Springer-Verlag, New York, 1974, Reprint of the 1960 edition, Undergraduate Texts in Mathematics. 1

[15] Jeffrey M. Lee, *Manifolds and differential geometry*, Graduate Studies in Mathematics, vol. 107, American Mathematical Society, Providence, RI, 2009. 1

[16] Balmohan V. Limaye, *Functional analysis*, second ed., New Age International Publishers Limited, New Delhi, 1996. 61

[17] Saunders Mac Lane, *Categories for the working mathematician*, second ed., Graduate Texts in Mathematics, vol. 5, Springer-Verlag, New York, 1998. 1

[18] Jerrold E. Marsden, Anthony J. Tromba, and Alan Weinstein, *Basic multivariable calculus*, Springer-Verlag, 1993. 1

[19] Shigeyuki Morita, *Geometry of differential forms*, Translations of Mathematical Monographs, vol. 201, American Mathematical Society, Providence, RI, 2001, Translated from the two-volume Japanese original (1997, 1998) by Teruko Nagase and Katsumi Nomizu, Iwanami Series in Modern Mathematics. 1

[20] James R. Munkres, *Topology*, Prentice Hall, Inc., Upper Saddle River, NJ, 2000, Second edition of [MR0464128]. 1, 25

[21] Andrew Pressley, *Elementary differential geometry*, second ed., Springer Undergraduate Mathematics Series, Springer-Verlag London Ltd., London, 2010. 1

[22] Charles Chapman Pugh, *Real mathematical analysis*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 2002. 1

[23] H. L. Royden, *Real analysis*, third ed., Macmillan Publishing Company, New York, 1988. 46

[24] Walter Rudin, *Principles of mathematical analysis*, third ed., McGraw-Hill Book Co., New York, 1976, International Series in Pure and Applied Mathematics. 1

[25] ———, *Real and complex analysis*, third ed., McGraw-Hill Book Co., New York, 1987. 46

[26] George F. Simmons, *Introduction to topology and modern analysis*, Robert E. Krieger Publishing Co. Inc., Melbourne, Fla., 1983, Reprint of the 1963 original. 1

[27] Michael Spivak, *A comprehensive introduction to differential geometry. Vol. I-V*, second ed., Publish or Perish Inc., Wilmington, Del., 1979. 1

[28] Terence Tao, *Analysis. I*, second ed., Texts and Readings in Mathematics, vol. 37, Hindustan Book Agency, New Delhi, 2009. 1

[29] ———, *Analysis. II*, second ed., Texts and Readings in Mathematics, vol. 38, Hindustan Book Agency, New Delhi, 2009. 1

[30] John A. Thorpe, *Elementary topics in differential geometry*, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1994, Corrected reprint of the 1979 original. 1

[31] Charles A. Weibel, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics, vol. 38, Cambridge University Press, Cambridge, 1994. 108