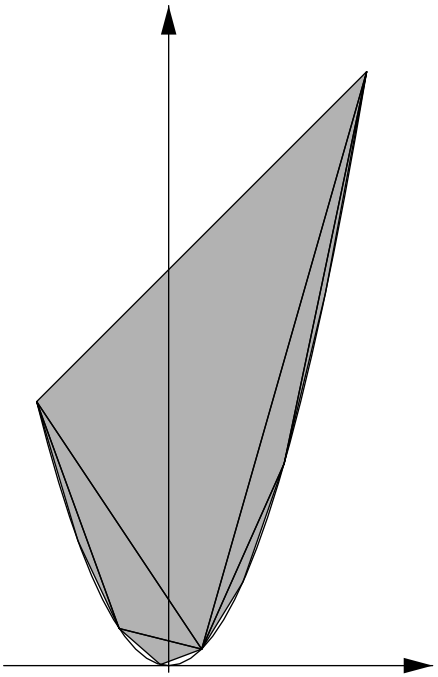
Calculus of One Variable



Jerry Shurman
Reed College

Contents

Preface xi

1 The Parabola 1
1.1 The Parabola in Euclidean Geometry and in Algebra 1
1.1.1 The Geometric Defining Property 1
1.1.2 The Algebraic Defining Equation 2
1.2 Quadrature of the Parabola 6
1.2.1 The Problem 6
1.2.2 The First Inscribed Triangle and Its Key Property 7
1.2.3 Adding More Triangles 8
1.2.4 Archimedes's Evaluation of a Sum 10
1.2.5 Solution of the Problem 12
1.3 Tangent Slopes of the Parabola 13
1.3.1 Difference-Quotient and Secant Slope 13
1.3.2 The Calculation Algebraically and Geometrically 14
1.3.3 The Inscribed Triangle Again 17
1.3.4 The Reflection Property of the Parabola 18
1.4 Summary 20

2 The Rational Power Function 21
2.1 Preliminaries 22
2.1.1 Assumptions About the Number System 22
2.1.2 The Finite Geometric Sum Formula 24
2.2 The Rational Power Function 25
2.2.1 Definition of the Rational Power Function 26
2.2.2 Increasing/Decreasing Behavior 31
2.3 Integration of a Particular Rational Power Function 33
2.3.1 The Problem 33
2.3.2 Intuitive Vocabulary 33

2.3.3 The Idea to be Demonstrated 34
2.3.4 The Problem Again, and the Pending Calculation 36
2.3.5 Tools To Be Used 37
2.3.6 The Geometric Partition 38
2.3.7 The Intervals and Their Widths 40
2.3.8 The Inner Box-Areas 41
2.3.9 The Sum of the Inner Box-Areas 41
2.3.10 The Limiting Value 43
2.4 Differentiation of the Rational Power Function 44
2.4.1 The Problem 44
2.4.2 The Calculation 45
2.4.3 A Fundamental Observation 47
2.5 Integration of the Rational Power Function 48
2.5.1 The Normalized Case 48
2.5.2 The General Case 51
2.6 Summary 54

3 Sequence Limits and the Integral 55
3.1 Sets, Functions, and Sequences 56
3.1.1 Sets 56
3.1.2 Functions 61
3.1.3 Sequences 66
3.1.4 Previous Examples 67
3.2 The Limit of a Real Sequence 67
3.2.1 Absolute Value and Distance 67
3.2.2 The Archimedean Property of the Real Number System 72
3.2.3 Definition of Sequence Limit 73
3.2.4 Basic Sequence Limit Rules 77
3.2.5 Irrelevance of Finite Index-Shifts 85
3.2.6 Uniqueness of the Limit 86
3.2.7 Progenitive Sequence Limit Rules 87
3.2.8 Geometric Series 95
3.2.9 Order Sequence Limit Rules 96
3.3 Integrability 99
3.3.1 The Previous Examples Revisited 99
3.3.2 Definition of Integrability 105
3.3.3 Monotonicity and Integrability 108
3.3.4 A Basic Property of the Integral 109
3.3.5 Piecewise Monotonicity and Integrability 111
3.3.6 Progenitive Integral Rules 116
3.4 Summary 118

4 Function Limits and the Derivative 119
4.1 The Limit of a Function 120
4.1.1 Definition of Function Limit 120
4.1.2 Basic Function Limit Rules 124
4.1.3 Progenitive Function Limit Rules 126
4.1.4 Order Function Limit Rules 128
4.2 The Derivative 130
4.2.1 Definition of the Derivative 130
4.2.2 A Consequence Worth Noting Immediately 131
4.2.3 The Derivative and the Tangent Line 131
4.2.4 A Basic Derivative: the Power Function Revisited 133
4.2.5 Progenitive Derivative Rules 136
4.3 Summary 143

5 The Logarithm Function 145
5.1 Definition and Properties of the Logarithm 145
5.1.1 Integration With Out-of-Order Endpoints 145
5.1.2 The Fundamental Theorem of Calculus 146
5.1.3 Definition of the Logarithm 147
5.1.4 The Key Property of the Logarithm 147
5.1.5 Proof of the Key Property: A Generality 148
5.1.6 Proof of the Key Property: A Specific Argument 150
5.1.7 Proof of the Key Property: End of the Proof 152
5.1.8 Further Properties of the Logarithm 152
5.2 Logarithmic Growth 154
5.3 Differentiation of the Logarithm 157
5.4 Integration of the Logarithm 162
5.4.1 Another Summation Formula 162
5.4.2 The Normalized Case 163
5.4.3 The General Case 165
5.4.4 The Fundamental Theorem of Calculus Again 169
5.5 Signed Integration in General 170
5.5.1 The Integral Revisited 170
5.5.2 Progenitive Integral Rules Revisited 172
5.5.3 The Area Between Two Curves 176

6 The Exponential Function 181
6.1 Continuity 182
6.1.1 Definition of Continuity 182
6.1.2 Continuity and Integrability 184
6.1.3 The Intermediate Value Theorem 185
6.1.4 Applications of the Intermediate Value Theorem 188

6.2 Definition and Properties of the Exponential Function 190
6.2.1 Definition and Basic Properties 190
6.2.2 Raising to Powers Revisited 193
6.3 Exponential Growth 195
6.4 Differentiation of the Exponential 196
6.5 Integration of the Exponential 200
6.6 The Exponential as a Limit of Powers 203
6.6.1 The Description 203
6.6.2 An Interpretation: Compound Interest 205

7 The Cosine and Sine Functions 209
7.1 The Circumference of the Unit Circle 209
7.2 Definition of the Cosine and the Sine 210
7.3 Identities for the Cosine and the Sine 211
7.3.1 Basic Identities 211
7.3.2 Angle Sum and Difference Formulas 213
7.3.3 Double and Half Angle Formulas 214
7.3.4 Product Formulas 214
7.3.5 Difference Formulas 215
7.4 Differentiation of the Cosine and the Sine 215
7.5 Integration of the Cosine and the Sine 219
7.5.1 Imaginary Exponential 219
7.5.2 The Integrals Via the Imaginary Exponential 221
7.6 Other Trigonometric Functions 224
7.7 Inverse Trigonometric Functions 226

8 Polynomial Approximation and Series Representation 233
8.1 The Finite Binomial Theorem 234
8.2 Preliminaries for the Pending Calculations 238
8.2.1 An Alternative Notation 238
8.2.2 The Power Function Integral With Endpoint 0 239
8.3 The Logarithm 242
8.4 The Exponential 246
8.4.1 A Precalculation 246
8.4.2 The Calculation 248
8.5 The Cosine and the Sine 252
8.6 The Power Function 255
8.6.1 The Polynomial and the Remainder 255
8.6.2 The Infinite Binomial Theorem 257
8.7 Summary 261

9 Theory and Applications of the Derivative 263
9.1 Optimization 263
9.1.1 The Extreme Value Theorem 264
9.1.2 Conditions for Optimization 266
9.1.3 Optimization Story-Problems 268
9.2 The Mean Value Theorem and Its Consequences 280
9.2.1 Statement of the Theorem 280
9.2.2 Consequences of the Mean Value Theorem 283
9.3 Curve Sketching 285
9.4 Related Rates Story-Problems 293

10 Integration via Antidifferentiation 299
10.1 The Fundamental Theorem of Calculus 299
10.1.1 Indefinite Integrals, Antiderivatives 299
10.1.2 The Fundamental Theorem, Part I 303
10.1.3 The Fundamental Theorem, Part II 310
10.2 Basic Antidifferentiation 314
10.3 Antidifferentiation by Forward Substitution 317
10.3.1 The Forward Substitution Formula 317
10.3.2 What the Formula Says and Why It Is True 317
10.3.3 Using the Formula in its Variable-Free Form 317
10.3.4 Improvement: the Formula With Variables 318
10.3.5 Second Improvement: the Procedure Instead of the Formula 320
10.3.6 Basic Forward Substitution Formulas 321
10.3.7 Forward Substitution in Integrals 323
10.4 Antidifferentiation by Inverse Substitution 327
10.4.1 The Inverse Substitution Formula and Why It Is True . 327
10.4.2 The Formula With Variables 327
10.4.3 The Procedure 328
10.4.4 Inverse Substitution in Integrals 330
10.5 Antidifferentiation by Parts 332

A Assumptions About the Real Number System 343

B Partial List of Named Ideas and Results 345

List of Symbols 351

Index 353

Preface

These are course notes for Mathematics 111 at Reed College. They are written for serious liberal arts students who want to understand calculus beyond memorizing formulas and procedures. The prerequisite is three years of high school mathematics, including algebra, euclidean geometry, analytic geometry, and (ideally) trigonometry. To profit from these notes, the student needn't be a math genius or possess large doses of the computational facilities that calculus courses often select for. But the student does need sufficient algebra skills, study habits, energy, and genuine interest to concentrate an investment in the subject matter.

I have tried to put enough verbal exposition in these notes to make at least portions of them readable outside of class. And I have tried to keep the calculations short, tidy, and lightly notated, in the hope of rendering them comprehensible stories that incur belief, rather than rituals to endure. To the extent that the notes are readable, I hope to use classtime *discussing* their contents rather than conform to the model of the instructor transcribing a lecture onto the blackboard from which the students transcribe it into their notebooks in turn. The goal is that the students leave the course not having taken my word about the results, but feeling truly viscerally that the results are inevitable.

Exigencies dictate that Math 111 simultaneously serve students who have taken a calculus course already and students who haven't. These notes attempt to do so in two ways,

- by rebalancing the weight of explanation between mathematical symbols and natural language,
- and by presenting the computations of calculus as little more than end-products of algebra that one could imagine naturally working out for oneself with some nudges in the right direction.

The presentation is meant to *defamiliarize* calculus for those who have seen it already, by undoing any impression of the subject as technology to use without understanding, while making calculus *familiar* to a wide range of readers, by which I mean comprehensible in its underlying mechanisms. Thus the notes will pose different challenges to students with prior calculus experience and to students without it. For students in the first group, the task is to consider the subject anew rather than fall back on invoking rote techniques. For students with no prior calculus, the task is to gain facility with the techniques as well as the ideas.

These notes address three subjects:

- *Integration*. What is the area under a curve? More precisely, what is a procedure to calculate the area under a curve?
- *Differentiation*. What is the tangent line to a curve? And again, whatever it is, how do we calculate it?
- *Approximation*. What is a good polynomial approximation of a function, how do we calculate it, and what can we say about the accuracy of the approximation?

Part of the complication here is that *area under a curve* and *tangent line to a curve* are geometric notions, but we want to calculate them using analytic methods. Thus the interface between geometry and analysis needs discussion.

The basic pedagogy is to let ideas emerge from calculations. In succession, these notes define, integrate, and differentiate

- the rational power function, $f(x) = x^\alpha$ where α is a rational number, meaning a ratio of whole numbers,
- the logarithm function, $f(x) = \ln(x)$,
- the exponential function, $f(x) = \exp(x) = e^x$,
- the cosine and sine functions, $f(x) = \cos(x)$ and $f(x) = \sin(x)$.

The integrals are computed without using the Fundamental Theorem of Calculus. Integrating the power function leads to the idea that an integral is not only an area, but more specifically an area that is well approximated from below and from above by suitable sums of box-areas. Although the geometrically natural idea is to integrate nonnegative-valued functions from a left endpoint to a right endpoint, the logarithm leads to the idea of integrating a function that could be negative between endpoints that need not be in order. The logarithm also illustrates the idea of defining a function as an integral and then studying its properties as such. Similarly, the exponential function illustrates the idea of defining and then studying a function as the inverse of another, and it suggests the idea of characterizing a function by a differential equation.

With the power function, the logarithm, the exponential, and the cosine and sine integrated and differentiated, we then find approximating polynomials for these functions and estimate the accuracy of the approximations.

Essentially all of the program just sketched can be carried out convincingly (if not "fully rigorously") based on only one small but versatile piece of technology, the finite geometric sum formula. This formula reduces our area calculations, limits of sums of many terms, to limits of quotients of two terms. In fancier language, the formula reduces integration to differentiation. This phenomenon is perhaps unsurprising because the Fundamental Theorem of Calculus says that integration and differentiation are closely related. But whereas the Fundamental Theorem is often taught as a procedure that circumvents computing integrals directly, a goal of these notes is to see differentiation emerge repeatedly from actual integration. Students who learn to integrate only by using the Fundamental Theorem risk gaining little appreciation of what integration really is, an appreciation worth having if only because the Fundamental Theorem is irrelevant to so much integration in applications of mathematics.

Calculus does at some point require the technical machinery of limits. Limits will be treated lightly *after* they are used informally. The general limit definition, with its delicate interaction between quantifiers, deserves its due respect, but first working informally with specific examples is meant to help the reader tangibly appreciate its economy and finesse.

The last two chapters of these notes, on applications of the derivative and on the Fundamental Theorem of Calculus, are traditional. In the footsteps of so many before us, we will maneuver ladders around corners, drain conical swimming pools, and generate blizzards of antiderivatives.

These notes are based on a set of notes by Ray Mayer. The motivation for creating a new set of notes was that when this project began, the other set of notes was not available in electronic form. That situation has now changed, and the reader of these notes is encouraged to look at Ray Mayer's notes as well.

August 2007

Jerry Shurman
Reed College
Portland, OR

1

The Parabola

This chapter uses the parabola to illustrate ideas from calculus quickly and informally. Section 1.1 characterizes the parabola geometrically and algebraically. Section 1.2 computes the area of the region between a parabola and a chord joining two of its points. The computation proceeds by systematically filling the region with triangles, adding ever smaller triangles at each step. The individual triangle-areas are calculable, and the sum of the areas after each step takes a form that lets us determine the value to which it tends the more steps we carry out. This value is the desired area. The parabolic area-calculation is our first example of *integration*, a fundamental process of calculus. Section 1.3 computes the slope of a tangent line to the parabola, first by algebra and then again by geometry. Here the idea is that although the tangent line passes through only one point of the parabola, we can approximate it by secant lines that pass through two parabola points (such lines cut the parabola, i.e., they section it, ergo their name), and the secant slopes are easy to understand. As the second parabola point nears the first, the value tended to by the secant slopes is the tangent slope. The parabolic tangent slope calculation is our first example of *differentiation*, another fundamental process of calculus.

1.1 The Parabola in Euclidean Geometry and in Algebra

1.1.1 The Geometric Defining Property

Working in the euclidean plane and using cartesian coordinates, consider a horizontal line called the **directrix**, set one quarter of a unit down from the x-axis,

$$D = \text{the points } (x, -1/4) \text{ for all values of } x.$$

We abbreviate the description of the directrix by writing

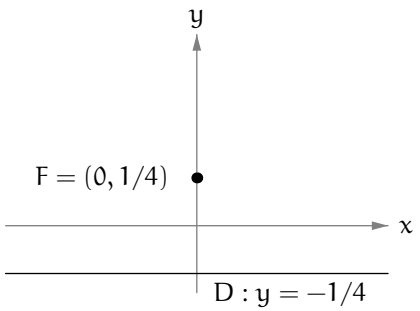


Figure 1.1. Directrix and focus

$$D : y = -1/4.$$

Consider also a point called the **focus**, set one quarter of a unit up the y-axis,

$$F = (0, 1/4).$$

(See figure 1.1.) The parabola is defined geometrically as the locus of all points P that are equidistant from the directrix and from the focus,

$$PD = PF.$$

(See figure 1.2.)

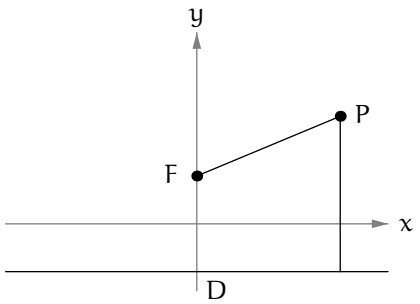


Figure 1.2. Point equidistant from the directrix and the focus

1.1.2 The Algebraic Defining Equation

To translate the geometric condition defining the parabola into an algebraic condition, note that for any point $P = (x, y)$, the square of the distance from P

to the directrix D is the square of the difference of the y-coordinates,

$$PD^2 = (y + 1/4)^2. \tag{1.1}$$

Also, the Pythagorean Theorem (exercise 1.1.1) says that the square of the distance from P to the focus F is

$$PF^2 = (y - 1/4)^2 + x^2. \tag{1.2}$$

(See figure 1.3.) The left sides of (1.1) and (1.2) are equal by the geometric definition PD = PF of the parabola. Therefore the right sides are equal,

$$(y + 1/4)^2 = (y - 1/4)^2 + x^2,$$

and because $(y \pm 1/4)^2 = y^2 \pm y/2 + 1/16$, we may cancel y^2 and $1/16$, and add $y/2$ to both sides, to get simply

$$y = x^2.$$

That is, the parabola is the graph of the squaring function $f(x) = x^2$. The parabola is shown in figure 1.4.

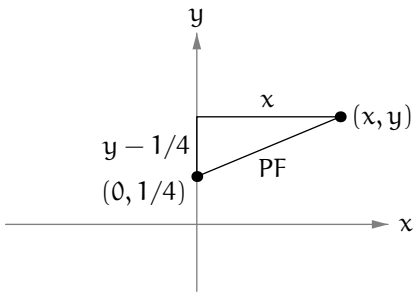


Figure 1.3. Point-to-focus distance by the Pythagorean Theorem

The algebraic equation $y = x^2$ of the parabola is normalized by the choice to place the focus and the directrix one quarter of a unit away from the x-axis. For any positive number r, suppose that instead the focus and the directrix are

$$F : (x, y) = \left(0, \frac{1}{4r}\right), \qquad D : y = -\frac{1}{4r}.$$

Introduce new variables that dilate the original variables by a factor of r,

$$\tilde{x} = rx, \quad \tilde{y} = ry.$$

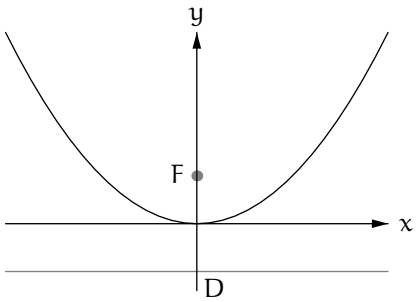


Figure 1.4. The parabola

In the $(\tilde{x}, \tilde{y})$-coordinate system, the focus is

$$F : (\tilde{x}, \tilde{y}) = \left(r \cdot 0, r \cdot \frac{1}{4r} \right) = (0, 1/4),$$

and the directrix is

$$D : \tilde{y} = r \cdot \left(-\frac{1}{4r} \right) = -1/4.$$

That is, in the $(\tilde{x}, \tilde{y})$-coordinate system, the focus and the directrix are back in their normalized positions where we have already studied them, and so the equation of the parabola is

$$\tilde{y} = \tilde{x}^2.$$

Returning to the (x, y)-coordinate system, because $\tilde{y} = ry$ and $\tilde{x} = rx$ the parabola with focus $F = (0, 1/(4r))$ and directrix $D : y = -1/(4r)$ therefore has equation $ry = (rx)^2$, or

$$y = rx^2.$$

If instead r is negative then the parabola opens down instead of up. Similarly, exchanging the roles of x and y to obtain an equation

$$x = ry^2$$

describes a parabola that opens to the right if r is positive, or to the left if r is negative. More generally, the equations

$$y - c = r(x - b)^2, \qquad x - b = r(y - c)^2$$

describe parabolas that are shifted a horizontal distance b and a vertical distance c. For the first of these, the focus and directrix are

$$F = \left(b, c + \frac{1}{4r} \right), \qquad D : y = c - \frac{1}{4r}, \tag{1.3}$$

and for the second they are

$$F = \left(b + \frac{1}{4r}, c\right), \qquad D : x = b - \frac{1}{4r}. \tag{1.4}$$

Exercises

1.1.1. Consider a right triangle with sides a and b and hypotenuse c. The Pythagorean Theorem states that

$$a^2 + b^2 = c^2,$$

i.e., *the square of the hypotenuse is the sum of the squares of the other two sides*. Explain why the shaded region in the right side of figure 1.5 is a square. Then explain why the figure proves the theorem. (Your argument should involve labeling some lengths and angles in the figure.)

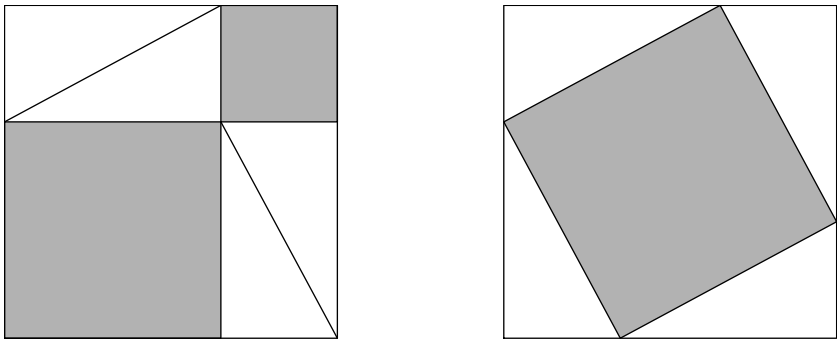


Figure 1.5. Proof of the Pythagorean Theorem

1.1.2. What is the equation of the parabola with focus $F = (1, 2)$ and directrix $D : x = 3$?

1.1.3. Consider the parabola with equation

$$y = Ax^2 + Bx + C, \quad A \neq 0.$$

Where are its focus and its directrix? Solve this problem by using one of (1.3) or (1.4).

1.1.4. Explain why the equation

$$y^2 + 2xy + x^2 - \sqrt{2}\,y + \sqrt{2}\,x = 0$$

describes a parabola. What are its focus and its directrix? It may help to consider the change of variables

$$\tilde{x} = \frac{x+y}{\sqrt{2}}, \quad \tilde{y} = \frac{-x+y}{\sqrt{2}}.$$

(See figure 1.6.)

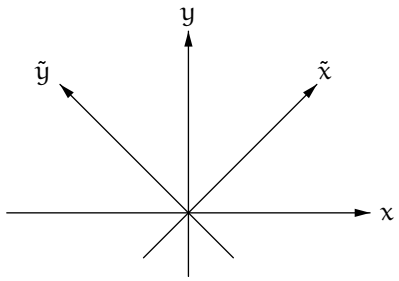


Figure 1.6. Rotated coordinate system

1.2 Quadrature of the Parabola

The earliest example of integration goes back not to Newton (1642–1727) or Leibniz (1646–1716) or Barrow (1630–1677) or Descartes (1596–1650), but to Archimedes (c. 287 BC–c. 212 BC).

1.2.1 The Problem

Let a *and* b *be numbers with* $a < b$. *Find the area between the parabola* $y = x^2$ *and its chord from* (a, a^2) *to* (b, b^2). (See figure 1.7.) Note that the word *area* informally has two different meanings that can easily blur together in this context. First, area means the shaded portion in figure 1.7, i.e., a *region*. But second, area means a *number* that somehow measures the planar size of the region on a linear scale, despite the fact that the region itself lies in the plane rather than on the line. For the parabola problem, we are tacitly assuming that indeed some number is the measure of the shaded region.

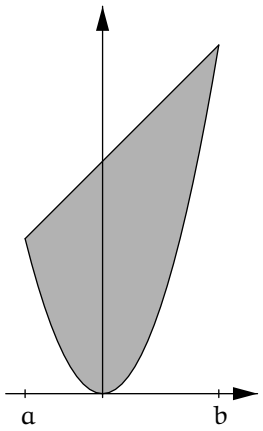


Figure 1.7. Region between a parabola and its chord

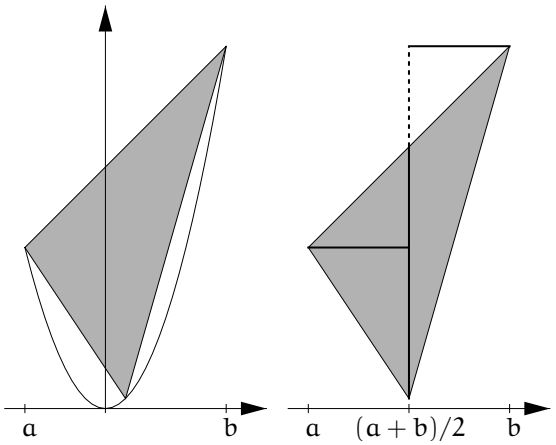


Figure 1.8. First triangle, subdivided vertically

1.2.2 The First Inscribed Triangle and Its Key Property

A first approximation to the region is a triangle with its left and right vertices above a and b and its middle vertex above the midpoint $(a + b)/2$. This triangle can be subdivided down its middle into two triangles sharing a vertical "base" over $(a+b)/2$ from $((a+b)/2)^2$ to $(a^2+b^2)/2$ (these y-coordinates are the square of the horizontal midpoint $(a+b)/2$ and the vertical midpoint of the horizontal endpoints' squares a^2 and b^2), each subtriangle having horizontal "height" $(b-a)/2$ (this is the distance from the midpoint $(a+b)/2$ to either endpoint a or b). (See figure 1.8.) It follows that the area of the triangle is

$$A_{tri} = \left[\frac{a^2+b^2}{2} - \left(\frac{a+b}{2}\right)^2\right]\frac{b-a}{2}, \tag{1.5}$$

and by some algebra (exercise 1.2.1) this works out to

$$A_{tri} = \frac{1}{8}(b-a)^3. \tag{1.6}$$

So the area of the triangle in the left part of figure 1.8 is one eighth of its width cubed. Now (1.6) allows us to make a crucial observation:

The area of the triangle depends only on the width of the triangle.

That is, (1.6) shows that the area depends on the difference $b-a$ but not on a and b individually, so long as the x-coordinate of the third vertex is their average $(a+b)/2$.

Exercise

1.2.1. Carry out the algebra that leads from (1.5) to (1.6).

1.2.3 Adding More Triangles

This observation that the area of a triangle inscribed in the parabola depends only on the triangle's width, provided that the x-coordinate of its middle vertex is the average of the x-coordinates of the left and right vertices, says that very different-looking triangles inscribed in the parabola will have the same area.

In particular, if we add two more triangles to fill in some of the missing space in the left part of figure 1.9, as shown in the right part of the figure, then even though the two new triangles are not congruent, they have the same area, one eighth of their width cubed. Because their width is one half the width of the first triangle, their areas are one eighth the area of the first triangle,

$$A'_{tri} = \frac{1}{8}\left(\frac{b-a}{2}\right)^3 = \frac{1}{8}(b-a)^3 \cdot \frac{1}{8} = A_{tri}\cdot\frac{1}{8}.$$

Adding the two new triangles, each having one eighth the area of the original triangle, adds to the original area a new factor of one quarter the original area, making the total area of the three triangles

$$S_2 = A_{tri} + 2A'_{tri} = A_{tri}\left[1+\frac{1}{4}\right].$$

We are calling this quantity S_2 because it is the triangle-area sum after the second generation of adding triangles. Naturally, S_1 is just A_{tri} itself.

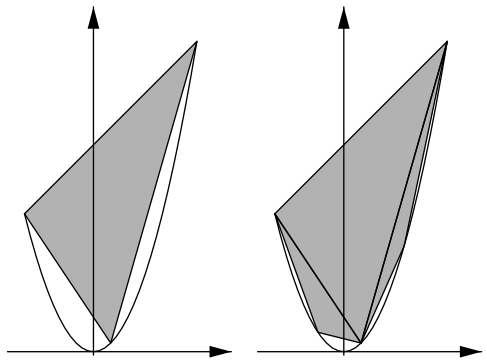


Figure 1.9. Filling in with two more triangles

Next add four more triangles, each half as wide as the two just added. (See figure 1.10. In the figure it is not at all visually suggestive to the author of these notes that the four new triangles all have the same area, and it is hard to tell the difference between three of the four new triangles and the parabolic region that they partially fill.) Thus we add twice as many triangles as at the previous step, each with one eighth the area of the ones added at the previous step, so that the new contribution to the area is one fourth the contribution of the previous step, which in turn was one fourth the area of the original triangle. So after three generations of adding triangles, the area of the seven triangles is

$$S_3 = A_{\text{tri}} + 2A'_{\text{tri}} + 4A''_{\text{tri}} = A_{\text{tri}} \left[1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 \right].$$

By the same sort of reasoning (exercise 1.2.2), adding eight more triangles gives fourth-generation area

$$S_4 = A_{\text{tri}} + 2A'_{\text{tri}} + 4A''_{\text{tri}} + 8A'''_{\text{tri}} = A_{\text{tri}} \left[1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^3 \right], \quad (1.7)$$

and so on. After n generations of adding triangles, each step has added twice as many triangles as the previous step, each triangle having one eighth the area of those added at the previous step, contributing in total a quarter of the previous contribution. That is, after n generations the area is

$$S_n = A_{\text{tri}} \left[1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots + \left(\frac{1}{4}\right)^{n-1} \right]. \quad (1.8)$$

The next task is to evaluate the sum in (1.8).

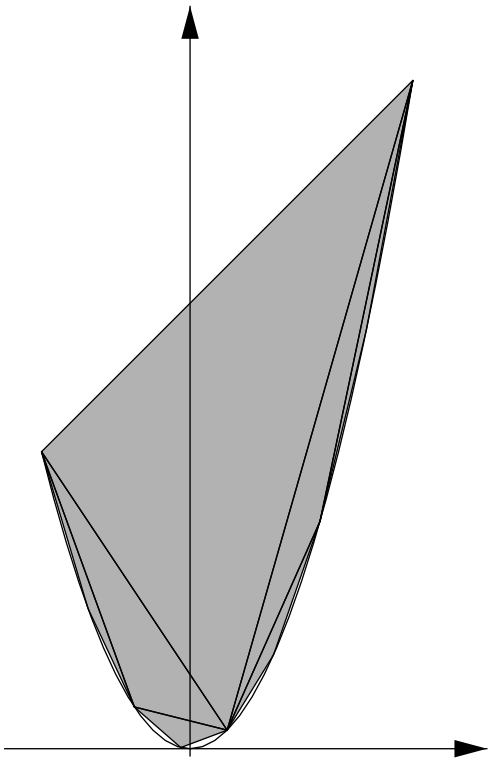


Figure 1.10. Four more triangles

Exercise

1.2.2. Explain carefully why equation (1.7) is correct.

1.2.4 Archimedes's Evaluation of a Sum

As mentioned, the calculation being shown here is due to Archimedes, but his mathematical environment was purely geometric whereas we have made heavy use of cartesian coordinates and algebra. In particular, Archimedes used a geometric argument to evaluate the sum in (1.8). For the argument when $n = 4$, consider the unit square shown in figure 1.11. The largest, lightest L-shaped region has area three quarters. Each successive, darker L-shaped region has linear dimensions half as big as its predecessor's, and hence

area one fourth of its predecessor's. That is, the total area of the four L-shaped pieces is

$$\frac{3}{4}\left[1+\frac{1}{4}+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^3\right].$$

Also, the small, dark square in the upper left corner has sides $(1/2)^4$ and hence area $(1/4)^4$. And the total area of the square is 1. Thus

$$\frac{3}{4}\left[1+\frac{1}{4}+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^3\right]+\left(\frac{1}{4}\right)^4=1,$$

and the geometric argument has given a tidy form to the sum of the first four powers of 1/4 (the zeroth power through the third power),

$$1+\frac{1}{4}+\left(\frac{1}{4}\right)^2+\left(\frac{1}{4}\right)^3=\frac{4}{3}\left(1-\left(\frac{1}{4}\right)^4\right).$$

The argument generalizes immediately to the sum of the first n powers of 1/4,

$$1+\frac{1}{4}+\left(\frac{1}{4}\right)^2+\cdots+\left(\frac{1}{4}\right)^{n-1}=\frac{4}{3}\left(1-\left(\frac{1}{4}\right)^n\right). \tag{1.9}$$

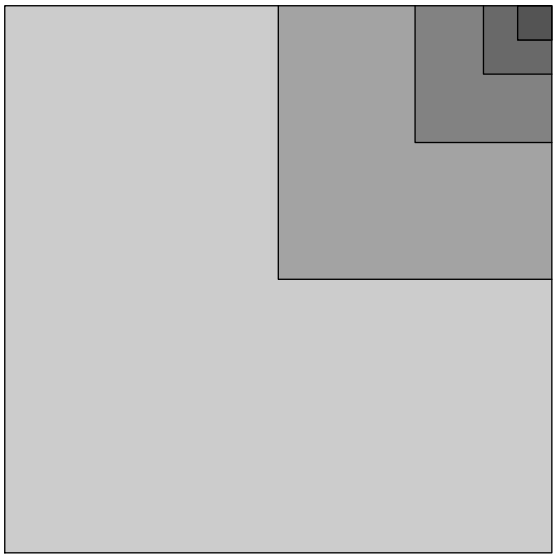


Figure 1.11. Subdivided unit square

1.2.5 Solution of the Problem

Formulas (1.8) and (1.9) combine to show that after n generations of adding triangles, the total area is

$$S_n = A_{\text{tri}} \cdot \frac{4}{3}\left(1 - \left(\frac{1}{4}\right)^n\right).$$

As n grows very large, the triangles fill up the region whose area we want, and so the limiting value of S_n is the region's area. But also, as n grows very large, $(1/4)^n$ gets very close to 0, and so the limiting value of the area is

$$\boxed{S_{\text{lim}} = A_{\text{tri}} \cdot \frac{4}{3}.}$$

That is, the area between the parabola and the chord is four-thirds times the area of the first inscribed triangle. This is how Archimedes formulated the solution. Because $A_{\text{tri}} = \frac{1}{8}(b-a)^3$, another formulation is

$$\boxed{S_{\text{lim}} = \frac{1}{6}(b-a)^3.} \tag{1.10}$$

The reasoning just given to obtain the boxed formulas for S_{lim} calls for scrutiny. To discuss what happens as n grows very large is not to say that any finite number of triangles fill up the parabolic region whose area we want, nor is it to say that the areas of any finite number of triangles sum to $(4/3)A_{\text{tri}}$. And we should be deeply skeptical about treating "infinity" as a number. A more sophisticated formulation of what is being said is that by our choice of configuration,

> *the parabolic region is* **exactly** *the region that the triangles* **tend toward** *filling*,

and that by our calculation,

> *the number* $(4/3)A_{\text{tri}}$ *is* **exactly** *the number that the sums of the triangle-areas* **tend toward** *reaching*,

and that

> *as the triangles tend toward filling the parabolic region, the sums of the triangle-areas must tend toward reaching the area of the parabolic region*,

and therefore

> *the area of the parabolic region must be* $(4/3)A_{\text{tri}}$.

This discussion is not at all satisfactory. It will be refined over the course of these notes.

Exercises

1.2.3. As in the section, let a and b be numbers with $a < b$. Find the area under the parabola and above the x-axis from a to b by using formula (1.10) and the formula for the area of a trapezoid.

1.2.4. Let n denote a positive integer. The larger n is, the closer the quantity $1 + 1/n$ is to 0. Does this mean that the quantity tends to 0? Explain.

1.2.5. The line segment from $(0,0)$ to $(1,0)$ and then the line segment from $(1,0)$ to $(1,1)$ can together be viewed as the one-step staircase from $(0,0)$ to $(1,1)$. Sketch this staircase. The line segments from $(0,0)$ to $(1/2,0)$ to $(1/2,1/2)$ to $(1,1/2)$ to $(1,1)$ give the two-step staircase. Sketch it as well. Similarly, we can make the n-step staircase for any n. What is the total length of the n-step staircase, independently of n? To what shape do the n-step staircases tend as n grows? What is the length of this shape? Do the lengths of the staircases tend to the length of the shape that they tend to?

1.2.6. Raise some criticisms of the "more sophisticated formulation" given at the end of the section.

1.3 Tangent Slopes of the Parabola

1.3.1 Difference-Quotient and Secant Slope

The squaring function is

$$f(x) = x^2.$$

This is the function whose graph is the parabola. For any fixed x, the quantity

$$\frac{s^2 - x^2}{s - x}, \quad s \neq x$$

is a **difference-quotient** and a **secant slope**. *Difference-quotient* means *quotient of differences*, i.e., the numerator of the previous display is the difference $f(s) - f(x)$ of output-values of f, while the denominator $s - x$ is the difference of the corresponding input-values. Meanwhile, a secant line of the parabola is a line through two parabola points, and a *secant slope* is the slope of a secant line. For the interpretation of the difference-quotient as a secant slope, see figure 1.12. The problem is: *To what value does the difference-quotient* (*secant slope*) *tend as* s *tends to* x? The limiting value is called, for reasons to be explained soon, the **tangent slope** of the parabola at (x, x^2).

What makes the problem subtle is that the numerator $s^2 - x^2$ and the denominator $s - x$ of the difference-quotient both tend to 0 individually as s

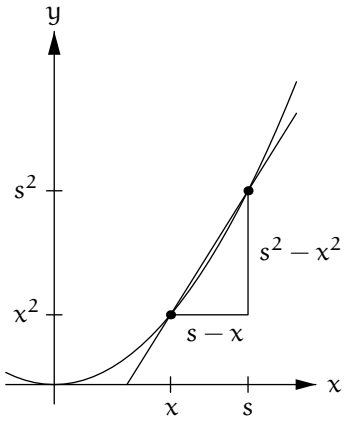


Figure 1.12. Secant lines through $(1, 1)$

tends to x. If instead they both tended to nonzero values then naturally we would guess that the limiting value of the quotient exists, because it should be the quotient of the limiting values. But the fact that the numerator and the denominator both tend to 0, and the fact that the quotient $0/0$ is undefined (this point will be discussed in chapter 2) doesn't *preclude* the possibility that the difference-quotient tends to some well defined value. The issue is that determining the value will require a little analysis.

1.3.2 The Calculation Algebraically and Geometrically

For any fixed real number x, compute that because $s^2 - x^2 = (s + x)(s - x)$,

$$\frac{s^2 - x^2}{s - x} = s + x \quad \text{for } s \neq x.$$

As s tends to x, $s + x$ tends to $2x$. It's that simple. And so:

$$\boxed{\frac{s^2 - x^2}{s - x} \quad \text{tends to } 2x \text{ as } s \text{ tends to } x.}$$

Geometrically, the result is:

$$\boxed{\text{The tangent slope of the parabola } y = x^2 \text{ at the point } (x, x^2) \text{ is } 2x.}$$

This is understood to mean that as s tends to x, the secant line of the parabola through (x, x^2) and (s, s^2) tends to the tangent line to the parabola at (x, x^2), and therefore the number $2x$ tended to by the secant slopes must be the

tangent slope. Like the argument about quadrature, this argument is open to many criticisms—for example, the phrase *tends to* apparently now applies to lines in addition to applying already to regions and to numbers—but we make do with it for the time being.

We can also obtain the tangent slope of the parabola from a direct geometric argument, with no *tends to* in its reasoning. Let x be fixed, and consider the tangent line to the parabola at (x, x^2). For any value s, moving along the parabola from (x, x^2) to (s, s^2) changes the vertical coordinate by

$$s^2 - x^2.$$

(This distance is the gray portion of the y-axis in figure 1.13.) On the other hand, moving from (x, x^2) along the tangent line to the point with first coordinate s changes the vertical coordinate by

$$m(s - x), \quad \text{where } m \text{ is the slope of the tangent line.}$$

(This is the other gray distance in figure 1.13.) We want to find m in terms of x. Because the parabola is convex (i.e., it bends up everywhere), the tangent line lies below the parabola everywhere except at the point of tangency. The vertical distance formulas in the previous two displays show that:

$$\text{Given } x, \text{ we want } m \text{ such that } s^2 - x^2 \geq m(s - x) \text{ for all } s.$$

Equivalently, because $s^2 - x^2$ factors as $(s + x)(s - x)$:

$$\text{Given } x, \text{ we want } m \text{ such that } (s + x)(s - x) \geq m(s - x) \text{ for all } s.$$

By algebra, this is:

$$\text{Given } x, \text{ we want } m \text{ such that } (s + x - m)(s - x) \geq 0 \text{ for all } s.$$

Guided either by hindsight or insight, we see that the correct choice is

$$m = 2x,$$

because then $(s - x)(s + x - m) = (s - x)^2$, and indeed

$$(s - x)^2 \geq 0 \quad \text{for all } s.$$

That is, the tangent slope to the parabola $y = x^2$ at the point (x, x^2) is $2x$, as we already computed above.

Note the differences between the two ways of computing the tangent slope of the parabola. The calculus argument is very quick, and it *tells* us the answer, but it relies on the nebulous notion of *tends to*. By contrast, the geometric argument stands on firmer footing, but it requires us to *know* the answer somehow and then confirm it.

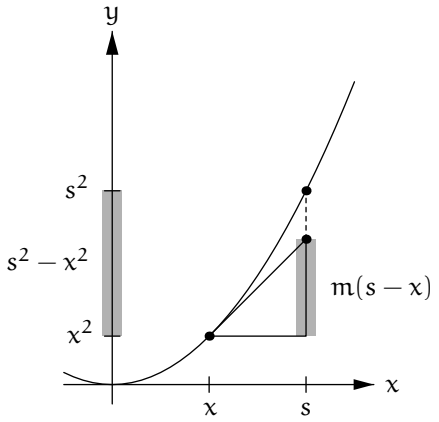


Figure 1.13. Height differences along the parabola and along a tangent line

Exercise

1.3.1. (a) Make a rough sketch of the cubic curve $y = x^3$. Your sketch should show that the curve is convex (bends up) for $x > 0$, is concave (bends down) for $x < 0$, and inflects at $x = 0$.

(b) Here is a geometric argument similar to the one just given for the parabolic curve $y = x^2$, but for the cubic curve instead. *Let x be fixed, and consider the tangent line to the cubic curve at (x, x^3). For any value s, moving along the curve to (s, s^3) changes the vertical coordinate by $s^3 - x^3$, and moving along the tangent line to the point with first coordinate s changes the vertical coordinate by $m(s-x)$, where m is the tangent slope. We want to find m in terms of x. By a little algebra, the difference between the two height-changes is*

$$s^3 - x^3 - m(s-x) = (s^2 + sx + x^2 - m)(s-x).$$

Divine inspiration tells us to consider

$$m = 3x^2.$$

Doing so makes the difference between the two height-changes

$$(s^2 + sx + x^2 - 3x^2)(s-x) = (s+2x)(s-x)^2. \tag{1.11}$$

By the geometry of the cubic curve, as described in part (a)...

Complete the argument that $m = 3x^2$ is the correct choice by explaining why the height difference on the right side of (1.11) is behaving appropriately.

Be clear about the respective roles and behaviors of x and s. The solution is not a simple mimicry of what's in the text—the cubic curve requires a more careful analysis than the parabola. You may assume that x is positive. (But if you want to be thorough then the case where x is negative can be handled similarly, and then the case where $x = 0$ can be handled in a special way.)

(c) What does the height difference on the right side of (1.11) tell us about where the cubic curve and its tangent line at (x, x^3) meet?

(d) Proceed similarly to parts (a) through (c) but with the quartic curve $y = x^4$.

1.3.3 The Inscribed Triangle Again

Recall that the quadrature of the parabola began with an inscribed triangle, its left vertex at $(x, y) = (a, a^2)$ and its right vertex at $(x, y) = (b, b^2)$. (The triangle is shown in the left parts of figure 1.8 and figure 1.9, and it is the large triangle in figure 1.10—see pages 7, 9, and 10.) The slope of the parabolic secant chord between these two vertices is

$$\text{slope between the left and right vertices} = \frac{b^2 - a^2}{b - a} = a + b.$$

Recall also that the triangle's third vertex has x-coordinate $(a+b)/2$. By our formula that the tangent slope to the parabola at any point (x, x^2) is $2x$, we have in particular (because $2 \cdot (a + b)/2 = a + b$),

$$\text{tangent slope to the parabola at the third vertex} = a + b.$$

That is, the inscribed triangle now has a geometric characterization: *The middle vertex is the point where the tangent line to the parabola is parallel to the line through the left and right vertices.* (See figure 1.14.) And in fact this characterization applies to all of the triangles in the quadrature of the parabola.

Exercise

1.3.2. Let a and b be real numbers with $a < b$. Consider two points on the parabola, $P = (a, a^2)$ and $Q = (b, b^2)$. For any number c between a and b, consider also a third point on the parabola, $R = (c, c^2)$. Thus the inscribed triangle used by Archimedes for the quadrature of the parabola occurs when in particular c is the average $(a + b)/2$. Give a geometric argument that of all triangles PQR where P and Q are the fixed points just mentioned, and R is some third point between them on the parabola, Archimedes chose the triangle of greatest area, i.e., the triangle that fills as much as possible of the

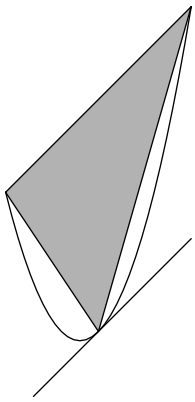


Figure 1.14. Tangent line parallel to secant chord

region between the parabola and the chord PQ. (Hint: Triangle-area is one half of base times height. View PQ is the common base of all the triangles in question. The tangent line to the parabola at Archimedes's choice of R lies below the parabola except at R, and the section has explained that this tangent line is parallel to the chord PQ. Your argument should be based on these ideas and make no reference to the (x, y)-coordinate system.)

1.3.4 The Reflection Property of the Parabola

Let

$$P = (x, x^2)$$

be a point on the parabola, where x is any number. The geometric definition of the parabola is $PD = PF$, and PD is the vertical distance $x^2 + 1/4$ from the point to the directrix, so that also

$$PF = x^2 + 1/4.$$

Consider the point vertically above P, also at distance $x^2 + 1/4$ from P,

$$Q = (x, 2x^2 + 1/4)$$

(see figure 1.15). The slope from F to Q is the y-coordinate difference divided by the x-coordinate difference,

$$m = \frac{2x^2 + 1/4 - 1/4}{x - 0} = 2x.$$

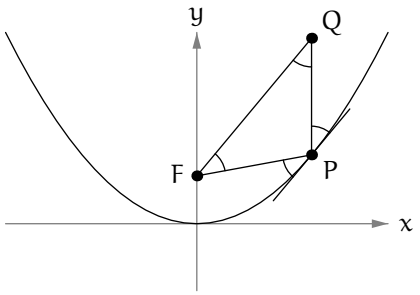


Figure 1.15. Equal angles

This is also the tangent slope of the parabola at P. By Euclidean geometry, the segments QP and PF therefore form the same angle with the tangent line. (Again see figure 1.15.) This gives the reflection property of the parabola: every vertical ray reflects in the parabola to a ray through the focus. The name *focus* for this point is now self-evident. (See figure 1.16.) This property of the parabola is used to construct telescopes and listening devices, and it is sometimes demonstrated in science museum exhibits.

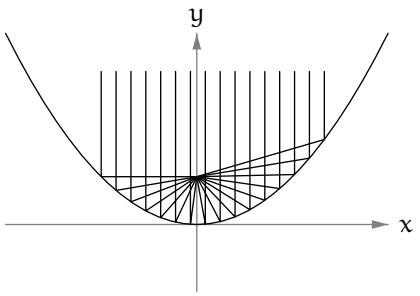


Figure 1.16. Reflected rays meet at the focus

Exercise

1.3.3. Explain why the four angles in figure 1.15 are equal.

1.4 Summary

Most of the work with the parabola in this chapter was geometric or algebraic. What gives the material aspects of calculus as well is that we determined the precise values that approximations tended to as they became ever more accurate. However, the entities involved in the calculations, and the reasoning about them, require closer scrutiny. The calculus that we have done so far is only provisional.

2

The Rational Power Function

The *rational power function* is the function

$$f_\alpha(x) = x^\alpha \quad \text{for } x > 0,$$

where the exponent α is a rational number, meaning a ratio of whole numbers, e.g., $\alpha = 3$ or $\alpha = -2$ or $\alpha = 3/2$ or $\alpha = -17/3$. While x^α is easy to understand when α is a positive whole number—it is $x \cdot x \cdots x$ (α times)—the meaning of x^α for negative whole α, or fractional α, or negative fractional α is less clear.

Section 2.1 lays out some ideas preliminary to studying the rational power function. Basic assumptions about the real number system are stated informally, and then a ubiquitously useful formula is introduced, the *finite geometric sum formula*. Section 2.2 defines the rational power function f_α and shows that for positive values of α the function is always climbing, while for negative values of α the function is always falling. Section 2.3 finds the area under the graph of the specific rational power function $f_{2/3}$ from $x = 1$ to $x = 8$. The process here is integration, and the exposition tries to convey a conceptual sense of it along with the details as it unfolds. Section 2.4 computes the derivative of the rational power function, carrying out the calculation in several steps from a normalized special case to full generality. The derivative calculation reproduces some of the ending work of the integration in the previous section, suggesting a connection between derivatives and integrals. Section 2.5 exploits this connection to calculate the integral of the general power function f_α (excepting the case $\alpha = -1$) between general endpoints $x = a$ and $x = b$.

2.1 Preliminaries

2.1.1 Assumptions About the Number System

Among the many tacit assumptions permeating chapter 1 were assumptions about numbers. We need to proceed from some consensus about how numbers behave. Thus:

We assume that there is a system of **real numbers.**

The assumed real number system has properties that should be familiar. Specifically:

- *The real number system subsumes the rational number system.* An **integer** is a whole number such as 0, 1, −1, 2, −2, 3, ... A **rational number** is a ratio p/q where p and q are integers and q is not 0. But q can be 1, so that the rational numbers subsume the integers. All rational numbers are real, but not all real numbers are rational.
- *Real numbers can be added, subtracted, multiplied, and divided, all subject to the usual rules of algebra.* Division by 0 is prohibited.
- *Every real number is finite.* "Infinity" is not a real number.
- *Real numbers can be compared.* Given any two real numbers, either the first one given is the lesser, or the two are equal, or the first one given is the greater. In particular, the **positive** real numbers are the ones that are greater than 0, and the **negative** real numbers are the ones that are less than 0.
- *The real numbers can be interpreted as the points of a line.* By convention, greater numbers are located to the right of lesser ones. Under this interpretation, the rational numbers are only some of the points of a line, and the rest of the real numbers somehow fill the holes. Any segment of the line having positive length contains both rational and irrational numbers.

Here are some comments about these assumptions. Each point receiving a comment is first repeated in italics.

We assume that there is a system of real numbers. This assumption does not say what a real number is. Nor does the mere act of using the word *real* cause anything to exist. In fact, the notion of the real number system has been understood in different ways at different times, and the current orthodoxy may well no longer be accepted a generation from now. These matters are beyond the scope of this course.

A rational number is a ratio p/q *where* p *and* q *are integers and* q *is not* 0. But 3/2, 15/10, $(-60)/(-40)$, and so on are all the same rational number, so really each p/q is only a *name* of a rational number, and each

rational number has infinitely many names. The usual choice of name for a rational number is the one where p and q are in lowest terms (i.e., they have no common factor greater than 1) and q is positive—for example, $-5/3$ rather than $15/(-9)$.

Not all real numbers are rational. The standard example of an irrational number is the square root of 2. The argument is that if the square root of 2 takes the form p/q then $p^2/q^2 = 2$, and so $p^2 = 2q^2$. But p and q each take the form of a power of 2 times an odd number (i.e., $p = 2^e\tilde{p}$ where $\tilde{p}$ is odd, and $q = 2^f\tilde{q}$ where $\tilde{q}$ is odd), so p^2 and q^2 each take the form of an *even* power of 2 times an odd number (i.e., $p^2 = 2^{2e}\tilde{p}^2$ and $\tilde{p}^2$ is odd, and $q^2 = 2^{2f}\tilde{q}^2$ and $\tilde{q}^2$ is odd), so p^2 is divisible by an even number of 2's while $2q^2$ is divisible by an odd number of 2's. Therefore p^2 can't equal $2q^2$, and the assumption that the square root of 2 takes the form p/q is untenable. However, this argument relies on a tacit assumption that an integer factors *in only one way* as a power of 2 times an odd number. The tacit assumption is true, but its proof takes a nontrivial effort.

Subject to the usual rules of algebra. The reciprocal of a sum is not the sum of the reciprocals:

$$\text{In general,} \quad \frac{1}{a+b} \quad \text{is not} \quad \frac{1}{a}+\frac{1}{b}.$$

For example, try $a = b = 1$: $1/(1+1) = 1/2$ is not $1/1+1/1 = 2$. But students whose algebra is rusty sometimes slip up on this point.

Division by 0 is prohibited. This prohibition is not arbitrary or capricious.

- One explanation is that to divide by a number b is to multiply by its reciprocal, the number b' such that $bb' = 1$. But 0 has no reciprocal because $0b' = 0 \neq 1$ for all b', and so division by 0 makes no sense.
- A second explanation begins by observing that to say that $a/b = c$ is to say that c is the number such that $a = bc$. So for $b = 0$, to say that $a/0 = c$ is to say that c is the number such that $a = 0c$. If $a \neq 0$ then *no* such c exists, while if $a = 0$ then *any* c will work, i.e., all values are equally plausible choices for $0/0$, and so no one value can be preferred. Nonetheless, the particular choices of 0 or 1 as the definition of $0/0$ are often put forward as somehow being natural and not contradicting this second explanation.
- But a third explanation shows that any definition of $0/0$ leads quickly to nonsense. The general rule for adding fractions is

  $$\frac{a}{b}+\frac{c}{d} = \frac{ad+bc}{bd},$$

  so that, if $0/0$ is to have meaning, for any number a,

$$\frac{a}{1}+\frac{0}{0}=\frac{a\cdot 0+1\cdot 0}{1\cdot 0}=\frac{0}{0}.$$

Subtract $0/0$ from both sides to get $a/1=0$, i.e., $a=0$. That is, if $0/0$ is defined then all numbers must be 0.

Truly, division by 0 is a bad idea, even when the numerator is 0 as well.

Infinity is not a real number. There exist extensions of the real number system that contain the symbols ∞ and $-\infty$ and rules such as $a+\infty=\infty$ and $a-\infty=-\infty$ for all finite numbers a. But these extensions compromise the integrity of the original system's algebra, requiring vigilance for cases and leaving new operations undefined, such as $\infty-\infty$. In this context, note that the equality $a/b+0/0=0/0$ from the previous comment suggests that ∞ or $-\infty$ rather than 0 or 1 might be a plausible definition of $0/0$. Predictably, either of these definitions leads to its own set of problems.

In fact, our assumptions so far about the real number system are inadequate for calculus. The process of working through calculus examples will illustrate the various additional assumptions that the subject requires.

Exercises

2.1.1. Argue similarly to the text that the square root of 3 is irrational. Why doesn't the argument apply to the square root of 4? In general, for what positive integers n is the square root of n rational?

2.1.2. Are there numbers a and b for which is it true that "by accident" $1/(a+b)$ does equal $1/a+1/b$?

2.1.2 The Finite Geometric Sum Formula

For any real number $r\neq 1$ and any positive integer n, the sum of the first consecutive n powers of r (starting at the 0th power $r^0=1$) is

$$\boxed{1+r+r^2+\cdots+r^{n-1}=\frac{r^n-1}{r-1}.}$$

This formula is the *finite geometric sum formula*. It reduces a sum of many terms to a quotient of two terms. The quantity whose powers we are summing is denoted r because it is the *ratio* of each pair of consecutive terms: 1 and r, r and r^2, and so on. When convenient (especially when $-1<r<1$, i.e., when 1 is larger in magnitude than r and its powers), we make the numerator and the denominator of the fraction positive by writing instead

$$\boxed{1 + r + r^2 + \cdots + r^{n-1} = \frac{1 - r^n}{1 - r}.}$$

The two ways of writing the finite geometric sum formula have the exact same content. To prove the formula, we may verify instead that the left side times the right side denominator equals the right side numerator,

$$(1 + r + r^2 + \cdots + r^{n-1})(1 - r) = 1 - r^n, \tag{2.1}$$

and this follows (exercise 2.1.3) from multiplying out the left side of (2.1).

Exercises

2.1.3. Verify formula (2.1) by multiplying out its left side.

2.1.4. Use an appropriate choice of r to show that the finite geometric sum formula reproduces Archimedes's calculation of the sum in (1.9) on page 11.

2.1.5. Use the geometric sum formula to show that for any real number $r \neq 1$ and any positive integers n and m,

$$\frac{r^n - 1}{r^m - 1} = \frac{1 + r + r^2 + \cdots + r^{n-1}}{1 + r + r^2 + \cdots + r^{m-1}}.$$

2.2 The Rational Power Function

The following notation is convenient to have at hand:

$$\begin{aligned}
\mathcal{Z} &= \text{the integers,} \\
\mathcal{Z}_{\geq 1} &= \text{the positive integers,} \\
\mathcal{Z}_{\geq 0} &= \text{the nonnegative integers,} \\
\mathcal{Z}_{\leq -1} &= \text{the negative integers,} \\
\mathcal{Q} &= \text{the rational numbers,} \\
\mathcal{R}_{>0} &= \text{the positive real numbers.}
\end{aligned}$$

All of the symbols just introduced are names of **sets**. *Set* means *collection of elements*. We take the notion of a set as something that will be comprehensible in our context. In fact set theory leads to slippery issues very quickly (see exercise 2.2.1), but what matters to us here is that the *paradigm* and the *notation* of set theory are tremendously helpful for organizing one's thoughts in the process of doing mathematics.

Also, the following symbol is ubiquitous in mathematics:

"$\in$" means *in* or *is in* or *is an element of*.

And similarly,

"$\notin$" means *not in* or *is not in* or *is not an element of*.

Thus, $1/2 \in \mathcal{Q}$ (read *one-half is an element of* $\mathcal{Q}$) because indeed $1/2$ is a rational number, but $1/2 \notin \mathcal{Z}$ (read *one-half is not in* $\mathcal{Z}$) because $1/2$ is not an integer. Note: we do not write "$\mathcal{Z} \in \mathcal{Q}$". Every integer is indeed a rational number, so that $\mathcal{Z}$ is a *subset* of $\mathcal{Q}$, but the symbol "$\in$" denotes element containment, not set containment. That is, the symbol "$\in$" is understood to have an individual element to its left and a set containing the element to its right. The language and notation of set theory will be discussed further in chapter 3.

2.2.1 Definition of the Rational Power Function

For any rational number $\alpha \in \mathcal{Q}$, the αth power function is denoted f_α,

$$\boxed{f_\alpha(x) = x^\alpha \quad \text{for positive real numbers } x \in \mathcal{R}_{>0}.}$$

The symbol-string "x^α" is easy enough to write down, but it is only notation. *Writing* x^α *does not address the question of what—if anything—raising a positive real number* x *to a rational power* α *actually means*. We approach the question systematically.

First, for any positive integer α, define for any positive real number x,

$$x^\alpha = x \cdot x \cdots x \ (\alpha \text{ times}) \quad \text{for } \alpha \in \mathcal{Z}_{\geq 1}.$$

Thus for example, exploiting the fact that 1 is multiplicatively neutral,

$$\begin{aligned} x^3 &= 1 \cdot x \cdot x \cdot x, \\ x^2 &= 1 \cdot x \cdot x, \\ x^1 &= 1 \cdot x, \end{aligned}$$

and this pattern extends naturally to the definition

$$x^0 = 1.$$

So now we have x^α for any nonnegative integer α, as compared to any positive integer α as at the beginning of this paragraph.

Second, consider a negative integer α, so that $-\alpha$ is a positive integer. For example, if $\alpha = -3$ then $-\alpha = 3$. The idea is shown in the definition $x^{-3} = 1/x^3$; here we already understand x^3, and $1/x^3$ denotes the multiplicative

inverse of x^3, i.e., the number whose product with x^3 equals 1. In general, define for all positive real numbers x,

$$x^\alpha = 1/x^{-\alpha} \quad \text{for } \alpha \in \mathcal{Z}_{\leq -1}.$$

Repeating, if α is a negative integer, so that $-\alpha$ is a positive integer and we understand $x^{-\alpha}$ for all positive real numbers x, then x^α is the number whose product with $x^{-\alpha}$ is 1. Still x^α is positive even though α is negative.

Third, let α be the reciprocal of a positive integer, so that $1/\alpha$ is itself a positive integer. For example, if $\alpha = 1/5$ then $1/\alpha = 5$. The idea is shown in the definition of $x^{1/5}$ as the unique positive number y such that $y^5 = x$. In general, define for all positive real numbers x,

$$x^\alpha = \text{ the unique positive number } y \text{ such that } y^{1/\alpha} = x \quad \text{for } 1/\alpha \in \mathcal{Z}_{\geq 1}.$$

That is,

$$x^{1/n} = \text{ the unique positive number } y \text{ such that } y^n = x \quad \text{for } n \in \mathcal{Z}_{\geq 1}.$$

For the just-displayed definition of $x^{1/n}$ to make sense, there must be *at least* one suitable y, and there must be *at most* one such y. For now we assume that these conditions do hold, so that indeed a unique y exists, making the definition sensible. This y is called the *positive nth root of* x. For example, $4^{1/2}$ unambiguously means 2, even though -2 also squares to 4. The definition of $x^{1/n}$ as the unique positive nth root of x relies on an assumption about the real number system beyond those that we have already made:

- *Every positive real number has a unique positive nth root for any positive integer n.*

Fourth, to finish defining the rational power function, let $\alpha = p/q$ be any rational number whatsoever, where p is an integer and q is a positive integer. Define for any positive real number x,

$$x^\alpha = (x^{1/q})^p \quad \text{for } \alpha = p/q \in \mathcal{Q},\ p \in \mathcal{Z},\ q \in \mathcal{Z}_{\geq 1}.$$

One can show (see exercise 2.2.2 for a partial proof) that if also $\alpha = p'/q'$ where p' is an integer and q' is a positive integer then $(x^{1/q'})^{p'} = (x^{1/q})^p$, and so the definition of x^α is independent of how the rational exponent α is represented. This completes the definition of the power function. Under the definition, x^α is positive for all positive real x and all rational α.

The relevant body of algebra in this context is *the laws of exponents*. These state that for any positive real numbers x and y, and for any rational numbers α and β,

$$x^\alpha x^\beta = x^{\alpha+\beta}, \qquad (x^\alpha)^\beta = x^{\alpha\beta} = (x^\beta)^\alpha, \qquad x^\alpha y^\alpha = (xy)^\alpha.$$

One can show that the laws of exponents for rational powers are consequences of our definition of raising a positive real number to a rational power. To illustrate with one example of the law $x^{\alpha+\beta} = x^{\alpha}x^{\beta}$, consider any positive real number x, recall that we assume the existence of a unique positive real 10th root of x, note the preliminary calculation that for any positive y we have $y^{14}(1/y^{15}) = y^{14}y(1/y)(1/y^{15}) = y^{15}(1/y)(1/y^{15}) = 1/y$, and compute (justify the steps),

$$\begin{aligned}
x^{7/5}x^{-3/2} &= x^{14/10}x^{-15/10} \\
&= (x^{1/10})^{14}(x^{1/10})^{-15} \\
&= (x^{1/10})^{14}(1/(x^{1/10})^{15}) \\
&= 1/x^{1/10} \\
&= (x^{1/10})^{-1} \\
&= x^{-1/10} \\
&= x^{(14-15)/10} \\
&= x^{14/10-15/10} \\
&= x^{7/5-3/2} \\
&= x^{7/5+(-3/2)}.
\end{aligned}$$

Similarly with the law $x^{\alpha}y^{\alpha} = (xy)^{\alpha}$, for any positive real numbers x and y, note that the preliminary calculation $(x^{1/2}y^{1/2})^2 = x^{1/2}y^{1/2}x^{1/2}y^{1/2} = x^{1/2}x^{1/2}y^{1/2}y^{1/2} = (x^{1/2})^2(y^{1/2})^2 = xy$ shows that $x^{1/2}y^{1/2} = (xy)^{1/2}$, and then compute

$$\begin{aligned}
x^{-3/2}y^{-3/2} &= (x^{1/2})^{-3}(y^{1/2})^{-3} \\
&= \frac{1}{(x^{1/2})^3} \cdot \frac{1}{(y^{1/2})^3} \\
&= \frac{1}{x^{1/2} \cdot x^{1/2} \cdot x^{1/2} \cdot y^{1/2} \cdot y^{1/2} \cdot y^{1/2}} \\
&= \frac{1}{x^{1/2}y^{1/2} \cdot x^{1/2}y^{1/2} \cdot x^{1/2}y^{1/2}} \\
&= \frac{1}{(x^{1/2}y^{1/2})^3} \\
&= (x^{1/2}y^{1/2})^{-3} \\
&= ((xy)^{1/2})^{-3} \\
&= (xy)^{-3/2}.
\end{aligned}$$

As these painstaking examples show, the thing to appreciate is that the laws of exponents are *uniform*; that is, even though the definition of the power

function proceeded by cases, the laws of exponents work the same way regardless of whether each of α and β is a nonnegative integer, a negative integer, a positive rational number that is not an integer, or a negative rational number that is not an integer. A full confirmation of the laws of exponents is symbolically dense, so we don't dwell on it, but we are very glad that the laws work uniformly across all cases.

If α is a nonnegative integer then the given definition of x^α can be extended to all real numbers x, not only positive ones (e.g., we understand $x^3 = x \cdot x \cdot x$ for any x, and we considered the squaring function $f(x) = x^2$ for any x in chapter 1). Note that $0^n = 0$ for $n \in \mathcal{Z}_{\geq 1}$ but $0^0 = 1$. Similarly, if α is a negative integer then the given definition of x^α can be extended to all nonzero real numbers x (e.g., we understand $x^{-3} = 1/x^3$ for any $x \neq 0$). And if $\alpha = 1/n$ for some positive integer n then the definition of $x^\alpha = x^{1/n}$ can be extended to $0^\alpha = 0$ always (because $0^n = 0$, i.e., the nth root of 0 is 0) and also to negative values of x if n is odd (e.g., if $y^3 = 5$ then $(-y)^3 = -5$ because 3 is odd, so that the cube root of -5 is the negative of the cube root of 5). Finally, if $\alpha = p/q$ where p is an integer and q is a positive integer and the fraction p/q is in lowest terms, then the definition $x^\alpha = (x^{1/q})^p$ can be extended to all x if p is nonnegative and q is odd, to all nonzero x if p is negative and q is odd, to all nonnegative x if p is nonnegative and q is even, and to all positive x if p is negative and q is even. This highly nonuniform multitude of cases is bewildering, to say the least. To avoid considering cases in analyzing the αth power function for general rational α, we have simplified our lives by insisting that its inputs be positive, and *we will generally restrict our analysis of the power function to positive inputs*. But the reader should be aware that by standard convention, the inputs to the αth power function are in fact taken to be

- all real numbers if α is a nonnegative integer (for example, $f_3(x) = x^3$ is defined for all x),
- all nonzero real numbers if α is a negative integer (for example, $f_{-2}(x) = x^{-2}$ is defined for all $x \neq 0$),
- all nonnegative real numbers if α is a nonnegative rational number that is not an integer (for example, $f_{3/2}(x) = x^{3/2}$ is defined for all $x \geq 0$),
- all positive real numbers if α is a negative rational number that is not an integer (for example, $f_{-2/5}(x) = x^{-2/5}$ is defined for all $x > 0$).

Later in these notes, we will define the power function for an arbitrary *real* exponent α, i.e., the exponent α will no longer be restricted to rational values.

The last point to be made in this section is that certain particular power functions will arise frequently through these notes, and so the reader should learn to recognize them:

f_0 is the constant function 1,	$f_0(x) = 1$ for all x,
f_1 is the identity function,	$f_1(x) = x$,
f_{-1} is the reciprocal function,	$f_{-1}(x) = 1/x$.

Similarly, the reader should be quickly able to recognize f_2 as the squaring function, $f_{1/2}$ as the square root function, and so on.

Exercises

2.2.1. Let S be the set whose elements are the sets that do not contain themselves as an element. Does the set S contain itself as an element?

2.2.2. Suppose that a positive rational number α takes the forms $\alpha = p/q$ and $\alpha = p'/q'$ where $p, p', q, q' \in \mathcal{Z}_{\geq 1}$. Let x be a positive real number. We want to show that

$$(x^{1/q})^p = (x^{1/q'})^{p'}.$$

(a) An assumption in the section says that it suffices to show instead that

$$\left((x^{1/q})^p\right)^{qq'} = \left((x^{1/q'})^{p'}\right)^{qq'}.$$

Explain.

(b) Without quoting the laws of exponents, explain why the definition of raising a real number to a positive integer exponent and then the definition of raising a real number to the reciprocal of a positive integer imply that

$$\left((x^{1/q})^p\right)^{qq'} = (x^{1/q})^{pqq'} = \left((x^{1/q})^q\right)^{pq'} = x^{pq'},$$

and similarly

$$\left((x^{1/q'})^{p'}\right)^{qq'} = (x^{1/q'})^{p'qq'} = \left((x^{1/q'})^{q'}\right)^{p'q} = x^{p'q}.$$

(c) Explain why the quantities on the right sides of the two displays in part (b) are equal. This completes the argument.

2.2.3. Let x be a positive real number, and let α and β be rational numbers. The symbol-string

$$x^{\alpha^\beta}$$

has two plausible interpretations. Explain. Show by example that the two interpretations can give different values. Which interpretation is the preferred one? Why?

2.2.2 Increasing/Decreasing Behavior

A function f is called **strictly increasing** if for any two input values s and x with $s > x$, also $f(s) > f(x)$; that is, larger input values yield larger output values. Visually the idea is that the graph of a strictly increasing function rises as it moves to the right. Returning to formulas, an equivalent statement is that f is strictly increasing if for any distinct input values s and x (*distinct* means that $s \neq x$), the input-difference $s - x$ and the output-difference $f(s) - f(x)$ have the same sign. Yet a third way to say this is that f is strictly increasing if for any distinct input values s and x, the quotient $(f(s) - f(x))/(s - x)$ is positive, whether it is a quotient of two positive numbers when $s > x$ or of two negative numbers when $s < x$. Similarly, f is **strictly decreasing** if for any distinct input values s and x, the quotient $(f(s)-f(x))/(s-x)$ is negative. We now show that

> *The power function* f_α *for any rational number* α *is strictly increasing if* α *is positive and strictly decreasing if* α *is negative.*

Here it is understood that f_α is taking only positive input values, even though some power functions f_α can take other inupts as well. Having a computer plot various power functions illustrates this statement visually, but establshing it symbolically is a worthwhile intellectual improvement over taking a few computer figures as God-given and as telling the full story. To establish the statement, we need to determine the sign of the quotient $(s^\alpha - x^\alpha)/(s - x)$ where $s \neq x$.

The argument begins by establishing a normalized case of the statement, the case where $x = 1$. Let s be a positive real number other than 1. Let α be a positive integer. By the finite geometric sum formula,

$$\frac{s^\alpha - 1}{s - 1} = 1 + s + s^2 + \cdots + s^{\alpha-1},$$

and the quotient on the left side is positive because the sum on the right side is a nonempty sum of positive terms.

Next let α be a negative integer, so that $-\alpha$ is a positive integer. By algebra,

$$\frac{s^\alpha - 1}{s - 1} = -s^\alpha \frac{s^{-\alpha} - 1}{s - 1}.$$

Because $-s^\alpha$ is negative and because the quotient on the right side is positive by the previous paragraph, altogether the quotient on the left side is negative.

Now let $\alpha = p/q$ where p is a nonzero integer and q is a positive integer. To work easily with this rational exponent, introduce the helper variable $\tilde{s} = s^{1/q}$. Thus $s^\alpha = \tilde{s}^p$ and $s = \tilde{s}^q$, and so

$$\frac{s^\alpha - 1}{s - 1} = \frac{\tilde{s}^p - 1}{\tilde{s}^q - 1} = \frac{\tilde{s}^p - 1}{\tilde{s} - 1} \cdot \frac{\tilde{s} - 1}{\tilde{s}^q - 1}.$$

Because p is a nonzero integer the previous two paragraphs show that the first quotient on the right side has same sign as p, and because q is a positive integer the penultimate paragraph shows that the second quotient on the right side is positive. Altogether, noting that $\alpha = p/q$ has the same sign as p, we have for $\alpha \neq 0$,

$$\frac{s^\alpha - 1}{s - 1} \text{ has the same sign as } \alpha.$$

For the general case, consider any two distinct positive real numbers s and x. Again introduce a helper variable, $\tilde{s} = s/x$, so that $s = \tilde{s}x$ and $\tilde{s}$ is a positive real number other than 1. Compute,

$$\frac{s^\alpha - x^\alpha}{s - x} = \frac{(\tilde{s}^\alpha - 1)x^\alpha}{(\tilde{s} - 1)x} = \frac{\tilde{s}^\alpha - 1}{\tilde{s} - 1} x^{\alpha - 1}.$$

The rightmost quotient has the same sign as α by the normalized case, and $x^{\alpha-1}$ is positive. Thus the leftmost quotient has the same sign as α, and so we have our result:

$$\text{The rational power function } f_\alpha \text{ is } \begin{cases} \text{strictly increasing} & \text{if } \alpha > 0, \\ \text{strictly decreasing} & \text{if } \alpha < 0. \end{cases}$$

In the remaining case $\alpha = 0$, the power function f_0 is the constant function 1.

Exercise

2.2.4. (a) A nonnormalized proof of this section's result begins by computing for any positive real numbers s and x and any positive integer n (using the laws of exponents and the finite geometric sum formula),

$$\begin{aligned} s^n - x^n &= x^n((s/x)^n - 1) \\ &= x^n((s/x) - 1)(1 + (s/x) + (s/x)^2 + \cdots + (s/x)^{n-1}) \\ &= (s - x)x^{n-1}(1 + (s/x) + (s/x)^2 + \cdots + (s/x)^{n-1}). \end{aligned}$$

Continue this calculation to establish a slight generalization of the finite geometric sum formula, the *difference of powers formula*: For any positive real numbers s and x, and any positive integer n,

$$s^n - x^n = (s - x)(x^{n-1} + sx^{n-2} + s^2x^{n-3} + \cdots + s^{n-1}). \tag{2.2}$$

(b) Write out the difference of powers formula for $n = 2$, $n = 3$, and $n = 4$.

2.3 Integration of a Particular Rational Power Function

2.3.1 The Problem

Find the area under the graph of the function

$$f(x) = x^{2/3}$$

from $x = 1$ *to* $x = 8$. The situation is shown in figure 2.1. Because f is the power function f_α where $\alpha = 2/3$ is positive, f is indeed strictly increasing as shown in the figure.

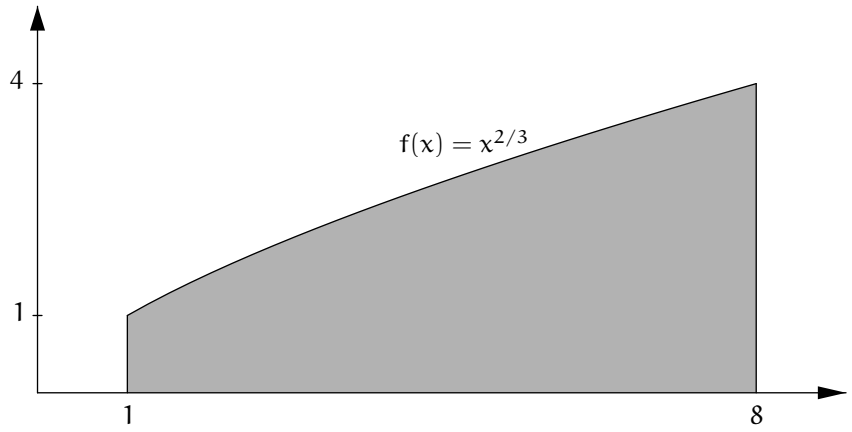


Figure 2.1. Area under a curve

Recall that we believe that raising any real number between 1 and 8 to the two-thirds power is a meaningful thing to do. Specifically, it is understood to mean taking the cube root and then squaring. Squaring a number is noncontroversial, because it is a special case of multiplying two numbers, something that we have assumed we can do. Taking cube roots—and taking nth roots in general—is not part of basic algebra, but it has been appended to the list of things that we assume we may take for granted.

As with the quadrature of the parabola, a tacit assumption in our new problem that can easily pass unnoticed is the assumption that indeed there is an area-number to be found.

2.3.2 Intuitive Vocabulary

Call a positive real number

- **large** if it is very far away from 0,
- **small** if it is very close to 0,
- **medium-sized** if it is moderately far away from or close to 0.

Yes, these terms are too qualitative to be mathematically precise. So the language will need to be understood from context. The point is that being able to track the qualitative sizes of various quantities facilitates insight into the computations of calculus. Imprecision is not acceptable mathematical methodology, but precision is guided by insight. Furthermore, *complete* precision is virtually never attainable, and it is not the goal in and of itself. Precision and insight complement each other in strengthening our understanding.

2.3.3 The Idea to be Demonstrated

The idea is:

> *Computing a medium-sized number can require calculations that use large and small numbers en route. Calculus encodes methods for doing so.*

Indeed, a *calculus* is a stone or pebble. The pebbles of mathematical calculus are the intermediate small numbers that generate a final medium-sized one.

The end-results of calculus conceal the intermediate steps in which the large and the small are manipulated to obtain a medium-sized answer. Operationally the concealment is a convenience because the calculations are detailed, but in practice it means that students often learn to apply calculus mechanically, substituting values into its formulas, without appreciating its finesse. A goal of this set of notes is that students do gain some sense of these matters.

In some situations, rules about numbers are plausibly obvious, but for other scenarios there are no rules. For the following discussion, recall that our informal taxonomy of large, small, and medium-sized applies only to positive numbers.

- (*Situations with plausible rules.*) The sum of two large numbers is a large number. Similarly for two medium-sized numbers and for two small numbers. The sum of a large number and any other positive number is again large. The sum of a medium-sized number and a small number is again medium-sized. The product of a small number and a medium-sized number is small. The product of a medium-sized number and a large number is large. And so on.

  These rules are plausible only at the level of intuition because (again) the terms *large*, *small*, and *medium-sized* are imprecise. To illustrate the imprecision, if the sum of two small numbers is again small, then the sum

of three numbers should be small too, because the sum of the first two small numbers is small and then the threefold sum is the sum of the small twofold sum and the small third number. But by iterating this reasoning, the sum of a thousand small numbers is small, or of a million, and the intuition is no longer valid. The intuition of calculus is fragile because the actual computations can be delicate.

- (*A situation with no rule.*) The sum of many small numbers can be large, medium-sized, or small. For example,

$$10^{-1000} + 10^{-1000} + \cdots + 10^{-1000} \ (10^{2000} \text{ times}) \ = 10^{1000},$$

while

$$10^{-1000} + 10^{-1000} + \cdots + 10^{-1000} \ (10^{1000} \text{ times}) \ = 1,$$

and

$$10^{-1000} + 10^{-1000} + \cdots + 10^{-1000} \ (10^{500} \text{ times}) \ = 10^{-500}.$$

In fact, numbers such as 10^{1000} and 10^{-500} are unimaginably large and small in any sort of physical terms. There are some 10^{77} elementary particles in the universe, and $77 \cdot 13 = 1001$, so 10^{1000} elementary particles would make roughly one-tenth of a universe of universes of universes of universes of universes of universes of universes of universes of universes of universes of universes of universes of universes, a quantity hopelessly beyond physical comprehension. But this is of no consequence, because we are treating numbers as purely platonic entities, not as descriptions of physical quantities.

One master concept of calculus that we will study, the **integral**, *comes—in its simplest form—from sums of ever more, ever-smaller numbers.*

Thus although this bullet says that we do not know in general how such sums behave, the ones that arise in calculus from reasonable situations will behave well in the sense of producing medium-sized answers as they should. We have already seen an example of this in chapter 1, where the finite geometric sum

$$1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots + \left(\frac{1}{4}\right)^{n-1} = \frac{4}{3}\left(1 - \left(\frac{1}{4}\right)^n\right).$$

visibly tends to $4/3$ when we add more and more terms by letting n grow.

- (*Another situation with no rule.*) The quotient of two small numbers can be large, or medium-sized, or small. Indeed, the calculations

$$\frac{10^{-500}}{10^{-1000}} = 10^{500}, \quad \frac{10^{-1000}}{10^{-1000}} = 1, \quad \frac{10^{-1000}}{10^{-500}} = 10^{-500}$$

provide examples.

The other master concept of calculus that we will study, the **derivative**, *comes from quotients of two ever-smaller numbers.*

This bullet says that such quotients can behave wildly, but again the ones that arise in calculus from reasonable situations will produce medium-sized answers. We have already seen an example of this in chapter 1, where the difference-quotient

$$\frac{s^2 - x^2}{s - x}, \quad s \neq x$$

is a ratio of terms that both grow small as s tends to x, but because the difference-quotient is also $s + x$, it is medium-sized, and it visibly tends to $2x$ as s tends to x.

- (*Not-really-another situation with no rule.*) The product of a small number and a large number could be small, medium-sized, or large. This is nothing new because the product can be interpreted as a quotient of two small numbers, or as the quotient of two large numbers. Specifically, if a is small and b is large then the reciprocals a^{-1} and b^{-1} are large and small, and $ab = a/b^{-1} = b/a^{-1}$.

To repeat, a clear understanding of calculus is an understanding of how to compute medium-sized quantities using very large and very small numbers correctly en route. The intermediate steps will require care because their workings are not immediately transparent to our intuition.

Exercise

2.3.1. Describe more situations with plausible rules.

2.3.4 The Problem Again, and the Pending Calculation

Recall the problem: *Find the area under the graph of the function*

$$f(x) = x^{2/3}$$

from $x = 1$ *to* $x = 8$. We are going to approximate the area by calculating the areas of many boxes, as shown in figure 2.2. Here are some features to observe about the figure:

- The region in question is roughly a trapezoid, so our eventual answer should be roughly the corresponding trapezoid-area, the base times the average of the heights, $(8-1)\cdot(1+4)/2 = 17.5$. But because the graph is concave (i.e., it bulges up in the middle, at least according to the computer that drew the figure), the true answer will be a little larger than this.

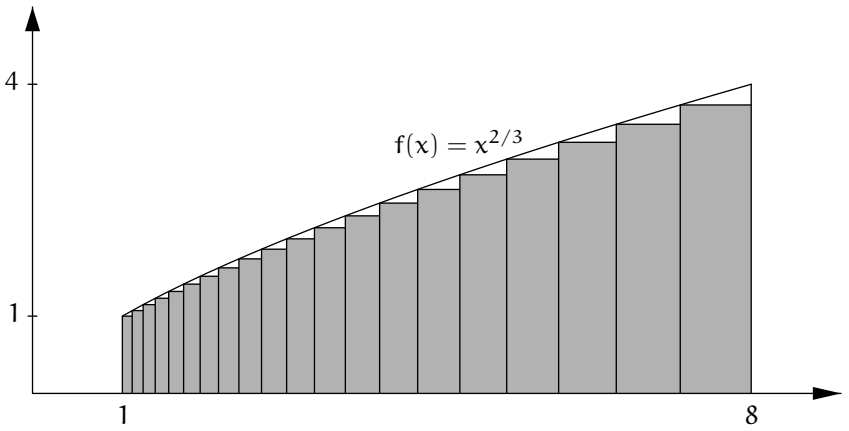


Figure 2.2. Box-areas under a curve

- Each box-height is determined by the value of the function over the left endpoint of the box-base.
- The boxes do not all have the same width, but their widths seem to be regular in some way, and because the graph of the height-determining function $f(x) = x^{2/3}$ is also regular, the box-areas appear regular in turn. The visual regularity of the box-widths and box-areas will soon be explained symbolically.
- Figure 2.2 shows twenty boxes in particular, but the idea is to calculate for n boxes where n is a general-purpose symbol, and then at the end of the calculation, let n grow very large. Although right-more boxes are wider, if n increases enough then plausibly even the rightmost box will grow narrow, and so the boxes will tend toward filling all of the region under the graph.

We will obtain a formula for the sum of the box-areas. Initially, the formula will be a sum of many small numbers, and so its nature will be unclear. But patient calculation will manipulate the formula into an expression that involves only medium-sized numbers, making it easy to understand. Only then will we let the number of boxes grow very large and see to what number the sum of their areas tends.

2.3.5 Tools To Be Used

- *The laws of exponents*. Again, these state that for any positive real numbers x and y, and for any rational numbers α and β,

$$x^\alpha x^\beta = x^{\alpha+\beta}, \qquad (x^\alpha)^\beta = x^{\alpha\beta} = (x^\beta)^\alpha, \qquad x^\alpha y^\alpha = (xy)^\alpha.$$

- *The finite geometric sum formula.* Again, the formula is that for any real number $r \neq 1$ and any positive integer n,

$$1 + r + r^2 + \cdots + r^{n-1} = \frac{r^n - 1}{r - 1},$$

or

$$1 + r + r^2 + \cdots + r^{n-1} = \frac{1 - r^n}{1 - r}.$$

- *Algebra.* As mentioned already, the idea is to calculate for n boxes where n is a general-purpose symbol. Consequently, various other quantities in the calculation will have to be represented by symbols as well rather than numbers, because they depend on n. Only at the end of the calculation, when we let n grow very large, will the symbols that we are working with finally yield an actual number as the answer.

Working through the calculation will also require patience, attention-span, persistence, and study-skills. Because the problem being solved is nontrivial, the solution is larger than bite-sized, perhaps too much to process in one reading. Even for several readings, having a pen and scratch paper at hand to keep track of the main quantities in play may be helpful.

2.3.6 The Geometric Partition

We return to the problem of finding the area under the graph of the function $f(x) = x^{2/3}$ from $x = 1$ to $x = 8$. Again see figure 2.2. Throughout the following calculation, one fundamental quantity is driving everything else:

$$\boxed{\text{The number of boxes is } n.}$$

Thus

$$n \text{ is large.}$$

As already explained, the figure shows twenty boxes but the idea is to calculate for a generic number of boxes, and then only after the calculation yields its result, the number of boxes will then grow very large. Introduce the following quantity:

$$\boxed{\text{The first partition point is } s = 8^{1/n}.}$$

That is, the first partition point s is a real number—dependent on the number of boxes—that is greater than 1. In figure 2.2, s is the right endpoint of the base of the leftmost box. To rephrase the definition:

The first partition point is the positive number s such that $s^n = 8$.

Note that $1^n = 1$, while 2^n is large when n is large. So s lies between 1 and 2, and the more boxes there are, the closer s tends to 1 from the right. Thus (exercise 2.3.2)

$$s - 1 \text{ is small.}$$

Divide the x-axis from $x = 1$ to $x = 8$ into n intervals having the partition points

$$\begin{aligned} x_0 &= s^0 = 1, \\ x_1 &= s^1 = s, \\ x_2 &= s^2, \\ x_3 &= s^3, \\ &\vdots \\ x_{n-1} &= s^{n-1}, \\ x_n &= s^n = 8. \end{aligned}$$

That is, using the symbol j to serve as a counter:

$$\boxed{\text{The partition points are } x_j = s^j \quad \text{for } j = 0, \ldots, n.}$$

This partition of the x-axis from $x = 1$ to $x = 8$ is a *geometric partition* (see figure 2.3, in which $n = 10$), as compared to a *uniform partition*, where all intervals have the same width. The geometric partition will lead nicely to a geometric sum in our pending area-calculation. It does so for reasons that rely on the function $f(x) = x^{2/3}$ of our example being a rational power function. The choice of a geometric partition rather than a uniform partition to solve our integration problem is guided by hindsight, an example of the artfulness of calculus.

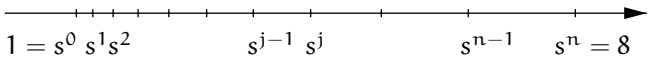


Figure 2.3. A geometric partition

Exercise

2.3.2. This exercise quantifies the assertion that if $s = 8^{1/n}$ then $s-1$ is small when n is large. More generally, let b be any real number greater than 1, and let $s = b^{1/n}$. Here b is fixed but the positive integer n varies.

(a) Use the fact that the power function $f_{1/n}$ is increasing to explain why $s > 1$.

(b) Explain why

$$s - 1 = \frac{b - 1}{1 + s + s^2 + \cdots + s^{n-1}}.$$

(c) Explain why

$$s - 1 < \frac{b - 1}{n},$$

and therefore $s - 1$ is small.

(d) If $0 < b < 1$ (instead of $b > 1$) then what is the nature of $s - 1$ when n is large?

2.3.7 The Intervals and Their Widths

The intervals determined by the geometric partition are

$$\begin{aligned} I_0 &= \text{ the x-axis from } x_0 \text{ to } x_1, \\ I_1 &= \text{ the x-axis from } x_1 \text{ to } x_2, \\ I_2 &= \text{ the x-axis from } x_2 \text{ to } x_3, \\ &\vdots \\ I_{n-1} &= \text{ the x-axis from } x_{n-1} \text{ to } x_n. \end{aligned}$$

That is:

$$\text{The intervals are } I_j = \text{ the x-axis from } x_j \text{ to } x_{j+1} \quad \text{for } j = 0, \ldots, n-1.$$

Their widths are

$$\begin{aligned} \Delta x_0 &= x_1 - x_0 = s - 1, \\ \Delta x_1 &= x_2 - x_1 = s^2 - s = (s-1)s, \\ \Delta x_2 &= x_3 - x_2 = s^3 - s^2 = (s-1)s^2, \\ &\vdots \\ \Delta x_{n-1} &= x_n - x_{n-1} = s^n - s^{n-1} = (s-1)s^{n-1}, \end{aligned}$$

That is:

$$\boxed{\text{The interval-widths are } \Delta x_j = (s-1)s^j \quad \text{for } j = 0, \ldots, n-1.}$$

Thus the jth interval-width is the product of the small number $s - 1$ with a medium-sized number s^j. This symbolic regularity in the formula for the

interval-widths corresponds to the geometric regularity of the widths in figure 2.2. Because s is greater than 1, the formula $\Delta x_j = (s-1)s^j$ shows that the intervals are getting wider as j increases, but even the greatest width, $(s-1)s^{n-1}$, is less than $(s-1)s^n = (8^{1/n}-1)\cdot 8$, and as n gets large this becomes a product of a small number and a medium-sized number, i.e., it becomes small.

2.3.8 The Inner Box-Areas

The base of the jth box is Δx_j. The height of the jth box is the value of the function $f(x) = x^{2/3}$ at the left endpoint of the jth interval,

$$f(x_j) = x_j^{2/3} = (s^j)^{2/3} \quad \text{for } j = 0, \ldots, n-1.$$

Thus the area of the jth box is

$$\Delta x_j \cdot f(x_j) = (s-1)s^j(s^j)^{2/3} \quad \text{for } j = 0, \ldots, n-1.$$

But by the laws of exponents,

$$s^j(s^j)^{2/3} = (s^j)^{5/3} = (s^{5/3})^j.$$

And so:

$$\boxed{\text{The inner box-areas are } (s-1)(s^{5/3})^j \quad \text{for } j = 0, \ldots, n-1.}$$

Thus the jth box-area is the product of the small number $s-1$ with a medium-sized number $(s^{5/3})^j$. As with the interval-widths, the symbolic regularity in the formula for the inner box-areas corresponds to the geometric regularity of the areas in figure 2.2.

2.3.9 The Sum of the Inner Box-Areas

Recall that we have n boxes and that $s = 8^{1/n}$. The sum of the inner box-areas is

$$\begin{aligned} S_n &= (s-1)\cdot\left[(s^{5/3})^0 + (s^{5/3})^1 + (s^{5/3})^2 + \cdots + (s^{5/3})^{n-1}\right] \\ &= (s-1)\cdot\left[1 + (s^{5/3}) + (s^{5/3})^2 + \cdots + (s^{5/3})^{n-1}\right]. \end{aligned}$$

This is a small number, $s-1$, times a sum of many medium-sized numbers, $1 + s^{5/3} + (s^{5/3})^2 + \cdots + (s^{5/3})^{n-1}$. So it is a small number times a large number, and as such, it does not have an obvious size. But, as anticipated, the happy choice of a geometric partition of the x-axis has reduced the sum

of inner box-areas to a finite geometric sum. Specifically, the sum in square brackets is a finite geometric sum with ratio $r = s^{5/3}$, and so by the finite geometric sum formula,

$$S_n = (s-1) \cdot \frac{(s^{5/3})^n - 1}{s^{5/3} - 1}.$$

So we have collapsed the sum, with its many terms, to a quotient of only two terms. But still the factor $s-1$ out front is small, as is the denominator $s^{5/3} - 1$ of the fraction. On the other hand, the numerator is 31, because $s^n = 8$ and so $(s^{5/3})^n = (s^n)^{5/3} = 8^{5/3} = 32$. After rearranging, the sum of the inner box-areas is

$$S_n = 31 \cdot \frac{s-1}{s^{5/3} - 1} \qquad \text{where } s = 8^{1/n}. \tag{2.3}$$

The 31 is quintessentially medium-sized, but the numerator and the denominator of the fraction are both small.

We introduce a helper variable to eliminate the fractional exponent 5/3 from our expression for S_n. Let

$$\tilde{s} = s^{1/3} = 2^{1/n}.$$

So $\tilde{s}$ is slightly bigger than 1. That is,

$$\tilde{s} - 1 \text{ is small.}$$

Because $s = \tilde{s}^3$ and $s^{5/3} = \tilde{s}^5$, the sum of the inner box-areas is now

$$S_n = 31 \cdot \frac{\tilde{s}^3 - 1}{\tilde{s}^5 - 1}, \quad \tilde{s} = 2^{1/n},$$

which rewrites as

$$S_n = 31 \cdot \frac{\left(\dfrac{\tilde{s}^3 - 1}{\tilde{s} - 1}\right)}{\left(\dfrac{\tilde{s}^5 - 1}{\tilde{s} - 1}\right)}, \quad \tilde{s} = 2^{1/n}.$$

Rewriting S_n this way may not seem to help matters, because $\tilde{s}^3 - 1$, $\tilde{s} - 1$, and $\tilde{s}^5 - 1$ are all small. But it sets up the finite geometric sum formula twice more (because $\tilde{s} \neq 1$), expanding sums now rather than collapsing them:

$$\boxed{\text{The inner box-area sum is } S_n = 31 \cdot \frac{1 + \tilde{s} + \tilde{s}^2}{1 + \tilde{s} + \tilde{s}^2 + \tilde{s}^3 + \tilde{s}^4}, \quad \tilde{s} = 2^{1/n}.}$$

And because $\tilde{s}$ is close to 1, the numerator and the denominator of the fraction are now medium-sized. Our prescient choice to use the geometric partition, and then our patient effort of

- deriving a long geometric sum,
- collapsing it to a quotient,
- rearranging the quotient,
- and finally expanding two short geometric sums in the numerator and the denominator of the quotient

have eliminated all large or small numbers from the formula for the sum of the box-areas.

2.3.10 The Limiting Value

Finally the calculation can give a meaningful medium-sized answer. As the number n of boxes grows very large, the helper variable $\tilde{s} = 2^{1/n}$ will tend to 1, and so the sum of inner box-areas will tend to an easily calculable number,

$$S_n \text{ tends to } 31 \cdot \frac{1+1+1}{1+1+1+1+1} = \frac{93}{5} = 18.6.$$

Because the boxes are filling up the region under this curve, this number must be the area. And indeed, it is slightly larger than the original underestimate of 17.5. Summarizing,

The area under the graph of $f(x) = x^{2/3}$ from $x = 1$ to $x = 8$ is 18.6.

Or, introducing some notation,

$$\int_1^8 f = 18.6 \quad \text{where } f(x) = x^{2/3}.$$

That is, *the integral sign "$\int$" is simply shorthand for the area under the graph.*

This is calculus.

Exercises

2.3.3. Show that in a calculation similar to the one in the section but using outer boxes rather than inner boxes gives the following result:

The jth outer box-area is $s^{2/3}$ times the jth inner box-area, $j = 0, \ldots, n-1$.

Therefore the sum of the outer box-areas is $s^{2/3}$ times the sum of the inner box-areas. To what value does $s^{2/3}$ tend as the number n of boxes grows? To what value does the sum of the outer box-areas consequently tend?

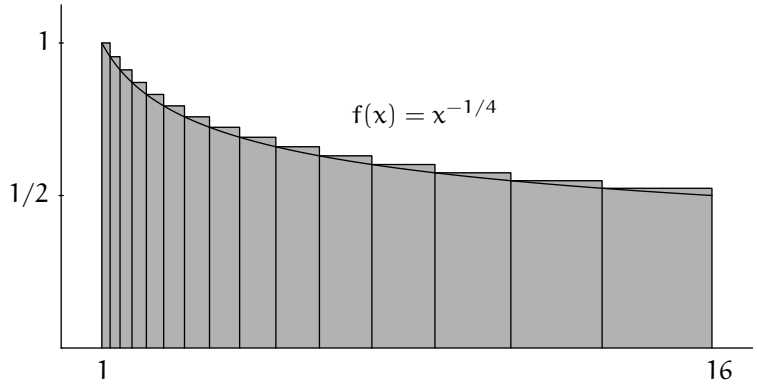


Figure 2.4. Boxes for exercise 2.3.4

2.3.4. Find the area under the graph of the function $f(x) = x^{-1/4} = 1/x^{1/4}$ from $x = 1$ to $x = 16$, using outer boxes. (See figure 2.4. Here the power function $f = f_\alpha$ is strictly decreasing because $\alpha = -1/4$ is negative.) Your writeup should review the ideas of the section.

2.3.5. Find the area under the graph of the function $f(x) = x^{-7/4}$ from $x = 1$ to $x = 16$. (For this function, the picture still looks qualitatively like figure 2.4 because of the negative exponent, but the algebra involves a new wrinkle: your calculations should lead you to an expression involving the quantity $1/(s^{-3/4} - 1)$, different from the example in the section and from the previous exercise because of the negative exponent. However, note that

$$\frac{1}{s^{-3/4} - 1} = -\frac{s^{3/4}}{s^{3/4} - 1},$$

and now the calculation can continue as before.)

2.3.6. Try to apply the same ideas to find the area under the graph of the function $f(x) = 1/x = x^{-1}$ from $x = 1$ to $x = 10$. This time, sum thing goes wrong and the process breaks down. Where does it do so, and why? For what rational exponents α will the power function $f(x) = x^\alpha$ lead to this breakdown?

2.4 Differentiation of the Rational Power Function

2.4.1 The Problem

Recall that for any rational number α, the αth power function is denoted f_α,

$$f_\alpha(x) = x^\alpha \quad \text{for positive real numbers } x \in \mathcal{R}_{>0}.$$

The problem is: *For any $x > 0$, find the limiting value of the difference-quotient*

$$\frac{f_\alpha(s) - f_\alpha(x)}{s - x}$$

as s tends to x. As with the squaring function in chapter 1, the numerator in the display is the difference of the output values of the power function f_α at an input $s \neq x$ and at x itself, and the denominator is the difference of the corresponding input values s and x.

2.4.2 The Calculation

Consider the special case that α is a nonnegative integer. Let $s \in \mathcal{R}_{>0}$ be a positive number other than 1. Repeating the first display of section 2.2.2 we have by the finite geometric sum formula,

$$\frac{s^\alpha - 1}{s - 1} = 1 + s + s^2 + \cdots + s^{\alpha-1},$$

except that now the sum on the right side is understood to be 0 if $\alpha = 0$. As mentioned above, the left side of the previous display is a difference-quotient. On the other hand, the right side of the display is an α-fold sum. If the input-difference $s - 1$ is very small then the summands all tend to 1, and so:

For $\alpha \in \mathcal{Z}_{\geq 0}$, the limiting value of $\dfrac{f_\alpha(s) - f_\alpha(1)}{s - 1}$ as s tends to 1 is α.

(The argument just given supports the statement in the case $\alpha = 0$ if we understand *the summands all tend to* 1 to be vacuous in that case.) So far the boxed result holds only if α is a nonnegative integer. The goal of this section is to show that the boxed result holds when α is any rational number whatsoever, and then to generalize the "1" in the formula to any positive number $x \in \mathcal{R}_{>0}$.

Suppose next that $\alpha \in \mathcal{Z}_{\leq -1}$ is a negative integer. Thus now $-\alpha$ is a positive integer. That is, the boxed result holds with $-\alpha$ in place of α, and we want to re-establish the boxed result for α itself. Repeating the second display from section 2.2.2,

$$\frac{s^\alpha - 1}{s - 1} = -s^\alpha \frac{s^{-\alpha} - 1}{s - 1}.$$

If s tends to 1 then so does $s^\alpha = 1/s^{-\alpha}$. And the quotient on the right side of the display tends to $-\alpha$ by the previous calculation. So the entire right

side tends to $-(-\alpha) = \alpha$. That is, the boxed result has been extended to all integers:

$$\boxed{\text{For } \alpha \in \mathcal{Z}, \text{ the limiting value of } \quad \frac{f_\alpha(s) - f_\alpha(1)}{s-1} \quad \text{as } s \text{ tends to } 1 \text{ is } \alpha.}$$

Now let $\alpha = p/q$ where p and q are integers with q nonzero. As in section 2.2.2, introduce the helper variable $\tilde{s} = s^{1/q}$, so that $s^\alpha = \tilde{s}^p$ and $s = \tilde{s}^q$. Repeating the third display of section 2.2.2,

$$\frac{s^\alpha - 1}{s-1} = \frac{\tilde{s}^p - 1}{\tilde{s}^q - 1} = \frac{\tilde{s}^p - 1}{\tilde{s} - 1} \cdot \frac{\tilde{s} - 1}{\tilde{s}^q - 1}.$$

As s tends to 1, so does $\tilde{s}$. By the results already established in this section, the quotients on the right side of the previous display tend respectively to p and $1/q$ as s tends to 1. Thus their product tends to p/q, i.e., it tends to α. Now the boxed result has been extended to all rational numbers:

$$\boxed{\text{For } \alpha \in \mathcal{Q}, \text{ the limiting value of } \quad \frac{f_\alpha(s) - f_\alpha(1)}{s-1} \quad \text{as } s \text{ tends to } 1 \text{ is } \alpha.}$$

Finally, replace the normalized value 1 by any positive number $x \in \mathcal{R}_{>0}$. Introduce the helper $\tilde{s} = s/x$, so that $s = \tilde{s}x$ and so that $\tilde{s} \neq 1$ whenever $s \neq x$. Repeating the fifth display of section 2.2.2

$$\frac{s^\alpha - x^\alpha}{s-x} = \frac{(\tilde{s}^\alpha - 1)x^\alpha}{(\tilde{s}-1)x} = \frac{\tilde{s}^\alpha - 1}{\tilde{s}-1} x^{\alpha-1}.$$

If s tends to x then $\tilde{s} = s/x$ tends to 1, and so the last quotient in the previous display tends to α; consequently the entire quantity in the previous display tends to $\alpha x^{\alpha-1}$. Thus the boxed result has been extended from $x = 1$ to any positive real number $x \in \mathcal{R}_{>0}$:

$$\boxed{\text{For } \alpha \in \mathcal{Q}, \text{ the limiting value of } \quad \frac{f_\alpha(s) - f_\alpha(x)}{s-x} \quad \text{as } s \text{ tends to } x \text{ is } \alpha x^{\alpha-1}.}$$

This completes the argument. Again introducing some notation, the conclusion is

$$\boxed{\text{For } \alpha \in \mathcal{Q}, \quad f'_\alpha = \alpha f_{\alpha-1}.}$$

That is, *the prime is simply shorthand for the limiting value of difference-quotients*.

Note that when $\alpha = 2$, we recover the formula for the tangent slope of the parabola: the derivative of the squaring function $f_2(x) = x^2$ for all $x > 0$ is the function $2f_1(x) = 2x$ for all $x > 0$. Similarly, the derivative of the identity function $f_1(x) = x$ for all $x > 0$ is the constant function $f_0(x) = 1$ for

all $x > 0$, and the derivative of the constant function $f_0(x) = 1$ for all $x > 0$ is the constant function $0 \cdot f_{-1}(x) = 0$ for all $x > 0$. The reader should understand these last two facts in terms of tangent slopes (exercise 2.4.3).

Because α is a general rational number in this section, it can be replaced by any other rational number, such as a rational number that happens to be named $\alpha + 1$ instead. In this case, now specializing again to $x = 1$:

$$\text{The limiting value of } \frac{s^{\alpha+1} - 1}{s - 1} \text{ as } s \text{ tends to } 1 \text{ is } \alpha + 1.$$

We will use this formula freely through the rest of this chapter.

Exercises

2.4.1. Is there a rational power function f_α whose derivative is f_{-1}? Is there a rational power function f_α whose derivative is any constant multiple of f_{-1}?

2.4.2. The last boxed result in the section took four steps to derive. Rederive it in three steps instead by using the difference of powers formula (2.2) from exercise 2.2.4.

2.4.3. (a) Graph the function $f_1(x) = x$ for all $x > 0$. For any x, what is the tangent slope to the graph at the point $(x, f(x))$? How does your answer relate to our general formula $f'_\alpha = \alpha f_{\alpha-1}$?

(b) Graph the function $f_0(x) = 1$ for all $x > 0$. For any x, what is the tangent slope to the graph at the point $(x, f(x))$? How does your answer relate to our general formula $f'_\alpha = \alpha f_{\alpha-1}$?

2.4.3 A Fundamental Observation

The area calculation in section 2.3 reduced the problem of studying an *integral*—the limiting value of sums of many small terms—to the problem of studying the limiting value of quotients of two small terms. Specifically, computing the area under the graph of the power function

$$f_{2/3}(x) = x^{2/3}$$

from $x = 1$ to $x = 8$ led to equation (2.3) on page 42, now slightly rewritten,

$$S_n = 31 \Big/ \frac{s^{5/3} - 1}{s - 1}, \quad s = 8^{1/n}.$$

Here S_n is the sum of the inner box-areas for n boxes, and the question was to what value S_n tends as the number n of boxes grows.

As n grows, s tends to 1, and so the area calculation is reduced to the *derivative* calculation of section 2.4. The normalized case of that calculation says that

$$\text{the limiting value of } \frac{s^{5/3}-1}{s-1} \text{ as } s \text{ tends to } 1 \text{ is } 5/3,$$

and so now we can finish the integration more quickly than we did in section 2.3,

$$\text{the limiting value of } S_n \text{ is } 31\big/(5/3) = 18.6.$$

Indeed, a rereading of section 2.4 and then section 2.3 from equation (2.3) to the end shows that the general derivative calculation encompasses the end-calculation of the integral as a special case.

On the face of things, the original integration problem is unrelated to any derivative, and yet the calculation reduced to a derivative: not the derivative of the original power function $f_{2/3}$, but of a different power function $f_{5/3}$ instead. Computing the derivative thus enabled us to compute the integral. A result called the **Fundamental Theorem of Calculus** will tell us that this was no fluke. Integral-values will come from derivative-values under a wide range of circumstances.

2.5 Integration of the Rational Power Function

The solution of a slightly more general integration problem than the one in section 2.3 should be understandable now. The only new issue is that the left endpoint 1, the right endpoint 8, and the power 2/3 will become general symbols a, b, and α. Thus the problem is: *Let a and b be real numbers with $0 < a < b$, and let $\alpha \neq -1$ be a rational number. Find the area under the graph of the function*

$$f_\alpha(x) = x^\alpha$$

from $x = a$ to $x = b$. If $\alpha > 0$ then f_α is strictly increasing, while if $\alpha < 0$ then f_α is strictly decreasing, but this will turn out to be irrelevant. The odd-seeming restriction that $\alpha \neq -1$ will emerge naturally from the pending calculation.

2.5.1 The Normalized Case

First consider the case where the left endpoint is still 1 and the right endpoint is b where $b > 1$. As before, let n be the number of boxes, and let $s = b^{1/n}$, i.e., s is the positive number such that $s^n = b$. As shown in exercise 2.3.2, $s - 1$ is small. The points of the relevant geometric partition are again

$$x_j = s^j \quad \text{for } j = 0, \dots, n,$$

the intervals determined by the geometric partition are

$$I_j = \text{the } x\text{-axis from } x_j \text{ to } x_{j+1} \quad \text{for } j = 0, \dots, n-1,$$

and their widths are

$$\Delta x_j = (s-1)s^j \quad \text{for } j = 0, \dots, n-1.$$

The base of the jth box is Δx_j. The height of the jth box is the value of the function f_α over the left endpoint of the jth interval,

$$f_\alpha(x_j) = x_j^\alpha = (s^j)^\alpha \quad \text{for } j = 0, \dots, n-1.$$

Thus the area of the jth box is

$$\Delta x_j \cdot f_\alpha(x_j) = (s-1)s^j(s^j)^\alpha = (s-1)(s^{\alpha+1})^j \quad \text{for } j = 0, \dots, n-1.$$

The sum of the box-areas is consequently

$$S_n = (s-1) \cdot \left[1 + (s^{\alpha+1}) + (s^{\alpha+1})^2 + \cdots + (s^{\alpha+1})^{n-1}\right].$$

Because $\alpha \neq -1$, the ratio $s^{\alpha+1}$ in the geometric sum is not 1, and so the finite geometric sum formula applies,

$$\begin{aligned} S_n &= (s-1) \cdot \left[1 + (s^{\alpha+1}) + (s^{\alpha+1})^2 + \cdots + (s^{\alpha+1})^{n-1}\right] \\ &= (s-1) \cdot \frac{(s^{\alpha+1})^n - 1}{s^{\alpha+1} - 1} \\ &= ((s^n)^{\alpha+1} - 1) \cdot \frac{s-1}{s^{\alpha+1} - 1} \\ &= (b^{\alpha+1} - 1) \Big/ \frac{s^{\alpha+1} - 1}{s-1}, \quad s = b^{1/n}. \end{aligned}$$

The derivative calculation in section 2.4 at $x = 1$ shows that therefore:

$$\text{The limiting value of } S_n \text{ as } n \text{ gets large is } \frac{b^{\alpha+1} - 1}{\alpha+1}.$$

We may instead use the right endpoints of the intervals to determine the box-heights, and this gives the same answer with no need to repeat the calculation, as follows. Now the box-heights are

$$f(x_{j+1}) = f(s^{j+1}) = (s^{j+1})^\alpha = s^\alpha(s^j)^\alpha = s^\alpha f(s^j) = s^\alpha f(x_j),$$

which is to say that each right endpoint box-area is simply the corresponding left endpoint box-area multiplied by s^α. Consequently the nth sum of right endpoint box-areas is this multiple of the nth sum of left endpoint box-areas,

$$T_n = s^\alpha S_n.$$

And because s^α tends to 1 (exercise 2.5.1), T_n tends to the same limiting value as S_n as the number of boxes grows.

If $\alpha > 0$, so that f_α is strictly increasing, then the boxes whose areas sum to S_n lie beneath the graph of f_α from 1 to b, and so the values S_n are all less than the area under the graph; similarly the values T_n are all greater than the area if $\alpha > 0$. And if $\alpha < 0$, so that f_α is strictly decreasing, then conversely. In either case, the common value tended to by S_n and T_n must be the area trapped between them. Summarizing, for any rational number $\alpha \neq -1$ and any real number $b > 1$,

The area under the graph of $f_\alpha(x) = x^\alpha$ from $x = 1$ to $x = b$ is $\dfrac{b^{\alpha+1} - 1}{\alpha + 1}$.

And in more mathematical notation,

$$\int_1^b f_\alpha = \frac{b^{\alpha+1} - 1}{\alpha + 1}, \quad \alpha \in \mathcal{Q},\ \alpha \neq -1,\ b > 1. \tag{2.4}$$

The end of the integration argument, as presented here, has improved over its prior incarnations in sections 1.2 (quadrature of the parabola) and 2.3 (integration of $f_{2/3}$ from 1 to 8). In those sections, the argument was that as n grows, the triangles fill the parabolic region, or the boxes fill the region under the power function's graph. Now the argument is less reliant on geometry and more on numbers: as n grows, box-area sums too small to be the desired area and box-area sums too large to be the desired area tend to a common value, and so this value must be the desired area. This point is important. So far we have been using the terms *area* and *integral* roughly as synonyms, but a better approximation to the right idea is that:

> *An integral is an area that is the common limiting value of box-area sums that are at most big enough and box-area sums that are at least big enough.*

This language will be made quantitative at the end of the next chapter.

Exercises

2.5.1. Let $b > 1$ be a real number. Let n be a positive integer and let $s = b^{1/n}$. Let $\alpha = p/q$ be a rational number, with p an integer and q a positive integer. This exercise shows that s^α tends to 1 as n grows.

(a) Explain why $s^\alpha = \tilde{b}^{1/n}$ where $\tilde{b} = b^\alpha$.

(b) Explain why exercise 2.3.2 (page 39) now completes the argument.

2.5.2. What happens in the calculation of $\int_1^b f_\alpha$ when $\alpha = 0$?

2.5.2 The General Case

The calculation so far has been normalized in that its left endpoint is 1. To change the left endpoint to an arbitrary positive real number a, we first give a geometric argument using boxes to establish the following proposition.

Proposition 2.5.1 (Scaling Result for the Power Function). *Let b and c be real numbers with $1 \leq b$ and $c > 0$. Let α be any rational number, including the possibility $\alpha = -1$. Then*

$$\int_c^{bc} f_\alpha = c^{\alpha+1} \int_1^b f_\alpha.$$

See figure 2.5, in which $\alpha = 2/3$. The figure shows the scaled interval $[c, bc]$ lying entirely to the right of the original interval $[1, b]$, but in general this need not be the case: the scaled interval can also lie to the right of the original interval but with overlap, or to the left of the original interval but with overlap, or entirely to the left of the original interval (exercise 2.5.3).

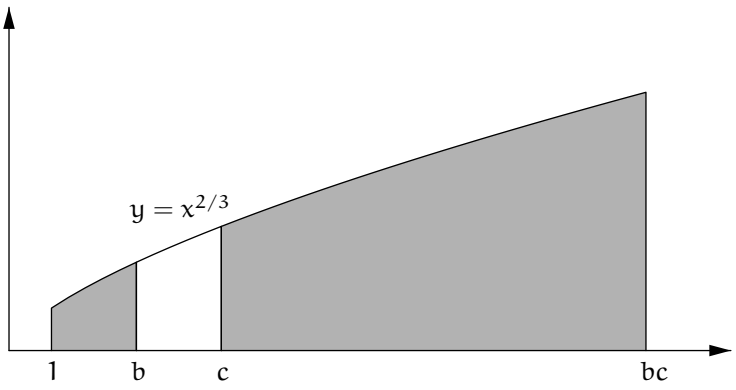


Figure 2.5. The right area is $c^{5/3}$ times the left one

In our present context of integrating the power function f_α we are assuming that $\alpha \neq -1$. The proposition is being presented as a self-contained entity because we will refer to it again in chapter 5 for the case $\alpha = -1$, to which it applies as well.

To establish the proposition, let $b \geq 1$, let n be a positive integer, let $s = b^{1/n}$, and recall our machinery from the normalized calculation—the partition points x_j of $[1, b]$, the interval-widths Δx_j, the heights $f_\alpha(x_j)$ over the

left endpoints, and the heights $f_\alpha(x_{j+1})$ over the right endpoints, culminating in the box-area sums S_n and $T_n = s^\alpha S_n$. The left part of figure 2.6 illustrates the boxes whose areas sum to S_n in a case where $\alpha > 0$. Here S_n is less than the true area under the graph from 1 to b, while T_n is greater than the true area. If instead $\alpha < 0$ then S_n is greater than the true area and T_n is less than it. So far this discussion has only repeated ideas from the normalized calculation.

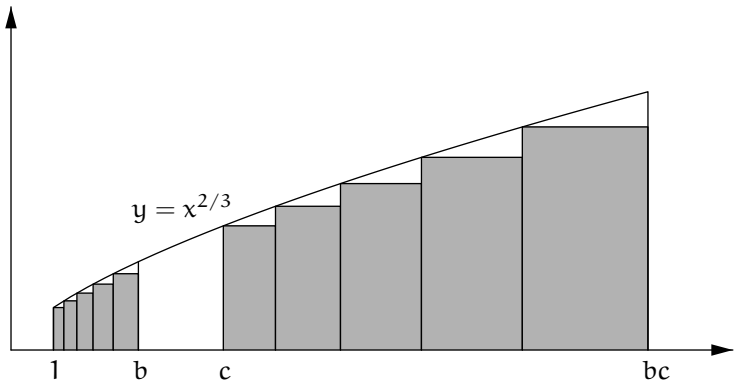


Figure 2.6. The right box-area sum is $c^{5/3}$ times the left one

Now let $c > 0$. Scale the partition of $[1, b]$ by c to get a partition of $[c, bc]$,

$$\tilde{x}_j = cx_j \quad \text{for } j = 0, \ldots, n.$$

These partition points determine subintervals of width

$$\Delta\tilde{x}_j = c\,\Delta x_j,$$

heights over left endpoints

$$f_\alpha(\tilde{x}_j) = f_\alpha(cx_j) = c^\alpha f_\alpha(x_j),$$

and heights over right endpoints

$$f_\alpha(\tilde{x}_{j+1}) = f_\alpha(cx_{j+1}) = c^\alpha f_\alpha(x_{j+1}).$$

The box-areas are now $\Delta\tilde{x}_j \cdot f_\alpha(\tilde{x}_j) = c^{\alpha+1}\Delta x_j \cdot f_\alpha(x_j)$ and $\Delta\tilde{x}_j \cdot f_\alpha(\tilde{x}_{j+1}) = c^{\alpha+1}\Delta x_j \cdot f_\alpha(x_{j+1})$, and the box-area sums are now

$$\widetilde{S}_n = c^{\alpha+1}S_n \quad \text{and} \quad \widetilde{T}_n = c^{\alpha+1}T_n. \tag{2.5}$$

The new boxes whose areas sum to $\widetilde{S}_n$ are shown in the right part of figure 2.6. If $\alpha > 0$ then $\widetilde{S}_n$ is too small to be the true area under the graph of f_α from c to bc, and so on, just as before.

Let n grow large. Then on the one hand, making no reference to the explicit formulas for S_n and T_n:

> *The fact that S_n and T_n trap the area under the graph of f_α from 1 to b between them, and the facts that $T_n = s^\alpha S_n$ and s^α tends to 1, combine to show that S_n and T_n tend to the same limiting value, that value being the area.*

(For the reader who is justifiably uneasy with the argument just displayed in italics: it will be shored up at the end of the next chapter.) Because S_n and T_n tend to $\int_1^b f_\alpha$, (2.5) shows that $\widetilde{S}_n$ and $\widetilde{T}_n$ both tend to $c^{\alpha+1}\int_1^b f_\alpha$. But on the other hand, the geometry underlying $\widetilde{S}_n$ and $\widetilde{T}_n$ shows that the common value that they tend to must be the area trapped between them, $\int_c^{bc} f_\alpha$. Consequently, Proposition 2.5.1 holds:

$$\text{If } b \geq 1 \text{ and } c > 0 \text{ then } \int_c^{bc} f_\alpha = c^{\alpha+1}\int_1^b f_\alpha. \tag{2.6}$$

This completes the proof of the proposition. To review, the crux is that for the power function f_α, scaling a box horizontally by the factor c scales it vertically by c^α, giving an area-scaling factor of $c^{\alpha+1}$. This observation doesn't depend on the boxes arising from the geometric partition in particular, or on our being able to put the box-area sum into a tidy form. Instead, it depended on the admittedly hand-waving argument displayed in italics above. Again: we will return to that argument in the following chapter.

With Proposition 2.5.1 in hand, we can complete the integration of the power function. Again let $\alpha \neq -1$ be a rational number, and let $0 < a \leq b$. Then

$$\begin{aligned}
\int_a^b f_\alpha &= \int_a^{(b/a)a} f_\alpha && \text{by basic algebra} \\
&= a^{\alpha+1}\int_1^{b/a} f_\alpha && \text{by (2.6) with } a \text{ for } c \text{ and } b/a \text{ for } b \\
&= a^{\alpha+1}\,\frac{(b/a)^{\alpha+1}-1}{\alpha+1} && \text{by the normalized result (2.4)} \\
&= \frac{b^{\alpha+1}-a^{\alpha+1}}{\alpha+1} && \text{by algebra.}
\end{aligned}$$

That is,

$$\boxed{\int_a^b f_\alpha = \frac{b^{\alpha+1}-a^{\alpha+1}}{\alpha+1}, \quad \alpha \in \mathcal{Q},\ \alpha \neq -1,\ 0 < a \leq b.} \tag{2.7}$$

This completes the integration of the power function, excluding the special case of the reciprocal function f_{-1}. We will return to the integral of the reciprocal function in chapter 5.

Exercise

2.5.3. Given a number $b > 1$, give conditions on the positive number c such that

(a) $[c, bc]$ lies entirely to the right of $[1, b]$,
(b) $[c, bc]$ lies partially to the right of $[1, b]$ but with overlap,
(c) $[c, bc]$ lies partially to the left of $[1, b]$ but with overlap,
(d) $[c, bc]$ lies entirely to the left of $[1, b]$.

2.6 Summary

Having discussed the parabola very informally in the previous chapter and then the power function somewhat informally in this chapter, we now have raised enough questions to make a closer discussion of calculus necessary, and we now have worked through enough examples for the discussion to be comprehensible.

3

Sequence Limits and the Integral

A *limit* is a value that is tended to, whether it is attained or not. The unexplained notion of *tends to*, which has served as the workhorse for finishing off arguments in the previous two chapters, needs to be made quantitative. A description of how to do so emerged in the nineteenth century, long after Newton and Leibniz. It has since stimulated generations of calculus students, for better or for worse.

The definition of limit should be enlightening once the student understands it by parsing it, seeing it used, and learning to use it. But because the definition involves two diagnostic quantities that interact in a delicate way, and because working with the definition requires skill with symbol-manipulation and with language in concert with geometric intuition, coming to terms with it can take some time, a resource in scant supply during a calculus course. Hence:

> *The student is encouraged to engage with the limit arguments in this chapter lightly and to taste.*

Said engagement should give some sense of why the definition of limit captures the right idea, and some sense of what form an argument using the definition should take. But it is much more important to understand the results—and their uses—than to understand every detail of every argument that the results hold.

Section 3.1 discusses preliminary matters: sets, functions, and sequences. A *sequence* is a special kind of function, naturally viewed as a list of data, such as the lists of successive area-approximations that we generated during the course of integrating the power function. Section 3.2 defines the *limit of a sequence*, the value to which the data are tending. With the limit of a sequence defined, we can prove basic results about sequence limits, and we can make inferences about unknown sequence limits in terms of known ones,

leading to more results. Section 3.3 uses the results of this chapter to redo some of the limit calculations from chapters 1 and 2 more satisfyingly. Once we see the methods work, it becomes clear that their scope extends beyond the particular instance of the power function. A precise and manipulable definition of the integral becomes natural to write down, and results about integration become natural to prove. As a payoff on our investment in definitions, the language suddenly, unexpectedly, carries us farther in clarity and results, with no more computational effort. The serendipitous economy of ideas is pleasing.

Still, the definition of limit given in this chapter is neither the alpha nor the omega of the idea. Calculus flourished for centuries before this definition evolved. We should not be so arrogant as to presume that the great mathematicians of the seventeenth and eighteenth centuries couldn't understand their work without the nineteenth century definition of limit. Indeed, a 1980 text called *Calculus Unlimited* by Jerrold Marsden and Alan Weinstein develops the subject matter with no recourse to limits at all. Nor should we believe that the nineteenth century definition of limit is the end of the story. It prominently features the phrase *there exists*, whose meaning is still in contention. Does something exist only if we know an algorithm to compute it, or does it exist if its nonexistence seems untenable, i.e., does it exist abstractly as compared to computationally? Do the two different notions of existence lead to different bodies of mathematics? Sadly, a traditional first calculus course has no time for these questions, but the student should be aware that they are serious ones. A 2001 text called *Computable Calculus* by Oliver Aberth develops calculus using only computability.

In keeping with this chapter's attempt to be more technical mathematically than chapters 1 and 2, the writing conventions here will be different. Definitions and propositions will be numbered, and proofs will be delineated. The change in style is not formalism for formalism's sake, but an attempt to lay the ideas out clearly.

3.1 Sets, Functions, and Sequences

3.1.1 Sets

As discussed in section 2.2, a set is a collection of elements. A set is often described by listing its elements in curly braces,

$$S = \{\text{elements of } S\}.$$

The order in which the elements are listed is irrelevant, as are repeat listings of the same element. Thus

$$\{2,3\} = \{3,2\} = \{2,2,3\} = \text{the set with elements 2 and 3.}$$

Some ubiquitous sets in mathematics are

$$\begin{aligned}
\mathcal{Z} &= \{\text{integers}\} = \{0, 1, -1, 2, -2, 3, \dots\},\\
\mathcal{Z}_{\geq 0} &= \{\text{natural numbers}\} = \{0, 1, 2, 3, \dots\},\\
\mathcal{Z}_{\geq 1} &= \{\text{positive integers}\} = \{1, 2, 3, \dots\},\\
\mathcal{Z}_{\leq -1} &= \{\text{negative integers}\} = \{-1, -2, -3, \dots\},\\
\mathcal{Q} &= \{\text{rational numbers}\},\\
\mathcal{R} &= \{\text{real numbers}\},\\
\mathcal{R}_{\geq 0} &= \{\text{nonnegative real numbers}\},\\
\mathcal{R}_{>0} &= \{\text{positive real numbers}\},\\
\mathcal{R}^2 &= \{\text{points in the plane}\}.\\
\emptyset &= \text{the empty set} = \text{the set containing no elements.}
\end{aligned}$$

The left curly brace reads *the set of* or *the set*, so that, for example, the first line in the previous display reads altogether,

$$\mathcal{Z} \textit{ is the set of integers, the set } 0,\ 1,\ -1,\ 2,\ -2,\ 3,\ \dots.$$

The empty set is not 0, nor is it $\{0\}$. In curly braces notation,

$$\emptyset = \{\}.$$

Perhaps the reason that the empty set is often confused with 0 is that the number of elements in the empty set is 0. However, a set is not the same thing as the number of its elements.

Sets are often defined by conditions. In this context, a colon ":" reads *such that*. So, for example, the notation

$$\mathcal{R}^2 = \{(x, y) : x, y \in \mathcal{R}\}.$$

reads

$\mathcal{R}^2$ *is the set of ordered pairs* (x, y) *such that* x *and* y *are real numbers.*

Ordered pair means a pair with one of its elements designated as the first of the two. Because $\mathcal{R}^2$ was defined a moment ago as the set of points in the plane, the last two displays give the appearance of a redefinition. However, the reader is assumed to be familiar with the representation of points in the plane as ordered pairs of numbers, so that the last two displays only rephrase the definition of $\mathcal{R}^2$ rather than revise it. From now on, the terms *point in the plane* and *ordered pair of real numbers* will be taken as synonyms.

(An *unordered pair* of numbers would be, for instance, the set $\{2,3\}$, which is also $\{3,2\}$. That is, viewing the pair of numbers 2 and 3 as a set does not connote that one of them is innately the first. By contrast, the notation $(2,3)$ expressly means the number-pair with 2 in its first position and 3 in its second. It is in fact possible to formulate the notion of ordered pair in terms of set theory rather than as a new primitive. Consider the sets $\{\{2,3\},2\}$ and $\{\{2,3\},3\}$. Each of these sets has another set as one of its elements and a number its other element. Both of them can be understood to specify the unordered pair $\{2,3\}$ and then to specify in addition which of 2 and 3 should be taken as the first element of the corresponding ordered pair. Similarly, ordered triples such as $(2,3,4)$, ordered quadruples, and so on can all be defined purely in terms of set theory, but once this is done, continuing to drag the resulting cumbersome notation around is silly.)

Sets defined by conditions also arise from the fact that analytic geometry describes geometrical objects by equalities and inequalities. The reader is assumed to be familiar with such representations. So, for example, the set

$$R = \{(x,y) \in \mathcal{R}^2 : 1 \le x \le 8,\ 0 \le y \le x^{2/3}\}$$

is the region between the x-axis and the graph of the power function $f_{2/3}$ from $x = 1$ to $x = 8$, depicted back in figure 2.1 on page 33. The last comma in the previous display is read *and*, and so the display reads altogether,

> R *is the set of points* (x,y) *in the plane such that* $1 \le x \le 8$ *and* $0 \le y \le x^{2/3}$.

Here it is understood that "$1 \le x \le 8$" means that $1 \le x$ and $x \le 8$, and similarly for any concatenation of inequalities. Inequalities should be concatenated only if they point in the same direction, i.e., something like "$1 \le x > y$" is poor style that can lead to calculation errors.

Another type of set that is defined by conditions is the *interval*. There are nine types of interval (!):

Definition 3.1.1 (Intervals). *Let* a *and* b *be real numbers with* $a \le b$. *Define the following subsets of* $\mathcal{R}$*:*

$$\begin{aligned}
(a,b) &= \{x \in \mathcal{R} : a < x < b\},\\
(a,b] &= \{x \in \mathcal{R} : a < x \le b\},\\
[a,b) &= \{x \in \mathcal{R} : a \le x < b\},\\
[a,b] &= \{x \in \mathcal{R} : a \le x \le b\},\\
(a,\infty) &= \{x \in \mathcal{R} : a < x\},\\
[a,\infty) &= \{x \in \mathcal{R} : a \le x\},\\
(-\infty,b) &= \{x \in \mathcal{R} : x < b\},\\
(-\infty,b] &= \{x \in \mathcal{R} : x \le b\},\\
(-\infty,\infty) &= \mathcal{R}.
\end{aligned}$$

A subset of $\mathcal{R}$ is called an **interval** *if it is a set of one of these nine types.*

So, for example, (a,b) is the set of real numbers x such that $a < x < b$. Note that $[a,a] = \{a\}$ and $(a,a) = \emptyset$, showing that a set consisting of just one point is an interval and so is the empty set.

The use of the symbols "∞" and "$-\infty$" in the notation for some types of interval is traditional, and it does uniformize the notations for the nine types. But it is pedagogically regrettable because, as already mentioned, ∞ and $-\infty$ are not real numbers. However, note that they occur only on the *left* sides of the above equalities; that is, truly they are nothing but shorthand notation to describe the sets on the right sides of the equalities, where they do not appear. And in the notation, they always occur adjacent to a parenthesis, never a square bracket, so even the shorthand notation does not suggest that an interval ever contains ∞ or $-\infty$. In any case, alternative notations for the fifth through eighth types of interval are $\mathcal{R}_{>a}$, $\mathcal{R}_{\ge a}$, $\mathcal{R}_{<b}$, and $\mathcal{R}_{\le b}$, while the ninth type of interval really needs no notation because it is simply $\mathcal{R}$.

Definition 3.1.2 (Endpoints). *If* I *is a nonempty interval of one of the first four types in Definition 3.1.1 then its* **endpoints** *are* a *and* b. *If* I *is an interval of one of the next two types then its one endpoint is* a. *If* I *is an interval of one of the next two types then its one endpoint is* b. *The interval* $(-\infty,\infty)$ *has no endpoints. The empty interval has no endpoints.*

Definition 3.1.3 (Open and Closed Intervals). *An interval is* **closed** *if it contains all of its endpoints. An interval is* **open** *if it contains none of its endpoints.*

Note that the mathematical usages of *open* and *closed* need not be exclusive or exhaustive. That is, an interval can conceivably be open, closed, neither, or both. Exercise 3.1.1 is to determine which of the nine types of interval are open and/or closed.

The wording of the previous definition (and of many definitions to come in these notes) deserves a comment. The reader who is parsing grammar with exceptional care could raise the following point:

> *The definition says that if an interval contains all of its endpoints then it is closed.*

But:

> *The definition does* **not** *say that if an interval is closed then it contains all of its endpoints.*

So, if we are told that an interval is closed, are we being told anything at all? Yes, we are. Within the definition, we can not yet talk about an interval being closed implying anything until we have first assigned meaning to the notion of an interval being closed. But it is a convention of mathematics that as soon as the meaning is assigned, the *if* tacitly evolves into an *if and only if*. An interval that contains all of its endpoints is closed, and a closed interval contains all of its endpoints. That is:

> Once the definition is stated, saying that an interval is closed is *synonymous* with saying that it contains all of its endpoints.

A similar discussion applies to every further definition in these notes that takes the form "A if B": once we are done reading the definition, to say that A holds is to say that B holds, and conversely.

Two sets whose elements are not numbers have already figured tacitly through these notes. The first is

$$\mathcal{P} = \{\text{polygons in the plane}\}.$$

Elements of $\mathcal{P}$ are not numbers or individual planar points, but instead they are planar regions. The reader should look back at figure 1.9, figure 1.10, figure 2.2, and figure 2.4 (pages 9, 10, 37, and 44) to see that these figures show elements of $\mathcal{P}$ and that these elements are highly relevant to integration. As for the second particular set to be aware of, define a subset of the plane to be **bounded** if some box contains it, and define

$$\mathcal{B} = \{\text{bounded subsets of the plane}\}.$$

So elements of $\mathcal{B}$ are planar regions too. The reader should look at figure 1.7 and figure 2.1 (pages 7 and 33) to see that these figures show nonpolygonal elements of $\mathcal{B}$, regions whose areas we wanted to find. No figure in these notes can accurately depict a planar set that is not an element of $\mathcal{B}$ because the entire figure will be contained in a box: the page.

Exercise

3.1.1. Let a and b be real numbers.

(a) Assume that $a < b$. For each of the first four types of interval described in Definition 3.1.1, is the relevant interval of that type for such a and b open? Is it closed?

(b) Now drop the assumption that $a < b$. For each of the next five types of interval described in Definition 3.1.1, is the relevant interval of that type open? Is it closed?

(c) Is the one-point interval $[a, a]$ open? Is it closed?

(d) Is the empty interval open? Is it closed?

3.1.2 Functions

The reader is presumed to have some experience with functions. The basic idea is that a function receives inputs and produces outputs. We will notate functions as follows:

$$f : A \longrightarrow B.$$

Here A and B are sets, and f is a rule or a process that assigns to each element of A an element of B. The set A is called the **domain** of the function, and the set B is a set called the **codomain** of the function. Thus the domain consists of all legal inputs to f, and the codomain consists of all potential outputs of f. Strictly speaking, the function consists of all three data, A, B, and f, but we often abbreviate it to f. The notation $f : A \longrightarrow B$ reads f *is a function from* A *to* B, or f *maps* A *to* B, or f *from* A *to* B, or various other phrasings along these lines.

For example, consider the function

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \qquad f(x) = x^2. \tag{3.1}$$

This is the squaring function that takes all real numbers as its inputs and is understood to produce real numbers as its outputs. Note, however, that because the square of any real number is nonnegative, not all points of the codomain $\mathcal{R}$ are actual outputs of f. This is an example of what was meant a moment ago in describing the codomain as all *potential* outputs of f: all squares are real numbers, but in fact not all real numbers are squares. Because a function strictly depends on its domain, its codomain, and its rule, the squaring function that takes all real numbers as its inputs and is understood to produce nonnegative real numbers as its outputs,

$$f : \mathcal{R} \longrightarrow \mathcal{R}_{\geq 0}, \qquad f(x) = x^2,$$

is not quite the same object as the squaring function in (3.1). This distinction is admittedly pedantic, but on the other hand, the squaring function from the integers,

$$f : \mathcal{Z} \longrightarrow \mathcal{R}, \qquad f(x) = x^2$$

is more compellingly different from (3.1) because it has a different domain, i.e., a different set of legal inputs. Thus although $f(2) = 4$ and $f(-2) = 4$ as for the squaring function (3.1) with domain $\mathcal{R}$, now $f(1/2)$ is not defined because $1/2 \notin \mathcal{Z}$.

The **range** of a function is the set of its actual outputs, a subset of the codomain,

$$f(A) = \{f(x) : x \in A\} = \{\text{outputs of } f\}.$$

Any function can have its codomain pruned down to its range and be rewritten

$$f : A \longrightarrow f(A).$$

Under many circumstances we don't bother doing this when specifying a function, because the purpose of the codomain is only to give us *some* sense of where the outputs of f are to be found. In fact, when the description of a function's domain and rule make its codomain clear, the codomain can go unmentioned. For example, saying "Let $f(x) = x^2$ for $x \in \mathcal{R}$" describes the squaring function on real inputs, whose outputs are understood to be real numbers or nonnegative real numbers, it doesn't particularly matter. (But saying only "Let $f(x) = x^2$" is too vague, unless the context has clearly established that the intent here is for x to vary through the real numbers.) On the other hand, one circumstance where it is worth tightening the codomain down to the range is when we want to follow f by a second function,

$$g : B \longrightarrow C,$$

i.e., when we want the outputs of f to serve as inputs to g. Doing so is sensible only when the range $f(A)$ of f is a subset of the domain B of g. For example, if we want the squaring function $f = f_2$ and the square root function $g = f_{1/2}$ to undo the effects of one another, we need to take care to specify their domains and codomains compatibly,

$$f : \mathcal{R}_{\geq 0} \longrightarrow \mathcal{R}_{\geq 0} \quad \text{and} \quad g : \mathcal{R}_{\geq 0} \longrightarrow \mathcal{R}_{\geq 0}.$$

A function strictly depends on its domain, its codomain, and its rule, but not on its typography. For example, consider the two functions

$$f : \mathcal{Q} \longrightarrow \mathcal{Q}, \qquad f(x) = x^2 - 1$$

and

$$g : \mathcal{Q} \longrightarrow \mathcal{Q}, \qquad g(y) = (y+1)(y-1).$$

These two functions are the same function even though their shared rule has been given two names (f and g) and described by different formulas with different variables ($x^2 - 1$ and $(y+1)(y-1)$).

A function is not the same thing as its graph. Let

$$f : A \longrightarrow B$$

be a function. The **graph of** f is a set of ordered pairs,

$$\text{graph}(f) = \{(x, y) : x \in A,\ y \in B,\ y = f(x)\} = \{(x, f(x)) : x \in A\}.$$

If the domain and codomain of f are subsets of $\mathcal{R}$, then the graph of f can be identified with a subset of the plane. The language here is entirely consistent with the description in chapter 1 of the parabola as the graph of the squaring function $f(x) = x^2$, other than the fact that the chapter 1 description didn't bother making explicit mention of the domain and codomain of f. Again, *the graph of a function is not the function itself*. Students often refer to a graph as a function, and this is understandable because the graph describes the function visually, but for purposes of reasoning the distinction is worth retaining.

An **algebraic function** is any function that can be built up from a finite succession of additions, subtractions, multiplications, divisions, and roots. Thus a typical algebraic function is

$$f(x) = \left(\frac{(x^2+1)^{1/2} - x}{x+1} \right)^{1/3}.$$

The tacit understanding here is that the domain of f is the set of real numbers x for which the formula is sensible. That is, the domain of f is the set of real numbers x which when substituted into the formula do not lead to any square roots or cube roots of nonpositive numbers, or to a divide by 0. For the function f in the display, the domain is all real numbers $x > -1$.

Not all functions are algebraic. A nonalgebraic function whose domain is a suitable subset of $\mathcal{R}$ (the meaning of *suitable* isn't worth going into in detail right now) and whose codomain is $\mathcal{R}$ is called **transcendental**. We have not yet seen any transcendental functions in these notes, but some examples for the reader who may have seen them elsewhere are the logarithm, the exponential function, and the trigonometric functions.

The notion of a *rule* as part of a function calls for some explanation. Just as a function is not a graph, *a function is not a formula*. The term function (*functio*) was introduced into mathematics by Leibniz, and its meaning has changed ever since. During the seventeenth century the ideas of function and

curve were usually thought of as being the same, and a curve was often thought of as the path of a moving point. By the eighteenth century the idea of function was associated with *analytic expression*. Leonard Euler (1707–1783) gave the following definition:

> A function of a variable quantity is an analytic expression composed in any way whatsoever of the variable quantity and numbers or constant quantities.
> Hence every analytic expression, in which all component quantities except the variable z are constants, will be a function of that z; Thus $a + 3z$; $az - 4z^2$; $az + b\sqrt{a^2 - z^2}$; c^z; etc. are functions of z

The use of the notation "$f(x)$" to represent the value of f at x was introduced by Euler in 1734. Our contemporary notion of a function as a *rule* is different from Euler's notion unless every analytic expression is understood it to produce output-values from input-values, and every rule or process that produces output-values from input-values is understood to have an analytic expression. If rules or processes are not the same thing as analytic expressions, then the next question is just what rules/processes are sensible. Must we be able to carry them out? What does *carry them out* mean?

Neither the inputs nor the outputs of a function need even be numerical. For examples of nonnumerical output, let $\mathcal{P}$ denote the set of all polygons in the plane, and let a and b be real numbers with $a < b$. The process in section 1.2 of starting with one triangle inscribed in the parabola with its left and right endpoints over a and b, then adding two more smaller triangles, then four more smaller-yet triangles, and so on, defines a function based on the original endpoints a and b, whose input is the generation-number and whose output is not a number at all, but rather is the corresponding polygonal amalgamation of triangles,

$$p_{a,b} : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{P}.$$

That is, $p_{a,b}(n)$ is the nth generation polygonal approximation of the region whose area we wanted to compute. Similarly, the process in section 2.3 of computing the area under the graph of the power function $f_{2/3}$ from $x = 1$ to $x = 8$ defines a function taking the number of boxes to the polygonal amalgamation of boxes,

$$p : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{P}.$$

That is, $p(n)$ is the polygon consisting of n boxes, shown in figure 2.2 for $n = 20$ (page 37).

For an example of nonnumerical input, let $\mathcal{B}$ denote the set of bounded subsets of the plane, introduced on page 60. We would like an *area-function*

$$\mathrm{Ar} : \mathcal{B} \longrightarrow \mathcal{R}_{\geq 0}$$

that assigns to each bounded subset of the plane a nonnegative real number to be considered its area. Certainly area-functions exist (we could simply assign areas in some random way), but the question is whether area-functions having good properties exist. (Going into detail about *good properties* would take us too far afield, but they are very basic: the area of a box is its base times its height, the area of two nonoverlapping sets is the sum of their areas, and so on.) Similarly, let $\mathcal{B}_3$ denote the set of all bounded subsets of 3-dimensional space. (A subset of space is bounded if it sits inside some 3-dimensional box.) A *volume-function*

$$\mathrm{Vol} : \mathcal{B}_3 \longrightarrow \mathcal{R}_{\geq 0}$$

would assign to each bounded subset of space a nonnegative real number to be considered its volume, and would have good properties. Remarkably, *area-functions exist but volume-functions do not.*

The invocation that area functions exist will ease our lives considerably. Let a and b be real numbers with $a \leq b$. Let M be a nonnegative real number. Consider a function

$$f : [a, b] \longrightarrow [0, M].$$

The region under the graph of f, a subset of the plane, is

$$R = \{(x, y) \in \mathcal{R}^2 : a \leq x \leq b,\ 0 \leq y \leq f(x)\},$$

and by our invocation, it *has* an area,

$$\mathrm{Ar}_a^b(f) = \mathrm{Ar}(R).$$

Indeed, there may be *more* than one plausible area if f is strange enough—too strange to draw or even to imagine visually, so this point is best not dwelled on—and if we switch our choice of area function, however we are "choosing" one in the first place. The reader would be thoroughly justified in objecting that the previous sentence amounts to speaking in tongues rather than mathematics, but the real point is that we are entering into a social contract: *The question of integrating* f *is not whether an area under its graph exists—it abstractly does, granting our invocation of area functions—but whether the area is a suitable limit of box-area sums.* The invocation of area is an expedient that lets us finesse the existence question. Yes, the existence question is important, but a one-semester calculus course has no time to address it, especially because, as mentioned at the beginning of the chapter, the mathematical meaning of *existence* is a live, arguable issue.

Exercise

3.1.2. Sketch the graphs of the following functions:

(a) $f(x) = (x-1)^2$ for all $x \in [0,4]$,
(b) $g(x) = (x-2)^2$ for all $x \in [-1,3]$,
(c) $h(x) = x^2 - 1$ for all $x \in [-2,2]$,
(d) $k(x) = x^2 - 2^2$ for all $x \in [-2,2]$.

3.1.3 Sequences

A *sequence* is a list of data. More specifically, a sequence is one datum per generation, where there is a starting generation and then an endless succession of generations thereafter. Formally, a sequence is a function whose domain is the positive integers,

$$f : \mathcal{Z}_{\geq 1} \longrightarrow S.$$

The domain $\mathcal{Z}_{\geq 1}$ is the set of generations. The codomain S is often a subset of the real numbers, but it need not always be. For instance, we recently mentioned the sequence of polygons arising from Archimedes's quadrature of the parabola in chapter 1, and the sequence of polygons arising from the integration of the power function in chapter 2, both sequences of the form

$$f : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{P}.$$

Any sequence

$$f : \mathcal{Z}_{\geq 1} \longrightarrow S$$

can be described by listing its outputs, consonantly with the idea of a sequence as a list,

$$(f(1), f(2), f(3), \ldots).$$

Sequences are usually written this way, with the domain and codomain tacit. Furthermore, sequences tend to have names such as s or x or a rather than f. And finally, to streamline the notation, outputs are denoted s_n rather than $s(n)$, or x_n, or a_n. Thus a typical sequence is written

$$(s_1,\ s_2,\ s_3,\ \ldots).$$

More briefly, we write

$$(s_n)_{n\geq 1}$$

or

$$(s_n)_{n=1}^{\infty}$$

even though (yet again) ∞ is not a number—here the notation is meant to convey that the terms of the sequence go on and on. Once the context is clear, even the notation

$$(s_n)$$

will do, so long as we understand what is happening: n is varying through $\mathcal{Z}_{\geq 1}$, and the sequence is a corresponding list of values s_n.

3.1.4 Previous Examples

In section 1.2, Archimedes's quadrature of the parabola led to the triangle-area sums

$$\begin{aligned} S_1 &= A_{\text{tri}}, \\ S_2 &= A_{\text{tri}}\,[1 + 1/4]\,, \\ S_3 &= A_{\text{tri}}\left[1 + 1/4 + (1/4)^2\right], \\ S_4 &= A_{\text{tri}}\left[1 + 1/4 + (1/4)^2 + (1/4)^3\right], \end{aligned}$$

and in general for $n \in \mathcal{Z}_{\geq 1}$,

$$S_n = A_{\text{tri}}\left[1 + 1/4 + (1/4)^2 + \cdots + (1/4)^{n-1}\right].$$

And Archimedes's evaluation of the finite geometric sum with ratio $r = 1/4$ gave a closed form (ellipsis-free) expression for the sequence entries, so that the sequence of triangle-area sums was in fact

$$(S_n) = \Big(A_{\text{tri}} \cdot (4/3)\big(1 - (1/4)^n\big)\Big)_{n\geq 1}. \tag{3.2}$$

In section 2.5, integrating the rational power function f_α (where $\alpha \neq -1$) from 1 to b gave rise to the sequence of box-area sums (see page 49)

$$(S_n) = \left((b^{\alpha+1} - 1)\Big/\frac{s_n^{\alpha+1} - 1}{s_n - 1}\right)_{n\geq 1} \quad \text{where } s_n = b^{1/n}. \tag{3.3}$$

In the previous two chapters, we made assertions about the limiting behaviors of sequences (3.2) and (3.3). Later in this chapter we will be able to substantiate the assertions, as well as other matters from the end of chapter 2.

3.2 The Limit of a Real Sequence

3.2.1 Absolute Value and Distance

To describe quantitatively the idea of two real numbers being near each other, regardless of which is the larger, we first describe the idea of one real number being near 0, regardless of whether it is positive or negative. The definition innately must be casewise:

Definition 3.2.1 (Absolute Value). *The* **absolute value** *function is*

$$|\ |: \mathcal{R} \longrightarrow \mathcal{R}_{\geq 0}, \qquad |x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x < 0. \end{cases}$$

So, for example, $|5| = 5$ and $|-1/10| = 1/10$. A number is near 0 if its absolute value is small. It is worth pausing to convince oneself that indeed the casewise formula for the absolute value function always yields a nonnegative real value, so that designating the codomain to be $\mathcal{R}_{\geq 0}$ makes sense.

Cases are a nuisance to drag around, and so our short-term program is to use the casewise definition of the absolute value to establish a collection of absolute value properties that no longer make direct reference to cases. Once that is done, absolute values can be manipulated by using the properties with no further reference to the underlying cases, and indeed, with no further thought of them.

Proposition 3.2.2 (Basic Absolute Value Properties). *Let x and y be real numbers. Then*

(1) $|x| = 0$ *if and only if* $x = 0$.
(2) $-|x| \leq x \leq |x|$.
(3) $|xy| = |x| \cdot |y|$. *In particular,* $|-x| = |x|$ *because* $-x = x \cdot (-1)$.
(4) *If* $y \neq 0$ *then* $\left|\frac{1}{y}\right| = \frac{1}{|y|}$, *and so by* (3), *also* $\left|\frac{x}{y}\right| = \frac{|x|}{|y|}$.

Proof. (Sketch.) For instance, to verify (2), note that

$$\text{if } x \geq 0 \text{ then } x = |x|, \text{ and so } -|x| = -x \leq 0 \leq x = |x|,$$

and

$$\text{if } x < 0 \text{ then } x = -|x|, \text{ and so } -|x| = x < 0 < -x = |x|.$$

Verifying the first statement in (3) requires checking four cases, because x and y can each be nonnegative or negative independently of the other. Four cases amount to one small nuisance, but as explained a moment ago, the point is that after they are checked once and only once, we never have to think about them again. The reader is encouraged to verify enough of Proposition 3.2.2 to convince himself or herself that the entire proposition can be verified in a similar fashion. □

(The symbol "□" at the end of the previous line denotes the end of a proof.)

Theorem 3.2.3 (Triangle Inequality). *For all real numbers x and y,*

$$|x + y| \leq |x| + |y|, \tag{3.4}$$
$$|x - y| \leq |x| + |y|, \tag{3.5}$$
$$\big||x| - |y|\big| \leq |x + y|, \tag{3.6}$$
$$\big||x| - |y|\big| \leq |x - y|. \tag{3.7}$$

The first inequality (3.4) of Theorem 3.2.3 is the *Basic Triangle Inequality*. The four inequalities can be gathered together as the statement that for all real numbers x and y,

$$\big||x|-|y|\big| \leq |x \pm y| \leq |x|+|y|. \tag{3.8}$$

The reader should beware that (3.8) does *not* say that $|x-y| \leq |x|-|y|$ in general, and the reader should further beware that even after one hears this and understands it in the abstract, a frequent calculation error is to write some specific version of the false inequality nonetheless.

Proof. For all x and y in $\mathcal{R}$ we have by Proposition 3.2.2 (2),

$$-|x| \leq x \leq |x| \quad \text{and} \quad -|y| \leq y \leq |y|,$$

and so adding the inequalities gives $-|x|-|y| \leq x+y \leq |x|+|y|$, or

$$-(|x|+|y|) \leq x+y \leq |x|+|y|.$$

If $x+y \geq 0$ then $|x+y| = x+y$, and so the right inequality in the previous display becomes $|x+y| \leq |x|+|y|$. If $x+y < 0$ then $|x+y| = -(x+y)$, i.e., $x+y = -|x+y|$, and hence the left inequality gives $-(|x|+|y|) \leq -|x+y|$. In either case we have the Basic Triangle Inequality (3.4),

$$|x+y| \leq |x|+|y|.$$

The other inequalities (3.5) through (3.7) are consequences of (3.4) and are left as an exercise. □

We introduce the symbol " $\iff$ " as shorthand for *if and only if*. That is, the symbol " $\iff$ " between two statements means that the statement to its left is true exactly when the statement to its right is true.

> **In your writing, do not use the symbol " $\iff$ " to mean anything other than *if and only if*. This, and nothing else, is its meaning.**

Again let $x \in \mathcal{R}$ be a real number, and let $p \in \mathcal{R}_{>0}$ be a positive real number. Then

$$|x| < p \iff -p < x < p$$

and

$$|x| \leq p \iff -p \leq x \leq p. \tag{3.9}$$

To establish (3.9), argue that if $x \geq 0$ then because $|x| = x$ and because the statement "$-p \leq x$" is true (because $-p < 0 \leq x$),

$$|x| \le p \iff x \le p \iff -p \le x \le p.$$

If $x < 0$ then because $|x| = -x$ and because "$x \le p$" is true (because $x < 0 < p$), and because multiplying each side of an inequality by -1 switches its direction,

$$|x| \le p \iff -x \le p \iff -p \le x \iff -p \le x \le p.$$

Thus (3.9) holds regardless of whether $x \ge 0$ or $x < 0$.

Proposition 3.2.4 (Relation Between Absolute Values and Intervals). *Let $a \in \mathcal{R}$ and let $p \in \mathcal{R}_{>0}$. Then for all $x \in \mathcal{R}$,*

$$|x - a| < p \iff a - p < x < a + p,$$

and

$$|x - a| \le p \iff a - p \le x \le a + p.$$

The two statements rephrase as assertions that certain sets are intervals,

$$\{x \in \mathcal{R} : |x - a| < p\} = (a - p, a + p)$$

and

$$\{x \in \mathcal{R} : |x - a| \le p\} = [a - p, a + p].$$

Proof. For the second statement of the proposition, use (3.9) and recall that adding the same quantity to both sides of an inequality preserves the inequality,

$$|x - a| \le p \iff -p \le x - a \le p \iff a - p \le x \le a + p.$$

The first statement of the proposition has virtually the same proof. And the third and fourth statements of the proposition are rephrasings of the first two. □

The geometric distance between two real numbers x and y on the number line is the absolute value of their difference, $|x - y|$. So, for example, Proposition 3.2.4 says that the set of numbers whose distance from a is smaller than p is the interval centered at a extending distance p in both directions, $(a - p, a + p)$. This is exactly as our visual intuition tells us that it should be, and it is easiest to remember by seeing the relevant picture in one's mind. But the fact that it follows readily from our definitions by analytic arguments sends a reassuring message about our methodology.

The following result sometimes provides the punchline of an argument. Its point is that to show that two quantities are equal we need only show that they lie arbitrarily close to each other.

Proposition 3.2.5 (Strong Approximation Lemma). *Let ℓ and ℓ' be real numbers. Suppose that*

$$|\ell' - \ell| < \varepsilon \quad \textit{for every positive number } \varepsilon.$$

Then $\ell' = \ell$.

Proof. Either $|\ell' - \ell|$ is positive or it is zero. But the given condition implies that

$$|\ell' - \ell| \neq \varepsilon \quad \text{for every positive number } \varepsilon.$$

So $|\ell' - \ell| = 0$. Consequently $\ell' - \ell = 0$, i.e., $\ell' = \ell$. □

Exercises

3.2.1. Prove inequalities (3.5) through (3.7) of Theorem 3.2.3. Prove them by showing that they are consequences of (3.4), not be repeating the effort of proving (3.4) three more times.

3.2.2. Let x and y be nonzero. In each of (3.4) through (3.7), under what conditions on the signs of x and y does equality hold?

3.2.3. Describe each of the four sets below in terms of intervals. A set may require more than one interval for its description. (You may do this problem by inspection.)

(a) $A_1 = \{x \in \mathcal{R} : |x - 1/2| < 3/2\}$,
(b) $A_2 = \{x \in \mathcal{R} : |x + 1/2| \leq 3/2\}$,
(c) $A_3 = \{x \in \mathcal{R} : |3/2 - x| < 1/2\}$,
(d) $A_4 = \{x \in \mathcal{R} : |3/2 + x| \geq 3/2\}$.

3.2.4. Sketch the graphs of the following functions from $\mathcal{R}$ to $\mathcal{R}$ defined by the following equations (no explanations are needed for this problem):

(a) $f_1(x) = |x|$,
(b) $f_2(x) = |x - 2|$,
(c) $f_3(x) = |x| - |x - 2|$,
(d) $f_4(x) = |x| + |x - 2|$,
(e) $f_5(x) = x^2 - 1$,
(f) $f_6(x) = |x^2 - 1|$,
(g) $f_7(x) = |x^2 - 1|^2$.

3.2.5. Let f_1 through f_7 be the functions described in the previous exercise. By looking at their graphs, express each of the following six sets in terms of intervals.

(a) $S_1 = \{x \in \mathcal{R} : f_1(x) < 1\}$,

(b) $S_2 = \{x \in \mathcal{R} : f_2(x) < 1\}$,
(c) $S_3 = \{x \in \mathcal{R} : f_3(x) < 1\}$,
(d) $S_4 = \{x \in \mathcal{R} : f_4(x) < 3\}$,
(e) $S_5 = \{x \in \mathcal{R} : f_5(x) < 3\}$,
(f) $S_6 = \{x \in \mathcal{R} : f_6(x) < 3\}$.

Also, let $S_7 = \{x \in \mathcal{R} : f_7(x) < 1/2\}$. Represent S_7 graphically on a number line.

3.2.2 The Archimedean Property of the Real Number System

Any positive real number, however large, is exceeded by some positive integer:

Proposition 3.2.6 (Archimedean Property of the Real Number System). *Let $x \in \mathcal{R}_{>0}$ be any positive real number. There exists a positive integer $N \in \mathcal{Z}_{\geq 1}$ such that $N > x$.*

The reader may feel that the Archimedean Property is self-evident and hardly deserves its own name. But in fact there are number systems other than the real number system (which, again, is not innately extant, much less unique or preferred among number systems, just because it is named *real*) in which the property does not hold. Indeed, early attempts at reasoning about calculus made reference to *infinitesimals*, quantities that we now think of as positive numbers so small that their reciprocals exceed all positive integers, this happening in a *hyper-real* number system that subsumes the reals. These ideas of *non-standard analysis* were made rigorous by Abraham Robinson only as recently as 1960. A freshman calculus text based on Robinson's infinitesimals, written by H. Jerome Keisler, is online at

`http://www.math.wisc.edu/~keisler/calc.html`

Here is an attempt to prove the Archimedean Property rather than assume it: Suppose that some positive real number x exceeds all the positive integers,

$$x > N \quad \text{for all } N \in \mathcal{Z}_{\geq 1}.$$

Then surely there is a *least* x at least as big as all the positive integers. Consider the positive real number $x - 1$. Because it is less than x, it is less than some positive integer, i.e.,

$$x - 1 < N \quad \text{for some } N \in \mathcal{Z}_{\geq 1}.$$

Consequently,

$$x < N + 1 \quad \text{for some } N \in \mathcal{Z}_{\geq 1}.$$

But $N + 1$ is again a positive integer, so that x is not at least as big as all positive integers after all. Thus the supposition that some positive real number exceeds all the positive integers must be false.

However, rather than prove the Archimedean Property, this argument shows only that it follows from any assumption about the real number system that makes valid the *Then surely there is a* least x... statement in the previous paragraph.

3.2.3 Definition of Sequence Limit

Recall that a sequence is a function whose domain is $\mathcal{Z}_{\geq 1}$,

$$s : \mathcal{Z}_{\geq 1} \longrightarrow S,$$

often viewed as a list of data,

$$(s_n) = (s_1,\ s_2,\ s_3,\ \ldots).$$

In particular, a **real sequence** is a sequence whose codomain is $\mathcal{R}$, i.e., a list of numerical values. From now on, all sequences in this chapter will be real sequences, and so usually they will simply be called *sequences* without bothering to say *real sequence* each time.

A real sequence *converges* to a limit ℓ if the terms of the sequence approach ℓ and stay near ℓ, as closely as desired, although they may or may not actually reach ℓ, and they may or may not stay at ℓ should they reach it. Figure 3.1 depicts convergence for a sequence viewed as a function,

$$s : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{R},$$

and figure 3.2 depicts the convergence of the same sequence viewed as a list of data,

$$(s_n) = (s_1,\ s_2,\ s_3,\ \ldots).$$

The following definition captures quantitatively and concisely the above-mentioned notion of *approach ℓ and stay near ℓ, as closely as desired, although they may or may not actually reach ℓ, and they may or may not stay at ℓ should they reach it.*

Definition 3.2.7 (Limit of a Real Sequence, Convergent Sequence, Divergent Sequence). *Let (s_n) be a real sequence, and let ℓ be a real number. We say that (s_n) is a* **convergent sequence** *with* **limit** ℓ *if the following condition holds.*

> For every positive real number $\varepsilon > 0$,
> there exists a positive integer N such that
> for all integers $n \geq N$, $|s_n - \ell| < \varepsilon$.

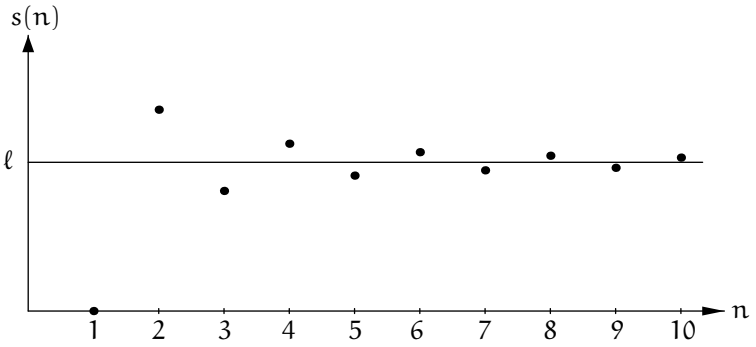


Figure 3.1. A convergent sequence, viewed as a function

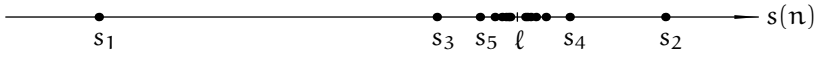


Figure 3.2. A convergent sequence, viewed as a list of data

When the condition holds, the fact that (s_n) *has limit* ℓ *is notated*

$$\lim_n (s_n) = \ell.$$

In this case we also say that (s_n) **converges to** ℓ. *If the sequence* (s_n) *does not converge then it is* **divergent**.

The typography "$\lim_n$" is short for "$\lim_{n\to\infty}$", but as usual we try to eschew the symbol ∞ because "infinity" is not part of the number system that we are using in our study of calculus.

The grammar of Definition 3.2.7 is sophisticated. Again, the idea of the definition is that a sequence (s_n) has limit ℓ if the terms of the sequence eventually get close to ℓ and stay close to ℓ. The numbers ε and N, and the interaction between them, are the mathematical machinery that together *quantify* the idea. To work successfully with Definition 3.2.7, one needs some sense of how the quantification indeed captures the idea, and one also needs some practice with the symbol-based language of the quantification. For this reason, the first few results that we will prove with the definition are meant to be simple and obvious-sounding: their point isn't to be earth-shattering, but to demonstrate what the definition says and how it works. The intent is

that as the examples accrue, the student will see that the definition encodes a natural idea in a way that is sensible, usable, and versatile. Nonetheless, every calculus teacher understands that for the student, coming to grips with one's first arguments with the definition of limit poses the double challenge of parsing the definition's grammar in general and isolating the key particular of each situation at hand. One gets better at this with time and experience.

A few words about mathematical proof may be useful here. Proofs in mathematics are not alienating formalisms, or at least they shouldn't be. The reader may have heard the maxim that *the exception proves the rule*. Because a mathematical proof is meant to establish a rule in all cases, with no exceptions, the maxim doesn't sound sensible in our context. But it is. The word *prove* is a variant of *probe*, and *rule* connotes a principle that holds often but not always; to say that *the exception probes the rule* is to say that knowing when a principle can break down informs us about the principle's scope—its extent and its limitations. In mathematics, we often prove a statement to the effect that if certain conditions A hold then some other condition B follows. Proving such a statement *teaches* us, because the argument will show us *how* conditions A lead to condition B. Condition B may well fail without conditions A in place—an exception that probes the rule that condition B holds. A mathematical proof is a persuasive explication of quantitative causality, with all the terms in that phrase negotiated socially with an eye to increasing our understanding and passing it on.

Definition 3.2.7 is illustrated in figures 3.3 and 3.4. In both figures the idea is that no matter how narrow the gray zone is, all but finitely many of the dots lie in it. A narrower gray zone may exclude more dots, but always only finitely many. Both figures are oversimplified in that they show a sequence with each successive term getting closer to the limit. A convergent sequence can behave more coyly, repeatedly approaching its limit and then backing away, until eventually it approaches the limit and stays close.

In complement to geometry, another way to understand Definition 3.2.7 is to interpret it as legislating a sort of adversarial process.

To argue that a sequence has limit ℓ, we first allow someone who doubts this to demand how close the terms of the sequence must get and stay to ℓ. That is, the skeptic provides the *error tolerance* $\varepsilon > 0$, which can be very small but must be positive. Once ε is specified, the skeptic has to be quiet as we consult with the sequence. If, in response to the given ε, we can come up with a *starting index* $N \in \mathcal{Z}_{\geq 1}$ such that the terms of the sequence from that index onward,

$$(s_N,\ s_{N+1},\ s_{N+2},\ \ldots),$$

are all within ε of ℓ, then we have successfully responded to the skeptic. To say that that the sequence has limit ℓ is to say that we can so respond to

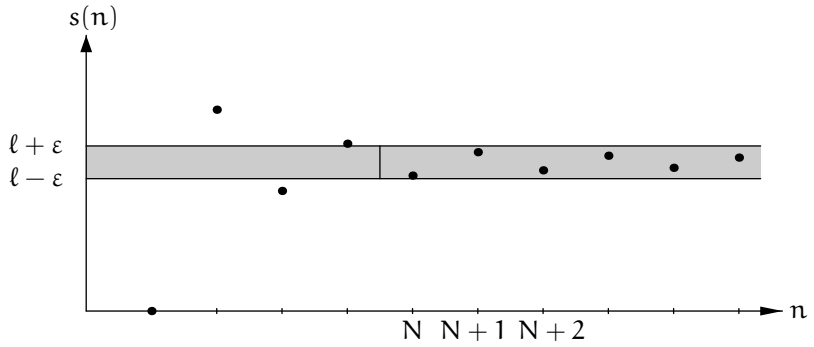


Figure 3.3. The definition of limit, viewing a sequence as a function

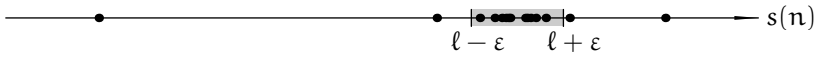


Figure 3.4. The definition of limit, viewing a sequence as a list of data

the skeptic's specified positive error tolerance with a corresponding starting index, no matter how small the error tolerance is.

On the other hand, to argue that a sequence does not have limit ℓ, it is we who play the role of the skeptic. After sizing up the sequence, we cleverly prescribe an error tolerance $\varepsilon > 0$ for which there is no starting index N, i.e., the terms of the sequence will never get and stay within distance ε of ℓ. In this case, we bear the onus of demonstrating that no starting index exists in response to our prescribed error tolerance.

All of this said, geometric and dramatic understanding of Definition 3.2.7 are ultimately developmental stages en route to a symbolic understanding of it, and to an appreciation that its versatility is astonishing. Indeed, we will see that it is so well designed that it can substantiate the disparate phenomena to be described in Proposition 3.2.8, Example 3.2.9, Example 3.2.10, Proposition 3.2.13, Proposition 3.2.14, Proposition 3.2.15, Proposition 3.2.19, Proposition 3.2.21, and Proposition 3.2.22.

Exercise

3.2.6. For each of the sequences below, calculate the first few terms, and make a guess as to whether or not the sequence converges. In some cases you will need to use a calculator. Try to explain the basis for your guess.

(a) $(s_n) = (1,\ 1+1/2,\ 1+1/2+1/3,\ 1+1/2+1/3+1/4,\ \ldots)$.
(b) $(s_n) = (1,\ 1-1/2,\ 1-1/2+1/3,\ 1-1/2+1/3-1/4,\ \ldots)$.
(c) $(s_n) = (1,\ 1+1/2^2,\ 1+1/2^2+1/3^2,\ 1+1/2^2+1/3^2+1/4^2,\ \ldots)$.
(d) $(s_n) = (1,\ 1+1/3,\ 1+1/3+1/3^2,\ 1+1/3+1/3^2+1/3^3,\ \ldots)$.
(e) $(s_n) = ((1+1/1)^1,\ (1+1/2)^2,\ (1+1/3)^3,\ (1+1/4)^4,\ \ldots)$.

3.2.4 Basic Sequence Limit Rules

Here are some examples of how to use Definition 3.2.7 to prove beginning results. None of the statements in the following proposition should be the least bit surprising.

Proposition 3.2.8 (Basic Sequence Limits).

(1) (Constant Sequence Rule.) *Let c be any real number. Consider the sequence s whose terms are $s_n = c$ for all $n \in \mathcal{Z}_{\geq 1}$,*

$$(s_n) = (c,\ c,\ c,\ \ldots) = (c).$$

This sequence's limit is c,

$$\boxed{\lim_n(c) = c \quad \text{for } c \in \mathcal{R}.}$$

(2) ($1/n$ Rule.) *Consider the sequence s whose terms are $s_n = 1/n$,*

$$(s_n) = (1,\ 1/2,\ 1/3,\ \ldots) = (1/n).$$

This sequence's limit is 0,

$$\boxed{\lim_n(1/n) = 0.}$$

(3) ($1/n^\alpha$ Rule.) *Let α be a positive rational number. Consider the sequence s whose terms are $s_n = 1/n^\alpha$,*

$$(s_n) = (1,\ 1/2^\alpha,\ 1/3^\alpha,\ \ldots) = (1/n^\alpha).$$

This sequence's limit is 0,

$$\boxed{\lim_n(1/n^\alpha) = 0 \quad \text{for } \alpha \in \mathcal{Q}_{>0}.}$$

(4) (nth Root Rule.) *Let* b *be a positive real number. Consider the sequence* s *whose terms are* $s_n = b^{1/n}$,

$$(s_n) = (b,\ b^{1/2},\ b^{1/3},\ \ldots) = (b^{1/n}).$$

This sequence's limit is 1,

$$\boxed{\lim_n (b^{1/n}) = 1 \quad \textit{for } b \in \mathcal{R}_{>0}.}$$

(5) (nth Power Rule.) *Let* r *be a real number such that* $|r| < 1$ $(-1 < r < 1)$. *Consider the sequence* s *whose terms are* $s_n = r^n$,

$$(s_n) = (1,\ r,\ r^2,\ \ldots) = (r^n).$$

This sequence's limit is 0,

$$\boxed{\lim_n (r^n) = 0 \quad \textit{for } |r| < 1.}$$

The sequence in the nth Power Rule has domain $\mathcal{Z}_{\geq 0}$ rather than $\mathcal{Z}_{\geq 1}$, but this is not a serious issue. More generally, the terms of a sequence can start at any index n rather than at $n = 1$, such as

$$(s_n)_{n \geq 17} = (s_{17},\ s_{18},\ s_{19},\ \ldots)$$

or

$$(s_n)_{n=-5}^{\infty} = (s_{-5},\ s_{-4},\ s_{-3},\ \ldots).$$

As before, the idea is that a sequence is one datum per generation where there is a starting generation (e.g., 17 or -5 as in the two examples just given) and then an endless succession of generations thereafter. We could insist that the initial generation always be indexed 1, but this would lead to notational contortions in situations such as the nth Power Rule where the initial generation clearly warrants the index 0 instead. Strictly speaking, the definitions of a sequence and of a sequence limit should be phrased to take into account the freer indexing scheme, but doing so is notationally onerous to no substantive purpose, especially in an environment whose grammar is already so symbol-heavy. In the case of sequence limits, we care only about the long-term behavior of the sequence anyhow, and so fussing about a finite shift in its indexing, or about aberrant behavior on the part of a small number of early terms, is patently irrelevant. We will soon quantify this *Irrelevance of Finite Index-Shifts*.

Note that each of the $1/n^\alpha$ rule, the nth root rule, and the nth power rule involves a free parameter—α or b or r—but a limit that is independent of the parameter. In chapter 1 we used the nth power rule with $r = 1/4$, and in chapter 2 we used the nth root rule first with $b = 8$ and then with general $b > 1$.

Proof. (1) To argue that $\lim_n(s_n) = c$ when $s_n = c$ for all $n \in \mathcal{Z}_{\geq 1}$, let any positive error tolerance $\varepsilon > 0$ whatsoever be given. Then the appropriate starting index in response to ε is simply $N = 1$. Indeed, because $s_n = c$ for all n, we have $|s_n - c| = 0$ for all n, and so certainly

$$\text{for all integers } n \geq 1, \quad |s_n - c| < \varepsilon.$$

Thus Definition 3.2.7 is satisfied.

(2) To argue that $\lim_n(1/n) = 0$, again let an error tolerance $\varepsilon > 0$ be given. Note that $|1/n - 0| = 1/n$ for $n \in \mathcal{Z}_{\geq 1}$. So we need to find a starting index $N \in \mathcal{Z}_{\geq 1}$ such that

$$\text{for all integers } n \geq N, \quad 1/n < \varepsilon.$$

By algebra, the previous display is equivalent to

$$\text{for all integers } n \geq N, \quad n > 1/\varepsilon,$$

and to show this, it suffices to show instead that some suitable starting index N satisfies

$$N > 1/\varepsilon,$$

because then also $n > 1/\varepsilon$ for all $n \geq N$. Because ε is positive and presumably small, $1/\varepsilon$ is positive and presumably big. However, no matter how big $1/\varepsilon$ is, the Archimedean Property of the real number system says that there exists some positive integer $N > 1/\varepsilon$. This completes the argument that the sequence $(1/n)$ has limit 0.

(3) Let α be a positive rational number. To argue that $\lim_n(1/n^\alpha) = 0$, again let $\varepsilon > 0$ be given. We want to find a corresponding $N \in \mathcal{Z}_{\geq 1}$ such that

$$\text{for all integers } n \geq N, \quad 1/n^\alpha < \varepsilon.$$

By a little algebra, the previous display is equivalent to

$$\text{for all integers } n \geq N, \quad n > 1/\varepsilon^{1/\alpha},$$

and to show this, it suffices to show instead that some suitable starting index N satisfies

$$N > 1/\varepsilon^{1/\alpha},$$

because then also $n > 1/\varepsilon^{1/\alpha}$ for all $n \geq N$. If ε is a small positive real number and α is a small positive rational number then $1/\varepsilon^{1/\alpha}$ is *very* big. Nonetheless, citing the Archimedean Property of the real number system completes the argument, as in the proof of (2).

(4) Let b be a positive real number. We need to argue that $\lim_n(b^{1/n}) = 1$.

First, if $b = 1$ then the sequence $(b^{1/n})$ is the constant sequence (1), and the result follows from the Constant Sequence Rule.

Second, if $b > 1$ then also $b^{1/n} > 1$ for each positive integer n, and so $|b^{1/n} - 1| = b^{1/n} - 1$ for each positive integer n. As shown in exercise 2.3.2,

$$b^{1/n} - 1 = \frac{b-1}{1 + b^{1/n} + \cdots + b^{(n-1)/n}} < \frac{b-1}{n}.$$

Now let an error tolerance $\varepsilon > 0$ be given. If we can produce a positive integer N such that $(b-1)/N < \varepsilon$ then it will follow that

$$\text{for all integers } n \geq N, \quad |b^{1/n} - 1| = b^{1/n} - 1 < \frac{b-1}{n} \leq \frac{b-1}{N} < \varepsilon,$$

completing the argument in the case $b > 1$. As for producing a positive integer N such that $(b-1)/N < \varepsilon$, this inequality is equivalent to $N > (b-1)/\varepsilon$, and by the Archimedean Property there is such an N. So the argument for $b > 1$ is indeed complete.

Third, if $0 < b < 1$ then let $\tilde{b} = 1/b > 1$. So also $b = 1/\tilde{b}$, and in fact $b^{1/n}$ and $\tilde{b}^{1/n}$ are inverses for all n. Because $0 < b^{1/n} < 1$ and $\tilde{b}^{1/n} > 1$ for all n,

$$|b^{1/n} - 1| = 1 - b^{1/n} = b^{1/n}(\tilde{b}^{1/n} - 1) < \tilde{b}^{1/n} - 1.$$

And because $\tilde{b} > 1$, we know from a moment ago that $\lim_n(\tilde{b}^{1/n}) = 1$. Now let $\varepsilon > 0$ be given. Then there exists some starting index N such that

$$\text{for all integers } n \geq N, \quad \tilde{b}^{1/n} - 1 < \varepsilon.$$

It follows from the previous two displays that

$$\text{for all integers } n \geq N, \quad |b^{1/n} - 1| < \varepsilon.$$

This completes the argument.

(5) Let r be a real number such that $|r| < 1$. We need to argue that $\lim_n(r^n) = 0$.

Let an error tolerance $\varepsilon > 0$ be given. Because $|r^n - 0| = |r|^n$, we need to show that for some positive integer N, we have $|r|^n < \varepsilon$ for all $n \geq N$. By (4), the sequence $(\varepsilon^{1/n})$ has limit 1, and so, because $|r| < 1$, there is a positive integer N such that $|r| < \varepsilon^{1/n}$ for all $n \geq N$. Consequently $|r|^n < \varepsilon$ for all $n \geq N$, and this is the desired result. □

Here we proved (5) quickly in consequence of having established (4). Alternatively, it is possible to establish (5) independently, and then prove (4) quickly in consequence. The main idea is that for $|r| < 1$ and $0 < b < 1$, the nth powers $|r|^n$ drop below b exactly when the nth roots $b^{1/n}$ climb above $|r|$. For example, suppose that $\lim_n(r^n) = 0$ for every real number r such that

$|r| < 1$. Consider any real number b such that $0 < b < 1$. For every r such that $|r| < 1$ we have $|r|^n < b$ for all large enough n (here *large enough* depends on r), and so $|r| < b^{1/n} < 1$ for all large enough n. Because $|r|$ can be as close to 1 as desired, this shows that $\lim_n(b^{1/n}) = 1$.

Although the proof of Proposition 3.2.8 has been written out in considerable length, once the reader digests its ideas, most of them should not seem difficult. The Constant Sequence Rule is an instant consequence of the definition of limit, and so are the $1/n$ Rule and the $1/n^\alpha$ Rule once one is aware of the Archimedean property of the real number system. The proof of the nth Root Rule relies on the finite geometric sum formula, and it was already carried out in exercise 2.3.2, The nth Power Rule follows from the nth Root Rule, or conversely as discussed in the previous paragraph. Quick mechanical proofs of the nth Root Rule and the nth Power Rule will become available (in exercise 5.1.4 to follow) once we have the logarithm.

Example 3.2.9. For an example of a divergent sequence, consider

$$(n) = (1,\ 2,\ 3,\ \ldots).$$

To see that (n) is divergent, suppose instead that it has a limit ℓ, and prescribe the error tolerance $\varepsilon = 1/3$. For any candidate positive integer N to serve as a suitable starting index, we would need to have

$$|n - \ell| < 1/3 \quad \text{for all } n \geq N,$$

so that in particular, letting $n = N$ and then letting $n = N + 1$,

$$|N - \ell| < 1/3 \quad \text{and} \quad |N + 1 - \ell| < 1/3,$$

It is intuitively impossible that N and $N + 1$, which are distance 1 away from each other, could both be within distance $1/3$ of ℓ. To quantify the impossibility, rewrite the previous display to emphasize the distance of ℓ from N and from $N + 1$ rather than their distances from ℓ

$$|\ell - N| < 1/3 \quad \text{and} \quad |\ell - (N + 1)| < 1/3,$$

and then note that consequently, by Proposition 3.2.4,

$$\ell \in (N - 1/3, N + 1/3) \quad \text{and} \quad \ell \in (N + 2/3, N + 4/3),$$

which is nonsense. Because the assumption that $\lim_n(n) = \ell$ for some ℓ has led to a contradiction, the sequence (n) has no limit.

Example 3.2.10. For another example of a divergent sequence, consider

$$(s_n)_{n\geq 0} = ((-1)^n)_{n\geq 0} = (1,\ -1,\ 1,\ -1,\ \ldots).$$

Suppose that (s_n) has limit ℓ. Let $\varepsilon = 1/2$. Then all terms s_n for large enough n lie within $1/2$ of ℓ. Because there are terms $s_n = -1$ and terms $s_n = 1$ for arbitrarily large n, we have $|-1-\ell| < 1/2$ and $|1-\ell| < 1/2$, or

$$|\ell - (-1)| < 1/2 \quad \text{and} \quad |\ell - 1| < 1/2,$$

and so by Proposition 3.2.4,

$$\ell \in (-3/2, -1/2) \quad \text{and} \quad \ell \in (1/2, 3/2),$$

which is nonsense. So the supposition that (s_n) has a limit is unsustainable. The sequence (s_n) *diverges by oscillation*.

Example 3.2.11. The equalities

$$\begin{aligned} 4 &= 2+2 \\ 6 &= 3+3 \\ 8 &= 3+5 \\ 10 &= 3+7 \\ 12 &= 5+7 \\ 14 &= 3+11 \end{aligned}$$

show that each even integer from 4 to 14 is the sum of two prime numbers. The *Goldbach Conjecture* (GC) states that in fact *every* even integer at least 4 is the sum of two primes. The conjecture dates back to the 18th century, and to this day nobody has shown a proof or a counterexample. Define a sequence

$$(g_n)_{n\geq 2}$$

as follows:

$$g_n = \begin{cases} 1 & \text{if each of } 4,\ 6,\ \ldots,\ 2n \text{ is the sum of two primes,} \\ 0 & \text{if not.} \end{cases}$$

Thus the sequence (g_n) begins

$$(1,\ 1,\ 1,\ 1,\ 1,\ 1,\ \ldots),$$

and either it says at 1 forever, or at some point it changes to 0 and then stays at 0 forever. Is (g_n) convergent, and if so then what is its limit? The sequence is constructed so that by definition its limit is

$$\lim_n(g_n) = \begin{cases} 1 & \text{if GC is true,} \\ 0 & \text{if GC is false,} \end{cases}$$

but is this a satisfying answer? Playing the role of the skeptic in the framework of Definition 3.2.7, we set $\varepsilon = 1/2$ and request a corresponding starting index N. The best response that an advocate of the convergence of (g_n) can give is to define, conditionally on GC being false, N_0 as the smallest integer at least 2 such that $2N_0$ is not the sum of two primes, and then to say that the starting index is

$$N = \begin{cases} 2 & \text{if GC is true,} \\ N_0 & \text{if GC is false.} \end{cases}$$

Again, is this answer satisfactory, or even meaningful? What if GC is neither provable nor disprovable from the usual starting assumptions about arithmetic (whatever these may be)? The questions here are questions of logic and philosophy, not the subject-matter of these notes, but this example is meant to show that even the beautifully-crafted grammar of Definition 3.2.7 does not answer all questions about sequence limits.

Exercises

3.2.7. (a) Recall the factorial function, denoted by an exclamation mark,

$$1! = 1, \quad 2! = 2 \cdot 1, \quad 3! = 3 \cdot 2 \cdot 1, \quad \dots .$$

Consider the sequence

$$(s_n) = (1/n!) = (1/1!,\ 1/2!,\ 1/3!,\ \dots).$$

Neither the $1/n$ Rule nor the $1/n^\alpha$ Rule (Proposition 3.2.8 (2) and (3)) applies to this sequence, but does either of them suggest anything about it? Explain.

(b) Consider the sequence

$$(s_n) = (n^{1/n}) = \lim_n(1,\ 2^{1/2},\ 3^{1/3},\ \dots).$$

Explain why the nth Root Rule (Proposition 3.2.8 (4)) does not apply directly to this sequence. Does the rule suggest anything about the sequence? Using suitable computing power, calculate some terms of (s_n) and then make a conjecture about its long-term behavior.

(c) Consider the sequence

$$(s_n) = ((n!)^{1/n}) = \lim_n(1,\ 2^{1/2},\ 6^{1/3},\ \dots).$$

Explain why the nth Root Rule (Proposition 3.2.8 (4)) does not apply directly to this sequence. Does the rule suggest anything about the sequence? Using suitable computing power, calculate some terms of (s_n) and then make a conjecture about its long-term behavior.

(d) Consider the sequence

$$(s_n) = (n^{10}/(1.1)^n) = (1/1.1,\ 2^{10}/(1.1)^2,\ 3^{10}/(1.1)^3,\ \ldots).$$

Using suitable computing power, calculate some terms of (s_n) and then make a conjecture about its long-term behavior.

3.2.8. (a) Explain why the sequence (r^n) is convergent for $r = 1$. What is its limit?

(b) Explain why the sequence (r^n) is divergent for $r = -1$.

3.2.9. The argument given in the text that (n) is divergent showed that the error tolerance $\varepsilon = 1/3$ has no corresponding starting index N. Show that $\varepsilon = 1/2$ also works as the error tolerance in the given argument. On the other hand, $\varepsilon = 2/3$ does not work in the given argument, but it works in a modified argument. Provide the modification.

3.2.10. Let (s_n) be a real sequence, and let ℓ be a real number. Suppose that as n gets ever larger, s_n gets ever nearer to ℓ. That is, suppose that for all $n, m \in \mathcal{Z}_{\geq 1}$,

$$\text{if } n > m \text{ then } |s_n - \ell| < |s_m - \ell|.$$

It does not follow that (s_n) converges to ℓ. Provide a counterexample.

3.2.11. Take the following result as given: *Let x be a nonnegative real number and let n be a nonnegative integer. Then*

$$(1+x)^n \geq 1 + nx.$$

Indeed, the expansion of $(1+x)^n = (1+x)(1+x)\cdots(1+x)$ contains the term 1 and n occurrences of the term x along with other nonnegative terms.

(a) Let a fixed real number $b > 1$ be given. Use the result just given to show that given any $\varepsilon > 0$, if $n > (b-1)/\varepsilon$ then $b^{1/n} - 1 < \varepsilon$. This gives another proof of Proposition 3.2.8(4) when $b > 1$.

(b) Let a fixed real number r such that $0 < |r| < 1$ be given, and set $x = 1/|r| - 1$ so that $|r| = 1/(1+x)$. Use the result given at the beginning of this exercise to show that given any $\varepsilon > 0$ with $\varepsilon < 1$, if $n > (1/\varepsilon - 1)/x$ then $|r|^n < \varepsilon$. This gives another proof of Proposition 3.2.8(5) when $r \neq 0$.

3.2.5 Irrelevance of Finite Index-Shifts

A brief discussion will quantify the earlier comment that a finite shift in a sequence's indexing is irrelevant to its limiting behavior.

Definition 3.2.12 (Index-Translate of a Sequence). *Let*

$$(s_n) = (s_1,\ s_2,\ s_3,\ \ldots)$$

be a real sequence, and let $p \in \mathcal{Z}_{\geq 1}$. *Then the sequence*

$$(s_{p+n}) = (s_{p+1},\ s_{p+2},\ s_{p+3},\ \ldots)$$

is called an **index-translate** *of* (s_n).

For example, the sequence

$$\left(\frac{1}{9},\ \frac{1}{16},\ \frac{1}{25},\ \cdots\right) = \left(\frac{1}{n^2}\right)_{n\geq 3} = \left(\frac{1}{(n+2)^2}\right)_{n\geq 1}$$

is an index-translate of the sequence

$$\left(1,\ \frac{1}{4},\ \frac{1}{9},\ \cdots\right) = \left(\frac{1}{n^2}\right)_{n\geq 1}.$$

Proposition 3.2.13 (Index-Translation Rule for Sequences). *Let* (s_n) *and* (t_n) *be real sequences, where* (t_n) *is an index-translate of* (s_n). *Then* (s_n) *converges and has limit* ℓ *if and only if* (t_n) *converges and has the same limit* ℓ. *That is, the two sequences converge or diverge together, and if they both converge then they have the same limit.*

Proof. We have

$$(s_n) = (s_1,\ s_2,\ s_3,\ \ldots)$$

and, for some positive integer p,

$$(t_n) = (s_{p+1},\ s_{p+2},\ s_{p+3},\ \ldots).$$

Suppose that (t_n) converges and has limit ℓ. We need to show that also (s_n) converges and has limit ℓ. So, let $\varepsilon > 0$ be given. We need to find a suitable starting index N for (s_n) in response to ε. On the other hand, we *know* that there is a suitable starting index M for (t_n) in response to ε. That is,

$$\text{for all } n \geq M, \quad |t_n - \ell| < \varepsilon.$$

Because $t_n = s_{p+n}$ for all n, the previous display rewrites as

$$\text{for all } n \geq M, \quad |s_{p+n} - \ell| < \varepsilon,$$

or

$$\text{for all } n \geq M + p, \quad |s_n - \ell| < \varepsilon,$$

Thus the appropriate starting index for (s_n) in response to ε is $N = M + p$.

Now suppose that (s_n) converges and has limit ℓ. We need to show that also (t_n) converges and has limit ℓ. Doing so is exercise 3.2.12 □

For example,

$$\lim_n \left(\frac{1}{(n+2)^2} \right) = \lim_n \left(\frac{1}{n^2} \right) = 0 \quad \text{by the } 1/n^\alpha \text{ rule.}$$

Exercises

3.2.12. Complete the proof of Theorem 3.2.13.

3.2.13. Let (s_n) and (t_n) be real sequences. Suppose that $\lim_n(s_n) = \ell$, and suppose that

$$|t_n - \ell| \leq |s_n - \ell| \quad \text{for all } n \in \mathcal{Z}_{\geq 1}.$$

Show that consequently $\lim_n(t_n) = \ell$.

3.2.6 Uniqueness of the Limit

Definition 3.2.7 (page 73) has a subtle worrisome feature: its wording allows the possibility of a convergent sequence having more than one limit.

Common sense dictates that a sequence can have at most one limit, but because our notion of limit is encoded as the grammar of Definition 3.2.7, we can not prove that a sequence has at most one limit by appealing to common sense. In fact, what common sense really dictates is that if some sequence has more than one limit under Definition 3.2.7, then the definition is foolishly posed. On the other hand, a light, graceful argument that a sequence can have at most one limit under Definition 3.2.7 would be evidence that the definition has been well formulated to capture the right ideas in the right way. Here is the argument.

Proposition 3.2.14 (Uniqueness of Limits). *Let (s_n) be a real sequence, and let ℓ and ℓ' be real numbers. Suppose that*

$$\lim_n(s_n) = \ell \quad \textit{and} \quad \lim_n(s_n) = \ell'.$$

Then $\ell' = \ell$.

The idea of the proof is that because the varying terms of the sequence (s_n) get arbitrarily close to ℓ and to ℓ', necessarily the fixed numbers ℓ and ℓ' must be arbitrarily close to each other, making them equal.

Proof. First note that for any positive integer n whatsoever, the Triangle Inequality gives

$$|\ell' - \ell| = |(s_n - \ell) - (s_n - \ell')| \leq |s_n - \ell| + |s_n - \ell'|. \tag{3.10}$$

Next let $\varepsilon > 0$ be given. Then also $\varepsilon/2 > 0$. (This seemingly pointless observation is a small piece of artfulness, guided by hindsight, that will pay off below.) Because $\lim_n(s_n) = \ell$, in response to the error tolerance $\varepsilon/2$ there is a starting index $M \in \mathcal{Z}_{\geq 1}$ such that

$$|s_n - \ell| < \varepsilon/2 \quad \text{for all } n \geq M. \tag{3.11}$$

Similarly, because $\lim_n(s_n) = \ell'$ there is a starting index $M' \in \mathcal{Z}_{\geq 1}$ such that

$$|s_n - \ell'| < \varepsilon/2 \quad \text{for all } n \geq M'. \tag{3.12}$$

Let N be the larger of M and M'. For $n = N$, (3.10) and (3.11) and (3.12) combine to give

$$|\ell' - \ell| \leq |s_N - \ell| + |s_N - \ell| < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

And now the nice little point is that keeping only the quantities at the two extreme ends of the previous display gives an inequality that makes no reference to the index N or to any sequence entry s_N that went into establishing it,

$$|\ell' - \ell| < \varepsilon.$$

Because this inequality holds for all $\varepsilon > 0$, we have $\ell' = \ell$ by the Strong Approximation Lemma. □

In the proof just given, we gained insight by simplifying, specifically by letting go of helper facts after they helped us, rather than doggedly insisting that every specific must continue to matter. This is a small instance of mathematical elegance.

3.2.7 Progenitive Sequence Limit Rules

Thanks to Proposition 3.2.8, we have five specific sequence limits in hand. But in addition to computing the limits of *particular* sequences, we can also compute the limits of *combinations* of sequences, assuming that we already

know the limits of the sequences individually. That is, in addition to computing limits from scratch, we can compute limits *progenitively*

The combinations of real sequences involved are as follows. Consider two sequences

$$s, t : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{R}.$$

Let $c \in \mathcal{R}$ be any number. Then the sequences

$$s \pm t, cs, st : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{R}$$

are defined as follows:

$$\begin{aligned} (s \pm t)_n &= s_n \pm t_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1}, \\ (cs)_n &= c \cdot s_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1}, \\ (st)_n &= s_n t_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1}. \end{aligned}$$

These sequences are the *sum/difference* of s and t, a *constant multiple* of s, and the *product* of s and t. Also, if $t_n \neq 0$ for all $n \in \mathcal{Z}_{\geq 1}$ then the sequences

$$1/t, s/t : \mathcal{Z}_{\geq 1} \longrightarrow \mathcal{R}$$

are defined to be

$$(1/t)_n = 1/t_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1}$$

and

$$(s/t)_n = s_n/t_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1}.$$

These sequences are the *reciprocal* of t and the *quotient* of s and t. The following result gives the limits of these newly-defined sequences in terms of the limits of s and t.

Proposition 3.2.15 (Progenitive Sequence Limit Rules). *Consider two real sequences s and t. Let $c \in \mathcal{R}$ be any number. Suppose that $\lim_n s = \ell$ and $\lim_n t = m$. Then*

(1) (Sum/Difference Rule.) $\lim_n(s \pm t)$ *exists and is $\ell \pm m$. That is,*

$$\boxed{\lim_n(s_n \pm t_n) = \lim_n(s_n) \pm \lim_n(t_n), \quad \textit{if both limits on the right exist.}}$$

(2) (Constant Multiple Rule.) $\lim_n cs$ *exists and is $c\ell$. That is,*

$$\boxed{\lim_n(cs_n) = c \cdot \lim_n(s_n), \quad \textit{if the limit on the right exists.}}$$

(3) (Product Rule.) $\lim_n st$ *exists and is ℓm. That is,*

$$\boxed{\lim_n(s_n t_n) = \lim_n(s_n) \cdot \lim_n(t_n), \quad \textit{if both limits on the right exist.}}$$

(4) (Reciprocal Rule.) *If* $t_n \neq 0$ *for all* $n \in \mathcal{Z}_{\geq 1}$ *and* $m \neq 0$ *then* $\lim_n 1/t$ *exists and is* $1/m$. *That is,*

$$\boxed{\lim_n \left(\frac{1}{t_n}\right) = \frac{1}{\lim_n(t_n)}, \quad \begin{array}{l} \textit{if each } t_n \textit{ is nonzero and the limit} \\ \textit{on the right exists and is nonzero.} \end{array}}$$

(5) (Quotient Rule.) *If* $t_n \neq 0$ *for all* $n \in \mathcal{Z}_{\geq 1}$ *and* $m \neq 0$ *then* $\lim_n s/t$ *exists and is* ℓ/m. *That is,*

$$\boxed{\lim_n \left(\frac{s_n}{t_n}\right) = \frac{\lim_n(s_n)}{\lim_n(t_n)}, \quad \begin{array}{l} \textit{if each } t_n \textit{ is nonzero and both limits} \\ \textit{on the right exist and } \lim_n(t_n) \textit{ is nonzero.} \end{array}}$$

The Difference Rule is a consequence of the Sum Rule and the Constant Multiple Rule, and so perhaps it doesn't deserve its own name. The fussy condition in the Reciprocal Rule and the Quotient Rule that $t_n \neq 0$ for all $n \in \mathcal{Z}_{\geq 1}$ can handwaved away: if $\lim_n(t_n) \neq 0$ then necessarily all t_n are nonzero past some starting index, and finite index-shifts are irrelevant to limits.

Proof. (Sketch.) (1) To prove the Sum Rule by arguing that $\lim_n(s_n + t_n) = \lim_n(s_n) + \lim_n(t_n)$ provided both limits on the right exist, let $\ell = \lim_n(s_n)$ and let $m = \lim_n(t_n)$. First note that for any positive integer n,

$$\begin{aligned} |(s+t)_n - (\ell + m)| &= |s_n + t_n - \ell - m| \\ &= |s_n - \ell + t_n - m| \\ &\leq |s_n - \ell| + |t_n - m|. \end{aligned}$$

Now let any error tolerance $\varepsilon > 0$ be given for $s + t$. Prescribe the error tolerance $\varepsilon/2$ for s to get a starting index N_s such that

$$\text{for all } n \geq N_s, \quad |s_n - \ell| < \varepsilon/2.$$

Similarly, there is a starting index N_t such that

$$\text{for all } n \geq N_t, \quad |t_n - m| < \varepsilon/2.$$

Let N be the larger of N_s and N_t. Then, using the previous three displays,

$$\text{for all } n \geq N, \quad |(s+t)_n - (\ell + m)| < \varepsilon.$$

Thus N is a suitable starting index in response to ε.

(2) The argument for the Constant Multiple Rule is very similar. The key calculation is

$$|(cs)_n - c\ell| = |c|\,|s_n - \ell|.$$

Let $\varepsilon > 0$ be given. For all large enough n, we have $|s_n - \ell| < \varepsilon/(|c|+1)$, and so $|(cs)_n - c\ell| < \varepsilon$. Using the denominator $|c|+1$ avoids dividing by 0 even when $c = 0$.

(3) For the Product Rule, the argument is a bit more elaborate. This time the key calculation is not purely mechanical: an auxiliary term is subtracted and added back before things arrange themselves nicely,

$$\begin{aligned} |(st)_n - \ell m| = |s_n t_n - \ell m| &= |s_n t_n - s_n m + s_n m - \ell m| \\ &\le |s_n t_n - s_n m| + |s_n m - \ell m| \\ &= |s_n|\,|t_n - m| + |s_n - \ell|\,|m|. \end{aligned}$$

Let $\varepsilon > 0$ be given. For large enough n, we have simultaneously that s_n is so close to ℓ and t_n is so close to m that plausibly the right side is less than ε. The details of the argument work as follows. We can ensure that for all large enough n,

- $|s_n| < |\ell| + 1$ and $|t_n - m| < \varepsilon/(2(|\ell|+1))$, so that $|s_n|\,|t_n - m| < \varepsilon/2$;
- $|s_n - \ell| < \varepsilon/(2(|m|+1))$, so that $|s_n - \ell|\,|m| < \varepsilon/2$.

It follows that for all large enough n,

$$|(st)_n - \ell m| \le |s_n|\,|t_n - m| + |s_n - \ell|\,|m| < \varepsilon.$$

(4) For the Reciprocal Rule, the key calculation is again mechanical,

$$|(1/t)_n - 1/m| = \left|\frac{1}{t_n} - \frac{1}{m}\right| = \left|\frac{m - t_n}{m t_n}\right| = \frac{|t_n - m|}{|m|\,|t_n|}\,.$$

For large enough n, simultaneously $|t_n - m| < \varepsilon|m|^2/2$ and $|t_n| > |m|/2$, and so

$$\frac{|t_n - m|}{|m|\,|t_n|} < \frac{\varepsilon|m|^2/2}{|m|\,|m|/2} = \varepsilon.$$

Concatenating the previous two displays gives the desired result.

(5) The Quotient Rule follows from the Product Rule and the Reciprocal Rule. □

For example, consider the limit

$$\lim_n \left(\frac{n^3 - 2n^2}{3n^3 + 4}\right).$$

The Quotient Rule does not apply immediately, because the limits of the numerator and denominator do not exist. (As always, ∞ is not a number.)

However, factor the highest power of n out of the numerator and denominator of each term of the sequence,

$$\frac{n^3-2n^2}{3n^3+4} = \frac{n^3(1-2/n)}{n^3(3+4/n^3)} = \frac{1-2/n}{3+4/n^3} \quad \text{for all } n \in \mathcal{Z}_{\geq 1}.$$

It is now clear what the limit is.

$$\lim_n \left(\frac{n^3-2n^2}{3n^3+4}\right) = \lim_n \left(\frac{1-2/n}{3+4/n^2}\right) = \frac{1-2\cdot 0}{3+4\cdot 0} = \frac{1}{3}.$$

The second equality in the last display comes from applying the Constant Sequence Rule, the $1/n$ Rule, the Constant Multiple Rule, and the Difference Rule in the numerator, the Constant Sequence Rule, the $1/n^\alpha$ Rule with $\alpha = 2$, the Constant Multiple Rule, and the Sum Rule in the denominator, and finally the Quotient Rule. The main feature of this example is that the original numerator and denominator had the same highest power of n, and the limit was the ratio of the relevant coefficients.

Earlier we contended with the subtle issue of the uniqueness of limits. The *existence* of limits, or lack thereof, is another subtle point that raises possible misapplications of the various sequence rules.

Example 3.2.16. The Product Rule says that a sequence that is the product of two convergent sequences is again convergent. But the converse is not true. That is, the product of two sequences, at least one of which diverges, still can converge, although it may well diverge.

For a very easy example, the product of any divergent sequence whatsoever of nonzero real numbers with its reciprocal sequence is the constant sequence (1), as nicely convergent as can be. For more interesting examples, consider the sequences

$$(s_n) = ((-1)^n(1/2)^n) \quad \text{and} \quad (t_n) = ((-1)^n(1/2)^{1/n}).$$

Each of these sequences is a product,

$$(s_n) = ((-1)^n)\cdot((1/2)^n) \quad \text{and} \quad (t_n) = ((-1)^n)\cdot((1/2)^{1/n}).$$

And we know that the sequence $((-1)^n)$ diverges. But as just explained, it does not follow that the sequences (s_n) and (t_n) diverge in consequence. Indeed, note that

$$(s_n) = ((-1/2)^n) = (r^n) \quad \text{where } r = -1/2,$$

and so $\lim_n(s_n)$ exists and is 0 by the nth Power Rule. But on the other hand, we know that $\lim_n((1/2)^{1/n}) = 1$ by the nth Root Rule, and so the terms of (t_n) tend ever more closely to alternating between 1 and -1. Thus $\lim_n(t_n)$ does not exist.

Example 3.2.17. Let r be a real number such that $|r| < 1$. Consider the sequence s whose terms are $s_n = r^n$,

$$(s_n) = (1,\ r,\ r^2,\ \ldots),$$

and consider a constant multiple of the sequence, $t = rs$,

$$(t_n) = (r,\ r^2,\ r^3,\ \ldots),$$

By the Constant Multiple Rule,

$$\lim_n(t_n) = r \cdot \lim_n(s_n).$$

But also, because (t_n) is an index-translate of (s_n), the Irrelevance of Finite Index-Shifts gives

$$\lim_n(t_n) = \lim_n(s_n).$$

Therefore

$$\lim_n(s_n) = r \cdot \lim_n(s_n),$$

and because $r \neq 1$, this gives

$$\lim_n(s_n) = 0,$$

which is precisely the nth Power Rule,

$$\lim_n(r^n) = 0.$$

So apparently the earlier proof of the nth Power Rule was unnecessary.

However, there must be a flaw in the reasoning here. The argument used only the assumption that $r \neq 1$, not that $|r| < 1$. So it purports to show, for example, that the sequence for $r = -1$,

$$(1,\ -1,\ 1,\ -1,\ \ldots),$$

has limit 0, which it does not. Furthermore, the argument purports to show that the sequence for $r = 2$,

$$(1,\ 2,\ 4,\ 8,\ \ldots),$$

also has limit 0, which it most certainly does not. The flaw in the reasoning is the assumption that $\lim_n(s_n)$ exists at all. What the argument has correctly shown is that *if* $\lim_n(r^n)$ exists *and* $r \neq 1$ then $\lim_n(r^n) = 0$.

Example 3.2.18. Consider the sequence (s_n) defined by the rules

$$\left\{\begin{array}{l} s_1 = 1, \\ s_2 = 1, \\ s_n = \dfrac{1+s_{n-1}}{s_{n-2}} \quad \text{for } n > 2. \end{array}\right\} \tag{3.13}$$

Thus

$$s_3 = \frac{1+1}{1} = 2, \quad s_4 = \frac{1+2}{1} = 3,$$

and so on. Note that s_n is positive for each $n \in \mathcal{Z}_{\geq 1}$. Let

$$\ell = \lim_n (s_n).$$

By the Quotient Rule, the Sum Rule, and the Index-Translation Rule, also

$$\ell = \frac{1 + \lim_n (s_n)}{\lim_n (s_n)} = \frac{1+\ell}{\ell}.$$

Thus $\ell^2 = 1 + \ell$, so that by the Quadratic Formula

$$\ell = \frac{1 \pm \sqrt{5}}{2}.$$

Because each s_n is positive, ℓ must be the positive root. (We will formalize this assertion in the Inequality Rule for sequences, Proposition 3.2.21, later in this chapter.) That is, the sequence (s_n) has limit

$$\ell = \frac{1+\sqrt{5}}{2}.$$

This example has the same flaw in its reasoning as the previous one, and its conclusion is flat-out wrong (exercise 3.2.16).

Exercises

3.2.14. Find the following limits, or explain why they don't exist.

(a) $\lim_n \left(7 + 6/n + 8/\sqrt{n}\right)$.

(b) $\lim_n \left(\dfrac{4+1/n}{5+1/n}\right)$.

(c) $\lim_n \left(\dfrac{3n^2+n+1}{1+3n+4n^2}\right)$.

(d) $\lim_n \left(\dfrac{(2+1/n)^2+4}{(2+1/n)^3+8}\right)$.

(e) $\lim_n \left(\dfrac{(2+1/n)^2-4}{(2+1/n)^3-8}\right)$.

(f) $\lim_n \left(\dfrac{8n^3 + 13n}{17 + 12n^3} \right)$.

(g) $\lim_n \left(\dfrac{8(n+4)^3 + 13(n+4)}{17 + 12(n+4)^3} \right)$.

(h) $\lim_n \left(\dfrac{n+1}{n^2+1} \right)$.

3.2.15. (a) Consider the following argument: *The constant sequence* (0) *is*

$$\begin{aligned}(0) &= (0,\ 0,\ 0,\ 0,\ \ldots) \\ &= (1-1,\ -1+1,\ 1-1,\ -1+1,\ \ldots) \\ &= (1,\ -1,\ 1,\ -1,\ \ldots) + (-1,\ 1,\ -1,\ +1,\ \ldots).\end{aligned}$$

But both of the last two sequences diverge by oscillation, and so the constant sequence (0) *diverges*. The argument must be wrong because $\lim_n(0) = 0$ by the Constant Sequence Rule. What is the flaw in the reasoning?

(b) Consider the following argument: *The constant sequence* (1) *has limit*

$$\begin{aligned}\lim_n(1) &= \lim_n(1,\ 1,\ 1,\ 1,\ \ldots) \\ &= \lim_n(1-0,\ 2-1,\ 3-2,\ 4-3,\ \ldots) \\ &= \lim_n(1,\ 2,\ 3,\ 4,\ \ldots) - \lim_n(0,\ 1,\ 2,\ 3,\ \ldots) \\ &= \infty - \infty \\ &= 0.\end{aligned}$$

But also $\lim_n(1) = 1$ *by the Constant Sequence Rule, and so* $0 = 1$. What is the flaw in the reasoning?

3.2.16. List the first ten terms of the sequence (3.13). What is the sequence's long-term behavior? Explain the flaw in the reasoning that sequence's limit is $(1+\sqrt{5})/2$.

3.2.17. Similarly to the sequence (3.13), consider the sequence

$$\left\{ \begin{array}{ll} s_1 = 1, & \\ s_{n+1} = \dfrac{s_n^2 + 2}{2s_n} & \text{for } n \geq 1. \end{array} \right\}$$

Assuming that this sequence has a limit ℓ, what is ℓ? Compute some terms of the sequence and use them to conjecture its actual behavior.

3.2.8 Geometric Series

Definition 3.2.19 (Geometric Series). *Let r be a real number. The sequence*

$$(s_n) = (1,\ 1+r,\ 1+r+r^2,\ \ldots) = (1+r+\cdots+r^{n-1})_{n\geq 1}$$

is the **geometric series with ratio** r.

It is crucial here to understand that the *terms* of the geometric series are ever-longer *sums*, specifically, ever-longer finite geometric sums.

Proposition 3.2.20 (Geometric Series Formula). *Let r be a real number such that $|r| < 1$. Then the geometric series with ratio r converges, and its limit is*

$$\lim_n (1+r+r^2+\cdots+r^{n-1}) = \frac{1}{1-r}.$$

Proof. By the finite geometric sum formula,

$$1+r+r^2+\cdots+r^{n-1} = \frac{1-r^n}{1-r} \quad \text{for } r \neq 1.$$

Because in fact $|r| < 1$, various sequence rules give the result immediately. □

The Geometric Sum Formula is often written as follows:

$$\boxed{1+r+r^2+\cdots+r^n+\cdots = \frac{1}{1-r} \quad \text{for } |r| < 1.}$$

Note the second "$+\cdots$" on the left side of the equality, connoting that the sum does not stop after any finite number of terms. That is, the formula is giving the value of an *infinite sum*, understood to be the limit of finite sums having more and more terms.

When $r = 1/4$, the Geometric Series Formula encodes the end-calculation of Archimedes's quadrature of the parabola from section 1.2

Exercises

3.2.18. (a) Let $(s_n) = (1+9/10+(9/10)^2+\cdots+(9/10)^{n-1})$. Find $\lim_n(s_n)$.

(b) Let $(s_n) = (1-9/10+(9/10)^2-\cdots+(-1)^{n-1}(9/10)^{n-1})$. Find $\lim_n(s_n)$.

3.2.19. (a) The "infinite decimal"

$$0.111\ldots = \lim_n(0.1,\ 0.11,\ 0.111,\ \ldots)$$

Is naturally viewed as a certain rational number. What rational number? Explain.

(b) Similarly, what rational number is $0.123123123\ldots$?

3.2.9 Order Sequence Limit Rules

Proposition 3.2.21 (Inequality Rule for Sequences). *Let* (s_n) *and* (t_n) *be convergent sequences. Suppose that*

$$s_n \leq t_n \quad \textit{for all } n \in \mathcal{Z}_{\geq 1}.$$

Then

$$\lim_n(s_n) \leq \lim_n(t_n).$$

Proof. Introduce a sequence $(u_n) = (t_n - s_n)$. Because both (s_n) and (t_n) converge, $\lim_n(u_n)$ exists and equals $\lim_n(t_n) - \lim_n(s_n)$. And so it suffices to prove that because $u_n \geq 0$ for all $n \in \mathcal{Z}_{\geq 1}$, also $\lim_n(u_n) \geq 0$.

For each $\varepsilon > 0$ there is a starting index N such that

$$\text{for all } n \geq N, \quad u_n < \lim_n(u_n) + \varepsilon.$$

But each $u_n \geq 0$, i.e., $0 \leq u_n$ for all n, and concatenating this with the previous display gives

$$\text{for all } n \geq N, \quad 0 < \lim_n(u_n) + \varepsilon.$$

The constants 0 and $\lim_n(u_n)$ and ε are independent of the generation-count n, so in fact we have a relation among constants,

$$0 < \lim_n(u_n) + \varepsilon,$$

or

$$-\varepsilon < \lim_n(u_n).$$

That is, $\lim_n(u_n)$ is greater than each negative number, no matter how close the negative number is to 0. Therefore $\lim_n(u_n) \geq 0$. □

The most common use of the Inequality Rule is in situations where

$$0 \leq t_n \quad \text{for all } n,$$

and we conclude that

$$0 \leq \lim_n(t_n).$$

Indeed, the proof of the rule proceeded by reducing it to this case.

Proposition 3.2.22 (Squeezing Rule for Sequences). *Let* (s_n), (t_n), *and* (u_n) *be three real sequences. Suppose that*

$$s_n \leq u_n \leq t_n \quad \textit{for all } n \in \mathcal{Z}_{\geq 1}.$$

Suppose further that (s_n) *and* (t_n) *both converge to the same limit* ℓ. *Then* (u_n) *also converges to* ℓ.

The nice point here is existence: the Squeezing Rule says that the middle sequence *has* a limit, and furthermore the limit is the shared limit ℓ of the outer sequences. If the middle sequence were already known to have a limit, then the limit would be ℓ by two applications of the Inequality Rule. The Squeezing Rule improves on the Inequality Rule in that it does not require us to know that the middle sequence has a limit. But on the other hand, it requires bounds on a sequence from both sides.

Proof. Let $\varepsilon > 0$ be given. For some starting index N_s,

$$\text{for all } n \geq N_s, \quad \ell - \varepsilon < s_n.$$

And for some starting index N_t,

$$\text{for all } n \geq N_t, \quad t_n < \ell + \varepsilon.$$

Let N be the larger of N_s and N_t. Then by the previous two displays and the hypothesis that $s_n \leq u_n \leq t_n$ for all $n \in \mathcal{Z}_{\geq 1}$,

$$\text{for all } n \geq N, \quad \ell - \varepsilon < u_n < \ell + \varepsilon.$$

That is,

$$\text{for all } n \geq N, \quad |u_n - \ell| < \varepsilon.$$

□

By the Irrelevance of Finite Reindexing, the Inequality Rule holds if instead:

$$\text{For some } N \in \mathcal{Z}_{\geq 1}, \quad s_n \leq t_n \quad \text{for all } n \geq N.$$

And the Squeezing Rule holds if instead:

$$\text{For some } N \in \mathcal{Z}_{\geq 1}, \quad s_n \leq u_n \leq t_n \quad \text{for all } n \geq N.$$

But the more simply stated versions are tidier to prove.

Exercises

3.2.20. Exercise 3.2.7 asks about the limits of four sequences whose limits do not follow from the five basic sequence limit rules. One of those four sequences has a limit that now can be found quickly by using the Squeezing Rule. Which sequence is it, and how does the Squeezing Rule tell us its limit?

3.2.21. For each of the statements to follow: if the statement is true then justify it by means of limit rules; if the statement is false then give a counterexample.

(a) Let (s_n) be a convergent real sequence. If $s_n > 0$ for all $n \in \mathcal{Z}_{\geq 1}$ then $\lim_n(s_n) > 0$.

(b) Let (s_n) and (t_n) be real sequences. If $\lim_n(s_n) = 0$ then $\lim_n(s_n t_n) = 0$.

(c) Let (s_n) be a real sequence. If $\lim_n(s_n^2) = 1$ then either $\lim_n(s_n) = 1$ or $\lim_n(s_n) = -1$.

(d) Let (s_n) and (t_n) be real sequences. If $\lim_n(s_n t_n) = 0$ then either $\lim_n(s_n) = 0$ or $\lim_n(t_n) = 0$.

3.2.22. (a) Suppose that we know the $1/n$ Rule (Proposition 3.2.8 (2)) but do not know the $1/n^\alpha$ Rule (Proposition 3.2.8 (3)). Suppose further that $\alpha \in \mathcal{Q}$ and $\alpha > 1$. Use a result or results from this section to establish the $1/n^\alpha$ Rule, i.e., $\lim_n(1/n^\alpha)$ exists and is 0. Where does the argument require $\alpha > 1$?

(b) Again suppose that we know the $1/n$ Rule but do not know the $1/n^\alpha$ Rule. This time suppose further that $\alpha \in \mathcal{Q}$ and $0 < \alpha < 1$. Explain why $N\alpha > 1$ for some $N \in \mathcal{Z}_{\geq 1}$. Now reason as follows. By N applications of the Product Rule, the Nth power of the limit is the limit of the Nth powers,

$$(\lim_n(1/n^\alpha))^N = \lim_n\left((1/n^\alpha)^N\right),$$

and the calculation continues,

$$\begin{aligned}
(\lim_n(1/n^\alpha))^N &= \lim_n\left((1/n^\alpha)^N\right) && \text{as just explained}\\
&= \lim_n(1/n^{N\alpha}) && \text{by algebra}\\
&= 0 && \text{by part (a), because } N\alpha > 1.
\end{aligned}$$

Therefore, $\lim_n(1/n^\alpha) = 0$ as well, and so the $1/n^\alpha$ Rule holds for $0 < \alpha < 1$ in consequence of the $1/n$ Rule as well. But there is a flaw in the reasoning here. What is it?

3.2.23. Consider the following variant of the Squeezing Rule: *Let* (s_n), (t_n), *and* (u_n) *be three real sequences. Suppose that*

$$s_n \leq u_n \leq t_n \quad \textit{for all } n \in \mathcal{Z}_{\geq 1}$$

and that

$$\lim_n(t_n - s_n) = 0.$$

Then (u_n) *also converges to the common limit* ℓ *of* (s_n) *and* (t_n), *which exists because* $\lim_n(t_n) - \lim_n(s_n) = \lim_n(t_n - s_n) = 0$. Is this variant correct? Either prove it, or explain the flaw in the reasoning and provide a counterexample.

3.3 Integrability

3.3.1 The Previous Examples Revisited

Section 3.1 recalled two sequences from chapters 1 and 2. The first was the sequence of triangle-area sums arising from Archimedes's quadrature of the parabola,

$$(S_n) = \Big(A_{\text{tri}} \cdot (4/3)\big(1 - (1/4)^n\big)\Big)_{n \geq 1}.$$

Now we can analyze this sequence quantitatively. By the Constant Sequence Rule and the nth Power Rule,

$$\lim_n(1) = 1 \quad \text{and} \quad \lim_n((1/4)^n) = 0,$$

and so by the Difference Rule,

$$\lim_n(1 - (1/4)^n) = 1,$$

and then by the Constant Multiple Rule,

$$\lim_n (S_n) = A_{\text{tri}} \cdot \frac{4}{3}.$$

Of course this is the value that we already obtained for the limiting value, but now it is on a much firmer footing.

The second sequence was a sequence of box-area sums arising from the integration of the rational power function,

$$(S_n) = \left((b^{\alpha+1} - 1) \Big/ \frac{s_n^{\alpha+1} - 1}{s_n - 1}\right)_{n \geq 1} \quad \text{where } s_n = b^{1/n}.$$

Here $b > 1$, and $\alpha \in \mathcal{Q}$ but $\alpha \neq -1$. We know by the nth Root Rule and because $b > 1$ that

- $s_n \in \mathcal{R}_{>0}$ for all $n \in \mathcal{Z}_{\geq 1}$,
- $\lim_n(s_n) = 1$,
- $s_n \neq 1$ for each $n \in \mathcal{Z}_{\geq 1}$.

These properties of (s_n) are all that we need to carry out an analysis that formalizes the derivative calculation in section 2.4. Because the analysis will be cited again in the next chapter, we isolate it. The general symbol α in the following proposition is not the specific α of the ambient discussion that has been broken off momentarily in order to establish the proposition.

Proposition 3.3.1. *Let $\alpha \in \mathcal{Q}$ be any rational number. Let (s_n) be any sequence of positive real numbers such that $\lim_n(s_n) = 1$ but $s_n \neq 1$ for each $n \in \mathcal{Z}_{\geq 1}$. For each $n \in \mathcal{Z}_{\geq 1}$, let*

$$t_n = \frac{s_n^\alpha - 1}{s_n - 1}.$$

Then

$$\lim_n(t_n) = \alpha.$$

Proof. As explained in section 2.4,

$$t_n = \begin{cases} 1 + s_n + s_n^2 + \cdots + s_n^{\alpha-1} & \text{if } \alpha \in \mathcal{Z}_{\geq 0}, \\ -\dfrac{1}{s_n^{-\alpha}} \cdot \dfrac{s_n^{-\alpha} - 1}{s_n - 1} & \text{if } \alpha \in \mathcal{Z}_{\leq -1}, \\ \dfrac{\tilde{s}_n^p - 1}{\tilde{s}_n - 1} \cdot \dfrac{\tilde{s}_n - 1}{\tilde{s}_n^q - 1} \quad \text{where } \tilde{s}_n = s_n^{1/q} & \text{if } \alpha = p/q,\ p \in \mathcal{Z},\ q \in \mathcal{Z}_{\geq 1}. \end{cases}$$

In the case $\alpha \in \mathcal{Z}_{\geq 0}$, because $\lim_n(1) = 1$ by the Constant Sequence Rule, and because we know that $\lim_n(s_n) = 1$, many applications of the Product Rule and then the Sum Rule give

$$\lim_n(t_n) = \lim_n \left(1 + s_n + s_n^2 + \cdots + s_n^{\alpha-1}\right) = \alpha.$$

Here the sum, and therefore the limit, are understood to be 0 if $\alpha = 0$.

Next consider the case $\alpha \in \mathcal{Z}_{\leq -1}$, so that $-\alpha \in \mathcal{Z}_{\geq 1}$. In this case,

$$t_n = -\frac{1}{s_n^{-\alpha}} \cdot \frac{s_n^{-\alpha} - 1}{s_n - 1}.$$

Because $\lim_n(s_n) = 1$, finitely many applications of the Product Rule say that $\lim_n(s_n^{-\alpha}) = 1$, and then, because each $s_n^{-\alpha}$ is nonzero and 1 is nonzero, the Reciprocal Rule gives $\lim_n(1/s_n^{-\alpha}) = 1/1 = 1$, and so by the argument from a moment ago for $\alpha \in \mathcal{Z}_{\geq 1}$, and by the Constant Multiple Rule and the Product Rule,

$$\lim_n(t_n) = -1 \cdot 1 \cdot (-\alpha) = \alpha.$$

Thus $\lim_n(t_n) = \alpha$ for all $\alpha \in \mathcal{Z}$.

For general $\alpha = p/q \in \mathcal{Q}$ where $p \in \mathcal{Z}$ and $q \in \mathcal{Z}_{\geq 1}$, the formula for t_n is

$$t_n = \frac{\tilde{s}_n^p - 1}{\tilde{s}_n - 1} \cdot \frac{\tilde{s}_n - 1}{\tilde{s}_n^q - 1} \text{ where } \tilde{s}_n = s_n^{1/q}.$$

By the finite geometric sum formula, and by the fact that each $\tilde{s}_n \in \mathcal{R}_{>0}$ because each $s_n \in \mathcal{R}_{>0}$,

$$|\tilde{s}_n - 1| = \left| \frac{s_n - 1}{1 + \tilde{s}_n + \tilde{s}_n^2 + \cdots + \tilde{s}_n^{q-1}} \right| = \frac{|s_n - 1|}{1 + \tilde{s}_n + \tilde{s}_n^2 + \cdots + \tilde{s}_n^{q-1}} \leq |s_n - 1|.$$

And so because $\lim_n(s_n) = 1$, also $\lim_n(\tilde{s}_n) = 1$ by exercise 3.2.13 (page 86). Now the formula

$$\lim(t_n) = p/q = \alpha$$

follows in this case from the previous two cases, the Reciprocal Rule, and the Product Rule. □

With the proposition proved, let the symbol α again take on its meaning from the sequence (S_n) and apply the proposition with $\alpha + 1$ as the α of the proposition. That is, now $\alpha \neq -1$ again, and if we let

$$t_n = \frac{s_n^{\alpha+1} - 1}{s_n - 1}, \quad n \in \mathcal{Z}_{\geq 1},$$

then the sequence that we want to analyze is

$$(S_n) = \left(\frac{b^{\alpha+1} - 1}{t_n} \right).$$

The proposition, the Reciprocal Rule, and the Constant Multiple Rule give the desired result,

$$\lim_n(S_n) = \frac{b^{\alpha+1} - 1}{\alpha + 1}.$$

The following consequence of the Squeezing Rule was tacitly used twice in chapter 2.

Proposition 3.3.2 (The Trapped Quantity is the Common Limit). *Let (s_n) and (t_n) be real sequences, and let u be a real number. Suppose that*

$$s_n \leq u \leq t_n \quad \textit{for all } n \in \mathcal{Z}_{\geq 1},$$

and suppose that (s_n) and (t_n) converge to the same limit,

$$\lim_n(s_n) = \lim_n(t_n) = \ell.$$

Then $u = \ell$.

Proof. Consider the constant sequence $(u_n) = (u)$, each of whose terms u_n is the number u. By the Constant Sequence Rule, $\lim_n(u) = u$. Also, we are given that

$$s_n \leq u_n \leq t_n \quad \text{for all } n \in \mathcal{Z}_{\geq 1},$$

so that by the Squeezing Rule, $\lim_n(u) = \ell$. Thus $u = \ell$. □

Proposition 3.3.2 is useful when u is some unknown area that we want to find and ℓ is the common limit of two sequences of box-area sums, one too small to be the area and one too big. Our first tacit use of the proposition was in section 2.5, where in addition to the sequence (S_n) from a moment ago, a second sequence appeared,

$$(T_n) = (s_n^\alpha S_n), \quad s_n = b^{1/n}.$$

For this sequence, the box-heights were determined by the values of the power function f_α over the right endpoints of their bases, rather than the left endpoints. And because (using the nth Root Rule for the last step in the display to follow)

$$\lim_n(s_n^\alpha) = \lim_n((b^{1/n})^\alpha) = \lim_n((b^\alpha)^{1/n}) = 1,$$

it follows that (assuming $\alpha \neq -1$)

$$\lim_n(T_n) = \lim_n(S_n) = \frac{b^{\alpha+1}-1}{\alpha+1}.$$

Let

$$\mathrm{Ar}_1^b(f_\alpha)$$

denote the area under the graph of f_α from 1 to b, a constant. Because the power function is increasing for $\alpha > 0$ and decreasing for $\alpha < 0$, we have

$$\begin{cases} S_n \le \mathrm{Ar}_1^b(f_\alpha) \le T_n & \text{for all } n \in \mathcal{Z}_{\ge 1}, \text{ if } \alpha > 0, \\ T_n \le \mathrm{Ar}_1^b(f_\alpha) \le S_n & \text{for all } n \in \mathcal{Z}_{\ge 1}, \text{ if } \alpha < 0. \end{cases}$$

In either case, Proposition 3.3.2 now gives firm footing to the familiar result that consequently the normalized power function area is

$$\mathrm{Ar}_1^b(f_\alpha) = \frac{b^{\alpha+1}-1}{\alpha+1}, \quad b > 1,\ \alpha \in \mathcal{Q},\ \alpha \neq -1.$$

And because the area is the limit of box-area sums from above and below, Definition 3.3.6 to follow (page 106) will say that it is in fact an integral,

$$\int_1^b f_\alpha = \frac{b^{\alpha+1}-1}{\alpha+1}, \quad b > 1,\ \alpha \in \mathcal{Q},\ \alpha \neq -1.$$

The case $\alpha = -1$, where we don't know that $\lim_n(S_n)$ and $\lim_n(T_n)$ exist, much less have a common value for which we have a formula, is more subtle. We will return to it shortly.

Our second tacit use of Proposition 3.3.2 was in computing the non-normalized power function integral, also in section 2.5. Given b and c with $1 \le b$ and $c > 0$, we defined the sequences

$$(\widetilde{S}_n) = c^{\alpha+1}(S_n)$$

and

$$(\widetilde{T}_n) = c^{\alpha+1}(T_n).$$

By the Constant Multiple Rule, and the fact that $\lim_n(S_n) = \lim_n(T_n) = \int_1^b f_\alpha$,

$$\lim_n(\widetilde{S}_n) = \lim_n(\widetilde{T}_n) = c^{\alpha+1}\int_1^b f_\alpha.$$

But by the geometry that led to $(\widetilde{S}_n)$ and $(\widetilde{T}_n)$, also

$$\begin{cases} \widetilde{S}_n \le \mathrm{Ar}_c^{bc}(f_\alpha) \le \widetilde{T}_n & \text{for all } n \in \mathcal{Z}_{\ge 1}, \text{ if } \alpha > 0, \\ \widetilde{T}_n \le \mathrm{Ar}_c^{bc}(f_\alpha) \le \widetilde{S}_n & \text{for all } n \in \mathcal{Z}_{\ge 1}, \text{ if } \alpha < 0. \end{cases}$$

And so it follows from Proposition 3.3.2 that

$$\mathrm{Ar}_c^{bc}(f_\alpha) = c^{\alpha+1}\int_1^b f_\alpha.$$

Furthermore, as the common limit of box area sums from above and below, the area will acquire the status of an integral in Definition 3.3.6, and so the previous equation rewrites, as in (2.6) (page 53),

$$\int_c^{bc} f_\alpha = c^{\alpha+1}\int_1^b f_\alpha.$$

This result holds for all $\alpha \in \mathcal{Q}$ such that $\lim_n(S_n) = \lim_n(T_n)$. For $\alpha \neq -1$, the limits were shown to be equal by the simple expedient of evaluating them. But for $\alpha = -1$ they are not yet established, and it turns out that they don't evaluate to anything yet in our ken. So in the case $\alpha = -1$, we need to argue that the limits exist and are equal even though we can't find a formula for them. This is the last point that this section will discuss.

Recall the verbal argument on page 53, which was made for all values of α:

> *The fact that* S_n *and* T_n *trap the area under the graph of* f_α *from* 1 *to* b *between them, and the facts that* $T_n = s^\alpha S_n$ *and* s^α *tends to* 1*, combine to show that* S_n *and* T_n *tend to the same limiting value, that value being the area.*

With sequence limit results in hand, we now can quantify the reasoning. For convenience, assume that $\alpha > 0$. Let

$$A = \mathrm{Ar}_1^b(f_\alpha)$$

denote the area under the graph of f_α from 1 to b, a constant. Then we have the following information:

(1) $S_n \le A \le T_n$ for all $n \in \mathcal{Z}_{\ge 1}$.
(2) $T_n = (b^\alpha)^{1/n} S_n$ for all $n \in \mathcal{Z}_{\ge 1}$.

From (1), then (2), and then (1) again,

$$0 \le T_n - S_n = ((b^\alpha)^{1/n} - 1)S_n \le ((b^\alpha)^{1/n} - 1)A \quad \text{for all } n \in \mathcal{Z}_{\ge 1}.$$

That is,

$$0 \le T_n - S_n \le ((b^\alpha)^{1/n} - 1)A \quad \text{for all } n \in \mathcal{Z}_{\ge 1}.$$

By various sequence limit rules, including the Squeezing Rule, it follows (exercise 3.3.1) that

$$\lim_n (T_n - S_n) \textit{ exists} \text{ and equals } 0. \tag{3.14}$$

But also from (1),

$$0 \le A - S_n \le T_n - S_n \quad \text{for all } n \in \mathcal{Z}_{\ge 1},$$

and so again by the Squeezing Rule, using (3.14),

$$\lim_n (A - S_n) \textit{ exists} \text{ and equals } 0. \tag{3.15}$$

Because A is constant, the Constant Sequence Rule says that also

$$\lim_n (A) \textit{ exists} \text{ and equals } A. \tag{3.16}$$

Now, note that (using square brackets rather than parentheses to group two numbers without overloading the sequence notation)

$$S_n = A - [A - S_n] \quad \text{for all } n \in \mathcal{Z}_{\ge 1},$$

and so the definition of the difference of two sequences gives

$$(S_n) = (A) - (A - S_n).$$

According to the Difference Rule, now (3.15) and (3.16) give

$$\lim_n (S_n) \textit{ exists} \text{ and equals } A = \mathrm{Ar}_1^b(f_\alpha).$$

And because

$$(T_n) = (T_n - S_n) + (S_n),$$

(3.14) and the previous display combine in turn via the Sum Rule to give

$$\lim_n (T_n) \textit{ exists} \text{ and equals } \mathrm{Ar}_1^b(f_\alpha).$$

The argument in italics is now fully quantified, for all values of α.

Note the finesse of the argument: deftly using the Squeezing Rule twice to show that auxiliary limits exist subtly but inexorably cornered our desired limits until they were forced to exist as well, and to be equal. And again, the reasoning was carried out with no recourse to explicit formulas, meaning that it should apply in contexts beyond the power function in particular. The remainder of this chapter will expand its scope.

Exercise

3.3.1. Show that (3.14) follows from the display immediately preceding it.

3.3.2 Definition of Integrability

Definition 3.3.3 (Lower Sum, Upper Sum). *Let* a *and* b *be real numbers with* $a \leq b$. *Let* M *be a nonnegative real number. Consider a function*

$$f : [a, b] \longrightarrow [0, M].$$

The region under the graph of f*, a subset of the plane, is*

$$R = \{(x, y) \in \mathcal{R}^2 : a \leq x \leq b,\ 0 \leq y \leq f(x)\}.$$

and it has an area,

$$\mathrm{Ar}_a^b(f) = \mathrm{Ar}(R).$$

Suppose that a number S *is a sum of finitely many box-areas, where the base of each box lies on the* x*-axis, the top of each box (at least as high as the base) lies under the graph of* f*, the overlap of any two boxes is at most a vertical line segment, and the bases combine to cover the* x*-axis from* a *to* b. *Then* S *is a* **lower sum for** $\mathrm{Ar}_a^b(f)$. *Suppose that a number* T *is a sum of finitely many box-areas, where the boxes satisfy the same conditions except that their tops lie over the graph of* f. *Then* T *is an* **upper sum for** $\mathrm{Ar}_a^b(f)$.

We saw lower sums and upper sums throughout the integration of the rational power function. The sums S_n were lower sums and the sums T_n were upper sums only for $\alpha > 0$; unfortunately, the S_n were upper sums and the T_n were lower sums for $\alpha < 0$, but this is only a notational irritant of no consequence.

Retaining the terminology of the definition, because any lower sum for $\mathrm{Ar}_a^b(f) = \mathrm{Ar}(R)$ is the area of a polygon that is a subset of R, and any upper sum for $\mathrm{Ar}_a^b(f)$ is the area of a polygon that is a superset of R, the following result is automatic from the basic properties of area.

Proposition 3.3.4 (Basic Property of Lower and Upper Sums). *Let* $a \leq b$*, and let* $M \geq 0$. *Consider a function*

$$f : [a, b] \longrightarrow [0, M].$$

Let S *be any lower sum for* $\mathrm{Ar}_a^b(f)$*, and let* T *be any upper sum for* $\mathrm{Ar}_a^b(f)$. *Then*

$$S \leq \mathrm{Ar}_a^b(f) \leq T.$$

Now we can generalize the recent argument that a minimal good property of lower and upper sums, that the limit of their differences is zero, has further good consequences, that the lower and the upper sums themselves have limits, that the two limits are equal, and that they equal the area.

Proposition 3.3.5 (Bootstrap Result for Lower and Upper Sums). *Let* $a \leq b$*, and let* $M \geq 0$*. Consider a function*

$$f : [a, b] \longrightarrow [0, M].$$

Suppose that a sequence (S_n) *of lower sums for* $\mathrm{Ar}_a^b(f)$ *and a sequence* (T_n) *of upper sums for* $\mathrm{Ar}_a^b(f)$ *satisfy the condition*

$$\lim_n (T_n - S_n) \text{ exists and equals } 0.$$

Then $\lim_n(S_n)$ *and* $\lim_n(T_n)$ *both exist, and*

$$\lim_n(S_n) = \lim_n(T_n) = \mathrm{Ar}_a^b(f).$$

The proof (exercise 3.3.2) is similar to the argument recently given in the special case of the power function. It requires the Squeezing Rule once to make a limit exist, and then basic and progenitive results to reach the desired conclusions.

Definition 3.3.6 (Integral). *Let* $a \leq b$*, and let* $M \geq 0$*. Consider a function*

$$f : [a, b] \longrightarrow [0, M].$$

If there exist a sequence (S_n) *of lower sums for* $\mathrm{Ar}_a^b(f)$*, and a sequence* (T_n) *of upper sums for* $\mathrm{Ar}_a^b(f)$*, such that*

$$\lim_n (T_n - S_n) \text{ exists and equals } 0$$

then f *is* **integrable** *from* a *to* b*. The* **integral** *of* f *from* a *to* b *is the area under its graph,*

$$\int_a^b f = \mathrm{Ar}_a^b(f).$$

Equivalently,

$$\int_a^b f = \lim_n(S_n) = \lim_n(T_n),$$

because by Proposition 3.3.5 both limits exist and equal $\mathrm{Ar}_a^b(f)$*.*

Note how neatly the definitions and propositions of this section quantitatively capture our earlier description of the integral in natural language, displayed in italics on page 50.

Certainly, if there exist a sequence (S_n) of lower sums for $\mathrm{Ar}_a^b(f)$ and a sequence (T_n) of upper sums for $\mathrm{Ar}_a^b(f)$ such that

$$\lim_n(S_n) \text{ and } \lim_n(T_n) \text{ both exist, and they are equal}$$

then f is integrable from a to b. The point of Proposition 3.3.5 is that these conditions follow from the seemingly-weaker conditions required in Definition 3.3.6. But it is perfectly fine to establish these conditions instead.

To review some of the ideas, again let $f : [a, b] \longrightarrow [0, M]$ be a function.

- *Does* $\mathrm{Ar}_a^b(f)$ *exist?* Yes, always. The reader should be aware that many calculus courses treat *all* existence issues as obvious, perhaps not even raising them, whereas many beginning real analysis courses derive existence results from a property of the real number system called *completeness*. In contrast to both of these approaches, our method is to invoke the existence of area functions but then derive further consequences of the invocation carefully.
- *Are there sequences* (S_n) *and* (T_n) *of lower and upper sums for* $\mathrm{Ar}_a^b(f)$ *both with* $\mathrm{Ar}_a^b(f)$ *as their limit?* Sometimes. The Bootstrapping Result shows that there are such sequences if there are sequences of lower and upper sums such that $\lim(T_n - S_n) = 0$, and the previous paragraph (starting *Certainly...*) observed that the converse holds as well. Under these circumstances, we view f as integrable. Thus, integrability means not that the area exists, but that the area is the limit of suitable box-area sums.
- *When* f *is integrable, can we put the common limit of* (S_n) *and* (T_n) *into some convenient form, such as an expression in terms of functions that we already know?* Not always. In the case of the power function f_α, we can do so for all $\alpha \neq -1$ but not for $\alpha = -1$.
- *When* f *is integrable but the area under its graph does not take a convenient form that we already understand, what good does the integrability do us?* We can still study the integral as a limit in order to learn more about its properties. To know a function's properties *is* to understand it. For example, in chapter 5 we will study the logarithm as an integral.

Exercise

3.3.2. Prove Proposition 3.3.5.

3.3.3 Monotonicity and Integrability

Definition 3.3.7 (Monotonic Function). *Let $a \le b$, and let $M \ge 0$. Consider a function*

$$f : [a, b] \longrightarrow [0, M].$$

The function f is **increasing** *if for all $x_1, x_2 \in [a, b]$ with $x_2 > x_1$, also $f(x_2) \ge f(x_1)$. The function f is* **decreasing** *if for all $x_1, x_2 \in [a, b]$ with $x_2 > x_1$, also $f(x_2) \le f(x_1)$. The function f is* **monotonic** *if it is increasing or it is decreasing.*

Thus a function is increasing if its graph, traversed from left to right, is everywhere rising or level, never falling. The *or level* distinguishes between an increasing function and a *strictly* increasing function as discussed earlier. And similarly, a function is decreasing if its graph is everywhere falling or level, never rising.

Theorem 3.3.8 (Monotonic Functions are Integrable). *Let $a \le b$, and let $M \ge 0$. Let*

$$f : [a, b] \longrightarrow [0, M]$$

be monotonic. Then f is integrable.

Proof. Now we use a uniform partition rather than a geometric one. The relevant partition-widths are

$$\delta_n = \frac{b-a}{n} \quad \text{for each } n \in \mathcal{Z}_{\ge 1},$$

and the partition points are

$$x_0 = a, \quad x_1 = a + \delta_n, \quad x_2 = a + 2\delta_n, \quad \ldots, \quad x_n = a + n\delta_n = b.$$

Assume that f is increasing. Then (exercise 3.3.3 (a)) the quantity

$$S_n = \delta_n(f(x_0) + f(x_1) + \cdots f(x_{n-1}))$$

is a lower sum for $\mathrm{Ar}_a^b(f)$, and the quantity

$$T_n = \delta_n(f(x_1) + f(x_2) + \cdots f(x_n))$$

is an upper sum. In their difference, nearly all the terms cancel,

$$T_n - S_n = \delta_n(f(x_n) - f(x_0)),$$

and because $x_0 = a$ and $x_n = b$, their difference is in fact

$$T_n - S_n = \delta_n(f(b) - f(a)).$$

Consequently, by various sequence limit rules (exercise 3.3.3 (b)),

$$\lim_n (T_n - S_n) \text{ exists and equals } 0.$$

This shows that f is integrable from a to b, as desired.

The proof when f is decreasing is virtually identical. □

Because the power function is monotonic, Theorem 3.3.8 encompasses it, even though the theorem's proof used uniform lower and upper sums rather than the geometric ones that we used earlier to analyze the power function. The uniform lower and upper sums used to prove the theorem do not readily *compute* the integral of the power function, but they do reconfirm its *existence*.

Exercise

3.3.3. In the proof of Theorem 3.3.8:

(a) Explain why S_n is a lower sum for $\mathrm{Ar}_a^b(f)$ and T_n is an upper sum.

(b) Explain why $\lim_n (T_n - S_n)$ exists and equals 0.

3.3.4 A Basic Property of the Integral

Proposition 3.3.9 (Horizontal Additivity of the Integral). *Let a, b, and c be real numbers with $a \le b \le c$. Let $M \ge 0$ be a positive real number. Let*

$$f : [a, c] \longrightarrow [0, M]$$

be a function. Then

$$\int_a^c f \text{ exists} \Longleftrightarrow \int_a^b f \text{ exists and } \int_b^c f \text{ exists},$$

and when the various integrals exist,

$$\int_a^c f = \int_a^b f + \int_b^c f.$$

Proof. The region under the graph of f from a to c is a bounded subset of the plane, and so it has an area. Similarly for the regions from a to b and from b to c. By the fact that area has basic sensible properties,

$$\mathrm{Ar}_a^c(f) = \mathrm{Ar}_a^b(f) + \mathrm{Ar}_b^c(f).$$

Granting momentarily that the three integrals in the proposition exist, it follows that

$$\int_a^c f = \mathrm{Ar}_a^c(f) = \mathrm{Ar}_a^b(f) + \mathrm{Ar}_b^c(f) = \int_a^b f + \int_b^c f.$$

This gives the last equality in the proposition. Thus what needs to be proved is that if $\int_a^c f$ exists then so do $\int_a^b f$ and $\int_b^c f$, and conversely.

Suppose that $\int_a^c f$ exists. This means that there are sequences of lower sums S_n for $\mathrm{Ar}_a^c(f)$, and sequences of upper sums T_n for $\mathrm{Ar}_a^c(f)$, such that

$$\lim_n (T_n - S_n) = 0.$$

For each n, if the boxes whose areas sum to S_n include a box whose base straddles the intermediate point b, then subdivide that box into two by adding a vertical line segment at b. This has no effect on S_n because the areas of two subboxes just created total the area of the box that was subdivided. And similarly for the upper sums T_n. That is, we may assume that each S_n and each T_n is the sum of box-areas for boxes whose bases comprise the x-axis from a to b and then more boxes whose bases comprise the x-axis from b to c. The sums decompose accordingly,

$$S_n = S'_n + S''_n \text{ and } T_n = T'_n + T''_n \quad \text{for } n \in \mathcal{Z}_{\geq 1}.$$

Here each S'_n is a lower sum for $\mathrm{Ar}_a^b(f)$, each S''_n is a lower sum for $\mathrm{Ar}_b^c(f)$, and similarly for T'_n and T''_n, so that $S'_n \leq T'_n$ and $S''_n \leq T''_n$. Because

$$T_n - S_n = (T'_n - S'_n) + (T''_n - S''_n),$$

if follows that

$$0 \leq T'_n - S'_n \leq T_n - S_n \quad \text{and} \quad 0 \leq T''_n - S''_n \leq T_n - S_n,$$

and so by the Squeezing Rule for sequences,

$$\lim_n (T'_n - S'_n) = 0. \quad \text{and} \quad \lim_n (T''_n - S''_n) = 0.$$

Thus $\int_a^b f$ and $\int_b^c f$ exist.

Now suppose that $\int_a^b f$ and $\int_b^c f$ exist. This means that there are sequences of lower sums S'_n for $\mathrm{Ar}_a^b(f)$, sequences of lower sums S''_n for $\mathrm{Ar}_b^c(f)$, sequences of upper sums T'_n for $\mathrm{Ar}_a^b(f)$, and sequences of upper sums T''_n for $\mathrm{Ar}_b^c(f)$, such that

$$\lim_n (T'_n - S'_n) = 0 \quad \text{and} \quad \lim_n (T''_n - S''_n) = 0.$$

For each n, the sum $S_n = S'_n + S''_n$ is a lower sum for $\mathrm{Ar}_a^c(f)$ and the sum $T_n = T'_n + T''_n$ is an upper sum for $\mathrm{Ar}_a^c(f)$. Because

$$T_n - S_n = (T'_n - S'_n) + (T''_n - S''_n),$$

if follows that

$$\lim (T_n - S_n) = 0.$$

Thus $\int_a^c f$ exists. □

3.3.5 Piecewise Monotonicity and Integrability

Definition 3.3.10 (Piecewise Monotonic Function). *Let* a *and* b *be real numbers with* $a \le b$*, and let* $M \ge 0$ *be a real number. A function*

$$f : [a, b] \longrightarrow [0, M]$$

is called **piecewise monotonic** *if there is a partition of* $[a, b]$*,*

$$a = x_0 < x_1 < \cdots < x_n = b,$$

such that f *is monotonic on each interval* $[x_{j-1}, x_j]$ *for* $j = 1, \ldots, n$*.*

For example, the absolute value function is piecewise monotonic on $[-1, 1]$, but it is not monotonic there.

Proposition 3.3.11 (Piecewise Monotonic Functions are Integrable). *Let* a *and* b *be real numbers with* $a \le b$*, and let* $M \ge 0$ *be a real number. Any piecewise monotonic function*

$$f : [a, b] \longrightarrow [0, M]$$

is integrable from a *to* b*.*

This follows from Theorem 3.3.8 and Proposition 3.3.9.

Example 3.3.12. We give two functions, neither of which is piecewise monotonic, but one of which is integrable. For any nonnegative integers k and ℓ, define two sets of points, both subsets of the interval $[0, 1]$,

$$P_k = \left\{ \frac{0}{2^k}, \frac{1}{2^k}, \frac{2}{2^k}, \ldots, \frac{2^k - 1}{2^k}, \frac{2^k}{2^k} \right\},$$
$$Q_\ell = \left\{ \frac{0}{3^\ell}, \frac{1}{3^\ell}, \frac{2}{3^\ell}, \ldots, \frac{3^\ell - 1}{3^\ell}, \frac{3^\ell}{3^\ell} \right\}.$$

That is, the points of P_k are spaced across $[0, 1]$ in uniform steps of size $1/2^k$, and similarly for Q_ℓ with step-size $1/3^\ell$. For a point to lie simultaneously in some P_k and some Q_ℓ requires

$$\frac{a}{2^k} = \frac{b}{3^\ell}, \quad 0 \le a \le 2^k, \quad 0 \le b \le 3^\ell,$$

or

$$3^\ell a = 2^k b, \quad 0 \le a \le 2^k, \quad 0 \le b \le 3^\ell.$$

Because positive integers factor *uniquely* into prime powers, the only solutions are $a = b = 0$, i.e., the common point is the left endpoint 0, and $a = 2^k$, $b =$

3^ℓ, i.e., the common point is the right endpoint 1. Excluding the endpoints, there is no overlap among the sets P_k and Q_ℓ.

Now consider two functions,

$$f : [0,1] \longrightarrow [0,1],$$

where

$$f(x) = \begin{cases} 1 & \text{if } x \in P_k \text{ for some } k, \\ 0 & \text{if } x \notin P_k \text{ for all } k, \end{cases}$$

and

$$g : [0,1] \longrightarrow [0,1],$$

where

$$g(x) = \begin{cases} 1/2^k & \text{if } k \text{ is the smallest integer such that } x \in P_k, \\ 0 & \text{if } x \notin P_k \text{ for all } k. \end{cases}$$

An approximation of the graph of f is shown in figure 3.5. Rather than check whether a point $x \in [0,1]$ lies in P_k for all $k \in \mathcal{Z}_{\geq 0}$, the figure was generated by checking only up to $k = 6$. A similar approximation of the graph of g is shown in figure 3.6. The figure shows why g is called the *ruler function*.

To see that neither f nor g is piecewise monotonic, note that any subinterval $[a,b]$ of $[0,1]$ having positive width contains points $i/2^k$ and $(i+1)/2^k$ consecutive in P_k for some k, and then a point $j/3^\ell$ of Q_ℓ for some ℓ such that $i/2^k < j/3^\ell < (i+1)/2^k$. Because $f(i/2^k) > 0$ and $f((i+1)/2^k) > 0$ while $f(j/3^\ell) = 0$, f is not monotonic on $[a,b]$. And similarly for g.

Because any subinterval $[a,b]$ of $[0,1]$ having positive width contains a point $i/2^k$ of P_k for some k, and a point $j/3^\ell$ of Q_ℓ for some ℓ, it follows that every lower sum S and every upper sum T for $\mathrm{Ar}_0^1(f)$ must satisfy

$$S = 0, \quad T \geq 1.$$

Therefore there are no sequences (S_n) of lower sums and (T_n) of upper sums satisfying the condition $\lim_n(T_n - S_n) = 0$ that is necessary for f to be integrable (see Definition 3.3.6). That is,

$$\int_0^1 f \text{ does not exist.}$$

On the other hand, g *is* integrable. Every lower sum S for $\mathrm{Ar}_0^1(g)$ is 0, and so by Definition 3.3.6, the question is whether a sequence (T_n) of upper sums has limit 0. It does. The idea is to cover finitely many high spikes efficiently with very narrow boxes, so that covering the infinitely many remaining low

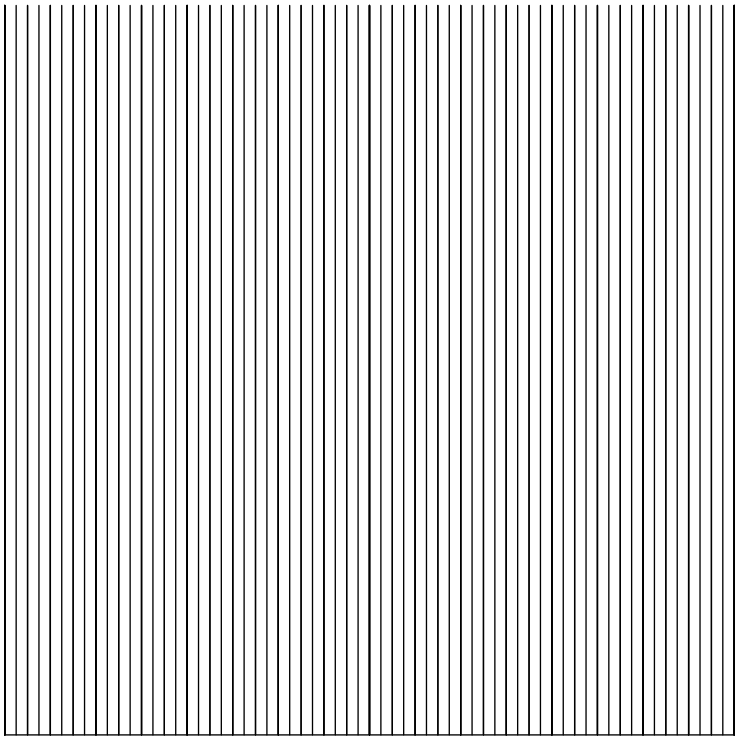


Figure 3.5. Approximation of the function that is 1 at values $i/2^k$

spikes inefficiently still produces a small upper sum. Quantitatively, given any $\varepsilon > 0$, there is a positive integer k such that

$$\frac{1}{2^{k+1}} < \frac{\varepsilon}{2}.$$

Let $w = \varepsilon/(2(2^k + 1))$, a positive value. Then

$$(2^k + 1)w + \frac{1}{2^{k+1}} < \varepsilon.$$

Cover the 2^k+1 spikes over the points $0/2^k$, $1/2^k$, $\ldots$, $2^k/2^k$ of P_k with boxes of width w and height 1. Cover the remainder of the graph of g with boxes of height $1/2^{k+1}$ and total width less than 1. This gives an upper sum T such that

$$T < (2^k + 1)w + \frac{1}{2^{k+1}} < \varepsilon.$$

(Figure 3.7 shows the rectangles for such an upper sum T that is slightly bigger than $1/8$.) Because ε is arbitrarily, we can create a sequence (T_n) of

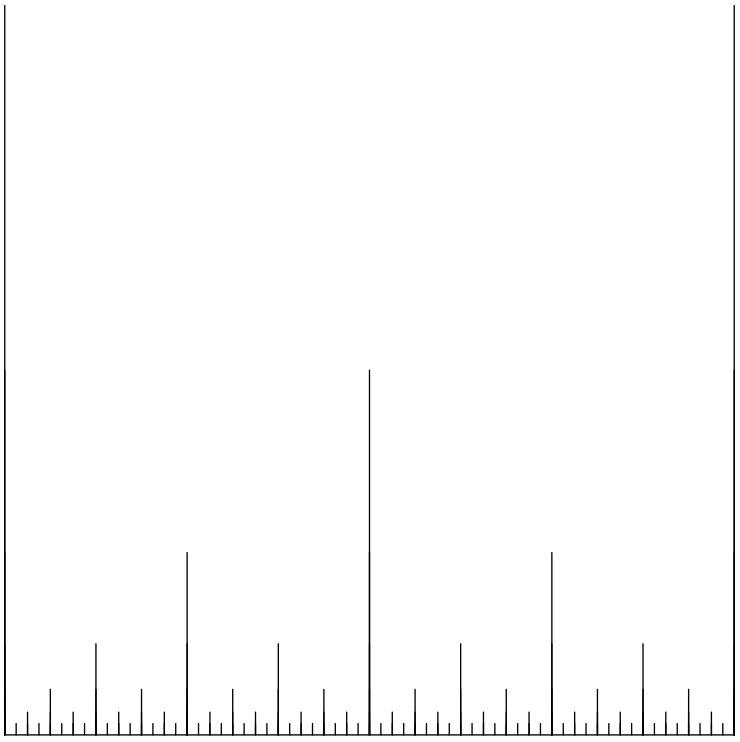


Figure 3.6. Approximation of the function that is $1/2^k$ at values $i/2^k$

such upper sums with limit 0. Therefore,

$$\int_0^1 g = 0.$$

Exercises

3.3.4. Consider two functions,

$$f : [0,2] \longrightarrow [0,1], \quad f(x) = \begin{cases} x & \text{if } 0 \le x \le 1, \\ x-1 & \text{if } 1 < x \le 2, \end{cases}$$

and

$$g : [0,2] \longrightarrow [0,1], \quad g(x) = \begin{cases} x & \text{if } 0 \le x < 1, \\ x-1 & \text{if } 1 \le x \le 2. \end{cases}$$

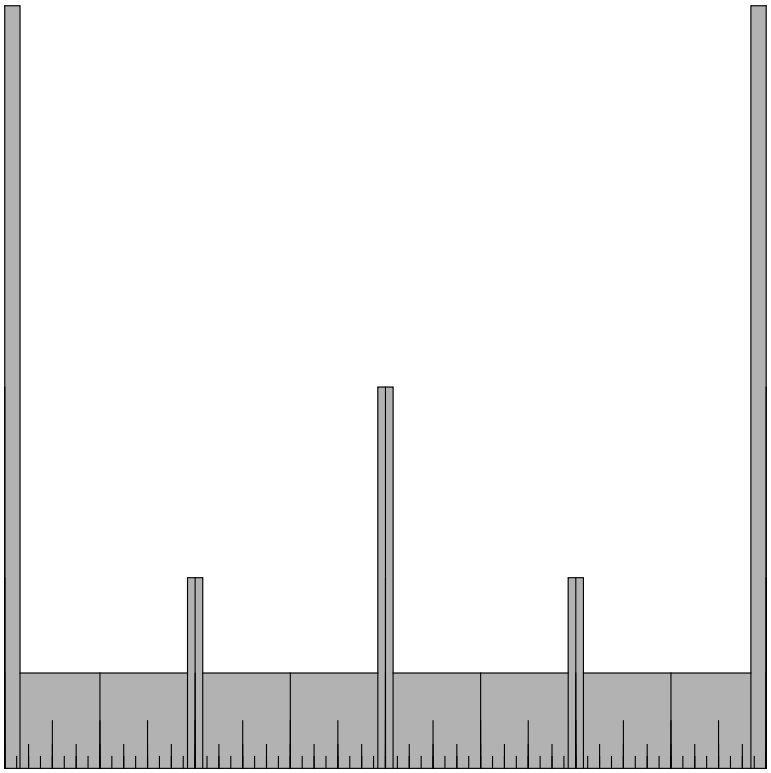


Figure 3.7. Small upper sum for the area under the ruler function

(a) Graph f and g side by side, in a way that shows the difference between them.

(b) Is f piecewise monotonic? Is g? Is any function

$$h : [0,2] \longrightarrow \mathcal{R}, \quad h(x) = \begin{cases} x & \text{if } 0 \le x < 1, \\ c & \text{if } x = 1, \\ x-1 & \text{if } 1 < x \le 2 \end{cases}$$

piecewise monotonic?

(c) Figure 3.8 shows two arrangements of boxes. Explain why the sum of box-areas arising from one arrangement is a lower sum for $\mathrm{Ar}_0^2(f)$ or for $\mathrm{Ar}_0^2(g)$, and the sum of box-areas arising from the other arrangement is an upper sum for $\mathrm{Ar}_0^2(f)$ or for $\mathrm{Ar}_0^2(g)$, but it is not the case that the figure shows a lower–upper sum pair for either f or g.

(d) Draw two more arrangements of boxes, similar to the figure, so that the figure and your picture give you a lower–upper sum pair for f and a

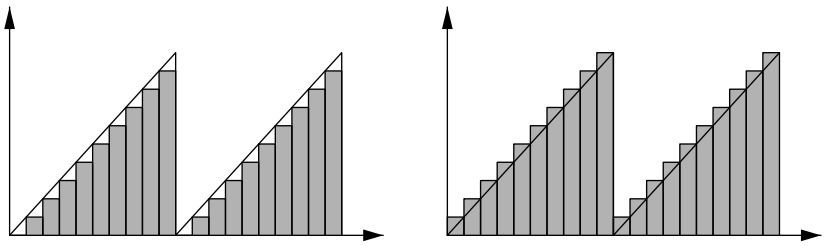


Figure 3.8. Box-arrangements

lower–upper sum pair for g. Using these figures, discuss qualitatively whether f and g are integrable, and if so, whether their integrals are equal.

3.3.5. Show by example that if $f, g : [a, b] \longrightarrow [0, M]$ are piecewise monotonic then their sum $f+g : [a, b] \longrightarrow [0, 2M]$ need not again be piecewise monotonic. (An explanation that involves some mixture of formulas, pictures, and words is fine.)

3.3.6 Progenitive Integral Rules

Proposition 3.3.13 (Progenitive Integral Rules). *Consider two integrable functions*

$$f : [a, b] \longrightarrow [0, M], \qquad \tilde{f} : [a, b] \longrightarrow [0, \widetilde{M}].$$

Then the function

$$f + \tilde{f} : [a, b] \longrightarrow [0, M + \widetilde{M}], \quad (f + \tilde{f})(x) = f(x) + \tilde{f}(x)$$

is integrable, and

$$\int_a^b (f + \tilde{f}) = \int_a^b f + \int_a^b \tilde{f}.$$

Let $c \in \mathcal{R}_{\geq 0}$ be a nonnegative real number. Then the function

$$cf : [a, b] \longrightarrow [0, cM], \quad (cf)(x) = c \cdot f(x)$$

is integrable, and

$$\int_a^b (cf) = c \int_a^b f.$$

Proof. There exist a sequence (S_n) of lower sums for $\mathrm{Ar}_a^b(f)$ and a sequence of (T_n) of upper sums for $\mathrm{Ar}_a^b(f)$ such that

$$\lim_n (T_n - S_n) = 0.$$

And there exist similar sequences $(\widetilde{S}_n)$ and $(\widetilde{T}_n)$ for $\mathrm{Ar}_a^b(\tilde{f})$ with

$$\lim_n (\widetilde{T}_n - \widetilde{S}_n) = 0.$$

Consequently, $(S_n + \widetilde{S}_n)$ is a sequence of lower sums for $\mathrm{Ar}_a^b(f + \tilde{f})$ (exercise 3.3.6) and $(T_n + \widetilde{T}_n)$ is a sequence of upper sums for $\mathrm{Ar}_a^b(f + \tilde{f})$, and by the Sum Rule for sequences,

$$\begin{aligned}\lim_n ((T_n + \widetilde{T}_n) - (S_n + \widetilde{S}_n)) &= \lim_n ((T_n - S_n) + (\widetilde{T}_n - \widetilde{S}_n)) \\ &= \lim_n (T_n - S_n) + \lim_n (\widetilde{T}_n - \widetilde{S}_n) \\ &= 0 + 0 = 0.\end{aligned}$$

Thus $\int_a^b (f + \tilde{f})$ exists, and its value is

$$\int_a^b (f + \tilde{f}) = \lim_n (S_n + \widetilde{S}_n) = \lim_n (S_n) + \lim_n (\widetilde{S}_n) = \int_a^b f + \int_a^b \tilde{f}.$$

The second part of the proposition is proved similarly (exercise 3.3.7). □

In connection with Proposition 3.3.13, it deserves note that the formula

$$\mathrm{Ar}_a^b(f + \tilde{f}) = \mathrm{Ar}_a^b(f) + \mathrm{Ar}_a^b(\tilde{f})$$

is not geometrically immediate. The problem is that the region under the graph of $f+\tilde{f}$ does not naturally decompose into two pieces with one congruent to the area under the graph of f and the other similar but for $\tilde{f}$.

Proposition 3.3.14 (Inequality Rule for Integrals). *Consider two integrable functions*

$$f, g : [a, b] \longrightarrow [0, M]$$

such that

$$f \leq g,$$

meaning that $f(x) \leq g(x)$ *for all* $x \in [a, b]$. *Then*

$$\int_a^b f \leq \int_a^b g.$$

Proof. This follows from the fact that area has sensible properties, because

$$\int_a^b f = \mathrm{Ar}_a^b(f) \leq \mathrm{Ar}_a^b(g) = \int_a^b g.$$

□

Exercises

3.3.6. The proof of Proposition 3.3.13 tacitly cites the following assertion: *Consider two functions*

$$f : [a, b] \longrightarrow [0, M], \qquad \tilde{f} : [a, b] \longrightarrow [0, \widetilde{M}],$$

and consider their sum,

$$f + \tilde{f} : [a, b] \longrightarrow [0, M + \widetilde{M}], \quad (f + \tilde{f})(x) = f(x) + \tilde{f}(x).$$

If S *is a lower sum for* $\mathrm{Ar}_a^b(f)$*, and* $\widetilde{S}$ *is a lower sum for* $\mathrm{Ar}_a^b(\tilde{f})$*, then* $S + \widetilde{S}$ *is a lower sum for* $\mathrm{Ar}_a^b(f + \tilde{f})$. While this assertion is correct, it is not quite automatic.

(a) For convenience, let $a = 0$ and $b = 1$, and let $f(x) = x$ and $\tilde{f}(x) = x$. Draw the graph of f, and then draw three boxes whose areas add up to a lower sum S for $\mathrm{Ar}_0^1(f)$. Separately, draw the graph of $\tilde{f}$, and then draw four boxes whose areas add up to a lower sum $\widetilde{S}$ for $\mathrm{Ar}_0^1(\tilde{f})$. Make the breakpoints that determine the bases of the four boxes be different from those that determine the three boxes from a moment ago (except for the breakpoints 0 and 1, of course).

(b) Draw a graph that shows the function $f + \tilde{f}$. Although the seven boxes from part (a) have total area $S + \widetilde{S}$, it may not be possible to show that this total area is a lower sum for $f + \tilde{f}$ by fitting the seven boxes nonoverlappingly under the graph of $f + \tilde{f}$, stacking the four boxes on top of the three. Use natural language and a picture to demonstrate this problem clearly, explaining its cause.

(c) Explain how to fix the problem, illustrating the solution with another picture.

3.3.7. Prove the second part of Proposition 3.3.13.

3.4 Summary

The notion of a sequence limit leads to a more precise understanding of the integral than we could attain in chapters 1 and 2. In the next chapter, the related notion of a *function limit* will similarly clarify the derivative.

4

Function Limits and the Derivative

A function f has limit ℓ at the point x if its output-values $f(s)$ approach ℓ as its input-values s approach x *continuously*. To say that s approaches x continuously is to say that s approaches x sequentially in any way whatsoever, except that s should never actually reach x. The function limit will thus be a common value of sequence limits, the limits of the output-sequences $(f(s_n))$ corresponding to all suitable input-sequences (s_n) approaching x. A bit more fancifully, to say f has limit ℓ as its input *glides* to x is to say that f has limit ℓ as its input *hops* to x, no matter how the hopping occurs. This chapter begins by defining function limits and establishing some of their properties.

After the examples of the first two chapters, the theory has amassed during the most recent chapter and will continue to do so during this one. So the reader should periodically step back from details in order to appreciate the cumulative arrangement of the ideas. Once the previous chapter's definition of sequence limit was in place, it led to basic results and progenitive results. This chapter's definition of function limit will be phrased in terms of the definition of sequence limit, and then it too will lead to basic and progenitive results, based on their counterparts for sequences. All of this is carried out in section 4.1. With function limits in place, the derivative can be be defined as a particular function limit. Basic derivative results and progenitive derivative results thus follow from corresponding function limit results, as shown in section 4.2.

So far, the only specific derivative that we know is that of the power function, but we will compute other specific derivatives in the chapters to follow.

4.1 The Limit of a Function

4.1.1 Definition of Function Limit

As just mentioned, the suitable input-sequences for the definition of function limit are those sequences that tend to a point but never reach it. For convenience, we name the phenomenon.

Definition 4.1.1 (Approaches, Approachable). *Let* (s_n) *be a real sequence, and let* x *be a real number. Then* (s_n) **approaches** x *if*

$$\lim_n(s_n) = x \quad \textit{but} \quad s_n \neq x \textit{ for each index } n.$$

Let A *be a set of real numbers, and let* x *be a real number. Then* x *is* **approachable from** A *if some sequence* (s_n) *in* A *approaches* x.

Whether a point x is approachable from a set A is in general independent of whether x is an element of A. That is, there are situations where x is approachable from A and lies in A, where x is approachable from A but does not lie in A, where x is not approachable from A but lies in A, and where x is not approachable from A and does not lie in A. Exercise 4.1.1 asks for examples.

Definition 4.1.2 (Limit of a Function). *Let* A *be a subset of* $\mathcal{R}$*, and let*

$$f : A \longrightarrow \mathcal{R}$$

be a function. Let $x \in \mathcal{R}$*, and let* $\ell \in \mathcal{R}$*. Then* f **has limit** ℓ **as** s **goes to** x*, notated*

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

if

(1) *The point* x *is approachable from* A.
(2) *For every sequence* (s_n) *in* A *that approaches* x, $\lim_n(f(s_n)) = \ell$.

> **To be clear about the notation, observe that the symbol-string "$\lim_n$" refers to a *sequence* limit, whereas "$\lim_{s \to x}$" refers to a *function* limit.**

As in the introduction to this chapter, Definition 4.1.2 encodes the notion of the input s approaching x continuously as encompassing all ways that s can approach x sequentially, and the function limit is the common value of all the corresponding output-sequence limits, if conditions are suitable and a common value exists. In natural language:

The function limit of f *at* x *is the output-value that the behavior of* f **near** x *suggests that* f *should take* **at** x.

Note that this natural language description uses the symbols f and x but not s. The reader is cautioned that if a calculation of $\lim_{s\to x} f(s)$ in some particular instance seems to give an answer involving the symbol s then something has gone wrong.

On the other hand, although Definition 4.1.2 of $\lim_{s\to x} f(s)$ captures the value that $f(s)$ tends to as s tends to x, the definition makes no reference to $f(x)$ itself. Indeed, x need not even be in the domain of f. That is:

The function limit of f *at* x *is* **independent** *of the value of* f *at* x, *and even of whether* $f(x)$ *exists.*

So there is a range of possibilities:

$$\text{If } \lim_{s\to x} f(s) \text{ exists then } \begin{cases} f(x) \text{ could exist and equal } \lim_{s\to x} f(s), \\ f(x) \text{ could exist and not equal } \lim_{s\to x} f(s), \\ f(x) \text{ could fail to exist.} \end{cases}$$

The definition's insistence on sequences (s_n) that approach x but never reach it prevents any accidental reference to $f(x)$.

The definition is of most interest to us when the domain A of f excludes a point x where we want to know what value f *should* take. For example, let x be any real number, let

$$\mathcal{R}_{\neq x} = \{s \in \mathcal{R} : s \neq x\},$$

and consider the function

$$f : \mathcal{R}_{\neq x} \longrightarrow \mathcal{R}, \qquad f(s) = \frac{s^2 - x^2}{s - x}.$$

Then also this function is

$$f : \mathcal{R}_{\neq x} \longrightarrow \mathcal{R}, \qquad f(s) = s + x,$$

but even though the formula $s + x$ is sensible for $s = x$, the function f is not defined there. (Figure 4.1 shows the graph of f.) Nonetheless, x is approachable from $\mathcal{R}_{\neq x}$, e.g., by the sequence $(s_n) = (x + 1/n)$, and so the first condition of Definition 4.1.2 is met. Furthermore, for any sequence (s_n) that approaches x we have $\lim_n(s_n) = x$ by Definition 4.1.1, and so the Sum Rule rule for sequences and the Constant Sequence Rule combine to give

$$\lim_n(f(s_n)) = \lim_n(s_n + x) = 2x.$$

Thus the second condition of Definition 4.1.2 is also satisfied (with $\ell = 2x$), and so we have established a function limit,

$$\lim_{s \to x} f(s) = 2x.$$

Visually, the idea is that the limit machinery has plugged the gap in figure 4.1. This little argument has essentially repeated the derivation of the tangent slope of the parabola in section 1.3, but now using the more precise language at hand to buttress the ideas. Note that the calculated value $2x$ of $\lim_{s \to x} f(s)$ does not contain the symbol s, as remarked after Definition 4.1.2.

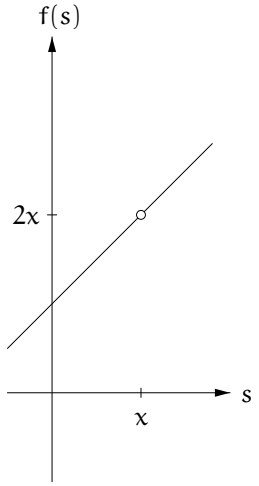


Figure 4.1. The graph of a difference-quotient function

The next result illustrates the idea that universalizing over all sequence-approaches captures continuous approach.

Proposition 4.1.3 (Persistence of Inequality). *Let A be a subset of $\mathcal{R}$, let $f : A \longrightarrow \mathcal{R}$ be a function, and let x be a point of A that is approachable from A. Suppose that*

$$\lim_{s \to x} f(s) > 0 \quad \textit{and} \quad f(x) > 0.$$

Then for all $s \in A$ close enough to x, also $f(s) > 0$.

Proof. If $f(s) > 0$ for all $s \in A$ such that $|s - x| < 1$ then we are done. Otherwise there exists some $s_1 \in A$ such that

$$|s_1 - x| < 1 \quad \text{and} \quad f(s_1) \leq 0.$$

And $s_1 \neq x$ because $f(x) > 0$. If $f(s) > 0$ for all $s \in A$ such that $|s - x| < |s_1 - x|/2$ then we are done. Otherwise, because $|s_1 - x|/2 < 1/2$, there exists some $s_2 \in A$ such that

$$|s_2 - x| < 1/2 \quad \text{and} \quad f(s_2) \leq 0.$$

And $s_2 \neq x$ because $f(x) > 0$. If $f(s) > 0$ for all $s \in A$ such that $|s - x| < |s_2 - x|/2$ then we are done. Otherwise, because $|s_2 - x|/2 < 1/4$, there exists some $s_3 \in A$ such that

$$|s_3 - x| < 1/4 \quad \text{and} \quad f(s_3) \leq 0.$$

And $s_3 \neq x$ because $f(x) > 0$. Continue in this fashion. Unless the process proves the proposition after finitely many steps, it produces a sequence (s_n) that approaches x but (because $f(s_n) \leq 0$ for each n) fails to satisfy the condition $\lim_n(f(s_n)) > 0$. This contradicts the hypothesis that $\lim_{s\to x} f(s) > 0$, and so the process must prove the proposition after finitely many steps. □

We end this section with one more remark. Let f be a function and let ℓ be a real number. Because

$$|f(s) - \ell| = \big||f(s) - \ell| - 0\big|,$$

it follows immediately that

$$\lim_{s\to x} f(s) \text{ exists and equals } \ell \iff \lim_{s\to x} |f(s) - \ell| \text{ exists and equals } 0. \tag{4.1}$$

Like the Strong Approximation Lemma (page 71), this principle can be handy to have available in isolated form for the sake of smoothing out the endgames of arguments. As a special case,

$$\lim_{s\to x} f(s) \text{ exists and equals } 0 \iff \lim_{s\to x} |f(s)| \text{ exists and equals } 0. \tag{4.2}$$

Exercises

4.1.1. (a) Find a subset A of the real numbers and a real number x such that x is approachable from A and lies in A.

(b) Find a subset A of the real numbers and a real number x such that x is approachable from A but does not lie in A.

(c) Find a subset A of the real numbers and a real number x such that x is not approachable from A but lies in A.

(d) Find a subset A of the real numbers and a real number x such that x is not approachable from A and does not lie in A.

4.1.2. Let x be any positive real number. Let $\mathcal{R}_{\neq x} = \{s \in \mathcal{R} : s \neq x\}$. Consider the function

$$f : \mathcal{R}_{\neq x} \longrightarrow \mathcal{R}, \qquad f(s) = \frac{s^3 - x^3}{s - x}.$$

Does $\lim_{s \to x} f(s)$ exist, and if so, what is it? Explain. (The difference of powers formula on page 32 may help.)

4.1.2 Basic Function Limit Rules

Two examples of function limits are eminently believable but still deserve to be stated clearly.

Proposition 4.1.4 (Basic Function Limits). *Let A be a subset of $\mathcal{R}$, and let $x \in \mathcal{R}$ be approachable from A. Consider the functions*

$$f_0 : A \longrightarrow \mathcal{R}, \qquad f_0(s) = 1 \text{ for all } s$$

and

$$f_1 : A \longrightarrow \mathcal{R}, \qquad f_1(s) = s.$$

Then

$$\lim_{s \to x} f_0(s) = 1$$

and

$$\lim_{s \to x} f_1(s) = x.$$

Less pedantically, the limits in the proposition are written

$$\boxed{\lim_{s \to x} 1 = 1}$$

and

$$\boxed{\lim_{s \to x} s = x.}$$

Note that here the power functions f_0 and f_1 have an arbitrary subset of $\mathcal{R}$ (not $\mathcal{R}_{>0}$) as their domain. This point was discussed on page 29.

Proof. For any sequence (s_n) in A that approaches x,

$$\lim_n (f_0(s_n)) = \lim_n (1,\ 1,\ 1,\ \ldots) = 1$$

and

$$\lim_n (f_1(s_n)) = \lim_n (s_n) = x.$$

□

For another basic function limit, recall (from page 29) that the domain of the power function f_α is

$$\begin{cases} \mathcal{R} & \text{if } \alpha \in \mathcal{Z}_{\geq 0}, \\ \mathcal{R}_{\neq 0} & \text{if } \alpha \in \mathcal{Z}_{\leq -1}, \\ \mathcal{R}_{\geq 0} & \text{if } \alpha \in \mathcal{Q}_{\geq 0} \text{ but } \alpha \notin \mathcal{Z}_{\geq 0}, \\ \mathcal{R}_{>0} & \text{if } \alpha \in \mathcal{Q}_{<0} \text{ but } \alpha \notin \mathcal{Z}_{\leq -1}. \end{cases}$$

Note that in for all $\alpha \in \mathcal{Q}$, 0 is approachable from the domain of f_α, even though 0 actually lies in the domain only for $\alpha \in \mathcal{Q}_{\geq 0}$.

Proposition 4.1.5 (Limit of the Power Function at Zero). *Let $\alpha \in \mathcal{Q}_{>0}$ be any positive rational number. Then*

$$\lim_{s \to 0} f_\alpha(s) = 0.$$

Let $\alpha \in \mathcal{Q}_{<0}$ be any negative rational number. Then

$$\lim_{s \to 0} f_\alpha(s) \text{ does not exist.}$$

Finally,

$$\lim_{s \to 0} f_0(s) = 1.$$

Proof. Let A denote the domain of f_α.

Suppose that $\alpha \in \mathcal{Q}_{>0}$. Consider any sequence (s_n) in A that approaches 0. Let an arbitrary $\varepsilon > 0$ be given. Define in turn $\tilde{\varepsilon} = \varepsilon^{1/\alpha}$. Because $\lim_n(s_n) = 0$, there is a starting index N such that:

$$\text{For } n \geq N, \quad |s_n| < \tilde{\varepsilon}.$$

It follows that

$$\text{For all } n \geq N, \quad |f_\alpha(s_n)| = |s_n^\alpha| = |s_n|^\alpha < \tilde{\varepsilon}^\alpha = \varepsilon.$$

Here s_n can be negative only if $\alpha \in \mathcal{Z}_{\geq 1}$, in which case $|(-1)^\alpha| = 1$ and so $|s_n^\alpha| = |(-|s_n|)^\alpha| = |(-1)^\alpha |s_n|^\alpha| = ||s_n|^\alpha| = |s_n|^\alpha$, giving the second equality in the display. For positive s_n, the second equality in the display is trivial. The "$<$" in the display follows from its counterpart in the previous one, because f_α is strictly increasing on $\mathcal{R}_{>0}$—here is where we use the condition that $\alpha > 0$. With the display justified, note that it says that $\lim_n(f_\alpha(s_n)) = 0$. And because the argument applies to any sequence (s_n) in A that approaches 0, we have shown that $\lim_{s \to 0} f_\alpha(s) = 0$.

Now suppose that $\alpha \in \mathcal{Q}_{<0}$, so that $-\alpha \in \mathcal{Q}_{>0}$. For any positive integer n,

$$f_\alpha(1/n) = (1/n)^\alpha = 1/(1/n^{-\alpha}).$$

By the $1/n^\alpha$ Rule for sequences, where the α in the rule is the $-\alpha$ here, $\lim_n(1/n^{-\alpha}) = 0$, and so $f_\alpha(1/n)$ grows without bound as n grows. Thus $\lim_{s\to 0} f_\alpha(s)$ does not exist.

Finally, the result that $\lim_{s\to 0} f_0(s) = 1$ is the first part of the previous proposition. □

Exercise

4.1.3. Discuss the meaning, the existence, and the value of $\lim_{s\to x} s^2$.

4.1.3 Progenitive Function Limit Rules

Let A be a subset of $\mathcal{R}$, and consider two functions

$$f, g : A \longrightarrow \mathcal{R}.$$

Let $c \in \mathcal{R}$ be any number. Then the functions

$$f \pm g, cf, fg : A \longrightarrow \mathcal{R}$$

are defined as follows:

$$\begin{aligned} (f \pm g)(x) &= f(x) \pm g(x) \quad \text{for all } x \in A, \\ (cf)(x) &= c \cdot f(x) \quad \text{for all } x \in A, \\ (fg)(x) &= f(x)g(x) \quad \text{for all } x \in A. \end{aligned}$$

As with sequences, these functions are the sum/difference of f and g, a constant multiple of f, and the product of f and g. Also, consider the subset of A where g is nonzero,

$$A' = \{x \in A : g(x) \neq 0\}.$$

Then the functions

$$1/g, f/g : A' \longrightarrow \mathcal{R}$$

are defined to be

$$(1/g)(x) = 1/g(x) \quad \text{for all } x \in A'$$

and

$$(f/g)(x) = f(x)/g(x) \quad \text{for all } x \in A'.$$

These functions are the reciprocal of g and the quotient of f and g.

For example, given two rational power functions on the positive real numbers,

$$f_\alpha, f_\beta : \mathcal{R}_{>0} \longrightarrow \mathcal{R},$$

their product is

$$(f_\alpha f_\beta)(x) = f_\alpha(x) f_\beta(x) = x^\alpha x^\beta = x^{\alpha+\beta}, \quad x \in \mathcal{R}_{>0},$$

which is to say that their product is

$$f_\alpha f_\beta = f_{\alpha+\beta} : \mathcal{R}_{>0} \longrightarrow \mathcal{R}.$$

The following result gives the limits of the newly-defined functions above in terms of the limits of f and g.

Theorem 4.1.6 (Progenitive Function Limit Rules). *Let A be a subset of $\mathcal{R}$, and consider two functions*

$$f, g : A \longrightarrow \mathcal{R}.$$

Let $c \in \mathcal{R}$ be any number. Suppose that the point $x \in \mathcal{R}$ is approachable from A. Suppose that $\lim_{s\to x} f(s)$ and $\lim_{s\to x} g(s)$ exist. Then

(1) (Sum/Difference Rule.) $\lim_{s\to x}(f \pm g)(s)$ *exists and is*

$$\boxed{\lim_{s\to x}(f \pm g)(s) = \lim_{s\to x} f(s) \pm \lim_{s\to x} g(s).}$$

(2) (Constant Multiple Rule.) $\lim_{s\to x}(cf)(s)$ *exists and is*

$$\boxed{\lim_{s\to x}(cf)(s) = c \cdot \lim_{s\to x} f(s).}$$

(3) (Product Rule.) $\lim_{s\to x}(fg)(s)$ *exists and is*

$$\boxed{\lim_{s\to x}(fg)(s) = \lim_{s\to x} f(s) \cdot \lim_{s\to x} g(s).}$$

(4) (Reciprocal Rule.) *Let $A' = \{s \in A : g(s) \neq 0\}$. If x is approachable from A' and $\lim_{s\to x} g(s) \neq 0$ then $\lim_{s\to x}(1/g)(s)$ exists and is*

$$\boxed{\lim_{s\to x}(1/g)(s) = 1/\lim_{s\to x} g(s).}$$

(5) (Quotient Rule.) *Let $A' = \{s \in A : g(s) \neq 0\}$. If x is approachable from A' and $\lim_{s\to x} g(s) \neq 0$ then $\lim_{s\to x}(f/g)(s)$ exists and is*

$$\boxed{\lim_{s\to x}(f/g)(s) = \lim_{s\to x} f(s)/\lim_{s\to x} g(s).}$$

The reader is reminded that the boxed formulas apply only when the limits on their rights sides exist (and are nonzero when necessary).

Proof. (Sketch.) Each of these follows immediately from the corresponding progenitive rule for sequence limits. However, the yoga of the grammar is elaborate enough to call for an example. So, let

$$\ell = \lim_{s \to x} f(s), \qquad m = \lim_{s \to x} g(s).$$

Let (s_n) be any sequence in A that approaches x. Then from Definition 4.1.2,

$$\lim_n (f(s_n)) = \ell, \qquad \lim_n (g(s_n)) = m.$$

To prove the Sum Rule for functions, note that by the Sum Rule for sequences, also

$$\lim_n (f(s_n) + g(s_n)) = \ell + m,$$

which is to say, by definition of the function $f + g$,

$$\lim_n ((f + g)(s_n)) = \ell + m.$$

Because this last display is valid for every sequence (s_n) in A that approaches x, Definition 4.1.2 now gives the Sum Rule for functions,

$$\lim_{s \to x} (f + g)(s) = \ell + m.$$

The other parts of the proposition are proved virtually identically. □

For example, consider any rational function

$$f : A \longrightarrow \mathcal{R}, \quad f = g/h,$$

where g is a polynomial and h is a polynomial other than the zero polynomial, and $A = \{x \in \mathcal{R} : h(x) \neq 0\}$. Because h has only finitely many roots, any point $x \in A$ is approachable from A. In consequence of the proposition,

$$\lim_{s \to x} f(s) = f(x).$$

4.1.4 Order Function Limit Rules

Proposition 4.1.7 (Inequality Rule for Functions). *Let A be a subset of $\mathcal{R}$, and consider two functions*

$$f, g : A \longrightarrow \mathcal{R}.$$

Suppose that the point $x \in \mathcal{R}$ is approachable from A, and that the limits $\lim_{s \to x} f(s)$ and $\lim_{s \to x} g(s)$ exist. Suppose further that

$$f(s) \leq g(s) \quad \textit{for all } s \in A.$$

Then

$$\lim_{s \to x} f(s) \leq \lim_{s \to x} g(s).$$

Proof. Take any sequence (s_n) in A approaching x. Then $\lim_n(f(s_n))$ exists and equals $\lim_{s\to x} f(s)$, and $\lim_n(g(s_n))$ exists and equals $\lim_{s\to x} g(s)$. Because $f(s_n) \leq g(s_n)$ for each n, the Inequality Rule for Sequences gives $\lim_n(f(s_n)) \leq \lim_n(g(s_n))$. That is, $\lim_{s\to x} f(s) \leq \lim_{s\to x} g(s)$ as desired. □

Proposition 4.1.8 (Squeezing Rule for Functions). *Let A be a subset of $\mathcal{R}$, and consider three functions*

$$f, g, h : A \longrightarrow \mathcal{R}.$$

Suppose that the point $x \in \mathcal{R}$ is approachable from A, and that the limits $\lim_{s\to x} f(s)$ and $\lim_{s\to x} g(s)$ exist and are equal to a common value ℓ. Suppose further that

$$f(s) \leq h(s) \leq g(s) \quad \textit{for all } s \in A.$$

Then $\lim_{s\to x} h(s)$ also exists and equals ℓ.

Proof. Take any sequence (s_n) in A approaching x. Then $\lim_n(f(s_n))$ exists and equals $\lim_{s\to x} f(s) = \ell$, and $\lim_n(g(s_n))$ exists and equals $\lim_{s\to x} g(s) = \ell$. Because $f(s_n) \leq h(s_n) \leq g(s_n)$ for each n, also $\lim_n(h(s_n)) = \ell$ by the Squeezing Rule for Sequences. This argument holds for all sequence (s_n) in A approaching x, and so $\lim_{s\to x} h(s) = \ell$ as desired. □

Analogues of various remarks about the Inequality Rule and the Squeezing Rule for Sequences apply to these rules for functions as well. First, if $g(s) \geq 0$ for all $s \in A$ and $\lim_{s\to x} g(s)$ exists then also $\lim_{s\to x} g(s) \geq 0$. Second, again a point of the Squeezing Rule is that the middle limit exists. And third, the hypothesis that $f(s) \geq g(s)$ (or $f(s) \leq h(s) \leq g(s)$) for *all* $s \in A$ can be weakened: the inequalities need to hold only for all $s \in A$ *close enough to* x. But similarly to reindexing sequences, we can shrink the domains of functions, and in particular A can be made small enough so that the additional fussing required to quantify and track the *close enough to* x through statements and proofs is not worth the effort.

The basic function limit rules and progenitive function limit rules have been established by working with Definition 4.1.2 of a function limit as the common value of all relevant sequence limits. But now that some function limit rules are in place, the idea is to use them whenever we can do so rather than making unnecessary further arguments that refer back to the definition and universalize over sequences. The process here is similar to establishing properties of the absolute value that make no reference to cases and then using the properties with no further reference to the casewise definition. And it is similar to reasoning about sequences by recourse to the sequence limit

rules rather than the definition of a sequence limit. Often in mathematics, we want our definition to give rise quickly to desirable, handy consequences that we can then use and build on further, with little need to work directly with the definition thereafter.

4.2 The Derivative

4.2.1 Definition of the Derivative

Definition 4.2.1 (Derivative). *Let*

$$f : A \longrightarrow \mathcal{R}$$

be a function. Let x *be a point of* A *that is approachable from* A*. Let*

$$A_{\neq x} = \{s \in A : s \neq x\}.$$

Consider an auxiliary function

$$g : A_{\neq x} \longrightarrow \mathcal{R}, \qquad g(s) = \frac{f(s) - f(x)}{s - x}.$$

If the function limit

$$\lim_{s \to x} g(s) = \lim_{s \to x} \frac{f(s) - f(x)}{s - x}$$

exists, then its value is the **derivative of** f **at** x*, denoted* $f'(x)$*. That is,*

$$f'(x) = \lim_{s \to x} \frac{f(s) - f(x)}{s - x}, \quad \textit{if the limit exists.}$$

If $f'(x)$ *exists then* f *is* **differentiable at** x*.*

Exercise

4.2.1. Let $\mathcal{R}_{\neq 0} = \{s \in \mathcal{R} : s \neq 0\}$. Consider the function

$$g : \mathcal{R}_{\neq 0} \longrightarrow \mathcal{R}, \quad g(s) = \frac{|s| - |0|}{s - 0}.$$

(a) Define a sequence in $\mathcal{R}_{\neq 0}$, $(s_n) = (1/n)$. Does $\lim_n(s_n)$ exist? If so, what is it?

(b) Define a sequence in $\mathcal{R}_{\neq 0}$, $(s_n) = (-1/n)$. Does $\lim_n(s_n)$ exist? If so, what is it?

(c) Is the absolute value function differentiable at 0? Explain.

4.2.2 A Consequence Worth Noting Immediately

Proposition 4.2.2. *Let A be a subset of $\mathcal{R}$, let x be a point of A, and suppose that the function*

$$f : A \longrightarrow \mathcal{R}$$

is differentiable at x. Then

$$\lim_{s \to x} f(s) = f(x).$$

Exercise

4.2.2. Use progenitive function limit rules to prove Proposition 4.2.2. The argument could start from the fact that for $s \neq x$,

$$f(s) - f(x) = \frac{f(s) - f(x)}{s - x} \cdot (s - x),$$

and the argument should address the existence issue.

4.2.3 The Derivative and the Tangent Line

The notion of a tangent slope as a limit of secant slopes is unsatisfying. Yes, the formula

$$f'(x) = \lim_{s \to x} \frac{f(s) - f(x)}{s - x}$$

encodes the idea of the *derivative* as the limit of the secant slopes, but calling the left side of the display the *tangent slope* finesses the question of what the tangent slope really is conceptually, of why the limit gives us something that we already care about rather than something on which we bestow a geometrically suggestive name to make ourselves care about it. Perhaps the tangent line is somehow the limit of the secant lines, and perhaps it follows that the tangent slope is the limit of the secant slopes, but any such argument is well beyond the scope of our grammar.

As a better formulation, a recharacterization of the derivative definition—really just a small algebraic rearrangement—gives us the quantitative language to show that the derivative has a property that captures the idea of tangent slope *analytically*.

Proposition 4.2.3 (Recharacterization of the Derivative). *Let A be a subset of $\mathcal{R}$. Let $f : A \longrightarrow \mathcal{R}$ be a function. Consider a point $x \in A$ that is approachable from A. Then for any real number ℓ,*

$$f'(x) \text{ exists and equals } \ell \iff \lim_{s \to x} \frac{f(s) - f(x) - \ell(s - x)}{s - x} = 0.$$

Proof. Given ℓ, compute for any $s \in A$ such that $s \neq x$,

$$\frac{f(s)-f(x)}{s-x} - \ell = \frac{f(s)-f(x)-\ell(s-x)}{s-x}.$$

The result follows immediately. □

To apply the proposition, set up its environment by taking a subset A of $\mathcal{R}$, a function $f : A \longrightarrow \mathcal{R}$, and a point $x \in A$ that is approachable from A. For any real number ℓ, the function

$$L_\ell : \mathcal{R} \longrightarrow \mathcal{R}, \quad L_\ell(s) = f(x) + \ell(s-x)$$

is the function whose graph is the line through the point $(x, f(x))$ having slope ℓ. The quantity

$$f(s) - L_\ell(s) = f(s) - f(x) - \ell(s-a), \quad s \in A$$

is the vertical distance between the graph of f and L_ℓ over points $s \in A$. Assuming that $\lim_{s\to x} f(s) = f(x)$, we have

$$\lim_{s\to x}(f(s) - L_\ell(s)) = \lim_{s\to x}(f(s)-f(x)) - \lim_{s\to x}\ell(s-a) = 0 - 0 = 0.$$

That is, the vertical distance over s between the graph and the line goes to 0 as s moves toward x. But by the proposition, *only when* $f'(x)$ *exists and* ℓ *is set to* $f'(x)$ *do we also have*

$$\lim_{s\to x} \frac{f(s) - L_\ell(s)}{s-x} = 0.$$

That is:

> *Only when* $f'(x)$ *exists is there a line through* $(x, f(x))$ *that fits the graph of* f *so well that the vertical distance from the graph to the line tends to* 0 *faster than* $s-x$ *as* s *tends to* x. *The line is unique, and its slope is* $f'(x)$.

In other words, we now have an analytic description of the tangent line to the graph of f at the point $(x, f(x))$. It is *the best-fitting line*, in the sense just quantified in the displayed text. These ideas lead onward to a geometric description of the tangent line, to be presented in the following exercise.

Exercise

4.2.3. Let A be a subset of $\mathcal{R}$, let $f : A \longrightarrow \mathcal{R}$ be a function, and suppose that f is differentiable at the point x of A. For any real number ℓ, define a function (as in the section)

$$L_\ell : \mathcal{R} \longrightarrow \mathcal{R}, \quad L_\ell(s) = f(x) + \ell(s - x).$$

(a) Show that

$$\lim_{s \to x} \left(\frac{f(s) - f(x) - \ell(s - x)}{s - x} \right) = f'(x) - \ell.$$

(b) Suppose that $\ell < f'(x)$, so that $f'(x) - \ell > 0$. What does the limit in (a) say about the vertical difference $f(s) - L_\ell(s)$ for all s close enough to x such that $s > x$? What does this say about the graphs of L_ℓ and f in relation to one another near $(x, f(x))$? What does the limit in (a) say about the vertical difference $f(s) - L_\ell(s)$ for all s close enough to x such that $s < x$? What does this say about the graphs of L_ℓ and f in relation to one another near $(x, f(x))$?

(c) Now suppose that $\ell > f'(x)$, so that $f'(x) - \ell < 0$. Now what holds for the graphs of L_ℓ and f in relation to one another near $(x, f(x))$?

(d) Let x be a point of A such that there exist other points s of A as close to x as desired with $s < x$, and there exist other points s of A as close to x as desired with $s > x$. Complete the following sentence by replacing each of X, Y, Z and W by "above" or "below": *This exercise has shown that* $f'(x)$ *is plausibly the tangent slope of the graph of* f *at* $(x, f(x))$ *in the geometric sense that any line through* $(x, f(x))$ *with shallower slope cuts the graph there from* X *to* Y *moving left to right, while any line through* $(x, f(x))$ *with steeper slope cuts the graph from* Z *to* W *moving left to right.*

(e) What needs to be said, similarly to (d), if x is a point of A that is approachable from A only from the right? Same question but with *left* instead of *right*.

4.2.4 A Basic Derivative: the Power Function Revisited

To apply the definition of the derivative to the power function from chapter 2,

$$f_\alpha : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad f_\alpha(x) = x^\alpha \qquad (\text{where } \alpha \in \mathcal{Q}),$$

first let

$$\mathcal{R}^+_{\neq 1} = \{s \in \mathcal{R}_{>0} : s \neq 1\},$$

and consider the function

$$g : \mathcal{R}^+_{\neq 1} \longrightarrow \mathcal{R}, \qquad g(s) = \frac{f_\alpha(s) - f_\alpha(1)}{s - 1}.$$

Consider any sequence (s_n) in $\mathcal{R}^+_{\neq 1}$ that approaches 1. This is exactly the sort of sequence required by Proposition 3.3.1 (page 100), which says that for any such sequence,

$$\lim_n (g(s_n)) = \alpha.$$

That is, f'_α exists at $x = 1$ and is $f'_\alpha(1) = \alpha$.

For general $x \in \mathcal{R}_{>0}$ rather than $x = 1$, define

$$\mathcal{R}^+_{\neq x} = \{s \in \mathcal{R}_{>0} : s \neq x\},$$

and consider the function

$$h : \mathcal{R}^+_{\neq x} \longrightarrow \mathcal{R}, \qquad h(s) = \frac{f_\alpha(s) - f_\alpha(x)}{s - x}.$$

Then by a little algebra as on page 46 (exercise 4.2.4(a)),

$$h(s) = x^{\alpha-1} g(t) \quad \text{where } t = s/x. \tag{4.3}$$

It follows (exercise 4.2.4(b)) that, as we found less formally in section 2.4,

$$\lim_{s \to x} h(s) = \alpha x^{\alpha-1}.$$

That is, for each $x \in \mathcal{R}_{>0}$, $f'_\alpha(x)$ exists and is $\alpha f_{\alpha-1}(x)$. And so:

$$\textit{For } \alpha \in \mathcal{Q},\ f'_\alpha = \alpha f_{\alpha-1}.$$

Because the power function is differentiable, it follows as a special case of Proposition 4.2.2 (page 131) that also:

$$\textit{For } \alpha \in \mathcal{Q} \textit{ and } x \in \mathcal{R}_{>0},\ \lim_{s \to x} f_\alpha(s) = f_\alpha(x).$$

In the previous paragraph, if α is an integer, then the argument applies to any nonzero $x \in \mathcal{R}$, not only to $x \in \mathcal{R}_{>0}$. If α is a nonnegative integer then a straightforward argument shows directly that $f'_\alpha(0) = \alpha f_{\alpha-1}(0)$, provided that we understand $0 \cdot f_{-1}$ to take the value 0 at $x = 0$. (exercise 4.2.5). If $\alpha \in \mathcal{Q}_{\geq 0}$ is a nonnegative rational number that isn't an integer, and $s \in \mathcal{R}_{>0}$, then

$$\frac{f_\alpha(s) - f_\alpha(0)}{s - 0} = \frac{s^\alpha}{s} = s^{\alpha-1},$$

and so, according to Proposition 4.1.5, $f'_\alpha(0) = 0$ if $\alpha > 1$ but $f'_\alpha(0)$ does not exist if $0 < \alpha < 1$. In sum:

Proposition 4.2.4 (Derivative of the Power Function). *Let α be a rational number. Take the domain of f_α to be*

$$\begin{cases} \mathcal{R} & \textit{if } \alpha \in \mathcal{Z}_{\geq 0}, \\ \mathcal{R}_{\neq 0} & \textit{if } \alpha \in \mathcal{Z}_{\leq -1}, \\ \mathcal{R}_{\geq 0} & \textit{if } \alpha \in \mathcal{Q}_{\geq 0} \textit{ but } \alpha \notin \mathcal{Z}_{\geq 0}, \\ \mathcal{R}_{>0} & \textit{if } \alpha \in \mathcal{Q}_{<0} \textit{ but } \alpha \notin \mathcal{Z}_{\leq -1}. \end{cases}$$

Then the formula

$$\boxed{f'_\alpha = \alpha f_{\alpha-1},}$$

holds everywhere on the domain of f_α *unless* $0 \le \alpha < 1$*, in which case it holds everywhere on the domain of* f *except at* $x = 0$*. For* $\alpha = 0$*, we have* $f'_0(0) = 0$*. For* $0 < \alpha < 1$*,* $f'_\alpha(0)$ *does not exist.*

Because $0 \cdot f_{-1}(x) = 0$ for all $x \neq 0$, sometimes the boxed formula is understood to encompass the case $\alpha = 0$ and $x = 0$, i.e., $0 \cdot f_{-1}(0)$ is understood to mean 0. This is not strictly correct: $f^{-1}(0)$ does not exist. What is correct is that $\lim_{s\to 0} 0 \cdot f_{-1}(s) = 0$, and so in some sense $0 \cdot f_{-1}$ *should* take the value 0 at $x = 0$.

At the purely procedural level, the boxed formula says that *to differentiate the power function, bring the power down in front and reduce it by one in the exponent*. This is indeed the procedure, but the reader should be aware that there even though the procedure is easy, there is substance to the result.

Corollary 4.2.5. *Let* α *be a rational number, and let* x *be any element of the domain of* f_α*. Then*

$$\lim_{s\to x} f_\alpha(s) = f_\alpha(x).$$

Proof. This follows from the proposition unless $0 < \alpha < 1$ and $x = 0$. These exceptional cases are covered by Proposition 4.1.5 (page 125). □

We end the section with one more derivative result. Exercise 4.2.7 is to give a sketch of the proof.

Proposition 4.2.6 (Derivative of the Absolute Value Function Away From Zero). *Let* $\mathcal{R}_{\neq 0} = \{x \in \mathcal{R} : x \neq 0\}$*. Consider the function*

$$f : \mathcal{R}_{\neq 0} \longrightarrow \mathcal{R}, \quad f(x) = |x|.$$

This function is differentiable and its derivative is

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

Exercises

4.2.4. (a) Establish (4.3).

(b) Explain carefully why consequently $\lim_{s\to x} h(s) = \alpha x^{\alpha-1}$.

4.2.5. Let α be a nonnegative integer. Show that $f'_\alpha(0) = \alpha f_{\alpha-1}(0)$, provided that we understand $0 \cdot f_{-1}(0)$ to be 0.

4.2.6. Explain (an informal explanation is fine) why it follows from Proposition 4.2.4 and the various function limit results in this chapter that for any algebraic function $f : A \longrightarrow \mathcal{R}$ (see page 63) and any $x \in A$ that is approachable from A,

$$\lim_{s \to x} f(s) = f(x).$$

4.2.7. Sketch a proof of Proposition 4.2.6. Your argument should not involve any detail-work, but rather it should explain why for $x > 0$ the issue reduces to the derivative of a power function while for $x < 0$ it reduces to the derivative of the negative of a power function. For now, cite the fact (to be proved in the next section) that the derivative of the negative is the negative of the derivative.

4.2.5 Progenitive Derivative Rules

We have computed only one derivative so far, the derivative of the power function. Soon we will compute other specific derivatives. But in addition to computing the derivatives of particular functions, we can also compute the derivatives of combinations of functions progenitively, assuming that we already know the derivatives of the functions individually.

Theorem 4.2.7 (Progenitive Derivative Rules). *Let A be a subset of $\mathcal{R}$, and consider two functions*

$$f, g : A \longrightarrow \mathcal{R}.$$

Let $c \in \mathcal{R}$ be any number. Suppose that f' and g' exist on A. Also, let $A' = \{x \in A : g(x) \neq 0\}$ and suppose that f' and g' exist on A'. Then

(1) (Sum/Difference Rule.) $(f \pm g)'$ *exists on* A *and is* $(f \pm g)' = f' \pm g'$. *That is,*

$$\boxed{(f \pm g)'(x) = f'(x) \pm g'(x) \quad \textit{for all } x \in A.}$$

(2) (Constant Multiple Rule.) $(cf)'$ *exists on* A *and is* $(cf)' = cf'$. *That is,*

$$\boxed{(cf)'(x) = c \cdot f'(x) \quad \textit{for all } x \in A.}$$

(3) (Product Rule.) $(fg)'$ *exists on* A, *and is* $(fg)' = fg' + f'g$. *That is,*

$$\boxed{(fg)'(x) = f(x)g'(x) + f'(x)g(x) \quad \textit{for all } x \in A.}$$

(4) (Reciprocal Rule.) $(1/g)'$ *exists on* A' *and is* $(1/g)' = -g'/g^2$. *That is,*

$$\boxed{(1/g)'(x) = -\frac{g'(x)}{g(x)^2} \quad \textit{for all } x \in A'.}$$

(5) (Quotient Rule.) $(f/g)'$ *exists on* A' *and is* $(f/g)' = (f'g - fg')/g^2$. *That is,*

$$\boxed{(f/g)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \quad \textit{for all } x \in A'.}$$

Regarding the hypotheses of the theorem, in fact f' and g' are guaranteed to exist on A', but showing this right now would take us too far afield. As with the progenitive function limit rules, the reader is reminded that the boxed formulas apply only when the derivatives on their right sides exist.

Proof. (1) To prove the Sum Rule, compute for any $x \in A$ that

$$\begin{aligned}\frac{(f+g)(s) - (f+g)(x)}{s-x} &= \frac{f(s) + g(s) - f(x) - g(x)}{s-x} \\ &= \frac{f(s) - f(x)}{s-x} + \frac{g(s) - g(x)}{s-x}.\end{aligned}$$

The quotients on the right side have limits $f'(x)$ and $g'(x)$ individually as s tends to x, and so their sum has limit $f'(x) + g'(x)$.

(2) The proof of the Constant Multiple Rule is very similar to the proof of the Sum/Difference Rule.

(3) To prove the Product Rule, we first need to recall that by Proposition 4.2.2 (page 131), for any $x \in A$, because $f'(x)$ exists, $\lim_{s\to x} f(s)$ exists and is $f(x)$. Now compute that

$$\begin{aligned}\frac{(fg)(s) - (fg)(x)}{s-x} &= \frac{f(s)g(s) - f(x)g(x)}{s-x} \\ &= \frac{f(s)g(s) - f(s)g(x) + f(s)g(x) - f(x)g(x)}{s-x} \\ &= f(s)\,\frac{g(s) - g(x)}{s-x} + \frac{f(s) - f(x)}{s-x}\,g(x).\end{aligned}$$

By various function limit rules, the right side has the desired limit as s tends to x.

(4) For the Reciprocal Rule, suppose that $x \in A'$. Then

$$\frac{(1/g)(s) - (1/g)(x)}{s-x} = \frac{1/g(s) - 1/g(x)}{s-x} = \frac{g(x) - g(s)}{s-x} \cdot \frac{1}{g(s)g(x)}.$$

As s tends to x, the first quotient has limit $-g'(x)$ and the second has limit $1/g(x)^2$, giving the result. Because $g'(x)$ exists, $\lim_{s\to x} g(s) = g(x)$ (again by Proposition 4.2.2) and so because $g(x)$ is nonzero, also $g(s)$ is nonzero for s-values close to x. That is, we needn't worry about dividing by 0 somewhere in this argument.

(5) Finally, the Quotient Rule follows from the Product Rule and the Reciprocal Rule because $f/g = f \cdot 1/g$. □

The most important progenitive derivative rule is called the Chain Rule. Given two functions where the codomain of the first is the domain of the second,

$$f : A \longrightarrow B \quad \text{and} \quad g : B \longrightarrow C,$$

their **composition** is

$$g \circ f : A \longrightarrow C, \qquad (g \circ f)(x) = g(f(x)).$$

Actually, the composition is sensible as long as the range $f(A)$ of A is a *subset* of the domain B of g, because then we may respecify the codomain of f to be B.

For example, given two rational power functions on the positive real numbers,

$$f_\alpha, f_\beta : \mathcal{R}_{>0} \longrightarrow \mathcal{R}_{>0},$$

their composition is

$$(f_\beta \circ f_\alpha)(x) = f_\beta(f_\alpha(x)) = (x^\alpha)^\beta = x^{\alpha\beta}, \quad x \in \mathcal{R}_{>0},$$

which is to say that their composition is

$$f_\beta \circ f_\alpha = f_{\alpha\beta} : \mathcal{R}_{>0} \longrightarrow \mathcal{R}_{>0}.$$

The Chain Rule says that the derivative of the composition at x, i.e., $(g \circ f)'(x)$, is the product of $g'(f(x))$ and $f'(x)$, assuming that these both exist. A second recharacterization of the derivative helps to prove the Chain Rule smoothly.

Proposition 4.2.8 (Second Recharacterization of the Derivative). *Consider a function $f : A \longrightarrow \mathcal{R}$. Let $x \in A$ be approachable from A. Then for any real number ℓ,*

$$f'(x) \text{ exists and equals } \ell$$

if and only if there is a function $q : A \longrightarrow \mathcal{R}$ such that

$$f(s) = f(x) + \big(\ell + q(s)\big) \cdot (s - x) \quad \text{where } \lim_{s\to x} q(s) = 0 \text{ and } q(x) = 0.$$

Proof. Suppose that $f'(x)$ exists and equals ℓ. Define

$$q(s) = \begin{cases} \dfrac{f(s)-f(x)}{s-x} - \ell & \text{if } s \neq x, \\ 0 & \text{if } s = x. \end{cases}$$

Then $f(s) = f(x) + \big(\ell + q(s)\big)\cdot(s-x)$ for $s \neq x$ and for $s = x$. Because $f'(x) = \ell$ it follows that $\lim_{s\to x} q(s) = 0$. And $q(x) = 0$, so q has the desired properties.

Conversely, suppose that such a function q exists. Then for all $s \in A$ such that $s \neq x$ we have

$$\frac{f(s)-f(x)}{s-x} = \ell + q(s),$$

and so

$$\lim_{s\to x} \frac{f(s)-f(x)}{s-x} = \lim_{s\to x} (\ell + q(s)) = \ell.$$

That is, $f'(x)$ exists and equals ℓ. □

Now the Chain Rule can be proved.

Theorem 4.2.9 (Chain Rule). *Let A, B, and C be subsets of $\mathcal{R}$. Consider two functions*

$$f : A \longrightarrow B \quad \textit{and} \quad g : B \longrightarrow C.$$

Let x be a point of A. Suppose that f is differentiable at x and g is differentiable at $f(x)$. Then the composition $g \circ f$ is differentiable at x, and its derivative there is

$$\boxed{(g\circ f)'(x) = g'(f(x))\cdot f'(x).}$$

Consequently, if f is differentiable on A and g is differentiable on $f(A)$, then $g \circ f$ is differentiable on A with derivative

$$\boxed{(g\circ f)' = (g'\circ f)\cdot f' : A \longrightarrow \mathcal{R}.}$$

Proof. Introduce the symbols $\ell = f'(x)$ and $y = f(x)$ and $m = g'(y) = g'(f(x))$. We know that for $s \in A$,

$$f(s) = f(x) + \big(\ell + q(s)\big)\cdot(s-x), \quad \lim_{s\to x} q(s) = 0 = q(x), \tag{4.4}$$

and that for $t \in B$,

$$g(t) = g(y) + \big(m + r(t)\big)\cdot(t-y), \quad \lim_{t\to y} r(t) = 0 = r(y). \tag{4.5}$$

And we need to show that for $s \in A$,

$$(g \circ f)(s) = (g \circ f)(x) + \big(m\ell + \tilde{q}(s)\big) \cdot (s - x), \quad \lim_{s \to x} \tilde{q}(s) = 0 = \tilde{q}(x).$$

Compute for $s \in A$ that by the definition of composition and then by (4.4),

$$(g \circ f)(s) = g(f(s)) = g\big(f(x) + (\ell + q(s)) \cdot (s - x)\big).$$

That is, $(g \circ f)(s) = g(t)$ where $t = y + (\ell + q(s)) \cdot (s - x)$. Now (4.5) gives

$$\begin{aligned}(g \circ f)(s) &= g(y) + \big(m + r(t)\big) \cdot (t - y) \\ &= g(y) + \big(m + r(t)\big) \cdot (\ell + q(s)) \cdot (s - x) \\ &= g(y) + \big(m\ell + r(t)\ell + (m + r(t))q(s)\big) \cdot (s - x).\end{aligned}$$

So, because $g(y) = g(f(x)) = (g \circ f)(x)$, we have

$$(g \circ f)(s) = (g \circ f)(x) + (m\ell + \tilde{q}(s))(s - x),$$

where

$$\tilde{q}(s) = r(t)\ell + \big(m + r(t)\big)q(s), \quad t = y + (\ell + q(s)) \cdot (s - x). \tag{4.6}$$

What needs to be shown is that $\lim_{s \to x} \tilde{q}(s) = 0$ and $\tilde{q}(x) = 0$. We have from (4.4)

$$\lim_{s \to x} q(s) = 0,$$

so that consequently, using the definition of t in (4.6),

$$\lim_{s \to x} t = y,$$

and so, because $s \to x$ now gives $t \to y$ or $t = y$, we have from (4.5)

$$\lim_{s \to x} r(t) = 0,$$

and so altogether, using the definition of $\tilde{q}$ in (4.6),

$$\lim_{s \to x} \tilde{q}(s) = 0.$$

Also $\tilde{q}(x) = r(y)\ell + (m + r(y))q(x) = 0\ell + (m + 0)0 = 0$. This completes the argument. □

A seemingly more natural way to go about proving the Chain Rule is by writing for $s \neq x$ in A,

$$\frac{(g \circ f)(s) - (g \circ f)(x)}{s - x} = \frac{g(f(s)) - g(f(x))}{f(s) - f(x)} \cdot \frac{f(s) - f(x)}{s - x},$$

or, letting $t = f(s)$ and $y = f(x)$,

$$\frac{(g\circ f)(s)-(g\circ f)(x)}{s-x}=\frac{g(t)-g(y)}{t-y}\cdot\frac{f(s)-f(x)}{s-x}.$$

Because $f'(x)$ exists, Proposition 4.2.2 says that $\lim_{s\to x} t = y$. And so, seemingly,

$$\begin{aligned}\lim_{s\to x}\frac{(g\circ f)(s)-(g\circ f)(x)}{s-x} &= \lim_{s\to x}\left(\frac{g(t)-g(y)}{t-y}\cdot\frac{f(s)-f(x)}{s-x}\right)\\ &= \lim_{t\to y}\frac{g(t)-g(y)}{t-y}\cdot\lim_{s\to x}\frac{f(s)-f(x)}{s-x}\\ &= g'(y)\cdot f'(x) = g'(f(x))\cdot f(x).\end{aligned}$$

The problem with this argument is that it assumes that $f(s)\neq f(x)$, i.e., that $t\neq y$. The exceptional case when $s\neq x$ but $f(s)=f(x)$ needs to be addressed separately, causing clutter in the seemingly more natural approach. Our method of proving the Chain Rule skirts this issue by working in a setting that doesn't involve denominators.

Exercises

4.2.8. (a) Prove the Constant Multiple Rule.

(b) Prove the Quotient Rule.

4.2.9. (a) Consider the function

$$f_1:\mathcal{R}\longrightarrow\mathcal{R},\qquad f_1(x)=x.$$

Prove from scratch that its derivative is

$$f_1'=f_0:\mathcal{R}\longrightarrow\mathcal{R},\qquad f_0(x)=1.$$

(b) Use only the result from (a) and the Product Rule for derivatives to prove that because $f_2=f_1\cdot f_1$, it follows that $f_2'=2f_1$.

(c) Use only the results from (a) and (b) and the Product Rule to prove that $f_3'=3f_2$. Convince yourself that this process extends to the rule $f_n'=nf_{n-1}$ for all $n\in\mathcal{Z}_{\geq 1}$.

(d) Use the result from (a) and the Reciprocal Rule for derivatives to prove that because $f_{-n}=1/f_n$ for all $n\in\mathcal{Z}_{\geq 1}$ (now restricting the domain to the set of nonzero real numbers), it follows that $f_{-n}'=-nf_{-n-1}$. That is, the derivative of the power function for *integer* exponents follows from the result in (a) and progenitive derivative rules,

$$f_n'=nf_{n-1}\quad\text{for all } n\in\mathcal{Z}.$$

(e) Imagine that we know $f_{1/2}$ to be differentiable, but we don't know its derivative. Differentiate both sides of the relation

$$f_1 = f_{1/2} \cdot f_{1/2},$$

applying the Product Rule to the right side to obtain an expression involving $f_{1/2}$ and the extant-but-unknown derivative $f'_{1/2}$. Solve the resulting equality to find $f'_{1/2}$. Note, however, that this process does not prove that $f_{1/2}$ is differentiable.

4.2.10. Let $p(x)$ be a polynomial. Explain why differentiating $p(x)$ some finite number of times yields the zero function.

4.2.11. Let n be a nonnegative integer and let

$$h(x) = x + x^2 + x^3 + \cdots + x^n.$$

Repeatedly apply the Sum Rule for derivatives and the power function derivative formula to get

$$h'(x) = 1 + 2x + 3x^2 + \cdots + nx^{n-1}.$$

But also, factoring out a power of x and then using the finite geometric sum,

$$h(x) = (1 + x + x^2 + \cdots + x^{n-1})x = \frac{x^n - 1}{x - 1}x, \quad x \neq 1.$$

Thus $h(x) = f(x) \cdot g(x)$ where $f(x) = x^n - 1$ and $g(x) = x/(x-1)$, and so the Product Rule for derivatives gives

$$h'(x) = f'(x)g(x) + f(x)g'(x), \quad x \neq 1.$$

Show that this formula says

$$1 + 2x + 3x^2 + \cdots + nx^{n-1} = \frac{nx^n}{x-1} - \frac{x^n - 1}{(x-1)^2}, \quad x \neq 1.$$

4.2.12. Let $f : \mathcal{R} \longrightarrow \mathcal{R}$ be any function. What can be said about the compositions $f_0 \circ f$, $f_1 \circ f$, $f \circ f_0$, and $f \circ f_1$, where $f_0 : \mathcal{R} \longrightarrow \mathcal{R}$ and $f_1 : \mathcal{R} \longrightarrow \mathcal{R}$ are the usual power functions?

4.2.13. Let a, b, c, and d be real numbers, with c and d positive. Let α and β be rational numbers. Define

$$f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad f = \frac{a + bf_\alpha}{c + df_\beta}.$$

Compute the derivative f'.

4.2.14. (a) Let a and b be positive real numbers. Let α and γ be rational numbers. Define

$$f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad f = (a + bf_\alpha)^\gamma.$$

Compute the derivative f'.

(b) Let a, b, c, and d be positive real numbers. Let α, β, and γ be rational numbers. Define

$$f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad f = \left(\frac{a + bf_\alpha}{c + df_\beta}\right)^\gamma.$$

Compute the derivative f'.

4.3 Summary

The underlying concepts of sequence limit and function limit provide us with an environment to study integrals and derivatives. So far we have studied the calculus of only one function, the rational power function. The next three chapters will define, integrate, and differentiate more functions: the logarithm, the exponential, the cosine and the sine.

5

The Logarithm Function

The logarithm is defined as an area, an integral of the reciprocal function f_{-1}. In section 2.5 we integrated the rational power function f_α for $\alpha \neq -1$, but exercise 2.3.6 showed that our methods for doing so fail for the reciprocal function. That is, the methods of chapter 2 will not calculate the logarithm. In fact, the logarithm is not an algebraic function, and so hoping for it to have a formula that we can study algebraically is futile. Nonetheless, we can analyze its properties, and we can differentiate and integrate it. In the process, we will expand our notion of integration, no longer requiring that the endpoints of integration be in increasing order, and no longer requiring that the function being integrated be nonnegative.

Section 5.1 defines the logarithm and establishes its important properties. Section 5.2 shows that although the logarithm grows without bound, it grows very slowly. Section 5.3 computes the derivative of the logarithm, and section 5.4 integrates the logarithm. Section 5.5 generalizes ideas that have emerged during the course of the chapter to extend integration to functions that aren't necessarily nonnegative and to endpoints that aren't necessarily in increasing order, and then re-establishes some of our earlier integration results.

5.1 Definition and Properties of the Logarithm

5.1.1 Integration With Out-of-Order Endpoints

Definition 5.1.1 (Integral With Out-of-Order Endpoints). *Let* a *and* b *be real numbers with* $b < a$. *Let* $M \geq 0$ *be any nonnegative real number. Let*

$$f : [b, a] \longrightarrow [0, M]$$

be an integrable function. Then the integral of f *from* a *to* b *is defined to be*

$$\int_a^b f = -\int_b^a f.$$

That is, if the endpoints a *and* b *are out of order then the integral from* a *to* b *is the negative of the integral from* b *to* a*, where* b *and* a *are in order.*

For example, let $\alpha \in \mathcal{Q}$, $\alpha \neq -1$, and let $0 < b \leq a$. Let M be the maximum of $f_\alpha(a)$ and $f_\alpha(b)$, so that we can consider the function

$$f_\alpha : [b, a] \longrightarrow [0, M].$$

Its integral from a to b is, by Definition 5.1.1 and then (2.7) (page 53),

$$\int_a^b f_\alpha = -\int_b^a f_\alpha = -\left(\frac{a^{\alpha+1} - b^{\alpha+1}}{\alpha+1}\right) = \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1}.$$

That is, the formula in (2.7) holds regardless of whether a and b are in order,

$$\int_a^b f_\alpha = \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1} \quad \alpha \in \mathcal{Q},\ \alpha \neq -1,\ a, b \in \mathcal{R}_{>0}. \tag{5.1}$$

The formula is *symbolically robust*, even though it no longer represents area (which is always nonnegative) when $b < a$.

Exercise

5.1.1. Let a and b be real numbers with $a \leq b$. Consider two integrable functions

$$f : [a, b] \longrightarrow [0, M], \qquad g : [a, b] \longrightarrow [0, N],$$

and let $c \in \mathcal{R}_{\geq 0}$ be a nonnegative real number. According to Proposition 3.3.13 (page 116), the functions $f + g$ and cf are integrable, and

$$\int_a^b (f + g) = \int_a^b f + \int_a^b g, \qquad \int_a^b (cf) = c\int_a^b f.$$

Show that these formulas still hold if instead $b < a$ and the functions f and g have domain $[b, a]$.

5.1.2 The Fundamental Theorem of Calculus

As another comment on formula (5.1), again given $\alpha \in \mathcal{Q}$, $\alpha \neq -1$, consider the function

$$F : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \quad F = f_{\alpha+1}/(\alpha+1).$$

Then $F' = f_\alpha$, and the formula rewrites as

$$\int_a^b f_\alpha = F(b) - F(a), \quad \alpha \in \mathcal{Q},\ \alpha \neq -1,\ a, b \in \mathcal{R}_{>0}. \tag{5.2}$$

This is an instance of the *Fundamental Theorem of Calculus*. The theorem says that under appropriate circumstances, the integral of the derivative is the difference of the derivatand's values at the endpoints.

5.1.3 Definition of the Logarithm

We now use calculus to *define* a function.

Definition 5.1.2 (Logarithm). *The* **logarithm function**,

$$\ln : \mathcal{R}_{>0} \longrightarrow \mathcal{R},$$

is defined as follows: For any $x \in \mathcal{R}_{>0}$,

$$\ln(x) = \int_1^x f_{-1}.$$

In words, the logarithm of x *is the integral of the reciprocal function from* 1 *to* x.

Definitions 5.1.1 and 5.1.2 combine to say that $\ln(x)$ is positive for $x > 1$ but negative for $0 < x < 1$. Figure 5.1 shows a portion of the $y = 1/x$ graph together with the corresponding portion of the $y = \ln(x)$ graph. The key to understanding the figure is to realize that the *height* of the $y = \ln(x)$ graph is the *area* under the $y = 1/x$ graph from 1 to x if $x \geq 1$, and the *negative* of the area if $0 < x < 1$. Thus the $y = \ln(x)$ graph lies below the x-axis to the left of the vertical line $x = 1$.

5.1.4 The Key Property of the Logarithm

Theorem 5.1.3 (The Logarithm Converts Multiplication to Addition). *For all positive real numbers* a *and* b,

$$\ln(ab) = \ln(a) + \ln(b).$$

Proving this theorem spot-on requires some care, and it also requires moving from geometric intuition to algebraic intuition, guided by natural language. We begin with a generality before proceeding to specifics. This tidies the exposition.

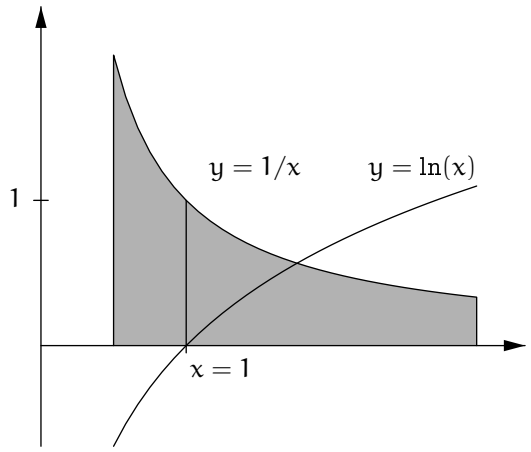


Figure 5.1. The reciprocal function and the logarithm function

5.1.5 Proof of the Key Property: A Generality

Let I be a nonempty interval in $\mathcal{R}$, let $M \geq 0$ be any nonnegative real number, and consider any function

$$f : I \longrightarrow [0, M],$$

integrable or not. Let a and b be any two points of I. If a and b are in order, i.e., if $a \leq b$, then we have a good visual sense of the area under the curve $y = f(x)$ from a to b. As before, let $\mathrm{Ar}_a^b(f)$ denote this area. Thus

For $a \leq b$, $\mathrm{Ar}_a^b(f)$ is the area over $[a, b]$ and under the graph of f.

This notion of $\mathrm{Ar}_a^b(f)$ is geometrically intuitive, but it is symbolically fragile because it relies on the side condition $a \leq b$. It would be unsustainable to verify the side condition every time that we make reference to a quantity $\mathrm{Ar}_a^b(f)$. But for now, we must. (Certainly we may *not* assume that whenever two letters represent numbers, the earlier letter in the alphabet is guaranteed to represent the smaller of the two numbers.) So, to make the expression $\mathrm{Ar}_a^b(f)$ symbolically robust, extend its definition as follows:

$$\text{If } a > b \text{ then } \mathrm{Ar}_a^b(f) = -\mathrm{Ar}_b^a(f). \tag{5.3}$$

For integrable functions, this repeats Definition 5.1.1, but the definition here is more general because f need not be integrable. This extended definition of $\mathrm{Ar}_a^b(f)$ may not feel entirely clear or natural geometrically, but it is utterly platonic symbolically. An immediate consequence of the definition is that

$$\mathrm{Ar}_b^a(f) = -\mathrm{Ar}_a^b(f) \quad \text{for all } a, b \in I. \tag{5.4}$$

Indeed, if $a \geq b$ then (5.4) simply repeats (5.3), while if $a < b$ then (5.4) follows from (5.3) with the roles of a and b reversed (exercise 5.1.2).

A consequence of (5.4) is that

$$\text{For all } a, b, c \in I, \quad \mathrm{Ar}_a^b(f) + \mathrm{Ar}_b^c(f) = \mathrm{Ar}_a^c(f). \tag{5.5}$$

The *all* in (5.5) connotes that the equality holds regardless of the order of a, b, and c. Proving this amounts to verifying that all six cases reduce back to the natural case where $a \leq b \leq c$, when (5.5) holds for geometric reasons (see figure 5.2) because our area function, whatever it is, has sensible basic properties. Life is too short to carry out all five other cases, so we content ourselves with working one of them, chosen at random, to get a sense of all five arguments. For example, suppose that $b \leq c \leq a$. Then

$$\begin{aligned}
&\mathrm{Ar}_a^b(f) + \mathrm{Ar}_b^c(f) \\
&\qquad = -\mathrm{Ar}_b^a(f) + \mathrm{Ar}_b^c(f) && \text{by (5.4)} \\
&\qquad = -(\mathrm{Ar}_b^c(f) + \mathrm{Ar}_c^a(f)) + \mathrm{Ar}_b^c(f) && \text{by (5.5) with } b, c, a \text{ for } a, b, c \\
&\qquad = -\mathrm{Ar}_c^a(f) && \text{because two of the terms cancel} \\
&\qquad = \mathrm{Ar}_a^c(f) && \text{by (5.4).}
\end{aligned}$$

To repeat, nothing in this proof of (5.5) is related to integrability.

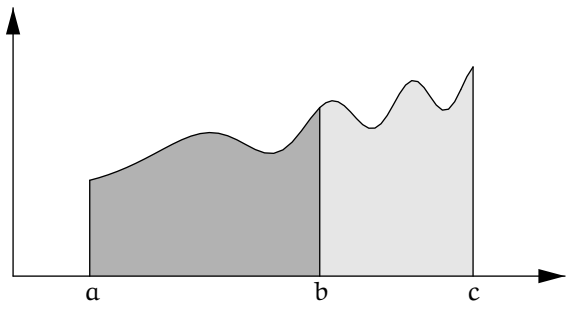


Figure 5.2. $\mathrm{Ar}_a^b(f) + \mathrm{Ar}_b^c(f) = \mathrm{Ar}_a^c(f)$: the obvious case

Exercises

5.1.2. Carefully explain the statement in the text that if $a < b$ then (5.4) follows from (5.3) with the roles of a and b reversed.

5.1.3. Choose another ordering of a, b, and c, and reduce the equality in (5.5) to the case that the three points are in order.

5.1.6 Proof of the Key Property: A Specific Argument

Now we return to studying areas under the graph of one particular function, the reciprocal function on the positive real numbers,

$$f_{-1} : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad f(x) = 1/x.$$

If we restrict f to any closed subinterval $I = [A, B]$ or $I = [A, \infty)$ of its domain, then necessarily $A > 0$ and we can set $M = 1/A$ to get

$$f_{-1} : I \longrightarrow [0, M].$$

Thus the restriction of the reciprocal function to I falls under the previous general discussion. However, the power function f_{-1} is monotonic, and hence, beyond the previous generalities, it is integrable. For this reason, we write "$\int$" rather than "Ar" in the following discussion. As in Definition 5.1.1, for any positive a and b,

$$\int_a^b f_{-1} = -\int_b^a f_{-1} \quad \text{if } a > b. \tag{5.6}$$

And (5.5) implies

$$\text{For all } a, b, c \in \mathcal{R}_{>0}, \quad \int_a^b f_{-1} + \int_b^c f_{-1} = \int_a^c f_{-1}. \tag{5.7}$$

Recall that Proposition 2.5.1 (page 51) showed that for any $\alpha \in \mathcal{Q}$,

$$\text{If } b \geq 1 \text{ and } c > 0 \text{ then } \int_c^{bc} f_\alpha = c^{\alpha+1} \int_1^b f_\alpha.$$

The idea was that scaling a box horizontally by the factor c scales it vertically by c^α, thus giving an area-scaling factor of $c^{\alpha+1}$. In the special case $\alpha = -1$, i.e., in the case of the reciprocal function, the area-scaling factor is trivial,

$$\text{If } b \geq 1 \text{ and } c > 0 \text{ then } \int_c^{bc} f_{-1} = \int_1^b f_{-1}. \tag{5.8}$$

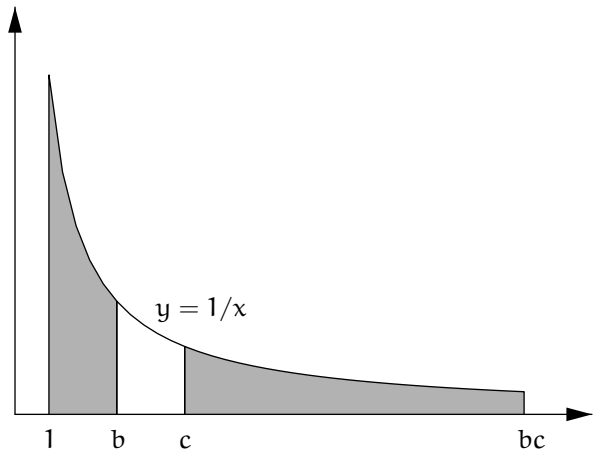


Figure 5.3. Equal areas

So the mnemonic for (5.8) is that we may "cancel" the c. (See figure 5.3. As before, the scaled interval $[c, bc]$, which the figure shows lying entirely to the right of the original interval $[1, b]$, can also lie to the right of the original interval but with overlap, or to the left of the original interval but with overlap, or entirely to the left of the original interval.)

If instead, $0 < b < 1$, while still $c > 0$, then symbolic reasoning shows that the equality in (5.8) still holds, courtesy of (5.6),

$$\begin{aligned}
\int_c^{bc} f_{-1} &= -\int_{bc}^{c} f_{-1} && \text{by (5.6)} \\
&= -\int_{bc}^{(1/b)bc} f_{-1} && \text{and note that } 1/b > 1 \\
&= -\int_1^{1/b} f_{-1} && \text{by (5.8), "canceling" the } bc \\
&= -\int_b^{(1/b)b} f_{-1} && \text{by (5.8), "uncanceling" a factor of } b \\
&= \int_1^{b} f_{-1} && \text{by (5.6) again, because } (1/b)b = 1.
\end{aligned}$$

Because the formula holds regardless of whether $b \geq 1$ or $0 < b < 1$, we have shown:

$$\text{For all } b, c \in \mathcal{R}_{>0}, \quad \int_c^{bc} f_{-1} = \int_1^{b} f_{-1}. \tag{5.9}$$

Again, we have extended a formula to make it symbolically robust, even though the generalized formula is no longer so intuitive geometrically in all cases.

5.1.7 Proof of the Key Property: End of the Proof

Using formulas (5.7) and (5.9), compute that for all $a, b \in \mathcal{R}_{>0}$,

$$\begin{aligned}\int_1^{ab} f_{-1} &= \int_1^a f_{-1} + \int_a^{ab} f_{-1} \quad \text{by (5.7), with } 1,\ a,\ ab \text{ as } a,\ b,\ c \\ &= \int_1^a f_{-1} + \int_1^b f_{-1} \quad \text{by (5.9), "canceling" the } a.\end{aligned}$$

In sum, we have established that

$$\text{For all } a, b \in \mathcal{R}_{>0}, \quad \int_1^{ab} f_{-1} = \int_1^a f_{-1} + \int_1^b f_{-1}. \tag{5.10}$$

But the natural logarithm is by definition

$$\ln : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \qquad \ln(x) = \int_1^x f_{-1}.$$

That is, formula (5.10) is precisely our desired result:

$$\text{For all } a, b \in \mathcal{R}_{>0}, \quad \ln(ab) = \ln(a) + \ln(b).$$

This completes the proof of Theorem 5.1.3.

5.1.8 Further Properties of the Logarithm

The remaining standard properties of the logarithm follow from its definition and its key property of converting multiplication to addition.

Theorem 5.1.4 (Properties of the Logarithm).

(1) $\ln(1) = 0$.

(2) *For all positive real numbers* x *and* $\tilde{x}$,

$$\ln(x\tilde{x}) = \ln(x) + \ln(\tilde{x}).$$

(3) *For all positive real numbers* x

$$\ln(1/x) = -\ln(x).$$

(4) *For all positive real numbers* x *and* $\tilde{x}$,

$$\ln(x/\tilde{x}) = \ln(x) - \ln(\tilde{x}).$$

(5) *For all positive real numbers* x *and all rational numbers* α*,*

$$\ln(x^\alpha) = \alpha \ln(x).$$

Proof. (1) follows from the definition of the logarithm: $\ln(1) = \int_1^1 f_{-1}$ is the area under the graph of the reciprocal function from $x = 1$ to $x = 1$, i.e., it is 0.

(2) is Theorem 5.1.3.

For (3), compute that

$$\begin{aligned} \ln(x) + \ln(1/x) &= \ln(x \cdot 1/x) && \text{by (2)} \\ &= \ln(1) \\ &= 0 && \text{by (1).} \end{aligned}$$

That is, $\ln(1/x)$ is the additive inverse of $\ln(x)$, as desired.

For (4), use (2) and (3),

$$\ln(x/\tilde{x}) = \ln(x \cdot 1/\tilde{x}) = \ln(x) + \ln(1/\tilde{x}) = \ln(x) - \ln(\tilde{x}),$$

as asserted.

For (5), first let $n \in \mathcal{Z}_{\geq 1}$ be a positive integer. Then

$$\begin{aligned} \ln(x^n) &= \ln(x \cdot x \cdots x) \\ &= \ln(x) + \ln(x) + \cdots + \ln(x) && \text{by repeated application of (2)} \\ &= n \ln(x). \end{aligned}$$

Note also that the formula $\ln(x^n) = n \ln(x)$ for $n = 0$ reduces to $0 = 0$, which certainly is true. Next let $n \in \mathcal{Z}_{\leq -1}$ be a negative integer. Because $-n \in \mathcal{Z}_{\geq 1}$,

$$\begin{aligned} \ln(x^n) &= \ln((1/x)^{-n}) \\ &= -n \ln(1/x) && \text{because } -n \in \mathcal{Z}_{\geq 1} \\ &= -n \cdot (-\ln(x)) && \text{by (3)} \\ &= n \ln(x). \end{aligned}$$

Finally, let $\alpha = p/q$ where p and q are integers with $q > 0$. For any $x \in \mathcal{R}_{>0}$, let $\tilde{x} = x^{1/q}$. Then $x^\alpha = \tilde{x}^p$ and $x = \tilde{x}^q$, and so

$$\ln(x^\alpha) = \ln(\tilde{x}^p) = p \ln(\tilde{x}), \qquad \ln(x) = \ln(\tilde{x}^q) = q \ln(\tilde{x}),$$

and consequently

$$\ln(x^\alpha) = \frac{p}{q} \ln(x) = \alpha \ln(x).$$

This completes the argument. □

With the properties of the logarithm established, we can revisit the proof of the nth Root Rule for sequences, and we can prove the nth Power Rule for sequences directly rather than by reference to the nth Root Rule (exercise 5.1.4).

Exercise

5.1.4. (a) Let $b > 1$ be a real number, and let $\varepsilon > 0$ be a real number. Show that for any positive integer n,

$$b^{1/n} - 1 < \varepsilon \iff n > \frac{\ln(b)}{\ln(1+\varepsilon)}.$$

Note that this fact gives another version of the part of the proof of Proposition 3.2.8 (4) that covers the case $b > 1$ (see page 80). Using a computer, choose various values $b > 0$ and $\varepsilon > 0$ (with ε presumably small) to compare the new starting index $N > \ln(b)/\ln(1+\varepsilon)$ from this exercise against the starting index $N > (b-1)/\varepsilon$ in the original proof.

(b) Let r be a nonzero real number such that $|r| < 1$, and let $\varepsilon > 0$ be a real number. Show that for any positive integer n,

$$|r|^n < \varepsilon \iff n > \frac{\ln(\varepsilon)}{\ln(|r|)}.$$

Note that this fact gives another proof of Proposition 3.2.8 (5) (see page 80).

5.2 Logarithmic Growth

Because 2 is bigger than 1, it follows that $\ln(2)$ is positive. And by the properties of the logarithm,

$$\begin{aligned} \ln(4) &= 2\ln(2), \\ \ln(8) &= 3\ln(2), \\ \ln(16) &= 4\ln(2), \end{aligned}$$

and so on. It follows that as n grows large, so does $\ln(2^n)$. However, the sequence

$$(2^n) = (2,\ 2^2,\ 2^3,\ 2^4,\ \ldots)$$

is doubling at each generation, growing faster and faster, whereas the corresponding sequence of logarithms,

$$(\ln(2^n)) = (\ln(2),\ 2\ln(2),\ 3\ln(2),\ 4\ln(2),\ \ldots) = \ln(2)\cdot(n)$$

is growing steadily in proportion to the generation-count. Both sequences get large with n, but apparently at different rates.

Moving from a generation-count n to a continuous variable x, any $x \geq 1$ lies between 2^n and 2^{n+1} for some n. Thus the ratio $\ln(x)/x$ is the quotient

of two large numbers when x is large, so that we do not immediately know how it behaves as x grows, but we suspect that

$$\boxed{\frac{\ln(x)}{x} \text{ tends to } 0 \text{ as } x \text{ gets large.}}$$

In more suggestive notation, albeit involving a verboten symbol, we suspect that

$$\boxed{\lim_{x\to\infty} \frac{\ln(x)}{x} = 0.}$$

(To make the notation sensible, define for any function $f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}$,

$$\lim_{x\to\infty} f(x) = \lim_{s\to 0} f(1/s).$$

That is, the left limit exists if the right limit does, in which case it takes its value from the right limit.) To quantify our sense that the boxed statement holds, first see figure 5.4. Recall that $\ln x$ is the area under the $y = 1/x$ curve from 1 to x, the darker shaded area in the figure. On the other hand, x itself is the area of the box in the figure, the remainder of which is shaded more lightly. As the box grows rightward, the darker shaded area becomes negligible as a portion of the total shaded area.

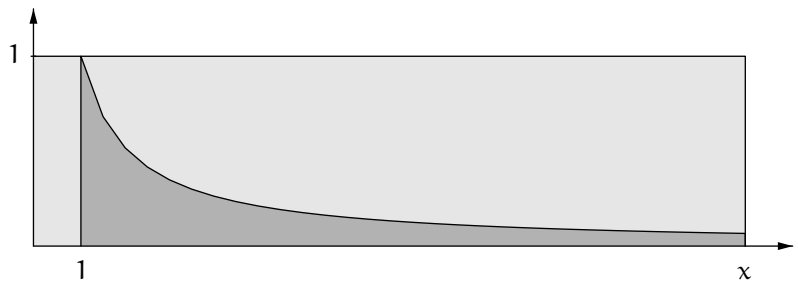


Figure 5.4. $\ln(x)$ as a portion of x

To further quantify the argument, see figure 5.5. Again, $\ln(x)$ is the area from 1 to x. Given $\varepsilon > 0$ (and also $\varepsilon < 1$: as usual, ε is a small positive number), this area is less than the area of the two boxes in the figure, the excess area being shown as lighter gray. Thus, for any $x > 2/\varepsilon$, which presumably is large, further enlarging the two boxes by extending each of them leftward to the y-axis gives the estimate

$$\ln x < \frac{2}{\varepsilon} + \frac{x\varepsilon}{2}.$$

It follows that

$$\frac{\ln(x)}{x} < \frac{1}{x} \cdot \frac{2}{\varepsilon} + \frac{\varepsilon}{2},$$

and so, noting that if $x > (2/\varepsilon)^2$ then $1/x < (\varepsilon/2)^2$,

$$\frac{\ln(x)}{x} < \varepsilon \qquad \text{for all } x > (2/\varepsilon)^2.$$

Here $x > (2/\varepsilon)^2$ presumably is *very* large. But such large x gives $\ln(x)/x < \varepsilon$ where $\varepsilon > 0$ was given to us, and so the desired conclusion follows.

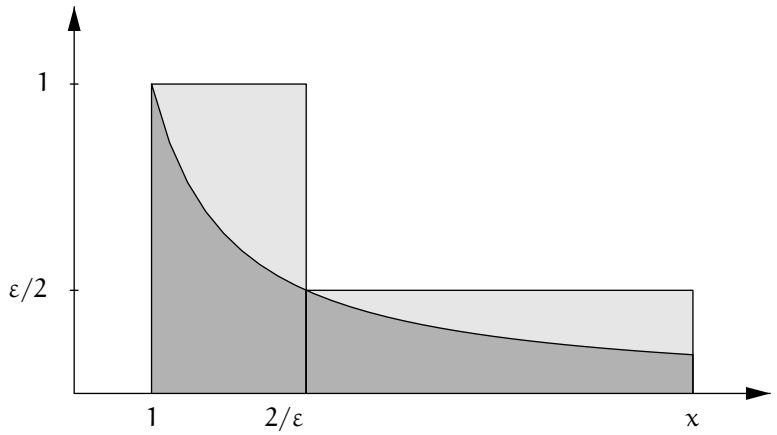


Figure 5.5. $\ln(x)$ is less than the two box-areas

Although the notion "$x \to \infty$" is outside the scope of our formal arguments, we can restrict the previous discussion from the continuous variable x to a discrete variable n to get a formal result:

Proposition 5.2.1 (The Logarithm Grows Slowly). *We have the sequence limit*

$$\lim_n \frac{\ln(n)}{n} = 0.$$

Exercises

5.2.1. Let n be a positive integer. Draw a figure that shows clearly why boxes of base 1 on the x-axis and heights $1, 1/2, 1/3, \ldots$ have their tops above the graph of the reciprocal function f_{-1}. Explain why this shows that

$$1 > \ln(2),$$
$$1 + \frac{1}{2} > \ln(3),$$
$$1 + \frac{1}{2} + \frac{1}{3} > \ln(4),$$

and in general

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} > \ln(n+1).$$

This shows that the sum $1 + 1/2 + 1/3 + \cdots + 1/n$ grows without bound as n grows.

5.2.2. (a) Use a box to show that $\ln(2) < 1$.

(b) Use a box to show that $1/2 < \ln(2)$. Explain why it follows that $1 < \ln(4)$.

(c) By its definition, the logarithm function is strictly increasing. Therefore there is one, and only one, number e such that $\ln(e) = 1$. Parts (a) and (b) of this exercise have shown that $2 < e < 4$. Use boxes to show that in fact $e < 3$, by showing that $1 < \ln(3)$. Explain carefully whether using inner or outer boxes is the correct choice for the argument, and why. How many boxes does your argument require?

5.2.3. (a) How does $\ln(x^{10})/x$ behave as x grows large?

(b) How does $\ln(x)/x^{1/10}$ behave as x grows large?

5.3 Differentiation of the Logarithm

Theorem 5.3.1 (Derivative of the Logarithm). *The logarithm function*

$$\ln : \mathcal{R}_{>0} \longrightarrow \mathcal{R},$$

is differentiable on its entire domain, and its derivative is the reciprocal function,

$$\ln' = f_{-1} : \mathcal{R}_{>0} \longrightarrow \mathcal{R}.$$

That is,

$$\ln'(x) = 1/x \quad \textit{for all } x \in \mathcal{R}_{>0}.$$

Consequently, also

$$\lim_{s \to x} \ln(s) = \ln(x) \quad \textit{for all } x \in \mathcal{R}_{>0}.$$

Proof. The bulk of the work is to establish the normalized result $\ln'(1) = 1$,

$$\lim_{s \to 1} \frac{\ln(s) - \ln(1)}{s - 1} = 1.$$

To study the limit in the display, consider first s-values greater than 1, and consider figure 5.6. The relative areas of the small box, the region under the graph of f_{-1}, from 1 to s, and the large box show that three positive quantities can be listed in order,

$$\frac{s-1}{s} \leq \ln(s) \leq s - 1 \quad \text{for } s > 1.$$

Therefore, dividing through by the positive quantity $s - 1$ gives

$$\frac{1}{s} \leq \frac{\ln(s)}{s-1} \leq 1 \quad \text{for } s > 1. \tag{5.11}$$

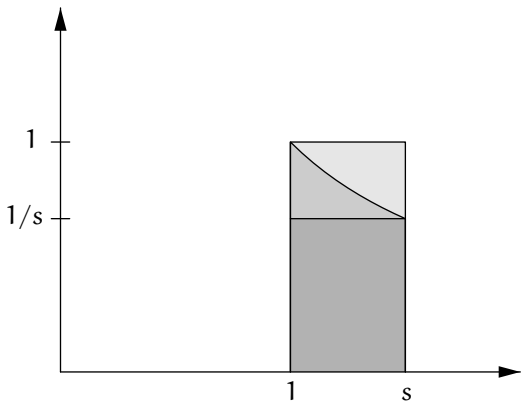


Figure 5.6. Bounds on $\ln(s)$ for $s > 1$

Next consider s-values less than 1, though of course still positive. Figure 5.7 shows that now we can list the same three quantities as before in increasing order, although this time they increase from the most negative to the least negative,

$$\frac{s-1}{s} \leq \ln(s) \leq s - 1 \quad \text{for } 0 < s < 1.$$

The fact that the same quantities bound $\ln(s)$ regardless of whether $s > 1$ or $0 < s < 1$ is an instance of symbolic robustness. But this time, dividing through by the negative quantity $s - 1$ reverses the inequalities,

$$1 \le \frac{\ln(s)}{s-1} \le \frac{1}{s} \quad \text{for } 0 < s < 1.$$

Recall that $\ln(1) = 0$. Inequality (5.11) and the previous display combine into a single estimate that covers the cases $s > 1$ and $0 < s < 1$,

$$\min\{1, 1/s\} \le \frac{\ln(s) - \ln(1)}{s-1} \le \max\{1, 1/s\} \quad \text{for all } s > 0 \text{ except } s = 1.$$

As s goes to 1, so does $1/s$, by the Reciprocal Rule for function limits (Theorem 4.1.6, part (4)). Also, the distances from $\min\{1, 1/s\}$ and $\max\{1, 1/s\}$ to 1 are at most the distance from $1/s$ to 1. From the estimate in the previous display and from these observations, the Squeezing Rule for functions (Theorem 4.1.8) gives

$$\lim_{s \to 1} \frac{\ln(s) - \ln(1)}{s-1} \text{ exists and equals } 1$$

That is, we have shown that $\ln'(1)$ exists and equals 1.

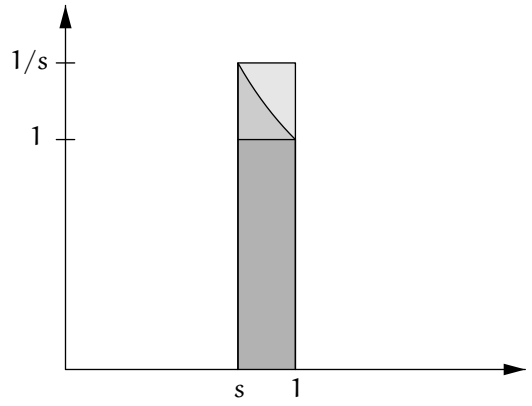


Figure 5.7. Bounds on the logarithm as s tends to 1 from the left

To finish the proof that $\ln'(x) = 1/x$ in general, let x be any positive real number. For any positive real number $s \neq x$, the logarithm properties $\ln(s) - \ln(x) = \ln(s/x)$ and $\ln(1) = 0$ combine with basic algebra to give

$$\begin{aligned} \frac{\ln(s) - \ln(x)}{s-x} &= \frac{1}{x} \cdot \frac{\ln(s/x) - \ln(1)}{s/x - 1} \\ &= \frac{1}{x} \cdot \frac{\ln(\tilde{s}) - \ln(1)}{\tilde{s} - 1} \quad \text{where } \tilde{s} = s/x. \end{aligned}$$

As s tends to x, $\tilde{s}$ tends to 1. And so by the previous calculation,

$$\text{For all } x \in \mathcal{R}_{>0}, \quad \lim_{s \to x} \frac{\ln(s) - \ln(x)}{s - x} = \frac{1}{x}.$$

This completes the proof of Theorem 5.3.1. □

The fact that $\ln' = f_{-1}$ on $\mathcal{R}_{>0}$ once again shows us a connection between integration and differentiation: the logarithm was defined by integrating the reciprocal function from startpoint 1 to a variable endpoint x, and the derivative of the logarithm at x is the reciprocal function there. More specifically, compute for any $a, b \in \mathcal{R}_{>0}$ that

$$\int_a^b f_{-1} = \int_a^1 f_{-1} + \int_1^b f_{-1} = \int_1^b f_{-1} - \int_1^a f_{-1} = \ln(b) - \ln(a).$$

Because $\ln' = f_{-1}$, this shows that formula (5.2) (page 147) extends to the case $\alpha = -1$ as well. Given $\alpha \in \mathcal{Q}$, define

$$F : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \quad F = \begin{cases} f_{\alpha+1}/(\alpha+1) & \text{if } \alpha \neq -1, \\ \ln & \text{if } \alpha = -1. \end{cases}$$

Then $F' = f_\alpha$ in all cases, and

$$\boxed{\int_a^b f_\alpha = F(b) - F(a), \quad \alpha \in \mathcal{Q},\ a, b \in \mathcal{R}_{>0}.}$$

Exercises

5.3.1. Let A be a subset of $\mathcal{R}$, let $\mathcal{R}_{\neq 0} = \{x \in \mathcal{R} : x \neq 0\}$, and let

$$f : A \longrightarrow \mathcal{R}_{\neq 0}$$

be any differentiable function. Prove that the function

$$g : A \longrightarrow \mathcal{R}, \quad g(x) = \ln(|f(x)|)$$

is differentiable and its derivative is

$$g'(x) = \frac{f'(x)}{f(x)}.$$

You will need to use the Chain Rule (Theorem 4.2.9, page 139) twice, and you will need to use Proposition 4.2.6 (page 135).

5.3.2. Find the derivatives of the following functions $f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}$.

(a) $f(x) = \ln(cx)$ where $c \in \mathcal{R}_{>0}$.
(b) $f(x) = \ln(x^\alpha)$ where $\alpha \in \mathcal{Q}$.
(c) $f(x) = x^\alpha \ln(cx^\beta)$ where $\alpha, \beta \in \mathcal{Q}$ and $c \in \mathcal{R}_{>0}$.
(d) $f(x) = (\ln(x))^2$. (Note that $(\ln(x))^2$ is not $\ln(x^2)$.)
(e) $f(x) = (\ln(x))^3$.
(f) $f(x) = (\ln(x))^n$ where $n \in \mathcal{Z}_{\geq 1}$.
(g) $f(x) = (\ln(x+1))^\alpha$ where $\alpha \in \mathcal{Q}$.
(h) $f(x) = \ln(\ln(x+3))$.

5.3.3. (a) The tangent line to the graph of the logarithm at $x = 1$ passes through the point $(1, 0)$ and has slope $1/1 = 1$, and so it is the graph of the function

$$t(x) = x - 1, \quad x \in \mathcal{R}_{>0}.$$

Explain why

$$t(x) = \int_1^x f_0 \quad \text{for } x \geq 1.$$

(b) Explain why $f_0(x) \geq f_{-1}(x)$ for all $x \geq 1$, with equality only for $x = 1$.

(c) Explain why parts (a) and (b) argue that that the tangent line to the graph of the logarithm at $x = 1$ lies above the graph to the right of $x = 1$.

(d) Recalling Definition 5.1.1, explain why also

$$t(x) = \int_1^x f_0 \quad \text{for } 0 < x < 1.$$

(e) Explain why $f_0(x) \leq f_{-1}(x)$ for all x such that $0 < x \leq 1$, with equality only for $x = 1$.

(f) Explain why despite the inequality in (e) pointing the other direction from the inequality in (b), parts (d) and (e) argue that that the tangent line to the graph of the logarithm at $x = 1$ lies above the graph to the left of $x = 1$ as well.

(g) Now let $x_0 \in \mathcal{R}_{>0}$ be any positive number, and consider the tangent line to the graph of the logarithm at x_0. Explain why it is the graph of the function

$$t(x) = \ln(x_0) + \frac{1}{x_0}(x - x_0).$$

Show that

$$t(x) - \ln(x) = \ln(x/x_0) - \left(\frac{x}{x_0} - 1\right).$$

Explain why therefore the earlier portions of this exercise show that $t(x) - \ln(x) \geq 0$, with equality only for $x = x_0$. That is, for any $x_0 \in \mathcal{R}_{>0}$, the tangent line to the graph of the logarithm at $x = x_0$ lies above the graph.

5.4 Integration of the Logarithm

The logarithm function is integrated similarly to the power function, but with some new wrinkles. The work requires evaluating a variant of the finite geometric sum and then noting that we already know how to evaluate a resulting limit, because it is a derivative.

Thus, we pre-emptively compute the sum that will arise in the calculation; this sum can be computed using calculus, as has already been done as an exercise and now will be repeated in the text, or using only algebra, as in an exercise to follow. With the sum in hand, we integrate the logarithm from 1 to b where $b > 1$. Finally we integrate the logarithm between general endpoints $a, b \in \mathcal{R}_{>0}$. This last calculation requires expanding our notion of integral, because part of the logarithm graph lies below the x-axis.

5.4.1 Another Summation Formula

To integrate the logarithm, we will need to evaluate the sum

$$\sigma(x) = 1 + 2x + 3x^2 + \cdots + nx^{n-1}, \quad x \neq 1.$$

Evaluating this sum was exercise 4.2.11, but we repeat the work here.

Let n be a nonnegative integer and let

$$h(x) = x + x^2 + x^3 + \cdots + x^n.$$

Repeatedly applying the Sum Rule for derivatives and the power function derivative formula gives

$$h'(x) = 1 + 2x + 3x^2 + \cdots + nx^{n-1} = \sigma(x), \quad x \neq 1.$$

But also, factoring out a power of x and then using the finite geometric sum,

$$h(x) = (1 + x + x^2 + \cdots + x^{n-1})x = \frac{x^n - 1}{x - 1}x, \quad x \neq 1.$$

Thus $h(x) = f(x) \cdot g(x)$ where $f(x) = x^n - 1$ and $g(x) = x/(x-1)$, and so the Product Rule for derivatives and then the Quotient Rule and other rules combine to give that for $x \neq 1$,

$$\begin{aligned} h'(x) &= f'(x)g(x) + f(x)g'(x) \\ &= \frac{nx^{n-1}x}{x-1} + (x^n - 1)\frac{x - 1 - x}{(x-1)^2} \\ &= \frac{nx^n}{x-1} - \frac{x^n - 1}{(x-1)^2}. \end{aligned}$$

Equate the two expressions for h' to evaluate $\sigma(x)$,

$$1 + 2x + 3x^2 + \cdots + nx^{n-1} = \frac{nx^n}{x-1} - \frac{x^n - 1}{(x-1)^2}, \quad x \neq 1. \tag{5.12}$$

The next exercise shows how to evaluate this sum without using calculus.

Exercise

5.4.1. A moment ago we used calculus to evaluate the sum

$$\sigma(x) = 1 + 2x + 3x^2 + \cdots + nx^{n-1}, \quad x \neq 1,$$

but in fact the sum does not require calculus. Instead, compute that

$$\begin{aligned}
\sigma(x) = 1 &+ x + x^2 + \cdots + x^{n-1} \\
&+ x + x^2 + \cdots + x^{n-1} \\
&\quad + x^2 + \cdots + x^{n-1} \\
&\qquad \ddots \quad \vdots \\
&\qquad\qquad + x^{n-1}.
\end{aligned}$$

That is,

$$\begin{aligned}
\sigma(x) = (1 &+ x + x^2 + \cdots + x^{n-1}) \\
&+ x(1 + x + \cdots + x^{n-2}) \\
&+ x^2(1 + \cdots + x^{n-3}) \\
&+ \cdots \\
&+ x^{n-1}.
\end{aligned}$$

Recall that $x \neq 1$, apply the finite geometric sum formula, and do some algebra to rederive (5.12).

5.4.2 The Normalized Case

Again let b be a real number greater than 1. To integrate the logarithm function from 1 to b, first use a geometric partition and right endpoints, thus creating an upper sum. Again n is the number of boxes, and $s = b^{1/n}$ so that $s^n = b$, and the jth interval-width is $(s-1)s^j$ for $j = 0, \ldots, n-1$. However, now

$$\text{the } j\text{th box-height is } \ln(s^{j+1}) = (j+1)\ln(s) \text{ for } j = 0, \ldots, n-1,$$

and so the sum of the box-areas is

$$\begin{aligned} T_n &= (s-1)\ln(s)\left[1s^0 + 2s^1 + 3s^2 + \cdots + ns^{n-1}\right] \\ &= (s-1)\ln(s)\left[1 + 2s + 3s^2 + \cdots + ns^{n-1}\right]. \end{aligned}$$

We just found the sum at the end of the previous display. By (5.12), and using the fact that $n\ln(s) = \ln(s^n)$,

$$T_n = (s-1)\ln(s)\left(\frac{ns^n}{s-1} - \frac{s^n-1}{(s-1)^2}\right) = s^n\ln(s^n) - (s^n-1)\frac{\ln(s)}{s-1},$$

and thus, because $s = b^{1/n}$ and so $s^n = b$,

$$T_n = b\ln(b) - (b-1)\frac{\ln(s)}{s-1}. \tag{5.13}$$

Now to find $\lim_n(T_n)$ remember that as n grows large, s goes to 1, and so because

$$\lim_{s\to 1}\frac{\ln(s)}{s-1} = \lim_{s\to 1}\frac{\ln(s)-\ln(1)}{s-1} = \ln'(1) = 1,$$

it follows from (5.13) that

$$\lim_n(T_n) = b\ln(b) - b + 1.$$

As with the rational power function, the end of thisn integral calculation uses a normalizied derivative value.

The sequence (S_n) of lower sums obtained by taking function values at left endpoints is such that $\lim_n(T_n - S_n) = 0$ (exercise 5.4.2). Therefore the logarithm function is integrable from 1 to b, and its integral is the limit $\lim_n(T_n)$,

$$\int_1^b \ln = b\ln(b) - b + 1, \quad b > 1.$$

This is the area under the logarithm curve from $x = 1$ to $x = b$ (see figure 5.8).

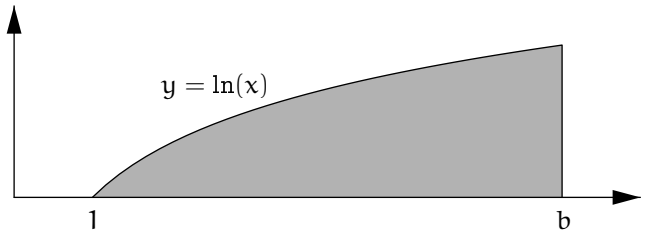


Figure 5.8. Area under the logarithm curve

Exercise

5.4.2. In the calculation just completed, each T_n is a sum of outer box-areas because we determined the box-heights by sampling the logarithm function over the right endpoints of the box-bases,

$$T_n = (s-1)\ln(s)\left[1s^0 + 2s^1 + 3s^2 + 4s^3 + \cdots + ns^{n-1}\right].$$

Show that if instead we use the left endpoints then the relevant sum of n inner box-areas is

$$S_n = (s-1)\ln(s)\left[0s^0 + 1s^1 + 2s^2 + 3s^3 + \cdots + (n-1)s^{n-1}\right].$$

Find $T_n - S_n$ and then explain why $\lim_n(T_n - S_n) = 0$. (It may be helpful to recall that $\ln(s) = \ln(b)/n$.)

5.4.3 The General Case

The opening idea of this chapter was to begin moving from geometric intuition to symbolic intuition by defining under suitable circumstances

$$\int_a^b f = -\int_b^a f \quad \text{if } a > b.$$

This definition worked perfectly well symbolically, but geometrically it called on us to expand our notion of the integral, thinking of it as a sort of *signed* area, depending on which direction we traverse the x-axis horizontally. Now, the fact that the logarithm function takes positive and negative values (as compared to the power function on $\mathcal{R}_{>0}$ which is always positive) is incentive to extend our notion of the integral as signed area *vertically* as well. That is, the closing idea of this chapter is to view area below the x-axis as negative area when the x-axis is traversed in the positive direction, i.e., from left to right. Because a negative times a negative is a positive, area below the x-axis is positive when the x-axis is traversed from right to left. The methodology that makes all the cases uniform is to think of the integral analytically as a limit of sums.

Let a and b be positive real numbers in order. That is, $0 < a \leq b$. We want to discuss the integral

$$\int_a^b \ln.$$

Although the logarithm could take positive and/or negative values on $[a, b]$, depending on the values of a and b, its output-values are in any case trapped between two numbers, $\ln(a)$ and $\ln(b)$. As shown in figure 5.9, it is geometrically natural to interpret the integral is the sum of two possible quantities,

a possible negative quantity, the negative of the area between the graph of the logarithm and the x-axis under the portion of the x-axis from a to the smaller of b and 1 (this term is present only if $0 < a < 1$),

and

a possible positive quantity, the area between the graph of the logarithm and the x-axis over the portion of the x-axis from the larger of a and 1 to b (this term is present only if $b > 1$).

Figure 5.9 shows a scenario where both terms are present because $a < 1 < b$.

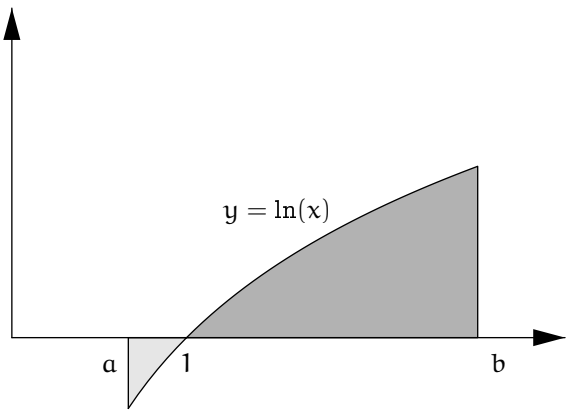


Figure 5.9. Integral of the logarithm as a sum of positive and negative areas

Putting aside the geometrically natural for a moment, we proceed in a way that is analytically natural. Because $0 < a \le b$, it follows that $b/a \ge 1$. Let (T_n) be the sequence of upper sums for $\mathrm{Ar}_1^{b/a}(\ln)$ used to compute $\int_1^{b/a} \ln$ in the previous subsection, where the b there is the b/a here. Thus for each $n \in \mathcal{Z}_{\ge 1}$ there are partition points

$$1 = s^0 < s^1 < \cdots < s^{n-1} < s^n = b/a \quad (\text{where } s = (b/a)^{1/n}),$$

and jth boxes for $j = 0, \ldots, n-1$ having bases and heights

$$\beta_j = (s-1)s^j, \qquad h_j = \ln(s^{j+1}).$$

The box-areas are

$$A_j = \beta_j h_j, \quad j = 0, \ldots, n-1,$$

and the box-area sum is

$$T_n = A_0 + \cdots + A_{n-1}.$$

Now scale the partition points by a to get new partition points as^j for $j = 0, \ldots, n$,

$$a = as^0 < as^1 < \cdots < as^{n-1} < as^n = b,$$

new box-bases and box-heights for $j = 0, \ldots, n-1$,

$$\tilde{\beta}_j = a\beta_j, \qquad \tilde{h}_j = \ln(as^{j+1}) = \ln(a) + \ln(s^{j+1}) = \ln(a) + h_j,$$

and hence new box-areas

$$\widetilde{A}_j = a\ln(a)\beta_j + aA_j, \quad j = 0, \ldots, n-1.$$

The new box-areas sum to

$$\widetilde{T}_n = \widetilde{A}_0 + \cdots + \widetilde{A}_{n-1} = a\ln(a)(\beta_0 + \cdots + \beta_{n-1}) + a(A_0 + \cdots + A_{n-1}).$$

The sum $\beta_0 + \cdots + \beta_{n-1}$ of the original box-bases is $b/a - 1$, and the second sum in the previous display is T_n, so we have shown that the new box-areas sum to

$$\widetilde{T}_n = a(b/a - 1)\ln(a) + aT_n = (b-a)\ln(a) + aT_n.$$

Consequently, our candidate analytic value of $\int_a^b \ln$ is

$$\begin{aligned}
\lim_n(\widetilde{T}_n) &= (b-a)\ln(a) + a\int_1^{b/a} \ln \\
&= (b-a)\ln(a) + a((b/a)\ln(b/a) - b/a + 1) \\
&= b\ln(a) - a\ln(a) + b\ln(b) - b\ln(a) - b + a \\
&= \big(b\ln(b) - b\big) - \big(a\ln(a) - a\big).
\end{aligned}$$

The question is whether this analytic work is compatible with the natural geometric interpretation of the integral $\int_a^b \ln$ when $0 < a \le b$. We cannot generally interpret $\widetilde{T}_n$ as an upper sum because the graph of the logarithm, and the boxes, can lie below the x-axis during some or all of the calculation. However, the upper sum idea *can* be salvaged by hoisting the graph and the boxes vertically to place the region of interest just above the x-axis. That is, rather than study the logarithm itself, we can study its vertical translate,

$$g : [a, b] \longrightarrow [0, \ln(b) - \ln(a)], \qquad g(x) = \ln(x) - \ln(a).$$

Subtracting $\ln(a)$ raises the graph when $\ln(a)$ is negative, as shown in figure 5.10. Letting the partition points $\tilde{x}_j = as^j$ of $[a, b]$ be as in the analytic work of the previous paragraph, each jth outer box for g over $[\tilde{x}_j, \tilde{x}_{j+1}]$ has

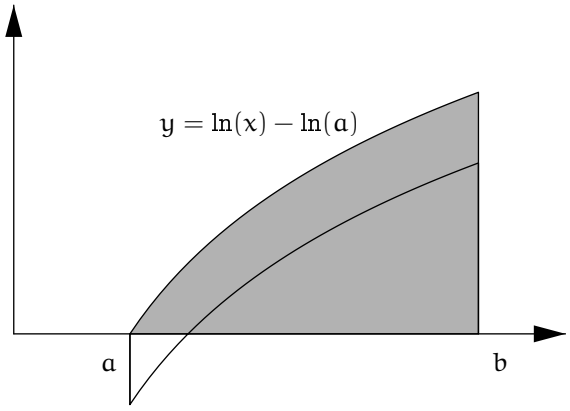


Figure 5.10. Hoisting the logarithm graph

base $\tilde{\beta}_j = a\beta_j$ as before, and height $\tilde{h}_j - \ln(a) = h_j$. Thus the resulting upper sum for g is aT_n, and in the limit,

$$\int_a^b g = a \int_1^{b/a} \ln .$$

Because we view $\int_a^b g$ as obtained by subtracting the signed rectangle-area $(b-a)\ln(a)$ from $\int_a^b \ln$, we obtain $\int_a^b \ln$ by reversing the process, i.e., adding the correction term $(b-a)\ln(a)$ to the computed value of $\int_a^b g$. Geometrically, we are making $\int_a^b \ln$ the sum of the dark area and the negative of the light area in the left side of figure 5.11, the correct value because it is sum of the dark area and the negative of the light area in the right side of the figure. Adding the correction term to the right side of the previous display gives

$$\int_a^b \ln = (b-a)\ln(a) + a \int_1^{b/a} \ln .$$

This is the same expression that we got for $\int_a^b \ln$ analytically above, and so without further ado we decree that its already-computed value is indeed the integral,

$$\int_a^b \ln = (b\ln(b) - b) - (a\ln(a) - a), \quad 0 < a \le b.$$

If instead, $0 < b < a$ then, naturally,

$$\begin{aligned} \int_a^b \ln = -\int_b^a \ln &= -\big((a\ln(a) - a) - (b\ln(b) - b)\big) && \text{by the previous display} \\ &= \big(b\ln(b) - b\big) - \big(a\ln(a) - a\big) && \text{again,} \end{aligned}$$

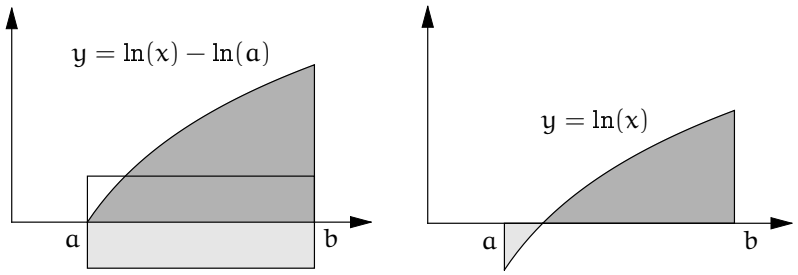


Figure 5.11. Signed area as difference of positive areas

so that, robustly,

$$\boxed{\int_a^b \ln = \big(b\ln(b) - b\big) - \big(a\ln(a) - a\big), \quad a, b \in \mathcal{R}_{>0}.}$$

Especially when $a = 1$, this formula extends the formula from the end of the previous subsection, now giving

$$\int_1^b \ln = b\ln(b) - b + 1, \quad b \in \mathcal{R}_{>0}.$$

Because a and b can be in either order, and because the logarithm function takes positive and negative values, we summarize how to track the boxed formula for the integral as a signed area. First, area between the x-axis and the graph of the logarithm is deemed positive when the graph lies above the axis, negative when it lies below, and the integral is the net signed area if $0 < a \leq b$. However, if $0 < b < a$ then the integral is the negative of the net positive area. All of this is easier to understand visually than verbally. But because the boxed formula is insensitive to the cases, symbolic understanding is truly the easiest of all in this context.

5.4.4 The Fundamental Theorem of Calculus Again

To end this section, consider the function

$$F : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \quad F(x) = x\ln(x) - x.$$

We have established the formula

$$\int_a^b \ln = F(b) - F(a), \quad a, b \in \mathcal{R}_{>0}.$$

By now the reader anticipates that the derivative of F is the function whose integral F is, i.e., $F' = \ln$. To see this, note first that F is built from the differentiable functions

$$f_1, \ln : \mathcal{R}_{>0} \longrightarrow \mathcal{R}.$$

Various results about derivatives confirm that indeed $F' = \ln$ (exercise 5.4.3).

Exercise

5.4.3. Complete the verification that the function F just given indeed has derivative $F' = \ln$.

5.5 Signed Integration in General

5.5.1 The Integral Revisited

The following definition captures the expanded notion of the integral that emerged during this chapter in the course of integrating the logarithm.

Definition 5.5.1. *Let* a *and* b *be real numbers with* $a < b$. *Let* L *and* M *be real numbers with* $L \leq M$. *Let*

$$f : [a, b] \longrightarrow [L, M]$$

be a function. If the vertically translated function with nonnegative outputs

$$g : [a, b] \longrightarrow [0, M - L], \quad g(x) = f(x) - L$$

is integrable from a *to* b, *then the integral of* f *from* a *to* b *is defined to be*

$$\int_a^b f = \int_a^b g + (b - a)L.$$

And then, as before, also the integral with out-of-order endpoints is defined to be

$$\int_b^a f = -\int_a^b f.$$

The first part of Definition 5.5.1 raises a problem. Its formula

$$\int_a^b f = \int_a^b g + (b - a)L$$

makes use of the datum L from the codomain of f, but this codomain can be chosen with considerable flexibility. If we keep the domain and the rule but change the codomain to get

$$f : [a, b] \longrightarrow [\widetilde{L}, \widetilde{M}],$$

then on the face of it, the definition could prescribe a different value of $\int_a^b f$. We need to ensure that this doesn't happen.

Geometrically, the issue is as follows. The first part of Definition 5.5.1 says to hoist the graph of f above the x-axis, compute the integral under the hoisted graph, and then adjust by an amount corresponding to the area gained by the hoisting. (Again, see figure 5.11.) What we need to verify is that hoisting the graph by a different amount and then making the corresponding different adjustment leads to the same answer. Phrasing the matter in these geometric terms makes the desired conclusion essentially inescapable: hoisting the graph higher adds a rectangular region between the x-axis and the hoisted graph, but the correction factor will be modified by an amount equal to the area of the added rectangular region, leading to the same value for the signed area as before. (See figure 5.12.)

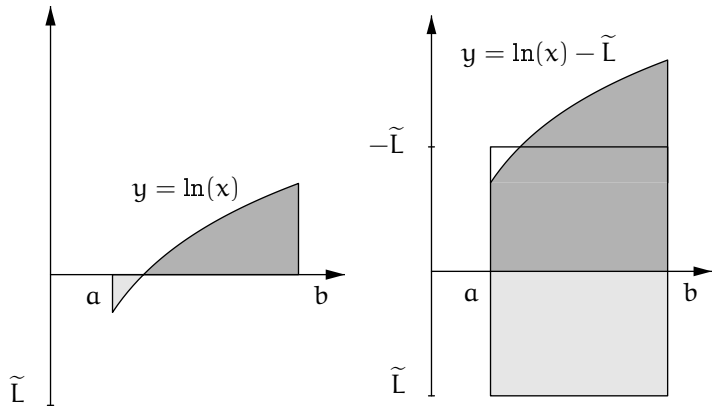


Figure 5.12. Same signed area as another difference of positive areas

To verify analytically that the definition makes sense, we proceed as follows. To repeat, for the original choice of codomain we have

$$f : [a, b] \longrightarrow [L, M],$$

and then

$$g : [a, b] \longrightarrow [0, M - L], \quad g(x) = f(x) - L,$$

and then

$$\int_a^b f = \int_a^b g + (b - a)L.$$

For the modified choice of codomain we have

$$f : [a, b] \longrightarrow [\widetilde{L}, \widetilde{M}] \quad \text{(same domain and same rule)},$$

and then

$$\tilde{g} : [a, b] \longrightarrow [0, \widetilde{M} - \widetilde{L}], \quad \tilde{g}(x) = f(x) - \widetilde{L},$$

and then also

$$\int_a^b f = \int_a^b \tilde{g} + (b - a)\widetilde{L}.$$

We want to show that the two definitions of $\int_a^b f$ are equal. By symmetry we may assume that $\widetilde{L} \geq L$. Thus for all $x \in [a, b]$,

$$g(x) - \tilde{g}(x) = (f(x) - L) - (f(x) - \widetilde{L}) = \widetilde{L} - L,$$

or

$$g = \tilde{g} + (\widetilde{L} - L).$$

Because g and $\tilde{g}$ are both nonnegative functions, and $\widetilde{L} - L$ is a nonnegative constant, the fact that area has sensible properties says that

$$\int_a^b g = \int_a^b \tilde{g} + (b - a)(\widetilde{L} - L).$$

Consequently, the first definition of $\int_a^b f$ as $\int_a^b g + (b - a)L$ becomes

$$\int_a^b f = \int_a^b \tilde{g} + (b - a)(\widetilde{L} - L) + (b - a)L = \int_a^b \tilde{g} + (b - a)\widetilde{L}.$$

This is the second definition of f, and we are done showing that Definition 5.5.1 is well defined when $a \leq b$. Because the definition of $\int_a^b f$ when $a > b$ is given in terms of $\int_a^b f$ when $a \leq b$, it is well defined too.

Exercise

5.5.1. (a) Draw a figure illustrating why it is geometrically obvious that $\int_{-1}^{1} f_1 = 0$.

(b) Carry out the process of computing $\int_{-1}^{1} f_1$ by hoisting.

5.5.2 Progenitive Integral Rules Revisited

With integration now defined for functions whose values could be negative, and for out-of-order endpoints, we need to revisit the results that (suppressing the fine print)

- the integral of a sum is the sum of the integrals, and

- the integral of a constant multiple is the constant multiple of the integral.

The proofs of the upgraded results should not repeat the earlier proofs for nonnegative functions, but rather should cite them and carry out only the new work required now for the upgrades. In other words, the required work is to push the previous results through the hoisting process that defines integrals of functions whose values could be negative.

Proposition 5.5.2 (Progenitive Integral Rules). *Consider two integrable functions*

$$f : [a, b] \longrightarrow [L, M], \qquad \tilde{f} : [a, b] \longrightarrow [\widetilde{L}, \widetilde{M}].$$

Then the function

$$f + \tilde{f} : [a, b] \longrightarrow [L + \widetilde{L}, M + \widetilde{M}], \quad (f + \tilde{f})(x) = f(x) + \tilde{f}(x)$$

is integrable, and

$$\int_a^b (f + \tilde{f}) = \int_a^b f + \int_a^b \tilde{f}.$$

Let $c \in \mathcal{R}_{\geq 0}$ be a nonnegative real number. Then the function

$$cf : [a, b] \longrightarrow [cL, cM], \quad (cf)(x) = c \cdot f(x)$$

is integrable, and

$$\int_a^b (cf) = c \int_a^b f.$$

Let $c \in \mathcal{R}_{<0}$ be a negative real number. Then the function

$$cf : [a, b] \longrightarrow [cM, cL], \quad (cf)(x) = c \cdot f(x)$$

is integrable, and

$$\int_a^b (cf) = c \int_a^b f.$$

Finally, the three equalities hold if instead $a > b$.

Proof. Define

$$g : [a, b] \longrightarrow [0, M - L], \quad g(x) = f(x) - L$$

and

$$\tilde{g} : [a, b] \longrightarrow [0, \widetilde{M} - \widetilde{L}], \quad \tilde{g}(x) = \tilde{f}(x) - \widetilde{L}.$$

Then their sum

$$g + \tilde{g} : [a, b] \longrightarrow [0, M + \widetilde{M} - L - \widetilde{L}]$$

is

$$(g+\tilde{g})(x) = (f+\tilde{f})(x) - (L+\widetilde{L}).$$

Therefore

$$\begin{aligned}\int_a^b (f+\tilde{f}) &= \int_a^b (g+\tilde{g}) + (b-a)(L+\widetilde{L}) \\ &= \int_a^b g + (b-a)L + \int_a^b \tilde{g} + (b-a)\widetilde{L} \\ &= \int_a^b f + \int_a^b \tilde{f}.\end{aligned}$$

(The integral on the right side of the top line exists and is the sum of the two integrals on the second line by Proposition 3.3.13 on page 116 because *they* exist and g and $\tilde{g}$ are nonnegative, and they exist because f and $\tilde{f}$ are given to be integrable.)

For $c \in \mathcal{R}_{\geq 0}$, define

$$cg : [a,b] \longrightarrow [0, c(M-L)], \quad (cg)(x) = (cf)(x) - cL.$$

Then

$$\begin{aligned}\int_a^b (cf) &= \int_a^b (cg) + (b-a)cL \\ &= c\int_a^b g + c(b-a)L \quad \text{by Proposition 3.3.13} \\ &= c\left(\int_a^b g + (b-a)L\right) \\ &= c\int_a^b f.\end{aligned}$$

To establish the result for $c \in \mathcal{R}_{<0}$, we may show that

$$\int_a^b (-f) = -\int_a^b f,$$

because any negative number is a positive multiple of -1, and we already have the result for $c \in \mathcal{R}_{>0}$. Recall that

$$\int_a^b f = \int_a^b g + (b-a)L \quad \text{where } g(x) = f(x) - L.$$

Also define

$$h : [a,b] \longrightarrow [0, M-L], \quad h(x) = -f(x) + M.$$

Then

$$\int_a^b (-f) = \int_a^b h - (b-a)M.$$

It follows that

$$\begin{aligned}\int_a^b f + \int_a^b (-f) &= \int_a^b g + (b-a)L + \int_a^b h - (b-a)M \\ &= \int_a^b (g+h) + (b-a)(L-M).\end{aligned}$$

But note that for any $x \in [a,b]$,

$$(g+h)(x) = f(x) - L - f(x) + M = M - L,$$

so that

$$\int_a^b (g+h) = (b-a)(M-L),$$

and therefore

$$\int_a^b f + \int_a^b (-f) = (b-a)(M-L) + (b-a)(L-M) = 0.$$

This is the desired result,

$$\int_a^b (-f) = -\int_a^b f.$$

Finally, the three results for $a > b$ follow from those for $a \leq b$ because of the basic definition that $\int_a^b f = -\int_b^a f$. □

Similarly, the Inequality Rule for integrals needn't require that the functions involved be nonnegative.

Proposition 5.5.3 (Inequality Rule for Integrals, Second Version). *Consider two integrable functions*

$$f, g : [a,b] \longrightarrow [L,M]$$

such that

$$f \leq g,$$

meaning that $f(x) \leq g(x)$ *for all* $x \in [a,b]$. *Then*

$$\int_a^b f \leq \int_a^b g.$$

Proof. We already have the result for $f - L$ and $g - L$, and the result follows immediately because

$$\int_a^b f = \int_a^b (f-L) + (b-a)L \leq \int_a^b (g-L) + (b-a)L = \int_a^b g.$$

□

5.5.3 The Area Between Two Curves

Proposition 5.5.4 (Area Between Two Curves As An Integral). *] Let* a *and* b *be real numbers with* $a \leq b$. *Consider two integrable functions*

$$f, g : [a, b] \longrightarrow \mathcal{R}.$$

The area between the graphs of f *and* g *is*

$$\int_a^b |g - f|.$$

Proof (Sketch of the proof.). For nonnegative functions $f, g : [a, b] \longrightarrow [0, M]$, define

$$\max\{f, g\},\ \min\{f, g\} : [a, b] \longrightarrow [0, M]$$

as follows. For any $x \in [a, b]$,

$$\max\{f, g\}(x) = \begin{cases} g(x) & \text{if } g(x) \geq f(x), \\ f(x) & \text{if } g(x) < f(x) \end{cases}$$

and

$$\min\{f, g\}(x) = \begin{cases} f(x) & \text{if } g(x) \geq f(x), \\ g(x) & \text{if } g(x) < f(x). \end{cases}$$

Thus for all $x \in [a, b]$,

$$\begin{aligned} \max\{f, g\}(x) - \min\{f, g\}(x) &= \begin{cases} g(x) - f(x) & \text{if } g(x) \geq f(x), \\ f(x) - g(x) & \text{if } g(x) < f(x) \end{cases} \\ &= |g(x) - f(x)|. \end{aligned}$$

Then the area is

$$\begin{aligned} \mathrm{Ar}_a^b(\max\{f, g\}) - \mathrm{Ar}_a^b(\min\{f, g\}) &= \int_a^b \max\{f, g\} - \int_a^b \min\{f, g\} \\ &= \int_a^b (\max\{f, g\} - \min\{f, g\}) \\ &= \int_a^b |g - f|. \end{aligned}$$

A subtle point here is that if f and g are integrable then so are $\max\{f, g\}$ and $\min\{f, g\}$, so that the areas in the argument are integrals as tacitly asserted. This will be obvious in the examples where we apply the proposition, and so we omit the general argument.

For bounded functions $f, g : [a, b] \longrightarrow [L, M]$, possibly taking negative values now, a hoisting argument reduces the problem to the nonnegative case. □

For example, consider the functions

$$f, g : [0,2] \longrightarrow \mathcal{R}$$

where

$$f(x) = x^2/2\,, \qquad g(x) = x^3 - 5x^2/2 + 2x.$$

Compute that their difference is

$$(g-f)(x) = x^3 - 3x^2 + 2x = x(x^2 - 3x + 2) = x(x-1)(x-2).$$

This shows that $f = g$ at $x = 0$, at $x = 1$, and at $x = 2$, while

$$g > f \text{ on } (0,1) \quad \text{and} \quad g < f \text{ on } (1,2).$$

(See figure 5.13.) Therefore the area between the graphs of f and g is

$$A = \int_0^1 (g-f) + \int_1^2 (f-g).$$

The first integral is

$$\int_0^1 (g-f) = \left(\frac{1^4 - 0^4}{4}\right) - 3\left(\frac{1^3 - 0^3}{3}\right) + 2\left(\frac{1^2 - 0^2}{2}\right) = \frac{1}{4}\,,$$

and the second is

$$\int_1^2 (f-g) = -\left(\frac{2^4 - 1^4}{4}\right) + 3\left(\frac{2^3 - 1^3}{3}\right) - 2\left(\frac{2^2 - 1^2}{2}\right) = \frac{1}{4}.$$

Thus the total area between the graphs is

$$A = \frac{1}{2}.$$

(In the first half of this calculation, we have applied the formula

$$\int_a^b f_\alpha = \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1}, \quad \alpha \neq -1$$

in the case where the endpoint a is 0. This extension has not yet been justified, and it is valid only for $\alpha \geq 0$, but we take it as granted for now because we will discuss it carefully in chapter 8.)

Exercises

5.5.2. Find the area enclosed by the curve whose equation is

$$y^2 + 2xy + 2x^2 = 1.$$

(See figure 5.14. The curve is an ellipse. Use the quadratic equation to solve the equation for y in terms of x, and then integrate between the two x-values for which there is one y-value.)

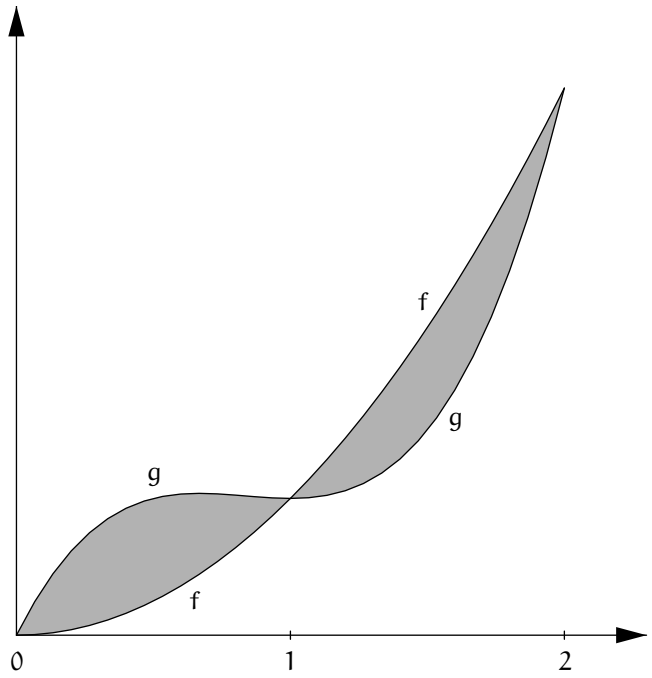


Figure 5.13. Area between two curves

5.5.3. (a) Find the area of the region shown between the curves $y = x(x-2)$ and $y = -x^3$ in figure 5.15.

(b) Find the area of the region shown between the curves $y = 8x^3/9 - 2x^2/9 - x$ and $y = 2x/3$ in figure 5.16.

(c) Find the area of the region shown between the curves $y = -(x-2)(x-3)$ and $y = (x-1)(x-2)(x-3)$ in figure 5.17.

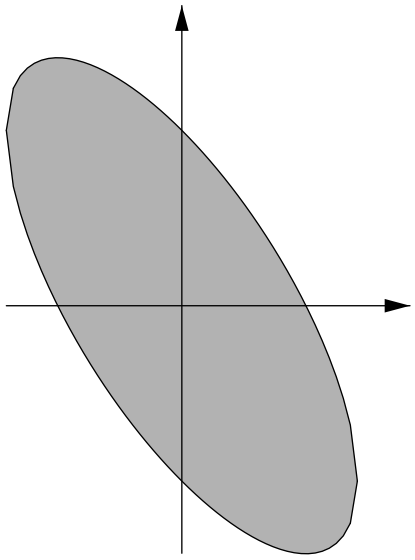


Figure 5.14. Figure for exercise 5.5.2

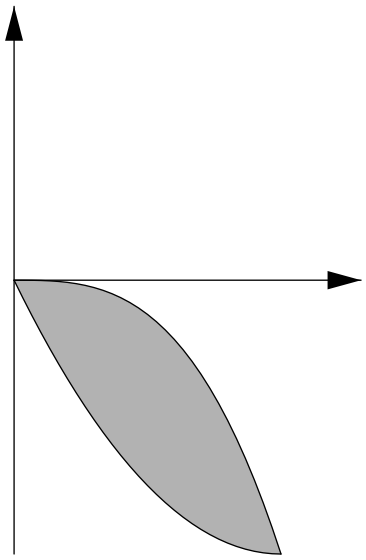


Figure 5.15. Figure for exercise 5.5.3 (a)

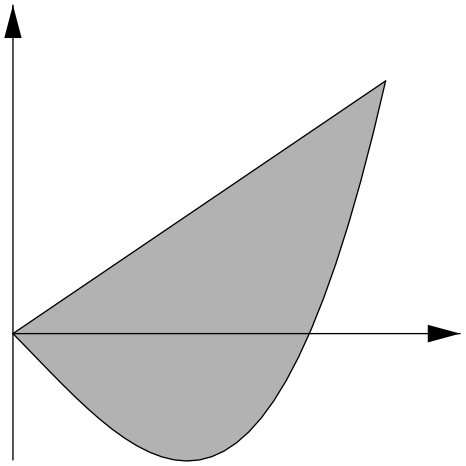


Figure 5.16. Figure for exercise 5.5.3 (b)

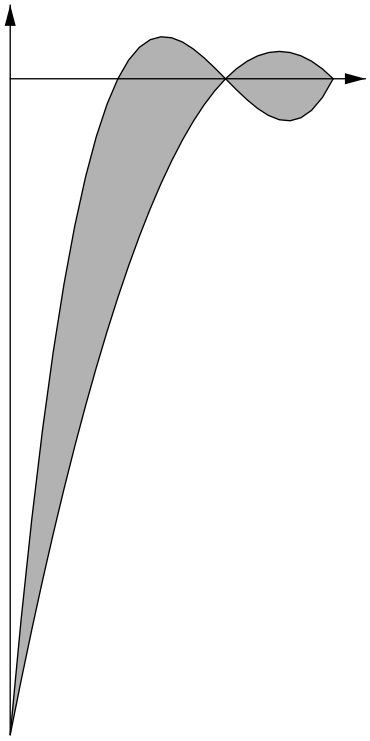


Figure 5.17. Figure for exercise 5.5.3 (c)

6

The Exponential Function

The exponential function is the most important function in mathematics. It can be described in various ways, all compatible. This chapter defines the exponential function as the inverse function of the logarithm, i.e., as the function that undoes the effect of the logarithm,

$$y = \exp(x) \quad \text{if and only if} \quad x = \ln(y).$$

The exponential function is also the unique differentiable function that is its own derivative and takes a normalized value at 0:

$$\text{If} \quad f' = f \text{ and } f(0) = 1 \quad \text{then} \quad f = \exp.$$

And the exponential function is a limit of ever higher powers of quantities ever closer to 1,

$$\exp(x) = \lim_n \left(\left(1 + \frac{x}{n} \right)^n \right).$$

Section 6.1 introduces the notion of a continuous function. The general principle that differentiable functions are continuous gives the continuity of the particular functions that we have differentiated in these notes. The Intermediate Value Theorem says that a continuous function whose domain is an interval cannot jump over feasible outputs: if the function assumes two values then it assumes every value between them as well. One consequence of the Intermediate Value Theorem is the existence of nth roots of positive real numbers, something that we invoked in chapter 2. Another consequence of the theorem is that every real number is a logarithm, and so the logarithm function has an inverse. Section 6.2 defines the exponential as this inverse function. The properties of the logarithm therefore give rise to corresponding properties of the exponential. These properties lead to a definition of raising any positive real number to any real exponent, whereas before we could raise a positive real number only to a rational exponent. Section 6.3 quantifies the

oft-cited fact that the exponential function grows very quickly. Section 6.4 shows that the exponential function is its own derivative. Section 6.5 integrates the exponential function. Section 6.6 shows that the exponential is a limit of powers as described above, and then gives an interpretation of the limit in terms of compound interest.

6.1 Continuity

6.1.1 Definition of Continuity

Definition 6.1.1 (Continuity). *Let A be a subset of $\mathcal{R}$, and let*

$$f : A \longrightarrow \mathcal{R}$$

be a function. Let x be a point of A. The function f is **continuous at** x *if*

$$\lim_{s \to x} f(s) = f(x).$$

The function f is **continuous on** A *if it is continuous at each $x \in A$.*

According to this definition, in order for $f : A \longrightarrow \mathcal{R}$ to be continuous at x, necessarily

- $x \in A$,
- x is approachable from A,
- and for every sequence (s_n) in A that approaches x, $\lim_n(f(s_n)) = f(x)$.

Especially, if $x \in A$ but x is not approachable from A then f cannot be continuous at x. So, for example, under our definition no function $f : \mathcal{Z} \longrightarrow \mathcal{R}$ can be continuous (exercise 6.1.1). The reader is alerted that under a different mathematical convention, more common than ours, the approachability condition is not required for continuity, and so *every* such function is continuous.

To show that a function $f : A \longrightarrow \mathcal{R}$ is discontinuous at a point $x \in A$ that is approachable from A, it suffices to find a sequence (s_n) in A that approaches x while the corresponding output sequence $(f(s_n))$ does not converge to $f(x)$. For example, take the function

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Consider the sequence $(s_n) = (1/n)$. The sequence approaches 0. The corresponding sequence of outputs,

$$(f(s_n)) = (1,\ 1,\ 1,\ \ldots),$$

has limit 1. And so, recalling that $f(0) = 0$,

$$(s_n) \text{ approaches } 0 \quad \text{but} \quad \lim_n(f(s_n)) \neq f(0).$$

Thus Definition 6.1.1 is not satisfied for $x = 0$, i.e., f is discontinuous at 0.

The following result says that most of the functions that we have worked with in these notes are continuous.

Proposition 6.1.2 (Differentiability Implies Continuity). *Let A be a subset of $\mathcal{R}$, and consider a function*

$$f : A \longrightarrow \mathcal{R}.$$

If f is differentiable on A then f is continuous on A.

Proof. This proposition only rephrases Proposition 4.2.2 on page 131 (exercise 6.1.3). □

Although all differentiable functions are continuous, not all continuous functions are differentiable. The simplest example is the absolute value function,

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = |x|.$$

We saw in exercise 4.2.1 (page 130) that f is not differentiable at 0. But it is continuous at 0, because for any sequence (s_n) in $\mathcal{R}_{\neq 0}$, to say that (s_n) approaches 0 is to say that $|s_n - 0|$ grows small as n grows large, i.e., $f(s_n)$ tends to 0, which is $f(0)$.

Extending the example of the absolute value function, Weierstrass used an analytic process of superimposing ever more, ever smaller corners to create a function $f : \mathcal{R} \longrightarrow \mathcal{R}$ that is continuous everywhere and differentiable nowhere. Its graph is somehow jagged no matter how closely we zoom in.

Proposition 6.1.3 (Continuity of the Power Function). *Let α be a rational number. The power function f_α is continuous on its domain.*

Proof. This proposition rephrases Corollary 4.2.5 on page 135. □

However, for another example of continuity without differentiability, let α be any rational number such that $0 < \alpha < 1$. Consider the αth power function, whose value at 0 is $f_\alpha(0) = 0$ (see the discussion on page 29),

$$f_\alpha : \mathcal{R}_{\geq 0} \longrightarrow \mathcal{R}, \quad f_\alpha(x) = x^\alpha.$$

This function is continuous at 0. However, Proposition 4.2.4 (page 134) says that $f'_\alpha(0)$ does not exist.

Proposition 6.1.2 says fairly broadly that the rational power functions f_α and the logarithm function are continuous on their domains. (The exception is that the proposition does not say that f_α is continuous at 0 for $0 < \alpha < 1$, but the previous paragraph has taken care of this case.) In particular, let $\mathcal{R}_{\neq 0} = \{x \in \mathcal{R} : x \neq 0\}$ and consider the reciprocal function

$$f_{-1} : \mathcal{R}_{\neq 0} \longrightarrow \mathcal{R}, \quad f(x) = 1/x.$$

As just remarked, f_{-1} is continuous on $\mathcal{R}_{\neq 0}$. And yet to graph f_{-1}, we must drastically *lift the pencil from the page* because $f_{-1}(x)$ is very negative for x a little less than 0, while $f_{-1}(x)$ is very positive for x a little greater than 0. This example shows that the common idea of a continuous function as one that can be graphed without lifting one's pencil is not entirely correct. The issue here is that the domain $\mathcal{R}_{\neq 0}$ of f_{-1} has two pieces, with a break at 0. The graph of f_{-1} on each piece of $\mathcal{R}_{\neq 0}$ *can* be drawn in one stroke. The continuity of f_{-1} means that its graph has no more breaks than its domain.

Exercises

6.1.1. Explain why no function $f : \mathcal{Z} \longrightarrow \mathcal{R}$ is continuous under Definition 6.1.1. A qualitative explanation is fine. For a more quantitative one, the choice $\varepsilon = 1/2$ could be helpful.

6.1.2. Consider the function

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

Show that f is not continuous at 0.

6.1.3. Explain with some care why Proposition 6.1.2 repeats Proposition 4.2.2.

6.1.2 Continuity and Integrability

Theorem 6.1.4 (Continuity Implies Integrability). *Let a and b be real numbers with $a \leq b$, and consider a function*

$$f : [a, b] \longrightarrow \mathcal{R}.$$

If f is continuous on $[a, b]$ then the integral $\int_a^b f$ exists.

Unfortunately, the proof of Theorem 6.1.4 is beyond our scope.

The functions that we know to be integrable are bounded piecewise monotonic functions. Such a function need not be continuous, and so the converse of Theorem 6.1.4 does not hold.

For an example of a bounded continuous function that is not piecewise monotonic, consider

$$f : [-1,1] \longrightarrow [-1,1], \quad f(x) = \begin{cases} x\sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The graph of f is shown in figure 6.1.

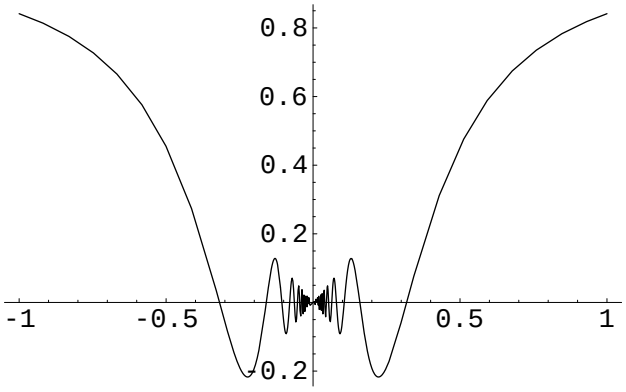


Figure 6.1. A bounded, continuous, but not piecewise monotonic function

6.1.3 The Intermediate Value Theorem

Even though the reciprocal function $f_{-1}(x) = 1/x$ is continuous, its range $\mathcal{R}_{\neq 0}$ has two pieces. However, this is because its domain $\mathcal{R}_{\neq 0}$ has two pieces already. If a function is continuous *and* its domain is an interval, then because its domain has no breaks, its range has no breaks either. The following result expresses this idea analytically.

Theorem 6.1.5 (Intermediate Value Theorem). *Let* a *and* b *be real numbers with* $a < b$. *Let the function*

$$f : [a,b] \longrightarrow \mathcal{R}$$

be continuous, and suppose that $f(a) < 0$ *and* $f(b) > 0$. *Then there exists some number* $c \in (a,b)$ *such that* $f(c) = 0$.

That is, if a continuous function whose domain is a closed interval transists from taking a negative output-value to taking a positive output-value, then it must take the output-value 0 *somewhere* in between. Note that the

conclusion of the theorem isn't simply "$f(c) = 0$," which in isolation would be meaningless because the hypotheses make no mention of a point c; rather the conclusion is that *there exists some* c such that $f(c) = 0$, although the theorem is mute on the precise whereabouts of c or how to find it. Again the phrase *there exists*, whose meaning is not truly settled, has arisen.

The *negative*, *positive*, and *zero* in Theorem 6.1.5 are normalizations. More generally we have

Corollary 6.1.6 (Intermediate Value Theorem, Second Version). *Let* a *and* b *be real numbers with* $a < b$. *Let the function*

$$f : [a, b] \longrightarrow \mathcal{R}$$

be continuous. Let v *be any value between* $f(a)$ *and* $f(b)$. *Then there is some number* $c \in (a, b)$ *such that* $f(c) = v$.

To reduce the corollary to the theorem, first replace the f in the corollary by $f_1 = f - v$. Then $f(c) = v$ if and only if $f_1(c) = 0$, and 0 lies between $f_1(a)$ and $f_1(b)$. Second, if $f_1(a) > 0$ and $f_1(b) < 0$ then replace f_1 by $f_2 = -f_1$. Now the hypotheses for the original theorem are met, and the conclusion of the original theorem gives the conclusion of the corollary, as desired. A tacit point here is that because f is continuous, so are f_1 and f_2. Exercise 6.1.5 is to draw pictures illustrating the argument given in this paragraph, and to explain the tacit point.

The corollary says that if a function is continuous on an interval, its output cannot *jump over* a value: if two numbers are output-values of the function then all numbers between them are output-values as well. This is the sense in which a continuous function is graphed without lifting the pencil from the page.

A loose rephrase of the Intermediate Value Theorem is:

> *Under appropriate conditions, to show that a function takes the value* 0 *somewhere, it suffices to show only that the function changes sign.*

We give an example. View the unit circle as a thin metal ring, and suppose that it has different temperatures at different points, but with the temperature varying continuously as one moves around the ring. Other than this, we assume nothing specific about the ring's temperature function. The claim is that despite knowing so little about the temperature function, we can assert that *somewhere on the ring, there is an antipodal pair of points that have the same temperature*. Indeed, for any angle s, let $p(s)$ denote the point of the ring at angle s, and let

$$g(s) = \text{temperature at } p(s) - \text{temperature at } p(s+\pi).$$

Here $p(s+\pi)$ is the point antipodal to $p(s)$. Thus we want to show that there exists some c such that $g(c) = 0$. If $g(0) = 0$ then we are done. Otherwise, because $p(2\pi) = p(s)$, we have

$$\begin{aligned} g(\pi) &= \text{temperature at } p(\pi) - \text{temperature at } p(2\pi) \\ &= \text{temperature at } p(\pi) - \text{temperature at } p(0) \\ &= -(\text{temperature at } p(0) - \text{temperature at } p(\pi)) \\ &= -g(0). \end{aligned}$$

This shows that $g(0)$ and $g(\pi)$ have opposite signs. Thus, the Intermediate Value Theorem says that for some c between 0 and π we have $g(c) = 0$, which is to say that the antipodal points $p(c)$ and $p(c+\pi)$ have the same temperature. To recapitulate, no matter how the temperature varies around the ring, provided only that it is continuous, the Intermediate Value Theorem says that some pair of antipodal points on the ring has the same temperature, but it tells us nothing about where that pair of points is to be found. This argument may provoke a mix of reactions from the reader, as it does for the author of these notes.

Here is an attempt to prove Theorem 6.1.5 (the normalized Intermediate Value Theorem) rather than assume it. There are f-inputs in $[a, b]$ such that the corresponding f-output is negative, such as a. The number b exceeds all such inputs. So surely there is a *least* value c that is at least as big as all such inputs. If $f(c) > 0$ then because $\lim_{s\to c} f(s) = f(c)$, necessarily $f(s) > 0$ for all s close enough to c (see Proposition 4.1.3 on page 122), and so some value $s < c$ is also at least as big as all f-inputs that produce negative outputs, contradicting the fact that c is the least such value. Similarly, if $f(c) < 0$ then necessarily $f(s) < 0$ for some value $s > c$, contradicting the fact that c is at least as big as all f-inputs that produce negative outputs. The only possibility remaining is that $f(c) = 0$.

However, rather than prove the Intermediate Value Theorem, this argument shows only that the theorem follows from any assumption about the real number system that makes valid the *Then surely there is a* least *value...* statement in the previous paragraph. The issue here is identical to the one that arose from the attempt to prove the Archimedean Property of the real number system back on page 72.

Exercises

6.1.4. Let $p(x) = x^3 - 3x + 1$. Use the Intermediate Value Theorem to show that there are at least three different numbers a, b, and c such that $p(a) = p(b) = p(c) = 0$.

6.1.5. (a) Illustrate the argument that the second version of the Intermediate Value Theorem follows from the normalized version.

(b) Let $f : [a, b] \longrightarrow \mathcal{R}$ be a continuous function. Show that for any real number h, also $f + h$ is continuous. Show that for any real number c, also cf is continuous.

6.1.6. Let $f : [0, 1] \longrightarrow [0, 1]$ be a continuous function such that $f(0) = 1$ and $f(1) = 0$. Draw a picture illustrating the situation. Geometrically, it is compelling that the graph of f must cross the 45-degree line $y = x$ at least once; that is, $f(c) = c$ for some $c \in (0, 1)$. Use the Intermediate Value Theorem to prove this. The argument requires an auxiliary (helper) function.

6.1.4 Applications of the Intermediate Value Theorem

For our first application of the Intermediate Value Theorem, we return to the subject of nth roots of positive real numbers. This topic was discussed starting on page 27, and the reader is encouraged to review the discussion there before continuing here.

Let $n \geq 2$ be an integer, and let $b > 1$ be a real number. Consider the power function

$$f_n : [1, b] \longrightarrow \mathcal{R}, \quad f_n(x) = x \cdots x \ (n \text{ times}).$$

We have argued that f_n is strictly increasing on $\mathcal{R}_{>0}$. Also, we have argued that f_n is differentiable and hence continuous on all of $\mathcal{R}$, so that its restriction here to $[1, b]$ is continuous as well. Neither of these arguments made any reference to the existence of nth roots. Note that because $b > 1$ and $n \geq 2$,

$$f(1) = 1 < b \qquad \text{and} \qquad f(b) = b^n > b.$$

By the Intermediate Value Theorem, there exists some number $c \in (1, b)$ such that $f(c) = b$. And there is only one such c because f is strictly increasing. That is, there is exactly one c such that $c^n = b$. In other words, c *is the unique positive* n*th root of* b, nicely served up to us by the theorem.

If $0 < b < 1$, then $1/b > 1$ and the argument just given produces the nth root c of $1/b$, and $1/c$ is the nth root of b. And if $b = 1$ then b is its own nth root. This covers all cases, and now our invocation of unique nth roots (page 27) is a consequence of whatever assumed property of the real number system will prove the Intermediate Value Theorem.

For our second application of the Intermediate Value Theorem, we know that the logarithm function is continuous on $\mathcal{R}_{>0}$, and we have shown in exercise 5.2.2 (page 157) that

$$\ln(2) < 1 < \ln(4).$$

The exercise then argued that consequently there is one and only one number e between 2 and 4 such that $\ln(e) = 1$. There is at most one such number because the logarithm is strictly increasing, but the fact that there is at least one such number was supportable for us earlier only at the level of intuition. Now it follows from the Intermediate Value Theorem. To repeat:

Definition 6.1.7 (The number e). *The unique real number x that satisfies the condition $\ln(x) = 1$ is denoted e. That is, e is defined by the property*

$$\ln(e) = 1.$$

Continuing to work with the logarithm function, let y be any positive real number. By the Archimedean property of the real number system, there is some positive integer n such that $n > y$. Because $\ln(1) = 0$ and $\ln(e^n) = n\ln(e) = n$, the logarithm function on the interval $[1, e^n]$ satisfies

$$\ln(1) < y \quad \text{and} \quad \ln(e^n) > y.$$

By the Intermediate Value Theorem,

$$\ln(x) = y \quad \text{for some } x \in (1, e^n).$$

That is, *every positive real number is a logarithm*. Similarly, if $b < 0$ then because $-b = \ln(x)$ for some x, it follows that $b = -\ln(x) = \ln(1/x)$. And of course, $0 = \ln(1)$. Because the logarithm function is strictly increasing, we have proved the following result.

Proposition 6.1.8 (The Logarithm is Bijective). *Each real number y takes the form $y = \ln(x)$ for exactly one positive number x. That is, the function*

$$\ln : \mathcal{R}_{>0} \longrightarrow \mathcal{R}$$

takes each value in its codomain exactly once.

Exercise

6.1.7. Let $b \in \mathcal{R}_{>0}$ be a positive real number. Consider the function

$$g : \mathcal{R} \longrightarrow \mathcal{R}, \quad g(x) = x^2 - b.$$

Thus the unique positive number c such that $g(c) = 0$ is the square root of b.

(a) For any positive real number s, the height of the graph of g over s is $g(s)$ (this could be negative) and the tangent slope of the graph is $g'(s) = 2s$. Use

these data and analytic geometry to show that the tangent line to the graph of g at $(s, g(s))$ meets the x-axis at

$$\tilde{s} = \frac{1}{2}\left(s + \frac{b}{s}\right).$$

Note that because b and s are positive, so is $\tilde{s}$.

(b) Choose any real number $s_1 > \sqrt{b}$, and then define a sequence recursively using the formula from (a),

$$s_{n+1} = \frac{1}{2}\left(s_n + \frac{b}{s_n}\right), \quad n \geq 1. \tag{6.1}$$

Show that for any $n \in \mathcal{Z}_{\geq 1}$, if $s_n^2 > b$ then consequently $s_{n+1}^2 > b$. Because $s_1^2 > b$, it follows (you need not explain this part analytically, but put some thought into it) that $s_n^2 > b$ for all $n \in \mathcal{Z}_{\geq 1}$. And then it further follows, because all the s_n-values are positive, that $s_n > \sqrt{b}$ for all $n \in \mathcal{Z}_{\geq 1}$.

(c) Show that $s_{n+1} \leq s_n$ for all $n \in \mathcal{Z}_{\geq 1}$. So (again, you needn't explain what follows, but put thought into it) the sequence $(s_1, s_2, s_3, \ldots)$ consists of entries that grow ever smaller, but the number $\sqrt{b}$ is less than or equal to all the s_n. Then surely there is a *greatest* number c less than or equal to all the s_n. This c is the limit of the sequence: the sequence elements s_n get ever closer to c as n grows, and if they don't get within $\varepsilon > 0$ of c then $c + \varepsilon$ is at most as big all the s_n, contradicting the fact that c is the greatest such number.

(d) Take the limit of both sides of (6.1), carefully explaining your use of various sequence limit rules, to conclude that $c = \sqrt{b}$. Thus the process in this exercise (a special case of *Newton's method*) computes $\sqrt{b}$.

(e) Use a computer to investigate how quickly the values s_n tend to $\sqrt{b}$ for various values of b and various starting approximations s_1 for each b.

6.2 Definition and Properties of the Exponential Function

6.2.1 Definition and Basic Properties

As discussed at the end of section 6.1.4, for each real number y there is exactly one positive real number x such that $y = \ln(x)$.

The names x and y, being mere symbols, can be interchanged: For each real number x there is exactly one positive real number y such that $x = \ln(y)$. The function that takes each real x to the corresponding positive real y is the *exponential function*. So earlier the logarithm was defined as an integral,

and now the exponential function is defined as the *inverse function* of the logarithm: it undoes the logarithm, and the logarithm undoes it. All of this takes us far beyond the idea of a function as an analytic expression.

Definition 6.2.1 (Exponential Function). *The* **exponential function**,

$$\exp : \mathcal{R} \longrightarrow \mathcal{R}_{>0},$$

is the inverse function of the logarithm. That is, the functions exp *and* ln *are related by the following property:*

$$\textit{For all } x \in \mathcal{R} \textit{ and all } y \in \mathcal{R}_{>0}, \quad y = \exp(x) \Longleftrightarrow x = \ln(y).$$

Because the exponential function and the logarithm function exchange the roles of x and y, the graph of the exponential function is obtained by reflecting the graph of the logarithm function through the line $y = x$ (exercise 6.2.1). Figure 6.2 shows (portions of) the graphs of the two functions.

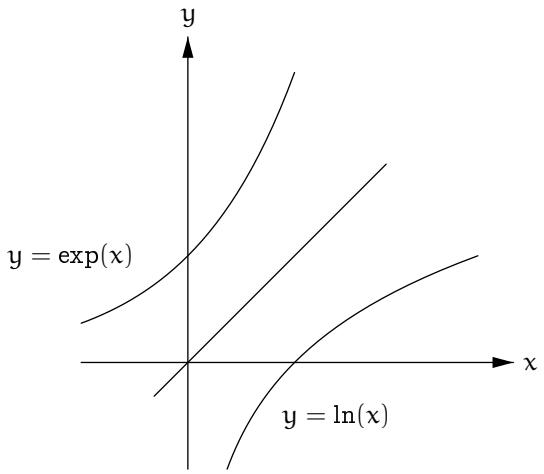


Figure 6.2. Graphs of the exponential and logarithm functions

We show that immediately in consequence of Definition 6.2.1,

$$\text{for all } x \in \mathcal{R}, \quad \ln(\exp(x)) = x, \tag{6.2}$$

and

$$\text{for all } y \in \mathcal{R}_{>0}, \quad \exp(\ln(y)) = y. \tag{6.3}$$

To establish (6.2), let $x \in \mathcal{R}$ and let $y = \exp(x)$. According to the defining property of the exponential function, $x = \ln(y)$, i.e., $x = \ln(\exp(x))$ as desired. That is, (6.2) follows from one direction across the double-headed arrow "$\Longleftrightarrow$" in the defining property of the exponential function. Naturally, (6.3) follows from the other (exercise 6.2.2).

Theorem 6.2.2 (Properties of the Exponential Function).

(1) $\exp(0) = 1$.

(2) *For all real numbers* x *and* $\tilde{x}$,

$$\exp(x + \tilde{x}) = \exp(x)\exp(\tilde{x}).$$

(3) *For all real numbers* x

$$\exp(-x) = 1/\exp(x).$$

(4) *For all real numbers* x *and* $\tilde{x}$,

$$\exp(x - \tilde{x}) = \exp(x)/\exp(\tilde{x}).$$

(5) *For all real numbers* x *and all rational numbers* α,

$$\exp(\alpha x) = (\exp(x))^{\alpha}.$$

Proof. These are all consequences of the corresponding properties of the logarithm, or (3) can be proved by using (1) and (2) as in the proof of the corresponding logarithm property. For example, to prove (2), let

$$y = \exp(x), \quad \tilde{y} = \exp(\tilde{x}),$$

so that

$$x = \ln(y), \quad \tilde{x} = \ln(\tilde{y}).$$

Then because $\ln(y) + \ln(\tilde{y}) = \ln(y\tilde{y})$,

$$\exp(x + \tilde{x}) = \exp(\ln(y) + \ln(\tilde{y}) = \exp(\ln(y\tilde{y})) = y\tilde{y} = \exp(x)\exp(\tilde{x}).$$

The remainder of the proof is exercise 6.2.3. □

Exercises

6.2.1. Explain why the geometric operation of reflecting a point $p = (x, y)$ in the plane through the 45-degree line $y = x$ is the same as the algebraic operation of exchanging the point's x- and y-coordinates. Note that $p = q + r$ where $q = \frac{x+y}{2}(1, 1)$ and $r = \frac{x-y}{2}(1, -1)$, and the reflection of p is $q - r$.

6.2.2. Show that (6.3) follows from the defining property of the exponential function.

6.2.3. Prove the rest of Theorem 6.2.2.

6.2.2 Raising to Powers Revisited

Recall that so far we understand raising any positive real number $b \in \mathcal{R}_{>0}$ to any rational exponent $\alpha \in \mathcal{Q}$. The exponential function provides us a mechanism to raise any $b \in \mathcal{R}_{>0}$ to any *real* exponent $x \in \mathcal{R}$, and to re-establish the laws of exponents in this context. The idea is that for $b \in \mathcal{R}_{>0}$ and $\alpha \in \mathcal{Q}$,

$$b^\alpha = \exp(\ln(b^\alpha)) = \exp(\alpha \ln(b)).$$

But in the previous display, *the right side makes sense with no reference to the fact that α is rational.* Thus the following definition is natural.

Definition 6.2.3 (Raising a Positive Real Number to a Real Power). *Let $b \in \mathcal{R}_{>0}$ be any positive real number, and let $x \in \mathcal{R}$ be any real number. Then b^x is defined to be*

$$b^x = \exp(x \ln(b)).$$

Again, this definition of exponentiation agrees with our previous notion of exponentiation when x is rational. Note that b^x is positive for $b \in \mathcal{R}_{>0}$ and $x \in \mathcal{R}$.

Immediately in consequence of Definition 6.2.3, Theorem 5.1.4(4) (page 152) no longer requires a rational exponent:

Proposition 6.2.4 (Enhanced Property of the Logarithm). *For all positive real numbers b and all real numbers x,*

$$\ln(b^x) = x \ln(b).$$

Also, to study a special case of Definition 6.2.3, recall the number e such that $\ln(e) = 1$. Because $\ln(e) = 1$, Definition 6.2.3 says that

$$e^x = \exp(x),$$

and this explains why the exponential function is often written e^x and called *e to the x*.

Returning from e^x to b^x for any positive real number b, our new notion of exponentiation satisfies the appropriate laws, as follows.

Proposition 6.2.5 (Laws of Real Exponents). *Let $b, \tilde{b} \in \mathcal{R}_{>0}$ be any positive real numbers. Let $x, \tilde{x} \in \mathcal{R}$ be any real numbers. Then*

(1) $b^0 = 1$ *and* $b^1 = b$.
(2) $b^x b^{\tilde{x}} = b^{x+\tilde{x}}$.
(3) $(b^x)^{\tilde{x}} = b^{x\tilde{x}}$.
(4) $(b\tilde{b})^x = b^x \tilde{b}^x$.

The only obstacle to proving the proposition is that its formulas are so familiar. But once one understands that the mechanism is Definition 6.2.3, the computations are easy.

Proof. (1) is immediate because the exponents 0 and 1 are rational numbers. We can also obfuscate matters and argue that $b^0 = \exp(0\ln(b)) = \exp(0) = 1$ by Theorem 6.2.2, and $b^1 = \exp(1\ln(b)) = \exp(\ln(b)) = b$ by (6.3), but doing so is gratuitous. For (3), compute that

$$(b^x)^{\tilde{x}} = \exp\left(\tilde{x}\ln(b^x)\right) = \exp(x\tilde{x}\ln(b)) = b^{x\tilde{x}}.$$

Parts (2) and (4) are exercise 6.2.5. □

Various properties that we have established for the rational power function extend to the function $f_\alpha(x) = x^\alpha$ for $x \in \mathcal{R}_{>0}$ where now α is real rather than rational. For example, for positive real s and t, with $u = t/s$,

$$t^\alpha - s^\alpha = s^\alpha(u^\alpha - 1) \quad \text{has the same sign as} \quad u^\alpha - 1,$$

and then, because

$$u^\alpha = \exp(\alpha\ln(u)),$$

if $\alpha > 0$ then $\alpha\ln(u)$ is positive for $u > 1$ and negative for $0 < u < 1$, and so $u^\alpha > 1$ for $u > 1$ and $0 < u^\alpha < 1$ for $0 < u < 1$, and so f_α on $\mathcal{R}_{>0}$ is increasing. Similarly, if $\alpha < 0$ then f_α on $\mathcal{R}_{>0}$ is decreasing. So as before, the power function f_α on $\mathcal{R}_{>0}$ is increasing if $\alpha > 0$ and decreasing if $\alpha < 0$.

Consequently, if $\alpha > 0$ then for any $\varepsilon > 0$ and any $x \in \mathcal{R}_{>0}$ we have

$$x^\alpha < \varepsilon \quad \text{if and only if} \quad x < \varepsilon^{1/\alpha}.$$

This shows that

$$\lim_{x\to 0^+} x^\alpha = 0 \quad \text{if } \alpha > 0.$$

That is, if $\alpha > 0$ then the function $f_\alpha(x) = x^\alpha$ for $x \in \mathcal{R}_{>0}$ extends continuously to $f_\alpha(0) = 0$.

Exercises

6.2.4. Prove Proposition 6.2.4.

6.2.5. Prove parts (2) and (4) of Proposition 6.2.5.

6.3 Exponential Growth

The first few terms of the sequence

$$(s_n) = \left(\frac{n^{100000000}}{1.00000001^n} \right)$$

are roughly, according to a computer,

$$\begin{aligned} s_1 &= 0.999999, \\ s_2 &= 3.684665 \times 10^{30102999}, \\ s_3 &= 2.964601 \times 10^{47712125}, \\ s_4 &= 1.357676 \times 10^{60205999}, \\ s_5 &= 2.713950 \times 10^{69897000}. \end{aligned}$$

These are enormous. On the other hand, the jumps in the powers of 10—from zero to 30 million to 48 million to 60 million to 70 million—seem to be slowing down, suggesting that perhaps the sequence is tending upward to some huge-but-finite value. In fact, $\lim_n(s_n) = 0$. (!)

Theorem 6.3.1 (Exponential Growth Dominates Polynomial Growth). *Exponential growth dominates polynomial growth in the sense that*

$$\boxed{\lim_{x\to\infty} \frac{x^a}{b^x} = 0 \quad \textit{for any } a > 0 \textit{ and } b > 1.}$$

Here, as on page 155, for any function $f : \mathcal{R}_{>0} \longrightarrow \mathcal{R}$ we define

$$\lim_{x\to\infty} f(x) = \lim_{s\to 0} f(1/s).$$

That is, the left limit exists if the right limit does, in which case it takes its value from the right limit.

Proof. We have already established that

$$\lim_{x\to\infty} \frac{\ln x}{x} = 0.$$

The result follows. For all large enough x we have

$$\frac{\ln x}{x} < \frac{\ln b}{a+1},$$

so that after cross-multiplying,

$$(a+1)\ln x < x \ln b,$$

and so by a property of the logarithm,

$$\ln(x^{a+1}) < \ln(b^x),$$

and so because the logarithm is increasing,

$$x^{a+1} < b^x,$$

and now dividing through by xb^x and noting that x^a/b^x is positive gives

$$0 < \frac{x^a}{b^x} < \frac{1}{x}.$$

Because $\lim_{x\to\infty} 1/x = 0$, we are done by squeezing. □

6.4 Differentiation of the Exponential

The main result of this section is as follows.

Theorem 6.4.1 (The Exponential Function is Its Own Derivative). *The exponential function is its own derivative,*

$$\boxed{\exp' = \exp.}$$

In consequence of the theorem, the exponential function is continuous.

The theorem is close to self-evident geometrically in consequence of the derivative of the logarithm being the reciprocal function. For any real number x, the point $(\exp(x), x)$ lies on the logarithm graph because $\exp(x) > 0$ and $\ln(\exp(x)) = x$, and we have:

The tangent slope to the logarithm graph at $(\exp(x), x)$ is $1/\exp(x)$.

Reflecting the logarithm graph through the $y = x$ line gives the exponential graph. And surely the tangent line of the reflected graph at the reflected point is the reflection of the tangent line to the original graph at the original point. (For example, this can be argued using the geometric characterization of the tangent line in exercise 4.2.3 on page 132.) Reflecting the line interchanges the roles of *rise* and *run*, so that the slope of the reflected line is the reciprocal of the original slope. In sum:

The tangent slope to the exponential graph at $(x, \exp(x))$ is $\exp(x)$.

That is, $\exp'(x) = \exp(x)$, and so this geometric argument supports the theorem and perhaps already proves it.

The Chain Rule also supports the theorem. For all $x \in \mathcal{R}$,

$$\ln(\exp(x)) = x,$$

and so taking derivatives gives

$$\ln'(\exp(x))\,\exp'(x) = 1,$$

or, because the derivative of the logarithm is the reciprocal,

$$\frac{\exp'(x)}{\exp(x)} = 1,$$

which is to say (again) that $\exp'(x) = \exp(x)$. However, the problem with this argument is that it assumes that the exponential function is differentiable. The *existence* of the derivative of the exponential function is the subtle issue here, not the value of the derivative once its existence is known. The just-given Chain Rule argument ignored this point, while the preceding geometric argument handwaved it. Our only method to show in a satisfactory way that the derivative exists is to work analytically and calculate it, and so our proof of the theorem will proceed by doing so.

Proof. (Proof that $\exp'$ exists and equals $\exp$.) First we establish the normalized case that $\exp'(0) = \exp(0)$, i.e., that

$$\lim_{s\to 0} \frac{\exp(s) - \exp(0)}{s - 0} = \exp(0).$$

Because $\exp(0) = 1$ this means showing that

$$\lim_{s\to 0} \frac{\exp(s) - 1}{s} = 1.$$

To establish this function limit is to establish that

$$\begin{aligned} &\text{for any sequence } (s_n) \text{ in } \mathcal{R} \text{ that approaches } 0, \\ &\lim_n \left(\frac{\exp(s_n) - 1}{s_n}\right) = 1. \end{aligned}$$

On the other hand, we know that $\ln'(1) = 1$, which is to say, recalling that $\ln(1) = 0$, that

$$\begin{aligned} &\text{for any sequence } (t_n) \text{ in } \mathcal{R}_{>0} \text{ that approaches } 1, \\ &\lim_n \left(\frac{\ln(t_n)}{t_n - 1}\right) = 1, \end{aligned}$$

and the Reciprocal Rule for sequence limits says that consequently

for any sequence (t_n) in $\mathcal{R}_{>0}$ that approaches 1,

$$\lim_n \left(\frac{t_n - 1}{\ln(t_n)}\right) = 1.$$

So, consider any sequence (s_n) in $\mathcal{R}$ that approaches 0. Let $t_n = \exp(s_n)$ for each n, a sequence in $\mathcal{R}_{>0}$. We show that (t_n) approaches 1. Let $\varepsilon > 0$ be given. Our goal is to make values t_n lie within ε of 1, so we may assume that $\varepsilon < 1$ by shrinking ε if necessary. Because $\ln(1-\varepsilon) < 0$ and $\ln(1+\varepsilon) > 0$ and $\lim_n(s_n) = 0$, there is a starting generation N such that $\ln(1-\varepsilon) < s_n < \ln(1+\varepsilon)$ for all $n \geq N$. Because the exponential function is increasing and $t_n = \exp(s_n)$ for all n, it follows that $1-\varepsilon < t_n < 1+\varepsilon$ for all $n \geq N$, showing that $\lim_n(t_n) = 1$. Furthermore, because $s_n \neq 0$ for all n, also $t_n \neq 1$ for all n, and so (t_n) approaches 1 as claimed. (See figure 6.3.) Now, because (t_n) approaches 1, we have that indeed

$$\lim_n \left(\frac{t_n - 1}{\ln(t_n)}\right) = 1.$$

And because $t_n = \exp(s_n)$ and $\ln(t_n) = s_n$ for each n, this is exactly our desired sequence limit,

$$\lim_n \left(\frac{\exp(s_n) - 1}{s_n}\right) = 1.$$

So we have our desired normalized function limit,

$$\lim_{s\to 0} \left(\frac{\exp(s) - 1}{s}\right) = 1,$$

which is to say that $\exp'(0) = 1$ as claimed.

For general but fixed $x \in \mathcal{R}$ and for any $s \in \mathcal{R}$, let $\tilde{s} = s - x$, so that $s = x + \tilde{s}$. Then

$$\begin{aligned}
\lim_{s\to x} \left(\frac{\exp(s) - \exp(x)}{s - x}\right) &= \lim_{\tilde{s}\to 0} \left(\frac{\exp(x+\tilde{s}) - \exp(x)}{\tilde{s}}\right) \\
&= \lim_{\tilde{s}\to 0} \left(\frac{\exp(x)\exp(\tilde{s}) - \exp(x)}{\tilde{s}}\right) \\
&= \exp(x) \lim_{\tilde{s}\to 0} \left(\frac{\exp(\tilde{s}) - 1}{\tilde{s}}\right) \\
&= \exp(x) \quad \text{by the normalized case.}
\end{aligned}$$

That is, $\exp'(x) = \exp(x)$ for every x. This completes the proof.

□

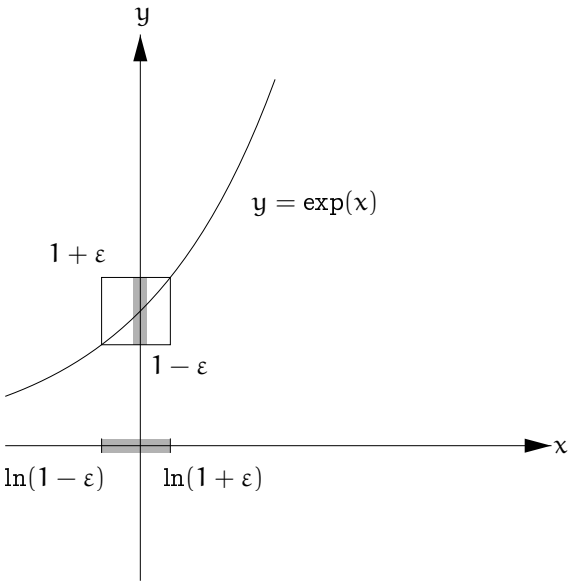


Figure 6.3. Convergence of (s_n) to 0 forces convergence of (t_n) to 1

Exercises

6.4.1. Let $\alpha \in \mathcal{R}$ be any real number, not necessarily rational. Define the corresponding power function

$$f_\alpha : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \quad f_\alpha(x) = \exp(\alpha \ln(x)).$$

Observe that if $\alpha \in \mathcal{Q}$ *is* rational then this function is the familiar rational power function f_α, other than the issue that its domain may now be smaller.

Show that f_α is differentiable and as before,

$$f'_\alpha = \alpha f_{\alpha-1}.$$

(Use the chain rule.)

6.4.2. For any positive real number b (abbreviating *base*), define

$$g_b : \mathcal{R} \longrightarrow \mathcal{R}, \quad g_b(x) = b^x = \exp(x \ln(b)).$$

Thus g_e is the exponential function. Show that g_b is differentiable on $\mathcal{R}$, and

$$g'_b = \ln(b) g_b.$$

That is, slightly abusing notation,

$$(b^x)' = \ln(b)\, b^x.$$

(Use the chain rule.)

6.4.3. We know that the exponential function satisfies the conditions

$$\exp' = \exp \quad \text{and} \quad \exp(0) = 1.$$

Suppose that some unknown function $f : \mathcal{R} \longrightarrow \mathcal{R}$ also satisfies the conditions

$$f' = f \quad \text{and} \quad f(0) = 1.$$

Define

$$g : \mathcal{R} \longrightarrow \mathcal{R}, \qquad g(x) = \exp(-x)f(x).$$

Show that $g' = 0$, i.e., $g'(x) = 0$ for all $x \in \mathcal{R}$. This fact *suggests* powerfully that that the function g itself must be some constant c. Granting this, find c by evaluating $g(0)$. What does this say about f?

6.4.4. For each of the following functions, determine the function's domain, and then differentiate the function on its domain.

(a) $f(x) = \exp(x)\ln(x)$.

(b) $f(x) = \exp(\ln(x) + 1/x)$

(c) $f(x) = x^a b^x$. (Here $a \in \mathcal{R}$ and $b \in \mathcal{R}_{>0}$ are constants.)

6.4.5. The hyperbolic cosine and the hyperbolic sine functions are

$$\cosh : \mathcal{R} \longrightarrow \mathcal{R}, \quad \cosh(x) = \frac{e^x + e^{-x}}{2}$$

and

$$\sinh : \mathcal{R} \longrightarrow \mathcal{R}, \quad \sinh(x) = \frac{e^x - e^{-x}}{2}.$$

(Their pronunciations rhyme with *bosh* and *grinch*.)

(a) Show that $\cosh' = \sinh$ and $\sinh' = \cosh$.

(b) Use the chain rule to compute $(\cosh^2 - \sinh^2)'$ without computing $\cosh^2$ or $\sinh^2$, putting your answer in as simple a form as you can. What does your answer suggest about $\cosh^2 - \sinh^2$?

(c) Sketch the graphs of cosh and sinh on one set of coordinate axes.

6.5 Integration of the Exponential

Let b be a positive real number. Because the exponential function is monotonic, its integral from 0 to b exists. Now see figure 6.4. Its light-shaded region

has area $\int_1^{\exp(b)} \ln$. The formula for the integral of the logarithm (see page 169) says that

$$\int_1^{\exp(b)} \ln = b\exp(b) - \exp(b) + 1.$$

And the entire shaded box in the figure has area $b\exp(b)$. It follows that the integral of the exponential from 0 to b is

$$\int_0^b \exp = b\exp(b) - (b\exp(b) - \exp(b) + 1) = \exp(b) - 1.$$

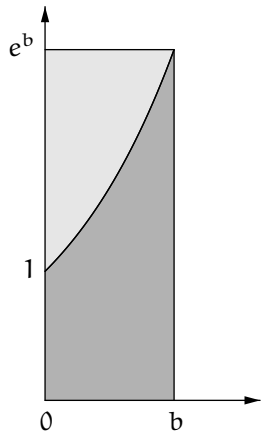


Figure 6.4. The integral of the exponential via the integral of the logarithm

It is also very easy to integrate the exponential with no reference to the integral of the logarithm. We show how to do so because the same method will be used in chapter 7 to integrate the cosine and sine functions. For a uniform partition of $[0, b]$ into n subintervals the lower sum is a finite geometric sum of length n,

$$\begin{aligned} S_n &= \frac{b}{n}\left[\exp(0) + \exp(b/n) + \exp(2b/n) + \cdots + \exp((n-1)b/n)\right] \\ &= \frac{b}{n}\left[1 + \exp(b/n) + (\exp(b/n))^2 + \cdots + (\exp(b/n))^{n-1}\right] \\ &= \frac{b}{n} \cdot \frac{(\exp(b/n))^n - 1}{\exp(b/n) - 1} \\ &= (\exp(b) - 1) \Big/ \frac{\exp(b/n) - \exp(0)}{b/n - 0}, \end{aligned}$$

and thus

$$\lim_n(S_n) = (\exp(b) - 1)/\exp'(0) = \exp(b) - 1.$$

Each corresponding upper sum T_n satisfies

$$T_n - S_n = \frac{b}{n}\left[\exp(b) - 1\right],$$

and so $\lim_n(T_n - S_n) = 0$. Therefore, again,

$$\int_0^b \exp = \exp(b) - 1.$$

Now the method of the previous paragraph can be used instead to rederive the integral of the logarithm with no work. Because the logarithm took some effort to integrate, there is a real gain of efficiency here. But integrating the logarithm this way would have deferred our learning the answer until now.

For the more general integral of the exponential function,

$$\int_a^b \exp, \quad a, b \in \mathcal{R},$$

first assume that $a < b$ and translate the boxes over $[0, b-a]$ by a. Note that $\exp(x + a) = \exp(a)\exp(x)$, and very quickly it follows that (exercise 6.5.1)

$$\int_a^b \exp = \exp(b) - \exp(a), \quad a \le b. \tag{6.4}$$

And then, finally, if instead $a > b$ then

$$\int_a^b \exp = -\int_b^a \exp = -(\exp(a) - \exp(b)) = \exp(b) - \exp(a).$$

That is,

Theorem 6.5.1 (Integral of the Exponential Function). *The integral of the exponential function is*

$$\boxed{\int_a^b \exp = \exp(b) - \exp(a), \quad a, b \in \mathcal{R}.}$$

Naturally this is another instance of the Fundamental Theorem of Calculus. Let $F = \exp$, so that $F' = \exp$. Then the formula says that

$$\int_a^b \exp = F(b) - F(a), \quad a, b \in \mathcal{R}.$$

Exercise

6.5.1. Let a and b be real numbers with $a \leq b$. Explain why lower sums $\widetilde{S}_n$ for $\mathrm{Ar}_a^b(\exp)$ satisfy

$$\widetilde{S}_n = e^a S_n,$$

where each S_n is in turn a lower sum for $\mathrm{Ar}_0^{b-a}(\exp)$, and similarly for upper sums. Show that formula (6.4) follows.

6.6 The Exponential as a Limit of Powers

6.6.1 The Description

Theorem 6.6.1 (The Exponential Function as a Limit of Powers). *For any real number* x,

$$\exp(x) = \lim_n \left(\left(1 + \frac{x}{n} \right)^n \right).$$

Before the proof, it deserves notice that the limit in the theorem is subtle. Given a fixed real number x, the question is to what value the quantity

$$\left(1 + \frac{x}{n} \right)^n$$

tends as n grows large. For that matter, does such a value even exist?

One argument proceeds as follows: *Because* x *is fixed,* $1+x/n$ *tends to* 1 *as* n *grows large, and so* $(1+x/n)^n$ *behaves like* 1^n. *Because* $1^n = 1$ *for all* n, *the limiting value of* $(1+x/n)^n$ *must be* 1.

A second argument reaches a drastically different conclusion: *Assume that* x *is positive. Because* $1+x/n$ *is greater than* 1 *for all* n, *it follows that* $(1+x/n)^n$ *is a high power of a quantity greater than* 1. *But the high powers of any quantity greater than* 1 *grow very large, and so the limiting value of* $(1+x/n)^n$ *must be infinite. Similarly, if* x *is negative then for large* n, $1+x/n$ *lies between* 0 *and* 1, *and so the limiting value of* $(1+x/n)^n$ *must be* 0. *Thus the limit is infinite for positive* x, *and* 0 *for negative* x. *And of course it is* 1 *for* $x = 0$.

These arguments, and their conclusions, are both incorrect. The problem with both of them is that to study the limiting behavior of $(1+x/n)^n$ as n grows large, we must take care that both occurrences of n in $(1+x/n)^n$ grow large *together*. The first argument let the denominator-n grow first and only thereafter let the exponent-n grow as well. The second argument made the complementary error. The falsity of the resulting conclusions illustrates yet again that to obtain the correct medium-sized answers in calculus, we need to manipulate small and large quantities carefully.

Proof. During the course of differentiating the logarithm, we established the bounds

$$\frac{s-1}{s} \le \ln(s) \le s-1, \quad s \in \mathcal{R}_{>0}.$$

Let $x \in \mathcal{R}$ be any real number, let $n \in \mathcal{Z}_{\ge 1}$ be any positive integer such that $n > |x|$, and replace s by $1 + x/n$ to get

$$\frac{x/n}{1+x/n} \le \ln\left(1+\frac{x}{n}\right) \le \frac{x}{n}, \quad n > |x|.$$

Multiply the inequalities through by n and then note that $n\ln(1+x/n) = \ln((1+x/n)^n)$ to get

$$\frac{x}{1+x/n} \le \ln\left(\left(1+\frac{x}{n}\right)^n\right) \le x, \quad n > |x|.$$

Because the exponential function is strictly increasing, the inequalities are preserved upon passing the quantities through it,

$$\exp\left(\frac{x}{1+x/n}\right) \le \left(1+\frac{x}{n}\right)^n \le \exp(x), \quad n > |x|. \tag{6.5}$$

By various sequence limit results (exercise 6.6.1),

$$\lim_n \left(\frac{x}{1+x/n}\right) = x,$$

and so, because the exponential function is continuous,

$$\lim_n \left(\exp\left(\frac{x}{1+x/n}\right)\right) = \exp(x),$$

To complete the argument, apply the Squeezing Rule for sequence limits to (6.5), showing that indeed

$$\lim_n \left(\left(1+\frac{x}{n}\right)^n\right) \text{ exists and equals } \exp(x) \text{ for all } x \in \mathcal{R}.$$

□

Exercise

6.6.1. Show that for any real number x,

$$\lim_n \left(\frac{x}{1+x/n}\right) = x.$$

6.6.2 An Interpretation: Compound Interest

The bank promises you an annual interest rate of x. For example, x could be some value such as $x = 0.05$, i.e., five percent, but to avoid being overspecific we view x as a generic positive real number.

You make a deposit d.

A year later the bank informs you that you now have your original deposit plus the interest on your deposit, said interest amounting to your deposit multiplied by the annual interest rate. That is, the bank tells you that the amount in your account is now

$$d(1+x).$$

This is not fair to you. The bank has been free to invest your deposit from the moment that you made it, and then further to invest any profits from the initial investment, and so on; but the bank has awarded you interest only on your deposit, despite earning money with your deposit all through the year.

It would be more fair for the bank to award you interest on your deposit halfway through the year, at half of the annual interest rate, and then to award you interest on your new larger amount of money at the end of the year, again at half of the annual interest rate. Thus at six months you would have your deposit plus your deposit times half the annual interest rate,

$$d(1+x/2),$$

and at the end of the year you would have your amount at six months plus your amount at six months times half the annual interest rate,

$$d(1+x/2)^2.$$

But this is not fair to you either. The bank has made money with your deposit continuously, but still it has awarded you interest only after six months and then compounded it only after another six months. On the brighter side, the end-year amount now satisfies

$$d(1+x/2)^2 = d(1+x+x^2/4) > d(1+x),$$

i.e., the situation is more fair to you now than it was after a single interest payment at the end of the year.

Similarly, if the bank awards and then compounds interest monthly then at the end of the year you have

$$d(1+x/12)^{12}.$$

Before we continue to analyze the situation, here is a small observation: For any positive quantity s and any integer $n > 1$, expanding the n-fold product

$$(1+s)^n = (1+s)(1+s)\cdots(1+s)$$

gives the term 1 (from multiplying all of the 1's together), gives the term ns (from the n different ways of multiplying one s and $n-1$ 1's together), and gives other terms, all of which are positive. Thus

$$(1+s)^n > 1+ns \quad \text{for } s>0 \text{ and } n>1.$$

Returning to the compound interest calculation we have, in consequence of the observation just made,

$$(1+x/12)^6 > 1+6x/12 = 1+x/2,$$

and so

$$d(1+x/12)^{12} = d\left((1+x/12)^6\right)^2 > d(1+x/2)^2.$$

That is, you have more at the end of the year if the bank compounds interest monthly than if the bank compounds interest only twice a year, just as you have more if the bank compounds interest twice a year than if the bank compounds interest only once a year.

If the bank compounds interest daily then at the end of the year your balance is

$$d(1+x/365)^{365}.$$

Presumably this is more than you have if the bank compounds interest monthly, but the relevant algebra to justify this fact is a bit hairier than the two comparisons that we we have made so far. (The problem is that 365 is not an integer multiple of 12, whereas 12 is an integer multiple of 2, and 2 is an integer multiple of 1.) Instead, note that if the bank compounds interest 12 times daily then because $12\cdot 365 = 4380$, at the end of the year you have

$$d(1+x/4380)^{4380}.$$

Furthermore, applying our small observation twice gives

$$(1+x/4380)^{12} > 1+x/365 \quad \text{and} \quad (1+x/4380)^{365} > 1+x/12,$$

so that

$$(1+x/4380)^{4380} > (1+x/365)^{365} \quad \text{and} \quad (1+x/4380)^{4380} > (1+x/12)^{12}.$$

In other words, if the bank compounds interest 12 times daily then at the end of the year then you have more than if the bank compounds interest daily *or* monthly. (Unsurprisingly, the daily compounding does yield more interest than the monthly, and this is easy to show using calculus rather than algebra. See exercise 6.6.2.)

Even if the bank compounds your interest 4380 times, this still isn't quite fair to you because the bank is making money continuously with your deposit. The correct scenario is for the bank to compound your interest continuously as well. Thus we are led to the exponential function: at the end of the year, your account value should be

$$d \lim_n ((1 + x/n)^n) = de^x.$$

Interpret an interest rate as a proportion factor relating the rate of increase of a quantity to the amount of the quantity. That is, viewing time t as an independent variable, the accruing value of your account is an unknown time-dependent function $f(t)$ such that

$$\begin{aligned} f(0) &= d \qquad \text{(initial condition)}, \\ f'(t) &= xf(t) \quad \text{(differential equation)}. \end{aligned}$$

Similar to exercise 6.4.3, the only function satisfying these conditions is

$$f(t) = de^{xt}.$$

Thus we recover our result that the exponential function arises naturally from continuously compounded interest. And more generally, it describes any quantity that increases in proportion to its amount.

Exercises

6.6.2. Recall that in the course of proving that the exponential is a limit of powers, we made use of the inequality

$$\frac{s}{1+s} \leq \ln(1+s), \quad s > -1.$$

Because the inequality was derived by comparing a box-area to an area under a curve, the inequality is strict (i.e., it is "$<$") unless $s = 0$.

(a) Let $x \in \mathcal{R}_{>0}$ be a fixed positive real number. Define a function of a variable t that also makes reference to the constant x,

$$f : \mathcal{R}_{\geq 1} \longrightarrow \mathcal{R}, \quad f(t) = \ln((1 + x/t)^t).$$

Show that

$$f'(t) = \ln(1+s) - \frac{s}{1+s}, \quad \text{where } s = \frac{x}{t}.$$

(b) Part (a) shows that f' is always positive by the inequality at the beginning of this exercise, because $s = x/t$ is positive. The fact that f' is

always positive *suggests* powerfully that f is strictly increasing. Granting this, explain why the function

$$g : \mathcal{R}_{\geq 1} \longrightarrow \mathcal{R}, \quad g(x) = \exp(f(x))$$

is strictly increasing as well.

(c) In particular,

$$g(1) < g(2) < g(3) < \cdots .$$

What does this say about your account balance at the end of the year if the bank compounds interest daily rather than monthly?

6.6.3. Let d and x be positive real numbers. What is the doubling time of the quantity

$$f : \mathcal{R}_{\geq 0}, \quad f(t) = de^{xt},$$

i.e., the t-value such that $f(t) = 2d$? How does the doubling time depend on the initial value d? If instead x is negative, what is the halving time of f?

7

The Cosine and Sine Functions

The basic trigonometric functions cosine and sine describe uniform oscillation, analogously to how the exponential function describes natural growth. Because oscillation is a more complicated phenomenon than natural growth, the properties of the cosine and sine are correspondingly more elaborate than those of the exponential. After establishing the properties of cosine and sine, we carry out our usual program of differentiating and integrating these functions.

Section 7.1 establishes the fact that the circumference of the unit circle is its diameter times its area, a fact relevant to the properties of cosine and the sine. Section 7.2 defines the cosine and the sine, and section 7.3 establishes some of their properties: basic identities, angle sum and difference formulas, double and half angle formulas, product and difference formulas. Using these properties, we can differentiate the cosine and the sine in section 7.4, and we can integrate them in section 7.5. Section 7.6 introduces other trigonometric functions, the tangent, cotangent, secant, and cosecant, and section 7.7 introduces their inverse functions. All of these functions are within our power to differentiate by using progenitive derivative rules, but for the most part we don't yet know how to integrate them.

7.1 The Circumference of the Unit Circle

Let π denote the area of the unit circle, and let c denote its circumference. We quickly review the proof that $c = 2\pi$. Figure 7.1 shows n triangles circumscribing the unit circle ($n = 8$ in the figure, and in general we assume that n is even). One triangle is shaded, and the altitude from its vertex at the center of the circle to the center of its base is shown. Because the altitude is the circle's radius, each triangle has height 1. Also, the triangle bases have total

length slightly larger than the circle circumference c. Meanwhile, figure 7.2 shows the same triangles rearranged into a rectangle whose base is half the sum of the triangle bases and whose height is 1. The area of the circumscribing polygon in figure 7.1 equals the area of the rectangle in figure 7.2. As n grows, the area of the circumscribing polygon in figure 7.1 tends to the area π of the unit circle, while the base of the rectangle in figure 7.2 tends to half the circle circumference c and its height is always 1 and so the area of the rectangle tends to $c/2$. In the limit, the equal areas are π and $c/2$, from which the desired result follows,

$$c = 2\pi.$$

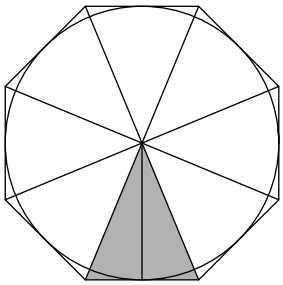


Figure 7.1. Triangles circumscribing the unit circle

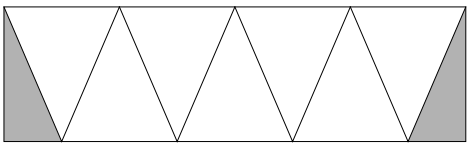


Figure 7.2. Rearrangement of the circumscribing triangles

7.2 Definition of the Cosine and the Sine

Let $s \in \mathcal{R}_{\geq 0}$ be a nonnegative real number. Starting at $(1, 0)$, proceed counterclockwise along the unit circle through arc length s, and let $P(s)$ denote the point thus reached. The x- and y-coordinates are respectively the **cosine** and the **sine** of s,

$$P(s) = (\cos(s), \sin(s)). \tag{7.1}$$

For $s \in \mathcal{R}_{<0}$, i.e., for negative s, proceeding counterclockwise along the unit circle through arc length s has the obvious interpretation of proceeding clockwise instead through arc length $-s$. Again let $P(s)$ denote the point thus reached, and extend formula (7.1) to this case as well. Thus we have defined functions

$$\cos, \sin : \mathcal{R} \longrightarrow [-1, 1].$$

So, for example, starting at $(1, 0)$ and then going one-quarter, one-half, three-quarters, and all the way around the unit circle counterclockwise gives the values

$$\begin{aligned}
&(\cos(0), \sin(0)) = (1, 0), \\
&(\cos(\pi/2), \sin(\pi/2)) = (0, 1), \\
&(\cos(\pi), \sin(\pi)) = (-1, 0), \\
&(\cos(3\pi/2), \sin(3\pi/2)) = (0, -1), \\
&(\cos(2\pi), \sin(2\pi)) = (1, 0).
\end{aligned}$$

The graphs of the cosine and sine functions (or rather, portions of the graphs) are shown in figure 7.3.

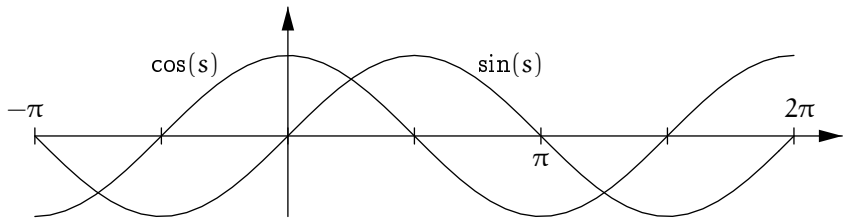


Figure 7.3. Cosine and sine

7.3 Identities for the Cosine and the Sine

7.3.1 Basic Identities

The following properties of sine and cosine are consequences of the definition. For all $s \in \mathcal{R}$,

$$\sin(s)^2 + \cos(s)^2 = 1, \tag{7.2}$$

$$\cos(s + 2\pi n) = \cos(s) \quad \text{for all } n \in \mathcal{Z}, \tag{7.3}$$

$$\sin(s + 2\pi n) = \sin(s) \quad \text{for all } n \in \mathcal{Z}, \tag{7.4}$$

$$\cos(-s) = \cos(s), \tag{7.5}$$

$$\sin(-s) = -\sin(s), \tag{7.6}$$

$$\cos(\pi/2 - s) = \sin(s), \tag{7.7}$$

$$\sin(\pi/2 - s) = \cos(s). \tag{7.8}$$

Property (7.2) holds by the Pythagorean Theorem because $P(s)$ lies on the unit circle. Properties (7.3) and (7.4) hold because the circle has circumference 2π. Properties (7.5) and (7.6) hold because proceeding from $(1, 0)$ clockwise rather than counterclockwise but through the same arc length s gives the same x-coordinate but the opposite y-coordinate. Properties (7.7) and (7.8) hold because starting at $(0, 1)$ (rather than at $(1, 0)$) and proceeding *clockwise* through arc length s along the unit circle is the reflection through the 45-degree line $y = x$ of starting at $(1, 0)$ and proceeding counterclockwise through arc length s. Thus the x-coordinate produced by the first process, $\cos(\pi/2 - s)$, must equal the y-coordinate produced by the second, $\sin(s)$, and similarly $\sin(\pi/2 - s) = \cos(s)$.

The reader should visually identify as many as possible of the basic properties of cosine and sine in figure 7.3. Also, anticipating a result to come, the reader should see that plausibly the tangent slope of the sine graph is the height of the cosine graph, and the tangent slope of the cosine graph is minus the height of the sine graph. That is, the graphs suggest (as we will show carefully soon) that $\sin' = \cos$ and $\cos' = -\sin$.

Exercise

7.3.1. Consider the functions $f, g, h : [0, \pi/2] \longrightarrow \mathcal{R}$ where

$$f(s) = (\sin(4s))^5, \qquad g(s) = (\sin(3s))^5, \qquad h(s) = (\cos(3s))^5,$$

and consider the quantities

$$A = \int_0^{\pi/2} f, \qquad B = \int_0^{\pi/2} g, \qquad C = \int_0^{\pi/2} h.$$

Arrange A, B, and C in increasing order. The idea here is not to compute precisely but to reason, with explanation, from rough sketches (which you should show) of the graphs of f, g, and h.

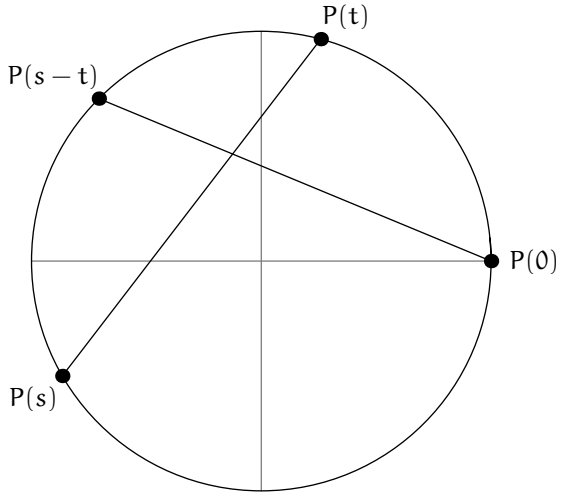


Figure 7.4. Geometry for the angle sum and difference formulas

7.3.2 Angle Sum and Difference Formulas

Let $s, t \in \mathcal{R}$ be any real numbers. As far as their cosines and sines are concerned, we may assume that in fact $s, t \in [0, 2\pi)$, and for the moment we further assume that $s > t$. Figure 7.4 shows the points $P(0) = (1, 0)$, $P(t) = (\cos(t), \sin(t))$, $P(s) = (\cos(s), \sin(s))$, $P(s-t) = (\cos(s-t), \sin(s-t))$. The arc length distance $s - t$ from $P(0)$ to $P(s - t)$ equals the arc length distance from $P(t)$ to $P(s)$, so the corresponding linear distances are equal as well, and hence so are their squares. By the Pythagorean Theorem, the equality of squares of distances is

$$(\cos(s - t) - 1)^2 + \sin^2(s - t) = (\cos(s) - \cos(t))^2 + (\sin(s) - \sin(t))^2,$$

or, after a little algebra that makes use of the first basic identity (7.2),

$$\cos(s - t) = \cos(s)\cos(t) + \sin(s)\sin(t).$$

Because this formula is unaffected by exchanging s and t, our initial assumption that $s > t$ (after both are translated to lie in $[0, 2\pi)$) is unnecessary, and the formula holds for all $s, t \in \mathcal{R}$. Altogether the angle sum and difference formulas are

$$\cos(s+t) = \cos(s)\cos(t) - \sin(s)\sin(t), \tag{7.9}$$
$$\cos(s-t) = \cos(s)\cos(t) + \sin(s)\sin(t), \tag{7.10}$$
$$\sin(s+t) = \sin(s)\cos(t) + \cos(s)\sin(t), \tag{7.11}$$
$$\sin(s-t) = \sin(s)\cos(t) - \cos(s)\sin(t). \tag{7.12}$$

We have established (7.10). The others follow from substitutions in (7.10) and the basic identities. The slight elaborateness of these formulas corresponds to the slightly complicated way that the oscillating x- and y-coordinates of a point moving around the circle are related to each other.

Exercise

7.3.2. Prove the other three angle sum and difference formulas.

7.3.3 Double and Half Angle Formulas

For all $s \in \mathcal{R}$,

$$\cos(2s) = \cos^2(s) - \sin^2(s) = 2\cos^2(s) - 1 = 1 - 2\sin^2(s), \tag{7.13}$$
$$\sin(2s) = 2\sin(s)\cos(s), \tag{7.14}$$
$$\cos^2(s/2) = \tfrac{1}{2}(1 + \cos(s)), \tag{7.15}$$
$$\sin^2(s/2) = \tfrac{1}{2}(1 - \cos(s)). \tag{7.16}$$

Here (7.13) and (7.14) follow from (7.9) and the first basic identity (7.2) and (7.11). Then (7.15) and (7.16) follow from (7.13).

Exercise

7.3.3. Prove the double and half angle formulas.

7.3.4 Product Formulas

For all $s, t \in \mathcal{R}$,

$$\cos(s)\cos(t) = \tfrac{1}{2}(\cos(s-t) + \cos(s+t)), \tag{7.17}$$
$$\sin(s)\sin(t) = \tfrac{1}{2}(\cos(s-t) - \cos(s+t)), \tag{7.18}$$
$$\cos(s)\sin(t) = \tfrac{1}{2}(\sin(s+t) - \sin(s-t)). \tag{7.19}$$

Here (7.17) follows from adding (7.9) and (7.10), and (7.18) follows from subtracting (7.9) from (7.10), and (7.19) follows from subtracting (7.12) from (7.11).

Exercise

7.3.4. Prove the product formulas.

7.3.5 Difference Formulas

For all $s, t \in \mathcal{R}$,

$$\cos(s) - \cos(t) = -2 \sin\left(\tfrac{s+t}{2}\right) \sin\left(\tfrac{s-t}{2}\right), \tag{7.20}$$

$$\sin(s) - \sin(t) = 2 \cos\left(\tfrac{s+t}{2}\right) \sin\left(\tfrac{s-t}{2}\right). \tag{7.21}$$

Here (7.20) follows from substituting $\frac{s+t}{2}$ for s and $\frac{s-t}{2}$ for t in (7.18). And (7.21) follows from the same substitutions in (7.19).

Exercises

7.3.5. Prove the difference formulas.

7.3.6. Complete the table of sines and cosines whose first half is

	0	$\pi/6$	$\pi/4$	$\pi/3$	$\pi/2$	$2\pi/3$	$3\pi/4$	$5\pi/6$	π
sin									
cos									

and whose second half is

	π	$7\pi/6$	$5\pi/4$	$4\pi/3$	$3\pi/2$	$5\pi/3$	$7\pi/4$	$11\pi/6$	2π
sin									
cos									

No justification of the table values is requested in your written work, but do derive them from the results of this course rather than quoting them from memory or elsewhere. To find $\sin(\pi/6)$ and $\cos(\pi/6)$, note that the triangle having vertices $(0,0)$, $(\cos(\pi/6), \sin(\pi/6))$, and $(\cos(-\pi/6), \sin(-\pi/6))$ is equilateral of side 1.

7.4 Differentiation of the Cosine and the Sine

The cosine and sine functions oscillate so regularly that one normalized derivative calculation—that $\sin'(0) = 1$—quickly gives the derivatives of the sine and cosine everywhere. To differentiate the sine at 0, the first step is to establish some estimates using geometry and algebra.

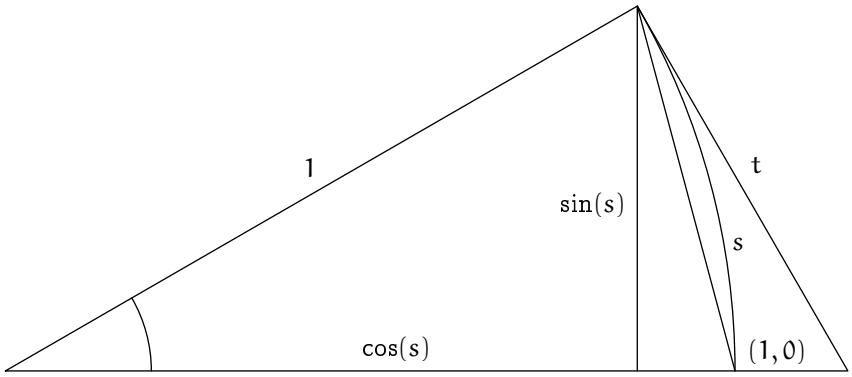


Figure 7.5. Analysis of sine and cosine

Lemma 7.4.1. *For all* $s \in (0, \pi/2)$,

$$0 \leq \sin(s) \leq s \leq \sin(s)/\cos(s).$$

The geometry of the arguments in the proof to follow is shown in figure 7.5.

Proof. Consider any $s \in (0, \pi/2)$. Because the point $(\cos(s), \sin(s))$ lies in the first quadrant of the cartesian plane, $0 \leq \sin(s)$.

For next inequality, note that $\sin(s)$ is the distance from $(\cos(s), 0)$ to $(\cos(s), \sin(s))$, which is less than the distance from $(1, 0)$ to $(\cos(s), \sin(s))$, and this is in turn less than the arc length from $(1, 0)$ to $(\cos(s), \sin(s))$, which is s. Thus $\sin(s) \leq s$.

For the third inequality, the right triangle in the figure with legs $\sin(s)$ and $\cos(s)$ is similar to the right triangle in the figure with legs t and 1, so that

$$t = \sin(s)/\cos(s).$$

The figure suggests strongly that also $t \geq s$. We invoke this as an assumption. (For Archimedes, it was a particular consequence of a more general assumption about curves that are *concave in the same direction*, the curves in this case being s and t each joined with its reflection through the horizontal axis, so that they start at a shared point and end at a shared point.) Thus $s \leq \sin(s)/\cos(s)$. □

The next proposition is self-evident geometrically, but we give a quick analytic proof.

Proposition 7.4.2 (Continuity of the Sine and Cosine Functions). *The sine and cosine functions are continuous at* 0*, meaning that*

$$\lim_{s \to 0} \sin(s) = 0, \qquad \lim_{s \to 0} \cos(s) = 1.$$

Proof. For the first limit, the lemma's inequalities $0 \le \sin(s) \le s$ for $0 < s < \pi/2$ give $0 \le |\sin(s)| \le |s|$ for $0 < |s| < \pi/2$, because $\sin(s)$ and s simply change their signs when s is negative rather than positive. Because $\lim_{s\to 0} |s| = 0$, the Squeeze Rule for function limits gives $\lim_{s\to 0} |\sin(s)| = 0$ in consequence, from which $\lim_{s\to 0} \sin(s) = 0$ as desired (see (4.2) on page 123).

For the second limit, again see figure 7.5. For $0 < |s| < \pi/2$, note that $|\cos(s) - 1|$ is the distance from $(1,0)$ to $(\cos(s),0)$, which is less than the distance from $(1,0)$ to $(\cos(s),\sin(s))$, and this is in turn less than the arc length from $(1,0)$ to $(\cos(s),\sin(s))$, which is $|s|$. That is, $|\cos(s) - 1| \le |s|$ for $0 < |s| < \pi/2$. The Squeeze Rule for function limits gives $\lim_{s\to 0} |\cos(s) - 1| = 0$, from which $\lim_{s\to 0} \cos(s) = 1$ as desired (see (4.1) on page 123).

□

Proposition 7.4.3 (Derivatives of the Sine and the Cosine at 0). *The sine and cosine are differentiable at 0, and*

$$\sin'(0) = 1, \qquad \cos'(0) = 0.$$

Proof. First we discuss the sine. Because $\sin(-s)/(-s) = \sin(s)/s$ for all nonzero s, we may consider the function

$$g : (0, \pi/2) \longrightarrow \mathcal{R}, \quad g(s) = \frac{\sin(s)}{s}$$

and show that $\lim_{s\to 0} g(s) = 1$. Note that s, $\sin(s)$, and $\cos(s)$ are positive for $s \in (0, \pi/2)$. The inequality $s \le \sin(s)/\cos(s)$ from the lemma gives $\cos(s) \le \sin(s)/s$, and the inequality $\sin(s) \le s$ from the lemma gives $\sin(s)/s \le 1$, so together

$$\cos(s) \le \frac{\sin(s)}{s} \le 1 \quad \text{for all } s \in (0, \pi/2).$$

Because $\lim_{s\to 0} \cos(s) = 1$ by the previous proposition, the Squeeze Rule for function limits shows that

$$\lim_{s\to 0} \frac{\sin(s)}{s} = 1.$$

That is,

$$\lim_{s\to 0} \frac{\sin(s) - \sin(0)}{s - 0} = 1,$$

which is to say that $\sin'(0)$ exists and equals 1.

Second we discuss the cosine, obtaining the desired result from the value $\sin'(0) = 1$ and from other work that we have done. By (7.16) (page 214),

$$\frac{\cos(s) - 1}{s} = \frac{-2\sin^2(s/2)}{s} = -\sin(s/2)\frac{\sin(s/2)}{s/2}.$$

As s goes to 0 so does $s/2$, and hence so does $\sin(s/2)$ by the previous proposition. And $\sin(s/2)/(s/2)$ goes to $\sin'(0) = 1$. Hence overall,

$$\lim_{s\to 0} \frac{\cos(s) - \cos(0)}{s - 0} = -0 \cdot 1 = 0,$$

which is to say that $\cos'(0)$ exists and equals 0.

For a more geometric proof of this, again see figure 7.5. In the figure, $1 - \cos(s)$ is the horizontal distance from $(1, 0)$ back to $(\cos(s), 0)$ (the reader should identify these points in the figure), and this is less than the horizaontal distance from the rightmost point in the figure back to $(\cos(s), 0)$. With reference to the quantity t in the figure, this latter distance is $t\sin(s)$ because the rightmost angle in the figure is $\pi/2 - s$, and so because $t = \sin(s)/\cos(s)$ this distance is $\sin^2(s)/\cos(s)$. So now we have for $0 < s < \pi/2$,

$$0 \le \frac{1 - \cos(s)}{s} \le \frac{\sin^2(s)}{s\cos(s)} = \frac{\sin(s)}{s} \cdot \frac{\sin(s)}{\cos(s)}.$$

As s goes to 0, the first quotient on the right side goes to $\sin'(0) = 1$ and the second to 0, and so the quotient $\frac{1-\cos(s)}{s} = \frac{-\cos(s)-(-\cos(0))}{s-0}$ is squeezed to 0. That is, $(-\cos)'(0)$ exists and is 0, from which $-\cos'(0) = 0$ by the constant multiple rule, and thus $\cos'(0) = 0$.

□

As mentioned, the general derivatives of cosine and sine are consequences of their regular oscillatory nature and of the normalized derivative value $\sin'(0) = 1$, which quickly gave $\cos'(0) = 0$ as well.

Theorem 7.4.4 (Derivatives of Cosine and Sine). *The cosine and the sine functions are differentiable on $\mathcal{R}$, and*

$$\cos' = -\sin, \qquad \sin' = \cos.$$

Proof. Fix any value $x \in \mathcal{R}$. Write any $s \in \mathcal{R}$ as $s = x + \tilde{s}$, so that $s - x = \tilde{s}$. Then

$$\begin{aligned}
\frac{\cos(s) - \cos(x)}{s - x} &= \frac{\cos(x + \tilde{s}) - \cos(x)}{\tilde{s}} \\
&= \frac{\cos(x)\cos(\tilde{s}) - \sin(x)\sin(\tilde{s}) - \cos(x)}{\tilde{s}} \\
&= \cos(x)\frac{\cos(\tilde{s}) - 1}{\tilde{s}} - \sin(x)\frac{\sin(\tilde{s})}{\tilde{s}}.
\end{aligned}$$

Thus, because s goes to x exactly when $\tilde{s}$ goes to 0, the normalized values $\cos'(0) = 0$ and $\sin'(0) = 1$ give

$$\lim_{s\to x}\frac{\cos(s)-\cos(x)}{s-x} = \cos(x)\lim_{\tilde{s}\to 0}\frac{\cos(\tilde{s})-1}{\tilde{s}} - \sin(x)\lim_{\tilde{s}\to 0}\frac{\sin(\tilde{s})}{\tilde{s}}$$
$$= \cos(x)\cdot 0 - \sin(x)\cdot 1 = -\sin(x).$$

In other words, $\cos' = -\sin$. The argument that $\sin' = \cos$ can be done virtually identically, or a nonrepetitive method is to use the identity $\sin(x) = \cos(\pi/2 - x)$ and the just-derived result $\cos' = -\sin$ and the chain rule and the identity $\sin(\pi/2 - x) = \cos(x)$ (exercise 7.4.1). □

This section has used trigonometry identities through the half angle formula (7.16), but not the difference formulas (7.20) and (7.21), which come from the product formulas (7.18) and (7.19). If we use the difference formulas then we don't need the normalized result $\cos'(0) = 0$, only $\sin'(0) = 1$; for example,

$$\frac{\cos(s)-\cos(x)}{s-x} = \frac{-2\sin(\frac{s+x}{2})\sin(\frac{s-x}{2})}{s-x} = -\sin(\tfrac{s+x}{2})\frac{\sin(\frac{s-x}{2})}{\frac{s-x}{2}},$$

and as s goes to x this goes to $-\sin(\frac{x+x}{2})\cdot 1 = -\sin(x)$.

Exercises

7.4.1. Carry out the argument that $\sin' = \cos$, by repeating the method of Theorem 7.4.4 or by following the suggested method at the end of the theorem's proof or by using the difference formula (7.21).

7.4.2. Find the derivatives of the following functions.
(a) $f(x) = \ln(\cos(x) + 2)$.
(b) $f(x) = \sin(4(x^3 + 2))$.
(c) $f(x) = \ln((\sin(x) + 1)/\cos(x))$ for $x \in (-\pi/2, \pi/2)$.

7.5 Integration of the Cosine and the Sine

7.5.1 Imaginary Exponential

To integrate the cosine and sine functions gracefully, once again via the finite geometric sum formula and a normalized derivative value, we use so-called *complex numbers*.

- First introduce the so-called *imaginary* number i, having the properties $i^2 = -1$ and $ir = ri$ for all $r \in \mathcal{R}$.
- A general complex number is a number $x + iy$ where $x, y \in \mathcal{R}$.

Just as using the word *real* does not cause anything to exist, nor does using the word *imaginary* prevent anything from existing, and nor does using the word *complex* make anything complicated; the complex number system is every bit as extant as the real number system and is a simple extension of it, both systems' adjectival names connoting only historical respect for the language of mathematics as it grappled with ideas. Indeed, one of the first mathematicians who solved a real number problem by using complex numbers used the words *putting aside the mental tortures involved* in describing his work.

One can check that addition, subtraction, multiplication, and division of complex numbers all work well, i.e., these operations satisfy the basic laws of algebra. The imaginary number i solves the polynomial equation $X^2 + 1 = 0$, which has no real solution. More generally, although a polynomial with real coefficients needn't have real roots, the *Fundamental Theorem of Algebra* says that every nonconstant polynomial with complex coefficients has complex roots (the coefficients could be real, and then the roots might be real). Thus, introducing i corrects an algebraic deficiency of the real number system.

Now further define the exponential of any *purely imaginary* complex number, meaning a real number multiple of i,

$$\boxed{e^{i\theta} = \cos(\theta) + i\sin(\theta), \quad \theta \in \mathcal{R}.} \tag{7.22}$$

A computation (exercise 7.5.1) shows that the addition laws for the cosine and sine functions, which we have established, two slightly complicated formulas that entwine the two functions,

$$\begin{aligned} \cos(\theta+\phi) &= \cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi) \\ \sin(\theta+\phi) &= \sin(\theta)\cos(\phi) + \cos(\theta)\sin(\phi), \end{aligned}$$

make the imaginary exponential work as we would hope, by taking sums to products, as expressed in one simple formula involving only one function,

$$e^{i(\theta+\phi)} = e^{i\theta}e^{i\phi}.$$

It follows that $e^{i2\theta} = e^{i(\theta+\theta)} = e^{i\theta}e^{i\theta} = (e^{i\theta})^2$, and then $e^{i3\theta} = e^{i(2\theta+\theta)} = e^{i2\theta}e^{i\theta} = (e^{i\theta})^2e^{i\theta} = (e^{i\theta})^3$, and in general

$$e^{ij\theta} = (e^{i\theta})^j, \quad j = 0, 1, 2, \ldots .$$

Here are perhaps the most fundamental entities from mathematics:

- Numbers
  - 1 and 0: basic to arithmetic; the multiplicative and additive identities

  - π: arises naturally in geometry as the area of the unit circle
  - e: arises naturally in analysis as the number whose logarithm is 1
  - i: arises naturally in algebra as a solution of $X^2 + 1 = 0$
- Operations
  - addition and multiplication: built into the real and complex number systems
  - exponentiation: relates addition and multiplication
- Relation
  - equality.

As a special case of the definition $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, set $\theta = \pi$ to get $e^{i\pi} = -1$, or

$$e^{i\pi} + 1 = 0.$$

In a marvel of efficiency, this comprises exactly the five fundamental numbers, the three fundamental operations, and the one fundamental relation, once each and nothing more.

Exercise

7.5.1. Show that Definition 7.22 and the addition laws for the cosine and sine functions give the law $e^{i(\theta+\phi)} = e^{i\theta}e^{i\phi}$ for all $\theta, \phi \in \mathcal{R}$.

7.5.2. Definition 7.22 expresses the imaginary exponential in terms the cosine and sine functions. Evaluate $(e^{i\theta} + e^{i(-\theta)})/2$ and $(e^{i\theta} - e^{i(-\theta)})/(2i)$ to show that also the cosine and sine functions can be expressed in terms of the imaginary exponential.

7.5.2 The Integrals Via the Imaginary Exponential

The cosine and sine functions are both bounded, because their values lie in $[-1, 1]$. Also, the cosine function is

$$\begin{aligned}
&\text{increasing on } [-\pi, 0] \text{ and decreasing on } [0, \pi],\\
&\text{decreasing on } [-2\pi, -\pi] \text{ and increasing on } [\pi, 2\pi],\\
&\text{increasing on } [-3\pi, -2\pi] \text{ and decreasing on } [2\pi, 3\pi],
\end{aligned}$$

and so on, while the sine function is

$$\begin{aligned}
&\text{increasing on } [-\pi/2, \pi/2],\\
&\text{decreasing on } [-3\pi/2, -\pi/2] \text{ and } [\pi/2, 3\pi/2],\\
&\text{increasing on } [-5\pi/2, -3\pi/2] \text{ and } [3\pi/2, 5\pi/2],
\end{aligned}$$

and so on. All of this is self-evident if one remembers the interpretation of the cosine and the sine as the coordinates of a point that moves around the circle.

Theorem 7.5.1 (Integrals of the Cosine and the Sine). *Let* a *and* b *be any real numbers. Then the integrals* $\int_a^b \cos$ *and* $\int_a^b \sin$ *exist, and their values are*

$$\boxed{\int_a^b \cos = \sin(b) - \sin(a)}$$

and

$$\boxed{\int_a^b \sin = \cos(a) - \cos(b).}$$

We will prove this nicely by using the imaginary exponential.

Proof. As usual, if we can prove the results for $a \leq b$ then they follow as well for $a > b$. So we assume that $a \leq b$, and in fact we assume that $a < b$ because the case $a = b$ is clear.

The sine and the cosine are bounded and piecewise monotonic. Any bounded monotonic function is integrable by Theorem 3.3.8 on page 108 (for nonnegative such functions) and the discussion in chapter 5 (extending the integral to bounded functions that can also take negative values). The sine and the cosine are therefore integrable by Proposition 3.3.9 on page 109. Also by the proposition, we may carry out the integral over an interval $[a, b]$ where the function (sine or cosine) is monotonic. First we address the normalized case $a = 0$.

Consider a uniform partition of $[0, b]$ into n subintervals, each having width b/n. The sum

$$\begin{aligned} S_n = \frac{b}{n}\left(\cos(\tfrac{0b}{n}) + \cos(\tfrac{1b}{n}) + \cos(\tfrac{2b}{n}) + \cdots + \cos(\tfrac{(n-1)b}{n}\right) \\ + i\frac{b}{n}\left(\sin(\tfrac{0b}{n}) + \sin(\tfrac{1b}{n}) + \sin(\tfrac{2b}{n}) + \cdots + \sin(\tfrac{(n-1)b}{n}\right) \end{aligned}$$

has limit

$$\lim_n S_n = \int_0^b \cos + i\int_0^b \sin .$$

Rearrange S_n and then use the imaginary exponential definition $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ to get

$$\begin{aligned}
S_n &= \frac{b}{n}\left(\cos(\tfrac{0b}{n})+\cos(\tfrac{1b}{n})+\cos(\tfrac{2b}{n})+\cdots+\cos(\tfrac{(n-1)b}{n})\right)\\
&\quad+i\frac{b}{n}\left(\sin(\tfrac{0b}{n})+\sin(\tfrac{1b}{n})+\sin(\tfrac{2b}{n})+\cdots+\sin(\tfrac{(n-1)b}{n})\right)\\
&=\frac{b}{n}\left(\begin{array}{l}\cos(\tfrac{0b}{n})+\cos(\tfrac{1b}{n})+\cos(\tfrac{2b}{n})+\cdots+\cos(\tfrac{(n-1)b}{n})\\ \quad+i\sin(\tfrac{0b}{n})+i\sin(\tfrac{1b}{n})+i\sin(\tfrac{2b}{n})+\cdots+i\sin(\tfrac{(n-1)b}{n})\end{array}\right)\\
&=\frac{b}{n}\left(e^{i0b/n}+e^{i1b/n}+e^{i2b/n}+\cdots+e^{i(n-1)b/n}\right),
\end{aligned}$$

and now, very similarly to the calculation beginning with *It is also very easy to integrate the exponential* on page 201 we have, by the property $e^{ij\theta}=(e^{i\theta})^j$, by the finite geometric sum formula, and by $(e^{i\theta})^n=(e^{in\theta})$, that this sum is

$$\begin{aligned}
S_n &= \frac{b}{n}\left(e^{i0b/n}+e^{i1b/n}+e^{i2b/n}+\cdots+e^{i(n-1)b/n}\right)\\
&=\frac{b}{n}\left(1+e^{ib/n}+(e^{ib/n})^2+\cdots+(e^{ib/n})^{n-1}\right)\\
&=\frac{b}{n}\cdot\frac{(e^{ib/n})^n-1}{e^{ib/n}-1}\\
&=\frac{b}{n}\cdot\frac{e^{ib}-1}{e^{ib/n}-1}.
\end{aligned}$$

Noting that $1=-i\cdot i$, this is

$$S_n=-i(e^{ib}-1)\Big/\frac{e^{ib/n}-e^0}{ib/n-0}.$$

As n grows large, the quotient $(e^{ib/n}-e^0)/(ib/n-0)$ in the previous display tends to $\exp'(0)=1$, and so overall the sum that we are evaluating has limit

$$\lim_n S_n=-i(e^{ib}-1)=-i(\cos(b)-1+i\sin(b))=\sin(b)+i(1-\cos(b)).$$

That is, because also $\lim_n S_n=\int_0^b\cos+i\int_0^b\sin$ as noted after the definition of S_n at the beginning of this paragraph,

$$\int_0^b\cos=\sin(b),\qquad\int_0^b\sin=1-\cos(b).$$

In general $\int_a^b\cos=\int_0^b\cos-\int_0^a\cos$ and similarly for the sine function, so the claimed results follow,

$$\int_a^b\cos=\sin(b)-\sin(a),\qquad\int_a^b\sin=\cos(a)-\cos(b).$$

□

The reader could reasonably be concerned that because the proof just given uses complex numbers, the finite geometric sum and normalized derivative value that it cites are unjustified variants of justified results. We gladly work with these variants because they are very natural and they let us make such a lovely argument. The variant finite geometric sum for complex numbers is established by exactly the same basic algebra that we used in section 2.1.2 to establish it for real numbers. As for the variant normalized derivative, see exercise 7.5.3(a).

Exercise

7.5.3. (a) Show that

$$\frac{e^{is}-1}{is} = -i\left(\frac{\cos(s)-1}{s} + i\frac{\sin(s)}{s}\right).$$

Use known normalized derivative values of the cosine and sine functions to explain why the two quotients on the right side of the equality go respectively to 0 and 1 as s goes to 0. Consequently the quotient on the left side goes to 1. This is the normalized derivative value that was used in the section with $s = b/n$.

(b) Show that for all $s, x \in \mathcal{R}$ with $s \neq x$,

$$\frac{e^{is}-e^{ix}}{i(s-x)} = \frac{e^{i(s-x)}-1}{i(s-x)}e^{ix}.$$

By part (a), with $s-x$ here playing the role of s there, the limit of this quotient, as the variable value s approaches the fixed value x, is e^{ix}. Explain how this shows that

$$\lim_{s\to x}\frac{\cos(s)-\cos(x)}{s-x} = -\sin(x) \quad \text{and} \quad \lim_{s\to x}\frac{\sin(s)-\sin(x)}{s-x} = \cos(x).$$

thus reproving Theorem 7.4.4.

7.6 Other Trigonometric Functions

Definition 7.6.1 (Tangent, Cotangent, Secant, Cosecant). *The* **tangent**, **cotangent**, **secant**, *and* **cosecant** *functions have the formulas*

$$\begin{aligned}
\tan(s) &= \sin(s)/\cos(s),\\
\cot(s) &= \cos(s)/\sin(s),\\
\sec(s) &= 1/\cos(s),\\
\csc(s) &= 1/\sin(s).
\end{aligned}$$

The domains of these functions are the largest subsets of $\mathcal{R}$ that avoid dividing by 0. For example, the domain of the tangent is

$$\operatorname{dom}(\tan) = \{s \in \mathcal{R} : s \neq \pm\pi/2,\ \pm 3\pi/2,\ \pm 5\pi/2,\ \ldots\}.$$

Example 7.6.2. We differentiate the tangent. Any point $s \in \operatorname{dom}(\tan)$ is also approachable from $\operatorname{dom}(\tan)$, and

$$\begin{aligned}
\tan'(s) &= \left(\frac{\sin}{\cos}\right)'(s) && \text{by definition of the tangent} \\
&= \left(\frac{\sin' \cdot \cos - \sin \cdot \cos'}{\cos^2}\right)(s) && \text{by the Quotient Rule} \\
&= \left(\frac{\cos^2 + \sin^2}{\cos^2}\right)(s) && \text{because } \sin' = \cos \text{ and } \cos' = -\sin \\
&= \frac{1}{\cos^2(s)} && \text{because } \cos^2 + \sin^2 = 1 \\
&= \sec^2(s).
\end{aligned}$$

That is,

$$\boxed{\tan' = \sec^2.}$$

Exercise 7.6.1 is to show that also

$$\boxed{\cot' = -\csc^2}$$

and

$$\boxed{\sec' = \tan \cdot \sec}$$

and

$$\boxed{\csc' = -\cot \cdot \csc.}$$

Exercises

7.6.1. What is the domain of the cotangent? The secant? The cosecant? Show that

$$\cot' = -\csc^2, \qquad \sec' = \tan \cdot \sec, \qquad \csc' = -\cot \cdot \csc.$$

7.6.2. (a) Describe the domain of the function given by the formula

$$f(x) = \ln(|\tan(x) + \sec(x)|).$$

Explain why f is differentiable and compute f'. The answer should simplify nicely.

(b) Similarly for

$$f(x) = -\ln(|\cot(x) + \csc(x)|).$$

(c) Similarly for

$$f(x) = \ln(|\tan(x/2)|).$$

7.7 Inverse Trigonometric Functions

Example 7.7.1 (The Inverse Cosine Function). Restrict the domain of the cosine to $[0, \pi]$. The resulting function

$$\cos : [0, \pi] \longrightarrow [-1, 1]$$

takes each value in its codomain exactly once, so it is invertible. The inverse function is the **arc-cosine**,

$$\arccos : [-1, 1] \longrightarrow [0, \pi].$$

A graph of the arc-cosine is shown in figure 7.6.

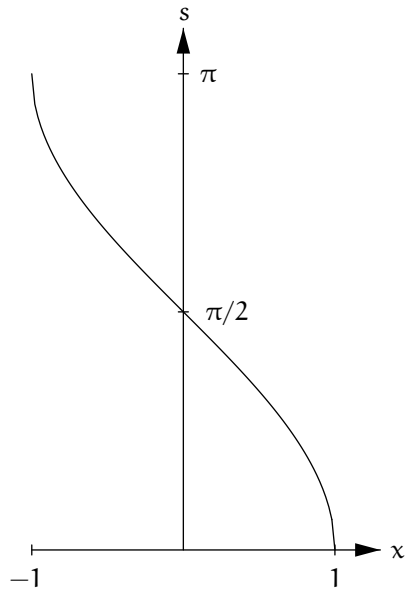


Figure 7.6. Graph of the arc-cosine function

Because we can integrate the cosine, we can integrate the arc-cosine by using the same method that let us integrate the exponential via the integral of the logarithm (or conversely) in section 6.5. Consider any value $b \in [-1, 1]$, let $c = \arccos(b)$, and consider figure 7.7. The dark-shaded area is $\int_{-1}^{b} \arccos$. The light-shaded area is $\int_{c}^{\pi} (b - \cos)$. The entire shaded area is $\pi(b + 1)$. Therefore,

$$\begin{aligned}\int_{-1}^{b} \arccos &= \pi(b+1) - \int_{c}^{\pi} (b - \cos) \\ &= \pi(b+1) - (\pi - c)b + \int_{c}^{\pi} \cos \\ &= \pi + bc + \sin(\pi) - \sin(c) \\ &= \pi + bc - \sin(c).\end{aligned}$$

But $c = \arccos(b)$ and $\sin(\arccos(b)) = \sqrt{1-b^2}$ (the general fact that

$$\sin(\arccos(x)) = \sqrt{1-x^2}\,, \quad x \in [-1,1]$$

is demonstrated by figure 7.8), and so

$$\int_{-1}^{b} \arccos = \pi + b\arccos(b) - \sqrt{1-b^2}.$$

(And we check that for $b = -1$ this gives 0 and for $b = 1$ this gives π, as it should.) It follows that more generally, for all $a, b \in [-1,1]$,

$$\boxed{\int_{a}^{b} \arccos = (b\arccos(b) - \sqrt{1-b^2}) - (a\arccos(a) - \sqrt{1-a^2}).}$$

We can also differentiate the arc-cosine. Note that

$$\cos(\arccos(x)) = x, \quad x \in [-1,1].$$

The Chain Rule says that consequently, for all $x \in [-1,1]$ such that the derivative $\arccos'(x)$ exists,

$$-\sin(\arccos(x)) \cdot \arccos'(x) = 1.$$

So for all $x \in [-1,1]$ such that the derivative $\arccos'(x)$ exists and furthermore $\sin(\arccos(x))$ is nonzero,

$$\arccos'(x) = -\frac{1}{\sin(\arccos(x))}\,.$$

But as already noted, $\sin(\arccos(x)) = \sqrt{1-x^2}$ for $x \in [-1,1]$. Thus $\sin(\arccos(x)) \neq 0$ for all $x \in (-1,1)$. Only the endpoints $x = \pm 1$ need to be excluded to avoid dividing by zero in the previous display. That is, *if* the derivative $\arccos'$ exists on $(-1,1)$ then it is

$$\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}\,, \quad -1 < x < 1.$$

But we still have to address the existence question.

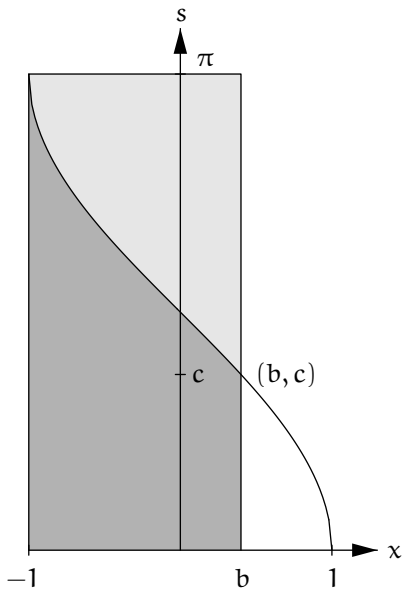


Figure 7.7. Integrating the arc-cosine

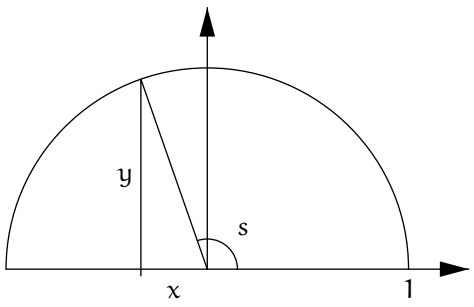


Figure 7.8. Figure to compute $\sin(\arccos(x))$

As with differentiating the exponential function, what needs to be shown for any $x \in (-1, 1)$ is that

$$\lim_{s \to x} \frac{\arccos(s) - \arccos(x)}{s - x} = -\frac{1}{\sqrt{1 - x^2}},$$

which is to say that

$$\text{for any sequence } (s_n) \text{ in } (-1,1) \text{ that approaches } x,$$
$$\lim_n \left(\frac{\arccos(s_n) - \arccos(x)}{s_n - x} \right) = -\frac{1}{\sqrt{1-x^2}}.$$

On the other hand, we know that $\cos'(\arccos(x)) = -\sqrt{1-x^2}$, which is to say that

$$\text{for any sequence } (t_n) \text{ in } (0,\pi) \text{ that approaches } \arccos(x),$$
$$\lim_n \left(\frac{\cos(t_n) - \cos(\arccos(x))}{t_n - \arccos(x)} \right) = -\sqrt{1-x^2},$$

and because $-\sqrt{1-x^2}$ is nonzero, and noting that $\cos(\arccos(x)) = x$, the Reciprocal Rule for sequence limits says that in fact

$$\text{for any sequence } (t_n) \text{ in } (0,\pi) \text{ that approaches } \arccos(x),$$
$$\lim_n \left(\frac{t_n - \arccos(x)}{\cos(t_n) - x} \right) = -\frac{1}{\sqrt{1-x^2}}.$$

So, consider any sequence (s_n) that approaches x in $(-1,1)$. Let $t_n = \arccos(s_n)$ for each n. We show that (t_n) approaches $\arccos(x)$. Let $\varepsilon > 0$ be given. Our goal is to make values t_n lie within ε of $\arccos(x)$, so we may assume that $\arccos(x) - \varepsilon > 0$ and $\arccos(x) + \varepsilon < \pi$ by shrinking ε if necessary. Consider the number $\rho_1 = \cos(\arccos(x) + \varepsilon)$, which is less than x, and the number $\rho_2 = \cos(\arccos(x) - \varepsilon)$, which is greater than x. Because $\lim_n(s_n) = x$, there is a starting generation N such that $\rho_1 < s_n < \rho_2$ for all $n \geq N$. Because the arc-cosine function is decreasing and $t_n = \arccos(s_n)$ for all n, it follows that $\arccos(x) - \varepsilon < t_n < \arccos(x) + \varepsilon$ for all $n \geq N$, showing that $\lim_n(t_n) = \arccos(x)$. Furthermore, because $s_n \neq x$ for all n, also $t_n \neq \arccos(x)$ for all n, and so (t_n) approaches $\arccos(x)$ as claimed.

Now, because (t_n) approaches $\arccos(x)$, we have

$$\lim_n \left(\frac{t_n - \arccos(x)}{\cos(t_n) - x} \right) = -\frac{1}{\sqrt{1-x^2}}.$$

And because $t_n = \arccos(s_n)$ and $\cos(t_n) = s_n$ for each n, this is exactly our desired sequence limit,

$$\lim_n \left(\frac{\arccos(s_n) - \arccos(x)}{s_n - x} \right) = -\frac{1}{\sqrt{1-x^2}}.$$

This completes the proof. In sum,

$$\boxed{\arccos'(x) = -\frac{1}{\sqrt{1-x^2}}, \quad -1 < x < 1.}$$

With the derivative of the arc-cosine in hand, we can see the Fundamental Theorem of Calculus in action once again. Define a function

$$F : [-1,1] \longrightarrow \mathcal{R}, \quad F(x) = x\arccos(x) - \sqrt{1-x^2}.$$

Then (exercise 7.7.1)

$$F'(x) = \arccos(x), \quad -1 < x < 1.$$

And so, as usual, for all $a, b \in (-1,1)$, glossing over the issue of the endpoints for now,

$$\int_a^b \arccos = F(b) - F(a), \quad F' = \arccos .$$

Similarly to the inverse cosine function, we can define the inverse sine,

$$\arcsin : [-1,1] \longrightarrow [-\pi/2, \pi/2],$$

the inverse tangent,

$$\arctan : \mathcal{R} \longrightarrow (-\pi/2, \pi/2),$$

and the inverse cotangent,

$$\operatorname{arccot} : \mathcal{R} \longrightarrow (0, \pi).$$

Exercises 7.7.2 and 7.7.3 are to discuss these functions, to calculate their derivatives, and to integrate the arc-sine.

Exercises

7.7.1. As in the section, let $F(x) = x\arccos(x) - \sqrt{1-x^2}$ for $x \in [-1,1]$. Show that $F' = \arccos$ on $(-1,1)$.

7.7.2. Restrict the domain of the sine to $[-\pi/2, \pi/2]$. The resulting function

$$\sin : [-\pi/2, \pi/2] \longrightarrow [-1,1]$$

takes each value in its codomain exactly once, so it is invertible. The inverse function is the **arc-sine**,

$$\arcsin : [-1,1] \longrightarrow [-\pi/2, \pi/2].$$

(a) Sketch a graph of the arc-sine.

(b) Explain why for all $y \in [-1,1]$ such that the derivative $\arcsin'(y)$ exists and $\cos(\arcsin(y))$ is nonzero,

$$\arcsin'(y) = \frac{1}{\cos(\arcsin(y))}\,.$$

(c) For generic $y \in [-1,1]$, draw a right triangle with vertices $(0,0)$, $(0,y)$, and (x,y) where the third vertex lies on the right half of the unit circle. Explain how this figure shows that for such y,

$$\cos(\arcsin(y)) = \sqrt{1-y^2}\,.$$

(d) Show that

$$\boxed{\arcsin'(y) = \frac{1}{\sqrt{1-y^2}}\,, \quad -1 < y < 1.}$$

(e) Compute $\int_a^b \arcsin$ for any $a, b \in [-1,1]$.

7.7.3. (a) Discuss the inverse tangent function

$$\arctan : \mathcal{R} \longrightarrow (-\pi/2, \pi/2)$$

and show that

$$\boxed{\arctan'(x) = \frac{1}{1+x^2}\,, \quad x \in \mathcal{R}.}$$

(b) Discuss the inverse cotangent function

$$\operatorname{arccot} : \mathcal{R} \longrightarrow (0,\pi)$$

and show that

$$\boxed{\operatorname{arccot}'(x) = -\frac{1}{1+x^2}\,, \quad x \in \mathcal{R}.}$$

8

Polynomial Approximation and Series Representation

The basic operations with real numbers are addition and multiplication. So the simplest functions are those that are evaluated by carrying out finitely many such operations. These functions are precisely the *polynomials*. It is natural to approximate the more complicated functions that we have studied—the power function for exponents other than nonnegative integers, the logarithm, the exponential, the trigonometric functions—by polynomials. And it is natural to investigate how good the approximations are, and whether the more complicated functions, despite not being polynomials, are somehow *limits* of polynomials.

This chapter derives approximating polynomials P_n of each degree n for the just-mentioned functions. For each such function f, for all x in a certain domain that may not be the full domain of f, the values $P_n(x)$ tend to $f(x)$ as n grows. Thus limits of polynomials provide a *uniform description* of the functions, despite the functions all being so different from each other. Also, estimates of how well the polynomials approximate their limit-functions provide a finite process to compute the functions to any desired accuracy. This makes the functions more tangible than they were previously.

Most of the arguments in this chapter are not hard to follow, but in a few places the calculations get detailed. The reader is encouraged to read those passages lightly rather than get bogged down.

Section 8.1 is somewhat of a warmup, expanding the polynomial $(1+x)^n$ in powers of x. Section 8.2 establishes preliminary results for the calculations to follow in the chapter—an alternative notation for the integral, better suited to computation, and a discussion of the integral of the power function when the left endpoint of integration is zero. Section 8.3 finds approximating polynomials and remainders for the logarithm, and section 8.4 does the same for the exponential, and section 8.5 does the same for the cosine and the sine. Finally, section 8.6 finds approximating polynomials and remainders for the

power function at $1+x$, i.e., for $(1+x)^\alpha$. The results here generalize the results of section 8.1.

8.1 The Finite Binomial Theorem

Let $n \in \mathcal{Z}_{\geq 0}$ be a nonnegative integer. The function

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = (1+x)^n$$

is a polynomial. To describe it in terms of nonnegative integer powers of x, introduce the notation

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!}, \quad k = 0, \ldots, n.$$

The numerator of the fraction is the product of k terms, starting at $n-0$ and decrementing to $n-(k-1)$. The denominator is also the product of k terms, from 1 to k, the **factorial** of k. When $k=0$ the numerator and the denominator are both understood to be 1 because a product of no terms naturally should be the multiplicatively neutral quantity 1, just as a sum of no terms is the additively neutral quantity 0.

For nonnegative integers n and k with $0 \leq k \leq n$, the binomial coefficients $\binom{n}{k}$ arrange themselves in a pleasing pattern. The first few are

$$\binom{0}{0} = \frac{1}{0!} = 1,$$

and

$$\binom{1}{0} = \frac{1}{0!} = 1, \quad \binom{1}{1} = \frac{1}{1!} = 1,$$

and

$$\binom{2}{0} = \frac{1}{0!} = 1, \quad \binom{2}{1} = \frac{2}{1!} = 2, \quad \binom{2}{2} = \frac{2\cdot 1}{2!} = 1,$$

and

$$\binom{3}{0} = \frac{1}{0!} = 1, \quad \binom{3}{1} = \frac{3}{1!} = 3, \quad \binom{3}{2} = \frac{3\cdot 2}{2!} = 3, \quad \binom{3}{3} = \frac{3\cdot 2\cdot 1}{3!} = 1.$$

This famous arrangement is *Pascal's Triangle*, in which each internal entry is the sum of the two entries above it on either side:

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \\
 & & & & & \text{etc.} & & & & &
\end{array}
$$

That is, the nth row's kth entry is $\binom{n}{k}$, where the top row is considered the 0th row and the left entry of each row is considered the 0th entry.

Theorem 8.1.1 (Finite Binomial Theorem). *Let $n \in \mathcal{Z}_{\geq 0}$ be a non-negative integer, and let $x \in \mathcal{R}$ be any real number. Then*

$$\boxed{(1+x)^n = 1 + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n, \quad x \in \mathcal{R},\ n \in \mathcal{Z}_{\geq 0}.}$$

A more compact formulation uses the Sigma-notation *for sums,*

$$\boxed{(1+x)^n = \sum_{j=0}^{n} \binom{n}{j}x^j, \quad x \in \mathcal{R},\ n \in \mathcal{Z}_{\geq 0}.}$$

The coefficients in the Finite Binomial Theorem form the appropriate row of Pascal's Triangle. For example,

$$(1+x)^5 = 1 + 5x + 10x^2 + 10x^3 + 5x^4 + x^5.$$

Because $\binom{n}{1} = n$, the Binomial Theorem confirms our observation in section 6.6.2 that $(1+x)^n > 1 + nx$ for any $x > 0$ and any $n > 1$.

The Finite Binomial Theorem is essentially a combinatorial statement, and so one can prove it (in the sense that a proof is a convincing argument rather than a formal ritual) by talking rather than manipulating symbols. Consider for example the x^3 term in $(1+x)^{10}$. Expand the tenfold product of two-term sums

$$(1+x)^{10} = (1+x)(1+x)\cdots(1+x)$$

by multiplying out all tenfold products that result from choosing a 1 or an x from each factor $1+x$. (There are $2^{10} = 1024$ such terms.) The coefficient of x^3 in the expansion of $(1+x)^{10}$ is the number of ways that we can make ten consecutive choices of an x or a 1 that result in three net x's and seven net 1's. The first x that we choose can be drawn from any of the ten factors $1+x$,

the second x can occur at any of the nine remaining factors, and the third x can occur at any of the eight factors then remaining in turn. Thus there are $10 \cdot 9 \cdot 8$ ways of choosing the three x's. But in this counting scheme, choosing the x's from, say, the second, fifth, and eighth factors is viewed as a separate event from choosing the x's from the fifth, second, and eighth factors, or from the second, eighth, and fifth factors, and so on. To eliminate overcounting, we must divide by the number of ways of arranging the labels (2, 5, and 8 in the example) of the three factors where we chose x. There are $3 \cdot 2 \cdot 1 = 3!$ such arrangements. So finally, the coefficient of x^3 in $(1+x)^{10}$ is

$$\frac{10 \cdot 9 \cdot 8}{3!} = \binom{10}{3}.$$

The argument with general nonnegative integers n and k (where $k \leq n$) in place of 10 and 3 is the same.

This section has expanded a certain polynomial (a nonnegative power of $1+x$) in powers of x, naturally obtaining an expansion with only finitely many terms. The rest of the chapter will expand nonpolynomial functions (see exercise 8.1.3) in powers of x as well, but the expansions will not terminate.

Exercises

8.1.1. Write out the binomial expansions for $(1+x)^3$ and for $(1+x)^4$, multiply them together, and confirm that you have obtained the binomial expansion for $(1+x)^7$.

8.1.2. Here is a third false argument about $(1+x/n)^n$, in the spirit of the two arguments on page 203. *According to the Binomial Theorem,*

$$\left(1+\frac{x}{n}\right)^n = 1 + \binom{n}{1}\frac{x}{n} + \binom{n}{2}\frac{x^2}{n^2} + \cdots + \binom{n}{n}\frac{x^n}{n^n}.$$

That is, because $\binom{n}{1} = n$, *it follows that* $(1+x/n)^n$ *equals* $1+x$ *plus finitely many terms, each of which is a constant times a negative power of* n. *Thus as* n *gets very large,* $(1+x/n)^n$ *tends toward the value* $1+x$.

This argument is incorrect. Explain at least one of its flaws.

8.1.3. (a) Explain why the derivative of any polynomial is another polynomial, and why the only polynomial that can equal its derivative, or equal the negative of the derivative of its derivative, is the zero polynomial.

(b) Why can't any of the functions ln, exp, cos, or sin be a polynomial?

8.1.4. With the Sigma-notation now introduced on page 235, albeit informally, we can write far more concise versions of computations that we have carried out.

(a) Identify the result early in chapter 2 that says that for $r \neq 1$,

$$\sum_{j=0}^{n-1} r^j = \frac{r^n - 1}{r - 1}.$$

(b) Identify the computation later in chapter 2 for $\alpha \in \mathcal{Q}$ other than -1, for $b > 1$, and with $s^n = b$, that

$$\begin{aligned}
(s-1)\sum_{j=0}^{n-1} s^j (s^j)^\alpha &= (s-1)\sum_{j=0}^{n-1} (s^{\alpha+1})^j && \text{by laws of exponents} \\
&= (s-1)\frac{(s^{\alpha+1})^n - 1}{s^{\alpha+1} - 1} && \text{by (a)} \\
&= (b^{\alpha+1} - 1) \Big/ \frac{s^{\alpha+1} - 1}{s-1} && \text{rearranging, } s^n = b \\
&\xrightarrow{n} \frac{b^{\alpha+1} - 1}{\alpha + 1} && \text{because } f'_{\alpha+1}(1) = \alpha + 1.
\end{aligned}$$

(c) Identify the calculation from chapter 5 that, again for $b > 1$ and with $s^n = b$,

$$\begin{aligned}
\sum_{j=0}^{n-1} (j+1)s^j &= \sum_{j=0}^{n-1} \sum_{k=0}^{j} s^j && \text{because } j + 1 = \sum_{k=0}^{j} 1 \\
&= \sum_{k=0}^{n-1} \sum_{j=k}^{n-1} s^j && \text{reversing the order of summation} \\
&= \sum_{k=0}^{n-1} s^k \sum_{j=0}^{n-k-1} s^j && \text{rewriting the inner sum} \\
&= \sum_{k=0}^{n-1} s^k \frac{s^{n-k} - 1}{s - 1} && \text{by (a)} \\
&= \frac{1}{s-1} \sum_{k=0}^{n-1} (s^n - s^k) && \text{by algebra} \\
&= \frac{ns^n}{s-1} - \frac{s^n - 1}{(s-1)^2} && \text{adding } s^n \text{ to itself } n \text{ times and by (a)} \\
&= \frac{nb}{s-1} - \frac{b-1}{(s-1)^2} && \text{because } s^n = b.
\end{aligned}$$

(d) Identify the calculation also from chapter 5 that, again for $b > 1$ and with $s^n = b$,

$$
\begin{aligned}
(s-1)\sum_{j=0}^{n-1} s^j \ln(s^{j+1}) &= (s-1)\ln(s)\sum_{j=0}^{n-1}(j+1)s^j && \text{by a logarithm property}\\
&= (s-1)\ln(s)\left(\frac{nb}{s-1}-\frac{b-1}{(s-1)^2}\right) && \text{by (c)}\\
&= b\ln(b)-(b-1)\frac{\ln(s)-\ln(1)}{s-1} && n\ln(s)=\ln(s^n)=\ln(b)\\
&\xrightarrow{n} b\ln(b)-b+1 && \text{because } \ln'(1)=1.
\end{aligned}
$$

(e) Identify the calculation from chapter 6 that for $b > 0$,

$$
\begin{aligned}
\frac{b}{n}\sum_{j=0}^{n-1} e^{jb/n} &= \frac{b}{n}\cdot\frac{e^b-1}{e^{b/n}-1} && \text{by (a)}\\
&= (e^b-1)\Big/\frac{e^{b/n}-e^0}{b/n-0} && \text{rearranging}\\
&\xrightarrow{n} e^b-1 && \text{because } \exp'(0)=1.
\end{aligned}
$$

(f) Identify the calculation from chapter 7 that for $b > 0$,

$$
\begin{aligned}
\frac{b}{n}\sum_{j=0}^{n-1}(\cos(jb/n)+i\sin(jb/n)) &= \frac{b}{n}\sum_{j=0}^{n-1} e^{ijb/n} && \text{by definition of } e^{ijb/n}\\
&= \frac{b}{n}\cdot\frac{e^{ib}-1}{e^{ib/n}-1} && \text{by (a)}\\
&= -i(e^{ib}-1)\Big/\frac{e^{ib/n}-e^0}{ib/n-0} && \text{rearranging}\\
&\xrightarrow{n} -i(e^{ib}-1) && \cos'(0)=0,\ \sin'(0)=1\\
&= \sin(b)+i(1-\cos(b)) && \text{by definition of } e^{ib}.
\end{aligned}
$$

8.2 Preliminaries for the Pending Calculations

8.2.1 An Alternative Notation

Definition 8.2.1 (New Notation for the Integral). *If a function f is integrable from a to b, we write*

$$\int_{x=a}^{b} f(x)$$

as a synonym for

$$\int_a^b f.$$

Because $\int_a^b f$ is a number that does not depend in any way on the symbol x that is present in the new notation, that symbol can be replaced by any other symbol not already in use. Thus other synonyms for $\int_a^b f$ are, for instance,

$$\int_{t=a}^{b} f(t), \quad \int_{x_1=a}^{b} f(x_1), \quad \int_{\bullet=a}^{b} f(\bullet)\,.$$

The *variable of integration*, meaning the x or the t or the x_1 or the $\bullet$ here, is called a *dummy variable* because its name does not affect the value of the integral. It comes into existence temporarily as we calculate, only to disappear when the calculation is complete. While our new notation is less streamlined than the old, its advantage for the purposes of this chapter is that having the variable of integration appear explicitly will let us keep track of events as we integrate functions that are themselves integrals of other functions, which are integrals in turn, and so on. The new notation facilitates computing.

One particular formula will be useful in the new notation, so we establish it immediately. For any $\alpha \in \mathcal{R}$ define a function

$$g : \mathcal{R}_{>-1} \longrightarrow \mathcal{R}, \quad g(x) = f_\alpha(1+x) = (1+x)^\alpha.$$

The graph of g is the graph of the power function f_α translated one unit to the left. So for any $x \geq 0$, $\mathrm{Ar}_1^{1+x}(f_\alpha) = \mathrm{Ar}_0^x(g)$, and for any x such that $-1 < x < 0$, $-\mathrm{Ar}_{1+x}^1(f_\alpha) = -\mathrm{Ar}_x^0(g)$. That is in all cases,

$$\int_1^{1+x} f_\alpha = \int_0^x g, \quad x > -1.$$

Rewrite the right side in our new notation to get the useful formula

$$\int_1^{1+x} f_\alpha = \int_{x_1=0}^{x} (1+x_1)^\alpha, \quad x > -1. \tag{8.1}$$

One last comment (for now) about notation: The reader with prior background in calculus has almost certainly seen an even more adorned form of writing the integral than the one just introduced here, to wit,

$$\int_a^b f = \int_{x=a}^{b} f(x)\,dx.$$

We will bring the dx into our notation later, when it too will help keep track of certain calculations, but for this chapter it is unnecessary.

8.2.2 The Power Function Integral With Endpoint 0

For any positive real number $b \in \mathcal{R}_{>0}$ and for any nonnegative exponent $\alpha \in \mathcal{R}_{\geq 0}$, the region under the graph of the αth power function from 0 to b,

$$R = \{(x, y) \in \mathcal{R}^2 : 0 \le x \le b,\ 0 \le y \le f_\alpha(x)\},$$

is a bounded subset of the plane, and so it has an area. (The relevant fact in play here are that for $\alpha \ge 0$, the power function f_α extends continuously from $\mathcal{R}_{>0}$ to $\mathcal{R}_{\ge 0}$. Specificaly, $f_\alpha(0) = 0$ for $\alpha > 0$ while $f_0(0) = 1$.) Furthermore, because the power function is monotonic on $[0, b]$, the area is an integral,

$$\int_0^b f_\alpha \quad \text{exists and is} \quad \mathrm{Ar}_0^b(f_\alpha).$$

So to find the integral, we need only to find the area.

Because the power function is nonnegative on $\mathcal{R}_{\ge 0}$, we have for any number a such that $0 < a \le b$,

$$\mathrm{Ar}_a^b(f_\alpha) \le \mathrm{Ar}_0^b(f_\alpha).$$

On the other hand, a box having base $[0, a]$ and height a^α shows (see figure 8.1) that also.

$$\mathrm{Ar}_0^b(f_\alpha) \le a^{\alpha+1} + \mathrm{Ar}_a^b(f_\alpha).$$

That is, remembering the explicit formula for $\mathrm{Ar}_a^b(f_\alpha)$,

$$\frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1} \le \mathrm{Ar}_0^b(f_\alpha) \le a^{\alpha+1} + \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1}.$$

Now let a tend to 0. Because $\alpha \ge 0$, certainly $\alpha+1 > 0$, so that $\lim_{a\to 0} a^{\alpha+1} = 0$. (Here it is understood that a is tending to 0 from the positive side.) Consequently, by the Squeezing Rule and various other limit rules,

$$\mathrm{Ar}_0^b(f_\alpha) = \frac{b^{\alpha+1}}{\alpha+1}.$$

That is,

$$\int_0^b f_\alpha = \frac{b^{\alpha+1}}{\alpha+1}, \quad \alpha \ge 0,\ b \ge 0.$$

Here we have extended the formula to $b = 0$, when it simply says that $0 = 0$.

(Before continuing, we make a brief digression. Even though we assumed that $\alpha \in \mathcal{R}_{\ge 0}$, it is striking that proceeding from the inequalities

$$\frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1} \le \mathrm{Ar}_0^b(f_\alpha) \le a^{\alpha+1} + \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1}$$

to the conclusion that

$$\mathrm{Ar}_0^b(f_\alpha) = \frac{b^{\alpha+1}}{\alpha+1}$$

required only that $\alpha+1 > 0$, i.e., it required only that $\alpha > -1$. This suggests, for example (letting $\alpha = -1/2$), that even though the graph of the function

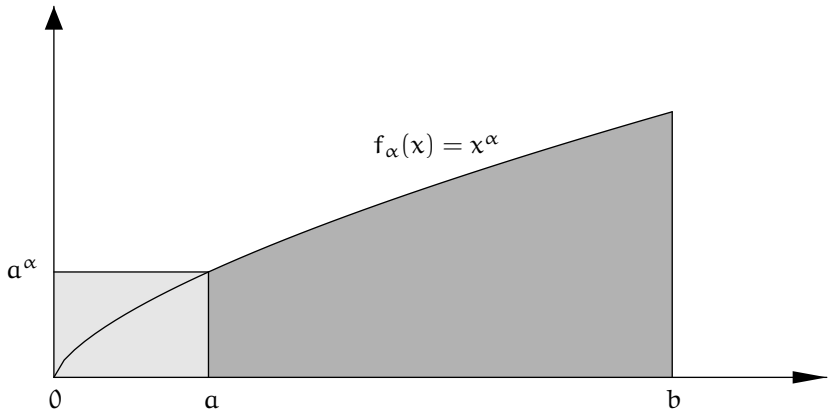


Figure 8.1. One more outer box

$$f_{-1/2} : \mathcal{R}_{>0} \longrightarrow \mathcal{R}, \quad f_{-1/2}(x) = 1/\sqrt{x}$$

has a vertical asymptote at $x = 0$, so that the region under the graph over the interval $(0, 1]$ is unbounded, apparently the very finite number 2 is a credible value for the region's area. The problem with this reasoning is that our underlying invocation that regions have areas has been made only for bounded regions. In our framework, the area of the unbounded region doesn't exist in the first place, and so the inequalities at the beginning of this paragraph are meaningless for negative values of α such as $\alpha = -1/2$. But this example encourages cautiously expanding the notion of area to apply to some unbounded regions. More specifically, the idea is that 2 is the area of the region under the graph of $f_{-1/2}$ over $(0, 1]$ in the sense that 2 is the least number that is at least as big as all areas of finite truncations of the region. The equation

$$\int_0^1 \frac{1}{\sqrt{x}} = 2$$

gives the value of an *improper integral*. We will not pursue this subject further.)

Specific examples of the power function integrals with left endpoint 0 are, using the new notation,

$$\int_{x_1=0}^{x} 1 = x, \qquad \int_{x_1=0}^{x} x_1 = \frac{x^2}{2}, \qquad \int_{x_1=0}^{x} x_1^2 = \frac{x^3}{3},$$

and in general,

$$\int_{x_1=0}^{x} x_1^n = \frac{x^{n+1}}{n+1}, \quad n \in \mathcal{Z}_{\geq 0},\ x \geq 0.$$

The previous display holds only for $x \geq 0$ so far, but because the exponent n is a nonnegative integer we want to extend it to negative x as well. Because $(-x_1)^n = (-1)^n x_1^n$, for negative x naturally

$$\int_{x_1=0}^{x} x_1^n = (-1)^{n+1} \int_{x_1=0}^{-x} x_1^n, \quad n \in \mathcal{Z}_{\geq 0},\ x < 0,$$

the extra power of -1 coming from the reversed direction of integration. Thus

$$\int_{x_1=0}^{x} x_1^n = (-1)^{n+1} \frac{(-x)^{n+1}}{n+1} = \frac{x^{n+1}}{n+1}, \quad n \in \mathcal{Z}_{\geq 0},\ x < 0.$$

And so predictably enough, the formula for $\int_0^x f_n$ is symbolically robust,

$$\int_{x_1=0}^{x} x_1^n = \frac{x^{n+1}}{n+1}, \quad n \in \mathcal{Z}_{\geq 0},\ x \in \mathcal{R}.$$

Exercise

8.2.1. Figure 8.1 shows a case where $\alpha > 0$. How would the figure change for $\alpha = 0$? Does this affect the argument in the text that $\int_0^b f_\alpha$ exists and equals $b^{\alpha+1}/(\alpha+1)$?

8.3 The Logarithm

By the definition of the logarithm and by the useful formula (8.1),

$$\ln(1+x) = \int_1^{1+x} f_{-1} = \int_{x_1=0}^{x} \frac{1}{1+x_1}, \quad x > -1.$$

Recall the finite geometric sum formula for any $n \in \mathcal{Z}_{\geq 0}$,

$$1 + r + r^2 + \cdots + r^{n-1} = \frac{1-r^n}{1-r}, \quad r \neq 1.$$

Rearrange the formula to get

$$\frac{1}{1-r} = 1 + r + r^2 + \cdots + r^{n-1} + \frac{r^n}{1-r}, \quad r \neq 1,$$

and then substitute $-x_1$ for r,

$$\frac{1}{1+x_1} = 1 - x_1 + x_1^2 - \cdots + (-1)^{n-1} x_1^{n-1} + (-1)^n \frac{x_1^n}{1+x_1}, \quad x_1 \neq -1. \tag{8.2}$$

Now integrate, letting x_1 vary from 0 to x, to obtain the logarithm as a polynomial and a remainder, using Proposition 5.5.2 (page 173),

$$\ln(1+x) = \int_{x_1=0}^{x} \frac{1}{1+x_1} = P_n(x) + R_n(x), \quad x > -1, \tag{8.3}$$

where $P_n(x)$ is obtained by integrating the polynomial in (8.2) term-by-term, carrying out power function integrals with left endpoint 0 (exercise 8.3.1(a)),

$$P_n(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n},$$

and the remainder is the integral of the rest of the right side of (8.2),

$$R_n(x) = (-1)^n \int_{x_1=0}^{x} \frac{x_1^n}{1+x_1}.$$

Equation (8.3) and sequence limit rules combine to show that for any x such that $x > -1$, the sequence of polynomials $(P_n(x))$ converges to $\ln(1+x)$ if and only if the sequence of remainders $(R_n(x))$ converges to 0 (exercise 8.3.1(b)). So the next question is how the remainder $R_n(x)$ behaves as n grows.

We address the question by estimating the remainder. The integral is being taken from $x_1 = 0$ to $x_1 = x$, where x is fixed through this discussion and now we stipulate that $-1 < x \le 1$. (The analysis will not work for $x > 1$.) If $0 \le x \le 1$ then throughout the integration process,

$$0 < \frac{1}{1+x_1} \le 1,$$

and so

$$|R_n(x)| \le \int_{x_1=0}^{x} x_1^n = \frac{x^{n+1}}{n+1}.$$

If $-1 < x < 0$ then throughout the integration,

$$0 < \frac{1}{1+x_1} \le \frac{1}{1+x},$$

and so

$$|R_n(x)| \le \frac{1}{1+x}\left|\int_{x_1=0}^{x} x_1^n\right| = \frac{1}{1+x}\left|\frac{x^{n+1}}{n+1}\right| = \frac{1}{1+x}\cdot\frac{|x|^{n+1}}{n+1}.$$

In both cases, letting C be the maximum of 1 and $1/(1+x)$,

$$|R_n(x)| \le C\frac{|x|^{n+1}}{n+1}.$$

Therefore by sequence limit rules (exercise 8.3.1(c)),

$$\lim(R_n(x)) = 0,$$

and we have shown that

$$\ln(1+x) = \lim_n (P_n(x)), \quad -1 < x \le 1.$$

Less formally,

$$\boxed{\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots, \quad -1 < x \le 1.}$$

And in Sigma-notation,

$$\boxed{\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n-1}\frac{x^n}{n}, \quad -1 < x \le 1.}$$

Note in particular the formula when $x = 1$,

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots.$$

(See exercise 8.3.2 for a geometric derivation of this formula.)

Beyond being pretty, the boxed formulas are shorthand for an algorithm to compute $\ln(1+x)$ (where $-1 < x \le 1$) to any desired accuracy by computing a polynomial, i.e., by carrying out finitely many additions and multiplications.. The idea is that given a desired accuracy, i.e., an error tolerance for our answer, the analysis that we just carried out lets us find a degree n so that $|R_n(x)|$ is smaller than the error tolerance. Thus $P_n(x)$ is as close to $\ln(1+x)$ as was desired.

Theorem 8.3.1 (Taylor Polynomial and Remainder for the Logarithm). *For any $x \in (-1, 1]$, and for any $n \in \mathcal{Z}_{\ge 0}$,*

$$\ln(1+x) = P_n(x) + R_n(x)$$

where

$$P_n(x) = \sum_{k=1}^{n} (-1)^{k-1}\frac{x^k}{k} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n}$$

and

$$|R_n(x)| \le \frac{|x|^{n+1}}{n+1} \max\left\{1, \ \frac{1}{1+x}\right\}.$$

Consequently, for any $x \in (-1, 1]$,

$$\ln(1+x) = \lim_n (P_n(x)) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^{n-1}\frac{x^n}{n} + \cdots.$$

For example, we use the theorem to estimate $\ln(1.1)$ by hand to within $1/500,000$. The nth degree polynomial approximation to $\ln(1.1)$ is

$$P_n(0.1) = (0.1) - \frac{(0.1)^2}{2} + \frac{(0.1)^3}{3} - \cdots + (-1)^{n-1}\frac{(0.1)^n}{n},$$

and the remainder satisfies

$$|R_n(0.1)| \le \frac{(0.1)^{n+1}}{(n+1)}\,.$$

The only symbolic variable is n, and the goal is to approximate $\ln(1.1)$ to within $1/500,000$. Set $n = 4$ in the previous display to get

$$|R_4(0.1)| \le \frac{1}{500,000}.$$

That is, the fourth degree Taylor polynomial

$$\begin{aligned} P_4(0.1) &= \frac{1}{10} - \frac{1}{200} + \frac{1}{3000} - \frac{1}{40000} \\ &= 0.10000000\cdots - 0.00500000\cdots + 0.00033333\cdots - 0.00002500\cdots \\ &= 0.09530833\cdots \end{aligned}$$

agrees with $\ln(1.1)$ to within $0.00000200\cdots$, so that

$$0.09530633\cdots \le \ln(1.1) \le 0.09531033\cdots\,.$$

Machine technology should confirm this.

The graphs of the natural logarithm and its first five Taylor polynomials are plotted from 0 to 2 in figure 8.2. (Here the functions are $\ln(x)$ and $P_n(x-1)$ so that the coordinate axes are in their usual position for the logarithm.) A good check of your understanding is to see if you can determine which graph is which in the figure (exercise 8.3.3).

Exercises

8.3.1. (a) Carry out the integration to obtain the polynomial $P_n(x)$ in the section.

(b) Explain how sequence limit rules show that $\lim_n(P_n(x)) = \ln(1+x)$ if and only if $\lim_n(R_n(x)) = 0$.

(c) Use sequence limit rules to show that $\lim_n(R_n(x)) = 0$ if $-1 < x \le 1$.

8.3.2. Figure 8.3 shows a partial geometric decomposition of $\ln(2)$.

(a) Identify the one box of width 1 in the figure and explain why its area is $1 - 1/2$.

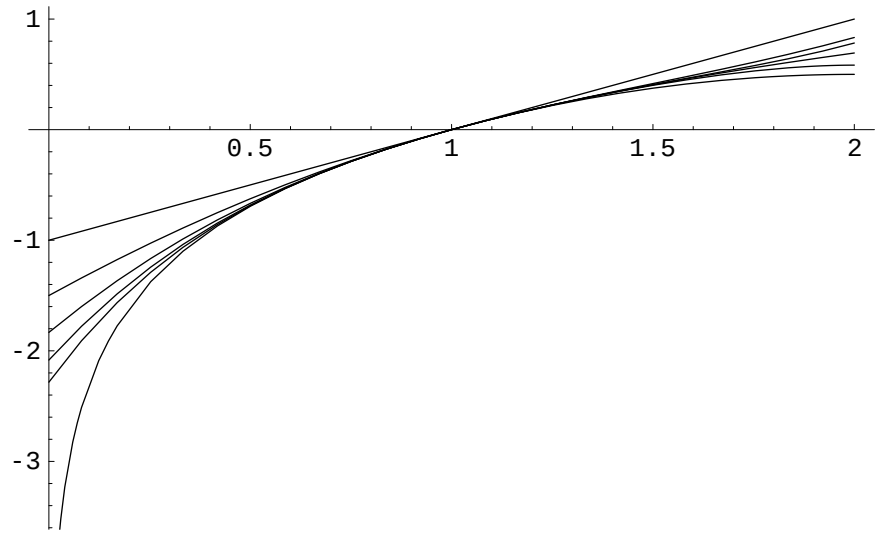


Figure 8.2. The natural logarithm and its Taylor polynomials

(b) Identify the one box of width $1/2$ in the figure and explain why its area is $1/3 - 1/4$.

(c) Identify the two boxes of width $1/4$ in the figure and explain why their areas are $1/5 - 1/6$ and $1/7 - 1/8$.

(d) Identify the four boxes of width $1/8$ in the figure and explain why their areas are $1/9 - 1/10$, $1/11 - 1/12$, $1/13 - 1/14$, and $1/15 - 1/16$.

(e) Identify the eight boxes of width $1/16$ in the figure and explain why the areas of the first two such boxes are $1/17 - 1/18$ and $1/19 - 1/20$.

8.3.3. In figure 8.2, identify the graphs of $P_1(x-1)$ through $P_5(x-1)$ and the graph of $\ln(x)$ near $x = 0$ and near $x = 2$.

8.4 The Exponential

8.4.1 A Precalculation

Define

$$I_0 : \mathcal{R} \longrightarrow \mathcal{R}, \quad I_0(x) = 1,$$

and then define for $n = 1, 2, 3, \ldots$

$$I_n : \mathcal{R} \longrightarrow \mathcal{R}, \quad I_n(x) = \int_0^x I_{n-1}.$$

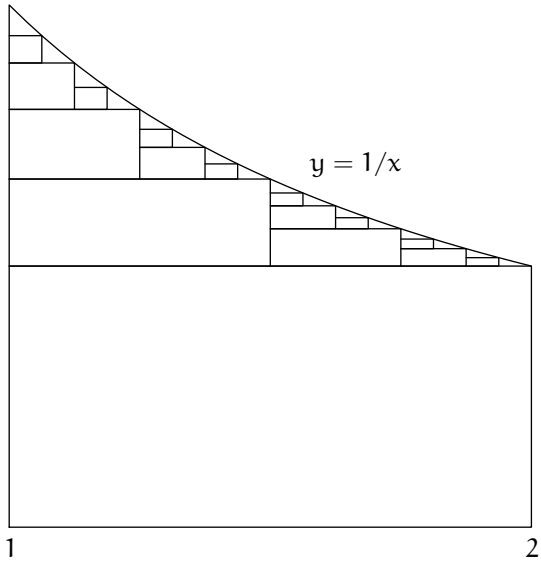


Figure 8.3. Partial geometric decomposition of $\ln(2)$

So, for example, $I_1(x) = \int_0^x 1 = x$. That is,

$$I_1 = f_1/1.$$

Consequently, $I_2(x) = \int_0^x f_1 = x^2/2$, or

$$I_2 = f_2/2.$$

Similarly, $I_3 = \int_0^x f_2/2 = x^3/(3 \cdot 2)$,

$$I_3 = f_3/3!.$$

Continuing in this vein shows that

$$I_n = f_n/n!, \quad n \in \mathcal{Z}_{\geq 0}.$$

(Recall that $0! = 1$ by convention, so that the formula for I_n covers the case $n = 0$.)

In our new notation, the calculations are written,

$$
\begin{aligned}
I_0(x) &= 1,\\
I_1(x) &= \int_{x_1=0}^{x} 1 = x,\\
I_2(x) &= \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} 1 = \int_{x_1=0}^{x} x_1 = \frac{x^2}{2},\\
I_3(x) &= \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2} 1 = \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} x_2 = \int_{x_1=0}^{x} \frac{x_1^2}{2} = \frac{x^3}{3\cdot 2},
\end{aligned}
$$

and in general,

$$
I_n(x) = \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\cdots\int_{x_n=0}^{x_{n-1}} 1 = \frac{x^n}{n!}, \quad n \in \mathcal{Z}_{\geq 0}.
$$

8.4.2 The Calculation

The method for expressing the exponential function as ever-higher degree polynomials plus the corresponding remainders is to use the basic identity

$$
\exp(x) = 1 + \int_0^x \exp, \quad x \in \mathcal{R}
$$

(a rearrangement of the formula $\int_0^x \exp = \exp(x) - \exp(0)$) over and over.

Fix any real number x for the duration of this discussion, and start from the basic identity, renotated,

$$
e^x = 1 + \int_{x_1=0}^{x} e^{x_1}.
$$

By the basic identity again, then Proposition 5.5.2, and then the precalculation,

$$
\begin{aligned}
e^x &= 1 + \int_{x_1=0}^{x} \left(1 + \int_{x_2=0}^{x_1} e^{x_2}\right)\\
&= 1 + \int_{x_1=0}^{x} 1 + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} e^{x_2}\\
&= 1 + x + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} e^{x_2}.
\end{aligned}
$$

Once more by the same process,

$$
\begin{aligned}
e^x &= 1 + x + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} \left(1 + \int_{x_3=0}^{x_2} e^{x_3}\right)\\
&= 1 + x + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} 1 + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2} e^{x_3}\\
&= 1 + x + \frac{x^2}{2!} + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2} e^{x_3}.
\end{aligned}
$$

Continuing this process through n iterations shows that

$$e^x = P_n(x) + R_n(x), \quad x \in \mathcal{R},$$

where $P_n(x)$ is an nth degree polynomial,

$$P_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!},$$

and $R_n(x)$ is the *remainder integral*,

$$R_n(x) = \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} e^{x_{n+1}}.$$

The task now is to analyze $R_n(x)$, because as with the logarithm, the sequence of polynomials $(P_n(x))$ converges to e^x if and only if $\lim_n(R_n(x)) = 0$.

If $x \geq 0$ then the integrand $e^{x_{n+1}}$ of $\mathcal{R}_n(x)$ lies between 1 and e^x. If $x < 0$ then the integrand lies between e^x and 1. In either case there is a constant C (specifically, C is the maximum of e^x and 1) such that all through the integration,

$$0 < e^{x_{n+1}} \leq C.$$

So if $x \geq 0$ then integrating this inequality $n+1$ times (see Proposition 3.3.14 on page 117) correspondingly bounds the remainder integral by a multiple of the precalculated integral,

$$0 \leq R_n(x) \leq C \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} 1 = C\frac{x^{n+1}}{(n+1)!}.$$

If $x < 0$ then all of the integrals in $R_n(x)$ have out-of-order endpoints, and

$$R_n(x) = (-1)^{n+1} \int_{x_1=x}^{0} \int_{x_2=x_1}^{0} \cdots \int_{x_{n+1}=x_n}^{0} e^{x_{n+1}},$$

and so

$$\begin{aligned}
|R_n(x)| &= \int_{x_1=x}^{0} \int_{x_2=x_1}^{0} \cdots \int_{x_{n+1}=x_n}^{0} e^{x_{n+1}} \\
&\leq C \int_{x_1=x}^{0} \int_{x_2=x_1}^{0} \cdots \int_{x_{n+1}=x_n}^{0} 1 \\
&= (-1)^{n+1} C \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} 1 \\
&= C\frac{(-x)^{n+1}}{(n+1)!}.
\end{aligned}$$

Combining the cases, we have shown that for all $x \in \mathcal{R}$,

$$|R_n(x)| \le C\frac{|x|^{n+1}}{(n+1)!}.$$

Because $|x|$ can be greater than 1, the quotient $|x|^{n+1}/(n+1)!$ can be a ratio of two large numbers, making its behavior initially unclear. But we can analyze it. The Archimedean property of the real number system says that there exists some positive integer m such that $m > 2|x|$. It follows that

$$\frac{|x|}{m+1} < \frac{1}{2}, \quad \frac{|x|}{m+2} < \frac{1}{2}, \quad \frac{|x|}{m+3} < \frac{1}{2},$$

and so on. Now consider the constant

$$K = \frac{|x|^m}{m!},$$

and note that (each line after the first in the following display making reference to the line before it)

$$\begin{aligned}
\frac{|x|^m}{m!} &= K,\\
\frac{|x|^{m+1}}{(m+1)!} &= \frac{|x|^m}{m!}\cdot\frac{|x|}{m+1} = K\cdot\frac{|x|}{m+1} \le K\cdot\frac{1}{2},\\
\frac{|x|^{m+2}}{(m+2)!} &= \frac{|x|^{m+1}}{(m+1)!}\cdot\frac{|x|}{m+2} \le K\cdot\frac{1}{2}\cdot\frac{1}{2} = K\cdot\left(\frac{1}{2}\right)^2,\\
\frac{|x|^{m+3}}{(m+3)!} &= \frac{|x|^{m+2}}{(m+2)!}\cdot\frac{|x|}{m+3} \le K\cdot\left(\frac{1}{2}\right)^2\cdot\frac{1}{2} = K\cdot\left(\frac{1}{2}\right)^3,
\end{aligned}$$

and in general,

$$\frac{|x|^{m+\ell}}{(m+\ell)!} \le K\cdot\left(\frac{1}{2}\right)^{\ell}, \quad \ell \in \mathcal{Z}_{\ge 0}.$$

The previous inequality rewrites as

$$\frac{|x|^{n+1}}{(n+1)!} \le K\cdot\left(\frac{1}{2}\right)^{n+1-m}, \quad n+1 \ge m.$$

And so,

$$|R_n(x)| \le CK2^{m-1}\cdot\left(\frac{1}{2}\right)^n, \quad n \ge m-1.$$

It follows that for any $x \in \mathcal{R}$,

$$\lim_n(R_n(x)) = 0,$$

and therefore

$$e^x = \lim(P_n(x)).$$

Less formally,

$$\boxed{e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots, \quad x \in \mathcal{R},}$$

or

$$\boxed{e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad x \in \mathcal{R}.}$$

Note in particular the formula when $x = 1$,

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} + \frac{1}{8!} + \cdots.$$

Theorem 8.4.1 (Taylor Polynomial and Remainder for the Exponential). *For any $x \in \mathcal{R}$, and for any $n \in \mathcal{Z}_{\geq 0}$,*

$$e^x = P_n(x) + R_n(x)$$

where

$$P_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$$

and (letting $\lceil x \rceil$ denote the smallest integer that is at least x)

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \max\left\{1,\ 3^{\lceil x \rceil}\right\}.$$

Consequently, for any $x \in \mathcal{R}$,

$$e^x = \lim_n(P_n(x)) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots.$$

Proof. Here the point is that we know

$$|R_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!} \max\{1,\ e^x\},$$

but because we don't know e^x, we loosen the bound a little more to get a bound that we can compute. □

Exercise

8.4.1. Estimate the accuracy to which $P_{10}(1)$ approximates e.

8.5 The Cosine and the Sine

Fix any real number x for the duration of this discussion. Recall that

$$\cos(x) + i\sin(x) = e^{ix}.$$

Thus the two identities

$$\cos(x) = 1 - \int_{x_1=0}^{x} \sin(x_1), \qquad \sin(x) = \int_{x_1=0}^{x} \cos(x_1)$$

encode more concisely as (using the fact that $-1 = i^2$)

$$e^{ix} = 1 + i\int_{x_1=0}^{x} e^{ix_1}.$$

It follows that

$$\begin{aligned} e^{ix} &= 1 + i\int_{x_1=0}^{x} \left(1 + i\int_{x_2=0}^{x_1} e^{ix_2}\right) \\ &= 1 + i\int_{x_1=0}^{x} 1 - \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} e^{ix_2} \\ &= 1 + ix - \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} e^{ix_2}. \end{aligned}$$

Again by the same process,

$$\begin{aligned} e^{ix} &= 1 + ix - \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} \left(1 + i\int_{x_3=0}^{x_2} e^{ix_3}\right) \\ &= 1 + ix - \int_{x_1=0}^{x}\int_{x_2=0}^{x_1} 1 - i\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2} e^{ix_3} \\ &= 1 + ix - \frac{x^2}{2} - i\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2} e^{ix_3}. \end{aligned}$$

And after one more iteration,

$$e^{ix} = 1 + ix - \frac{x^2}{2} - i\frac{x^3}{3!} + \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2}\int_{x_4=0}^{x_3} e^{ix_4}.$$

Continuing through n iterations shows that

$$e^{ix} = P_n(x) + R_n(x)$$

where $P_n(x)$ is an nth degree polynomial,

$$P_n(x) = 1 + ix + i^2\frac{x^2}{2} + i^3\frac{x^2}{3!} + \cdots + i^n\frac{x^n}{n!},$$

and $R_n(x)$ is the remainder integral,

$$R_n(x) = i^{n+1} \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} e^{ix_{n+1}}.$$

Now take the sum out to $2n+1$ rather than out to n, decompose the result into its real and imaginary parts, and revise the notation P_n to get for the cosine,

$$P_n(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} = \sum_{k=0}^{n} (-1)^k \frac{x^{2k}}{(2k)!}$$

and

$$R_n(x) = \pm \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{2n+2}=0}^{x_{2n+1}} \cos(x_{2n+2}),$$

and to get for the sine,

$$P_n(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

and because the x^{2n+2} term of the sine expansion is 0, in fact

$$R_n(x) = \pm \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{2n+3}=0}^{x_{2n+2}} \cos(x_{2n+3}).$$

Because the cosine function is bounded in absolute value by 1, this gives

Theorem 8.5.1 (Taylor Polynomial and Remainder for the Cosine and Sine). *For any $x \in \mathcal{R}$, and for any $n \in \mathcal{Z}_{\geq 0}$,*

$$\cos(x) = P_n(x) + R_n(x),$$

where

$$P_n(x) = \sum_{k=0}^{n} (-1)^k \frac{x^{2k}}{(2k)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!}$$

and

$$|R_n(x)| \leq \frac{|x|^{2n+2}}{(2n+2)!}.$$

Consequently, for any $x \in \mathcal{R}$,

$$\cos(x) = \lim_n (P_n(x)) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + \cdots.$$

For any $x \in \mathcal{R}$, and for any $n \in \mathcal{Z}_{\geq 0}$,

$$\sin(x) = P_n(x) + R_n(x),$$

where

$$P_n(x) = \sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!}$$

and

$$|R_n(x)| \le \frac{|x|^{2n+3}}{(2n+3)!}.$$

Consequently, for any $x \in \mathcal{R}$,

$$\sin(x) = \lim_n (P_n(x)) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \cdots.$$

Here the polynomial P_n for the cosine has degree $2n$ or $2n+1$ rather than n, and the polynomial P_n for the sine has degree $2n+1$ or $2n+2$. The reader who wants polynomial subscripts to match polynomial degrees is welcome to renotate them P_{2n} or P_{2n+1}, and P_{2n+1} or P_{2n+2}, and to renotate the remainders correspondingly as well.

Exercises

8.5.1. Explain how Theorem 8.5.1 shows that for positive x the approximation $\sin(x) \approx x$ is accurate to within $x^3/6$. It is understood that x is in radians rather than degrees. Switching now to degrees, and using a calculator to convert radians to degrees, show that the approximation $\sin x \approx x$ holds within 0.0005 for x up to 8 degrees, but not if x is 9 degrees.

8.5.2. For any value $s \in (0, \pi/2)$, let $a(s)$ be the area of the shaded region in the left half of figure 8.4, and let $b(s)$ be the area of the shaded region in the right half of the figure.

(a) Explain why $a(s) = (1/2)(1 - \cos(s))\sin(s)$. Approximate $1 - \cos(s)$ by a polynomial in s that includes all powers of s no higher than s^3. Do the same for $\sin(s)$. Use these two polynomials to do the same for $a(s)$.

(b) Explain why $b(s) = (1/2)s - (1/2)\cos(s)\sin(s)$. Approximate $\cos(s)$ and $\sin(s)$ by polynomials in s that include all powers of s no higher than s^3, and then do the same for $b(s)$.

(c) Evaluate

$$\lim_{s \to 0} \frac{a(s)}{b(s)}.$$

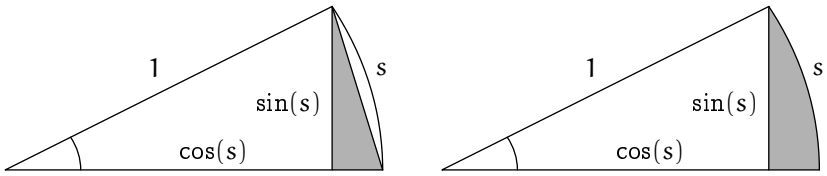


Figure 8.4. Two areas

8.6 The Power Function

8.6.1 The Polynomial and the Remainder

The method for expressing the power function as ever-higher degree polynomials plus the corresponding remainders is essentially the same as for the other functions in this chapter. (And a fairly general theory is little different from our specific examples.) This time, begin by recalling formula (2.7) from the end of chapter 2,

$$\int_a^b f_\alpha = \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha+1} \qquad \alpha \in \mathcal{R},\ \alpha \neq -1,\ 0 < a \leq b.$$

(This formula is valid for irrational exponents α as well as rational exponents α, although we haven't shown so.) The usual verification confirms that in fact the formula holds if a and b are out of order as well. Specialize to $a = 1$ and $b = 1 + x$ where $x > -1$, and replace α by $\alpha - 1$,

$$\int_1^{1+x} f_{\alpha-1} = \frac{(1+x)^\alpha - 1}{\alpha}, \quad \alpha \in \mathcal{R},\ \alpha \neq 0,\ x > -1.$$

Rewrite the formula as follows, switching to our new notation and citing the useful formula (8.1) (page 239),

$$(1+x)^\alpha = 1 + \alpha \int_{x_1=0}^{x} (1+x_1)^{\alpha-1}, \quad \alpha \in \mathcal{R},\ x > -1. \tag{8.4}$$

(For $\alpha = 0$, this formula doesn't follow from the previous one, but in this case it simply says that $1 = 1$.)

Fix any real number $x > -1$. Let $\alpha \in \mathcal{R}$ be any real number. By (8.4) once,

$$(1+x)^\alpha = 1 + \alpha \int_{x_1=0}^{x} (1+x_1)^{\alpha-1}.$$

By (8.4) again, and by the results cited in the calculation for the exponential function, the integral is

$$\begin{aligned}\int_{x_1=0}^{x}(1+x_1)^{\alpha-1} &= \int_{x_1=0}^{x}\left(1+(\alpha-1)\int_{x_2=0}^{x_1}(1+x_2)^{\alpha-2}\right)\\ &= \int_{x_1=0}^{x}1+(\alpha-1)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}(1+x_2)^{\alpha-2}\\ &= x+(\alpha-1)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}(1+x_2)^{\alpha-2},\end{aligned}$$

so that

$$(1+x)^{\alpha} = 1+\alpha x+\alpha(\alpha-1)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}(1+x_2)^{\alpha-2}.$$

Similarly, the double integral is

$$\begin{aligned}\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}(1+x_2)^{\alpha-2} &= \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\left(1+(\alpha-2)\int_{x_3=0}^{x_2}(1+x_3)^{\alpha-3}\right)\\ &= \int_{x_1=0}^{x}\int_{x_2=0}^{x_1}1+(\alpha-2)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2}(1+x_3)^{\alpha-3}\\ &= \frac{x^2}{2!}+(\alpha-2)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2}(1+x_3)^{\alpha-3},\end{aligned}$$

so that now

$$(1+x)^{\alpha} = 1+\alpha x+\frac{\alpha(\alpha-1)}{2!}x^2+\alpha(\alpha-1)(\alpha-2)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\int_{x_3=0}^{x_2}(1+x_3)^{\alpha-3}.$$

Reintroduce the binomial coefficient notation, but where now α can be any real number,

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}, \quad k\in\mathcal{Z}_{\geq 0}.$$

Unless α is a nonnegative integer, the binomial coefficient $\binom{\alpha}{k}$ is nonzero for all $k\in\mathcal{Z}_{\geq 0}$, and as soon as k exceeds α, $\binom{\alpha}{k}$ changes sign with each increment of k. After n iterations of the process, we get

$$(1+x)^{\alpha} = P_n(x)+R_n(x), \quad x>-1,$$

where similarly to the Finite Binomial Theorem,

$$P_n(x) = 1+\binom{\alpha}{1}x+\binom{\alpha}{2}x^2+\cdots+\binom{\alpha}{n}x^n$$

and the remainder is

$$R_n(x) = \alpha(\alpha-1)\cdots(\alpha-n)\int_{x_1=0}^{x}\int_{x_2=0}^{x_1}\cdots\int_{x_{n+1}=0}^{x_n}(1+x_{n+1})^{\alpha-n-1}.$$

We have derived the formula up to $n = 2$, and exercise 8.6.1 is to obtain it for $n = 3$. As usual, the sequence of polynomials $(P_n(x))$ converges to $(1+x)^\alpha$ if and only if the sequence of remainders $(R_n(x))$ has limit 0. If $\alpha \in \mathcal{Z}_{\geq 0}$ then the formulas for $P_n(x)$ and $R_n(x)$ show that $P_n(x) = P_\alpha(x)$ and $R_n(x) = 0$ for all $n \geq \alpha$. This situation is covered by the Finite Binomial Theorem. The situation where $\alpha \in \mathcal{R}$ but $\alpha \notin \mathcal{Z}_{\geq 0}$ will be discussed next.

Exercises

8.6.1. Carry out one more step of the process in the section to get from the formula $(1+x)^\alpha = P_2(x) + R_2(x)$ to the formula $(1+x)^\alpha = P_3(x) + R_3(x)$.

8.6.2. Recall that

$$\ln(1+x) = \int_{x_1=0}^{x} (1+x)^{-1}, \quad x > -1.$$

Also, specializing the work just done to the case $\alpha = -1$ gives an equality

$$(1+x_1)^{-1} = P_{n-1}(x_1) + R_{n-1}(x_1), \quad x > -1, \ n \geq 1.$$

Integrate the equality to re-obtain the degree-n polynomial approximation of $\ln(1+x)$ (i.e., the $P_n(x)$ for $\ln(1+x)$ is $\int_{x_1=0}^{x} P_{n-1}(x)$, integrating the $P_{n-1}(x_1)$ for $(1+x_1)^{-1}$), and to obtain an expression for the remainder that is valid for $x > -1$ rather than only for $|x| < 1$. (To make the notation work smoothly, write the $R_{n-1}(x_1)$ for $(1+x_1)^{-1}$ with outermost variable of integration x_2.)

8.6.2 The Infinite Binomial Theorem

Now take $\alpha \in \mathcal{R}$ but $\alpha \notin \mathcal{Z}_{\geq 0}$. That is, α is a real number other than a nonnegative integer. To review, we have the formula

$$(1+x)^\alpha = P_n(x) + R_n(x), \quad x > -1,$$

where $P_n(x)$ is the nth degree polynomial

$$P_n(x) = 1 + \binom{\alpha}{1}x + \binom{\alpha}{2}x^2 + \cdots + \binom{\alpha}{n}x^n,$$

and

$$R_n(x) = \binom{\alpha}{n+1}(n+1)! \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} (1+x_{n+1})^{\alpha-n-1},$$

so that

$$R_n(x) = \left| \binom{\alpha}{n+1} (n+1)! \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} (1+x_{n+1})^{\alpha-n-1} \right|.$$

The task is to analyze $R_n(x)$ as n grows large. The analysis here is more complicated than those earlier in this chapter, and so the reader should engage with it to taste.

Because n is growing large, we assume that $n > \alpha - 1$ (such n exists by the Archimedean property of the real number system), so that the exponent $\alpha - n - 1$ in the remainder integral is negative.

If $x \geq 0$ then the integrand $(1+x_{n+1})^{\alpha-n-1}$ lies between $(1+x)^{\alpha-n-1}$ and 1. If $x < 0$ then the integrand lies between 1 and $(1+x)^{\alpha-n-1}$. In either case there is a constant C (the maximum of 1 and $(1+x)^{\alpha-n-1}$) such that all through the integration,

$$0 < (1+x_{n+1})^{\alpha-n-1} \leq C.$$

So if $x \geq 0$ then integrating this inequality $n+1$ times bounds the remainder integral by a multiple of the precalculated integral,

$$\begin{aligned} |R_n(x)| &\leq C \left| \binom{\alpha}{n+1} \right| (n+1)! \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} 1 \\ &= C \left| \binom{\alpha}{n+1} \right| x^{n+1}. \end{aligned}$$

If $x < 0$ then all of the integrals in $R_n(x)$ have out-of-order endpoints, and so

$$\begin{aligned} |R_n(x)| &= \left| \binom{\alpha}{n+1} \right| (n+1)! \int_{x_1=x}^{0} \int_{x_2=x_1}^{0} \cdots \int_{x_{n+1}=x_n}^{0} (1+x_{n+1})^{\alpha-n-1} \\ &\leq C \left| \binom{\alpha}{n+1} \right| (n+1)! \int_{x_1=x}^{0} \int_{x_2=x_1}^{0} \cdots \int_{x_{n+1}=x_n}^{0} 1 \\ &= C \left| \binom{\alpha}{n+1} \right| (n+1)!(-1)^{n+1} \int_{x_1=0}^{x} \int_{x_2=0}^{x_1} \cdots \int_{x_{n+1}=0}^{x_n} 1 \\ &= C \left| \binom{\alpha}{n+1} \right| (-x)^{n+1}. \end{aligned}$$

Combining the cases, we have shown that for all $x > -1$,

$$|R_n(x)| \leq B_n(x), \quad \text{where } B_n(x) = C \left| \binom{\alpha}{n+1} \right| |x|^{n+1}.$$

Now impose the stronger condition that $|x| < 1$. Then more specifically,

$$|x| = 1 - 2\delta, \tag{8.5}$$

where $\delta = (1-|x|)/2 > 0$. Consider the sequence (whose origin will be explained in a moment)

$$(s_n) = \left(\left|\frac{\alpha - n}{n+1}\right|\right).$$

By sequence limit rules and the fact that the absolute value function is continuous,

$$\lim_n (s_n) = 1, \tag{8.6}$$

and so, because $1/(1-\delta) > 1$, there exists a starting index N such that

$$0 < s_n < \frac{1}{1-\delta} \quad \text{for all } n \geq N. \tag{8.7}$$

Now consider the ratio of the estimates of consecutive-generation remainders, the generation being comfortably large,

$$\frac{B_n(x)}{B_{n-1}(x)} = \frac{C\left|\binom{\alpha}{n+1}\right| |x|^{n+1}}{C\left|\binom{\alpha}{n}\right| |x|^n} = \left|\frac{\alpha - n}{n+1}\right| |x| = s_n|x|.$$

Thus the ratio is what gives rise to s_n. Let

$$r = \frac{1-2\delta}{1-\delta},$$

so that $0 \leq r < 1$. By (8.5) and (8.7),

$$\frac{B_n(x)}{B_{n-1}(x)} < r \quad \text{for all } n \geq N.$$

That is (starting at $n = N+1$ rather than at $n = N$ just to be tidy, and each line after the first in the following display making reference to the line before it),

$$\begin{aligned} B_{N+1}(x) &\leq rB_N(x), \\ B_{N+2}(x) &\leq rB_{N+1}(x) \leq r^2 B_N(x), \\ B_{N+3}(x) &\leq rB_{N+2}(x) \leq r^3 B_N(x), \end{aligned}$$

and in general,

$$B_{N+\ell}(x) \leq r^\ell B_N(x) \quad \text{for all } \ell \in \mathcal{Z}_{\geq 0}.$$

By sequence limit rules it follows that for any x such that $|x| < 1$,

$$\lim_n (B_n(x)) = 0,$$

and therefore that

$$\lim_n (R_n(x)) = 0.$$

Recall that

$$(1+x)^\alpha = P_n(x) + R_n(x), \quad |x| < 1,\ n \in \mathcal{Z}_{\geq 0}.$$

Because $\lim_n(R_n(x)) = 0$, it follows that $\lim(P_n(x)) = (1+x)^\alpha$. Less formally,

$$\boxed{(1+x)^\alpha = 1 + \binom{\alpha}{1}x + \binom{\alpha}{2}x^2 + \cdots + \binom{\alpha}{n}x^n + \cdots, \quad |x| < 1,\ \alpha \in \mathcal{R}.}$$

In Sigma-notation,

$$\boxed{(1+x)^\alpha = \sum_{n=0}^{\infty}\binom{\alpha}{n}x^n, \quad |x| < 1,\ \alpha \in \mathcal{R}.}$$

Theorem 8.6.1 (Taylor Polynomial and Remainder for the Power Function: the Binomial Theorem). *For any $x \in (-1,1)$, for any $\alpha \in \mathcal{R}$, and for any $n \in \mathcal{Z}_{\geq 0}$ such that $n > \alpha - 1$,*

$$(1+x)^\alpha = P_n(x) + R_n(x)$$

where

$$P_n(x) = \sum_{k=0}^{n}\binom{\alpha}{k}x^k = 1 + \binom{\alpha}{1}x + \binom{\alpha}{n}2x^2 + \cdots + \binom{\alpha}{n}x^n$$

and

$$|R_n(x)| \leq \left|\binom{\alpha}{n+1}\right| |x|^{n+1} \max\left\{1,\ (1+x)^{\alpha-n-1}\right\}.$$

Consequently, for any $x \in (-1,1)$ and for any $\alpha \in \mathcal{R}$,

$$(1+x)^\alpha = \lim_n(P_n(x)) = 1 + \binom{\alpha}{1}x + \binom{\alpha}{n}2x^2 + \cdots + \binom{\alpha}{n}x^n + \cdots.$$

The graphs of the square root and its first six Taylor polynomials are plotted from 0 to 4 in figure 8.5. (Here the functions are $\sqrt{x}$ and $P_n(x-1)$ so that the coordinate axes are in their usual position.) The figure shows that the polynomials are approximating the square root well from 0 to 2, but not for x beyond 2.

Exercises

8.6.3. Let α be a real number.

(a) Let $(\tilde{s}_n) = ((\alpha - n)/(n+1))$. Show that $\lim_n(\tilde{s}_n) = -1$. Explain why (8.6) follows.

(b) Confirm the calculation in the section that $B_n(x)/B_{n-1}(x) = s_n|x|$.

(c) Use sequence limit rules to explain why $\lim_n(B_n(x)) = 0$.

(d) Use sequence limit rules to explain why $\lim_n(R_n(x)) = 0$.

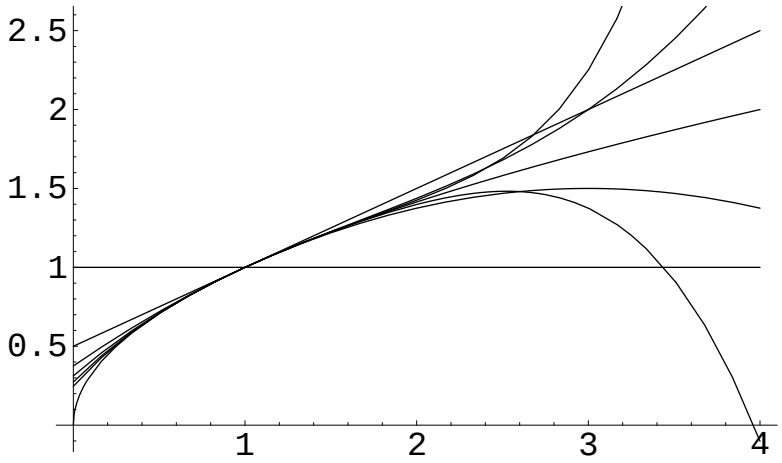


Figure 8.5. The square root and its Taylor polynomials

8.6.4. Use a second degree polynomial to approximate $\sqrt{1.2}$. Find a guaranteed accuracy of the approximation and thus find upper and lower bounds for $\sqrt{1.2}$.

8.6.5. (For the symbolically-inclined.) The identity

$$(1+x)^{\alpha} = e^{\alpha \ln(1+x)}, \quad |x| < 1, \ \alpha \in \mathcal{R}$$

now has a formal interpretation as an equality involving three infinite sums: on the left side of the equality, a sum of coefficients times powers of x; and on the right side of the equality, a second sum of coefficients times powers of a third sum. Does the formal interpretation seem to gives the same expression on both sides of the equality?

8.6.6. The remainder analysis for the logarithm and the power function required $|x| < 1$ (or $x = 1$ for the logarithm) to show that $\lim_n(R_n(x)) = 0$, but the analysis for the exponential function, the cosine, and the sine did not. Why is this? That is, what aspect of the analysis was different enough among the various functions to produce two pairs of different-flavored results?

8.7 Summary

The calculations that we carried out for various functions in this chapter have much in common. Once we have the Fundamental Theorem of Calculus, an exercise (exercise 10.1.7) will show how to describe all of the calcululations in one general encoding.

9

Theory and Applications of the Derivative

The derivative is a *local* construct, computed by zooming in on the behavior of a function about a single point. This chapter discusses how despite being local, the derivative can help as answer large-scale questions.

One such question, the problem of optimizing a function—making it as large or as small as possible—is discussed in section 9.1. Here the basic idea is intuitive geometrically: where the function is optimized, its graph should have a horizontal tangent, i.e., its derivative should be zero. Section 9.2 introduces the Mean Value Theorem and its consequences. This theorem gives a formula relating a function and its derivative, with no reference to the fact that the derivative is a limit. With the formula in hand, plausible statements about functions (e.g., the fact that if the derivative is positive, the function is strictly increasing) become easy to prove. Section 9.3 shows how to use calculus to sketch the graphs of functions. Computer graphing technology is effective and readily available nowadays, but nonetheless sometimes calculus can inform us about features of a graph that are not shown clearly by the machine-generated figure. Section 9.4 discusses problems of the following form: given two related quantities, and given the rate of change of one of them, what is the rate of change of the other? Such problems are known as related rates problems.

9.1 Optimization

To *optimize* a function is to make it as big as possible or as small as possible by suitably specifying its input-value. The following definition gives us language to discuss this subject.

Definition 9.1.1 (Minimum, Maximum, Extremum, Local Minimum, Local Maximum, Local Extremum). *Let* A *be a subset of* $\mathcal{R}$*, and let* $f : A \longrightarrow \mathcal{R}$ *be a function. Then*

- *A* **minimum of** f *is a function-value* $f(x)$ *such that* $f(x) \leq f(s)$ *for all* $s \in A$. *The function* f **has a minimum at** x *if* $f(x)$ *is a minimum of* f.
- *A* **maximum of** f *is a function-value* $f(x)$ *such that* $f(x) \geq f(s)$ *for all* $s \in A$. *The function* f **has a maximum at** x *if* $f(x)$ *is a maximum of* f.
- *An* **extremum of** f *is a minimum or a maximum of* f. *The function* f **has an extremum at** x *if* $f(x)$ *is an extremum of* f.
- *A* **local minimum of** f *is a function-value* $f(x)$ *such that* $f(x) \leq f(s)$ *for all* $s \in A$ *within some positive distance of* x. *The function* f **has a local minimum at** x *if* $f(x)$ *is a local minimum of* f.
- *A* **local maximum of** f *is a function-value* $f(x)$ *such that* $f(x) \geq f(s)$ *for all* $s \in A$ *within some positive distance of* x. *The function* f **has a local maximum at** x *if* $f(x)$ *is a local maximum of* f.
- *A* **local extremum of** f *is a local minimum or a local maximum of* f. *The function* f **has a local extremum at** x *if* $f(x)$ *is a local extremum of* f.

A minimum of a function is a local minimum but not necessarily conversely, and similarly for maximum and extremum. A function can have at most one minimum and at most one maximum (although the function can takes its minimum at many different inputs, and similarly for the maximum), but it can have many local minima and local maxima. Exercise 9.1.1 asks for examples of these phenomena. According to the definition, if f is constant, then its value is both a minimum and a maximum, and it has a minimum and a maximum at every input-value. To escape this counterintuitive linguistic circumstance, one can further define a **strict minimum** of f by changing "$f(x) \leq f(s)$ for all $s \in A$ within some positive distance of x" to "$f(x) < f(s)$ for all $s \in A$ within some positive distance of x except $s = x$" in the definition, and so on. (In general, an inequality is called *strict* if it precludes equality.)

9.1.1 The Extreme Value Theorem

For an arbitrary function f, there is no reason to believe that extrema exist to be found. But for a certain class of functions, optimization is guaranteed, at least in the abstract.

Theorem 9.1.2 (Extreme Value Theorem). *Let* f *be a continuous function whose domain is a closed, bounded interval,*

$$f : [a, b] \longrightarrow \mathcal{R}, \quad f \text{ continuous.}$$

Then f *assumes a maximum value and a minimum value.*

Like the Intermediate Value Theorem (page 185), the Extreme Value Theorem is an abstract existence theorem. That is, its conclusion is not that (for example) "$f(x) \leq f(s)$ for all $s \in [a, b]$," which in isolation would be meaningless because the hypotheses make no mention of any particular point x. Rather the conclusion is that *there exists some* x such that $f(x) \leq f(s)$ for all $s \in [a, b]$. But the theorem says nothing about where x is or how to find it.

The Extreme Value Theorem is not a calculus theorem: it makes no reference to derivatives, integrals, or infinite sums. Especially, *the function* f *in the theorem is not assumed to be differentiable*. On the other hand, the theorem does make use of limits, because limits are at the heart of continuity. Also, the theorem depends on a property of the real number system, a property equivalent—after some work—to the Completeness Property that we have invoked throughout these notes. Because the issues that arise in proving Theorem 9.1.2 are foundational, the proof is beyond our scope. This paragraph applies verbatim to the Intermediate Value Theorem as well. Although the Intermediate Value Theorem and the Extreme Value Theorem are separated by many pages in these notes, they are very similar in their roles as theorems that support the foundations of calculus, and they should be viewed as a pair.

In our context, what is more important anyway than proving the Extreme Value Theorem is that one can gain intuition about its content by convincing oneself via examples that its conclusion can fail unless all of the hypotheses are met. That is,

- A discontinuous function whose domain is a closed, bounded interval need not assume a maximum value or a minimum value.
- A continuous function whose domain is a closed but unbounded interval need not assume a maximum value or a minimum value.
- A continuous function whose domain is a bounded but non-closed interval need not assume a maximum value or a minimum value.
- A continuous function whose domain is a non-closed, unbounded interval need not assume a maximum value or a minimum value.

Exercise 9.1.2 asks for examples.

Exercises

9.1.1. Illustrate graphically examples of the following phenomena:

(a) A local minimum of a function need not be a minimum of the function.

(b) A function can take its maximum at many inputs.

(c) A function can have many local minima, all distinct.

(d) A function can have many local maxima, each taken at many inputs.

9.1.2. (a) Find a function $f : [0,1] \longrightarrow \mathcal{R}$ that assumes no maximum value and no minimum value. (Give a clear sketch of the graph of such an f, or a formula for such an f, with some explanation in either case.)

(b) Find a continuous function $f : [0,\infty) \longrightarrow \mathcal{R}$ that assumes no minimum value and no maximum value.

(c) Find a continuous function $f : (0,1] \longrightarrow \mathcal{R}$ that assumes no maximum value.

(d) Find a continuous function $f : (0,1] \longrightarrow \mathcal{R}$ that assumes no minimum value and no maximum value.

9.1.3. (a) Explain why for any continuous function $f : [a,b] \longrightarrow \mathcal{R}$, we may take the codomain of f to be some interval $[L,M]$ instead.

(b) May we take the range of f to be $[L,M]$?

9.1.2 Conditions for Optimization

Under suitable conditions, the conditions under which a function is locally optimized are very constrained.

Theorem 9.1.3 (Critical Point Theorem). *Consider a function on a closed, bounded interval,*

$$f : [a,b] \longrightarrow \mathcal{R}.$$

Suppose that f *has a local extremum at the point* $x \in [a,b]$*, i.e.,* $f(x)$ *is a local extremum of* f*. Then at least one of the following conditions holds:*

(a) f *is not differentiable at* x.
(b) x *is an endpoint of* $[a,b]$*, i.e.,* $x = a$ *or* $x = b$.
(c) f *is differentiable at* x *and* $f'(x) = 0$.

For an example of conditions (a) and (b), the absolute value function

$$|\ \ | : [-1,1] \longrightarrow \mathcal{R}$$

assumes its minimum at $x = 0$, where it is not differentiable, and the function assumes its maximum at the endpoints $x = -1$ and $x = 1$.

Corollary 9.1.4. *Let* $f : [a,b] \longrightarrow \mathcal{R}$ *be continuous, and let* f *be differentiable on* (a,b)*. The only possible points where its extrema can occur are* a*,* b*, and* $x \in (a,b)$ *such that* $f'(x) = 0$*.*

Proof (of the theorem). Suppose that $f(x)$ is a local maximum. Suppose also that f is differentiable at x, and that x is not an endpoint, i.e., that conditions (a) and (b) do not hold. We want to show that consequently condition (c) holds, i.e., that $f'(x) = 0$.

Because x is not the endpoint b of $[a, b]$, there exist values $s \in [a, b]$ such that $s > x$. For all such s close enough to x, we have $f(s) - f(x) \leq 0$ because $f(x)$ is a local maximum, and we have $s - x > 0$ because $s > x$, and so the quotient $(f(s) - f(x))/(s - x)$ is zero or negative. Thus, letting s approach x from the right,

$$f'(x) = \lim_{s \to x^+} \frac{f(s) - f(x)}{s - x} \leq 0.$$

Similarly letting s approach x from the left shows that $f'(x) \geq 0$. Thus $f'(x) = 0$. The case where f has a local minimum at x can be treated similarly, or by using the auxiliary function $g = -f$, which has a local maximum at x and has derivative $g' = -f'$. □

Example 9.1.5. Consider the function

$$f : [-2, 2] \longrightarrow \mathcal{R}, \quad f(x) = x^3 - 3x.$$

This function is differentiable with derivative

$$f' : [-2, 2] \longrightarrow \mathcal{R}, \quad f'(x) = 3x^2 - 3.$$

The values x such that $f'(x) = 0$ are therefore $x = -1$ and $x = 1$, and so any possible local extrema of f occur at these points or at the endpoints $x = -2$ and $x = 2$. Compute that

$$f(-2) = -2, \qquad f(-1) = 2, \qquad f(1) = -2, \qquad f(2) = 2.$$

Because f assumes a minimum by the Extreme Value Theorem, the minimum is $f(-2) = f(1) = -2$. Similarly, the maximum of f is $f(-1) = f(2) = 2$.

Example 9.1.6. Consider the function

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = \frac{1}{x^2 + 1}.$$

Then f is always positive. The denominator of $f(x)$ is smallest at $x = 0$, so that the maximum of f is $f(0) = 1$. The denominator of $f(x)$ grows without bound as $|x|$ grows without bound, so that the positive value $f(x)$ grows ever closer to 0 as $|x|$ grows without bound, and f has no minimum.

Exercises

9.1.4. Sketch the graph of a generic function $f : [a, b] \longrightarrow \mathcal{R}$. Then sketch the graph of the related function $g(s) = f(-s)$.

9.1.5. Find all local extrema of the following functions. For each local extremum, state whether it is also global.

(a) $f : [-2, 2] \longrightarrow \mathcal{R}$, $f(x) = x^4 - x^2$.

(b) $f : [-2, 2] \longrightarrow \mathcal{R}$, $f(x) = 4x^3 - 3x^4$.

9.1.6. Consider the function

$$f : [1, \infty) \longrightarrow \mathcal{R}, \quad f(x) = \ln(x)/x.$$

Note that $f(1) = 0$, that $f(x) \geq 0$ for all $x \in [1, \infty)$, and that $\lim_{x\to\infty} f(x) = 0$ (see page 155).

(a) Does the Extreme Value Theorem guarantee that f has a maximum?

(b) Show that the unique $x \in [1, \infty)$ such that $f'(x) = 0$ is $x = e$.

(c) Because $\lim_{x\to\infty} f(x) = 0$, there exists some value $b \geq 1$ such that $f(x) \leq f(e)/2$ for all $x \geq b$. Explain why f has a maximum on $[1, b]$, and why this maximum is also the maximum of f on $[1, \infty)$.

(d) Which is larger, $\ln(e)/e$ or $\ln(\pi)/\pi$? Explain. Consequently, which is larger, $\pi \ln(e)$ or $e \ln(\pi)$? $\ln(e^\pi)$ or $\ln(\pi^e)$? Which is larger, e^π or π^e?

9.1.3 Optimization Story-Problems

To solve an optimization story-problem, proceed as follows.

- Draw and label a figure.
- Write an equation for the quantity to optimize. If possible, express the quantity in terms of a single independent variable. Be aware of the domain of values for the variable.
- Typically the domain is an interval. If the interval is closed and bounded, evaluate the quantity at the endpoints. Otherwise analyze the quantity near the missing endpoints, or as the variable gets large or small.
- Find the values of the variable for which the derivative of the quantity is zero, and evaluate the quantity at each such value.
- Evaluate the quantity at points where its derivative fails to exist.

In practice, one sometimes gets a little casual with this procedure.

Example 9.1.7. *Optimize the product x_1x_2 of two nonnegative numbers that sum to 1.*

Experimentation suggests strongly that the answer is $x_1 = x_2 = 1/2$. For example,

$$\frac{1}{10}\cdot\frac{9}{10} = \frac{9}{100}, \quad \frac{2}{10}\cdot\frac{8}{10} = \frac{16}{100}, \quad \frac{3}{10}\cdot\frac{7}{10} = \frac{21}{100}, \quad \frac{4}{10}\cdot\frac{6}{10} = \frac{24}{100}, \quad \frac{5}{10}\cdot\frac{5}{10} = \frac{25}{100}.$$

To phrase the problem in terms of one variable, we want to optimize the function

$$f(x) = x(1-x), \quad x \in [0,1].$$

Because the domain of f is a closed, bounded interval, and because f is continuous, f assumes a maximum and a minimum. The endpoint value $f(0) = f(1) = 0$ is the minimum because f is nonnegative. We can find the maximum in various ways:

- Use algebra: Complete the square,

  $$f(x) = -x^2 + x - 1/4 + 1/4 = 1/4 - (x-1/2)^2,$$

  or symmetrize to get the same formula,

  $$f(x) = (1/2 + (x-1/2))(1/2 - (x-1/2)) = 1/4 - (x-1/2)^2,$$

  so because $(x-1/2)^2$ is nonnegative, f takes its maximum at $x = 1/2$, and the maximum is $1/4$.
- Use geometry: The graph of $f(x) = x(1-x) = x - x^2$ is a downward-opening parabola that passes through the x-axis at $x = 0$ and at $x = 1$, so its highest point is halfway between them horizontally, making its x-coordinate $1/2$. This highest point has height $1/2(1-1/2) = 1/4$.
- Use calculus: Because $f(x) = x - x^2$,

  $$f'(x) = 1 - 2x,$$

  and so $f'(x) = 0$ if and only if $x = 1/2$. By Corollary 9.1.4, the maximum of f therefore occurs at $x = 1/2$, and it is $f(1/2) = 1/4$.

Example 9.1.8. *Optimize the volume of a cylinder that sits inside a sphere.*

The situation is depicted in profile in figure 9.1. Let the cylinder have base-radius r and half-height a. Then its volume is

$$V = \pi r^2 \cdot 2a.$$

But because the cylinder fits in a sphere, which may as well be the unit sphere, we also have

$$a^2 + r^2 = 1.$$

The extreme cases of the geometry are cylinders that degenerate either to a line segment between the spherical poles or an equatorial disk, both having volume zero. And so, as a function of a, the volume is

$$f(a) = \pi(1-a^2) \cdot 2a = 2\pi(a - a^3), \quad a \in [0,1].$$

Because f is differentiable and its domain is a closed, bounded interval, it assumes a minimum and a maximum. Because the endpoint value $f(0) =$

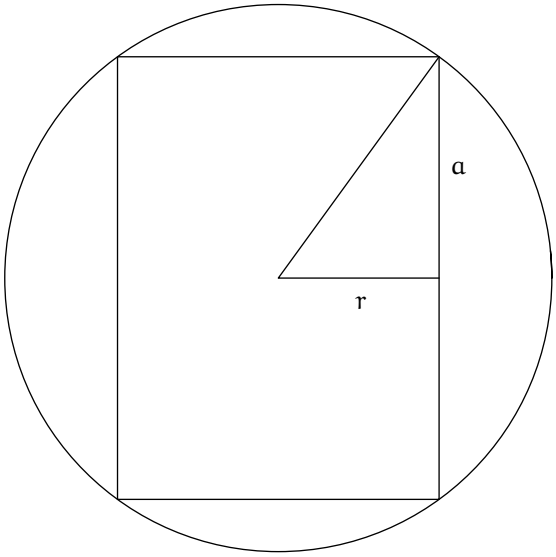


Figure 9.1. Cylinder inside sphere, seen in profile

$f(1) = 0$ is the minimum, the maximum value must be assumed somewhere where $f' = 0$. Compute that

$$f'(a) = 2\pi(1 - 3a^2),$$

and so the maximum occurs when $a^2 = 1/3$. Recall that $a^2 + r^2 = 1$, so that also $r^2 = 2/3$ for the maximum volume. That is, the proportions for the maximum volume satisfy $2a^2 = r^2$, or

$$r = \sqrt{2}\, a.$$

The cylinder's height is $h = 2a$, so the answer rewrites as

$$h = \sqrt{2}\, r.$$

Because this answer is phrased in terms of proportions, it does not depend on normalizing the sphere-radius to 1.

Example 9.1.9. *Optimize the surface-area of a cylinder having given volume.*

We may normalize the volume and phrase the answer in terms of proportions. Let the cylinder have radius r and height h. Then its surface area, encompassing the base, the top, and the side, is

$$A = 2\pi r^2 + 2\pi rh = 2\pi(r^2 + rh),$$

while the volume is

$$V = \pi r^2 h.$$

Normalize the volume to π, so that $r^2h = 1$ and thus $rh = 1/r$. Then the surface area is

$$A(r) = 2\pi(r^2 + 1/r).$$

Note that A is very large for small positive r (a very tall, thin cylinder) and for large positive r (a very wide, squat cylinder). To make A small, set its derivative to 0,

$$A'(r) = 2\pi(2r - 1/r^2) = 2\pi(2r^3 - 1)/r^2.$$

Thus $A'(r) = 0$ when $r^3 = 1/2$. But because $r^2h = 1$, consequently $r^3h = r$, or $r^3 = r/h$. And so the proportions of the optimal cylinder are

$$\frac{r}{h} = \frac{1}{2},$$

or $h = 2r$. The cylinder has diameter $d = 2r$, so in fact the proportions are

$$h = d.$$

That is, the cylinder sits tightly inside a cube.

Example 9.1.10. *Build the biggest configuration of* m *by* n *rectangular pens using a given amount of fencing, consisting of* m *rows of pens, having total length* x*, by* n *columns of pens, having total length* y*.*

Normalize to one unit of fencing, so that $(m+1)x + (n+1)y = 1$. The quantity that we want to optimize is

$$f(x) = xy = x \cdot \frac{1-(m+1)x}{n+1} = \frac{x-(m+1)x^2}{n+1},$$

with $0 \le x \le \frac{1}{m+1}$. Note that $f(0) = f(\frac{1}{m+1}) = 0$ while f is positive for the rest of its inputs. So, to maximize f, set its derivative to 0,

$$0 = f'(x) = \frac{1-2(m+1)x}{n+1},$$

giving $x = \frac{1}{2(m+1)}$. (Because $f(x)$ is quadratic, the optimizing value of x being the midpoint of the domain can be obtained in other ways, as in the first example above.) Consequently $y = \frac{1}{2(n+1)}$ and the largest enclosed area is $\frac{1}{4(m+1)(n+1)}$. The optimizing amounts of east–west and north–south fencing are equal, $x = \frac{m+1}{2(m+1)} = \frac{1}{2}$ and similarly $y = \frac{1}{2}$.

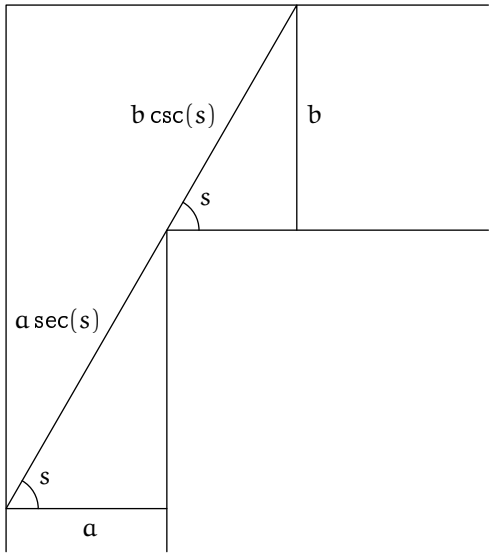


Figure 9.2. Moving a toothpick around a corner

Example 9.1.11. *An ant wants to move a toothpick around a 90-degree corner between a tunnel of width* a *and a tunnel of width* b*. Both tunnels are horizontal and have negligible height. How long a toothpick can fit around the corner?*

The situation is depicted in figure 9.2. For any angle $s \in (0, \pi/2)$, the longest toothpick that can fit into the corner at angle s has length

$$f(s) = a\sec(s) + b\csc(s).$$

The smallest value of f is length of the longest toothpick that will fit all the way around the corner.

The domain of f is a bounded interval, but it is missing both of its endpoints. However, note that $f(s)$ is very large for s slightly greater than 0 and for s slightly less than $\pi/2$, and this is consonant with our geometric intuition that the toothpick's fit is tightest somewhere in the middle of the process of getting it around the corner. So, because f is differentiable, we consider its derivative,

$$f'(s) = a\tan(s)\sec(s) - b\cot(s)\csc(s) = \frac{a\sin^3(s) - b\cos^3(s)}{\sin^2(s)\cos^2(s)}.$$

This derivative vanishes for

$$s = \arctan\left((b/a)^{1/3}\right),$$

and for this value of s we have

$$f(s) = a^{2/3}\sqrt{a^{2/3}+b^{2/3}} + b^{2/3}\sqrt{a^{2/3}+b^{2/3}} = \left(a^{2/3}+b^{2/3}\right)^{3/2}.$$

For example, if $a = 8$ and $b = 27$ then a toothpick of length

$$(8^{2/3}+27^{2/3})^{3/2} = 13\sqrt{13} \approx 47$$

will fit around the corner.

Example 9.1.12. *The bottom of a statue is* h *units higher than the viewer's eye. The top of the statue is* H *units higher than the viewer's eye. How far back should the viewer stand to maximize the vertical angle that she perceives the statue to fill?*

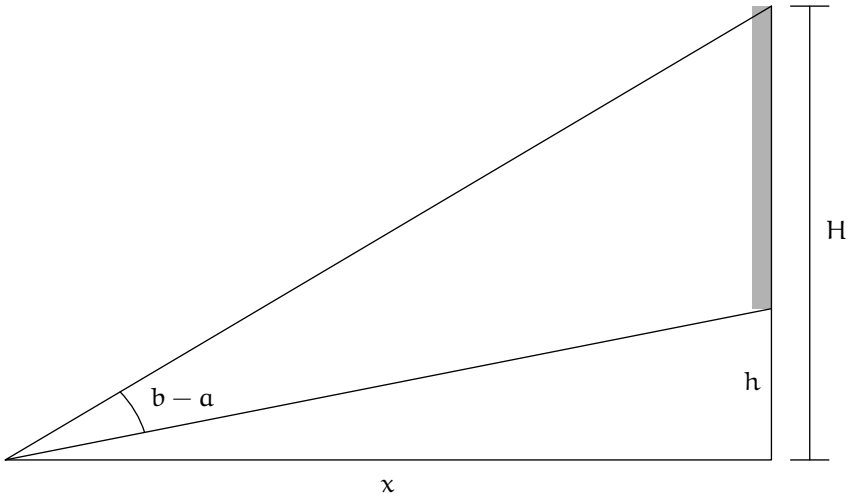


Figure 9.3. Viewing a statue

The situation is depicted in figure 9.3. Let the horizontal distance from the viewer's eye to the statue be x, a positive number. Then the angle in question is $b - a$ where $\tan(a) = h/x$ and $\tan(b) = H/x$. Thus we want to maximize the function

$$f(x) = \arctan(H/x) - \arctan(h/x), \quad x > 0.$$

For small positive x, $\arctan(H/x)$ and $\arctan(h/x)$ are close to $\pi/2$, so $f(x)$ is close to 0, and for large positive x, $\arctan(H/x)$ and $\arctan(h/x)$ are close to 0,

so again $f(x)$ is close to 0. Thus we search for x-values such that $f'(x) = 0$. Compute

$$\begin{aligned} f'(x) &= \frac{-H/x^2}{1+H^2/x^2} - \frac{-h/x^2}{1+h^2/x^2} = \frac{-H}{x^2+H^2} + \frac{h}{x^2+h^2} \\ &= \frac{-H(x^2+h^2)+h(x^2+H^2)}{(x^2+H^2)(x^2+h^2)}. \end{aligned}$$

The fraction vanishes exactly when its numerator vanishes, and its numerator is

$$-H(x^2+h^2)+h(x^2+H^2) = (h-H)x^2 - hH(h-H) = (x^2-hH)(h-H).$$

This vanishes only for $x = \sqrt{hH}$, and so we have the optimal distance. This distance is the *geometric mean* of h and H, meaning their multiplicative average, as compared to their *arithmetic mean* $(h+H)/2$. Incidentally, the relation $x^2 = hH$ shows that in figure 9.3 the two right triangles having side x are similar when x is optimal. Because the optimal horizontal distance is $\sqrt{hH}$, the optimal perceived vertical angle is consequently (invoking the identity $\arctan b - \arctan a = \arctan(\frac{b-a}{1+ab})$)

$$\arctan(H/\sqrt{hH}) - \arctan(h/\sqrt{hH}) = \arctan\left(\frac{H-h}{2\sqrt{hH}}\right).$$

Especially if $H = 2h$, which is to say that the statue is as high as its base, as for example is true for the Statue of Liberty, then the optimal perceived viewing angle is $\arctan\left(\frac{1}{2\sqrt{2}}\right) \approx 19.47°$. Similarly if $H = 3h$, which is to say that the statue is twice as high as its base, then the optimal perceived viewing angle is $30°$.

Example 9.1.13. *A particle travels through medium 1 at speed v and through medium 2 at speed w. If the particle travels from point A to point B (see figure 9.4) in the least possible amount of time, what is the relation between angles α and β?*

If, for example, v is greater than w, then one argument is that the particle should travel in medium 1 (where it moves faster) to the point on the boundary between the media just above point B, and then drop straight down to B, thus spending as little time as possible traveling slowly. But this strategy entails taking a long path from A to B. A second argument is that the particle should take the shortest path from A to B, the line segment joining them, regardless of the fact that in doing so it will traverse a longer path in medium 2, where it moves slowly. The correct answer will lie somewhere between the answers suggested by these two arguments.

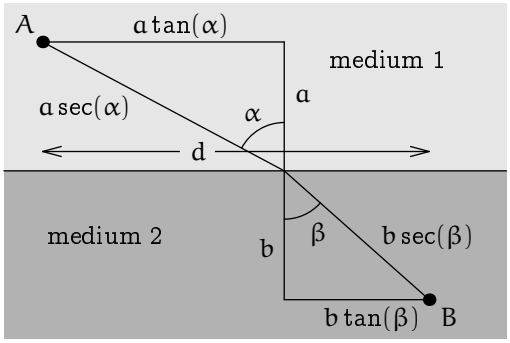


Figure 9.4. Geometry of Snell's Law

Because time is distance over speed, a little trigonometry shows that this problem is to minimize the time

$$t = \frac{a}{v}\sec(\alpha) + \frac{b}{w}\sec(\beta),$$

where the lateral distance traveled is

$$a\tan(\alpha) + b\tan(\beta) = d.$$

View α as the independent variable and β as a function of α. Differentiate the expression for the lateral distance to get

$$a\sec^2(\alpha) + b\sec^2(\beta)\cdot\beta' = 0.$$

Now differentiate t to get

$$t' = \frac{a}{v}\tan(\alpha)\sec(\alpha) + \frac{b}{w}\tan(\beta)\sec(\beta)\cdot\beta',$$

which rewrites as

$$t' = \frac{a}{v}\sin(\alpha)\sec^2(\alpha) + \frac{b}{w}\sin(\beta)\sec^2(\beta)\cdot\beta'.$$

But from the derivative of the lateral distance, $b\sec^2(\beta)\cdot\beta' = -a\sec^2(\alpha)$, and so

$$\begin{aligned} t' &= \frac{a}{v}\sin(\alpha)\sec^2(\alpha) - \frac{a}{w}\sin(\beta)\sec^2(\alpha) \\ &= a\sec^2(\alpha)\left(\frac{\sin(\alpha)}{v} - \frac{\sin(\beta)}{w}\right). \end{aligned}$$

That is, $t' = 0$ exactly when $\sin(\alpha)/v = \sin(\beta)/w$, or

$$\frac{\sin(\alpha)}{\sin(\beta)} = \frac{v}{w}.$$

This relation is called *Snell's Law*.

A second optimization method can sometimes solve problems more quickly and easily. Explaining why this second method works requires ideas from two-variable calculus, beyond the scope of these notes. Even so, we present it as a procedure and then revisit problems from earlier in this section to show it in action. The procedure is as follows.

To optimize a function f *of two variables, subject to a constraint* $g = c$ *where* g *is a second function of two variables, proceed as follows.*

1. *Compute four quantities*
   - A*: the derivative of* f *with respect to the first of its two variables,*
   - B*: the derivative of* f *with respect to the second of its two variables,*
   - C*: the derivative of* g *with respect to the first of its two variables,*
   - D*: the derivative of* g *with respect to the second of its two variables.*
2. *Study the relation* $AD = BC$ *and the constraint* $g = c$ *together, and see what they tell us.*

The product problem revisited. Recall the problem:

> *Find the biggest value of the product* x_1x_2 *where* x_1 *and* x_2 *are two nonnegative numbers that sum to* 1.

Here we have $f(x_1, x_2) = x_1x_2$ and $g(x_1, x_2) = 1$ where $g(x_1, x_2) = x_1 + x_2$. Thus

$$A = x_2, \quad B = x_1, \quad C = 1, \quad D = 1,$$

and the relation $AD = BC$ is

$$x_2 = x_1.$$

Because also $x_1 + x_2 = 1$, the solution obviously is $x_1 = x_2 = 1/2$ with product $1/4$, as above. This same method shows that the biggest value of the product $x_1^{\alpha_1}x_2^{\alpha_2}$ where x_1 and x_2 are two nonnegative numbers that sum to 1, and α_1 and α_2 are positive exponents, occurs at $x_1 = \alpha_1/(\alpha_1 + \alpha_2)$ and $x_2 = \alpha_2/(\alpha_1 + \alpha_2)$ and is $\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}/(\alpha_1 + \alpha_2)^{\alpha_1+\alpha_2}$. The reader can check that for $\alpha_1 = \alpha_2 = 1$ this solution matches the particular case that we have solved.

The cylinder problem. Recall the problem:

> *Find the largest cylinder that fits in a sphere.*

(See figure 9.1 on page 270.) Here we have $f(r,h) = \pi r^2 h$ and $g(r,h) = 1$ where $g(r,h) = r^2 + (h/2)^2$. Thus

$$A = 2\pi rh, \quad B = \pi r^2, \quad C = 2r, \quad D = h/2,$$

and the relation $AD = BC$ is

$$\pi rh^2 = 2\pi r^3$$

Cancel πr to get $h^2 = 2r^2$, or

$$h = \sqrt{2}\, r.$$

As above, these are the proportions of the biggest cylinder that fits in a sphere.

Another cylinder problem. Recall the problem:

> *Find the cylinder of least area having fixed volume.*

Here we have $f(r,h) = 2\pi r^2 + 2\pi rh$ and $g(r,h) = 1$ where $g(r,h) = \pi r^2 h$. Thus

$$A = 4\pi r + 2\pi h = 2\pi(2r + h), \quad B = 2\pi r, \quad C = 2\pi rh, \quad D = \pi r^2,$$

and the relation $AD = BC$ is

$$2\pi^2 r^2 (2r + h) = 4\pi^2 r^2 h,$$

or, after canceling $2\pi^2 r^2$,

$$2r + h = 2h,$$

or

$$2r = h.$$

So the height is the diameter, as above.

The fence problem. Recall the problem:

> *Build the biggest configuration of m by n rectangular pens using a given amount of fencing.*

Here we have $f(x,y) = xy$ and $g(x,y) = 1$ where $g(x,y) = (m+1)x + (n+1)y$. Thus

$$A = y, \quad B = x, \quad C = m + 1, \quad D = n + 1,$$

and the relation $AD = BC$ is

$$(n + 1)y = (m + 1)x.$$

This relation says to allocate half of one's fencing to east–west and the other half to north–south, as we know from above to be the solution.

The toothpick problem. Recall the problem:

An ant wants to move a toothpick around a 90-degree corner between a tunnel of width a and a tunnel of width b. Both tunnels are horizontal and have negligible height. How long a toothpick can fit around the corner?

(See figure 9.2 on page 272.) Let u be the portion of the toothpick in the width-a tunnel and let v be the portion of the toothpick in the width-b tunnel. When the toothpick is at angle θ we have $\cos(\theta) = a/u$ and $\sin(\theta) = b/v$. Thus the function to be optimized is $f(u,v) = u + v$ and the constraint condition is $g(u,v) = 1$ where $g(u,v) = a^2/u^2 + b^2/v^2$. We have

$$A = 1, \quad B = 1, \quad C = -2\frac{a^2}{u^3}, \quad D = -2\frac{b^2}{v^3},$$

and so the condition $AD = BC$ is $a^2/u^3 = b^2/v^3$. This condition raised to the power $2/3$ is $a^{4/3}/u^2 = b^{4/3}/v^2$, and this along with the constraint are

$$a^{-2/3}\frac{a^2}{u^2} = b^{-2/3}\frac{b^2}{v^2}, \qquad \frac{a^2}{u^2} + \frac{b^2}{v^2} = 1.$$

Use the first of these to substitute $b^{2/3}/a^{2/3} \cdot a^2/u^2$ for b^2/v^2 in the second,

$$\frac{a^{2/3} + b^{2/3}}{a^{2/3}}\frac{a^2}{u^2} = 1,$$

from which $u^2 = a^{4/3}(a^{2/3} + b^{2/3})$ and thus

$$u = a^{2/3}(a^{2/3} + b^{2/3})^{1/2}.$$

Symmetrically $v = b^{2/3}(a^{2/3}+b^{2/3})^{1/2}$, and so the optimal value is as above,

$$u + v = (a^{2/3} + b^{2/3})^{3/2}.$$

The statue problem. Recall the problem:

The bottom of a statue is h units higher than the viewer's eye. The top of the statue is H units higher than the viewer's eye. How far back should the viewer stand to maximize the vertical angle that she perceives the statue to fill?

(See figure 9.3 on page 273.) We optimize $f(a,b) = b - a$ subject to the constraint $g(a,b) = 0$ where $g(a,b) = (\tan a)/h - (\tan b)/H$, each of whose terms is $1/x$ where x is the horizontal distance in the figure. Remembering that $\tan' = \sec^2 = \tan^2 + 1$,

$$A = -1, \quad B = 1, \quad C = (\tan^2 a + 1)/h, \quad D = -(\tan^2 b + 1)/H.$$

Because $C = h(\tan^2 a)/h^2 + 1/h = h/x^2 + H/(hH)$ and similarly $D = -H/x^2 - h/(hH)$, the relation $AD = BC$ is

$$H/x^2 + h/(hH) = h/x^2 + H/(hH).$$

So $(H-h)/x^2 = (H-h)/(hH)$, giving $x^2 = hH$ and thus $x = \sqrt{hH}$ as above.

In fact the statue problem can be solved without calculus. To minimize $b-a$ we may minimize $\tan(b-a)$ instead, which by a trigonometry identity (exercise 9.1.14) and then the geometry of the problem and then some algebra is

$$\tan(b-a) = \frac{\tan(b)-\tan(a)}{1+\tan(b)\tan(a)} = \frac{H/x - h/x}{1+hH/x^2} = \frac{H-h}{x+hH/x}.$$

This is largest when its denominator $x + hH/x$ is smallest. And

$$x + \frac{hH}{x} = \sqrt{hH}\left(\frac{x}{\sqrt{hH}} + \frac{\sqrt{hH}}{x}\right).$$

In general for positive y we have $y + 1/y = (\sqrt{y} - 1/\sqrt{y})^2 + 2 \geq 2$ with equality if and only if $y = 1$. The previous display has $y = x/\sqrt{hH}$, so it is smallest when $x = \sqrt{hH}$, the same answer as twice before.

The light ray problem. Recall the problem:

> *Find the quickest path for a light ray to travel from point* A *at vertical distance* a *up in the air to point* B *at vertical depth* b *down in the water. The light ray travels at speed* v *in the air and at speed* w *in the water.*

(See figure 9.4 on page 275.) Here the related quantities are the angle of incidence α and the angle of refraction β. The function to be optimized is $f(\alpha,\beta) = (a/v)\sec(\alpha) + (b/w)\sec(\beta)$, and the constraining relation is $g(\alpha,\beta) = 1$ where $g(\alpha,\beta) = a\tan(\alpha) + b\tan(\beta)$. Thus

$$\begin{aligned} A &= (a/v)\tan(\alpha)\sec(\alpha), & B &= (b/w)\tan(\beta)\sec(\beta) \\ C &= a\sec^2(\alpha), & D &= b\sec^2(\beta), \end{aligned}$$

and because $\tan = \sin\cdot\sec$, these are

$$\begin{aligned} A &= (a/v)\sin(\alpha)\sec^2(\alpha), & B &= (b/w)\sin(\beta)\sec^2(\beta) \\ C &= a\sec^2(\alpha), & D &= b\sec^2(\beta). \end{aligned}$$

The relation $AD = BC$ is

$$(ab/v)\sin(\alpha)\sec^2(\alpha)\sec^2(\beta) = (ab/w)\sin(\beta)\sec^2(\beta)\sec^2(\alpha),$$

which cancels down to

$$\frac{\sin(\alpha)}{v} = \frac{\sin(\beta)}{w},$$

and this is Snell's law as above,

$$\frac{\sin(\alpha)}{\sin(\beta)} = \frac{v}{w}.$$

Exercises

9.1.7. Let α_1 and α_2 be positive real numbers. Optimize the weighted product $x_1^{\alpha_1} x_2^{\alpha_2}$ of two nonnegative numbers x_1 and x_2 that sum to 1.

9.1.8. Optimize the volume of a cone that sits inside a sphere. (Let r be the radius of the cone's circular base, and let h be the height from the cone's base to its vertex. Your answer should describe the proportions of the cone.)

9.1.9. Optimize the volume of a cylinder that sits inside a cone.

9.1.10. Rotate a right triangle of a given hypotenuse to form a cone of greatest volume.

9.1.11. Optimize the volume of a box created by cutting four small squares away from the corners of a large square and then folding up the resulting flaps.

9.1.12. Find the point(s) on the parabola $y = x^2$ that are nearest to the point $(0, 9/2)$.

9.1.13. Optimize the geometric mean $\sqrt{hH}$ where h and H are nonnegative numbers whose arithmetic mean is 1.

9.1.14. Use the identities $\sin(b - a) = \sin(b)\cos(a) - \cos(b)\sin(a)$ and $\cos(b-a) = \cos(b)\cos(a)+\sin(b)\sin(a)$ to show that $\tan(b-a) = (\tan(b) - \tan(a))/(1 + \tan(b)\tan(a))$.

9.2 The Mean Value Theorem and Its Consequences

9.2.1 Statement of the Theorem

Theorem 9.2.1 (Rolle's Theorem). *Suppose that a function*

$$f : [a, b] \longrightarrow \mathcal{R}$$

is continuous on $[a, b]$ *and differentiable on* (a, b)*, and suppose further that* $f(a) = f(b) = 0$*. Then there exists some value* $c \in (a, b)$ *such that* $f'(c) = 0$.

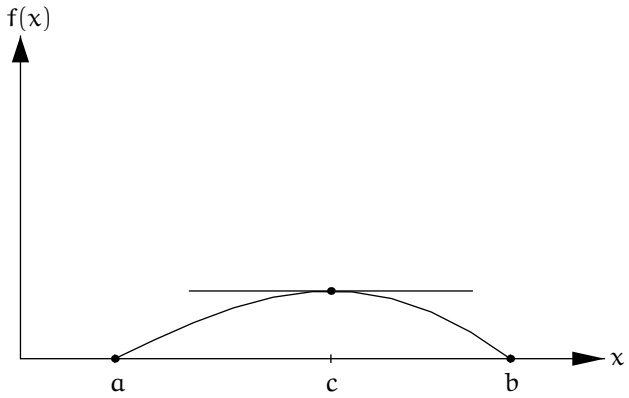


Figure 9.5. Rolle's Theorem

Rolle's Theorem is illustrated in figure 9.5.

Proof. If f is identically 0 then so is f', and any $c \in (a, b)$ will do.

Otherwise, note that f has a minimum and a maximum by the Extreme Value Theorem. Either the minimum or the maximum is nonzero, and so it occurs at a nonendpoint $c \in (a, b)$, where f is differentiable. By Corollary 9.1.4, $f'(c) = 0$. □

The proof of Rolle's Theorem relies on the Extreme Value Theorem, which we have not proved. Like the Extreme Value Theorem, Rolle's Theorem is an existence theorem: its conclusion is not that "$f'(c) = 0$," which in isolation would be meaningless because the hypotheses make no mention of a point c, but that *there exists some* c such that $f'(c) = 0$.

Theorem 9.2.2 (Mean Value Theorem). *Suppose that a function*

$$f : [a, b] \longrightarrow \mathcal{R}$$

is continuous on $[a, b]$ *and differentiable on* (a, b). *Then there exists some value* $c \in (a, b)$ *such that*

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

The Mean Value Theorem is illustrated in figure 9.6. Note also that in Archimedes's quadrature of the parabola, each inscribed triangle has its middle vertex over the point c from the Mean Value Theorem (cf. page 17 and

figure 1.14 on page 18). The condition that f needs to be continuous on the closed interval but need not be differentiable at the endpoints means, for example, that the Mean Value Theorem applies to a function such as the square root on $[0, 1]$ despite the fact that its graph has a vertical tangent at the origin.

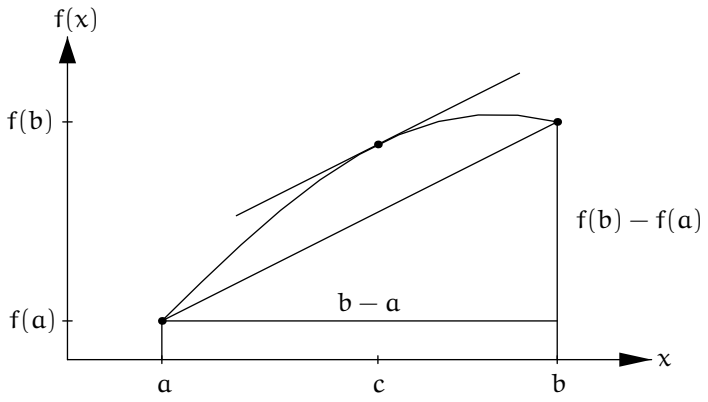


Figure 9.6. The Mean Value Theorem

Proof. Define auxiliary functions

$$g : [a, b] \longrightarrow \mathcal{R}, \quad g(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

and

$$h : [a, b] \longrightarrow \mathcal{R}, \quad h(x) = f(x) - g(x).$$

The graph of g is the line segment from $(a, f(a))$ to $(b, f(b))$, and the function h measures the vertical distance from the graph of g to the graph of f.

Because h meets the conditions for Rolle's Theorem, there exists some value $c \in (a, b)$ such that $h'(c) = 0$. But in general,

$$h'(x) = f'(x) - g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a},$$

so that the condition $h'(c) = 0$ is, as desired,

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

□

Exercise

9.2.1. (a) Sketch the graph of a function $f : [a, b] \longrightarrow \mathcal{R}$ for which there are exactly four possible choices of c in the Mean Value Theorem, i.e., there are exactly four points $c \in (a, b)$ such that $f'(c) = (f(b) - f(a))/(b - a)$.

(b) Sketch the graph of a function $f : [a, b] \longrightarrow \mathcal{R}$ for which there are infinitely many possible choices of c in the Mean Value Theorem, i.e., there are infinitely many points $c \in (a, b)$ such that $f'(c) = (f(b) - f(a))/(b - a)$, but also there are some choices of $c \in (a, b)$ that do not satisfy the condition of the Mean Value Theorem, i.e., there are some points $c \in (a, b)$ such that $f'(c) \neq (f(b) - f(a))/(b - a)$.

9.2.2 Consequences of the Mean Value Theorem

The Mean Value Theorem has a wealth of consequences. To rephrase, it says that *if* f *is continuous on* $[a, b]$ *and differentiable on* (a, b) *then*

$$\frac{f(b) - f(a)}{b - a} = f'(c) \quad \textit{for some } c \in (a, b).$$

To appreciate why this statement enables us to do things that we can't do without it, first note that it involves a sort of tradeoff. The drawback is that:

> *The statement involves a point* $c \in (a, b)$*, but we don't know the value of* c.

But on the other hand, the benefit is that:

> *The statement gives us a connection between the function* f *and its derivative* f' *with no reference to a limit.*

Because limits are elaborate, technical, and sometimes unwieldy, the benefit outweighs the drawback once we learn how to use the theorem despite not knowing c.

Here is an example. Let I be any interval in $\mathcal{R}$, and let

$$f : I \longrightarrow \mathcal{R}$$

be a differentiable function such that $f' = 0$ everywhere on I. As mentioned in exercise 6.4.3 (page 200), this strongly *suggests* that f is constant, but until now an easy proof was not accessible to us. The easy proof proceeds as follows. Let a and b be any two distinct points of I. We may assume that $a < b$. Restrict the domain of f to $[a, b]$. The resulting function satisfies the hypotheses of the Mean Value Theorem. Therefore,

$$f(b) - f(a) = f'(c)(b - a) \quad \text{for some } c \in (a, b).$$

We don't know where c is, but this doesn't matter because $f'(c) = 0$ for all c. That is,

$$f(a) = f(b).$$

Because a and b are arbitrary points of I it follows that f is constant.

It would be eminently reasonable for the reader to underwhelmed by an argument to support the patently obvious fact that if the derivative is always zero then the function is constant. However, the underlying issue is that the fact is patently obvious only because of our intuition that the set of real numbers geometrically forms an unbroken line, a *continuum*. The set of rational numbers also comes with a linear order, and as a subset of the line the rationals leave no gaps of positive length—that is, every real interval of positive length contains rational numbers. Algebraically, the rational numbers and the real numbers can be characterized indistinguishably: addition, subtraction, multiplication, and division work as they should. Nonetheless, if we go through the exercise of defining the concepts in these notes only in the restricted context of the rational numbers, then not all of the results continue to hold. In particular, the function

$$f : \mathcal{Q} \longrightarrow \mathcal{Q}, \quad f(x) = \begin{cases} 0 & \text{if } x^2 < 2, \\ 1 & \text{if } x^2 > 2 \end{cases}$$

is differentiable at each point $x \in \mathcal{Q}$, its derivative is 0 everywhere, and yet it is not a constant function. Thus, any argument that if the derivative is zero the function is constant must somehow rely on a property of the real number system that distinguishes it from the rational number system.

The next exercise is to derive more consequences of the Mean Value Theorem.

Exercise

9.2.2. (a) Let $f_1, f_2 : [a, b] \longrightarrow \mathcal{R}$ be differentiable functions such that $f_1' = f_2'$ on $[a, b]$. Show that $f_2 = f_1 + C$ for some constant C.

(b) Let $f : [a, b] \longrightarrow \mathcal{R}$ be a differentiable function such that $f' > 0$ on $[a, b]$. Show that f is strictly increasing on $[a, b]$.

(c) Let $f : [a, b] \longrightarrow \mathcal{R}$ be differentiable and strictly increasing. Must it be true that $f' > 0$ on $[a, b]$? Proof or counterexample.

(d) Let $f : [a, b] \longrightarrow \mathcal{R}$ be a differentiable function such that $f' \geq 0$ on $[a, b]$. Show that f is increasing on $[a, b]$.

(e) Let $f : [a, b] \longrightarrow \mathcal{R}$ be differentiable and increasing. Must it be true that $f' \geq 0$ on $[a, b]$? Proof or counterexample.

9.3 Curve Sketching

To sketch the graph of a function f with the help of calculus, here are some points to bear in mind.

- The formula for f may make clear where f is positive, negative, and zero, i.e., where the graph is above, below, or crossing the x-axis.
- If the formula for f has a denominator then f is undefined at x-values where the denominator is zero. If the numerator is nonzero for such x then f probably has a vertical asymptote at x. Check the sign of f at values slightly larger than x and slightly smaller than x to see whether the graph is rising very high or dropping very low on each side of the asymptote.
- Similarly, if the formula has a square root then f is defined only for x-values where the quantity under the square root is nonnegative, and so on.
- The graph may also have horizontal asymptotes or diagonal asymptotes. Horizontal asymptotes arise if $f(x)$ tends to a limit as $x \to +\infty$ or as $x \to -\infty$, and similarly for diagonal asymptotes if $f(x)/x$ tends to a limit.
- The formula for f' may make clear where f' is positive, negative, and zero, i.e., where the graph is rising, falling, or has a horizontal tangent.
- The formula for f'' may make clear where f'' is positive, negative, and zero, i.e., where the graph is convex (bending up), concave (bending down), or inflecting (switching bend-directions).
- At an x-value where f is undefined it is understood that f' and f'' are undefined as well, and similarly at an x-value where f' is undefined it is understood that f'' is undefined as well. At an x-value where f is defined but f' is not, the graph may have a corner or some other exotic behavior.

Computer graphing technology is so effective and so readily available that sketching curves with the help of calculus may feel like a pointless endeavor, especially because the computer can plot many points quickly and produce a figure that is accurate in shape and scale. However, sometimes calculus can tell us about key features of the graph that are hard to see in computer-generated plots, e.g., the precise location of local extrema, or points of inflection.

Example 9.3.1. Consider the function

$$f(x) = \frac{x^2}{x^2 - 1}, \quad x \in \mathcal{R},\ x \neq \pm 1.$$

Note that f is even (i.e., $f(-x) = f(x)$), so we need only study it for $x \geq 0$. Also,

- $f(0) = 0$.

- $\lim_{x\to+\infty} f(x) = 1$.
- $\lim_{x\to 1^+} f(x) = +\infty$ and $\lim_{x\to 1^-} f(x) = -\infty$.

(Here the second bullet is shorthand for f *tends to* 1 *as its inputs grow large and positive*, and the third bullet is shorthand for f *is large and positive at inputs slightly greater than* 1 and f *is large and negative at inputs slightly less than* 1.) Compute that the derivative of f is

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

Thus $f'(0) = 0$ and $f'(x) < 0$ for $0 < x < 1$ and for $1 < x$. Similarly, a bit of algebra shows that

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

Thus $f''(x) < 0$ for $0 \le x < 1$ and $f''(x) > 0$ for $1 < x$. We can present many of our observations in a table. The icons indicate whether the graph of f is rising or falling, and whether it is convex or concave.

	$0 < x < 1$	$1 < x$
f	$-$	$+$
f'	$-$	$-$
f''	$-$	$+$

A computer-generated plot of f (figure 9.7) shows the features that we have deduced analytically,

Example 9.3.2. Consider the function

$$f(x) = x^{1/3} + x^{-1/3} = x^{-1/3}(x^{2/3}+1), \quad x > 0.$$

This function should behave like $x^{-1/3}$ for x near 0, and like $x^{1/3}$ for large x. More specifically,

- $f(x) > 0$ for all $x > 0$.
- $\lim_{x\to 0^+} f(x) = +\infty$ and $\lim_{x\to+\infty} f(x) = +\infty$.

Compute that the derivative of f is

$$f'(x) = \frac{1}{3}x^{-2/3} - \frac{1}{3}x^{-4/3} = \frac{1}{3}x^{-4/3}(x^{2/3}-1), \quad x > 0.$$

Thus $f'(1) = 0$, and $f'(x) < 0$ for $0 < x < 1$, and $f'(x) > 0$ for $1 < x$. Similarly, a bit of algebra shows that

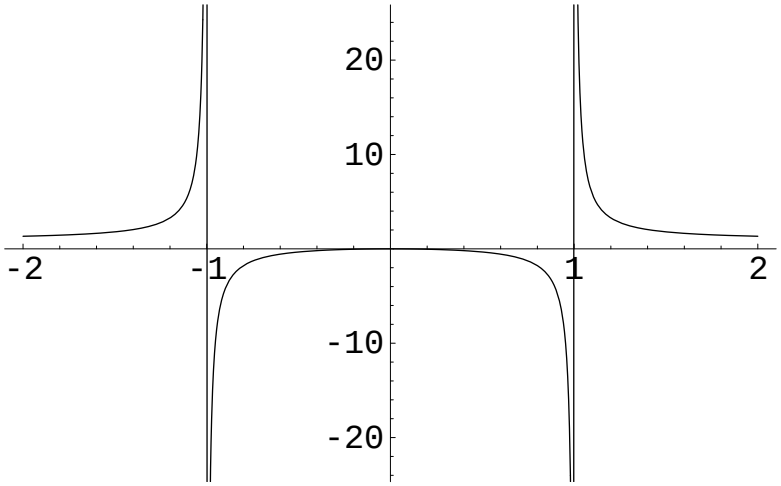


Figure 9.7. Graph of $f(x) = x^2/(x^2-1)$

$$f''(x) = \frac{2}{9}x^{-7/3}(2 - x^{2/3}), \quad x > 0.$$

Thus $f''(2^{3/2}) = 0$, and $f''(x) > 0$ for $0 < x < 2^{3/2}$, and $f''(x) < 0$ for $2^{3/2} < x$. We can present many of these observations in a table.

	$0 < x < 1$	$1 < x < 2^{3/2}$	$2^{3/2} < x$
f	$+$	$+$	$+$
f'	$-$	$+$	$+$
f''	$+$	$+$	$-$
	∪	∩	∩

A computer-generated plot of f (figure 9.8) shows some of the features that we have deduced analytically, but the transition from positive to negative curvature at $x = 2^{3/2} \approx 2.828$ is not really visible, nor is the asymptotic behavior $f(x) \sim x^{1/3}$ for large x.

Example 9.3.3. Consider the function

$$f(x) = 2\sin(x) - \sin(2x), \quad -\pi \le x \le \pi.$$

Note that f is odd (i.e., $f(-x) = -f(x)$), so we may study it on $[0, \pi]$ instead. In particular, $f(0) = f(\pi) = 0$. The derivative of f is (now suppressing the domain from the notation)

$$f'(x) = 2\cos(x) - 2\cos(2x).$$

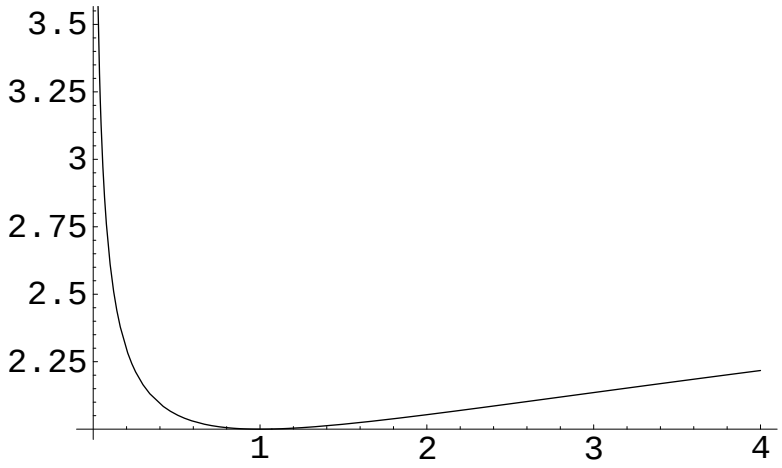


Figure 9.8. Graph of $f(x) = x^{1/3} + x^{-1/3}$

Recall that $\cos(2x) = 2\cos^2(x) - 1$. Therefore,

$$f'(x) = 2(\cos(x) - 2\cos^2(x) + 1) = -2(\cos(x) - 1)(2\cos(x) + 1).$$

Thus $f'(x) = 0$ if $x = 0, 2\pi/3$. And $f(2\pi/3) = \sqrt{3} + \sqrt{3}/2 = 3\sqrt{3}/2 \approx 2.6$. Because $-2(\cos(x)-1) \geq 0$ for all x, the sign of f' is determined by the sign of $2\cos(x)+1$, which is positive for $0 \leq x < 2\pi/3$ and negative for $2\pi/3 < x \leq \pi$. The second derivative of f is

$$f''(x) = -2\sin(x) + 4\sin(2x).$$

Recall that $\sin(2x) = 2\sin(x)\cos(x)$. Therefore,

$$f''(x) = -2\sin(x)(1 - 4\cos(x)).$$

Thus $f''(x) = 0$ at $x = 0, \pi, \arccos(1/4)$. Because $\cos(\pi/3) = 1/2$ and $\cos(\pi/2) = 0$, it follows that $\arccos(1/4)$ lies between $\pi/3$ and $\pi/2$. And a small calculation shows that $f(\arccos(1/4)) = 3\sqrt{15}/8 \approx 1.45$. Because $\sin(x)$ is positive for $0 < x < \pi$, while $1 - 4\cos(x)$ is positive for $0 \leq x < \arccos(1/4)$ and negative for $\arccos(1/4) < x \leq \pi$, we have the following table.

	$0 < x < \arccos(1/4)$	$\arccos(1/4) < x < 2\pi/3$	$2\pi/3 < x \leq \pi$
f	$+$	$+$	$+$
f'	$+$	$+$	$-$
f''	$+$	$-$	$-$

A computer-generated plot of f (figure 9.9) shows most of the features that we have deduced analytically, although the inflection points over $\pm \arctan(-1/4)$, where the graph changes from bending up to bending down, are not easy to pick out until the figure is enhanced to emphasize them.

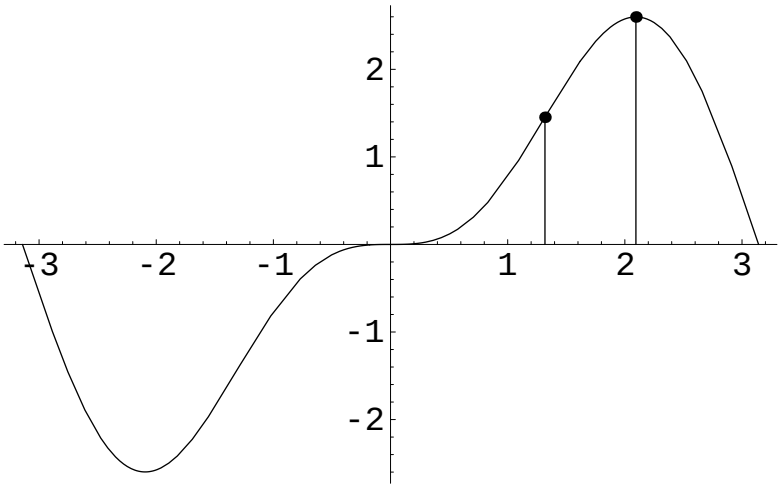


Figure 9.9. Graph of $f(x) = 2\sin(x) - \sin(2x)$

This example arises from the very beginnings of *Fourier analysis*, loosely the theory of expressing general functions as combinations of oscillations, similarly to how we expressed functions as combinations of powers (polynomials) in chapter 8. The weighted combination of the oscillations $\sin(x)$ and $\sin(2x)$ is approximating the 45-degree line identity function $f_1(x) = x$ on $[-\pi, \pi]$. For any positive integer n, the function

$$\begin{aligned} g_n(x) &= 2\sum_{k=1}^{n} (-1)^{k-1}\frac{1}{k}\sin(kx) \\ &= 2\left(\sin(x) - \frac{1}{2}\sin(2x) + \frac{1}{3}\sin(3x) - \cdots + (-1)^{n-1}\frac{1}{n}\sin(nx)\right) \end{aligned}$$

uses more oscillations to approximate the the line more closely. The graph of g_6 is shown in figure 9.10.

Example 9.3.4. Consider the function

$$f(x) = 2(x-1)^{5/3} + 5(x-1)^{2/3} = (x-1)^{2/3}(2x+3).$$

Here we take the domain of f to be the set of all real numbers, even though according to our formalism $f(x)$ is sensible only for $x \geq 1$. The idea is that

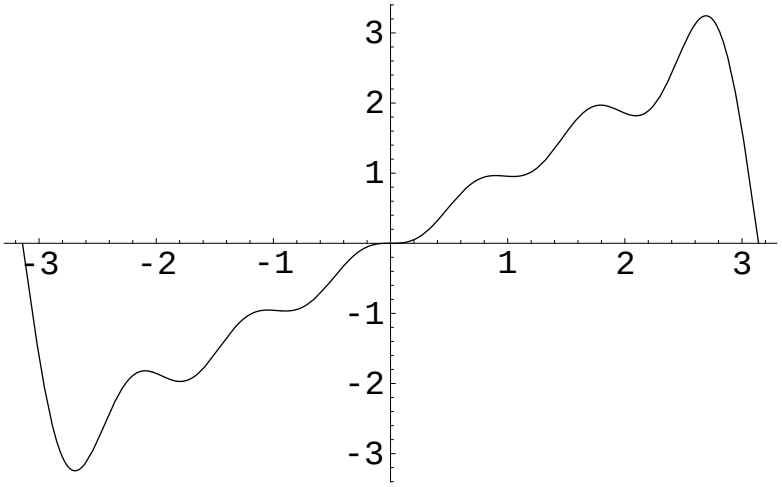


Figure 9.10. Graph of $g_6(x) = 2\sum_{k=1}^{6}(-1)^{k-1}\sin(kx)/k$

for $x < 1$, we can take $(x-1)^{1/3}$ as the negative number whose cube is $x-1$ (this number is $-(1-x)^{1/3}$), and then $f(x)$ is its square. Observe that

- $f(x) = 0$ for $x = 1$ and $x = -3/2$.
- $f(0) = 3$.
- $\lim_{x\to+\infty} f(x) = +\infty$ and $\lim_{x\to-\infty} f(x) = -\infty$.

Next compute that the derivative of f is

$$\begin{aligned} f'(x) &= \frac{10}{3}(x-1)^{2/3} + \frac{10}{3}(x-1)^{-1/3} \\ &= \frac{10}{3}(x-1)^{-1/3}x. \end{aligned}$$

Observe that

- $f'(1)$ is undefined, $\lim_{x\to1^-} f'(x) = -\infty$ and $\lim_{x\to1^+} f'(x) = \infty$.
- $f'(0) = 0$.
- $f'(x) > 0$ for $x < 0$, $f'(x) < 0$ for $0 < x < 1$, and $f'(x) > 0$ for $x > 1$.

The second derivative of f is

$$\begin{aligned} f''(x) &= \frac{20}{9}(x-1)^{-1/3} - \frac{10}{9}(x-1)^{-4/3} \\ &= \frac{10}{9}(x-1)^{-4/3}(2x-3). \end{aligned}$$

Observe that

- $f''(x)$ is undefined at $x = 1$ (naturally, because $f'(1)$ was already undefined).
- $f''(3/2) = 0$.
- $f''(x) < 0$ for $x < 1$ and $1 < x < 3/2$, and $f''(x) > 0$ for $x > 3/2$.

Along with the value $f(0) = 3$, note that

$$f(3/2) = (1/2)^{2/3} \cdot 6 = 6/2^{2/3} > 6/2 = 3.$$

We can present many of our observations in a table.

	$x < -3/2$	$-3/2 < x < 0$	$0 < x < 1$	$1 < x < 3/2$	$3/2 < x$
f	$-$	$+$	$+$	$+$	$+$
f'	$+$	$+$	$-$	$+$	$+$
f''	$-$	$-$	$-$	$-$	$+$

A computer-generated plot of f (figure 9.11) shows most of the features that we have deduced analytically, but it does not clearly show that the graph inflects at $(3/2, f(3/2))$.

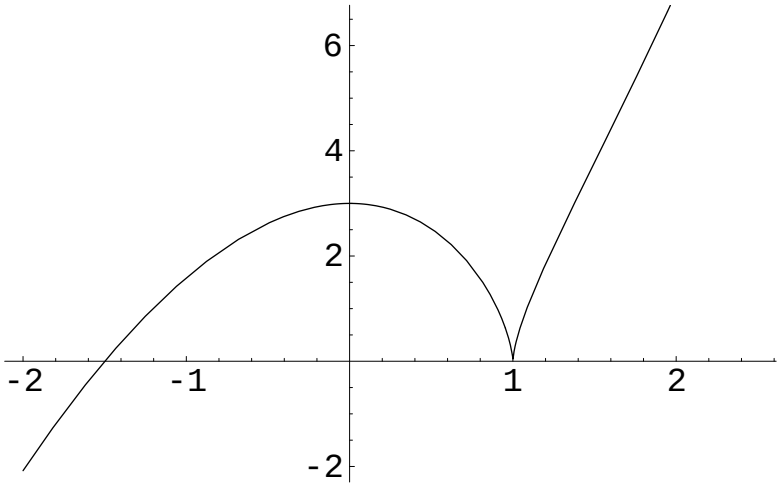


Figure 9.11. Graph of $f(x) = (x-1)^{2/3}(2x+3)$

Example 9.3.5 (Shape of the power function). The following table summarizes many of our observations about the power function $f_\alpha(x)$ on $\mathcal{R}_{>0}$,

extended to $x = 0$ when possible. In all cases the graph lies in the first coordinate quadrant and passes through the point $(1, 1)$. For the limits in the table, Proposition 4.1.5 says that

$$\lim_{s\to 0} f_\alpha(s) = \begin{cases} \infty & \text{if } \alpha < 0 \\ 1 & \text{if } \alpha = 0 \\ 0 & \text{if } \alpha > 0, \end{cases}$$

and then, because we define

$$\lim_{s\to\infty} f_\alpha(s) = \lim_{s\to 0} f_\alpha(1/s)$$

and then note that $f_\alpha(1/s) = (1/s)^\alpha = s^{-\alpha} == f_{-\alpha}(s)$, we also have

$$\lim_{s\to\infty} f_\alpha(s) = \lim_{s\to 0} f_{-\alpha}(s) = \begin{cases} 0 & \text{if } \alpha < 0 \\ 1 & \text{if } \alpha = 0 \\ \infty & \text{if } \alpha > 0. \end{cases}$$

The observations combine to show that for all $\alpha < 0$, the graph of the function looks qualitatively like the hyperbola-branch $y = 1/x$, that for all α strictly between 0 and 1, the graph looks qualitatively like the square root curve $y = \sqrt{x}$, and that for all $\alpha > 1$, the graph looks qualitatively like the parabola $y = x^2$. Of course, the graph of f_0 is the line $y = 1$ and the graph of f_1 is the line $y = x$.

	$\alpha < 0$	$\alpha = 0$	$0 < \alpha < 1$	$\alpha = 1$	$1 < \alpha$
f_α	$+$	$+$	$+$	$+$	$+$
$f_\alpha(1)$	1	1	1	1	1
$\lim_{s\to 0} f_\alpha(s)$	∞	1	0	0	0
$\lim_{s\to\infty} f_\alpha(s)$	0	1	∞	∞	∞
$f'_\alpha = \alpha f_{\alpha-1}$	$-$	0	$+$	1	$+$
$\lim_{s\to 0} f'_\alpha(s)$	$-\infty$	0	∞	1	0
$\lim_{s\to\infty} f'_\alpha(s)$	0	0	0	1	∞
$f''_\alpha = \alpha(\alpha-1)f_{\alpha-2}$	$+$	0	$-$	0	$+$

Exercises

9.3.1. Let $f = \ln$. What is f''? What is its sign? How does this relate to exercise 5.3.3 (page 161)?

9.3.2. Graph the following functions, giving some discussing critical points, asymptotes, convexity/concavity, and so on, as relevant.

(a) $f(x) = x^3/(1-x^2)$.
(b) $f(x) = (1-x^2)^2$.
(c) $f(x) = x/(1+x^2)$
(d) $f(x) = x + \sin(x)$.
(e) $f(x) = x\ln(x)$.

9.4 Related Rates Story-Problems

In a typical *related rates* problem, some time-dependent process involves two related quantities. At some moment, we presumably can measure one quantity, and we know its rate of change. The idea is to determine the rate of change of the second quantity at that moment. The technique is to differentiate the original relation between the quantities with respect to time, remembering to use the Chain Rule.

Example 9.4.1. *The bottom end of a ladder of length ℓ is being moved away from the wall at constant speed. At what speed is the top of the ladder sliding down the wall?*

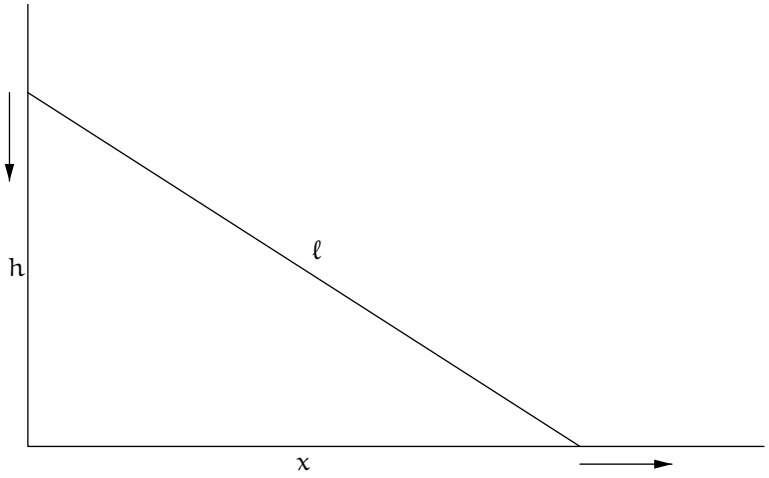


Figure 9.12. Ladder

Let $x(t)$ denote the time-dependent distance of the base of the ladder from the wall, and let $h(t)$ denote the height of the top of the ladder up the wall. The situation is depicted in figure 9.12. Then we have

$$x(t)^2 + h(t)^2 = \ell^2, \quad x(0) = 0, \quad x'(t) = 1.$$

(Here we normalize the *constant speed* given to us by the problem to 1 for convenience.) Differentiate with respect to time,

$$2x(t)x'(t) + 2h(t)h'(t) = 0.$$

That is, because $x'(t) = 1$,

$$h'(t) = -\frac{x(t)}{h(t)}.$$

Because $h(t) = \sqrt{\ell^2 - x(t)^2}$, we have (now suppressing t from the notation),

$$h' = -\frac{x}{\sqrt{\ell^2 - x^2}}\,.$$

Thus, at the beginning moment of the process, when the ladder's base is at the wall ($x = 0$), the horizontal motion of the ladder's base is not causing any vertical motion of the ladder's top down the wall. On the other hand, at the end-moment, when $x = \ell$, the top of the ladder instantaneously has *infinite* vertical velocity down the wall.

Example 9.4.2. *A pedestrian of height* h *walks away from a street light of height* H *at constant speed. At what speed is her shadow-length increasing?*

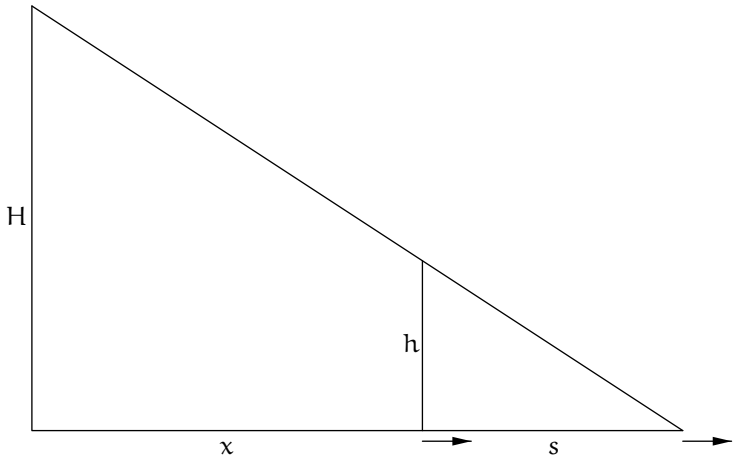


Figure 9.13. Street light, pedestrian, and shadow

Let x denote the pedestrian's horizontal distance from the base of the street light, and let s denote the pedestrian's shadow-length. The situation is depicted in figure 9.13. Similar triangles show that

$$\frac{s}{h} = \frac{x+s}{H}.$$

Differentiate with respect to time,

$$\frac{s'}{h} = \frac{x'+s'}{H},$$

or, after a little bit of algebra, and again normalizing to $x' = 1$,

$$s' = \frac{h}{H-h}.$$

The shadow-length is increasing at a constant rate.

Example 9.4.3. *A child is flying a kite at constant height. Wind is blowing the kite horizontally at constant speed. At what speed is string playing out through the child's hand?*

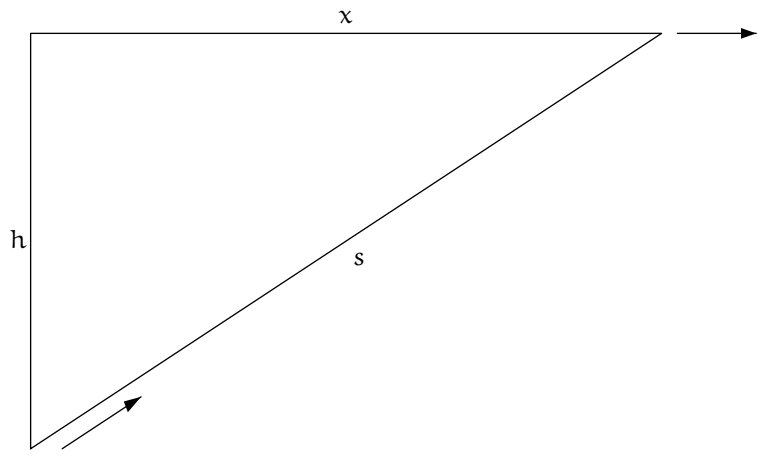


Figure 9.14. Kite

Let h denote the height of the kite, let x denote the horizontal distance from the kite to the child, and let s denote the length of string from the child to the kite. Unrealistically idealize the string as a line segment, so that

$$s^2 = x^2 + h^2.$$

The situation is depicted in figure 9.14. Take time-derivatives,

$$2ss' = 2xx'.$$

Normalizing to $x' = 1$, we have

$$s' = \frac{x}{s} = \frac{x}{\sqrt{x^2 + h^2}}.$$

Thus $s' = 0$ when the kite is directly over the child's head. Also, in the limit as ever more string is played out, s' tends to 1; this is sensible because the proportions of the triangle degenerate toward $s = x$ in the limit.

Example 9.4.4. *A rope is suspended over a pulley at height* y. *A weight is attached to one end of the pulley, and the other end of the rope is being pulled horizontally away from beneath the pulley at constant rate, lifting the weight. At what rate is the weight rising?*

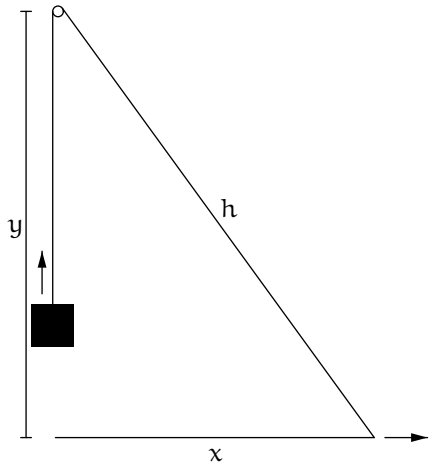


Figure 9.15. Weight and pulley

Let x denote the horizontal distance from point on the floor beneath the pulley to the end of the rope that is being pulled. Let h (for *hypotenuse*) denote the length of rope from the pulley to the end being pulled. Then

$$x^2 + y^2 = h^2.$$

The situation is depicted in figure 9.15. The weight is rising at the rate that rope is passing over the pulley, and rope is passing over the pulley at rate h', so we want to find h'. Differentiate the previous relation,

$$2xx' = 2hh'.$$

And so, normalizing to $x' = 1$,

$$h' = \frac{x}{h} = \frac{x}{\sqrt{x^2 + y^2}}.$$

The weight is initially rising at speed 0, and (assuming that the pulled end of the rope was initially beneath the pulley) it has reached the pulley when $h = 2y$ and thus $x = \sqrt{3}\,y$. The weight is rising at speed $\sqrt{3}/2$ when it reaches the pulley.

Example 9.4.5. *Water is being added to a conical tank of height* H *and radius* R *at constant rate. At what rate is the height of the water in the tank changing?*

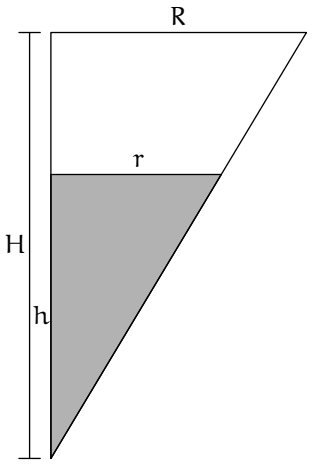


Figure 9.16. Half-profile of conical tank

Let h denote the height of the water in the tank, and r the radius. At any instant, the water in the tank forms a cone similar to the tank, so that

$$\frac{h}{H} = \frac{r}{R},$$

and therefore we can express h in terms of r, and r' in terms of h',

$$h = \frac{H}{R}r \quad \text{and} \quad r' = \frac{R}{H}h'.$$

The situation is depicted in figure 9.16. The volume of water in the tank is

$$V = \frac{1}{3}\pi r^2 h.$$

Therefore, by the product rule,

$$V' = \frac{1}{3}\pi(2rr'h + r^2h') = \frac{1}{3}\pi\left(2r\frac{R}{H}h'\frac{H}{R}r + r^2h'\right) = \pi r^2 h'.$$

But $V' = 1$, and so

$$h' = \frac{1}{\pi r^2}.$$

That is, perhaps obviously in hindsight, the water's rate of rising equals the reciprocal of the cross-sectional circular area of the tank at the water's height. At the beginning of the process, when $h = 0$, apparently h' is instantaneously infinite, and at the end, h' is $1/(\pi R^2)$.

Exercises

9.4.1. A spectator at a tennis match is sitting netside. The court has length ℓ and width w. A player standing at the middle of the baseline hits the ball perfectly horizontally, giving it velocity v. At what rate is the spectator's head swiveling as the ball moves?

9.4.2. A Ferris wheel has radius R, and it rotates at rate ω. It is nighttime. A lantern is suspended immediately above the top of the Ferris wheel. As you ride the Ferris wheel, at what rate is your shadow moving when you are at angle t from the top?

9.4.3. Point p moves along the x-axis at rate a, and point q moves along the y-axis at rate b. At what rate does the distance between them change?

9.4.4. A sphere of radius r has volume $V = 4\pi r^3/3$ and area $A = 4\pi r^2$. If an evaporating spherical drop of water is losing volume at a rate proportional to its area, show that it is losing area at a rate proportional to its radius.

9.4.5. A balloon rises vertically at constant rate, and car travels horizontally at a different constant rate. How is the distance between them changing?

9.4.6. One end of a rope is attached to the bow of a boat. The other end of the rope passes through a ring on a dock, distance h higher than the bow, at constant rate. At what rate does the bow move toward the dock?

10

Integration via Antidifferentiation

Each time that we have integrated a function f in these notes, the result was the difference of two values of a second function F whose derivative was f. The Fundamental Theorem of Calculus says that this phenomenon is general: If f is continuous on $[a, b]$ and $F' = f$ then

$$\int_a^b f = F(b) - F(a).$$

Thus the problem of integrating f is solved whenever we can *antidifferentiate* f, i.e., whenever we can find a function F whose derivative is f. This chapter establishes the Fundamental Theorem of Calculus in section 10.1 and then lays out some antidifferentiation techniques. Some basic antiderivatives are given in section 10.2. Section 10.3 explains how to find certain antiderivatives by a process called forward substitution, and section 10.4 explains a related process called inverse substitution. Section 10.5 presents a useful technique called antidifferentiation by parts.

10.1 The Fundamental Theorem of Calculus

10.1.1 Indefinite Integrals, Antiderivatives

To make the ideas clear, we begin by naming some of the phenomena just discussed in the chapter introduction. First,

Definition 10.1.1 (Indefinite Integral). *Let I be a nonempty interval in $\mathcal{R}$, and consider a function*

$$f : I \longrightarrow \mathcal{R}$$

such that $\int_a^b f$ *exists for all* $a, b \in I$. *An* **indefinite integral of** f *is a second function*

$$F : I \longrightarrow \mathcal{R}$$

such that

$$\int_a^b f = F(b) - F(a) \quad \textit{for all } a, b \in I.$$

So if we have an indefinite integral F of f then we can compute $\int_a^b f$ without going through a process involving sums and limits. Various indefinite integrals that we have found during the course of these notes by going through processes involving sums and limits, i.e., by integrating, are shown in figure 10.1.

f	F
f_α $(\alpha \neq -1)$ on $\mathcal{R}_{>0}$	$f_{\alpha+1}/(\alpha+1)$
f_{-1} on $\mathcal{R}_{>0}$	$\ln$
$\ln$	$f_1 \ln - f_1$
$\exp$	$\exp$
$\cos$	$\sin$
$\sin$	$-\cos$
$\arccos$	$f_1 \arccos - \sin \circ \arccos$

Figure 10.1. Indefinite integrals

The first goal of this chapter is to produce indefinite integrals, and therefore the means to integrate, without integrating. To prepare for doing so, it is very helpful to have some properties of indefinite integrals at hand, as follows.

Proposition 10.1.2 (Indefinite Integral Properties). *Let* I *be a nonempty interval in* $\mathcal{R}$, *let*

$$f, g : I \longrightarrow \mathcal{R}$$

be functions such that $\int_a^b f$ *and* $\int_a^b g$ *exist for all* $a, b \in I$, *and let* $c \in \mathcal{R}$ *be a constant.*

(a) *Suppose that* $F, G : I \longrightarrow \mathcal{R}$ *are respectively indefinite integrals of* f *and* g. *Then* $F + G$ *is an indefinite integral of* $f + g$ *and* cF *is an indefinite integral of* cf.

(b) *Suppose that* $F : I \longrightarrow \mathcal{R}$ *is an indefinite integral of* f. *Then a second function* $\widetilde{F} : I \longrightarrow \mathcal{R}$ *is also an indefinite integral of* f *if and only if* $\widetilde{F} = F + C$ *for some constant* C.

Proof. Exercise 10.1.1(a). □

Soon we will use the "if" part of Proposition 10.1.2(b), i.e.:

Given an indefinite integral F of f,
if $\widetilde{F} = F + C$ for some constant C
then $\widetilde{F}$ is also an indefinite integral of f.

So, given an indefinite integral F of f, we know how to *recognize some* other indefinite integrals of f from it, they being $F + C$ for any C. This is not substantive, but to establish it one does have to verify that if $\int_a^b f = F(b) - F(a)$ for all $a, b \in I$ then the same condition holds with $F + C$ in place of F.

Similarly to Definition 10.1.1, we formally name the second phenomenon from the chapter introduction.

Definition 10.1.3 (Antiderivative). *Let* I *be a nonempty interval in* $\mathcal{R}$, *and consider a function*

$$f : I \longrightarrow \mathcal{R}.$$

An **antiderivative of** f *is a second function*

$$F : I \longrightarrow \mathcal{R}$$

such that

$$F' = f.$$

Note that the antiderivative definition makes no reference to integration. As with the indefinite integral, it will be very helpful for us to have properties of antiderivatives. Because we anticipate that antiderivatives and indefinite integrals will turn out to be the same functions, it is no surprise that the antiderivative properties in the next proposition are exactly the same as the the indefinite integral properties in Proposition 10.1.2.

Proposition 10.1.4 (Antiderivative Properties). *Let* I *be a nonempty interval in* $\mathcal{R}$, *let*

$$f, g : I \longrightarrow \mathcal{R}$$

be functions, and let $c \in \mathcal{R}$ *be a constant.*

(a) *Suppose that* $F, G : I \longrightarrow \mathcal{R}$ *are respectively antiderivatives of* f *and* g. *Then* $F + G$ *is an antiderivative of* $f + g$ *and* cF *is an antiderivative of* cf.

(b) *Suppose that* $F : I \longrightarrow \mathcal{R}$ *is an antiderivative of* f. *Then a second function* $\widetilde{F} : I \longrightarrow \mathcal{R}$ *is also an antiderivative of* f *if and only if* $\widetilde{F} = F + C$ *for some constant* C.

Proof. Exercise 10.1.2(a). □

Soon we will use the "only if" part of Proposition 10.1.4(b), i.e.:

If F and $\widetilde{F}$ are antiderivatives of f
then $\widetilde{F} = F + C$ for some constant C.

So, given an antiderivative F of f, we know that *every* antiderivative of f takes the form $F + C$ for some C. This *is* substantive, its proof relying on the Mean Value Theorem.

Each time that we found an indefinite integral F for a function f, the indefinite integral was also an antiderivative of f, as shown in figure 10.2. The Fundamental Theorem of Calculus asserts that under reasonable conditions, this will always be so. That is, the Fundamental Theorem says that:

Under suitable conditions, integration reduces to antidifferentiation.

With this slogan in mind, we are motivated to study antidifferentiation, which is not innately of interest but is relatively easy, as a means to integration, which is very much of interest but is difficult.

The next result will help us prove the Fundamental Theorem of Calculus.

Proposition 10.1.5. *Let* I *be a nonempty interval in* $\mathcal{R}$, *and let* $f : I \longrightarrow \mathcal{R}$ *be a function. Suppose that some one function* $F : I \longrightarrow \mathcal{R}$ *is both an antiderivative of* f *and an indefinite integral of* f. *Then every antiderivative of* f *is an indefinite integral of* f.

Proof. Consider any antiderivative of $\widetilde{F}$ of f. Because F is an antiderivative of f, the "only if" part of Proposition 10.1.4(b) says, substantively, that $\widetilde{F} = F + C$ for some C. And so now, because F is an indefinite integral of f, the "if" part of Proposition 10.1.2(b) says, easily, that $\widetilde{F}$ is also an indefinite integral of f. □

So, given I and f as in the proposition, if we can produce just *one* function F that is both an antiderivative of f and an indefinite integral of f then we have many indefinite integrals of f.

Exercises

10.1.1. (a) Prove Proposition 10.1.2. For part (b), first show that if $\widetilde{F} = F + C$ then $\widetilde{F}$ inherits the indefinite integral property from F; second, for $\widetilde{F}$ an indefinite integral of f along with F, let $p \in I$ be a fixed point, then state and

f	F	F'
f_α $(\alpha \neq -1)$ on $\mathcal{R}_{>0}$	$f_{\alpha+1}/(\alpha+1)$	f_α
f_{-1} on $\mathcal{R}_{>0}$	ln	f_{-1}
ln	$f_1 \ln - f_1$	ln
exp	exp	exp
cos	sin	cos
sin	$-\cos$	sin
arccos	$f_1 \arccos - \sin \circ \arccos$	arccos

Figure 10.2. Indefinite integrals and their derivatives

explain the relation between $\widetilde{F}(x) - \widetilde{F}(p)$ and $F(x) - F(p)$ where x is any point of I, and conclude from the relation that $\widetilde{F} = F + C$ for some C.

(b) Suppose that $F, G : I \longrightarrow \mathcal{R}$ are respectively indefinite integrals of f and g. Need the product FG be an indefinite integral of the product fg?

10.1.2. (a) Prove Proposition 10.1.4. For part (b), first show that if $\widetilde{F} = F + C$ then $\widetilde{F}$ inherits the antiderivative property from F; second, for $\widetilde{F}$ an antiderivative of f along with F, use the Mean Value Theorem or one of its consequences to show that $\widetilde{F} = F + C$ for some C.

(b) Suppose that $F, G : I \longrightarrow \mathcal{R}$ are respectively antiderivatives of f and g. Need the product FG be an antiderivative of the product fg?

10.1.2 The Fundamental Theorem, Part I

The Fundamental Theorem of Calculus is really two theorems, each of which describes a sense in which differentiation and integration are inverse operations. The second of the two theorems is the one that relates integration and antidifferentiation, but we naturally begin with the first.

Theorem 10.1.6 (Fundamental Theorem of Calculus, Part I). *Let* I *be a nonempty interval in* $\mathcal{R}$*, and let* p *be a point of* I*. Let the function* $f : I \longrightarrow \mathcal{R}$ *be continuous. Define a function*

$$F : I \longrightarrow \mathcal{R}, \qquad F(x) = \int_p^x f.$$

Then F *is differentiable on* I *and*

$$F' = f.$$

Theorem 10.1.6 has several points of interest.

- Its constructed function F is an indefinite integral of its given function f, because for any points a and b of I,
$$F(b) - F(a) = \int_p^b f - \int_p^a f = \int_a^b f.$$
Thus, one interpretation of the first part of the Fundamental Theorem is that
a particular indefinite integral of f *is also an antiderivative of* f.
The pending second part of the Fundamental Theorem, that
every antiderivative of f *is also an indefinite integral of* f,
will follow immediately from this and Proposition 10.1.5.
- The theorem's statement that if $F(x) = \int_p^x f$ then F' exists and equals f, i.e., that the integral of a continuous function up to a variable endpoint differentiates back to the function, says loosely that differentiation inverts integration. It does not say that integration inverts differentiation.
- Geometrically, the idea of the theorem is that:
The rate at which the area of the region under a curve grows as an endpoint moves is the height of the curve over the moving point.
- Because the logarithm is defined as $\ln(x) = \int_1^x f_{-1}$, our calculation in chapter 5 of the derivative $\ln' = f_{-1}$ now amounts to proving a special case of Theorem 10.1.6.
- The theorem says that:
Every continuous function on an interval has an antiderivative, constructed by integration.
We may not be able to write the antiderivative without an integral sign, i.e., we may not be able to find a nice expression for $\int_p^x f$, but nonetheless its derivative is f. For example, the bell-shaped function
$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = \frac{1}{\sqrt{\pi}} e^{-x^2}$$
has as an antiderivative the *error function*,
$$\operatorname{erf} : \mathcal{R} \longrightarrow \left(-\frac{1}{2}, \frac{1}{2}\right), \quad \operatorname{erf}(x) = \int_0^x f,$$
even though we can't simplify the formula for *erf*. Like the logarithm, the error function, defined as an integral, needs its own name. It is part of the area under the bell-shaped curve that describes many limiting behaviors in probability. The factor $1/\sqrt{\pi}$ normalizes f so that the total area beneath its graph is 1. (See figure 10.3.)

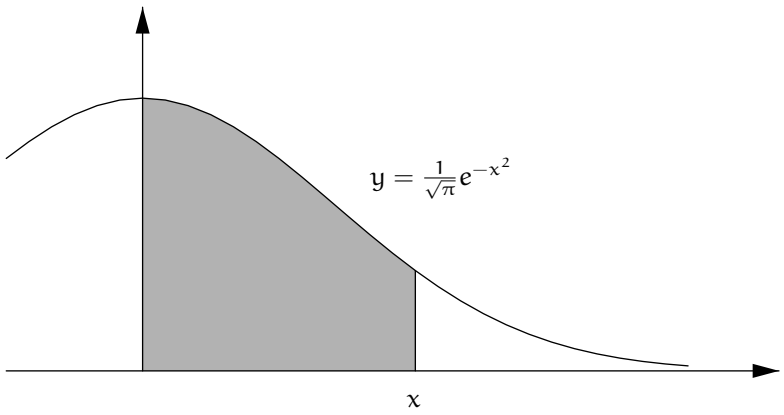


Figure 10.3. erf(x): area under a bell-shaped curve

The hypothesis in Theorem 10.1.6 that f is continuous warrants a quick remark. As we have discussed, every continuous function $f : [p, x] \longrightarrow \mathcal{R}$ (or $f : [x, p] \longrightarrow \mathcal{R}$ if $x < p$) is integrable, but the proof of this fact is technical and so we omitted it. The functions that we know to be integrable over closed, bounded intervals are the (not necessarily continuous) bounded, piecewise monotonic functions. We will state a weaker version of Theorem 10.1.6, hypothesizing such a continuous function f rather than an arbitrary continuous f, at the end of this section.

Proof (of Theorem 10.1.6). Recall the function in the statement of the theorem,

$$F : I \longrightarrow \mathcal{R}, \quad F(x) = \int_p^x f.$$

We want to show that

$$F' = f \quad \text{on } I,$$

showing in the process that F' exists on I. That is, we need to show that for any point x of I,

$$\lim_{s \to x} \frac{F(s) - F(x)}{s - x} = f(x).$$

Let x be any point of I, let s be any point of I with $s \neq x$, and compute

$$\frac{F(s) - F(x)}{s - x} = \frac{\int_p^s f - \int_p^x f}{s - x} = \frac{\int_x^s f}{s - x},$$

So what we need to show is that for our arbitrarily chosen point x in I,

$$\lim_{s \to x} \frac{\int_x^s f}{s - x} = f(x).$$

By the Extreme Value Theorem, because f is continuous, f on the interval $[\min\{x,s\}, \max\{x,s\}]$ has a minimum $f(x_m)$ and a maximum $f(x_M)$, although we don't know the minimizing input x_m or the maximizing input x_M. That is,

$$f(x_m) \le f(t) \le f(x_M) \quad \text{for all } t \in [\min\{x,s\}, \max\{x,s\}].$$

The minimizing input x_m and the maximizing input x_M of f between $\min\{x,s\}$ and $\max\{x,s\}$ depend on s. Assume that $x < s$. By the Inequality Rule for integrals (Proposition 5.5.3, page 175), it follows that

$$(s-x)f(x_m) \le \int_x^s f \le (s-x)f(x_M),$$

or, equivalently,

$$f(x_m) \le \frac{\int_x^s f}{s-x} \le f(x_M).$$

Now assume instead that $s < x$. The same argument gives

$$f(x_m) \le \frac{\int_s^x f}{x-s} \le f(x_M),$$

and multiplying the numerator and the denominator in the middle by -1 gives the same relation as in the case $x < s$,

$$f(x_m) \le \frac{\int_x^s f}{s-x} \le f(x_M).$$

This is, the previous display holds for all $s \neq x$ in I. Now let s go to x. Because x_m and x_M lie in $[\min\{x,s\}, \max\{x,s\}]$, as s goes to x also x_m and x_M are squeezed to x. Consequently, again because f is continuous, $f(x_m)$ and $f(x_M)$ go to $f(x)$ as s goes to x. Thus the quotient in the middle of the previous display is squeezed to the same value,

$$\lim_{s\to x} \frac{\int_x^s f}{s-x} = f(x).$$

This is exactly what we needed to show. □

To summarize, the Extreme Value Theorem shows that the derivative of the function F given by integrating a continuous function f from a fixed endpoint to a variable endpoint is the original function f at the variable endpoint.

As mentioned earlier, the functions that we know to be integrable are the bounded piecewise monotonic functions. The variant of Theorem 10.1.6 that assumes that f is a piecewise monotonic such function is

Theorem 10.1.7 (FTC I, Weaker Variant). *Let* I *be a nonempty interval in* $\mathcal{R}$*, and let* p *be a point of* I*. Let the function* $f : I \longrightarrow \mathcal{R}$ *be continuous and piecewise monotonic. Define a function*

$$F : I \longrightarrow \mathcal{R}, \qquad F(x) = \int_p^x f.$$

Then F *is differentiable on* I *and* $F' = f$*.*

This theorem can be proved with no reference to the Extreme Value Theorem (exercise 10.1.4).

Example 10.1.8. Consider the functions

$$f : \mathcal{R} \longrightarrow \mathcal{R}, \quad f(x) = \frac{1}{1+x^2}$$

and

$$F : \mathcal{R} \longrightarrow \mathcal{R}, \quad F(x) = \int_0^x f.$$

By Theorem 10.1.7, $F' = f$. But also the function $\arctan : \mathcal{R} \longrightarrow \mathcal{R}$ has derivative f, so that $F(x) = \arctan(x) + C$ for some constant C. Substitute $x = 0$ to get $C = 0$. That is, using the notation introduced in Definition 8.2.1 (page 238),

$$\arctan(x) = \int_{x_1=0}^x \frac{1}{1+x_1^2}, \quad x \in \mathcal{R}. \tag{10.1}$$

This formula is similar to the defining formula for the logarithm on page 242,

$$\ln(1+x) = \int_{x_1=0}^x \frac{1}{1+x_1}, \quad x > -1.$$

Just as the logarithm formula led in section 8.3 to an expression for $\ln(1+x)$ as a limit of polynomials when $-1 < x \leq 1$, formula (10.1) leads to such an expression for $\arctan(x)$ (exercise 10.1.6),

$$\boxed{\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} + \cdots, \quad -1 \leq x \leq 1,}$$

or

$$\boxed{\arctan(x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}, \quad -1 \leq x \leq 1.}$$

In particular, it gives the lovely formula

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} - \frac{1}{15} + \cdots,$$

similar to the formula for $\ln(2)$ on page 244.

Example 10.1.9. Exercise 10.1.5 will establish a generalization of Theorem 10.1.6 as follows. *Let the function* $h : I \longrightarrow J$ *from one interval in* $\mathcal{R}$ *to another be differentiable, and let* p *be a point of* J. *Let the function* $f : J \longrightarrow \mathcal{R}$ *be continuous. Define a function*

$$F : I \longrightarrow \mathcal{R}, \qquad F(x) = \int_p^{h(x)} f.$$

Then F *is differentiable on* I *and*

$$F'(x) = f(h(x))h'(x), \quad x \in I.$$

Using this result, we can revisit example 9.4.5 (page 297) in more generality. Consider a tank of height H having cross-sectional area $A(h)$ for each h from 0 to H. The volume of the tank up to a given height h is thus $V = \int_0^h A$. Suppose that water is being added to the tank at constant rate, which we normalize to 1. At what rate is the height of the water in the tank changing at any given time t?

To answer this question, let $h(t)$ denote the time-dependent height of the water in the tank, and note that the time-dependent volume of the water in the tank is

$$V(t) = \int_0^{h(t)} A.$$

By the extension of Theorem 10.1.6 given at the beginning of this example, here with V for F and t for x and A for f,

$$V'(t) = A(h(t))h'(t).$$

But we are assuming that $V'(t) = 1$ for all t, and so for all t,

$$h'(t) = 1/A(h(t)).$$

That is, the water's rate of rising equals the reciprocal of the cross-sectional area of the tank at the water's height. We saw this result for the special case of a conical tank in example 9.4.5, but now Theorem 10.1.6 shows that it holds for any tank shape whatsoever.

Exercises

10.1.3. Define a function

$$f : \mathcal{R} \longrightarrow [0, 1], \quad f(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0. \end{cases}$$

This function is bounded and monotonic, and so it is integrable over any interval $[a, b]$. Define a second function

$$F : \mathcal{R} \longrightarrow \mathcal{R}, \quad F(x) = \int_0^x f.$$

Is F differentiable? Does this contradict Theorem 10.1.6 or Theorem 10.1.7?

10.1.4. (a) Explain why to prove Theorem 10.1.7, it suffices to prove the theorem but with the stronger hypothesis that the continuous function f is monotonic rather than only piecewise monotonic. (To make things simpler, assume that the piecewise monotonicity involves only two pieces.)

(b) Assume that f is increasing. Show that with this assumption, Theorem 10.1.7 can be proved with no reference to the Extreme Value Theorem.

10.1.5. Let $f : \mathcal{R} \longrightarrow \mathcal{R}$ be continuous. Consider the function

$$G : \mathcal{R} \longrightarrow \mathcal{R}, \quad G(x) = \int_0^{x^2} f.$$

Theorem 10.1.6 does not apply here to say that $G' = f$, because the upper limit of integration in the integral that defines G is not x itself, but rather a function of x.

(a) Define two functions,

$$F : \mathcal{R} \longrightarrow \mathcal{R}, \quad F(x) = \int_0^x f$$

and

$$g : \mathcal{R} \longrightarrow \mathcal{R}, \quad g(x) = x^2.$$

Explain why G is a composition (which?) of F and g.

(b) Use the Chain Rule and Theorem 10.1.6 to differentiate G.

(c) Let $g, h : \mathcal{R} \longrightarrow \mathcal{R}$ be differentiable functions (g here is now general, no longer the specific g from earlier in this exercise) and consider the function

$$H : \mathcal{R} \longrightarrow \mathcal{R}, \quad H(x) = \int_{g(x)}^{h(x)} f.$$

Explain why H is differentiable, and compute H'. (The new wrinkle here is that now both endpoints of integration vary. Your solution should reduce the situation back to applications of the case of one variable endpoint, and then use the Chain Rule for those cases as in part (b).)

10.1.6. Similarly to the analysis of the logarithm in section 8.3, obtain the boxed formula for $\arctan(x)$ given in the section.

10.1.3 The Fundamental Theorem, Part II

Whereas Part I of the Fundamental Theorem is a statement about the derivative of the integral, our first statement of Part II is a statement about the integral of the derivative.

Theorem 10.1.10 (Fundamental Theorem of Calculus, Part II). *Let* I *be a nonempty interval in* $\mathcal{R}$*. Let the function* $f : I \longrightarrow \mathcal{R}$ *be continuous. Suppose that the function* $F : I \longrightarrow \mathcal{R}$ *is an antiderivative of* f*, which is to say that* $F' = f$*. Then for any points* a *and* b *of* I*,*

$$\int_a^b f = F(b) - F(a).$$

Thus the second part of the Fundamental Theorem says loosely that integration inverts differentiation, in that the integral of the derivative is the difference of the original function's values at the two endpoints. That is:

> *The integral of the rate of change of a function is the net change in the function.*

Note also that the second part of the Fundamental Theorem says that every antiderivative of f is also an indefinite integral of f.

Proof. Let p be any point of I, and define $\widetilde{F} : I \longrightarrow \mathcal{R}$ to be $\widetilde{F}(x) = \int_p^x f$. Theorem 10.1.6 says that $\widetilde{F}' = f$, which is to say that $\widetilde{F}$ is an antiderivative of f. And as discussed after the statement of Theorem 10.1.6, $\widetilde{F}$ is also an indefinite integral of f. So Proposition 10.1.5 says that our given antiderivative F of f is an indefinite integral of f, meaning precisely that $\int_a^b f = F(b) - F(a)$ for all $a, b \in I$, as desired.

An essentially identical proof, more self-contained rather than citing Proposition 10.1.5, is as follows. Let p be any point of I, and define $\widetilde{F} : I \longrightarrow \mathcal{R}$ to be $\widetilde{F}(x) = \int_p^x f$. Then $\widetilde{F}' = f$ by Theorem 10.1.6, and so, because also our given F is such that $F' = f$, it is a consequence of the Mean Value Theorem that $\widetilde{F} = F + C$ for some C. Thus for any points a and b of I,

$$\int_a^b f = \widetilde{F}(b) - \widetilde{F}(a) = (F(b) + C) - (F(a) + C) = F(b) - F(a).$$

Again this completes the proof. The use of the Mean Value Theorem here is precisely what proves the "only if" part of Proposition 10.1.4(b), and the computation just displayed is precisely what proves "if" part of Proposition 10.1.2(b). These two implications are what prove Proposition 10.1.5; this paragraph uses them directly rather than citing it. □

The proof just given that Part II of the Fundamental Theorem follows from Part I is essentially instant, but also Theorem 10.1.10 can be proved directly without reference to Theorem 10.1.6. Here is one version of the argument.

Fix points $a, b \in I$ with $a < b$. Choose any partition points

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

For $j = 1, \ldots, n$, restrict the domain of F to the jth subinterval $[x_{j-1}, x_j]$, and let $[x_{j-1}, x_j]$ and F play the roles of $[a, b]$ and f in the Mean Value Theorem. Then the theorem says that

$$F(x_j) - F(x_{j-1}) = (x_j - x_{j-1})F'(c_j) \quad \text{for some } c_j \in (x_{j-1}, x_j).$$

But $F' = f$, and so the previous display rewrites as

$$F(x_j) - F(x_{j-1}) = (x_j - x_{j-1})f(c_j) \quad \text{for some } c_j \in (x_{j-1}, x_j).$$

Let M_j be any number at least as big as all f-values on $[x_{j-1}, x_j]$. Then the previous equality implies the inequality

$$F(x_j) - F(x_{j-1}) \leq (x_j - x_{j-1})M_j.$$

Summing the left sides of this inequality for $j = 1, \ldots, n$ gives

$$(F(x_1)-F(x_0))+(F(x_2)-F(x_1))+\cdots+(F(x_{n-1})-F(x_{n-2})+(F(x_n)-F(x_{n-1})),$$

which telescopes to $F(b) - F(a)$. On the other hand, summing the right sides of the inequality for $j = 1, \ldots, n$ gives

$$(x_1 - x_0)M_1 + (x_2 - x_1)M_2 + \cdots + (x_{n-1} - x_{n-2})M_{n-1} + (x_n - x_{n-1})M_n.$$

This is an upper sum for $\mathrm{Ar}_a^b(f)$, and because the number n of partition points is arbitrary, as are the partition points x_j and the values M_j that exceed f on $[x_{j-1}, x_j]$, the upper sum is utterly general. So we have shown that

$$F(b) - F(a) \leq T \quad \text{for any upper sum } T \text{ for } \mathrm{Ar}_a^b(f).$$

Some sequence of upper sums converges to $\int_a^b$, and so it follows that

$$F(b) - F(a) \leq \int_a^b f.$$

Similarly $\int_a^b f \leq F(b) - F(a)$, and so we are done,

$$\int_a^b f = F(b) - F(a).$$

So far the argument has assumed that $f \geq 0$, but extending it to general continuous f is just a matter of passing it through the usual hoisting process. The key idea here is that loosely speaking, Part II of the Fundamental Theorem comes from the Mean Value Theorem. Note that the first proof that we gave of Part II tacitly used the Mean Value Theorem as well, along with quoting Part I. Unlike Part I, Part II cannot be proved without recourse to an abstract existence theorem even under simplifying conditions such as assuming that f is monotonic. There is a conceptual reason for this: Whereas Part I zooms in to measure the local rate of change of a quantity arising from large-scale synthesis (the total area of a region), Part II pulls the camera back to make an assertion about a quantity arising from large-scale synthesis of local information.

Introducing a little more notation will clarify how Part II of the Fundamental Theorem of Calculus sometimes reduces integration to antidifferentiation.

Definition 10.1.11 (Antiderivative Notation). *Let I be a nonempty interval in $\mathcal{R}$, and let $f : I \longrightarrow \mathcal{R}$ be a function. Then*

$$\int f \quad \textit{denotes any antiderivative of } f.$$

This definition needs to be read carefully. Recall that an antiderivative of f is a function whose derivative is f, even though the notation just introduced for an antiderivative is very similar to that for an integral. The difference is that in the antiderivative notation $\int f$, the integral sign is bare rather than adorned by limits of integration as it is in the integral notation $\int_a^b f$. Note that

$\int f$ is a function,

any of a family of functions differing from each other by constants, whereas

$\int_a^b f$ is a number.

Introduce one more piece of notation:

Definition 10.1.12 (Notation for Difference of Function-Values at Two Points). *For any function F whose domain includes the points a and b,*

$$F\Big|_a^b \quad \textit{is short for } F(b) - F(a).$$

With all of this notation in place, Part II of the Fundamental Theorem of Calculus can be rephrased.

Theorem 10.1.13 (Fundamental Theorem of Calculus, Part II, Rephrased). *Let* I *be a nonempty interval in* $\mathcal{R}$*. Suppose that the function* $f : I \longrightarrow \mathcal{R}$ *is continuous, and that* $\int f$ *is an antiderivative of* f*. Then for any points* a *and* b *of* I*,*

$$\int_a^b f = \left(\int f\right)\Big|_a^b.$$

That is, any antiderivative of f *is an indefinite integral of* f*.*

The last statement in the theorem is why the antiderivative notation was chosen to resemble an integral. To rephrase the theorem one more time, the integral of f from a to b is the difference of the antiderivative values at the endpoints. And hence, **to integrate it suffices to antidifferentiate**.

Exercises

10.1.7. Let I be a nonempty interval in $\mathcal{R}$, let n be a nonnegative integer, and let

$$f : I \longrightarrow \mathcal{R}$$

have $n+1$ continuous derivatives. (This means that the function $f^{(0)} = f$, the derivative $f^{(1)} = f'$, the second derivative $f^{(2)} = f''$, and so on up to the $(n+1)$st derivative $f^{(n+1)}$ exist and are continuous on I.) Let a and x be points of I.

(a) Explain why

$$f(x) = f(a) + \int_{x_1=a}^{x} f'(x_1).$$

(b) Assuming that $n \geq 1$, explain why

$$f'(x_1) = f'(a) + \int_{x_2=a}^{x_1} f''(x_2),$$

and therefore

$$f(x) = f(a) + f'(a)(x-a) + \int_{x_1=a}^{x}\int_{x_2=a}^{x_1} f''(x_2).$$

(c) Continue in this vein to explain why, if $n \geq 2$,

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \int_{x_1=a}^{x}\int_{x_2=a}^{x_1}\int_{x_3=a}^{x_2} f'''(x_3),$$

(d) Explain why in general,

$$f(x) = P_n(x) + R_n(x),$$

where the degree n approximating polynomial is

$$\begin{aligned} P_n(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \\ &= \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k, \end{aligned}$$

and the corresponding remainder is an $(n+1)$-fold iterated integral,

$$R_n(x) = \int_{x_1=a}^{x} \cdots \int_{x_{n+1}=a}^{x_n} f^{(n+1)}(x_{n+1}).$$

Note that this exercise unifies much of the work in chapter 8.

10.2 Basic Antidifferentiation

We expand the antiderivative notation introduced a moment ago.

Definition 10.2.1 (Antiderivative Notation With a Variable). *Let* I *be a nonempty interval in* $\mathcal{R}$*, and let* $f : I \longrightarrow \mathcal{R}$ *be a function. Then, letting* x *denote a variable,*

$$\int f(x)\,dx \text{ denotes any antiderivative of } f, \text{ viewed as a function of } x.$$

Every differentiation formula gives rise to an antidifferentiation formula. Some antidifferentiation formulas arising from the derivatives that we have computed during the course of these notes are shown in figure 10.4, using the notation just introduced. In each formula, the right side is *one* antiderivative of the left side, and the general antiderivative is the given specific one plus an arbitrary constant. For example, the first formula in the table perhaps should say instead,

$$\int x^\alpha\,dx = \frac{x^{\alpha+1}}{\alpha+1} + C, \quad \alpha \neq -1.$$

But to keep the notation lean, we omit the "+C" throughout the table, remembering to incorporate it as necessary when we compute. You should know the formulas in the table backward and forward.

Example 10.2.2. To calculate the antiderivative

$$\int x(x^3+1)^3\,dx,$$

$$\int x^{\alpha}\,dx = \frac{x^{\alpha+1}}{\alpha+1}, \quad \alpha \neq -1$$

$$\int \frac{1}{x}\,dx = \ln(|x|)$$

$$\int \ln(x)\,dx = x\ln(x) - x$$

$$\int e^{x}\,dx = e^{x}$$

$$\int \cos(x)\;dx = \sin(x)$$

$$\int \sin(x)\;dx = -\cos(x)$$

$$\int \sec^2(x)\;dx = \tan(x)$$

$$\int \csc^2(x)\;dx = -\cot(x)$$

$$\int \tan(x)\sec(x)\;dx = \sec(x)$$

$$\int \cot(x)\csc(x)\;dx = -\csc(x)$$

$$\int \frac{1}{1+x^2}\,dx = \arctan(x)$$

$$\int \frac{1}{\sqrt{1-x^2}}\,dx = \arcsin(x)$$

$$\int \sec(x)\,dx = \ln(|\tan(x)+\sec(x)|)$$

$$\int \csc(x)\,dx = -\ln(|\cot(x)+\csc(x)|) = \ln(|\tan(x/2)|)$$

Figure 10.4. Basic antidifferentiation formulas

use the Finite Binomial Theorem to expand $(x^3+1)^3$,

$$\begin{aligned}\int x(x^3+1)^3\,dx &= \int x\big((x^3)^3 + 3(x^3)^2 + 3(x^3) + 1\big)\,dx \\ &= \int (x^{10} + 3x^7 + 3x^4 + x)\,dx \\ &= \frac{x^{11}}{11} + \frac{3x^8}{8} + \frac{3x^5}{5} + \frac{x^2}{2} + C.\end{aligned}$$

Example 10.2.3. The antiderivative

$$\int \csc^2(3x)\,dx$$

isn't quite to be found in the figure 10.4 table because of the $3x$. Nonetheless, the table entry $\int \csc^2(x)\,dx = -\cot(x)$ suggests that a natural starting guess is $-\cot(3x)$. By the Chain Rule, the derivative of $-\cot(3x)$ is $3\csc^2(3x)$, and because constants pass through differentiation it follows that

$$\int \csc^2(3x)\,dx = -\frac{1}{3}\cot(3x) + C.$$

Example 10.2.4. The antiderivative

$$\int \frac{1}{a^2+x^2}\,dx \quad (\text{where } a > 0 \text{ is constant})$$

looks similar to the table entry $\int \frac{1}{1+x^2}\,dx = \arctan(x)$. To make the function whose antiderivative we want look more like the function in the table entry, rewrite it,

$$\frac{1}{a^2+x^2} = \frac{1}{a^2}\cdot\frac{1}{1+(x/a)^2}\,.$$

Thus the natural starting guess for our antiderivative is $\arctan(x/a)$. By the Chain Rule, the derivative of $\arctan(x/a)$ is $1/(1+(x/a)^2)\cdot(1/a)$, and so because constants pass through derivatives it follows that

$$\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan\left(\frac{x}{a}\right) + C.$$

Exercises

10.2.1. Verify all of the formulas in figure 10.4 by locating the text or exercise in these notes that differentiates each quantity on the right side.

10.2.2. Find the following antiderivatives.

(a) $\displaystyle\int x(1+\sqrt[3]{x})\,dx.$

(b) $\displaystyle\int \frac{e^{2x}+e^{3x}}{e^{4x}}\,dx.$

(c) $\displaystyle\int \sec(2x+3)\,dx.$

(d) $\displaystyle\int \frac{1}{\sqrt{a^2-x^2}}\,dx$ where $a > 0$ is constant.

10.3 Antidifferentiation by Forward Substitution

Because the emphasis will now be on calculating, we revert to a less formal writing style. All functions are assumed to be integrable and/or differentiable as necessary.

10.3.1 The Forward Substitution Formula

The forward substitution formula is

$$\boxed{\int \Big((f \circ g) \cdot g' \Big) = \Big(\int f \Big) \circ g.} \tag{10.2}$$

The natural first response to this formula is to have no idea what it says, much less how to use it. We will address these matters one at a time.

10.3.2 What the Formula Says and Why It Is True

Formula (10.2) says:

> *If* $\int f$ *is an antiderivative of* f *then the composition* $\big(\int f\big) \circ g$ *is in turn an antiderivative of* $(f \circ g) \cdot g'$.

That is:

> *Finding an antiderivative of a function of the form* $(f \circ g) \cdot g'$ *reduces to finding an antiderivative of the simpler function* f.

To establish the formula, it suffices to show that given an antiderivative $\int f$ of f, the derivative of the composition $\big(\int f\big) \circ g$ is $(f \circ g) \cdot g'$. Compute, using the Chain Rule and the fact that $(\int f)' = f$ by definition, that the derivative of the composition is indeed

$$\left[\left(\int f \right) \circ g \right]' = \left(\left(\int f \right)' \circ g \right) \cdot g' = (f \circ g) \cdot g'.$$

This proof is so simple because the forward substitution formula (10.2) uses the variable-free notation for functions, the uncluttered notation that is well suited to arguments.

10.3.3 Using the Formula in its Variable-Free Form

Actual calculational examples involve functions-as-formulas, so that their notation and the variable-free notation are at odds. Here is an example of translating the with-variable notation (as in Definition 10.2.1) of an antiderivative problem into the variable-free notation of formula (10.2).

Example 10.3.1. To find the antiderivative

$$\int e^{\tan(x)} \sec^2(x)\,dx \quad \text{(this is a function of the variable } x\text{)},$$

introduce (using variable-free notation)

$$f = \exp \quad \text{and} \quad g = \tan, \qquad \text{so that} \quad g' = \sec^2.$$

Then the antiderivative takes the desired form,

$$\int e^{\tan(x)} \sec^2(x)\,dx = \int \Big((f \circ g) \cdot g'\Big).$$

Here the notations are in conflict: the left side carries the information that the variable of the antiderivative is to be named x, while the right side makes no reference to the variable name. In using the variable-free notation, we now must also remember that in the final answer—a function defined by a formula—the variable is to be named x. With this detail filed away somewhere in our memories, note that the forward substitution formula (10.2) says that the antiderivative is instead

$$\int e^{\tan(x)} \sec^2(x)\,dx = \left(\int f\right) \circ g.$$

But

$$\int f = \int \exp = \exp + C,$$

so that

$$\left(\int f\right) \circ g = (\exp + C) \circ \tan = \exp \circ \tan \, + C.$$

That is, bringing the variable x back into the notation,

$$\int e^{\tan(x)} \sec^2(x)\,dx = e^{\tan(x)} + C.$$

10.3.4 Improvement: the Formula With Variables

It would be silly to keep working examples in the fashion of the previous paragraph. Rather than translate every example into variable-free notation, we should translate the one forward substitution formula into notation that incorporates variables. The result is (exercise 10.3.1)

$$\boxed{\int f(g(x))g'(x)\,dx = \int f(u)\,du \quad \text{where } u = g(x).} \tag{10.3}$$

Example 10.3.2. Using the boxed formula, the previous example can be reworked more succinctly, again with $f = \exp$ and $g = \tan$,

$$\begin{aligned}\int e^{\tan(x)} \sec^2(x)\,dx &= \int e^u\,du && \text{where } u = \tan(x)\\ &= e^u + C && \text{where } u = \tan(x)\\ &= e^{\tan(x)} + C.\end{aligned}$$

The basic mnemonic for forward substitution is:

See something and its derivative.

In the previous example, the *something* was $\tan(x)$ and *its derivative* was $\sec^2(x)$. That is, the *something* is the $g(x)$ in the forward substitution formula.

Example 10.3.3. Similarly, let $a > 0$ and consider the antiderivative

$$\int x\sqrt{a^2 - x^2}\,dx.$$

Here the something is $a^2 - x^2$, and its derivative is $-2x$. The problem doesn't quite give us the derivative, but the imperfect fit is easy to fix because constants pass through antidifferentiation,

$$\int x\sqrt{a^2 - x^2}\,dx = -\frac{1}{2}\int (-2x)\sqrt{a^2 - x^2}\,dx.$$

And so by forward substitution,

$$\begin{aligned}\int x\sqrt{a^2 - x^2}\,dx &= -\frac{1}{2}\int (-2x)\sqrt{a^2 - x^2}\,dx\\ &= -\frac{1}{2}\int g'(x)f(g(x))\,dx\\ &= -\frac{1}{2}\int f(u)\,du && \text{where } u = g(x)\\ &= -\frac{1}{2}\int \sqrt{u}\,du && \text{where } u = a^2 - x^2\\ &= -\frac{1}{2}\cdot\frac{2}{3}u^{3/2} + C && \text{where } u = a^2 - x^2\\ &= -\frac{1}{3}(a^2 - x^2)^{3/2} + C.\end{aligned}$$

Exercise

10.3.1. Explain carefully how each side of the forward substitution formula with variables (10.3) arises from its counterpart in the variable-free forward substitution formula (10.2).

10.3.5 Second Improvement: the Procedure Instead of the Formula

Forward substitution is easier in practice if it is viewed as a procedure rather than a formula, whether the formula has variables or not. The procedure is facilitated by a piece of notation due to Leibniz.

Definition 10.3.4 (Leibniz Notation for the Derivative). *Let* u *be a differentiable function of* x*. Then*

$$\frac{du}{dx} \text{ is a synonym for } u'(x).$$

Example 10.3.5. Returning to the antiderivative

$$\int x\sqrt{a^2 - x^2}\,dx,$$

the procedure is to make the substitution

$$u = a^2 - x^2 \quad (\text{the } \textit{something}).$$

Then

$$\frac{du}{dx} = -2x \quad (\text{its derivative}),$$

so that

$$du = -2x\,dx,$$

and consequently

$$x\,dx = -\frac{1}{2}\,du.$$

Therefore the antiderivative is

$$\begin{aligned}\int x\sqrt{a^2 - x^2}\,dx &= -\frac{1}{2}\int \sqrt{u}\,du \quad (\text{substituting})\\ &= -\frac{1}{2}\cdot\frac{2}{3}u^{3/2} + C\\ &= -\frac{1}{3}(a^2 - x^2)^{3/2} + C.\end{aligned}$$

This procedure works, and it has been learned by generations of calculus students. But it is not as self-evident as it appears. The problem is that *the Leibniz notation for the derivative,*

$$\frac{du}{dx},$$

is a single, indivisible symbol, while the separate notations dx *and* du *have not been given meanings at all unless they occur in conjunction*

with the integral sign as part of the antiderivative notation. So the idea that in general

$$\text{“if } u = g(x) \text{ then } du = g'(x)\,dx\text{”}$$

is in isolation a meaningless statement, much less a valid argument. But the Leibniz notation relentlessly suggests it as a valid ritual to practice during the course of antidifferentiation, and the notation has been designed so that the ritual is valid. In procedural terms, the mnemonic for forward substitution is:

$$\boxed{\textit{See } u \textit{ and } du.}$$

Example 10.3.6. Consider the antiderivative

$$\int \tan(x)\,dx = \int \frac{\sin x}{\cos x}\,dx.$$

Make the substitution $u = \cos x$. Then $du = -\sin x\,dx$, and so

$$\int \frac{\sin x}{\cos x}\,dx = -\int \frac{du}{u} = -\ln(|u|) + C = -\ln(|\cos x|) + C.$$

Example 10.3.7. Sometimes a workable substitution takes a little algebra to find. For instance, the antiderivative

$$\int \frac{(1-x)^{2/5}}{x^{12/5}}\,dx,$$

can be rewritten as

$$\int \frac{(1-x)^{2/5}}{x^{12/5}}\,dx = \int \left(\frac{1-x}{x}\right)^{2/5} \frac{1}{x^2}\,dx = \int \left(\frac{1}{x} - 1\right)^{2/5} \frac{1}{x^2}\,dx.$$

Let $u = 1/x - 1$. Then $du = -(1/x^2)\,dx$, and so we have

$$\int \frac{(1-x)^{2/5}}{x^{12/5}}\,dx = -\int u^{2/5}\,du = -\frac{5}{7}u^{7/5} = -\frac{5}{7}\left(\frac{1}{x} - 1\right)^{7/5} + C.$$

To summarize, the variable-free formulation is the right environment for justifying the forward substitution method, but the algorithm is the right way to apply it.

10.3.6 Basic Forward Substitution Formulas

Every basic antidifferentiation formula combines with the Chain Rule to give rise to a forward substitution antidifferentiation formula. These are shown in figure 10.5. As in figure 10.4, a “+C” is tacit in the right side of each formula.

$$\int (g(x))^{\alpha} g'(x)\,dx = \frac{(g(x))^{\alpha+1}}{\alpha+1}, \quad \alpha \neq -1$$

$$\int \frac{g'(x)}{g(x)}\,dx = \ln(|g(x)|)$$

$$\int \ln(g(x))g'(x)\,dx = g(x)\ln(g(x)) - g(x)$$

$$\int e^{g(x)} g'(x)\,dx = e^{g(x)}$$

$$\int \cos(g(x))\,g'(x)\,dx = \sin(g(x))$$

$$\int \sin(g(x))\,g'(x)\,dx = -\cos(g(x))$$

$$\int \sec^2(g(x))\,g'(x)\,dx = \tan(g(x))$$

$$\int \csc^2(g(x))\,g'(x)\,dx = -\cot(g(x))$$

$$\int \sec(g(x))\tan(g(x))\,g'(x)\,dx = \sec(g(x))$$

$$\int \csc(g(x))\cot(g(x))\,g'(x)\,dx = -\csc(g(x))$$

$$\int \frac{g'(x)}{1+g^2(x)}\,dx = \arctan(g(x))$$

$$\int \frac{g'(x)}{\sqrt{1-g^2(x)}}\,dx = \arcsin(g(x))$$

$$\int \sec(g(x))\,g'(x)\,dx = \ln(|\sec(g(x)) + \tan(g(x))|)$$

$$\int \csc(g(x))\,g'(x)\,dx = -\ln(|\csc(g(x)) + \cot(g(x))|)$$

Figure 10.5. Basic forward substitution formulas

Example 10.3.8. We will calculate $\int x^2 \sec^2(x^3+1)\,dx$. Compute that

$$\begin{aligned}\int x^2 \sec^2(x^3+1)\,dx &= \frac{1}{3}\int \sec^2(g(x))g'(x)\,dx \quad \text{where } g(x) = x^3+1\\ &= \frac{1}{3}\tan(x^3+1) + C,\end{aligned}$$

by the seventh formula in table 10.5.

Example 10.3.9. We will calculate $\int te^{t^2}\,dt$. The antiderivative is

$$\int te^{t^2}\,dt = \frac{1}{2}\int g'(t)e^{g(t)}\,dt \quad \text{where } g(t) = t^2$$
$$= \frac{1}{2}e^{t^2} + C.$$

Exercises

10.3.2. Find the following antiderivatives.

(a) $\displaystyle\int e^x \sin(e^x)\,dx.$

(b) $\displaystyle\int \frac{\sin(x)}{\sin(\cos(x))}\,dx.$

(c) $\displaystyle\int (3w^4 + w)^2(12w^3 + 1)\,dw.$

(d) $\displaystyle\int \frac{1}{x\sqrt{1-(\ln(x))^2}}\,dx.$

(e) $\displaystyle\int \frac{x}{1+x^2}\,dx.$

(f) $\displaystyle\int \frac{2}{1+w^2}\,dw.$

(g) $\displaystyle\int \sin^3(x)\,dx.$

(h) $\displaystyle\int \sin^4(x)\,dx.$

(i) $\displaystyle\int \frac{\sin(\ln(x))}{x}\,dx.$

(j) $\displaystyle\int \frac{\cos(\tan(\sqrt{x}))\sec^2(\sqrt{x})}{\sqrt{x}}\,dx.$

10.3.3. The last entry of Table 10.4 is $\int \csc(x)\,dx = -\ln(|\cot(x)+\csc(x)|) = \ln(|\tan(x/2)|)$, but the last entry of Table 10.5 is only $\int \csc(g(x))\,g'(x)\,dx = -\ln(|\csc(g(x))+\cot(g(x))|)$. Why not also $\ln(|\tan(g(x)/2)|)$?

10.3.7 Forward Substitution in Integrals

The forward substitution formula for integrals (as compared to antiderivatives) is

$$\boxed{\int_a^b \Big((f\circ g)\cdot g'\Big) = \int_{g(a)}^{g(b)} f.} \tag{10.4}$$

This formula follows from its counterpart (10.2) for antiderivatives and Part II of the Fundamental Theorem of Calculus (Theorem 10.1.10),

$$\begin{aligned}
\int_a^b \Big((f\circ g)\cdot g'\Big) &= \Big(\int \Big((f\circ g)\cdot g'\Big)\Big)\Big|_a^b && \text{by Theorem 10.1.10}\\
&= \Big(\Big(\int f\Big)\circ g\Big)\Big|_a^b && \text{by (10.2)}\\
&= \Big(\int f\Big)\Big|_{g(a)}^{g(b)} && \text{by definition of composition}\\
&= \int_{g(a)}^{g(b)} f && \text{by Theorem 10.1.10 again.}
\end{aligned}$$

The following notation is well suited to change of variable integral calculations.

Definition 10.3.10 (New Notation for the Integral). *If a function* f *is integrable from* a *to* b *then*

$$\int_{x=a}^{b} f(x)\,dx \text{ is a synonym for } \int_a^b f.$$

As explained after Definition 8.2.1 (page 238), the dummy variable x in this notation can be replaced by any other symbol not already in use. Using this notation, formula (10.4) is

$$\boxed{\int_{x=a}^{b} f(g(x))g'(x)\,dx = \int_{u=g(a)}^{g(b)} f(u)\,du.} \tag{10.5}$$

Example 10.3.11. To evaluate the integral

$$\int_{x=1}^{e} \frac{(\ln x)^2}{x}\,dx,$$

let $u = \ln(x)$. Then $du = dx/x$. Also if $x = 1$ then $u = 0$, and if $x = e$ then $u = 1$. Thus, by a straightforward application of (10.5),

$$\int_{x=1}^{e} \frac{(\ln x)^2}{x}\,dx = \int_{u=0}^{1} u^2\,du = \frac{u^3}{3}\Big|_0^1 = \frac{1}{3}.$$

Example 10.3.12. To evaluate

$$\int_{\pi^2}^{4\pi^2} \frac{\sin(\sqrt{x})}{\sqrt{x}}\,dx,$$

let $u = \sqrt{x}$. Then $du = dx/(2\sqrt{x})$, and so $dx/\sqrt{x} = 2\,du$. Also, if $x = \pi^2$ then $u = \pi$, and if $x = 4\pi^2$ then $u = 2\pi$. Thus

$$\begin{aligned}\int_{\pi^2}^{4\pi^2} \frac{\sin(\sqrt{x})}{\sqrt{x}}\,dx &= 2\int_{\pi}^{2\pi} \sin(u)\,du \\ &= -2\cos(u)\Big|_{\pi}^{2\pi} \\ &= -2(\cos(2\pi) - \cos(\pi)) \\ &= -2(1+1) = -4.\end{aligned}$$

Example 10.3.13. Let $a > 0$ be a constant. We will calculate

$$\int_0^a \frac{1}{a^2+x^2}\,dx.$$

From example 10.2.4 (page 316), an antiderivative of $1/(a^2+x^2)$ is

$$\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan\left(\frac{x}{a}\right).$$

So by Part II of the Fundamental Theorem of Calculus,

$$\int_0^a \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\arctan\left(\frac{x}{a}\right)\Big|_0^a = \frac{1}{a}\arctan(1) = \frac{\pi}{4a}.$$

(We did not use (10.5) for this example.)

Example 10.3.14. To evaluate

$$\int_{x=-1}^{1} \sqrt{1-x^2}\,dx,$$

Note that the upper half of the unit circle is the graph of the function $f(x) = \sqrt{1-x^2}$. Thus the integral is the area above the x-axis and below the upper half of the unit circle, i.e., it is $\pi/2$. (We did not use (10.5) for this example.)

Example 10.3.15. To evaluate

$$\int_{t=0}^{\pi/2} \sin^2(t)\,dt,$$

note that because $\sin(t) = \cos(\pi/2 - t)$, this integral is also

$$\int_{t=0}^{\pi/2} \cos^2(t)\,dt.$$

(See figure 10.6.) Therefore, *twice* the integral is

$$\int_{t=0}^{\pi/2} (\sin^2(t) + \cos^2(t))\,dt = \int_{t=0}^{\pi/2} 1\,dt = \pi/2.$$

Thus

$$\int_{t=0}^{\pi/2} \sin^2(t)\,dt = \int_{t=0}^{\pi/2} \cos^2(t)\,dt = \pi/4.$$

(We did not use (10.5) for this example.)

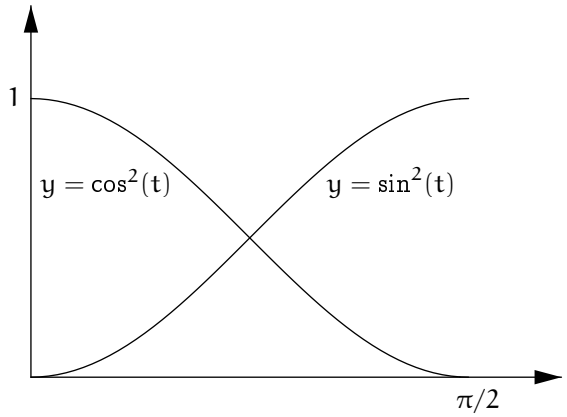


Figure 10.6. Same area under the graphs of $\cos^2$ and $\sin^2$

Example 10.3.16. To evaluate

$$\int_{x=-\ln(\sqrt{3})}^{\ln(\sqrt{3})} \frac{1}{e^x + e^{-x}}\, dx,$$

note that this integral is

$$\begin{aligned}\int_{x=-\ln(\sqrt{3})}^{\ln(\sqrt{3})} \frac{1}{e^x + e^{-x}}\, dx &= \int_{x=-\ln(\sqrt{3})}^{\ln(\sqrt{3})} \frac{e^x}{1 + (e^x)^2}\, dx \\ &= \int_{u=1/\sqrt{3}}^{\sqrt{3}} \frac{1}{1+u^2}\, du \quad \text{where } u = e^x \\ &= \arctan(u)\Big|_{1/\sqrt{3}}^{\sqrt{3}} = \frac{\pi}{3} - \frac{\pi}{6} = \frac{\pi}{6}.\end{aligned}$$

Exercise

10.3.4. Find the following integrals.

(a) $\displaystyle\int_{x=0}^{\sqrt[3]{\pi}} x^2 \sin(x^3)\, dx.$

(b) $\displaystyle\int_{x=\ln(e^2-1)}^{\ln(e^3-1)} \frac{e^x}{1+e^x}\, dx.$

(c) $\displaystyle\int_{x=1/3}^{e/3} \frac{\ln(3x)}{x}\, dx.$

(d) $\displaystyle\int_{x=0}^{\ln(\ln(2))/\ln(2)} 2^x\, dx.$

10.4 Antidifferentiation by Inverse Substitution

10.4.1 The Inverse Substitution Formula and Why It Is True

The forward substitution formula is a bit contrived, in that it produces antiderivatives only of functions having the special form $(f \circ g) \cdot g'$. That is, it works only when u and du are present. The inverse substitution formula is more general. In its variable-free form it says

$$\boxed{\int (f \circ g) = \left(\int (f \cdot h')\right) \circ g \qquad \text{where } h \text{ inverts } g.} \tag{10.6}$$

As with the initial presentation of the forward substitution formula, the meaning and the use of the inverse substitution formula are probably opaque at first glance.

The formula says that if h inverts g and $\int (f \cdot h')$ is an antiderivative of $f \cdot h'$ then the composition $\left(\int (f \cdot h')\right) \circ g$ is in turn an antiderivative of $f \circ g$. That is:

> *Finding an antiderivative of the composition* $f \circ g$ *reduces to finding an antiderivative of the product* $f \cdot h'$ *where* h *inverts* g.

And the inverse formula follows quickly from the forward formula. Indeed, the forward formula $\int \Big((f \circ g) \cdot g'\Big) = \big(\int f\big) \circ g$ with $f \cdot h'$ in place of f is

$$\int \Big(\big((f \cdot h') \circ g\big) \cdot g'\Big) = \left(\int (f \cdot h')\right) \circ g.$$

The right side of the previous display is the desired right side of the inverse substitution formula. On the left side, by the chain rule and then by the fact that because $h \circ g$ is the identity map its derivative is 1, the integrand is

$$\big((f \cdot h') \circ g\big) \cdot g' = (f \circ g) \cdot (h' \circ g) \cdot g' = (f \circ g) \cdot (h \circ g)' = f \circ g.$$

That is, the left side integrand is $f \circ g$ as desired.

As with forward substitution, we now want to rewrite the inverse substitution formula with variables, and then reduce it to a procedure.

10.4.2 The Formula With Variables

The inverse substitution formula with variables is

$$\boxed{\int f(g(x))\,dx = \int f(u)h'(u)\,du \quad \text{where } u = g(x) \text{ and } x = h(u).}$$

For example, consider the antiderivative

$$\int e^{\sqrt{x}}\,dx.$$

This is

$$\int e^{\sqrt{x}}\,dx = \int f(g(x))\,dx \quad \text{where} \quad f(u) = e^u \text{ and } g(x) = \sqrt{x}.$$

The inverse function of $u = \sqrt{x} = g(x)$ is $x = u^2 = h(u)$, whose derivative is $h'(u) = 2u$. Thus according to the inverse substitution formula,

$$\int e^{\sqrt{x}}\,dx = 2\int ue^u\,du \quad \text{where} \quad u = \sqrt{x}.$$

A plausible first guess for $\int ue^u\,du$ is ue^u. This guess has derivative $ue^u + e^u$, so that correcting it to $ue^u - e^u$ gives the correct antiderivative. Therefore, the original antiderivative is

$$\int e^{\sqrt{x}}\,dx = 2\sqrt{x}e^{\sqrt{x}} - 2e^{\sqrt{x}} + C.$$

10.4.3 The Procedure

The procedure is as follows. To compute the antiderivative

$$\int e^{\sqrt{x}}\,dx,$$

make the substitution

$$u = \sqrt{x}.$$

Then the inverse substitution is

$$x = u^2,$$

and so

$$dx = 2u\,du.$$

Therefore the antiderivative is

$$\int e^{\sqrt{x}}\,dx = 2\int e^u u\,du \quad \text{where} \quad u = \sqrt{x}.$$

And from here things go as before.

The inverse substitution procedure differs from the forward substitution procedure in that for inverse substitution, after determining the substitution $u = g(x)$, we indeed invert it by finding $x = h(u)$ and then express dx in terms

of u and du, i.e., $dx = h'(u)\,du$. Then we substitute u and dx. Unlike forward substitution, this doesn't require the problem to contain both u and du.

Inverse substitution doesn't really have a mnemonic counterpart to the *see* u *and* du slogan for forward substitution. The idea is to choose some g to make the function whose antiderivative we want have the form $f(g(x))$, and let $u = g(x)$. Invert g and differentiate the inverse h. Then the new function to antidifferentiate is $f(u)h'(u)$. If this is easier, then the inverse substitution has helped. But there are no general rules for choosing g well. A promising-looking inverse substitution can lead nowhere, and an outlandish-looking one can render a problem trivial. The closest thing to a mnemonic for inverse substitution is:

Choose u *and express* dx *in terms of* u *and* du.

Example 10.4.1. For a promising-looking inverse substitution that does no good, consider an antiderivative closely related to the error function introduced on page 304,

$$\int e^{-x^2}\,dx.$$

A plausible choice is $u = x^2$. Then $x = \sqrt{u}$, and so $dx = du/(2\sqrt{u})$. Thus the antiderivative is

$$\frac{1}{2}\int \frac{e^{-u}}{\sqrt{u}}\,du \quad \text{where } u = x^2,$$

but this is no better than what we started with. In fact, no amount of substitution will allow us to express this antiderivative in terms of functions that have been studied in this course. It has no such expression.

Example 10.4.2. For an unpromising-looking inverse substitution that does good, consider the antiderivative

$$\int \frac{dx}{\sqrt{1+\sqrt{x}}}.$$

With reckless abandon, let $u = \sqrt{1+\sqrt{x}}$. Now invert: $u^2 = 1 + \sqrt{x}$, and so $x = (u^2-1)^2$. It follows that

$$dx = 2(u^2-1)\cdot 2u\,du,$$

and so

$$\begin{aligned}\int \frac{dx}{\sqrt{1+\sqrt{x}}} &= 4\int \frac{(u^2-1)u\,du}{u} = 4\int (u^2-1)\,du \\ &= 4\left(\frac{u^3}{3} - u\right) + C = \frac{4}{3}(1+\sqrt{x})^{3/2} - 4(1+\sqrt{x})^{1/2} + C.\end{aligned}$$

Exercises

10.4.1. (a) Find $\displaystyle\int \frac{x^2+1}{(2x-3)^2}\,dx$.

(b) Let $p(x)$ denote a generic polynomial. Let a and b be real numbers, not both zero, and let n be a positive integer. Explain how an inverse substitution will evaluate $\displaystyle\int \frac{p(x)}{(ax+b)^n}\,dx$.

10.4.2. Find $\displaystyle\int \frac{\sqrt{1-x^2}}{x^2}\,dx$. (Let $x = \cos(u)$.)

10.4.4 Inverse Substitution in Integrals

The inverse substitution formula for integrals (as compared to antiderivatives) is

$$\boxed{\int_a^b (f\circ g) = \int_{g(a)}^{g(b)} (f\cdot h') \qquad \text{where } h \text{ inverts } g.}$$

This formula follows from its variable-free counterpart, similarly to the case of forward substitution. That is,

$$\begin{aligned}
\int_a^b (f\circ g) &= \Big(\int (f\circ g)\Big)\Big|_a^b && \text{by Theorem 10.1.10}\\
&= \left(\left(\int (f\cdot h')\right)\circ g\right)\Big|_a^b && \text{by (10.6)}\\
&= \Big(\int (f\cdot h')\Big)\Big|_{g(a)}^{g(b)} && \text{by definition of composition}\\
&= \int_{g(a)}^{g(b)} (f\cdot h') && \text{by Theorem 10.1.10 again.}
\end{aligned}$$

With variables, the inverse substitution formula for integrals is

$$\boxed{\int_{x=a}^{b} f(g(x))\,dx = \int_{u=g(a)}^{g(b)} f(u)h'(u)\,du \qquad \text{where } u = g(x) \text{ and } x = h(u).}$$

Example 10.4.3. To evaluate

$$\int_0^9 \frac{dx}{\sqrt{1+\sqrt{x}}}.$$

recall the substitution $u = \sqrt{1+\sqrt{x}}$ from example 10.4.2. If $x = 0$ then $u = 1$ and if $x = 1$ then $u = 2$. Thus, by the calculation in example 10.4.2 and by the inverse substitution formula for integrals,

$$\int_0^9 \frac{dx}{\sqrt{1+\sqrt{x}}} = 4\int_1^2 (u^2-1)\,du = 4\left(\frac{1}{3}u^3 - u\right)\Big|_1^2 = \frac{16}{3}.$$

Example 10.4.4. To find $\int_{t=-1}^{0} t\sqrt{t+1}\,dt$, let $u = t+1$. Then $t = u-1$, so $dt = du$. And if $t = -1$ then $u = 0$, while if $t = 0$ then $u = 1$. Hence

$$\int_{t=-1}^{0} t\sqrt{t+1}\,dt = \int_{u=0}^{1} (u-1)\sqrt{u}\,du = \int_{u=0}^{1} (u^{3/2} - u^{1/2})\,du$$
$$= \frac{2}{5}u^{5/2} - \frac{2}{3}u^{3/2}\Big|_0^1 = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15}.$$

We could also do this example by letting $u = \sqrt{t+1}$, so that $t = u^2 - 1$ and then $dt = 2\,du$ and so on.

Example 10.4.5. Let $a > 0$ be constant. To evaluate

$$\int_{x=-a}^{a} \sqrt{a^2 - x^2}\,dx,$$

let $u = x/a$, so that $x = au$. Then $\sqrt{a^2 - x^2} = a\sqrt{1-u^2}$ and $dx = a\,du$. Also, if $x = \pm a$ then $u = \pm 1$. Thus the integral is, citing example 10.3.14 (page 325) at the last step,

$$\int_{x=-a}^{a} \sqrt{a^2 - x^2}\,dx = a^2 \int_{u=-1}^{1} \sqrt{1-u^2}\,du = \frac{\pi a^2}{2}.$$

The answer is inevitable because the upper half of the circle of radius a has equation $y = \sqrt{a^2 - x^2}$ and the upper half of the disk of radius a has area $\pi a^2/2$.

Exercises

10.4.3. Find the following integrals.

(a) $\displaystyle\int_0^1 x^2(x^3+1)^3\,dx.$

(b) $\displaystyle\int_0^{3/2} \frac{1}{\sqrt{9-x^2}}\,dx.$

(c) $\displaystyle\int_0^1 x\sqrt{1-x}\,dx.$

10.4.4. Find the area of the region bounded by the ellipse $x^2/4 + y^2 = 1$ and the lines $x = \pm 1$.

10.4.5. Consider the following argument: *To evaluate the integral*

$$\int_{x=0}^{\pi} \cos^2(x)\,dx,$$

let $u = \sin(x)$. *Then* $\cos(x) = \sqrt{1-u^2}$ *and* $du = \cos(x)\,dx$. *Also, if* $x = 0$ *then* $u = 0$, *and if* $x = \pi$ *then* $u = 0$. *Thus the integral is*

$$\int_{x=0}^{\pi} \cos^2(x)\,dx = \int_{x=0}^{\pi} \cos(x)\,\cos(x)\,dx = \int_{u=0}^{0} \sqrt{1-u^2}\,u\,du = 0.$$

And so the inverse substitution procedure has shown that the integral is zero.

(a) This argument cannot be correct because the integral is visibly positive. What is wrong with it?

(b) Explain why the integral is also

$$\int_{x=0}^{\pi} \sin^2(x)\,dx.$$

Use this fact and the idea of example 10.3.15 (page 325) to find the integral.

10.5 Antidifferentiation by Parts

The formula for antidifferentiation by parts is

$$\boxed{\int fg' = fg - \int gf'.} \tag{10.7}$$

(Here fg' means f times the derivative of g, and similarly for gf'.) This formula follows immediately from antidifferentiating the Product Rule for derivatives,

$$(fg)' = fg' + gf'$$

to get

$$fg = \int fg' + \int gf'.$$

The corresponding formula for integration by parts is, naturally,

$$\boxed{\int_a^b fg' = fg\Big|_a^b - \int_a^b gf'.} \tag{10.8}$$

To use the formula, the idea is to write the function whose antiderivative we seek in the form fg', where the antiderivative of gf' is easier to find.

Example 10.5.1. We will calculate $\int_0^{\pi} x\sin(3x)\,dx$. The first step is to search for an antiderivative of $x\sin(3x)$. Let

$$f(x) = x, \quad g'(x) = \sin(3x),$$
$$f'(x) = 1, \quad g(x) = -\frac{1}{3}\cos(3x).$$

Then by the formula for antidifferentiation by parts

$$\begin{aligned}\int x\sin(3x)\,dx &= \int f(x)g'(x)\,dx \\ &= f(x)g(x) - \int f'(x)g(x)\,dx \\ &= -\frac{x}{3}\cos(3x) + \frac{1}{3}\int \cos(3x)\,dx \\ &= -\frac{x}{3}\cos(3x) + \frac{1}{9}\sin(3x) + C.\end{aligned}$$

Hence

$$\int_0^\pi x\sin(3x)\,dx = \left(-\frac{x}{3}\cos(3x) + \frac{1}{9}\sin(3x)\right)\Big|_0^\pi = -\frac{\pi}{3}\cos(3\pi) = \frac{\pi}{3}.$$

If instead we had proceeded by setting

$$f(x) = \sin(3x), \quad g'(x) = x,$$
$$f'(x) = 3\cos(3x), \quad g(x) = \frac{1}{2}x^2,$$

then by the formula for antidifferentiation by parts

$$\begin{aligned}\int x\sin(3x)\,dx &= \int f(x)g'(x)\,dx \\ &= f(x)g(x) - \int f'(x)g(x)\,dx \\ &= \frac{1}{2}x^2\sin(3x) - \frac{3}{2}\int x^2\cos(3x)\,dx.\end{aligned}$$

In this case the antiderivative $\int x^2\cos(3x)\,dx$ looks more complicated than the one we started with. When you antidifferentiate by parts, it is not always clear what you should take for f and for g'. If you find that things are starting to look more complicated rather than less complicated, you might try another choice for f and g'.

Example 10.5.2. To find $\int \sin(\sqrt{x})\,dx$, first carry out the inverse substitution $u = \sqrt{x}$. Then $x = u^2$, so that $dx = 2u\,du$. Thus

$$\int \sin(\sqrt{x})\,dx = \int \sin(u)\cdot 2u\,du = 2\int u\sin(u)\,du.$$

Now we can antidifferentiate by parts to find $\int u\sin(u)\,du$. Let

$$f(u) = u, \qquad g'(u) = \sin(u),$$
$$f'(u) = 1, \qquad g(u) = -\cos(u).$$

Then

$$\begin{aligned}\int u\sin(u)\,du &= \int f(u)g'(u)du \\ &= f(u)g(u) - \int f'(u)g(u)\,du \\ &= -u\cos(u) + \int \cos(u)\,du \\ &= -u\cos(u) + \sin(u) + C.\end{aligned}$$

Hence

$$\begin{aligned}\int \sin(\sqrt{x})\,dx &= 2\int u\sin u\,du \\ &= -2u\cos(u) + 2\sin(u) + C \\ &= -2\sqrt{x}\,\cos(\sqrt{x}) + 2\sin(\sqrt{x}) + C.\end{aligned}$$

Example 10.5.3. Antidifferentiation by parts is used to evaluate antiderivatives of the forms

$$\int x^n \sin(ax)\,dx, \qquad \int x^n \cos(ax)\,dx, \qquad \int x^n e^{ax}\,dx,$$

where n is a positive integer. All three antiderivatives can be reduced to antiderivatives of the forms

$$\int x^{n-1} \sin(ax)\,dx, \qquad \int x^{n-1} \cos(ax)\,dx, \qquad \int x^{n-1} e^{x}\,dx,$$

and so by applying the process n times we reduce the power of x down to x^0, which gives us antiderivatives that we can find easily. For example, for $\int x^n \sin(ax)\,dx$, let

$$f(x) = x^n, \qquad g'(x) = \sin(ax),$$
$$f'(x) = nx^{n-1}, \qquad g(x) = -\frac{1}{a}\cos(ax).$$

Then

$$\begin{aligned}\int x^n \sin(ax)\,dx &= \int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx \\ &= -\frac{x^n}{a}\cos(ax) + \frac{n}{a}\int x^{n-1}\,\cos(ax)\,dx.\end{aligned}$$

Example 10.5.4. We will calculate $\int \sin(\ln(x))\,dx$. To do so, let

$$\begin{aligned} f(x) &= \sin(\ln(x)), & g'(x) &= 1, \\ f'(x) &= \frac{\cos(\ln(x))}{x}, & g(x) &= x. \end{aligned} \tag{10.9}$$

Then

$$\begin{aligned} \int \sin(\ln(x))\,dx &= \int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx \\ &= x\sin(\ln(x)) - \int \cos(\ln(x))\,dx \end{aligned}$$

Next use the same technique to find an antiderivative of $\cos(\ln(x))$. Let

$$\begin{aligned} f(x) &= \cos(\ln(x)), & g'(x) &= 1, \\ f'(x) &= -\frac{\sin(\ln(x))}{x}, & g(x) &= x. \end{aligned}$$

Then

$$\begin{aligned} \int \cos(\ln(x))\,dx &= \int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx \\ &= x\cos(\ln(x)) + \int \sin(\ln(x))\,dx \end{aligned}$$

It may seem as though we are back where we started, but in fact the two calculations combine to give

$$\int \sin(\ln(x))\,dx = x\sin(\ln(x)) - x\cos(\ln(x)) - \int \sin(\ln(x))\,dx.$$

Thus

$$2\int \sin(\ln(x))\,dx = x\sin(\ln(x)) - x\cos(\ln(x)) + C,$$

and

$$\int \sin(\ln(x))\,dx = \frac{x}{2}(\sin(\ln(x)) - \cos(\ln(x))) + C.$$

Example 10.5.5. We already know from working a sum in chapter 5 that

$$\int \ln(x)\,dx = x\ln(x) - x + C.$$

Now we rederive the result using antidifferentiation by parts. Let

$$\begin{aligned} f(x) &= \ln(x), & g'(x) &= 1, \\ f'(x) &= 1/x, & g(x) &= x. \end{aligned}$$

Then

$$\begin{aligned}\int \ln(x)\,dx &= \int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx \\ &= x\ln(x) - \int 1\,dx \\ &= x\ln(x) - x + C.\end{aligned}$$

We naturally wonder whether this method of finding $\int \ln(x)\,dx$ is related to the summation method of chapter 5. It is, closely, as explained in exercise 10.5.4.

Theorem 10.5.6 (Antiderivative of the Inverse Function). *Let* I *and* J *be intervals in* $\mathcal{R}$*. Let the continuous function*

$$g : I \longrightarrow J$$

have inverse function

$$h : J \longrightarrow I.$$

Suppose that h *is differentiable and that* h' *is continuous. Recall that* f_1 *denotes the identity function. Then an antiderivative of* h *is*

$$\int h = f_1 \cdot h - \left(\int g\right) \circ h.$$

Thus the antiderivative of the inverse function can be expressed in terms of the inverse function and the antiderivative of the original function. With variables, the previous formula is

$$\int h(x)\,dx = xh(x) - \int g(u)\,du \quad \textit{where } u = h(x).$$

It can be shown that because g is continuous, so is its inverse function h. We omit the proof, but it deserves comment that the argument requires that I and J be intervals. (See exercise 10.5.6.) This is an instance where the full description of functions—including domains and codomains—is necessary to analyze a situation.

Proof. Compute, using the facts that $f_1' = f_0$ is the constant function 1 and that $g \circ h$ is the identity function f_1 because g and h are inverse functions to each other,

$$\begin{aligned}
\int h &= \int f_1' \cdot h \\
&= f_1 \cdot h - \int f_1 \cdot h' \qquad \text{antidifferentiating by parts} \\
&= f_1 \cdot h - \int (g \circ h) \cdot h' \\
&= f_1 \cdot h - \left(\int g\right) \circ h \qquad \text{by forward substitution.}
\end{aligned}$$

Alternatively, the reader may find the same calculation easier to read with variables, now noting that $g(h(x)) = x$,

$$\begin{aligned}
\int h(x)\,dx &= \int x' h(x)\,dx \\
&= xh(x) - \int xh'(x)\,dx \\
&= xh(x) - \int g(h(x))h'(x)\,dx \\
&= xh(x) - \int g(u)\,du \quad \text{where } u = h(x).
\end{aligned}$$

□

The formula in Theorem 10.5.6 says that the antiderivative of the inverse function equals a product minus a complementary antiderivative of the original function. This fact about antiderivatives dovetails perfectly, via the Fundamental Theorem of Calculus, with our earlier calculations of the integrals of the exponential and the arc-cosine (see pages 200 and 226): the integral of the inverse function equals a box-area minus a complementary integral of the original function.

As an example, to compute $\int \tan(x)\,dx$, note that $\tan(x) = \sin(x)/\cos(x)$, and so if we let $u = \cos(x)$ so that also $du = -\sin(x)\,dx$ then

$$\int \tan(x)\,dx = -\int \frac{du}{u} = -\ln(|u|) + C = -\ln(|\cos(x)|) + C.$$

Now we can also compute $\int \arctan(x)\,dx$,

$$\begin{aligned}
\int \arctan(x)\,dx &= x\arctan(x) - \int \tan(u)\,du \quad \text{where } u = \arctan(x) \\
&= x\arctan(x) + \ln(|\cos(\arctan(x))|) + C \\
&= x\arctan(x) - \frac{1}{2}\ln(1+x^2) + C.
\end{aligned}$$

Indeed, one can check this by computing

$$\left(x\arctan(x) - \frac{1}{2}\ln(1+x^2)\right)' = \arctan(x) + \frac{x}{1+x^2} - \frac{1}{2}\cdot\frac{2x}{1+x^2} = \arctan(x).$$

Exercises

10.5.1. Calculate the following antiderivatives.

(a) $\int xe^x \, dx$.

(b) $\int x^2 e^x \, dx$.

(c) $\int e^x \sin(x) \, dx$. (Integrate by parts twice, and then don't give up.)

(d) $\int \frac{x}{\sqrt{4-x^2}} \, dx$.

(e) $\int x\sqrt{4-x^2} \, dx$.

(f) $\int x^\alpha \ln(|x|) \, dx$, where $\alpha \in \mathcal{R}$. (Don't forget the case where $\alpha = -1$.)

(g) $\int x^2 \cos(2x) \, dx$.

(h) $\int x \ln(x) \, dx$.

10.5.2. Evaluate the following integrals.

(a) $\int_{x=0}^{1} x \arcsin(x) \, dx$.

(b) $\int_{x=1}^{4} \operatorname{arcsec}(\sqrt{x}) \, dx$. (The arc-secant function has domain $[1, \infty)$ and codomain $[0, \pi/2)$.)

10.5.3. (a) Antidifferentiate by parts to find an antiderivative of arccos.

(b) Antidifferentiate by parts to find an antiderivative of arctan.

10.5.4. This exercise describes *summation by parts*, the discrete analogue of antidifferentiation by parts, which is a continuous process.

(a) Consider two functions (sequences, in fact)

$$f, g : \mathcal{Z}_{\geq 0} \longrightarrow \mathcal{R},$$

and define for $k \in \mathcal{Z}_{\geq 1}$

$$\Delta f(k) = f(k+1) - f(k), \qquad \Delta g(k) = g(k) - g(k-1).$$

Show that for any $n \in \mathcal{Z}_{\geq 1}$,

$$\sum_{k=1}^{n} f(k)\Delta g(k) = f(k)g(k-1)\Big|_1^{n+1} - \sum_{k=1}^{n} \Delta f(k)g(k).$$

(It may be most convincing to write out each side of the equality and confirm that the same terms occur.) Note the similarity between this formula and formula (10.8) for integration by parts.

(b) Recall that in chapter 5 we integrated the logarithm by evaluating the sum

$$\sigma(x) = 1 + 2x + 3x^2 + \cdots + nx^{n-1}, \quad x \neq 1.$$

Using the notation of part (a), let

$$f(k) = k \qquad \text{and} \qquad g(k) = 1 + x + \cdots + x^{k-1},$$

including the case $g(0) = 0$. Show that

$$\sigma(x) = \sum_{k=1}^{n} f(k)\Delta g(k),$$

so that by part (a), in fact

$$\sigma(x) = f(n+1)g(n) - \sum_{k=1}^{n} \Delta f(k)g(k).$$

Evaluate this second expression for $\sigma(x)$ to rederive the sum as computed in exercise 5.4.1 (page 163).

10.5.5. What is wrong with the following argument? Let

$$f(x) = \frac{1}{x}, \qquad g'(x) = 1,$$
$$f'(x) = -\frac{1}{x^2}, \quad g(x) = x.$$

Then

$$\begin{aligned}\int \frac{1}{x}\,dx &= \int f(x)g'(x)\,dx = f(x)g(x) - \int f'(x)g(x)\,dx \\ &= 1 + \int \frac{1}{x}\,dx.\end{aligned}$$

If we subtract $\int \frac{1}{x}\,dx$ from both sides we obtain

$$0 = 1.$$

10.5.6. Consider the set

$$I = \{x \in \mathcal{R} : 0 \leq x < 1 \text{ or } 2 \leq x \leq 3 \text{ or } 4 < x \leq 5\}.$$

Note that I is not an interval. Let $J = [0, 3]$, an interval. Consider the function

$$g : I \longrightarrow J, \quad g(x) = \begin{cases} x & \text{if } 0 \leq x < 1, \\ x - 1 & \text{if } 2 \leq x \leq 3, \\ x - 2 & \text{if } 4 < x \leq 5. \end{cases}$$

The graph of g is shown in figure 10.7. Draw the graph of the inverse function

$$h : J \longrightarrow I.$$

Explain why g is continuous but h is not. Is h integrable? What would change in this exercise if the domain of g were extended to include $x = 1$ and $x = 4$?

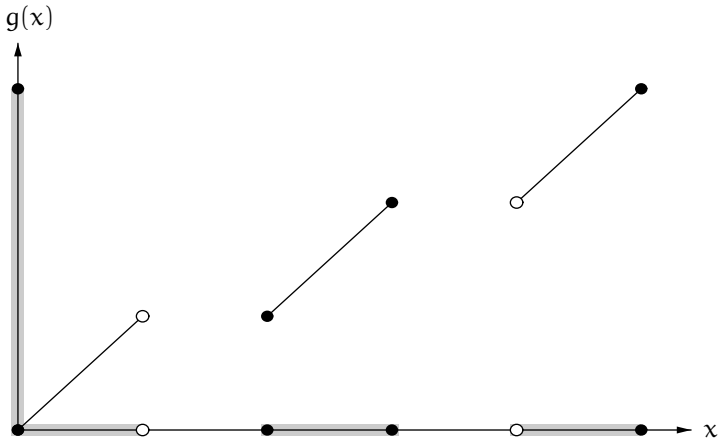


Figure 10.7. A continuous, invertible function with a discontinuous inverse

10.5.7. Let

$$I_n = \int_0^{\pi/2} \sin^n(x)\,dx, \quad n \in \mathcal{Z}_{\geq 0}.$$

(a) Evaluate I_0 and I_1.

(b) Show that

$$I_n = \frac{n-1}{n} I_{n-2}, \quad n \geq 2.$$

Use this formula to evaluate I_3 and I_4.

(c) Show that for n odd,

$$I_n = \frac{2 \cdot 4 \cdot 6 \cdots (n-1)}{3 \cdot 5 \cdot 7 \cdots n}.$$

Show that for n even,

$$I_n = \frac{1 \cdot 3 \cdot 5 \cdots (n-1)}{2 \cdot 4 \cdot 6 \cdots n} \cdot \frac{\pi}{2}.$$

10.5.8. Find reduction formula for the following antiderivatives.

(a) $\int x^n \cos(ax)\, dx$.

(b) $\int x^n e^{ax}\, dx$.

(c) $\int (\ln(x))^n\, dx$.

A

Assumptions About the Real Number System

We assume that there is a **real number system**, a set $\mathcal{R}$ that contains two distinct elements 0 and 1 and is endowed with the algebraic operations of addition,

$$+ : \mathcal{R} \times \mathcal{R} \longrightarrow \mathcal{R},$$

and multiplication,

$$\cdot : \mathcal{R} \times \mathcal{R} \longrightarrow \mathcal{R}.$$

The sum $+(a, b)$ is written $a + b$, and the product $\cdot(a, b)$ is written $a \cdot b$ or more briefly as ab.

The assumed algebraic properties of the real number system are as follows.

Theorem A.0.1 (Field Axioms for $(\mathcal{R}, +, \cdot)$). *The real number system, with its distinct 0 and 1 and with its addition and multiplication, is assumed to satisfy the following set of axioms.*

(a1) *Addition is associative:* $(x + y) + z = x + (y + z)$ *for all* $x, y, z \in \mathcal{R}$.
(a2) 0 *is an additive identity:* $x + 0 = x$ *for all* $x \in \mathcal{R}$.
(a3) *Existence of additive inverses: For each* $x \in \mathcal{R}$ *there exists* $y \in \mathcal{R}$ *such that* $x + y = 0$.
(a4) *Addition is commutative:* $x + y = y + x$ *for all* $x, y \in \mathcal{R}$.
(m1) *Multiplication is associative:* $x(yz) = (xy)z$ *for all* $x, y, z \in \mathcal{R}$.
(m2) 1 *is a multiplicative identity:* $1x = x$ *for all* $x \in \mathcal{R}$.
(m3) *Existence of multiplicative inverses: For each nonzero* $x \in \mathcal{R}$ *there exists* $y \in \mathcal{R}$ *such that* $xy = 1$.
(m4) *Multiplication is commutative:* $xy = yx$ *for all* $x, y \in \mathcal{R}$.
(d1) *Multiplication distributes over addition:* $(x + y)z = xz + yz$ *for all* $x, y, z \in \mathcal{R}$.

All of basic algebra follows from the field axioms. For example, additive and multiplicative inverses are unique, the cancellation law holds, $0 \cdot x = 0$ for all real numbers x, and so on.

Subtracting a real number from another is defined as adding the additive inverse. In symbols,

$$- : \mathcal{R} \times \mathcal{R} \longrightarrow \mathcal{R}, \qquad x - y = x + (-y) \quad \text{for all } x, y \in \mathcal{R}.$$

We also assume that $\mathcal{R}$ is an **ordered** field. This means that there is a subset $\mathcal{R}^+$ of $\mathcal{R}$ (the **positive** elements) such that the following axioms hold.

Theorem A.0.2 (Order Axioms).

(o1) *Trichotomy Axiom: For every real number* x*, exactly one of the following conditions holds:*

$$x \in \mathcal{R}^+, \qquad -x \in \mathcal{R}^+, \qquad x = 0.$$

(o2) *Closure of positive numbers under addition: For all real numbers* x *and* y*, if* $x \in \mathcal{R}^+$ *and* $y \in \mathcal{R}^+$ *then also* $x + y \in \mathcal{R}^+$*.*

(o3) *Closure of positive numbers under multiplication: For all real numbers* x *and* y*, if* $x \in \mathcal{R}^+$ *and* $y \in \mathcal{R}^+$ *then also* $xy \in \mathcal{R}^+$*.*

For all real numbers x and y, define "$x < y$" to mean "$y - x \in \mathcal{R}^+$." The definitions of "$x \leq y$" and "$x > y$" and "$x \geq y$" are analogous. The usual rules for inequalities then follow from the axioms.

Finally, we assume that the real number system is **complete**. Completeness can be phrased in various ways, all logically equivalent. The version of completeness that is currently in Ray Mayer's notes for Mathematics 112 is as follows.

Theorem A.0.3 (Completeness as a Binary Search Criterion). *Every binary search sequence in the real number system converges to a unique limit.*

Two other versions of completeness are phrased in terms of sequences and in terms of set-bounds:

Theorem A.0.4 (Completeness as a Monotonic Sequence Criterion). *Every bounded monotonic sequence in* $\mathcal{R}$ *converges to a unique limit.*

Theorem A.0.5 (Completeness as a Set-Bound Criterion). *Every nonempty subset of* $\mathcal{R}$ *that is bounded above has a least upper bound.*

Convergence is a concept of analysis, and therefore so is completeness. All three statements of completeness are existence statements.

B

Partial List of Named Ideas and Results

Chapter 2: The Rational Power Function

Finite geometric sum, page 24, page 25
Rational power function, page 26
Laws of exponents, page 27
Increasing/decreasing behavior of the rational power function, page 32
Rational power function derivative, page 46, page 46
Rational power function integral, page 53

Chapter 3: Sequence Limits and the Integral

Intervals (Definition 3.1.1), page 58
Endpoints (Definition 3.1.2), page 59
Open and closed intervals (Definition 3.1.3), page 59
Absolute value (Definition 3.2.1), page 67
Basic absolute value properties (Proposition 3.2.2), page 68
Triangle inequality (Theorem 3.2.3), page 68
Relation between absolute values and intervals (Proposition 3.2.4), page 70
Strong approximation lemma (Lemma 3.2.5), page 71
Archimedean property of the real number system (Proposition 3.2.6), page 72
Limit of a real sequence, convergent sequence, divergent sequence (Definition 3.2.7), page 73
Basic sequence limits (Proposition 3.2.8), page 77
Index-translate of a sequence (Definition 3.2.12), page 85
Index-translation rule for sequences (Proposition 3.2.13), page 85
Uniqueness of limits (Proposition 3.2.14), page 86
Progenitive sequence limit rules (Proposition 3.2.15), page 88

Geometric series (Definition 3.2.19), page 95
Geometric series formula (Proposition 3.2.20), page 95
Inequality rule for sequences (Proposition 3.2.21), page 96
Squeezing rule for sequences (Proposition 3.2.22), page 96
The trapped quantity is the common limit (Proposition 3.3.2), page 101
Lower sum, upper sum (Definition 3.3.3), page 105
Basic property of lower and upper sums (Proposition 3.3.4), page 105
Bootstrap result for lower and upper sums (Proposition 3.3.5), page 106
Integral (Definition 3.3.6), page 106
Monotonic function (Definition 3.3.7), page 108
Monotonic functions are integrable (Theorem 3.3.8), page 108
Horizontal additivity of the integral (Proposition 3.3.9), page 109
Piecewise monotonic function (Definition 3.3.10), page 111
Piecewise monotonic functions are integrable (Proposition 3.3.11), page 111
Progenitive integral rules (Proposition 3.3.13), page 116
Inequality rule for integrals (Proposition 3.3.14), page 117

Chapter 4: Function Limits and the Derivative

Approaches, approachable (Definition 4.1.1), page 120
Limit of a function (Definition 4.1.2), page 120
Persistence of inequality (Proposition 4.1.3), page 122
Basic function limits (Proposition 4.1.4), page 124
Limit of the power function at zero (Proposition 4.1.5), page 125
Progenitive function limit rules (Theorem 4.1.6), page 127
Inequality rule for functions (Proposition 4.1.7), page 128
Squeezing rule for functions (Proposition 4.1.8), page 129
Derivative (Definition 4.2.1), page 130
Differentiability implies continuity (Proposition 4.2.2), page 131
Derivative of the power function (Proposition 4.2.4), page 134
Continuity of the power function (Corollary 4.2.5), page 135
Derivative of the absolute value function away from zero (Proposition 4.2.6), page 135
Progenitive derivative rules (Theorem 4.2.7), page 136
Chain rule for derivatives (Theorem 4.2.9), page 139

Chapter 5: The Logarithm Function

Integral with out-of-order endpoints (Definition 5.1.1), page 145
Logarithm (Definition 5.1.2), page 147

The logarithm converts multiplication to addition (Theorem 5.1.3), page 147
Properties of the logarithm (Theorem 5.1.4), page 152
Logarithmic growth formula, page 155
The logarithm grows slowly (Proposition 5.2.1), page 156
Derivative of the logarithm (Theorem 5.3.1), page 157
Integral of the logarithm, page 169
Integral of a function with some negative outputs (Definition 5.5.1), page 170
Progenitive integral rules (Proposition 5.5.2), page 173
Inequality rules for integrals, second version (Proposition 5.5.3), page 175
Area between two curves as an integral (proposition 5.5.4), page 176

Chapter 6: The Exponential Function

Continuity (Definition 6.1.1), page 182
Differentiability implies continuity (Proposition 6.1.2), page 183
Continuity of the power function (Proposition 6.1.3), page 183
Continuity implies integrability (Theorem 6.1.4), page 184
Intermediate value theorem (Theorem 6.1.5), page 185
Intermediate value theorem, second version (Corollary 6.1.6), page 186
The number e (Definition 6.1.7), page 189
The logarithm is bijective (Proposition 6.1.8), page 189
Exponential function (Definition 6.2.1), page 191
Properties of the exponential function (Theorem 6.2.2), page 192
Raising a positive real number to a real power (Definition 6.2.3), page 193
Enhanced property of the logarithm (Proposition 6.2.4), page 193
Laws of real exponents (Proposition 6.2.5), page 193
Exponential growth dominates polynomial growth (Theorem 6.3.1), page 195
The exponential function is its own derivative (Theorem 6.4.1), page 196
Integral of the exponential function (Theorem 6.5.1), page 202
The exponential function as a limit of powers (Theorem 6.6.1), page 203

Chapter 7: The Cosine and Sine Functions

Basic identities for the cosine and sine ((7.2) through (7.8)), page 212
Angle sum and difference formulas ((7.9) through (7.12)), page 214
Double and half angle formulas ((7.13) through (7.16)), page 214
Product formulas ((7.17) through (7.19)), page 214
Difference formulas ((7.20) and (7.21)), page 215
Continuity of the sine and cosine functions (Proposition 7.4.2), page 216
Derivatives of the sine and the cosine at 0 (Proposition 7.4.3), page 217

Derivatives of the sine and the cosine (Theorem 7.4.4), page 218
Integrals of the cosine and the sine (Theorem 7.5.1), page 222
Tangent, cotangent, secant, cosecant (Definition 7.6.1), page 224
Derivatives of the tangent, cotangent, secant, cosecant, page 225
Inverse cosine function (Example 7.7.1), page 226
Integral of the inverse cosine function, page 227
Derivative of the inverse cosine function, page 229
Derivative of the inverse sine function, page 231
Derivative of the inverse tangent function, page 231
Derivative of the inverse cotangent function, page 231

Chapter 9: Theory and Applications of the Derivative

Minimum, maximum, extremum, local minimum, local maximum, local extremum (Definition 9.1.1), page 263
Extreme value theorem (Theorem 9.1.2), page 264
Critical point theorem (Theorem 9.1.3), page 266
Critical point theorem corollary (Corollary 9.1.4), page 266
Rolle's theorem (Theorem 9.2.1), page 280
Mean value theorem (Theorem 9.2.2), page 281

Chapter 10: Integration via Antidifferentiation

Indefinite integral (Definition 10.1.1), page 299
Indefinite integral properties (Proposition 10.1.2), page 300
Antiderivative (Definition 10.1.3), page 301
Antiderivative properties (Proposition 10.1.4), page 301
Fundamental theorem of calculus, part I (Theorem 10.1.6), page 303
Fundamental theorem of calculus!part I, weaker variant (Theorem 10.1.7), page 307
Fundamental theorem of calculus, part II (Theorem 10.1.10), page 310
Antiderivative notation (Definition 10.1.11), page 312)
Difference of function-values notation (Definition 10.1.12), page 312
Fundamental theorem of calculus, part II, rephrased (Theorem 10.1.13), page 313
Antiderivative Notation With a Variable (Definition 10.2.1), page 314
Basic antidifferentation formulas (Figure 10.4), page 315
Forward substitution formula for antiderivatives, page 317
Forward substitution formula for antiderivatives with variables, page 318
Leibniz notation for the derivative (Definition 10.3.4), page 320

Basic forward substitution formulas (Figure 10.5), page 322
Forward substitution formula for integrals, page 323
New notation for the integral (Definition 10.3.10), page 324
Forward substitution formula for integrals with variables, page 324
Inverse substitution formula for antiderivatives, page 327
Inverse substitution formula for antiderivatives with variables, page 327
Inverse substitution formula for integrals, page 330
Inverse substitution formula for integrals with variables, page 330
Antidifferentiation by parts formula, page 332
Integration by parts formula, page 332
Antiderivative of the inverse function (Theorem 10.5.6), page 336
Summation by parts (exercise 10.5.4), page 338

List of Symbols

(a, b) (open interval), 59
$(a, b]$ (half-open interval), 59
(a, ∞) (open half-infinite interval), 59
$(-\infty, b)$ (open half-infinite interval), 59
$(-\infty, b]$ (closed half-infinite interval), 59
$(-\infty, \infty)$ (the number line), 59
(s_n) (sequence), 66
$[a, b)$ (half-open interval), 59
$[a, b]$ (closed interval), 59
$[a, \infty)$ (closed half-infinite interval), 59

$\mathrm{Ar}_a^b(f)$ (area under a graph), 65
A_{tri} (area of first triangle inscribed in parabola), 8
$|\,|$ (absolute value), 67
$\binom{\alpha}{k}$ (binomial coefficient), 256
Ar (area-function), 64
arccos (inverse cosine function), 226
arccot (inverse cotangent function), 230
arcsin (inverse sine function), 230
arctan (inverse tangent function), 230
$\mathrm{Ar}(R)$ (area of a region), 65

$\mathcal{B}$ (the set of bounded subsets of the plane), 60

cos (cosine function), 211
cot (cotangent function), 224
csc (cosecant function), 224

D (parabola directrix), 1, 3
du/dx (Leibniz notation for derivative), 320

e (base of the natural logarithm, 189
$\emptyset$ (empty set), 57
erf (error function), 304
exp (exponential function), 191

F (parabola focus), 2, 3
f_0 (constant function 1), 30
f_1 (identity function), 30
$F|_a^b$ (difference of function-values), 312
f_α (rational power function), 21, 26
$f : A \longrightarrow B$ (f is a function from A to B), 61
f_{-1} (reciprocal function), 30
f' (derivative of f), 130
$f'(x)$ (derivative of f at x), 130
$f(A)$ (range of a function), 62

$\mathrm{graph}(f)$ (graph of a function), 63

$\iff$ (if and only if), 69
$\in$ (in, is in, is an element of), 26
$\int_a^b f$ (integral of f from a to b), 106
$\int f$ (antiderivative of f), 312
$\int f(x)\,dx$ (antiderivative of f), 314
$\int_{x=a}^b f(x)\,dx$ (integral notation with variable and infinitesimal), 324
$\int_{x=a}^b f(x)\,dx$ (integral notation with variable), 238

$\{\,\}$ (set notation), 56

lim (limit of a function), 120
$\lim_n$ (limit of a sequence), 74
ln (logarithm function), 147

$\binom{n}{k}$ (binomial coefficient), 234
$\notin$ (not in, is not in, is not an element of), 26

$\mathcal{P}$ (the set of polygons in the plane), 60
π (area of the unit circle), 209

$\mathcal{Q}$ (the rational numbers), 25, 57

$\mathcal{R}$ (the real numbers), 25, 57
$\mathcal{R}^2$ (the euclidean plane), 57
$\mathcal{R}_{>0}$ (the positive real numbers), 25, 57
$\mathcal{R}_{\geq 0}$ (the nonnegative real numbers), 57

S (lower sum for an area), 105
S_n (sum of areas of n generations of triangles inscribed in parabola), 9
S_n (sum of inner box-areas), 41
sec (secant function), 224
sin (sine function), 211

T (upper sum for an area), 105
tan (tangent function), 224

$\mathcal{Z}$ (the integers), 25, 57
$\mathcal{Z}_{\leq -1}$ (the negative integers), 25, 57
$\mathcal{Z}_{\geq 1}$ (the positive integers), 25, 57
$\mathcal{Z}_{\geq 0}$ (the nonnegative integers), 25, 57

Index

absolute value, 67
  basic properties, 68
  relation with intervals, 70
algebraic function, 63
antiderivative, 301
  notation, 312
  notation with variable, 314
  properties, 301
antidifferentiation by parts formula, 332
antidifferentiation formulas, 315
approach, 120
approachability of a point from a set, 120
arc-cosine function, 226
  derivative, 227
  integral, 226
arc-cotangent function, 230
  derivative, 231
arc-sine function, 230
  derivative, 231
arc-tangent function, 230
  derivative, 231
Archimedean property of the real number system, 72
area between two curves as an integral, 176
area-function, 64

basic function limits, 124
basic sequence limits, 77
  $1/n$ Rule, 77
  $1/n^{\alpha}$ Rule, 77
  nth Power Rule, 78
  nth Root Rule, 78
  Constant Sequence Rule, 77
Binomial Theorem, 260
  Finite, 235
bounded
  subset of the plane, 60
Brouncker's formula for $\ln(2)$, 244

Chain Rule for derivatives, 139
closed interval, 59
codomain of a function, 61
completeness of the real number system, 344
  as a binary search criterion, 344
  as a monotonic sequence criterion, 344
  as a set-bound criterion, 344
compound interest, 205
constant multiple of a sequence, 88
Constant Multiple Rule
  for derivatives, 136
  for function limits, 127
  for sequences, 88
Constant Sequence Rule, 77
continuity
  definition, 182
  of the power function, 183
continuity implies integrability, 184
convergent sequence, 73
cosine, 210
cosine and sine

angle sum and difference formulas, 214
basic identities, 212
derivatives, 218
difference formulas, 215
double and half angle formulas, 214
integrals, 222
product formulas, 214
Taylor polynomial and remainder, 253
Critical Point Theorem, 266

decreasing function, 108
derivative
definition, 130
Leibniz notation, 320
of the absolute value function away from zero, 135
of the logarithm, 157
of the power function, 134
recharacterization, 131
second recharacterization, 138
difference of powers formula, 32
difference of two sequences, 88
difference-quotient
for the parabola, 13
differentiability implies continuity, 183
differentiable function, 130
differentiation rules
progenitive, 136
directrix of a parabola, 1
divergent sequence, 73, 74
domain of a function, 61

e, 189
empty set, 57
endpoints of an interval, 59
error function, 304
exponential function
as a limit of powers, 203
definition, 191
integral of, 202
is its own derivative, 196
properties, 192
Taylor polynomial and remainder, 251
exponential growth dominates polynomial growth, 195
Extreme Value Theorem, 264
extremum of a function, 264
local, 264
strict, 264
strict, 264

field axioms, 343
Finite Binomial Theorem, 235
finite geometric sum formula, 24, 38
for $r = 1/4$, 11
focus of a parabola, 2
forward substitution antidifferentiation formulas, 322
forward substitution formula, 317
for integrals, 323
for integrals with variable and dx, 324
with variables, 318
function, 61
algebraic, 63
codomain, 61
continuous, 182
decreasing, 108
differentiable, 130
domain of, 61
graph, 63
increasing, 108
integrable, 106
limit of, 120
monotonic, 108
piecewise monotonic, 111
range, 62
strictly decreasing, 31
strictly increasing, 31
transcendental, 63
function limits
basic, 124
progenitive, 127
Fundamental Theorem of Calculus
Part I, 303
Part I, weaker variant, 307
Part II, 310
Part II, rephrased, 313

geometric partition, 39

geometric series, 95
  with ratio r, 95
geometric series formula, 95
Goldbach Conjecture, 82
graph of a function, 63

increasing function, 108
indefinite integral, 299
  properties, 300
index-translate of a sequence, 85
index-translation rule for sequences, 85
inequality
  strict, 264
Inequality Rule
  for sequences, 96
Inequality Rule for integrals, 117
integer, 22
integrability of piecewise monotonic functions, 111
integrable function, 106
integral
  definition, 106
  notation with variable, 238
  notation with variable and dx, 324
  of the logarithm, 162
  progenitive rules, 116
  with out-of-order endpoints, 145
integration
  signed, 170
  with out-of-order endpoints, 170
integration by parts formula, 332
Intermediate Value Theorem, 185
  and nth roots, 188
  second version, 186
interval, 58
  closed, 59
  endpoints of, 59
  open, 59
  relation with absolute value, 70
inverse cosine function, 226
  derivative, 227
  integral, 226
inverse cotangent function, 230
  derivative, 231
inverse function
  antiderivative, 336
inverse sine function, 230
  derivative, 231
inverse substitution formula, 327
  for integrals, 330
  with variable, 327
inverse tangent function, 230
  derivative, 231
Irrelevance of Finite Index-shifts, 85

large positive real number, 34
laws of exponents, 27, 37
laws of real exponents, 193
Leibniz notation for the derivative, 320
limit
  basic sequence limits, 77
  of a sequence, 73
  of the power function at zero, 125
limit of a function, 120
logarithm
  definition, 147
  derivative, 157
  integral, 162
  key property, 147
  properties, 152
  Taylor polynomial and remainder, 244
logarithmic growth, 154
lower and upper sums
  basic property, 105
  bootstrap result, 106
lower sum for the area under a graph, 105

maximum of a function, 264
  local, 264
    strict, 264
  strict, 264
Mean Value Theorem, 281
medium-sized positive real number, 34
minimum of a function, 264
  local, 264
    strict, 264
  strict, 264
monotonic function, 108
  integrability of, 108

nth Power Rule, 78
nth Root Rule, 78

$1/n^{\alpha}$ Rule, 77
$1/n$ Rule, 77
open interval, 59
order axioms, 344
ordered pair, 57

pair
  ordered, 57
  unordered, 58
parabola
  algebraic defining equation, 3
  difference-quotient, 13
  directrix, 1
  focus, 2
  geometric defining property, 2
  more general algebraic equations, 4
  quadrature of, 6
  secant slope, 13
  tangent slope, 13
partition
  geometric, 39
  uniform, 39
Persistence of Inequality, 122
piecewise monotonic function, 111
  integrability, 111
power function
  continuity of, 183
  limit at zero, 125
  Taylor polynomial and remainder, 260
Product Rule
  for derivatives, 136
  for function limits, 127
  for sequences, 88
progenitive, 88
progenitive derivative rules, 136
  Chain Rule, 139
  Constant Multiple Rule, 136
  Product Rule, 136
  Quotient Rule, 137
  Reciprocal Rule, 137
  Sum/Difference Rule, 136
progenitive function limit rules, 127
  Constant Multiple Rule, 127
  Inequality Rule, 128
  Product Rule, 127
  Quotient Rule, 127
  Reciprocal Rule, 127
  Squeezing Rule, 129
  Sum/Difference Rule, 127
progenitive integral rules, 116, 173
  Inequality Rule, 117
  Inequality Rule, second version, 175
progenitive sequence limits, 88
  Constant Multiple Rule, 88
  Inequality Rule, 96
  Product Rule, 88
  Quotient Rule, 89
  Reciprocal Rule, 89
  Squeezing Rule, 96
  Sum/Difference Rule, 88
Pythagorean Theorem, 5

quadrature of the parabola, 6
quotient of two sequences, 88
Quotient Rule
  for derivatives, 137
  for function limits, 127
  for sequences, 89

raising a positive real number to a real power, 193
range of a function, 62
rational number, 22
rational power function
  definition, 26
  differentiation of, 44
  integration of, 48
real number system, 343
  Archimedean property, 72
real sequence, 73
reciprocal of a sequence, 88
Reciprocal Rule
  for derivatives, 137
  for function limits, 127
  for sequences, 89
Rolle's Theorem, 280
ruler function, 112

scaling result for power functions, 51
secant slope
  for the parabola, 13
sequence, 66, 73
  approaches a point, 120

convergent, 73
divergent, 73, 74
index-translate of, 85
limit, 73
real, 73
uniqueness of limit, 86
sequence, index-translation rule for, 85
set, 25, 56
defined by conditions, 57
empty, 57
sine, 210
small positive real number, 34
Snell's Law, 276
Squeezing Rule
for sequences, 96
strict inequality, 264
strictly decreasing function, 31
strictly increasing function, 31
Strong Approximation Lemma, 71
sum of two sequences, 88
Sum/Difference Rule
for derivatives, 136
for function limits, 127
for sequences, 88
summation by parts, 338

tangent line
analytic description, 132
geometric description, 132
tangent slope
of the parabola, 13
Taylor polynomial and remainder
for cosine and sine, 253
for the exponential, 251
for the logarithm, 244
for the power function, 260
transcendental function, 63
triangle inequality, 68
basic, 69

uniform partition, 39
uniqueness of sequence limits, 86
unordered pair, 58
upper sum for the area under a graph, 105

volume-function, 65