Applied Calculus
Math 215

Karl Heinz Dovermann

Professor of Mathematics
University of Hawaii

July 5, 1999

This publication was typeset using $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX, the American Mathematical Society's TEX macro system, and LaTeX 2_ε. The graphics were produced with the help of *Mathematica*[1].

This book is dedicated to my wife Emily (Eun Hee) and my sons Christopher and Alexander.

This is a draft which will undergo further changes.

[1] *Mathematica* Version 2.2, Wolfram Research, Inc., Champaign, Illinois (1993).

Contents

Preface

These notes are written for a one-semester calculus course which meets three times a week and is, preferably, supported by a computer lab. The course is designed for life science majors who have a precalculus back ground, and whose primary interest lies in the applications of calculus. We try to focus on those topics which are of greatest importance to them and use life science examples to illustrate them. At the same time, we try of stay mathematically coherent without becoming technical. To make this feasible, we are willing to sacrifice generality. There is less of an emphasis on *by hand* calculations. Instead, more complex and demanding problems find their place in a computer lab. In this sense, we are trying to adopt several ideas from *calculus reform*. Among them is a more visual and less analytic approach. We typically explore new ideas in examples before we give formal definitions.

In one more way we depart radically from the traditional approach to calculus. We introduce differentiability as a local property without using limits. The philosophy behind this idea is that limits are the a big stumbling block for most students who see calculus for the first time, and they take up a substantial part of the first semester. Though mathematically rigorous, our approach to the derivative makes no use of limits, allowing the students to get quickly and without unresolved problems to this concept. It is true that our definition is more restrictive than the ordinary one, and fewer functions are differentiable in this manuscript than in a standard text. But the functions which we do not recognize as being differentiable are not particularly important for students who will take only one semester of calculus. In addition, in our opinion the underlying geometric idea of the derivative is at least as clear in our approach as it is in the one using limits.

More technically speaking, instead of the traditional notion of differentiability, we use a notion modeled on a Lipschitz condition. Instead of an ϵ-δ definition we use an explicit local (or global) estimate. For a function to be differentiable at a point x_0 one requires that the difference between the function and the tangent line satisfies a Lipschitz condition[2] of order 2 in $x - x_0$ for all x in an open interval around x_0, instead of assuming that this difference is $o(x - x_0)$.

This approach, which should be to easy to follow for anyone with a background in analysis, has been used previously in teaching calculus. The author learned about it when he was teaching assistant (Übungsgruppenleiter) for a course taught by Dr. Bernd Schmidt in Bonn about 20 years ago. There this approach was taken for the same reason, to find a less technical and efficient approach to the derivative. Dr. Schmidt followed suggestions which were promoted and carried out by Professor H. Karcher as innovations for a reformed high school as well as undergraduate curriculum. Professor Karcher had learned calculus this way from his teacher, Heinz Schwarze. There are German language college level textbooks by Kütting and Möller and a high school level book by Müller which use this approach.

Calculus was developed by Sir Isaac Newton (1642–1727) and Gottfried Wilhelm Leibnitz (1646–1716) in the 17th century. The emphasis was on differentiation and integration, and these techniques were developed in the quest for solving real life problems. Among the great achievements are the explanation of Kepler's laws, the development of classical mechanics, and the solutions of many important differential equations. Though very successful, the treatment of calculus in those days is not rigorous by nowadays mathematical standards.

In the 19th century a revolution took place in the development of calculus, foremost through the work of Augustin-Louis Cauchy (1789–1857) and Karl Weierstrass (1815–1897), when the modern idea of a function was introduced and the definitions of limits and continuous functions were developed. This elevated calculus to a mature, well rounded, mathematically satisfying theory. This also made calculus much more demanding. A considerable, mathematically challenging setup is required (limits) before one comes to the central ideas of differentiation and integration.

A second revolution took place in the first half of the 20th century with the introduction of generalized functions (distributions). This was stimulated by the development of quantum mechanics in the 1920ies and found is final mathematical form in the work of Laurent Schwartz in the 1950ies.

What are we really interested in? We want to introduce the concepts of differentiation and integration. The functions to which we like to apply these techniques are those of the first period. In this sense, we do not

[2]see page 42 of: A. Zygmund, *Trigonometric Series*, Vol I, Cambridge University Press, 1959, reprinted with corrections and some additions 1968.

need the powerful machine developed in the 19th century. Still, we like to be mathematically rigorous because this is the way mathematics is done nowadays. This is possible through the use of the slightly restrictive notion of differentiability which avoids the abstraction and the delicate, technically demanding notions of the second period.

To support the student's learning we rely extensively on examples and graphics. Often times we accept computer generated graphics without having developed the background to deduce their correctness from mathematical principles.

Calculus was developed together with its applications. Sometimes the applications were ahead, and sometimes the mathematical theory was. We incorporate applications for the purpose of illustrating the theory and to motivate it. But then we cannot assume that the students know already the subjects in which calculus is applied, and it is also not our goal to teach them. For this reason the application have to be rather easy or simplified.

Chapter 0

A Preview

In this introductory course about calculus you will learn about two principal concepts, differentiation and integration. We would like to explain them in an intuitive manner using examples. In Figure 1 you see the graph of a function. Suppose it represents a function which describes the size of a

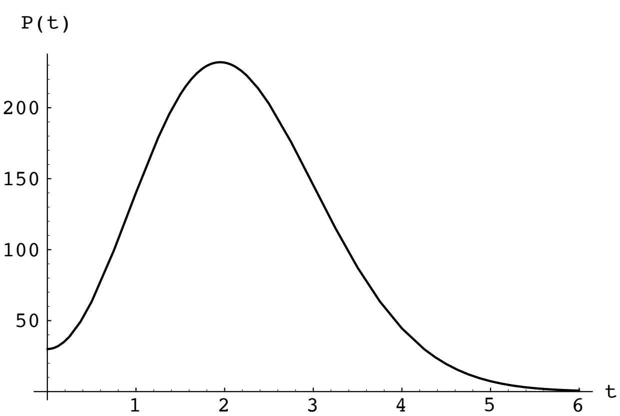

Figure 1: Yeast population as a function of time

population of live yeast bacteria in a bun of pizza dough. Abbreviating time by t (say measured in hours) and the size of the population by P (say measured in millions of bacteria), we denote this function by $P(t)$. You like to know at what rate the population is changing at some fixed time, say at time $t_0 = 4$.

- For a straight line, the rate of change is its slope.

We like to apply the idea of rate of change or slope also to the function $P(t)$, although its graph is certainly not a straight line.

What can we do? Let us try to replace the function $P(t)$ by a line $L(t)$, at least for values of t near t_0. The distance between the points $(t, P(t))$ and $(t, L(t))$ on the respective graphs is

$$E(t) = |P(t) - L(t)|. \tag{1}$$

This is the error which we make by using $L(t)$ instead of $P(t)$ at time t. We will require that this error is "small" in a sense which we will precise soon. If a line $L(t)$ can be found so that the error is small for all t in some open interval around t_0, then we call $L(t)$ the tangent line to the graph of P at t_0. The slope of the line $L(t)$ will be called the slope of the graph of $P(t)$ at the point $(t_0, P(t_0))$, or the rate of change of $P(t)$ at the time $t = t_0$.

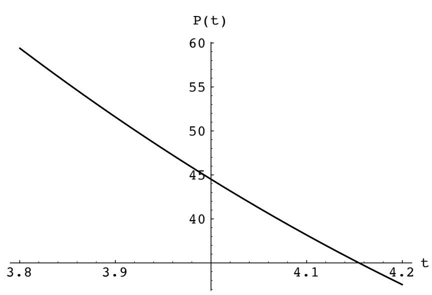

Figure 2: Zoom in on a point.

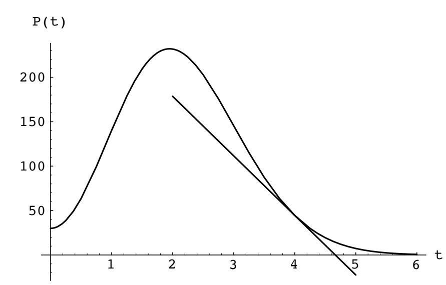

Figure 3: Graph & tangent line

Let us make an experiment. Put the graph under a microscope or, on your graphing calculator, zoom in on the point $(4, P(4))$ on the graph. This process works for the given example and most other functions treated in these notes. You see the zoom picture in Figure 2. Only under close scrutiny, you detect that the graph is not a line, but still bent. So, let us ignore this bit of bending and pretend that the shown piece of graph is a line. Actual measurements in the picture let you suggest that the slope of that line should be about -70. This translates into the statement that the population of the live bacteria decreases at a rate of roughly 70 million per hour. In Figure 3 we drew the actual tangent line to the graph of $P(t)$ at $t = 4$. A calculation based on the expression for $P(t)$, which you should be able to carry out only after having studied a good part of this manuscript, shows that the value of the slope of this line is about -67.0352. You may agree, that the geometric determination of the rate of change was quite accurate.

To some extent, it is up to us to decide the meaning of the requirement

- $|P(t) - L(t)|$ is small for all t near t_0.

One possible requirement[1], which it technically rather simple and which we will use, is:

- The exists a positive number A and an open interval (a, b) which contains t_0, such that

$$|P(t) - L(t)| \leq A(t - t_0)^2 \quad \text{for all } t \text{ in } (a, b). \tag{2}$$

The inequality in (2) dictates how close we require the graph of $P(t)$ to be to line $L(t)$. There may, or there may not, exist an interval and a number A such that the inequality holds for an appropriate line. If the line, the interval, and A exist, then the line is unique. Its slope is called the derivative of $P(t)$ at t_0, it is denoted by $P'(t_0)$, and we say that $P(t)$ is differentiable at t_0. Remembering that the rate of change of line $L(t)$ is its slope, we say

- If $P(t)$ is a function which is differentiable at t_0, then $P'(t_0)$ is, by definition, the rate at which $P(t)$ changes when $t = t_0$.

[1] In a standard treatment a weaker condition, which depends on the notion of limits, is imposed at this point. Our choice of requirement and our decision to avoid limits is based on the desire to keep the technicalities of the discussion at a minimum, and to make these notes as accessible as possible. Different interpretations of the word 'small' lead to different ideas about differentiability. More or fewer functions will be differentiable. The notion of the derivative, if it exists, is not effected by the choice of meaning for the word. On the other hand, the interpretation of the word 'small' has to imply the uniqueness of the derivative.

In due time we will explain all of this in more detail. You noticed that we need the idea of a line. When you look at (2) and see the square of the variable you can imagine that we need parabolas. So we review and elaborate on lines and parabolas in Chapter 1. We also introduce the, possibly, two most important functions in life science applications, the exponential function and the logarithm function.

Chapter 2 is devoted to the precise definition of the derivative and the exploration of related ideas. Relying only on the definition, we calculate the derivative for some basic functions. Then we establish the major rules of differentiation, which allow us to differentiate many more functions.

Chapter 3 is devoted to applications. We investigate the ideas of monotonicity and concavity and discuss the 1st and 2nd derivative tests for finding extrema of functions. In many applications of calculus one proceeds as follows. One finds a mathematical formulation for a problem which one encounters in some other context. One formulates the problem so that its solution corresponds to an extremum of its mathematical formulation. Then one resorts to mathematical tools for finding the extrema. Having found the solution for the mathematically formulated problem one draws conclusions about the problem one started out with.

E.g., look at a drop of mercury. Physical principles dictate that the surface area be minimized. You can derive mathematically that the shape of a body which minimizes the surface area, given a fixed volume, is a ball. This is roughly what you see. There is a slight perturbation due to the effect of gravity. This effect is much greater if you take a drop of water, for which the internal forced are not as strong as the ones in a drop of mercury.

Often calculus is used to solve differential equations. These are equations in which a relation between a function and its rate of change is given[2]. The unknown in the equation is the function. E.g., for some simple population models the equation (Malthusian Law)

$$P'(t) = aP(t)$$

is asserted. The rate at which the population changes ($P'(t)$) is proportional to the size of the population ($P(t)$). We solve this and some other population related differential equations. We will use both, analytical and numerical means.

The second principal concept is the one of the integral. Suppose you need to take a certain medication. Your doctor prescribes you a skin patch. Let

[2] In more generality, the relation may also involve the independent variable and higher derivatives.

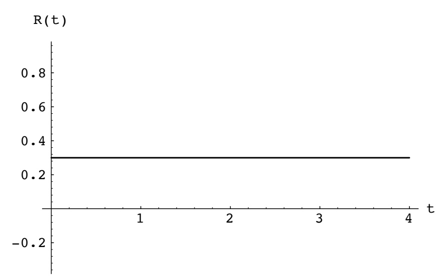

Figure 4: Constant Rate

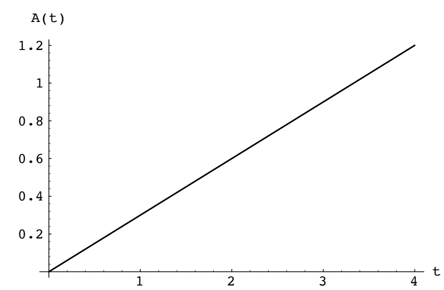

Figure 5: Amount absorbed

us say that the rate at which the medication is absorbed through the skin is a function $R(t)$, where R stands for rate and t for time. It is fair to say, that over some period of time $R(t)$ is constant, say .3 mg/hr. The situation is graphed in Figure 4. Over a period of three hours your body absorbs .9 mg of the medication. We multiplied the rate at which the medication is absorbed with the length of time over which this happened. Assuming that you applied the patch at time $t = 0$, the three hours would end at time $t = 3$. An interpretation of the total amount of medication which is absorbed between $t = 0$ and $t = 3$ is the area of the rectangle bounded by the line $t = 0$, the line $t = 3$, the x-axis, and the graph of the function $R(t) = .3$. Its side lengths are 3 and .3. In Figure 5 you see the function $A(t) = .3t$. It tells you, as a function of time, how much medication has been absorbed.

Suppose next that the medication is given orally in form of a pill. As the pill dissolves in the stomach, it sets the medication free so that your body can absorb it. The rate at which the medication is absorbed is proportional to the amount dissolved. As time progresses, the medication is moved through your digestive system, and decreasing amounts are available to being absorbed. A function which could represent the rate of absorption as a function of time is shown in Figure 6. We denote it once more by $R(t)$. Again you may want to find out how much medication has been absorbed within a given time, say within the first 4 hours after swallowing the pill. Set the time at which you took the pill as time $t = 0$. It should be reasonable to say (in fact a strong case can be made for this) that the amount of

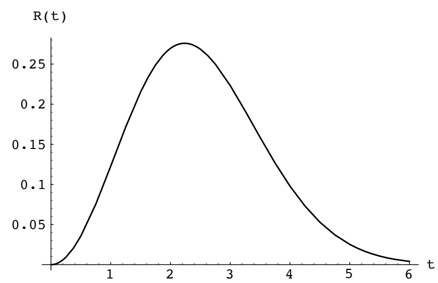

Figure 6: Time dependent rate

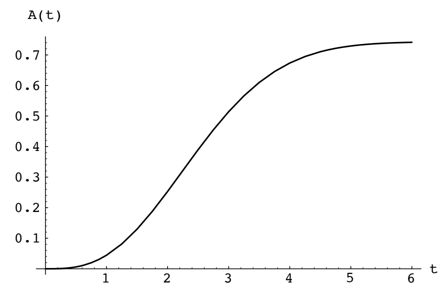

Figure 7: Amount absorbed

medication which has been absorbed between $t = 0$ and $t = T$ is the area under the graph of $R(t)$ between $t = 0$ and $t = T$. We denote this function by $A(T)$. Using methods which you will learn in this course, we found the function A. The graph is shown in Figure 7. You may find the value for $A(4)$ in the graph. A numerical calculation yields $A(4) = 0.6735$.

More generally, one may want to find the area under the graph of a function $f(x)$ between $x = a$ and $x = b$. To make sense out of this we first need to clarify what we mean when we talk about the area of a region, in particular if the region is not bounded by straight lines. Next we need to determine the areas of such regions. In fact, finding the area between the graph of a non-negative function f and the x-axis between $x = a$ and $x = b$ means to integrate f from a to b. Both topics are addressed in the chapter on integration.

The ideas of differentiation and integration are related to each other. If we differentiate the function shown in Figure 7 at some time t, then we get the function in Figure 6 at t. You will understand this after the discussion in Section 4.6. In this section we also discuss the Fundamental Theorem of Calculus, which is our principal tool to calculate integrals.

The two basic ideas of the rate of change of a function and the area below the graph of a function will be developed into a substantial body of mathematical results that can be applied in many situations. You are expected to learn about them, so you can understand other sciences where they are applied.

Chapter 1

Some Background Material

Introduction

In this chapter we review some basic functions such as lines and parabolas. In addition we discuss the exponential and logarithm functions for arbitrary bases. In a prior treatment you may only have been exposed to special cases.

Remark 1. Calculus (in one variable) is about functions whose domain and range are subsets of, or typically intervals in, the real line. So we will not repeat this assumption in every statement we make, unless we really want to emphasize it.

1.1 Lines

Lines in the plane occur in several contexts in these notes, and they are fundamental for the understanding of almost everything which follows. A typical example of a line is the graph of the function

$$y(x) = 2x - 3 \tag{1.1}$$

drawn in Figure 1.1. More generally, one may consider functions of the form

$$y(x) = mx + b \tag{1.2}$$

where m and b are real numbers. Their graphs are straight lines with slope m and y-intercept (the point where the line intersects the y axis) b. In the example the slope of the line is $m = 2$ and the y-intercept is $b = -3$. Even more generally than this, we have the following definition.

Definition 1.1. *A line consists of the points (x, y) in the $x-y$-plane which satisfy the equation*

$$ax + by = c \tag{1.3}$$

for some given real numbers a, b and c, where it is assumed that a and b are not both zero.

If $b = 0$, then we can write the equation in the form $x = c/a$, and this means that the solutions of the equation form a vertical line. The value for x is fixed, and there is no restriction on the value of y. Lines of this kind cannot be obtained if the line is specified by an equation as in (1.2). The line given by the equation $2x = 3$ is shown as the solid line in Figure 1.2.

If $a = 0$, then we can write the equation in the form $y = c/b$, and this means that the solutions of the equation form a horizontal line, the value for y is fixed, and there is no restriction on the value of x. The line given by the equation $2y = 5$ is shown as the dashed line in Figure 1.2.

If $b \neq 0$, then $ax + by = c$ translates into $y = -\frac{a}{b}x + \frac{c}{b}$, and the equation describes a line with slope $-a/b$ and y-intercept c/b.

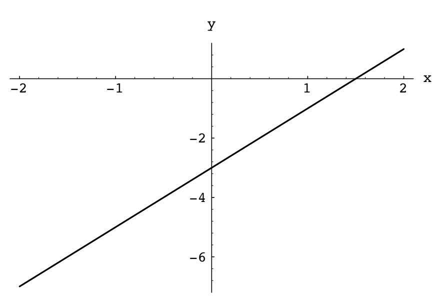

Figure 1.1: $y(x) = 2x - 3$

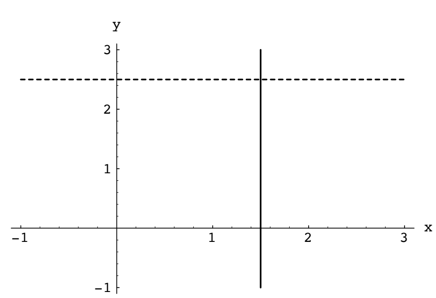

Figure 1.2: $2x = 3$ & $2y = 5$

Exercise 1. Sketch the lines $5x = 10$ and $3y = 5$.

Exercise 2. Sketch and determine the y-intercept and slope of the lines $3x + 2y = 6$ and $2x - 3y = 8$.

In application, we are often given the slope of a line and one of its points. Suppose the slope is m and the point on the line is (x_0, y_0). Then the line is given by the equation

$$y = m(x - x_0) + y_0.$$

Using functional notation, the line is the graph of the function

$$y(x) = m(x - x_0) + y_0. \tag{1.4}$$

To see this, observe that $y(x_0) = y_0$, so that the point (x_0, y_0) does indeed lie on the graph. In addition, you can rewrite the expression for the function in the form $y(x) = mx + (-mx_0 + b)$ to see that it describes a line with slope m. Its y-intercept is $-mx_0 + b$.

Example 1.2. The line with slope 3 through the point $(1, 2)$ is given by the equation

$$y = 3(x - 1) + 2. \quad \Diamond$$

Occasionally, we want to find the equation of a line through two distinct, given points (x_0, y_0) and (x_1, y_1). Assume that $x_0 \neq x_1$, otherwise the line is vertical. Set

$$y(x) = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0. \tag{1.5}$$

This is the point slope formula for a line through the point (x_0, y_0) with slope $\left[\frac{y_1 - y_0}{x_1 - x_0}\right]$. You should check that $y(x_1) = y_1$. This means that (x_1, y_1) is also a point on the line. In slope intercept form, the equation of the line is:

$$y(x) = \left[\frac{y_1 - y_0}{x_1 - x_0}\right] x + \left[-\frac{y_1 - y_0}{x_1 - x_0} x_0 + y_0\right].$$

Example 1.3. Find the equation of the line through the points $(x_0, y_0) = (1, -1)$ and $(x_1, y_1) = (3, 4)$.

Putting the points into the equation of the line, we find

$$y(x) = \left[\frac{4 - (-1)}{3 - 1}\right](x - 1) + (-1) = \frac{5}{2}x - \frac{7}{2}. \quad \Diamond$$

The line is shown in Figure 1.3. $\Diamond$

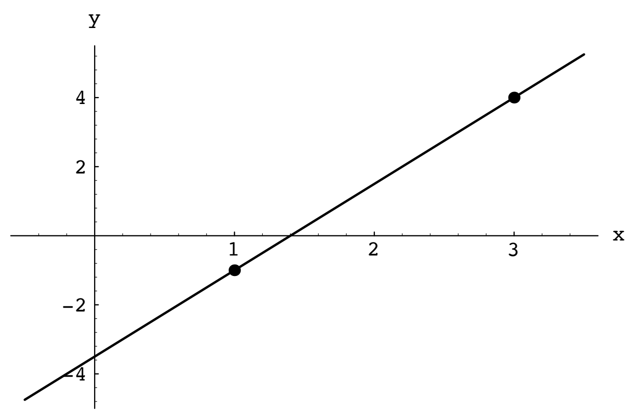

Figure 1.3: Line through $(1, -1)$ & $(3, 4)$

Summarizing the three examples, we ended up with three different ways to write down the equation of a non-vertical line, depending on the data which is given to us:

- **Intercept-Slope Formula:** We are given the y-intercept b and slope m of the line. The equation for the line is

$$y = mx + b.$$

- **Point-Slope Formula:** We are given a point (x_0, y_0) on the line and its slope m. The equation of the line is

$$y = m(x - x_0) + y_0.$$

- **Two-Point Formula:** We are given two points $(x_0.y_0)$ and (x_1, y_1) with different x-coordinate on the line. The equation of the line is

$$y(x) = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0.$$

Exercise 3. Suppose a line has slope 2 and $(2, 1)$ is a point on the line. Using the point $(2, 1)$, write down the point slope formula for the line and convert it into the slope intercept formula. Find the x and y-intercept for the line and sketch it.

Exercise 4. Find the point-slope and intercept-slope formula of a line with slope 5 through the point $(-1, -2)$.

Exercise 5. A line goes through the points $(-1, 1)$ and $(2, 5)$. Find the two point and slope intercept formula for the line. What is the slope of the line? Where does the line intersect the coordinate axes? Sketch the line.

Intersections of Lines

Let us discuss intersections of two lines. Consider the lines

$$l_1 : \ ax + by = c \quad \& \quad l_2 : \ Ax + By = C.$$

They intersect in the point (x_0, y_0) if this point satisfies both equations. I.e., to find intersection points of two lines we have to solve two equations in two unknowns simultaneously.

Example 1.4. Find the intersection points of the lines

$$2x + 5y = 7 \quad \& \quad 3x + 2y = 5.$$

Apparently, both equations hold if we set $x = 1$ and $y = 1$. This means that the lines intersect in the point $(1, 1)$. As an exercise you may verify that $(1, 1)$ is the only intersection point for these two lines. $\Diamond$

The lines $ax + by = c$ and $Ax + By = C$ are parallel to each other if

$$Ab = aB, \tag{1.6}$$

and in this case they will be identical, or they will have no intersection point.

Example 1.5. The lines

$$2x + 5y = 7 \quad \& \quad 4x + 10y = 14$$

are identical. To see this, observe that the second equation is just twice the first equation. A point (x, y) will satisfy one equation if and only if it satisfies the other one. A point lies on one line if and only if it lies on the other one. So the lines are identical. $\Diamond$

Example 1.6. The lines

$$2x + 5y = 7 \quad \& \quad 4x + 10y = 15$$

are parallel and have no intersection point.

To see this, observe that the first equation, multiplied with 2, is $4x + 10y = 14$. There are no numbers x and y for which $4x + 10y = 14$ and $4x + 10y = 15$ at the same time. Thus this system of two equations in two unknowns has no solution, and the two lines do not intersect. $\Diamond$

To be parallel also means to have the same slope. If the lines are not vertical ($b \neq 0$ and $B \neq 0$), then the condition says that the slopes $-a/b$ of the line l_1 and $-A/B$ of the line l_2 are the same. If both lines are vertical, then we have not assigned a slope to them.

If $Ab \neq aB$, then the lines are not parallel to each other, and one can show that they intersect in exactly one point. You saw an example above.

If $Aa = -bB$, then the lines intersect perpendicularly. Assuming that neither line is vertical ($b \neq 0$ and $B \neq 0$), the equation may be written as

$$\frac{a}{b} \times \frac{A}{B} = -1.$$

This means that the product of the slopes of the lines ($-a/b$ is the slope of the first line and $-A/B$ the one of the second line) is -1. The slope of one line is the negative reciprocal of the slope of the other line. This is the condition which you have probably seen before for two lines intersecting perpendicularly.

Example 1.7. The lines

$$3x - y = 1 \quad \& \quad x + 3y = 7$$

have slopes 3 and $-1/3$, resp., and intersect perpendicularly in $(x, y) = (1, 2)$. $\Diamond$

Exercise 6. Find the intersection points of the lines

$$l_1(x) = 3x + 4 \qquad \& \qquad l_2(x) = 4x - 5.$$

Sketch the lines and verify your calculation of the intersection point.

Exercise 7. Determine the slope for each of the following lines. For each pair of lines, decide whether the lines are parallel, perpendicular, or neither. Find all intersection points for each pair of lines.

$$\begin{aligned} l_1: \ 3x - 2y &= 7 \\ l_2: \ 6x + 4y &= 6 \\ l_3: \ 2x + 3y &= 3 \\ l_4: \ 6x - 4y &= 5 \end{aligned}$$

Exercise 8. Suppose a line $l(x)$ goes through the point $(1, 2)$ and intersects the line $3x - 4y = 5$ perpendicularly. What is the slope of the line? Find its slope point formula (use $(1, 2)$ as the point on the line) and its slope intercept formula. Sketch the line.

1.2 Parabolas and Higher Degree Polynomials

A parabola is the graph of a degree 2 polynomial, i.e., a function of the form

$$y(x) = ax^2 + bx + c \tag{1.7}$$

where a, b, and c are real numbers and $a \neq 0$. Depending on whether a is positive or negative the parabola will be open up- or downwards. Abusing language slightly, we say that $y(x)$ is a parabola. We will study parabolas in their own right, and they will be of importance to us in one interpretation of the derivative.

Typical examples of parabolas are the graphs of the functions

$$p(x) = x^2 - 2x + 3 \quad \text{and} \quad q(x) = -x^2 - x + 1$$

shown in Figures 1.4 and 1.5. The first parabola is open upwards, the second one downwards.

The x-intercepts of the graph of $p(x) = ax^2 + bx + c$ are also called *roots* or the *zeros* of $p(x)$. To find them we have to solve the quadratic equation

$$ax^2 + bx + c = 0.$$

The solutions of this equation are found with the help of the quadratic formula

$$p(x) = 0 \quad \text{if and only if} \quad x = \frac{1}{2a}\left[-b \pm \sqrt{b^2 - 4ac}\right]. \tag{1.8}$$

The expression $b^2 - 4ac$ under the radical is referred to as the *discriminant* of the quadratic equation. There are three cases to distinguish:

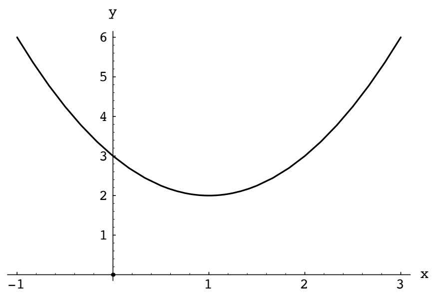

Figure 1.4: $y = x^2 - 2x + 3$

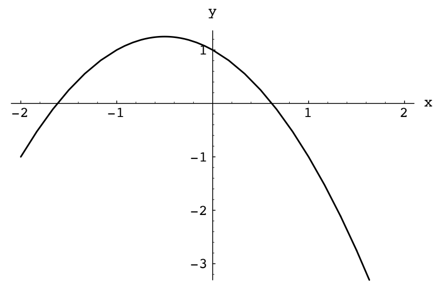

Figure 1.5: $y = -x^2 - x + 1$

- $p(x)$ has two distinct roots if the discriminant is positive.
- $p(x)$ has exactly one root if the discriminant is zero.
- $p(x)$ has no (real) root if the discriminant is negative.

Example 1.8. Find the roots of the polynomial $p(x) = 3x^2 - 5x + 2$.

According to the quadratic formula

$$3x^2 - 5x + 2 = 0 \quad \text{if and only if} \quad x = \frac{1}{6}\left[5 \pm \sqrt{25 - 24}\right].$$

So the roots of $p(x)$ are 1 and 2/3. $\Diamond$

Exercise 9. Find the roots of the following polynomials.

(1) $p(x) = x^2 - 5x + 2$ (3) $r(x) = 2x^2 - 12x + 18$

(2) $q(x) = 2x^2 + 3x - 5$ (4) $s(x) = -x^2 + 5x - 7$

Let us find the intersection points for two parabolas, say

$$p(x) = a_1x^2 + b_1x + c_1 \quad \text{and} \quad q(x) = a_2x^2 + b_2x + c_2.$$

To find their intersection points we equate $p(x)$ and $q(x)$. In other words, we look for the roots of

$$p(x) - q(x) = (a_1 - a_2)x^2 + (b_1 - b_2)x + (c_1 - c_2).$$

The highest power of x in this equation is at most 2 (this happens if $(a_1 - a_2) \neq 0$), and this means that it has at most two solutions.

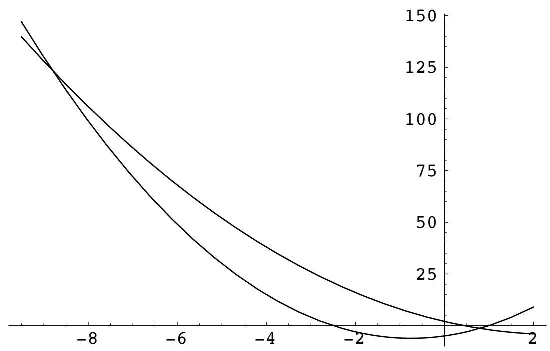

Figure 1.6: Intersecting parabolas

Example 1.9. Find the intersection points of the parabolas

$$p(x) = x^2 - 5x + 2 \quad \text{and} \quad q(x) = 2x^2 + 3x - 5.$$

We need to find the solutions of the equation

$$p(x) - q(x) = -x^2 - 8x + 7 = 0.$$

According to the quadratic equation, the solutions are

$$x = -\frac{1}{2}[8 \pm \sqrt{64 + 28}] = -4 \pm \sqrt{23}.$$

So the parabolas intersect at $x = -4 \pm \sqrt{23}$. You see the parabolas in Figure 1.6, and you can check that our calculation is correct. $\diamondsuit$

Exercise 10. Find the intersection points for each pair of parabolas from Exercise 9. Graph the pairs of parabolas and verify your calculation.

We will study how parabolas intersect in more detail in Section 2.5. Right now we like to turn our attention to a different matter. In Section 1.1 we used the slope-intercept and the point-slope formula to write down the equation of a line. The equation

$$y = mx + b = mx^1 + bx^0 \tag{1.9}$$

expresses y in powers of x. In the last term in (1.9) we added some redundant notation to make this point clear. When we write down the point slope formula of a line with slope m through the point (x_0, y_0),

$$y = m(x - x_0) + y_0 = m(x - x_0)^1 + y_0(x - x_0)^0,$$

then we expressed y in powers of $(x - x_0)$. The mathematical expression for this is that we *expanded* y in powers of $(x - x_0)$. We like to do the same for higher degree polynomials. We start out with an example.

Example 1.10. Expand the polynomial

$$y(x) = x^2 + 5x - 2 \tag{1.10}$$

in powers of $(x - 2)$.

Our task is to find numbers A, B, and C, such that

$$y(x) = A(x - 2)^2 + B(x - 2) + C. \tag{1.11}$$

Expanding the expression in (1.11) and gathering terms according to their power of x we find

$$\begin{aligned} y(x) &= A(x^2 - 4x + 4) + B(x - 2) + C \\ &= Ax^2 + (-4A + B)x + (4A - 2B + C) \end{aligned}$$

Two polynomials are the same if and only if their coefficients are the same. So, comparing the coefficients of y in (1.10) with those in our last expression for it, we obtains equations for A, B, and C:

$$\begin{aligned} A &= 1 \\ -4A + B &= 5 \\ 4A - 2B + C &= -2 \end{aligned}$$

These equations can be solved consecutively, $A = 1$, $B = 9$, and $C = 12$. So

$$y(x) = (x - 2)^2 + 9(x - 2) + 12.$$

We expanded $y(x)$ in powers of $(x - 2)$. $\Diamond$

Working through this example with general coefficients, we come up with the following formula:

$$y(x) = ax^2 + bx + c = A(x - x_0)^2 + B(x - x_0) + C. \tag{1.12}$$

where

$$\begin{aligned} A &= a \\ B &= 2ax_0 + b \\ C &= ax_0^2 + bx_0 + c = y(x_0) \end{aligned} \tag{1.13}$$

In fact, given any polynomial $p(x)$ and any x_0, one can expand $p(x)$ in powers of $(x - x_0)$. The highest power of x will be the same as the highest power of $(x - x_0)$. The process is the same as above, only it gets lengthier. On the computer you can do it in a jiffy.

Exercise 11. Expand $y(x) = x^2 - x + 5$ in powers of $(x - 1)$.

Exercise 12. Expand $y(x) = -x^2 + 4x + 1$ in powers of $(x + 2)$.

Exercise 13. Expand $y(x) = x^3 - 4x^2 + 3x - 2$ in powers of $(x - 1)$.

Exercise 14. Expand $p(x) = x^6 - 3x^4 + 2x^3 - 2x + 7$ in powers of $(x + 3)$.

What is the purpose of expanding a parabola in powers of $(x - x_0)$? Let us look at an example and see what it does for us. Consider the parabola

$$p(x) = 2x^2 - 5x + 7 = 2(x - 2)^2 + 3(x - 2) + 5.$$

The last two terms in the expansion form a line:

$$l(x) = 3(x - 2) + 5.$$

This line has an important property:

$$|p(x) - l(x)| = 2(x - 2)^2 \quad \text{and in particular,} \quad p(2) = l(2). \tag{1.14}$$

In the sense of the estimate suggested in (2) in the Preview, we found a line $l(x)$ which is close to the graph of $p(x)$ near $x = 2$. The constant A in (2) may be taken as 2 (or any number larger than 2), and the estimate holds for all x in $(-\infty, \infty)$ (or any interval).

Exercise 15. For each of the following parabolas $p(x)$ and points x_0, find a line $l(x)$ and a constant A, such that $|p(x) - l(x)| \leq A(x - x_0)^2$.

1. $p(x) = 3x^2 + 5x - 18$ and $x_0 = 1$.
2. $p(x) = -x^2 + 3x + 1$ and $x_0 = 3$.
3. $p(x) = x^2 + 3x + 2$ and $x_0 = -1$.

Let us do a higher degree example:

Example 1.11. Let $p(x) = x^4 - 2x^3 + 5x^2 - x + 3$ and $x_0 = 2$. Find a line $l(x)$ and a constant A, such that $|p(x) - l(x)| \leq A(x - x_0)^2$ for all x in the interval $I = (1, 3)$. (Note that the open interval I contains the point $x_0 = 2$.)

Expanding $p(x)$ in powers of $(x - 2)$ we find

$$p(x) = (x-2)^4 + 6(x-2)^3 + 17(x-2)^2 + 27(x-2) + 21.$$

Set $l(x) = 27(x - 2) + 21$. Then

$$\begin{aligned} |p(x) - l(x)| &= \left|(x-2)^4 + 6(x-2)^3 + 17(x-2)^2\right| \\ &= \left|(x-2)^2 + 6(x-2) + 17\right| (x-2)^2 \\ &\leq (1 + 6 + 17)(x-2)^2 \\ &\leq 24(x-2)^2. \end{aligned}$$

In the calculation we used the triangle inequality ((5.9) in Section 5.2 to get the first inequality. If $x \in (1, 3)$, then $|x - 2| < 1$ and $|x - 2|^k < 1$ for all $k \geq 1$. This helps you to verify the second inequality. So, with $A = 24$ and $l(x) = 27(x - 2) + 21$, we find that

$$|p(x) - l(x)| \leq A(x - x_0)^2$$

for all $x \in (1, 3)$. $\Diamond$

Exercise 16. Let $p(x) = 2x^4 + 5x^3 - 5x^2 - 3x + 7$ and $x_0 = 5$. Find a line $l(x)$ and a constant A, such that $|p(x) - l(x)| \leq A(x - x_0)^2$ for all x in the interval $I = (4, 6)$.

Remark 2. The general recipe (algorithm) for what we just did is as follows. Consider a polynomial

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0.$$

Pick a point x_0, and expand $p(x)$ in powers of x_0:

$$p(x) = C_n(x - x_0)^n + C_{n-1}(x - x_0)^{n-1} + \cdots + C_1(x - x_0) + C_0.$$

This can always be done, and we learned how to do this. Set

$$l(x) = C_1(x - x_0) + C_0.$$

Then

$$\begin{aligned}
|p(x) - l(x)| &= \left|C_n(x-x_0)^{n-2} + \cdots + C_3(x-x_0) + C_2\right|(x-x_0)^2 \\
&\le \left||C_n(x-x_0)^{n-2}| + \cdots + |C_3(x-x_0)| + |C_2|\right|(x-x_0)^2 \\
&\le (|C_n| + |C_{n-1}| + \cdots + |C_2|)\,(x-x_0)^2
\end{aligned}$$

for all $x \in I = (x_0 - 1, x_0 + 1)$. The details of the calculation are as follows. To get the equation, we took $|p(x) - l(x)|$ and factored out $(x - x_0)^2$. To get the first inequality we repeatedly used the triangle inequality, see (5.9) in Section 5.2. The last inequality follows as $(x - x_0)^k < 1$ if $k \ge 1$.

In summary, for $l(x) = C_1(x - x_0) + C_0$ and $A = (|C_n| + \cdots + |C_2|)$ we have seen that

$$|p(x) - l(x)| \le A(x - x_0)^2$$

for all $x \in (x_0 - 1, x_0 + 1)$. In the sense of our preview, and the upcoming discussion about derivatives, this means

- The rate of change of $p(x)$ at the point $(x_0, p(x_0))$ is C_1, the slope of the line $l(x)$.

Exercise 17. For each of the following polynomials $p(x)$ and points x_0, find the rate of change of $p(x)$ when $x = x_0$.

1. $p(x) = x^2 - 7x + 2$ and $x_0 = 4$.

2. $p(x) = 2x^3 + 3$ and $x_0 = 1$.

3. $p(x) = x^4 - x^3 + 3x^2 - 8x + 4$ and $x_0 = -1$.

Remark 3. You may have noticed, that we began to omit labels on the axes of graphs. One reason for this is, that we displayed more than one function in one graph, and that means that there is no natural name for the variable associated to the vertical axis.

Our general rule is, that we use the horizontal axis for the independent variable and the vertical one for the dependent one[1]. This is the rule which almost any mathematical text abides by. In some sciences this rule is reversed.

[1] If you like to review the terms independent and dependent variable, then we suggest that you read Section 5.3 on page 268.

1.3 The Exponential and Logarithm Functions

Previously you have encountered the expression a^x, where a is a positive real number and x is a rational number. E.g.,

$$10^2 = 100, \qquad 10^{1/2} = \sqrt{10}, \quad \text{and} \quad 10^{-1} = \frac{1}{10}$$

In particular, if $x = n/m$ and n and m are natural numbers, then a^x is obtained by taking the n-th power of a and then the m-root of the result. You may also say that $y = a^{m/n}$ is the unique solution of the equation

$$y^n = a^m.$$

By convention, $a^0 = 1$. To handle negative exponents, one sets $a^{-x} = 1/a^x$.

Exercise 18. Find exact values for

$$\left(\frac{1}{2}\right)^{-2} \qquad 4^{3/2} \qquad 3^{-1/2} \qquad 25^{-3/2}.$$

Exercise 19. Use your calculator to find approximate values for

$$3^{4.7} \qquad 5^{-.7} \qquad 8^{.1} \qquad .1^{-.3}.$$

Until now you may not have learned about irrational (i.e., not rational) exponents as in expressions like 10^π or $10^{\sqrt{2}}$. The numbers π and $\sqrt{2}$ are irrational. We like to give a meaning to the expression a^x for any positive number a and any real number x. A new idea is required which does not only rely on arithmetic. First, recall what we have. If $a > 1$ (resp., $0 < a < 1$) and x_1 and x_2 are two rational numbers such that $x_1 < x_2$, then $a^{x_1} < a^{x_2}$ (resp., $a^{x_1} > a^{x_2}$). We think of $f(x) = a^x$ as a function in the variable x. So far, this function is defined only for rational arguments (values of x). The function is monotonic. More precisely, it is increasing if $a > 1$ and decreasing if $0 < a < 1$.

Theorem-Definition 1.12. *Let a be a positive number, $a \neq 1$. There exists exactly one monotonic function, called the exponential function with base a and denoted by $\exp_a(x)$, which is defined for all real numbers x such that $\exp_a(x) = a^x$ whenever x is a rational number. Furthermore, $a^x > 0$ for all x, and so we use $(0, \infty)$ as the range[2] of the exponential function $\exp_a(x)$.*

[2] You may want to review the notion of the range of a function in Section 5.3 on page 268.

We will prove this theorem in Section 4.11. This will be quite easy once we have more tools available. Right now it would be a rather distracting tour-de-force. Never-the-less, the exponential function is of great importance and has many applications, so that we do not want to postpone its introduction. It is common, and we will follow this convention, to use the notation a^x for $\exp_a(x)$ also if x is not rational.

You can see the graph of an exponential function in Figures 1.7 and 1.8. We used $a = 2$ and two different ranges for x. In another graph, see Figure 1.9, you see the graph of an exponential function with a base a smaller than one. We can allowed $a = 1$ as the base for an exponential function, but $1^x = 1$ for all x, and we do not get a very interesting function. The function $f(x) = 1$ is just a constant function which does not require such a fancy introduction.

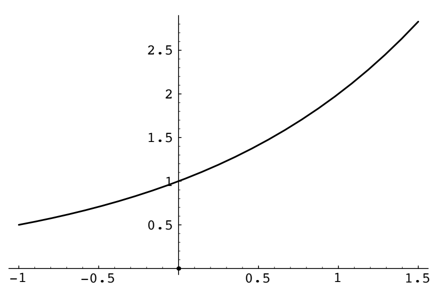

Figure 1.7: 2^x for $x \in [-1, 1.5]$

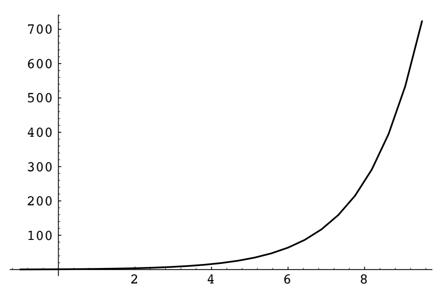

Figure 1.8: 2^x for $x \in [-1, 9.5]$

Let us illustrate the statement of Theorem 1.12. Suppose you like to find 2^π. You know that π is between the rational numbers 3.14 and 3.15. Saying that $\exp_2(x)$ is increasing just means that

$$2^{3.14} < 2^\pi < 2^{3.15}.$$

Evaluating the outer expressions in this inequality and rounding them down, resp., up, places 2^π between 8.81 and 8.88. In fact, if r_1 and r_2 are any two rational numbers, $r_1 < \pi < r_2$, then due to the monotonicity of the exponential function,

$$2^{r_1} < 2^\pi < 2^{r_2}.$$

The theorem asserts that there is at least one real number 2^π which satisfies these inequalities, and the uniqueness part asserts that there is only one number with this property, making 2^π unique.

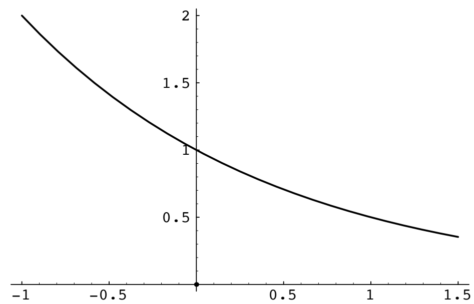

Figure 1.9: $(1/2)^x$

The arithmetic properties of the exponential function, also called the exponential laws, are collected in our next theorem. The theorem just says that the exponential laws, which you previously learned for rational exponents, also hold in the generality of our current discussion. You will derive the exponential laws from the logarithm laws later on in this section as an exercise.

Theorem 1.13 (Exponential Laws). *For any positive real number a and all real numbers x and y*

$$\begin{aligned} a^0 &= 1 \\ a^1 &= a \\ a^x a^y &= a^{x+y} \\ a^x/a^y &= a^{x-y} \\ (a^x)^y &= a^{xy} \end{aligned}$$

Some of the exponential laws can be obtained easily from the other ones. The second one holds by definition. Assuming the third one, one may deduce the first and third one. You are invited to carry out these deductions in the following exercises.

Exercise 20. Show: If $a \neq 0$, the $a^0 = 1$.

Although we did not consider an exponential function with base 0, it is common to set $0^0 = 1$. This is convenient in some general formulas. If $x \neq 0$, then $0^x = 0$.

Exercise 21. Assume $a^0 = 1$ and $a^x a^y = a^{x+y}$. Show $a^x/a^y = a^{x-y}$.

We need another observation about exponential functions, the proof of which we also postpone for a while (see Section 4.11).

Theorem 1.14. *Let a and b be positive real numbers and $a \neq 1$. There exists a unique (i.e., exactly one) real number x such that*

$$a^x = b.$$

You may make the uniqueness statement in the theorem more explicit by saying:

$$\text{If } a^x = a^y, \text{ then } x = y, \text{ or equivalently, if } x \neq y, \text{ then } a^x \neq a^y. \tag{1.15}$$

Let us consider some examples to illustrate the statement in the theorem. We assume that a and b are positive numbers and that $a \neq 1$. View the expression

$$a^x = b \tag{1.16}$$

as an equation in x. For a given a and b we want to (and the theorem says that we can) find a number x, so that the equation holds. E.g. if

$$\begin{array}{llll} a = 2 & \text{and} & b = 8, & \text{then} \quad x = 3. \\ a = 4 & \text{and} & b = 2, & \text{then} \quad x = 1/2. \\ a = 1/2 & \text{and} & b = 2, & \text{then} \quad x = -1. \\ a = \sqrt{2} & \text{and} & b = \pi, & \text{then} \quad x = 3.303. \end{array}$$

The value for x in the last example was obtained from a calculator and is rounded off.

Exercise 22. Solve the equation $a^x = b$ if

(1) $(a, b) = (10, 1000)$ (3) $(a, b) = (2, 4)$ (5) $(a, b) = (2, 1/4)$
(2) $(a, b) = (1000, 10)$ (4) $(a, b) = (4, 2)$ (6) $(a, b) = (100, .1)$.

For a given a ($a > 0$ and $a \neq 1$) and $b > 0$ we denote the unique solution of the equation in (1.16) by $\log_a(b)$. In other words:

Definition 1.15. *If a and b are positive numbers, $a \neq 1$, then $\log_a(b)$ is the unique number, such that*

$$a^{\log_a(b)} = b \quad \textit{or} \quad \exp_a(\log_a(b)) = b. \tag{1.17}$$

Here are some sample logarithms for the base 2:

$$\log_2 4 = 2 \quad \log_2 16 = 4 \quad \log_2(1/8) = -3 \quad \log_2 \sqrt{2} = 1/2$$

and for the base 10:

$$\log_{10} 1 = 0 \quad \log_{10} 100 = 2 \quad \log_{10}(1/10) = -1.$$

Your calculator will give you good approximations for at least $\log_{10}(x)$ for any $x > 0$.

Exercise 23. Find logarithms for the base 10:

(1) $\log_{10} 5$	(3) $\log_{10} \pi$	(5) $\log_{10} 25$
(2) $\log_{10} 100$	(4) $\log_{10}(1/4)$	(6) $\log_{10} 1$.

Mathematically speaking, we just defined a function. Let us express it this way.

Definition 1.16. *Let a be a positive number, $a \neq 1$. Mapping b to $\log_a(b)$ defines a function, called the* logarithm function with base a. *It is defined for all positive numbers, and its range is the set of real numbers.*

Part of the graph of $\log_2(x)$ is shown in Figure 1.10. In Figure 1.11 you see the graph of a logarithm function with base a less than 1.

We also like to see for every real number y that

$$\log_a(a^y) = y \quad \text{or} \quad \log_a(\exp_a(y)) = y. \tag{1.18}$$

Setting $b = a^y$ in (1.17) we have that

$$a^{\log_a(a^y)} = a^y.$$

The statement in (1.15) says that $\log_a(a^y) = y$.

Taken together, (1.17) and (1.18) say that for every $a > 0$, $a \neq 1$, we have

$$\begin{aligned} a^{\log_a(y)} &= y \quad \text{for all } y > 0 \text{ and} \\ \log_a(a^x) &= x \quad \text{for all } x \in (-\infty, \infty). \end{aligned}$$

This just means that

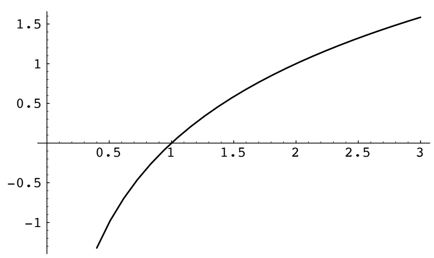

Figure 1.10: $\log_2(x)$

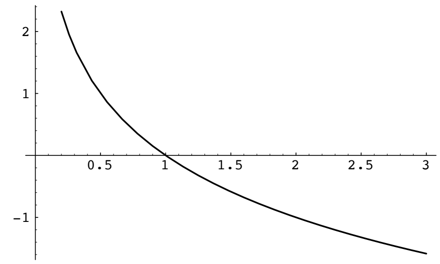

Figure 1.11: $\log_{(1/2)}(x)$

Theorem 1.17. *The exponential function* $\exp_a(x) = a^x$ *and the logarithm function* $\log_a(y)$ *are inverses*[3] *of each other.*

Using the same bases, we obtain the graph of the logarithm function by reflecting the one of the exponential function at the diagonal in the Cartesian plane. This is the general principle by which the graph of a function and its inverse are related. The role of the independent and dependent variables, and with this the coordinate axes, are interchanged. The graph of $\log_2(x)$, see Figure 1.10 is a reflection of the one in Figure 1.7. When you compare the two graphs, you need to take into account that the parts of the function shown are not quite the same and that there is a difference in scale. Once you make these adjustments you will see the relation.

Theorem 1.18. *Let a be a positive number, $a \neq 1$. The logarithm function* $\log_a$ *is monotonic. It is increasing if $a > 1$ and decreasing if $a < 1$. Suppose u and v are positive numbers. If* $\log_a(u) = \log_a(v)$*, then $u = v$, and equivalently, if $u \neq v$, then* $\log_a(u) \neq \log_a(v)$*.*

Proof. It is a general fact, that the inverse of an increasing function is increasing, and the inverse of a decreasing function is decreasing (see Proposition 5.25 on page 291). So the monotonicity statements for the logarithm functions follow from the monotonicity properties of the exponential functions (see Theorem 1.14) because these functions are inverses of each other.

[3]A quick review of the idea of inverse functions is given in Section 5.6 on page 286, and you are encouraged to read it in case you forgot about this concept.

Furthermore, $\log_a(u) = \log_a(v)$ implies that

$$u = a^{\log_a(u)} = a^{\log_a(v)} = v.$$

This verifies the remaining claim in the theorem. □

Corresponding to the exponential laws in Theorem 1.13 on page 22 we have the laws of logarithms. Some parts of the theorem are proved in Section 4.11. The other parts are assigned as exercises below.

Theorem 1.19 (Laws of Logarithms). *For any positive real number $a \neq 1$, for all positive real numbers x and y, and any real number z*

$$\begin{aligned}
\log_a(1) &= 0 \\
\log_a(a) &= 1 \\
\log_a(xy) &= \log_a(x) + \log_a(y) \\
\log_a(x/y) &= \log_a(x) - \log_a(y) \\
\log_a(x^z) &= z \log_a(x)
\end{aligned}$$

Because the exponential and logarithm functions are inverses of each other, their rules are equivalent. In the following exercises you are asked to verify this.

Exercise 24. Assume the exponential laws and deduce the laws of logarithms.

Exercise 25. Assume the laws of logarithms and deduce the exponential laws.

To show you how to solve this kind of problem, we deduce one of the exponential laws from the laws of logarithms. Observe that

$$\log_a(a^x a^y) = \log_a(a^x) + \log_a(a^y) = x + y = \log_a(a^{x+y}).$$

The first equation follows from the third equation in Theorem 1.19, and the remaining two equations hold because of the way the logarithm function is defined. Comparing the outermost expressions, we deduce from Theorem 1.18 the third exponential law:

$$a^x a^y = a^{x+y}.$$

Exercise 26. Assume that

$$\log_a 1 = 0 \quad \text{and} \quad \log_a(xy) = \log_a(x) + \log_a(y).$$

Show that

$$\log_a(x/y) = \log_a(x) - \log_a(y).$$

The Euler number e as base

You may think that $f(x) = 10^x$ is the easiest exponential function, at least you have no problems to find 10^n if n is an integer (a whole number). Later on you will learn to appreciate the use of a different base, the number e, named after L. Euler[4]. It is an irrational number, so the decimal expansion does not have a repeating block. Up to 50 decimal places e is

$$2.71828182845904523536028747135266249775724709369996. \tag{1.19}$$

A precise definition of e is given in Definition 4.61. The reason why $f(x) = e^x$ is such an interesting function will become clear in Theorem 2.12 on page 52 where it is stated that this function is its own derivative. If we talk about *the exponential function* then we mean the exponential function for this base. The inverse of this exponential function, the logarithm function for the base e, is called the *natural logarithm function.* It also has a very simple derivative, see Theorem 2.13 on page 52. For reference purposes, let us state the definitions formally. We graph these two functions on some reasonable intervals to make sure that you have the right picture in mind when we talk about them, see Figure 1.12 and Figure 1.13.

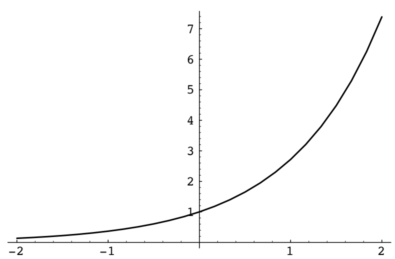

Figure 1.12: e^x for $x \in [-2, 2]$

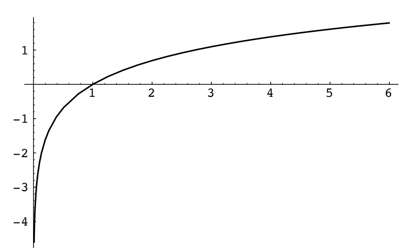

Figure 1.13: $\ln x$ for $x \in [.01, 6]$

Definition 1.20. *The exponential function is the exponential function for the base e. It is denoted by $\exp(x)$ or e^x. Its inverse is the natural logarithm function. It is denoted by $\ln(x)$.*

---

[4]Leonard Euler (1707–1783), one of the great mathematicians of the 18th century.

Exponential Functions grow fast.

Example 1.21 (Exponential Growth). It is not so apparent from the graph how fast the exponential function grows. You may remember the tale of the ancient king who, as payment for a lost game of chess, was willing to put 1 grain of wheat on the first square on the chess board, 2 on the second, 4 on the third, 8 on the forth, etc., doubling the number of grains with each square. The chess board has 64 squares, and that commits him to 2^{63} grains on the 64th square for a total of

$$2^{64} - 1 = 18,446,744,073,709,551,615$$

grains. In mathematical notation, you say that he puts

$$f(n) = 2^{n-1}$$

grains on the n-th square of the chess board. So, let us graph the function $f(x) = 2^x$ for $0 \leq x \leq 63$, see Figure 1.14. On the given scale in the graph, even an already enormous number like 2^{54}, cannot be distinguished from 0.

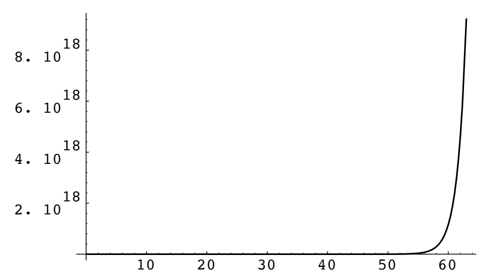

Figure 1.14: Graph of $f(x) = 2^x$

It is difficult to imagine how large these numbers are. The amount of grain which the king has to put on the chess board suffices to feed the current world population (of about 6 billion people) for thousands of years. $\Diamond$

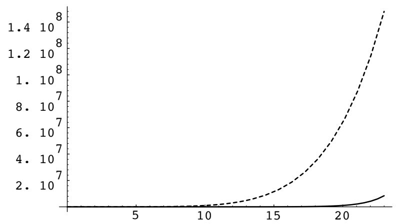

Figure 1.15: Compare 2^x and x^6.

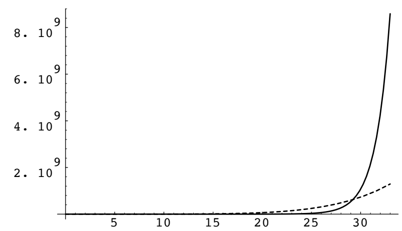

Figure 1.16: Compare 2^x and x^6.

Example 1.22 (Comparison with Polynomials). A different way of illustrating the growth of an exponential function is to compare it with the growth of a polynomial. In Figures 1.15 and 1.16 you see the graphs of an exponential function ($f(x) = 2^x$) and a polynomial ($p(x) = x^6$) over two different intervals, $[0, 23]$ and $[0, 33]$. In each figure, the graph of f is shown as a solid line, and the one of p as a dashed line. In the first figure you see that, on the given interval, the polynomial p is substantially larger than the exponential function f. In the second figure you see how the exponential function has overtaken the polynomial and begins to grow a lot faster. $\Diamond$

Other Bases

Finally, let us relate the exponential and logarithm functions for different bases. The result is, for any positive number a ($a \neq 1$),

Theorem 1.23.

$$a^x = e^{x \ln a} \quad \textit{and} \quad \log_a x = \frac{\ln x}{\ln a}.$$

Proof. This is seen quite easily. The first identity is obtained in the following way:

$$a^x = (e^{\ln a})^x = e^{x \ln a}.$$

To see the second identity, use

$$e^{\ln x} = x = a^{\log_a x} = (e^{\ln a})^{\log_a x} = e^{\ln a \log_a x}.$$

This means that $\ln x = (\ln a)(\log_a x)$, or $\log_a x = \frac{\ln x}{\ln a}$, as claimed. $\square$

Exponential Growth

Consider a function of the form

$$f(t) = Ce^{at}. \tag{1.20}$$

The constants C and a, and with this the function $f(t)$ itself, can be determined if we give the value of f at two points. We call a the *growth rate*[5]. We say that a function f grows exponentially if it has the form in (1.20).

Example 1.24. Suppose the function $f(t)$ grows exponentially, $f(0) = 3$, and $f(5) = 7$. Find the function f, its relative growth rate a, and the time t_0 for which $f(t_0) = 10$.

Solution: By assumption, the function is of the form $f(t) = Ce^{at}$. Substituting $t = 0$, we find

$$3 = f(0) = Ce^{a \cdot 0} = Ce^0 = C.$$

After having found $C = 3$, we substitute $t = 5$ into the expression of $f(t)$:

$$7 = f(5) = 3e^{5a}.$$

From this we deduce, using arithmetic and the fact that the natural logarithm function is the inverse of the exponential function, that

$$e^{5a} = 7/3 \qquad \& \qquad a = \frac{\ln(7/3)}{5} = .16946.$$

In particular, the growth rate of the function is (approximately) .16946, and $f(t) = 3e^{.16946t}$.

Finally, t_0 is determined by the equation

$$3e^{.16946t_0} = 10.$$

We calculate:

$$e^{.16946t_0} = 10/3 \quad \& \quad t_0 = \frac{\ln(10/3)}{.16946} = 7.105.$$

The value for t_0 is rounded off. $\diamond$

[5]Some texts call this number a the growth constant, others the relative growth rate. Actually, the rate of change of $f(t)$ at time t_0 is $af(t_0)$, so that the name relative growth rate (i.e., relative to the value to $f(t)$) is quite appropriate. Still, in the long run, you may get tired of having to say relative all the time, and with the exact meaning understood, you are quite willing to drop this adjective.

Exercise 27. Suppose the function $f(t)$ grows exponentially, $f(1) = 3$, and $f(4) = 7$. Find the function f, its relative growth rate a, and the time t_0 for which $f(t_0) = 10$.

Exercise 28. Suppose the function $f(t)$ grows exponentially, and $f(T) = 2f(0)$. Show that $f(t+T) = 2f(t)$ for any t.

Exercise 29. Suppose $f(t)$ describes a population of e-coli bacteria in a Petrie dish. You assume that the population grows exponentially. At time $t = 0$ you start out with a population of 800 bacteria. After three hours the population is 1900. What is the relative growth rate for the population? How long did it take for the population to double. How long does it take until the population has increased by a factor 4?

Remark 4. Some problems remain unresolved in this section. We still have justify our characterization of the exponential function in Theorem 1.12. We still have to prove two of the laws of logarithms from Theorem 1.19:

$$\log_a(xy) = \log_a(x) + \log_a(y) \quad \text{and} \quad \log_a(x^z) = z\log_a(x),$$

and we have to define the Euler number e. All of this will be done in Sections 4.11.

1.4 Use of Graphing Utilities

A word of caution is advised. We are quite willing to use graphing utilities, in our case *Mathematica*, to draw graphs of functions. We use these graphs to illustrate the ideas and concepts under discussion. They allow you to visualize situations and help you to understand them. For a number of reasons, no graphing utility is perfect and we cannot uncritically accept their output. When one of the utilities is pushed to the limit errors occur. Given any computer and any software, no matter how good they are, with some effort you can produce erroneous graphs. That is not their mistake, it only says that their abilities are limited.

In Figures 1.17 and 1.18 you see two graphs of the function

$$p(x) = (x+1)^6 = x^6 + 6x^5 + 15x^4 + 20x^3 + 15x^2 + 6x + 1.$$

Once we instructed the program to use the expression $(x+1)^6$ to produce the graph, and then we asked it to use the expanded expression. The outcome is remarkably different. Why? The program makes substantial round-off errors in the calculation. Which one is the correct graph? Calculus will

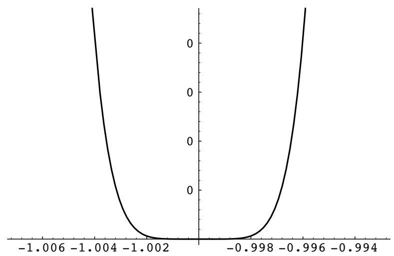

Figure 1.17: $p(x) = (x+1)^6$

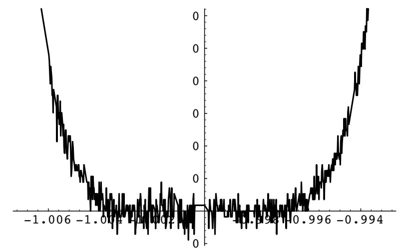

Figure 1.18: $p(x) = (x+1)^6$

tell you that the second graph cannot have come close to the truth. Is the first one correct? This is difficult to tell, particularly, as y values are indistinguishable. The program shows 0's at all ticks on this axis. True, the numbers are small, but they are certainly not zero. Still, the general shape of the graph in the first figure appears to be quite accurate.

On a smaller interval the results get even worse. You see what happens in Figures 1.19 and 1.20. The first graph is accurate in the sense that, given the scale shown on the axes, you should not see anything.

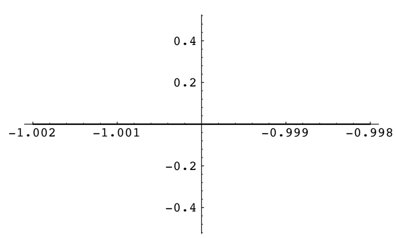

Figure 1.19: $p(x) = (x+1)^6$

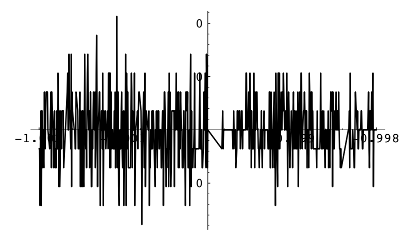

Figure 1.20: $p(x) = (x+1)^6$

When you use technology to assist you in graphing functions, then you have to make sure that the task does not exceed its abilities. Only experience and knowledge of the subject matter, in our case calculus, will help you. The process of using graphics is interactive. Graphs help you to understand calculus, but you need calculus to make sure the graphs are correct.

Chapter 2

The Derivative

The derivative is one of the most important tools in the study of graphs of functions, and with this the behavior of functions. Essentially, a function $f(x)$ is differentiable at a point x_0 if there is a line (the tangent line to the graph of f at x_0) which is close to the graph of the function for all x near x_0. The slope of this line will be called the derivative of f at x_0 and denoted by $f'(x_0)$. If the function is differentiable at all points in its domain, and with this $f'(x)$ is defined for all x in the domain of f, then we consider $f'(x)$ as a function and call it the derivative of $f(x)$.

We demonstrate this idea first with two examples. In the first example we use the exponential function to illustrate several ideas which enter into the general definition of the derivative. In the second example we take a geometric approach and interpret the tangent line in a special case. The geometry tells us what the derivative of the function is, or should be.

After these two examples we formally define the terms *differentiability* and *derivative*, see Definition 2.2 on page 43. The requirement that the difference between the graph and the tangent line is small is expressed analytically in terms of an inequality.

We give several geometric interpretations for this inequality. One interpretation places the function between an upper and a lower parabola on an open interval around the point at which the function is differentiable. The parabolas touch at this point. Another interpretation gives an estimate for the difference between the slope of the tangent line and secant lines through nearby points.

We calculate the derivatives of some basic functions based on the definition. For some other functions we provide the derivatives and postpone the calculation to a later point in this manuscript. We interpret the derivative at a point as the rate of change of the function at this point. Then we use the derivative to formulate and solve an easy, yet very important differential equation. A large part of this chapter is devoted to rules which allow us find the derivatives of composite functions, if the derivative of the constituents are known. We calculate many examples. We include a section on numerical methods for finding values (approximation by differentials) and zeros of functions (Newton's method), and on solving some differential equations (Euler's method). We close the chapter with a list of the rules of differentiation and a table of derivatives of important functions.

First Example

Consider the exponential function, $f(x) = e^x$, and the point $(1, e)$ on the graph. We would like to find the *tangent line* to the graph of f at this point. It is a straight line which close to the function near this point. A part of the graph of the exponential function is shown in Figure 2.1. In addition, we indicated the proposed tangent line.

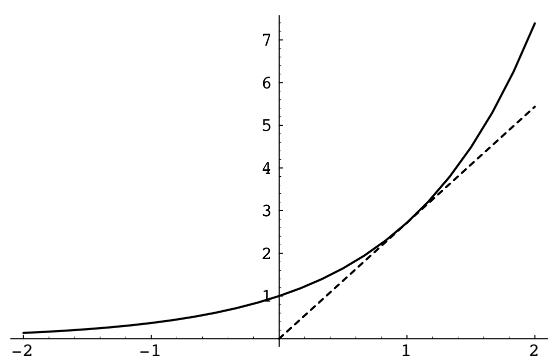

Figure 2.1: Graph of the Exponential Function

You can get a feeling for the tangent line by zooming in on the point. In Figure 2.2 you find a smaller piece of the graph of the exponential function. Take a ruler and see whether you can still distinguish the graph from a straight line. There is still a difference, but it is small. At least in the example we are rather successful. You should try this example and similar ones by yourself on a graphing calculator or a personal computer.

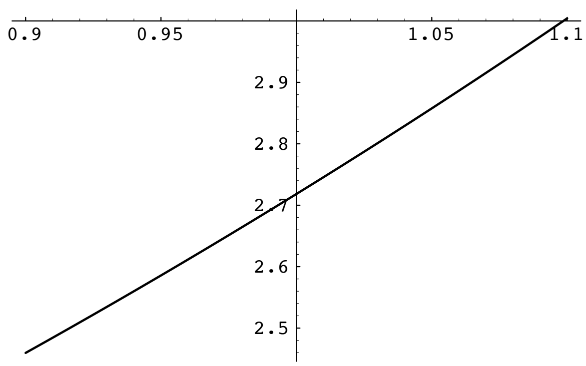

Figure 2.2: A Smaller Piece of the Graph of the Exponential Function

Our declared goal is to find a line which is close to the graph of $f(x) = e^x$ near the point $(1, e)$. In fact, the graph of f begins to look like a straight line when we look at it closely. So the graph of f itself suggests what line we ought to take. A measurement using Figure 2.2 suggests that, as x increases from .9 to 1.1, the values for $f(x)$ increase from 2.46 to 3.00. This means that the average rate of increase of f for x between .9 and 1.1 is, as precisely as we were able to measure it,

$$\frac{3.00 - 2.46}{1.1 - .9} = \frac{.54}{.2} = 2.7.$$

So the line which resembles the graph of f near $(1, e)$, and which we plan to call the tangent line, is supposed to go through the point $(1, e)$ and have slope 2.7, approximately. According to the point-slope formula of a line (see Section 1.1), such a line is given by the equation

$$l(x) = 2.7(x - 1) + e.$$

Let us denote the slope of the tangent line to the graph of f at the point $(x, f(x))$ by $f'(x)$. Later on we will call $f'(x)$ the derivative of f at x and interpret $f'(x)$ as the slope of graph of f at $(x, f(x))$. In the example you decided that $f'(1)$ is approximately 2.7. More exactly, $f'(1) = e$, as you will see later as a consequence of Theorem 2.12 on page 52. That means that the tangent line has the formula

$$l(x) = e(x-1) + e = ex.$$

Our goal is to find a line which is close to the graph, near a given point. So let us check how close $l(x)$ is to e^x if x is close to 1. In Table 2.1 you find the values of e^x and $l(x)$ for various values of x. You see that $e^x - l(x)$ is small, particularly for x close to 1. Let us compare $e^x - l(x)$ and $x - 1$ by taking their ratio $(e^x - l(x))/(x-1)$. As you see in the second last column of the table, even this quantity is small for x near 1. In other words, $e^x - l(x)$ is small compared to the distance of x from 1. Let say casually that $(x-1)^2$ is very small if $x - 1$ is small. The last column of the table suggests that $e^x - l(x)$ is roughly proportional to the very small quantity $(x-1)^2$.

x	e^x	$l(x)$	$e^x - l(x)$	$\frac{e^x - l(x)}{(x-1)}$	$\frac{e^x - l(x)}{(x-1)^2}$
2	7.389056	5.436564	1.952492	1.952492	1.952492
1.2	3.320117	3.261938	0.058179	0.290894	1.454468
1.1	3.004166	2.990110	0.014056	0.140560	1.405601
1.05	2.857651	2.854196	0.003455	0.069104	1.382079
1.01	2.745601	2.745465	0.000136	0.013637	1.363682

Table 2.1: Numerical Calculation for the Exponential Function

Exercise 30. Make a table like Table 2.1 for $f(x) = \ln x$ and $l(x) = x - 1$. More specifically, tabulate $f(x)$, $l(x)$, $f(x) - l(x)$, $(f(x) - l(x))/(x-1)$ and $(f(x) - l(x))/(x-1)^2$ for $x = 2$, 1.5, 1.2, 1.1 1.05 and 1.01.

Let us interpret the example geometrically. In Figure 2.3 you see the graph of the exponential function, which we denoted by $f(x)$. We used a solid line to draw it. There are two parabolas. One of them is open upwards and we call it $p(x)$, and the other one is open downwards and we call it $q(x)$.

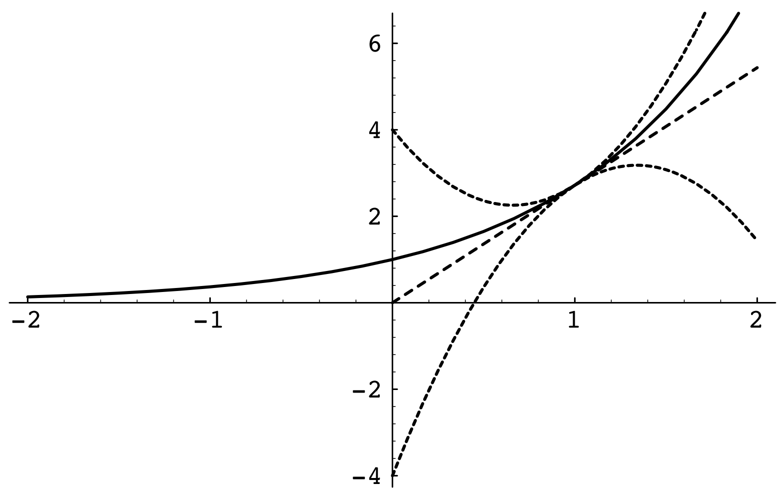

Figure 2.3: Exponential Function and Tangent Line between two Parabolas

We used short dashes to draw their graphs. These two parabola touch in the point $(1, e)$. In addition you see the line $l(x)$, which is our candidate for the being the tangent line.

On the shown interval, the graphs of $f(x)$ and $l(x)$ are above the graph of $q(x)$ and below the one of $p(x)$. In mathematical notation this is expressed as

$$q(x) \leq f(x) \leq p(x) \quad \text{and} \quad q(x) \leq l(x) \leq p(x).$$

One way of saying that $f(x)$ and $l(x)$ are close to each other near x_0 is to require that they are jointly in between two parabolas which touch (and do not cross each other) in the point $(x_0, f(x_0)) = (x_0, l(x_0))$. The 'hugging' behaviour of the parabolas shows that there is only little room in between them near x_0, and if $f(x)$ and $l(x)$ are both squeezed in between these parabolas, then the distance between $f(x)$ and $l(x)$ is small. As it turns out, only one line can be placed in between two parabolas as in the picture, and this line is the tangent line to the graph of $f(x)$ at $(x_0, f(x_0))$. The slope of the line $l(x)$ is e, so that the derivative of $f(x) = e^x$ at $x = 1$ is $f'(1) = e$. We will talk more about this geometric interpretation in Section 2.5.

Exercise 31. Use DfW (or any other accurate tool) to graph $f(x) = \ln x$, $l(x) = x - 1$, $p(x) = 2(x-1)^2 + x - 1$ and $q(x) = -2(x-1)^2 + x - 1$ on the interval $[.5, 1.5]$.

Second Example

Before we discuss the second example, let us think more about the tangent line. What is its geometric interpretation? Which line looks most like the graph of a function f near a point x. Sometimes (though not always) you can take a ruler and hold it against the graph. The edge of the ruler on the side of the graph gives you the tangent line. You find a line l which has the same value at x as f ($f(x) = l(x)$), and the line does not cross the graph of f (near x the graph of f is on one side of the line). This rather practical recipe for finding the tangent line of a differentiable function works for all functions in these notes at almost all points, see Remark 18 on page 164. It works in the previous example as well as in the one we are about to discuss.

For $x \in (-1, 1)$ we define the function

$$f(x) = y = \sqrt{1 - x^2}. \tag{2.1}$$

We like to use practical reasoning and a little bit of analytic geometry to show that

$$f'(x) = \frac{-x}{\sqrt{1 - x^2}}. \tag{2.2}$$

The function describes the upper hemisphere of a circle of radius 1 centered at the origin of the Cartesian coordinate system. To see this, square the equation and write it in the form $x^2 + y^2 = 1$, which is the equation of the circle. Thus we are saying that the slope of the tangent line to the circle at a point $(x, \sqrt{1 - x^2})$ in the upper hemisphere is $-x/\sqrt{1 - x^2}$. The circle and the tangent line are shown in Figure 2.4.

What is the slope of the tangent line to the circle at a point (x, y)? Your intuition is correct if you say that it is perpendicular to the radial line through the point $(0, 0)$, the origin of the Cartesian plane, and the point

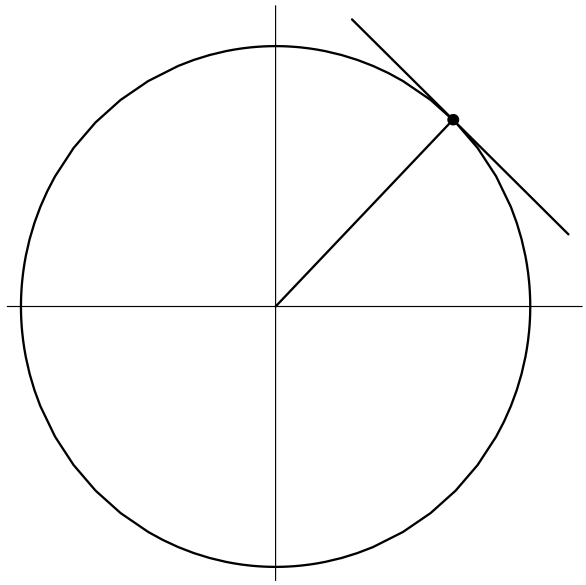

Figure 2.4: The radial line is perpendicular to the tangent line.

(x, y).[1] The slope of the radial line is y/x. In analytic geometry you (should have) learned that two lines intersect perpendicularly if the product of their slopes is -1. This means that the slope of the tangent line to the circle at the point (x, y) is $-x/y$. We called the slope of the tangent line to the graph of f at a point $(x, f(x))$ the derivative of f at x and we denoted it by $f'(x)$. Substituting $y = \sqrt{1 - x^2}$, we find that

$$f'(x) = -\frac{x}{y} = \frac{-x}{\sqrt{1 - x^2}}. \tag{2.3}$$

This is exactly the result predicted in the beginning of the discussion.

We will return to this example later (see Example 2.61) when we formally calculate the derivative of this specific function.

Exercise 32. In Figure 2.5 you see part of the graph of the function $f(x) = \sin x$. In this picture draw a line to resemble the graph near the point $(1, \sin 1)$. Determine the slope of the line which you drew. Write out the equation for this line in point slope form. Find $f'(1)$.

[1] You are encouraged to use geometric reasoning to come up with a justification of this statement. You may also measure the angle in the figure.

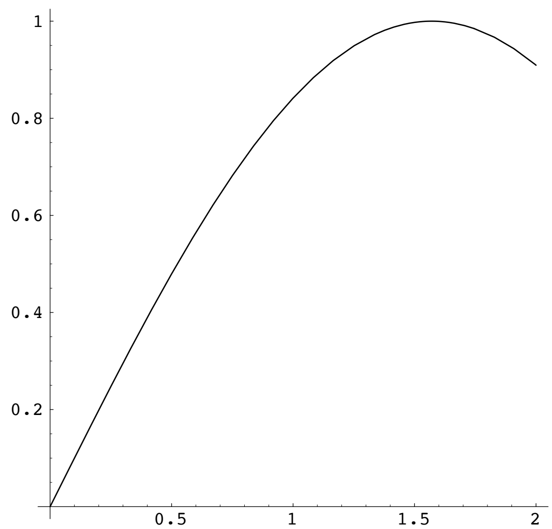

Figure 2.5: The graph of $\sin x$

Exercise 33. Use your graphing calculator, DfW, or any other means to get the graph of $f(x) = e^x$ near $x = 1/2$. You can also use one of the graphs of the exponential functions from these notes. Use the graph to estimate $f'(1/2)$, the slope of the tangent line at this point.

2.1 Definition of the Derivative

In both of the previous examples we were able to suggest the tangent line to the graph of a function at a point. In the first example we also discussed the idea of the tangent line being close to the graph. We discussed the idea numerically and in terms of a picture. The graph and the tangent line were squeezed between two parabolas. In our upcoming definition of differentiability, of the tangent line, and the derivative we will express 'closeness' analytically. We will use the absolute value. If you are not familiar with it, then you may want to read about it in Section 5.2. But, for the moment it suffices that you know that for any two real numbers a and b the absolute value of their difference, i.e., $|a-b|$, is the distance between these two points. Let us formalize the idea of an interior point.

Definition 2.1. *Let D be a subset of the real line, and x_0 an element in D. We say that x_0 is an* interior point *of D if D contains an open interval*

I and x_0 belongs to I.

Remember also that the domain of a function is the set on which it is defined.

Definition 2.2. *Let f be a function and x_0 an interior point of its domain. We say that f is* differentiable *at x_0 if there exists a line $l(x)$ and a real number A, such that*

$$|f(x) - l(x)| \leq A(x - x_0)^2 \tag{2.4}$$

for all x in some open interval which contains x_0.[2] *We call $l(x)$ the tangent line to the graph of f at x_0. We denote the slope of $l(x)$ by $f'(x_0)$ and call it the derivative of f at x_0. We also say that $f'(x_0)$ is the* slope *of the graph of f at x_0 and the* rate of change[3] *of f at x_0. To* differentiate *a function at a point means to find its derivative at this point.*

For this definition to make sense, it is important to observe that the derivative is unique (there is only one derivative of f at x_0), whenever it exists. This is stated in the following theorem. It is important to prove this theorem, but won't do this in these notes.

Theorem 2.3. *If f is as in Definition 2.2, and f is differentiable at x_0, then there exists only one line $l(x)$ for which* (2.4) *holds.*

Example 2.4. Consider the function $f(x) = \cos x$. We like to show

$$f'(0) = 0.$$

One can use elementary geometry to show (see (5.30)) that

$$|\cos x - 1| \leq \frac{1}{2}x^2$$

[2] To keep our approach simple, we have committed ourselves to the exponent 2 in (2.7). We could have taken any exponent α with $1 < \alpha \leq 2$. In fact, there won't be any essential change in our discussion of differentiability, with one exception. When we discuss the differentiability of inverse functions (see Section 2.11.4 and in particular Theorem 2.69), then we need that $f(x) = (bx + c)^{1/\alpha}$ is differentiable for those x for which $bx + c > 0$. This is somewhat more involved than the proof in the special case where $\alpha = 2$ where we consider $(bx + c)^{1/2} = \sqrt{bx + c}$, see Proposition 2.15. A disadvantage of using the exponent 2 is that the Fundamental Theorem of Calculus (see Theorem 4.32) will not be as generally applicable as we may want it to be. Still, we can take care of this matter when time comes.

[3] The interpretation of the derivative as rate of change has a concrete meaning which you understand better after reading Section 2.4.

for all $x \in (-\pi/4, \pi/4)$. So, setting $x_0 = 0$, $l(x) = 1$ and $A = 1/2$ we see that

$$|f(x) - l(x)| \leq A(x - x_0)^2$$

for all $x \in (-\pi/4, \pi/4)$. The slope of the line $l(x)$ is zero. So, according to our definition, f is differentiable at $x_0 = 0$ and $f'(0) = 0$. $\diamondsuit$

Exercise 34. Let $f(x) = \sin x$. Show that $f'(0) = 1$. Hint: Use the estimate $|\sin x - x| \leq x^2/2$ for all $x \in (-\pi/4, \pi/4)$ given in (5.30).

Let us explain how the requirement:

$$|f(x) - l(x)| \leq A(x - x_0)^2 \tag{2.5}$$

for all x in some open interval around x_0, expresses that the function $f(x)$ is close to its tangent line $l(x)$ on some interval around x_0. It does not hurt to take a numerical example. Suppose $A = 1$, it is merely a scaling factor anyway. If x is close to x_0, then $|x - x_0|$ is small, and $(x - x_0)^2$ is very small. If $|x - x_0| < .1$, then we are requiring that $|f(x) - l(x)| < .01$. If $|x - x_0| < .001$, then we are requiring that $|f(x) - l(x)| < .000001$.

Our first example is relevant and easy. We use it also to illustrate the idea and the definition of the derivative. Linear functions are functions whose graph is a line, and these are exactly the functions which are given by an equation of the form

$$l(x) = ax + b.$$

Example 2.5. Show that the linear function $l(x) = ax + b$ is differentiable everywhere, find its tangent line at each point on the graph, and show that

$$l'(x) = a,$$

for all real numbers x.

To make this general statement more concrete, you may replace the coefficients in the formula for $l(x)$ by numbers. Then you get special cases.

- If $l(x) = 3x + 5$, then $l'(x) = 3$.
- If $l(x) = c$, then $l'(x) = 0$.

A First Approach: Pick a point $(x_0, l(x_0))$ on the graph of $l(x)$. By design, the tangent line to a graph of $l(x)$ at $(x_0, l(x_0))$ is a line which is close to the graph of $l(x)$. Apparently there is a perfect choice, the line itself. So the tangent line to the graph of $l(x)$ at $(x_0, l(x_0))$ is $l(x)$. The derivative of a function at a point is, by definition, the slope of the tangent line at the point. In our case, this is the slope of the line itself. This slope is a, and we find $l'(x) = a$.

A Second Approach: Fix a point x_0. According to our definition, we need to find a line $t(x)$ (we use a different name for this line to distinguish it from our function $l(x)$) such that

$$|l(x) - t(x)| \leq A(x - x_0)^2 \tag{2.6}$$

for all x in some open interval around x_0. Setting $l(x) = t(x)$, the left hand side in (2.6) is zero, so that the inequality holds for any positive number A and all x. This means that $l(x)$ is differentiable at x_0, in fact at any x_0, and that $l(x)$ is its own tangent line. The slope of the line is a, and we find that $l'(x_0) = a$. $\diamondsuit$

Example 2.6. Show that polynomials are differentiable at each point x in $(-\infty, \infty)$. Find their derivatives.[4]

Solution: A polynomial is a function of the form

$$p(x) = c_n x^n + c_{n-1} x^{n-1} + \cdots + c_1 x + c_0.$$

Pick a point x_0. We like to show that $p(x)$ is differentiable at x_0, and find $p'(x_0)$.

We saw earlier, see Remark 2 in Section 1.2, that we can expand $p(x)$ in powers of x_0:

$$p(x) = C_n(x - x_0)^n + C_{n-1}(x - x_0)^{n-1} + \cdots + C_1(x - x_0) + C_0.$$

With

$$l(x) = C_1(x - x_0) + C_0 \quad \text{and} \quad A = (|C_n| + \cdots + |C_2|)\,.$$

we saw, that

$$|p(x) - l(x)| \leq A(x - x_0)^2$$

for all $x \in (x_0 - 1, x_0 + 1)$. This means that $l(x) = C_1(x - x_0) + C_0$ is the tangent line to the graph of $p(x)$ at the point $(x_0, p(x_0))$, and that $p'(x_0) = C_1$.

[4] We will find a more efficient way for differentiating a polynomial later.

Let us be more specific. Consider the polynomial

$$p(x) = 2x^4 - 5x^3 + 7x^2 - 3x + 1,$$

and expanded in powers of $(x - 2)$:

$$p(x) = 2(x - 4)^4 + 11(x - 2)^3 + 25(x - 2)^2 + 29(x - 2) + 15.$$

The tangent line to the graph of $p(x)$ at the point $(2, p(2)) = (2, 15)$ is $l(x) = 29(x - 2) + 15$. This line has slope 29, so that $p'(2) = 29$. $\diamondsuit$

Example 2.7. Find the derivative of the degree two polynomial

$$p(x) = ax^2 + bx + c$$

at the point x_0.

Solution: As an example, we earlier expanded degree 2 polynomials in powers of $(x - x_0)$. We found (see (1.13)) that

$$p(x) = ax^2 + bx + c = a(x - x_0)^2 + (2ax_0 + b)(x - x_0) + (ax_0^2 + bx_0 + c).$$

This means, the tangent line to the graph of $p(x)$ at the point $(x_0, p(x_0))$ is

$$l(x) = (2ax_0 + b)(x - x_0) + (ax_0^2 + bx_0 + c),$$

and the derivative (the slope of the tangent line) is $p'(x_0) = 2ax_0 + b$.

To give a numerical example, the tangent line to the graph of

$$p(x) = 5x^2 - 3x + 7$$

at $x_0 = 3$ is $l(x) = 27(x - 3) + 43$ and $p'(3) = 27$. $\diamondsuit$

Exercise 35. For the given polynomial $p(x)$ and point x_0, find the tangent line to the graph of $p(x)$ at the point $(x_0, p(x_0))$ and $p'(x_0)$.

1. $p(x) = 3x^2 - 4x + 3$
2. $p(x) = 7x^2 + 2x - 5$
3. $p(x) = x^3 - 3x^2 + 2x + 7$.

2.2 Differentiability as a Local Property

Differentiability is a local property of a function. This means, whether a function is differentiable at a point x_0 depends on the behaviour of the function on an open interval around x_0. It does not suffice to consider the function only at the point x_0, and it does not matter how the function looks like further away from x_0. The definition refers only to an open interval around x_0, and that interval can be chosen to be small, in fact, as small as we like as long as it contains points to the right and left of x_0. Typically, the estimate in (2.4) holds only on some interval around x_0, and not on the entire domain of the function.

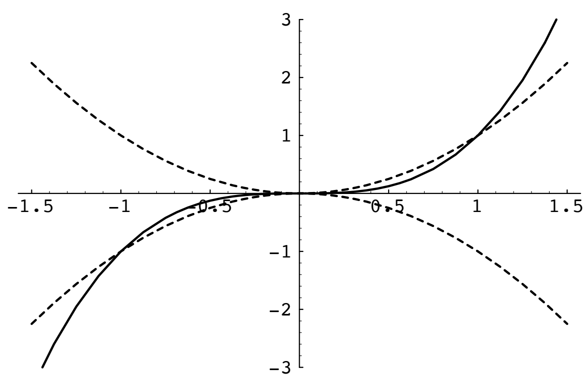

Figure 2.6: Estimates are local.

Let us illustrate this fact with an example. Consider the differentiability of the function $f(x) = x^3$ at $x_0 = 0$. We assert that the tangent line $l(x)$ to the graph of $f(x)$ at $(0, 0)$ is $l(x) = 0$. This is true because

$$|f(x) - l(x)| = |x^3 - 0| \leq (x - 0)^2$$

for all $x \in (-1, 1)$. We satisfied the requirement in Definition 2.2 with $A = 1$. On the other hand, for any $A > 0$ the inequality

$$|f(x) - l(x)| = A(x - 0)^2$$

holds only for $x \in [-A, A]$, and not for arbitrary x. You see this illustrated in Figure 2.6. The solid line is the graph of $f(x)$. The dashed lines are the graphs of $\pm x^2$. For the inequality to hold with $A = 1$, the solid line needs to be between the dashed lines, and this happens only for $x \in [-1, 1]$.

2.3 Derivatives of some Basic Functions

In this section we use the definition to find the derivatives of some basic functions. For some of them we can give a detailed argument, for others we have to postpone the justification. We collect the examples in Table 2.2 before we discuss them one by one. In Section 2.11 you will learn rules by which you can calculate the derivatives of composite functions. That will give you many more examples.

$y(x)$	$y'(x)$	Domain
$ax + b$	a	$x \in (-\infty, \infty)$
$\sin x$	$\cos x$	$x \in (-\infty, \infty)$
$\cos x$	$-\sin x$	$x \in (-\infty, \infty)$
e^{ax}	ae^{ax}	$x \in (-\infty, \infty)$
$\ln x$	$1/x$	$x \in (0, \infty)$
$ax^2 + bx + c$	$2ax + b$	$x \in (-\infty, \infty)$
$\sqrt{ax + b}$	$\frac{a}{2\sqrt{ax+b}}$	$x \in (-b/a, \infty)$ if $a > 0$
$\sqrt{ax + b}$	$\frac{a}{2\sqrt{ax+b}}$	$x \in (-\infty, -b/a)$ if $a < 0$

Table 2.2: Some Derivatives

We will encounter many functions which have a derivative at each point in their domain. This motivates the following definition.

Definition 2.8. *Let a function f be defined on an open interval (a, b). We say that f is* differentiable *on (a, b) (or differentiable for short) if the derivative $f'(x)$ exists for all $x \in (a, b)$. In this case we obtain a function f' which is defined for all $x \in (a, b)$ and which is called the* derivative *of f.*

If a function is defined on a union of open intervals, then we say that the function is differentiable if it is differentiable on each of the intervals.

Let us reformulate our Definition 2.2 in a less elegant but more practical way. Instead of saying "*for all x in some open interval around x_0*" we say "*for some $d > 0$ and all $x \in (x_0 - d, x_0 + d)$.*" Instead of asking for a line we ask for a number m, its slope, and use the line $l(x) = f(x_0) + m(x_0 - x)$. Then the definition reads this way:

Definition 2.9. *Let f be a function and x_0 an interior point of its domain. We say that f is* differentiable *at x_0 if there exist numbers m, A and $d > 0$, such that*

$$|f(x) - [f(x_0) + m(x - x_0)]| \leq A(x - x_0)^2 \tag{2.7}$$

for all x in the open interval $(x_0 - d, x_0 + d)$ around x_0.[5] *If $f(x)$ is differentiable at x_0, then the* tangent line *to the graph of f at x_0 is defined as the line given by the equation*

$$l(x) = f(x_0) + m(x - x_0). \tag{2.8}$$

We denote its slope m by $f'(x_0)$ and call it the derivative *of f at x_0. We also say that $f'(x_0)$ is the* slope *of the graph of f at x_0 and the* rate of change. *To* differentiate *a function at a point means to find its derivative at this point.*

We provide one more reformulation which makes some calculations look more elegant. For a fixed x_0 and any x we set

$$h = x - x_0. \tag{2.9}$$

With this notation the following three statements are equivalent:

(1) $x \in (x_0 - d, x_0 + d)$, (2) $h \in (-d, d)$, and (3) $|h| < d$.

The reformulation of Definition 2.9 using this notation looks as follows. Here $f(x)$ still denotes a function, and x_0 is assumed to be an interior point of its domain.

Definition 2.10. *We say that f is* differentiable *at x_0 if there exist numbers $f'(x_0)$, A and $d > 0$, such that*

$$|f(x_0 + h) - [f(x_0) + f'(x_0)h]| \leq Ah^2 \tag{2.10}$$

[5] We have to make sure that the left hand side of the inequality in (2.7) makes sense, i.e., that $f(x)$ is defined for all x in $(x_0 - d, x_0 + d)$. This can be assured by choosing d sufficiently small.

for all h for which $|h| < d$. The tangent line *to the graph of f at x_0 is defined as the line given by the equation*

$$l(x) = f(x_0) + f'(x_0)(x - x_0). \tag{2.11}$$

We call $f'(x_0)$ the derivative *of f at x_0.*

Example 2.11. Show that the derivative of the sine function is the cosine function, or, expressed in mathematical notation[6],

$$\sin' x = \cos x. \tag{2.12}$$

For this equation to hold, the angle x needs to be measured in radians.

Solution: We appeal to the definition of differentiability and the derivative as it is formulated in Definition 2.10. Fix a point x_0. Then we need to provide positive numbers A and d and show that

$$|\sin(x_0 + h) - [\sin x_0 + (\cos x_0)h]| \leq Ah^2 \tag{2.13}$$

for all h for which $|h| < d$. We do this for $A = 1$ and $d = \pi/4$.

As tools, we will use the inequalities (see (5.30))

$$|h - \sin h| \leq h^2/2 \qquad \text{and} \qquad |1 - \cos h| \leq h^2/2 \tag{2.14}$$

(they hold for $|h| < \pi/4$), and the trigonometric identity (see (5.19))

$$\sin(x + h) = \sin x \cos h + \sin h \cos x. \tag{2.15}$$

Furthermore, we need a few basic facts about absolute values. For any real numbers a, and b one has

$$\text{(i) } a \leq |a|, \quad \text{(ii) } |a + b| \leq |a| + |b|, \text{ and (iii) } |ab| = |a||b|.$$

In the first step of the upcoming calculation we use (2.15). The second step is basic arithmetic. In the third one we use the facts about working

[6]In some texts you will find this written as $(\sin x)' = \cos x$. We chose our notation in analogy with the symbol $f'(x)$. The notation does not really matter as long as it is interpreted correctly by the reader. Ambiguities and inconsistencies can be avoided if one writes: If $f(x) = \sin x$, then $f'(x) = \cos x$. This convention is used frequently, but it is somewhat wordy, so that a more compact expression is preferable. The problem arises, and our way of placing the prime to indicate the derivative does not work, if one tries to write down the derivative of a function like $f(x) = x^2$. The reader may and should not be bothered by this notational problem, though mathematicians will try hard to express themselves in a condensed manner, while staying precise and consistent.

with absolute values. In the forth step we use (2.14), and in the last one we use that $|\sin x_0|$ and $|\cos x_0|$ are ≤ 1.

$$\begin{aligned}
&|\sin(x_0+h) - [\sin x_0 + (\cos x_0)h]| \\
&\qquad = |\sin x_0 \cos h + \sin h \cos x_0 - \sin x_0 - (\cos x_0)h| \\
&\qquad = |\sin x_0(\cos h - 1) + \cos x_0(\sin h - h)| \\
&\qquad \leq |\sin x_0||\cos h - 1| + |\cos x_0||\sin h - h| \\
&\qquad \leq |\sin x_0|\frac{h^2}{2} + |\cos x_0|\frac{h^2}{2} \\
&\qquad \leq h^2.
\end{aligned}$$

This completes the verification of (2.13) with our chosen A and d. In particular, we have shown that the sine function is differentiable and that its derivative is the cosine function.

Let us look at the specific value $x_0 = \pi/4$. You can find the numerical value of $\cos(\pi/4)$ using elementary geometry, or you may look it up in Table 5.3 on page 280. Our formula says that

$$\sin'(\pi/4) = \cos(\pi/4) = \sqrt{2}/2.$$

The tangent line to the graph of the function $\sin x$ at $x_0 = \pi/4$ is given by the equation

$$l(x) = \frac{\sqrt{2}}{2}\left(x - \frac{\pi}{4}\right) + \frac{\sqrt{2}}{2}.$$

The slope of this tangent line (resp., the rate of change of $\sin x$ at the point $x_0 = \pi/4$) is $\sqrt{2}/2$. $\diamondsuit$

Exercise 36. Find the tangent line to the graph of $\sin x$ at the point $(\pi/6, 1/2)$.

Exercise 37. Show that

$$\cos' x = -\sin x.$$

Hint: The calculation is pretty much like the one in Example 2.11. Use $l(x) = -\sin(x_0)(x - x_0) + \cos x_0$ as the proposed tangent line to the graph of $\cos x$ at $(x_0, \cos x_0)$. Instead of the trigonometric identity (5.19), use the corresponding formula for $\cos x$ (see (5.21)):

$$\cos(x_0 + h) = \cos x_0 \cos h - \sin x_0 \sin h.$$

Theorem 2.12. *Let a and c be constants. The function $f(x) = ce^{ax}$ is differentiable at all x, and $f'(x) = ace^{ax}$. Furthermore, functions of the form $f(x) = ce^{ax}$ are the only functions which satisfy the equation $f'(x) = af(x)$.*

Exercise 38. Show that the exponential function $\exp x = e^x$ is its own derivative.

At this point we are not in the position to prove either statement in the theorem. For the time being we need to accept the theorem as a fact. In Example 2.70 we will show that the exponential function is differentiable, and that it is its own derivative. We will assume Theorem 2.13, which is stated next. The claim that multiples of this exponential are the only solutions of the equation $f(x) = ce^{ax}$ is shown in Section 3.2.

You may find it enlightening to review the data which we presented in Table 2.1. Essentially, we looked at numerical evidence that $f(x) = e^x$ is differentiable at $x = 1$, and that the tangent line at this point is $l(x) = ex$. The theorem says that $f'(1) = e$, and this is the slope of the line $l(x)$. The last column in the table gives evidence that

$$|e^x - l(x)| \leq 2(x-1)^2$$

for $1 \leq x \leq 2$. Using the formulation of differentiability as in (2.4), these two statements are consistent.

We defined the natural logarithm function $\ln x$ in Definition 1.20 on page 27. We will see later on (more precisely, we will use as definition, see Definition 4.58 and Theorem 4.59):

Theorem 2.13. *The natural logarithm function is differentiable at all the points in its domain $(0, \infty)$, and*

$$\ln'(x) = 1/x$$

Exercise 39. Find the tangent line to the graph of $\ln x$ at the point $(1, 0)$.

By now you may have gotten the impression that all functions are differentiable. This is not so.

Example 2.14. Show that the absolute values function (for a graph see Figure 2.7)

$$f(x) = |x| = \begin{cases} x & \text{for } x \geq 0 \\ -x & \text{for } x \leq 0 \end{cases}$$

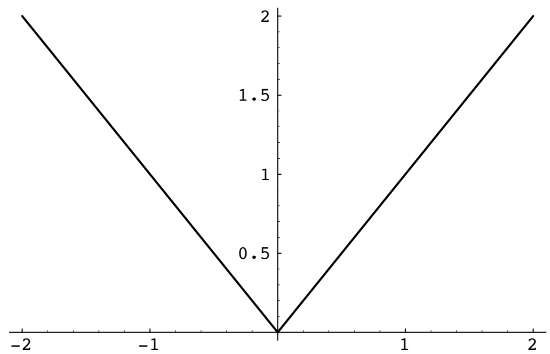

Figure 2.7: The absolute value function

is not differentiable at $x = 0$.

Solution: There is no potential tangent line which is close to the graph of $f(x)$ near $x = 0$ in the sense in which it has been specified in the definition of differentiability. Zooming in on the point $(0, 0)$ does not help, the picture remains the same.

You can give an analytical argument. If there is a tangent line, then it has to be of the form $l(x) = bx$, as it has to go through the point $(0, 0)$. Let A be any positive number. The estimate in Definition 2.10 applied in our context becomes

$$||h| - bh| \leq Ah^2. \tag{2.16}$$

For $h > 0$ this translated into $|1-b| \leq A|h|$, and for $h < 0$ into $|1+b| \leq A|h|$. If $b \neq 1$, then the first inequality is violated for some h of sufficiently small absolute value. If $b \neq -1$, then the second inequality is violated for some h of sufficiently small absolute value. That means that we cannot satisfy (2.16) for any number A and all h in some open interval around 0. So the absolute value function is not differentiable at $x = 0$. $\diamondsuit$

There is a last example of a function for which we like to find the derivative by hand. We formulate it as a Proposition. It is of intrinsic importance to our approach. It is essential to the proof of Theorem 2.69 on page 104.

Proposition 2.15. *The function $g(x) = \sqrt{bx+c}$ is defined for all real numbers x for which $bx + c \geq 0$. This function is differentiable for all x for which $bx + c > 0$, and the derivative is*

$$g'(x) = \frac{b}{2\sqrt{bx+c}}.$$

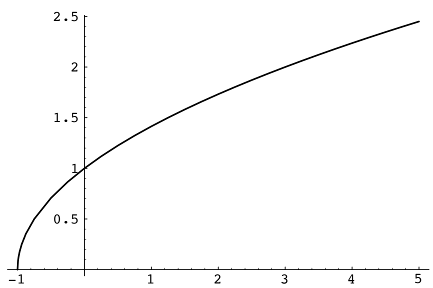

Figure 2.8: $g(x) = \sqrt{x+1}$

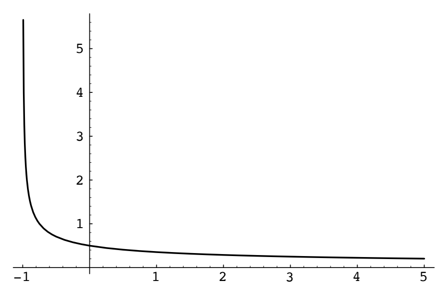

Figure 2.9: $g'(x) = 1/(2\sqrt{x+1})$

In Figure 2.8 you see the graph of the function $g(x) = \sqrt{x+1}$, and in Figure 2.9 the graph of its derivative $g'(x) = \frac{1}{2\sqrt{x+1}}$.

Let us give another concrete

Example 2.16. Find the domain and the derivative of the function $f(x) = \sqrt{5x-3}$.

Solution: The function is defined whenever $5x - 3 \geq 0$, and this means that $x \in [3/5, \infty)$. The derivative of the function is

$$f'(x) = \frac{5}{2\sqrt{5x-3}}.$$

The expression for the derivative holds for $x \in (3/5, \infty)$. $\diamondsuit$

The expression which defines $g(x)$ is a real number only if the term under the radical sign is non-negative, which means that we have to make the assumption that $bx+c \geq 0$. If $b > 0$, then this means that $g(x)$ is defined for all x in $[-c/b, \infty)$, and differentiable at all $x \in (-c/b, \infty)$. If $b < 0$, then this means that $g(x)$ is defined for all x in $(-\infty, -c/b]$, and differentiable at all x in $(-\infty, -c/b)$. If you have difficulties with the verification, then you may want to review the rules for calculating with inequalities from Section 5.2 on page 266. The borderline case in which $b = 0$ and $c \geq 0$ leads to a constant function with zero derivative. In case $b = 0$ and $c < 0$ the function is not defined, or, in other words, there is no x for which the function is defined.

Exercise 40. For each of the following functions, decide where the function is defined and where it is differentiable, and find the expression for the derivative.

$$(1)\ f(x) = \sqrt{2x+5} \quad (2)\ f(x) = \sqrt{-3x+4} \quad (3)\ f(x) = \sqrt{7x-2}.$$

Proof of Proposition 2.15. We only treat the case $g(x) = \sqrt{x}$. This special case, together with the chain rule, implies the general case, see Example 2.48.

We fix a value for $x > 0$. Using the formulation of the inequality which defines differentiability in (2.11)[7], we need to find positive numbers d and A, such that

$$|g(x+h) - [g(x) + g'(x)h]| \leq Ah^2$$

or, explicitly,

$$\left|\sqrt{x+h} - \left[\sqrt{x} + \frac{h}{2\sqrt{x}}\right]\right| \leq Ah^2 \tag{2.17}$$

whenever $|h| < d$.

It is a little tricky and takes some work to come up with values for A and d, and you are not expected to develop great skills at this. If you use

$$A = \frac{1}{2(\sqrt{x})^3}, \tag{2.18}$$

then we claim that the inequality in (2.17) holds as long as $x \in (0, \infty)$ and $|h| < d = x$. With this choice of d it is assured that $x + h \in (0, \infty)$ and that $g(x+h)$ is defined. This is all we will need. We hope that you can recognize the steps in the following calculation. It is a challenge.

[7] We use x instead of x_0.

$$\begin{aligned}
\left|\sqrt{x+h}-\left[\sqrt{x}+\frac{h}{2\sqrt{x}}\right]\right| &= \left|\sqrt{x+h}-\sqrt{x}-\frac{h}{2\sqrt{x}}\right| \\
&= \left|\frac{(x+h)-x}{\sqrt{x+h}+\sqrt{x}}-\frac{h}{2\sqrt{x}}\right| \\
&= |h|\left|\frac{1}{\sqrt{x+h}+\sqrt{x}}-\frac{1}{2\sqrt{x}}\right| \\
&= |h|\left|\frac{2\sqrt{x}-(\sqrt{x+h}+\sqrt{x})}{2\sqrt{x}(\sqrt{x+h}+\sqrt{x})}\right| \\
&= |h|\left|\frac{\sqrt{x}-\sqrt{x+h}}{2\sqrt{x}(\sqrt{x+h}+\sqrt{x})}\right| \\
&\leq |h|\left|\frac{\sqrt{x}-\sqrt{x+h}}{2x}\right| \\
&= |h|\left|\frac{x-(x+h)}{2x(\sqrt{x}+\sqrt{x+h})}\right| \\
&= h^2\left|\frac{1}{2x(\sqrt{x}+\sqrt{x+h})}\right| \\
&\leq h^2\left|\frac{1}{2x\sqrt{x}}\right| \\
&= Ah^2.
\end{aligned}$$

With this we verified (2.17) and completed the proof of the proposition in the stated special case. □

Exercise 41. Prove Proposition 2.15 directly for any $b > 0$ and c. Hint: One may use the road map of the calculation which we just went through. The expressions just get a bit bigger.

Using the formulation of the inequality which defines differentiability in (2.11), you need to find positive numbers d and A, such that

$$|g(x+h)-[g(x)+g'(x)h]| \leq Ah^2$$

or, explicitly,

$$\left|\sqrt{b(x+h)+c}-\left[\sqrt{bx+c}+\frac{bh}{2\sqrt{bx+c}}\right]\right| \leq Ah^2 \tag{2.19}$$

whenever $|h| < d$. Use

$$A = \frac{b^2}{2(\sqrt{bx+c})^3},$$

then the inequality in (2.19) holds as long as $x \in (-c/b, \infty)$ and $|h| < d = |x + c/b|$. With this choice of d it is assured that $x + h \in (-c/b, \infty)$ and that $g(x+h)$ is defined. Good luck!

2.4 Slopes of Secant Lines and Rates of Change

Let us compare the derivative with the result of another, geometric construction. Consider a function f which is defined on an open interval (a, b). Let x_0 be a point in the interval. For all $x \in (a, b)$, $x \neq x_0$, we can draw a line through the points $(x_0, f(x_0))$ and $(x, f(x))$. It is called the *secant line* through these two points. The slope of the secant line is

$$\frac{f(x) - f(x_0)}{x - x_0}.$$

and we call it the *average rate of change* of $f(x)$ over the interval with endpoints x_0 and x.

You see this idea illustrated in Figure 2.10. On the graph you see two points, $(x, f(x)$ and $(x_0, f(x_0))$. We also indicated x and $f(x)$ along the axes. The straight line is the secant line. Its slope is the average rate of change of the function over the interval $[x_0, x]$.

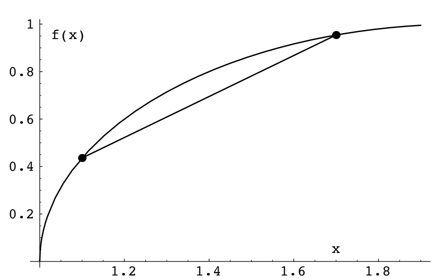

Figure 2.10: The function $f(x)$ and one secant line.

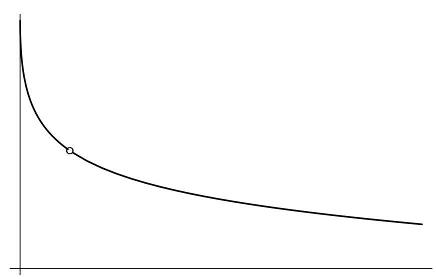

Figure 2.11: The slopes of secant lines, the function $g(x)$.

Exercise 42. Find the average slope of the function $\sin x$ over the interval $[\pi/6, \pi/3]$.

Exercise 43. Find the average slope of the function $\ln x$ over the interval $[2, 15]$.

Keeping x_0 fixed and allowing x to vary, we may consider the slope of the secant line through the points $(x_0, f(x_0))$ and $(x, f(x))$ as a function of x. The expression for this function is

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0}. \tag{2.20}$$

This function is defined for all $x \in (a, b)$ for which $x \neq x_0$. In Figure 2.11 you see the graph of this function, where $f(x)$ is the function shown in Figure 2.10. The little empty circle indicates where the function is not defined, i.e., where $x = x_0$.

Here is a concrete example. It is not the one shown in the figures.

Example 2.17. If $f(x) = x^2$, and we fix x_0, then

$$g(x) = \frac{x^2 - x_0^2}{x - x_0} = \frac{(x - x_0)(x + x_0)}{x - x_0} = x + x_0.$$

Exercise 44. Suppose $f(x) = x^3$, and you fix x_0. Simplify the expression for the function

$$g(x) = \frac{f(x) - f(x_0)}{x - x_0}. \tag{2.21}$$

Hint: Use long division to calculate $(x^3 - x_0^3)/(x - x_0)$.

Suppose now that f is differentiable at x_0. This means that there exists a positive number A such that

$$|f(x) - [f(x_0) + f'(x_0)(x - x_0)]| \leq A(x - x_0)^2 \tag{2.22}$$

for all x in some open interval I around x_0 (see Definition 2.2 on page 43). Dividing (2.22) by $x - x_0$ we find:

$$|g(x) - f'(x_0)| = \left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| \leq A|x - x_0| \tag{2.23}$$

for all $x \in I$, $x \neq x_0$.

This means that, for a differentiable function f, the slope of the tangent line at a point is approximately the slope of the secant line through a nearby point. More precisely, the inequality in (2.23) tells us how small the difference between the slope of the tangent line and the slopes of secant lines through points $(x_0, f(x_0))$ and $(x, f(x))$ must be as a function of $(x - x_0)$, as long $x \in I$. It cannot exceed $A|x - x_0|$. We summarize this discussion as

Theorem 2.18. [8] *Suppose f is a function which is differentiable at x_0. There exist positive numbers A and an open interval I around x_0 such that, when $g(x)$ is the slope of the secant line through $(x_0, f(x_0))$ and $(x, f(x))$, then*

$$|g(x) - f'(x_0)| \le A|x - x_0| \quad \textit{for all } x \in I,\ x \ne x_0.$$

We discuss an example to illustrate the theorem.

Example 2.19. Let $f(x) = x^2$. Draw the tangent line at the point to the graph of f at $(1, 1)$ and some secant lines through nearby points.

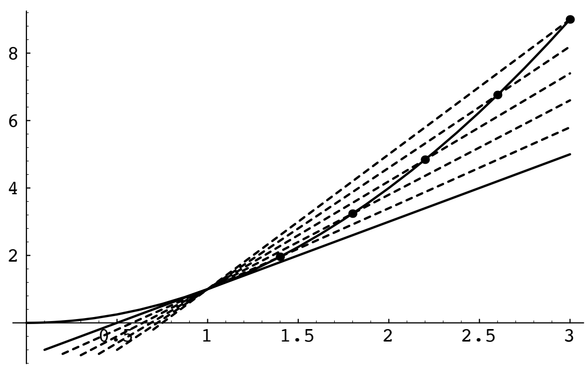

Figure 2.12: Tangent and Secant Lines

Solution: In Figure 2.12 you see the graph of $f(x) = x^2$ (solid line), the tangent line at the point $(1, 1)$ (solid line), and five secant lines (dashed lines). Each of them goes through the point $(1, f(1))$. In addition, they go through the points $(1.4, f(1.4))$, $(1.8, f(1.8))$, $(2.2, f(2.2))$, $(2.6, f(2.6))$, and $(3, f(3))$, respectively. You should recognize how the slopes of the secant lines are not that far from the slope of the tangent line, in particular as x gets closer to x_0. You might say, that the difference of the slopes is controlled by a function of the form $A|x - x_0|$. $\diamondsuit$

[8] In fact, a function is differentiable at x_0 if and only if there exists a number $f'(x_0)$ for which the conclusion in this theorem holds. With this we have found a another way to express that a function is differentiable at a point.

Let us consider an example where the derivative is interpreted as a rate of change, and where we can compare it with the slope of secant lines.

Example 2.20. At which rate does the volume of a cube change as we increase its side length?

Solution: You see the picture of a cube in Figure 2.13. A cube with side length a centimeter (cm) has a volume of $V(a) = a^3$ cubic centimeters (cm^3). We think of V as a function of a. You see part of this function graphed in Figure 2.14. If $a = 10$ cm then the volume is 1000 cm^3.

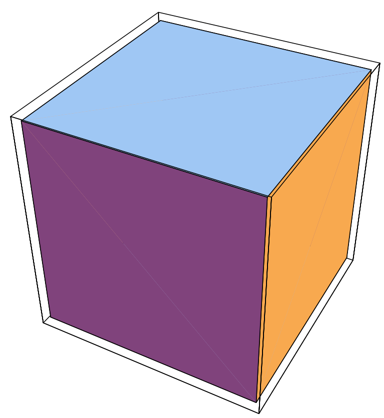

Figure 2.13: A cube

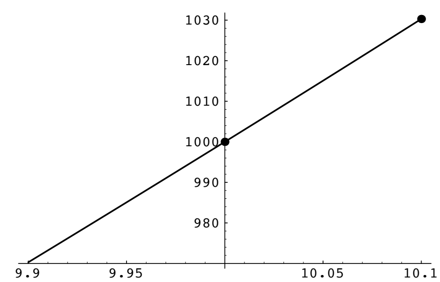

Figure 2.14: $V(a) = a^3$

What happens as we increase a by 1 millimeter, i.e., by .1 cm? Your calculator will show that $V(10.1) = 1030.301$ cm^3. This means that the volume increased by 30.301 cm^3. So, if we start out with a cube of side length 10 cm and increase the side length by .1 cm, then the volume increases by 30.301 cm^3. This translates into an average rate of change in volume (as a increases from 10 cm to 10.1 cm) of 30.301 cm^3 per .1 cm in side length, or of 303.01 cm^3 per 1 cm. Our calculation is illustrated in Figure 2.14. The two dots in Figure 2.14 represent the points $(10, 1000)$ and $(10.1, 1030.301)$. The slope of the secant line through the two dots has slope 303.01.

Let us compare this conclusion with the one derived from the derivative. In Example 2.42 on page 88 we will show[9] that $V'(a) = 3a^2$, so that $V'(10) = 300$. This means that the tangent line to the graph of V at $a = 10$ has slope 300. Interpreted in terms of rates of change this means that the volume

[9] Instead, you can also use Exercise 44 to arrive at the same conclusion. This takes some arithmetic skill.

of the cube increases at a rate of 300 cm^3 per 1 cm of side length at the moment the side length is 10.

In Figure 2.14 we have zoomed in on the point $(10, 1000)$ on the graph, and this means that you can barely see that the graph is not a straight line This also means that we cannot make visible anymore the difference between the secant line through the points $(10, 1000)$ and $(10.1, 1030.301)$ and the tangent line at $(10, 1000)$.

Consider a practical way of enlarging the cube. Add a layer of thickness .1 cm to three, non-opposing sides. That will add 30 cm^3 to its volume. The volume increases at a rate of of 30 cm^3 per .1 cm of thickness of the layer, or 300 cm^3 per 1 cm. Well, we made a mistake. After adding the layers to the sides, we do no have a cube anymore. Along some edges there will be a groove. The volume of these grooves will be .301 cm^3 if the thickness of the layer is .1 cm, or 3.01 cm^3 per 1 cm. The rate at which the volume of the groove changes with the thickness of the added layer is the difference between the rate of change and the average rate of change, the slope of the tangent line and the secant line. $\diamondsuit$

Exercise 45. Consider a ball of radius 10 cm.

1. Find the volume and the surface area of the ball. (You may consult your high school math book, or any other source.)

2. By how much does the volume of the ball change if its radius is increased to 10.1 cm?

3. What is the average rate of change in volume as its radius is increased from 10 cm to 10.1 cm?

4. At which rate does the volume of the ball change when its radius is 10 cm? Explain in practical terms why this rate coincides with the surface area of the ball.

2.5 Upper and Lower Parabolas

We would like to give a geometric definition of differentiability and the derivative. For this we first need to understand the geometry of intersecting and touching lines and parabolas.

Lines touching Parabolas

Consider a parabola $y(x)$ and a point x_0. In Section 1.2 we learned how to expand $y(x)$ in powers of $(x - x_0)$:

$$y(x) = A(x - x_0)^2 + B(x - x_0) + C.$$

We considered the line $l(x) = B(x - x_0) + C$ and wrote $y(x)$ in the form:

$$y(x) = A(x - x_0)^2 + l(x).$$

The following two properties are apparent:

1. The only intersection point of $y(x)$ and $l(x)$ is (x_0, y_0), i.e., $y_0 = y(x_0) = l(x_0)$ and $y(x) \neq l(x)$ if $x \neq x_0$.
2. The parabola lies on one side of the line, i.e., $y(x) \geq l(x)$ for all x or $y(x) \leq l(x)$ for all x.

Summarizing these properties we make the following

Definition 2.21. *Suppose $y(x)$ is a parabola and $l(x)$ a line. We say that they touch in (x_0, y_0) if 1 and 2 from above are satisfied.*

Example 2.22. Find the line $l(x)$ which touches the parabola

$$y(x) = x^2 - 2x + 3$$

in the point $(2, 3)$.

Using the formulas in (1.13) (see Section 1.2), we find the expansion of $y(x)$ in powers of $(x - 2)$:

$$y(x) = (x - 2)^2 + 2(x - 2) + 3.$$

Setting $l(x) = 2(x - 2) + 3 = 2x - 1$, we find the desired line. In Figure 2.15 you see the parabola, the line, and the point they have in common. $\diamondsuit$

With a little more work, one can show

Proposition 2.23. *Given a parabola $p(x)$ and a point (x_0, y_0) on it. There exists exactly one line $l(x)$ which touches the parabola in (x_0, y_0).*

Exercise 46. Find the line which touches the parabola

$$p(x) = 3x^2 - 5x + 2$$

in the point $(2, 4)$.

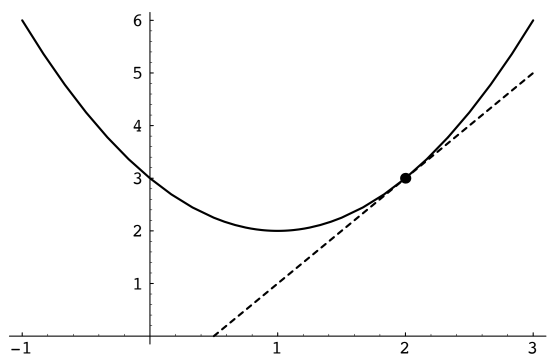

Figure 2.15: A line touching a parabola

Two Parabolas touching each other

Let us investigate how parabolas intersect. Suppose you are given two parabolas:

$$p(x) = a_1x^2 + b_1x + c_1 \quad \text{and} \quad q(x) = a_2x^2 + b_2x + c_2.$$

To find their intersection points we equate $p(x)$ and $q(x)$. In other words, we look for the roots of

$$p(x) - q(x) = (a_1 - a_2)x^2 + (b_1 - b_2)x + (c_1 - c_2).$$

The highest power of x in this equation is at most 2 (this happens if $(a_1 - a_2) \neq 0$), and this means that it has at most two solutions. We consider an example in which we encounter the behaviour which we are most interested in. You will study other possible intersection behaviour in the exercises.

Example 2.24. Investigate how the following two parabolas intersect:

$$p(x) = x^2 - 2x + 3 \quad \& \quad q(x) = -x^2 + 6x - 5.$$

We graphed the parabolas in Figure 2.16, and you can compare the following calculation with the picture. We find the intersection points of the

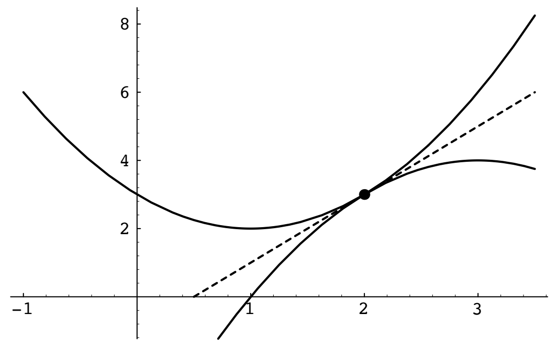

Figure 2.16: A line separating two parabolas

parabolas:

$$p(x) - q(x) = 2x^2 - 8x + 8 = 2(x-2)^2.$$

The parabolas intersect in exactly one point, $(2, 3)$. In fact, $p(x) - q(x) \geq 0$, so that $p(x) \geq q(x)$ for all x. Equality holds only for $x = 2$. Geometrically speaking, the two graphs touch in the point $(2, 3)$, but they do not cross.

Expanding $p(x)$ and $q(x)$ in powers of $(x - 2)$, we find

$$p(x) = (x-2)^2 + (2x-1) \quad \& \quad q(x) = -(x-2)^2 + (2x-1).$$

The lines which touch the parabolas $p(x)$ and $q(x)$ in the intersection point $(2, 3)$ are the same, namely $l(x) = 2x - 1$. This line separates the two parabolas in the sense that the parabola $p(x)$ lies above the line, and the parabola $q(x)$ lies below it. In Figure 2.16 the line $l(x)$ is shown as a dotted line. $\Diamond$

There are two essential features to the intersection behaviour in the example.

1. The parabolas $p(x)$ and $q(x)$ *touch* in (x_0, y_0), i.e. $p(x) \geq q(x)$ for all x, or $p(x) \leq q(x)$ for all x, and $p(x) = q(x)$ if and only if $x = x_0$.[10]

[10] We could have required the inequalities on some open interval which contains x_0

2. There is a line $l(x)$ which *separates* $p(x)$ and $q(x)$, i.e., $q(x) \leq l(x) \leq p(x)$ for all x, or $p(x) \leq l(x) \leq q(x)$ for all x.

With some effort one can show:

Proposition 2.25. *Suppose $p(x)$ and $q(x)$ are parabolas which intersect in the point (x_0, y_0). The parabolas touch in (x_0, y_0) and they are separated by a line if and only if one parabola is open upwards, one parabola is open downwards, and (x_0, y_0) is their only intersection point. The line which separates the parabolas is unique. It is the line which touches $p(x)$ and $q(x)$ in (x_0, y_0).*

For completeness sake, let us look at the other possible intersection behaviours. Instead of touching at an intersection point, the parabolas could cross. You probably have the right intuitive ideas what that means, but to give you the means of checking this property, we formalize the idea. The graphs of two functions $p(x)$ and $q(x)$ *cross* at x_0 if $p(x_0) = q(x_0)$ and $(p(a) - q(a))(p(b) - q(b)) < 0$ for all a in an interval (A, x_0) and b in (x_0, B). The intervals are assumed to be non-empty.

In the following exercise you can observe all of the different behaviours.

Exercise 47. Find the intersection points for each pair of parabolas. Decide for each intersection point whether the parabolas touch or cross. If the parabolas touch in an intersection point, decide whether there is a line separating the parabolas, and if so, find the equation of the separating line. Provide a sketch for the intersection behaviour of each pair of parabolas.

1. $p(x) = x^2 - x + 1$ and $q(x) = 2x^2 - 3x + 2$.
2. $p(x) = x^2 - 3x + 2$ and $q(x) = x^2 - 5x + 6$.
3. $p(x) = x^2 - 4x + 4$ and $q(x) = 2x^2 - 4x + 5$.
4. $p(x) = x^2 - 2x + 1$ and $q(x) = -x^2 + 2x + 3$.
5. $p(x) = x^2 - 3x + 3$ and $q(x) = -x^2 + 5x - 5$.

Exercise 48. Suppose two parabolas $p(x)$ and $q(x)$ intersect in the point (x_0, y_0) without crossing. Show that (x_0, y_0) is the only intersection point for these parabolas.

instead. One can show that under the given circumstances these two conditions are the same.

Differentiability via Upper and Lower Parabolas

We like to give another interpretation for the concept of differentiability, the tangent line, and the derivative. For a function $f(x)$ to be differentiable at an interior point x_0 of its domain, we asked for a line $l(x)$ and a (necessarily non-negative) number A, such that

$$|f(x) - l(x)| \leq A(x - x_0)^2$$

for all x in some open interval I around x_0. The inequality may be written in a different, equivalent form:

$$-A(x - x_0)^2 \leq f(x) - l(x) \leq A(x - x_0)^2.$$

After adding $l(x)$ everywhere, it reads

$$-A(x - x_0)^2 + l(x) \leq f(x) \leq A(x - x_0)^2 + l(x). \tag{2.24}$$

Note that $l(x)$ is the tangent line to the graph of f at the point $(x_0, f(x_0))$, and that it is given by the formula

$$l(x) = f'(x_0)(x - x_0) + f(x_0).$$

The left and right most terms in (2.24) are parabolas, and with the expression for $l(x)$ substituted we denote them by

$$\begin{aligned} q(x) &= -A(x - x_0)^2 + m(x - x_0) + f(x_0) \\ p(x) &= \quad A(x - x_0)^2 + m(x - x_0) + f(x_0). \end{aligned}$$

With this notation (2.24) reads

$$q(x) \leq f(x) \leq p(x). \tag{2.25}$$

The parabola $q(x)$ is open downwards and the parabola $p(x)$ is open upwards. They touch each other in the point $(x_0, f(x_0))$, and they are separated by the tangent line $l(x) = m(x - x_0) + f(x_0)$. Summarizing the above we have the following geometric formulation for the concept of differentiability. Expressed informally:

- A function is differentiable at an interior point x_0 of its domain, if, on some open interval around x_0, its graph is trapped between two parabolas which touch each other in the point $(x_0, f(x_0))$. The unique line which separates the parabolas is called the tangent line to the graph of f at x_0, and its slope is called the derivative of f at this point. This slope is denoted by $f'(x_0)$.

In formal mathematical language, this statement reads as follows:

Proposition 2.26. *Suppose f is a function and x_0 is an interior point of its domain. Then f is differentiable at x_0 if and only if there exist parabolas $p(x)$ and $q(x)$, one open upwards and one downwards, which touch each other in $(x_0, f(x_0))$ such that*

$$q(x) \leq f(x) \leq p(x)$$

for all x in some open interval around x_0.

Note that in this proposition the tangent line to the graph of f at x_0 is the unique line which separates the parabolas, and its slope is $f'(x_0)$, the derivative of $f(x)$ at x_0.

Strictly speaking, we have only shown the 'if' part of the proposition. We leave the 'only if' part to the motivated audience. The advantage of the proposition is that it expresses differentiability in a geometric way. It gives you a concrete picture which you can think about. It provides you with some intuition. Using the example of the exponential function $f(x) = e^x$, we illustrated the statement that f is differentiable at $x = 1$ in the language of the proposition in Figure 2.3 on page 39.

Let us illustrate the discussion with an example and draw the corresponding picture.

Example 2.27. Find the tangent line and upper and lower parabolas to the graph of $f(x) = \sin x$ at $x_0 = \pi/4$. Graph all of the above.

Solution: We learned that $\sin' x = \cos x$, so that $\cos(\pi/4) = \sqrt{2}/2$ is the slope to the tangent line in question. Noting that $\sin(\pi/4) = \sqrt{2}/2$, we find the point-slope formula for the tangent line:

$$l(x) = \frac{\sqrt{2}}{2}\left(x - \frac{\pi}{4}\right) + \frac{\sqrt{2}}{2}.$$

In our discussion in Example 2.11, we saw that for our function $f(x)$ and the line $l(x)$:

$$|f(x) - l(x)| \leq A(x - x_0)^2$$

with $A = 1$ and $|x - x_0| < \pi/4$. This means that we can trap the graph of $\sin x$ between the parabolas (see (2.24))

$$\begin{aligned} q(x) &= -A(x - x_0)^2 + f'(x)(x - x_0) + f(x_0) \\ &= -\left(x - \frac{\pi}{4}\right)^2 + \frac{\sqrt{2}}{2}\left(x - \frac{\pi}{4}\right) + \frac{\sqrt{2}}{2}, \end{aligned}$$

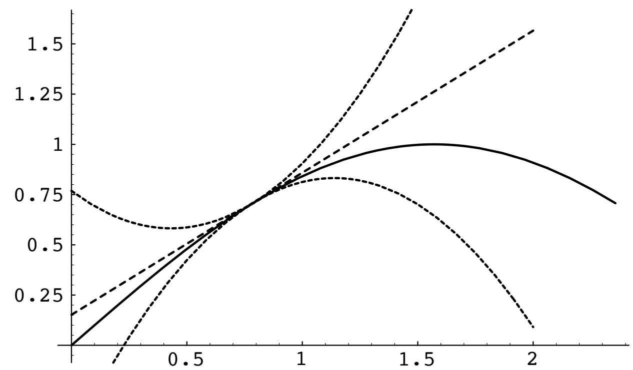

Figure 2.17: Sine Function and Tangent Line between two Parabolas

and

$$p(x) \quad = \quad \left(x - \frac{\pi}{4}\right)^2 + \frac{\sqrt{2}}{2}\left(x - \frac{\pi}{4}\right) + \frac{\sqrt{2}}{2}.$$

The graph of $f(x)$ (solid line), the tangent line (long dashes) and the parabolas (short dashes) are shown in Figure 2.17. $\diamondsuit$

Exercise 49. Let $f(x) = \cos x$ and $x_0 = \pi/6$.

1. Find the tangent line to the graph of f at the point $(x_0, f(x_0))$.

2. Find the parabolas $p(x)$ and $q(x)$ which touch at $(x_0, f(x_0))$, so that the graph of f is trapped in between them, at least as long as $|x - x_0| < \pi/4$.

3. Use technology to graph the functions f, l, p, and q accurately.

Exercise 50. Let $f(x) = \sqrt{3x+2}$ and $x_0 = 5$.

1. Find the tangent line to the graph of f at the point $(x_0, f(x_0))$.

2. Find the parabolas $p(x)$ and $q(x)$ which touch at $(x_0, f(x_0))$, so that the graph of f is trapped in between them. Hint: Use A from Exercise 41.

3. Use technology to graph all of the above accurately.

Using the formulation of differentiability from Proposition 2.26, we would like to given another interpretation of the statement that the distance between the graph and the tangent line is 'small'. You see in Figure 2.17, and in other pictures where we drew touching parabolas, how the parabolas 'hug' the line which separates them. On some interval around the point at which the parabolas touch, the graph of the function and the tangent line are 'squeezed' in between these two parabolas. This is how close the graph and the tangent line have to be to each other.

The intuitive, geometric picture to understand differentiability and the derivative may appeal to you. Still, there are benefits to the way in which these concepts are explained in Definition 2.9. The definition is formulated so that no other concepts have to be developed first. Without any other preparation you can just write it down. That definition is also analytic, and this means that you can manipulate it and use it in calculations.

2.6 Other Notations for the Derivative

There are different notations for the derivative of a function. Physicists will indicate a derivative with respect to time by a dot. E.g., if x is a function of time, then they will write $\dot{x}(t)$ instead of $x'(t)$. Leibnitz' notation for the derivative of a function f of a variable x is $\frac{df}{dx}$. We will use it frequently. Using this notation, Theorem 2.12 on page 52 translates into the statement:

$$\text{If } y(x) = e^x, \text{ then } \frac{dy}{dx} = y \text{ or } \frac{dy}{dx} = e^x.$$

A reformulation of Theorem 2.13 on page 52 is:

$$\text{If } y(x) = \ln x, \text{ then } \frac{dy}{dx} = \frac{1}{x}.$$

This notation is not always specific enough. The expression dy/dx stands for the derivative of y with respect to x, and that is a function. The expression does not tell where dy/dx is evaluated. To be specific about this aspect, it makes sense to write (compare Example 2.11 on page 50):

$$\text{If } y(x) = \sin x, \text{ then } \frac{dy}{dx}(x) = \cos x.$$

In this notation x plays two roles. It is the name of the variable of y as well as the name of the variable of the derivative of y. This in acceptable because it won't lead to confusion. Instead of $\frac{df}{dx}(x)$ we also write $\frac{d}{dx}f(x)$. This is particularly convenient if f stands for a larger expression as in

$$\frac{d}{dx}\sin x = \cos x \quad \text{or} \quad \frac{d}{dx}e^x = e^x.$$

Exercise 51. Find the derivatives

$$(1)\ \frac{d}{dx}\ln x \quad (2)\ \frac{d}{dx}\sqrt{8x-4} \quad (3)\ \frac{d}{dx}(4x^2-3x+5) \quad (4)\ \frac{d}{dx}4e^{3x}$$

2.7 Exponential Growth and Decay

We like to give an application of the concept of the derivative. Suppose you culture bacteria in a laboratory. You assume that the rate at which the population grows is proportional to the populations. This is called the *Malthusian Law*. To express this mathematically, let $A(t)$ be the number of bacteria in the sample at time t. Then you are asserting that

$$A'(t) = aA(t) \tag{2.26}$$

for some constant a.

This equation is an example of a *differential equation*. The unknown is a function, and the equation relates the functions and its derivative. More specifically, (2.26) is an ordinary first order linear differential equation with constant coefficients. A solution of a differential equation is a function which satisfies the equation.

In Theorem 2.12 on page 52 we stated that the only solutions of (2.26) are of the form $A(t) = ce^{at}$ for some constant c. We can find c by plugging $t = 0$ into the equation, $A(0) = c$. So c is the number of bacteria at time $t = 0$, and, setting $A(0) = A_0$,

$$A(t) = A_0e^{at}. \tag{2.27}$$

Next we can determine a, at least if we also know the population at another time t_1. So we suppose that

$$A_1 = A(t_1) = A_0e^{at_1}, \tag{2.28}$$

where A_1 is known. Then $A_1/A_0 = e^{at_1}$. Applying the natural logarithm function to both sides of the equation (it is the inverse of the exponential function, see Definition 1.20 on page 27) we find $\ln(A_1/A_0) = at_1$, and

$$a = \frac{1}{t_1}\ln(A_1/A_0). \tag{2.29}$$

Remark 5. We called the constant a the relative growth rate, or growth rate for short. Its physical dimension is "per time unit." If time is measured in hours, then the dimension of a is per hour. The units of A_0 in the example are bacteria, so the units of $A'(t)$ are bacteria per hour.

Let us at a numerical example.

Example 2.28. Suppose that in the beginning of your experiment you estimate that your culture contains 850 yeast bacteria. Ten minutes later the population has grown to 1200 bacteria. You assume that the same growth rate continues for 50 more minutes. What is the population 40 minutes after you started the culture? At which time do you expect that the culture contains 2500 bacteria?

Let us work out the answers. We set $t_0 = 0$ and $t_1 = 10$. By assumption, $A_0 = 850$ and $A_1 = 1200$. This means that

$$a = \frac{1}{t_1}\ln(A_1/A_0) = \frac{1}{10}\ln(1200/850) = .034484,$$

or that the population grows at a rate of 3.4484% per minute. The first question is about $A(40)$. Plugging in our data we find

$$A(40) = 850e^{.034484\times 40} = 3376.5,$$

so that you expect about 3376 bacteria in your culture 40 minutes into your experiment.

The second question is, at which time t do you expect that $A(t) = 2500$? This means that we have to solve the equation

$$A(t) = 850e^{.034484t} = 2500 \quad \text{or} \quad e^{.034484t} = 2500/850$$

for t. Applying the natural logarithm function to both sides of the equation you see that

$$.034484t = \ln(2500/850) \quad \text{or} \quad t = 31.2844 \text{ minutes.}$$

You should verify these calculations on your calculator. $\Diamond$

Exercise 52. Suppose a culture of yeast bacteria grows at a constant rate for one hour. Initially you have $3,000$ bacteria, and 15 minutes later you have $20,000$.

1. What is the growth rate of the population?

2. What is the population 40 minutes after you started the culture?

3. When will the population reach $1,000,000$?

More generally than above, we have:

Proposition 2.29. *The function $A(t) = A_0 e^{a(t-t_0)}$ is the unique solution of the initial value problem*

$$A'(t) = aA(t) \quad \text{and} \quad A(t_0) = A_0.$$

Exercise 53. Show the proposition. Hint: Modify the arguments which we used above.

Exercise 54. Suppose a given population doubles within an hour. What is the growth rate?

You may say, that by now you have an explicit formula for $A(t)$ and know everything about the population of bacteria at any time. Still, if you apply your conclusion to real life, then you should be aware that we only modeled population growth. If we apply the conclusions to real life, then we may have to be cautious.

To be more specific, let us consider two functions. Let $A(t)$ denote the size of the population in our mathematical model, and $B(t)$ the actual population. For certain purposes it is asserted, that we can identify these two functions. E.g., we make this assertion if we like to estimate the populations of a sizeable, homogeneous population of microbes grown under controlled and constant conditions. Then we think, and experimental evidence confirms, that the Malthusian law ($A' = aA$) describes the dynamics of population growth closely enough. For practical purposes, we think that we may identify $A(t)$ and $B(t)$.

In certain respects, the two functions are very different from each other. The value for $B(t)$ is always a natural number and the one of $A(t)$ typically is not. So $A(t)$ does not tell us the exact population at a given time. The function $B(t)$ gives the exact value for the size of the population at any time. It also shows when there is an increase (or decrease) of the size.

Given a and A_0, we know $A(t)$ exactly, but typically it is impossible to know $B(t)$ precisely. There is no way to keep track of the exact number of bacteria at all times in a population which is in the thousands.

In our examples we have seen how to find the growth rate a for a population [11]. Assuming that populations growth followed the Malthusian law,

[11] A fairly good estimate for the growth rate a in obtained by finding the average rate

we only needed to know the size of the population at two different times. The accuracy with which we know a depends on the accuracy with which we the size of the population at those two times.

In an experiment, the growth rate will depend on the food supply, the temperature, the size of the population, the concentration of chemicals produced by the population, and more factors. It is quite difficult to keep them, and with this the growth rate constant. If the growth rate depends on time and the population, then a is a function of these variable, and that substantially changes the solution for the differential equation for $A(t)$. The growth rate will also change with time, if the populations consists of several smaller populations, and each of them grows at a different rate. This occurs if the population is not homogeneous. Still, if we can keep the growth rate 'nearly' constant, then general mathematical theory tells us that $A(t)$ is 'rather close' to an exponential function. Only real life comparison between the mathematical model and the laboratory experience can (and does) confirm that the model fits reality well.

The essential statement is, that the Malthusian law

$$A'(t) = aA(t)$$

reflects essential elements of the dynamics of population growth. It says that the rate at which a population grows is proportional to the size of the population. Under idealized circumstances and over shorter periods of time one may assume that the growth rate a is constant, and this leads to the conclusion that the size of a population grows exponentially.

Outside the laboratory population growth is much more complicated. If you try to find a function which tells you about the human population in the future, then you need to take many more aspects into account. It is essential that you distinguish three periods in life, the time before the reproductive age, the reproductive age, and the time after this. The number of females is more important than the number of males. Social values and economical interest affect the rate of reproduction in specific parts of the population. Food supply, sanitary conditions, and medical care influence the survival rate of newborns. Progress in medicine, the supply of doctors and medications affect the life expectancy of individuals. National and ethnic values and legislation encourage or discourage reproduction. Most of these factors are difficult to measure and incorporate in a weighted fashion into

of change $b = (A(t_0) - A(t_1))/(t_0 - t_1)$ of the function $A(t)$ for t between t_0 and t_1. According to our discussion about slope of tangent lines and secant lines in Section 2.4, $A'(t_0) = aA(t_0)$ is close to b, at least if t_1 is close to t_0

the equation. They also change considerably with time. Still, the abstract statement made in Equation (2.26) remains true, but it does not apply to the population as a whole. It needs to be applied in a very differentiated way with a lot of attention to detail before one can hope to understand the growth of the human population.

Carbon-14 Dating

Let us consider the method of carbon-14 dating as another example of exponential growth, or better exponential decay. This method was discovered by Willard Libby around 1949. The situation is as follows. Cosmic rays bombard the atmosphere of the earth, and produce, through a complicated process, carbon-14 (^{14}C). It is called a radiocarbon as it decays radioactively. Living substances, like wood or bones, absorb carbon-14 during their life time. At the same time, radiocarbons disintegrate, and in the living substance it comes to an equilibrium where as much radiocarbon is absorbed as disintegrates. The concentration of ^{14}C is characteristic for the substance. It is also assumed that the bombardment with cosmic rays has been constant for a long period of time, so that the concentration in substances is independent of the time during which they were alive. (This has changed recently with the atmospheric tests of nuclear devices, which increased the concentration of radiocarbons in the atmosphere.) When the organism dies, no more radiocarbons are absorbed. Radiocarbons decay and change to non-radioactive substances. Physics and experience tell us that the number of ^{14}C molecules which decay in some time period is proportional to the number of molecules present. If $A(t)$ is the number of ^{14}C molecules in a sample, then $A'(t)$ is the rate at which $A(t)$ changes. If we call the proportionality factor $-a$, then we again end up with the equation

$$A'(t) = -aA(t).$$

We conclude that the solution for this differential equation is of the form

$$A(t) = A_0 e^{-at} \tag{2.30}$$

where A_0 is now the number of ^{14}C molecules in the sample at the time of death. We used $-a$, instead of a in the exponential growth example, to indicate that $A'(t)$ is negative. We call a the *rate of decay.* Once more, this explicit expression for $A(t)$ can be used to provide us with lots of information. All we need to know are A_0 and a.

Usually it is not feasible to count the number of radioactive molecules in a sample. It is much easier to measure the number of decays in one unit (e.g., one gram) of the substance per time unit (e.g., one minute). This number is proportional to the number of radioactive molecules in the sample. If we denote this number by $A(t)$, then this function still satisfies the differential equation in (2.30), only that now A_0 denotes the number of decays (in one unit of the substance per time unit) at time $t = 0$. Context dictates which meaning we will assign to $A(t)$.

For a radioactive substance it is typical to provide the *half-life*. This is the time in which half of the substance decays. For ${}^{14}C$ the half-life is about 5568 years. If you start out with 1mg of it, then after 5568 years only half of it is left. Knowing the half-life allows us to calculate the rate of decay. If T is the half-life for a radioactive substance, then

$$\frac{1}{2} = e^{-aT} \quad \text{or} \quad a = \frac{\ln 2}{T}. \tag{2.31}$$

For ${}^{14}C$ the value for a is about 0.000124488, as you may verify on your calculator. As we used years to measure time, this means that approximately 0.0124488% of the ${}^{14}C$ decays per year. The word 'approximate' refers to the fact that the rate of decay is approximately the amount which decays in one unit of time, as explained in Theorem 2.18 on page 59.

You need a second piece of information. It has been measured that one gram of living wood produces 6.68 ${}^{14}C$ disintegrations per minute, or, more precisely, that this was true for wood which died before nuclear testing began. This provides us with A_0, if we need it.

Example 2.30. Let us consider as example a piece of wood found in the burial chamber of the mummy (compare Example 5.16 on page 287). Suppose you measure 1.8 disintegrations of carbon-14 per gram and minute in the sample piece of wood. How old is the piece of wood?

Solution: If t_1 is the age of the piece of wood, and A_0 is the amount of radiocarbon in the wood at time $t_0 = 0$, then we are saying that

$$A(t_1) = 1.8 = A_0 e^{-\frac{\ln 2}{5568} t_1}. \tag{2.32}$$

We like to solve this equation for t_1. We divide the second equality by $A_0 = 6.68$, apply the natural logarithm, and find

$$\ln \frac{1.8}{6.68} = -\frac{\ln 2}{5568} t_1 \quad \text{or} \quad t_1 = -\frac{5568}{\ln 2} \ln\left(\frac{1.8}{6.68}\right) = 10,533.8, \tag{2.33}$$

as you should verify. This means that the piece of wood should be about 10,500 years old. You jump to the conclusion that the mummy is that old as well. Check your book on world history! $\diamond$

Exercise 55. A piece of wood weighing 7.4 grams produces 23.47 disintegrations of carbon-14 per minute. How old is it?

Exercise 56. You measure the number of radio active decays in a soil sample which was taken near Chernobyl after the radio active fallout settled. Now, 10 years after the accident, the sample shows 370 decays per minute. Records indicate that seven years ago (i.e., three years after the accident) the same sample produced 430 decays per minute. Assume that there is only one kind of radio active substance in the sample.

1. What is the half-life of the radio active substance in the sample?
2. How many decays would you have measured right after the accident?
3. How many more years will it take until the sample will only produce 25 decays per minute?

You may ask why the equation

$$A'(t) = aA(t)$$

is of such great importance. As we emphasized, it expresses that the rate of change of $A(t)$ is proportional to $A(t)$. This is the principal assumption made for many real life processes. It is assumed that, within limitations, this happens when your body absorbs an orally administered medication. This happens when your liver eliminates toxins from your blood. This is how a contagious disease spreads in a population (initially!). This is how the value of money diminishes with inflation. Typically other factors will also effect A', at least after some time. E.g., a substantial part of a population may develop an immunity to the disease. This will change the equation. Taking such changes into account makes the equation more complicated, and the solution will look quite different. We discuss one modification of the Malthusian law in the next section.

2.8 More Exponential Growth and Decay

More generally than in (2.26), consider the differential equation

$$f'(t) = af(t) + b, \tag{2.34}$$

where a and b are constants, and $a \neq 0$.

Theorem 2.31. *Functions of the form $f(t) = ce^{at} - b/a$ are solutions of the differential equation in* (2.34)*, and every solution of* (2.34) *is of this form. Here c denotes an arbitrary constant.*

We obtain a unique solution if we add an initial condition to the differential equation in (2.34).

Theorem 2.32. *The function*

$$f(t) = \left(y_0 + \frac{b}{a}\right) e^{a(t-t_0)} - \frac{b}{a}$$

is the unique solution of the initial value problem

$$f'(t) = af(t) + b \quad \textit{and} \quad f(t_0) = y_0.$$

Exercise 57. Work out the formula in Theorem 2.32 by using the conclusion of Theorem 2.31.

Proof of Theorem 2.31. Adding a constant to a function moves the graph vertically, and this does not change the derivative of the function. This is also implied by (2.37), which we discuss later. If $f(t) = ce^{at} - b/a$, then

$$f'(t) = \frac{d}{dt}\left(ce^{at} - \frac{b}{a}\right) = ace^{at} = a\left(f(t) + \frac{b}{a}\right) = af(t) + b,$$

so that $f(t)$ satisfies (2.34).

We show that every solution of (2.34) is of the form $ce^{at} - b/a$, for some constant c. Set

$$g(t) = f(t) + \frac{b}{a} \quad \text{resp.,} \quad f(t) = g(t) - \frac{b}{a}$$

Then

$$g'(t) = f'(t) = af(t) + b = a\left(g(t) - \frac{b}{a}\right) + b = ag(t).$$

According the Theorem 2.12, $g(t)$ is of the form ce^{at}, so that f is of the form claimed in the theorem. □

Let us apply these ideas to solve some problems. The important aspects are to translate the given information into a mathematical equation. The rest will be routine calculation.

Example 2.33. On graduation day your student loan has a balance of \$15,000. Interest is added at a rate of .5% per month, and you are repaying the loan at a rate of \$ 200.00 per month. How long will it take you to repay the loan?

Solution: As variable we use time, and we denote it by t. We measure time in months, because this is the way in which the information is given to us. We set $t = 0$ at the time of graduation. This is the time at which you start to repay the loan. Denote the balance of your loan at time t by $B(t)$. Let us determine $B'(t)$, the rate a which the balance of the account changes. The balance increases due to interest charges, and the rate at which this happens is $.005B(t)$. Secondly, the balance decreases at a rate of \$200.00 per month due to payments which you make. These two contributions determine how $B(t)$ changes, and we conclude that

$$B'(t) = .005B(t) - 200.$$

In addition we have that $B(0) = 15,000$. We apply Theorem 2.32 with $f(t) = B(t)$, $t_0 = 0$, $a = .005$, $b = -200$, and $y_0 = 15,000$. The conclusion of the theorem is that

$$B(t) = \left(15,000 + \frac{-200}{.005}\right) e^{.005t} - \frac{-200}{.005} = -25,000e^{.005t} + 40,000.$$

The problem asked us to find the time T for which $B(T) = 0$, i.e., the time at which you paid the loan in full. This provides us with the following equation for T:

$$0 = -25,000e^{.005T} + 40,000 \quad \text{or} \quad \frac{40}{25} = e^{.005T}.$$

Then

$$T = \frac{1}{.005} \ln\left(\frac{40}{25}\right) = 94.$$

In the final analysis, you repaid your loan in 94 months, or 7 years and 10 months. Your total payments were \$18,800, so that you paid the principal plus \$3,800 in interest. ◊

Exercise 58. You are saving money at a rate of \$1,000.00 per month towards the down payment of your family residence. Your bank pays interest at a rate of .6% per month. But, your spouse keeps spending the money at a rate of .2% of the account balance per month. (E.g., if the balance is \$40,000, then the spouse spends the money at a rate of \$80.00 per month.) How long will it take to accumulate a down payment of \$80,000?

Exercise 59. You are absorbing a medication at a rate of 3 mg per hour. (You can keep this rate constant with a skin patch.) The liver eliminates the medication at a rate of 4% per hour. I.e., if there are 30 mg in your body, then the liver eliminates the medication at a rate of 1.2 mg/hr. Denote by $A(t)$ the amount of medication in your body t hours after you started taking the medication.

1. Which differential equation does $A(t)$ satisfy?
2. Find $A(0)$ and $A(t)$ for any time t.
3. For which value of A is your intake of medication the same as the amount eliminated by the liver?
4. Which amount of medication in your body will not be exceeded?
5. How long does it take until the amount of medication in your body reaches 65 mg?

Example 2.34 (Newton's Law of Cooling). Suppose you have an object whose temperature is different from the temperature of its surroundings. With time, the temperature of the object will approach the one of its surroundings. We discuss how this happens, at least under idealized circumstances.

Think of the object as the coffee in your cup which you keep on your desk. You stir the coffee gently so that the temperature in the cup remains homogeneous and almost no energy is added through the process of stirring.[12] Denote the temperature of the object (your coffee) by T. It is a function of time, so that we write $T(t)$. Newton's law of cooling says that the rate at which the heat is transferred, and with this the rate of change of temperature of the coffee, is proportional to the temperature difference. If K is the temperature of the surroundings, then

$$T'(t) = a(T(t) - K). \tag{2.35}$$

[12]The physics of heat transfer changes substantially if you take a solid object, such as a turkey in the oven. The temperature in the solid will not be homogenous, the ourside warms up much faster than the inside. In addition, the specific heat (the amount of energy needed to increase the temperature of one unit of the material by one degree) varies. It is different for fat, protein, and bone. Furthermore, the specific heat is highly temperature dependent for substances like protein. That means, a in (2.35) depends on the temperature T. All of this leads to a significantly different development of the temperature inside a turkey as you roast it for your Thanksgiving dinner.

The differential equation in (2.35) is just the one in (2.34) is a slightly disguised form. After multiplying out the parentheses we get

$$T'(t) = aT(t) - aK,$$

so that the relation to the equation in (2.34) is made by setting $b = -aK$.

Let us work out a numerical example. At time $t = 0$, just after you poured the coffee into your cup, its temperature is 95 degrees Celsius. Five minutes later the temperature has dropped to 80 degree, while you stir it slightly and patiently. The room temperature is 25 degrees. You feel comfortable to start sipping the coffee once the temperature has dropped to 70 degrees.

1. Determine the function $T(t)$.

2. How much longer do you have to wait before you can start sipping your coffee?

Solution: To apply Theorem 2.32, we set $t_0 = 0$, $y_0 = 95$, and $K = 25$. Note that $-b/a = K$. Putting all of this into the formula for the solution of the initial value problem, we get that

$$T(t) = (95 - 25)e^{at} + 25 = 70e^{at} + 25.$$

To determine a we use that

$$T(5) = 80 = 70e^{5a} + 25,$$

and we conclude that $a = \frac{1}{5}\ln\left(\frac{55}{70}\right) \approx -.0482$. Using these data, Equation (2.35) says that the temperature of the coffee drops at a rate of about .048 degrees per minute for each degree of difference between the temperature of the coffee and the room temperature.

Having a numerical value for a gives us an explicit expression for the temperature T as a function of t:

$$T(t) = 70e^{-.0482t} + 25.$$

We like to find out the time t_1 for which

$$T(t_1) = 70e^{-.0482t_1} + 25 = 70.$$

Solving the equation for t_1, we find that $t_1 \approx 9.17$. That means we can start drinking the coffee 9.17 minutes after pouring it, or that we have to wait about another 4 minutes before we can enjoy it. $\Diamond$

Exercise 60. You buy coffee at a convenience store to drink on your way to school. Initially its temperature is 70 degrees. The temperature in your car is 28 degrees, and after 15 minutes the temperature of the coffee is 55 degrees. At which time will the temperature drop to 40 degrees?

Exercise 61. A chemical factory is located on the banks of a river. Down stream from the factory is a lake, and the river is the only contributor to the lake. Assume that the amount of water carried by the river is the same all year around, and the amount of water in the lake is 10 times the amount of water carried by the river per year. In negotiations which the EPA, the owner has agreed to an acceptable level of 2.5 mg per m^3 of a pollutant in the lake. After a major accident the level has risen to 15 mg per m^3. As a remedy, the factory owner proposes to reduce the emission of pollution so that the level of pollutant in the river is only 1.5 mg per m^3. It is assumed that the pollutant is distributed uniformly in the lake at any time.

1. Let $P(t)$ denote the amount of pollutant (measured in mg per m^3) in the lake at time t. Let $t_0 = 0$ be the time just after the accident and at which the clean-up strategy is implemented. State the initial value problem for $P(t)$.

2. Find the function $P(t)$.

3. At which time will the level of pollution be back to 2.5 mg per m^3?

2.9 Differentiability Implies Continuity

The basic observation made in this section is used in several places in this manuscript, but mostly within proofs. So you may consider it as a resource which you call upon whenever needed.

Differentiability of a function f at a point x_0 gives us good control over the values of the function at all points near x_0 in terms of $f(x_0)$ and $f'(x_0)$. Specifically, we have positive numbers A and d, such that the the estimate

$$\big|f(x) - [f(x_0) + f'(x_0)(x - x_0)]\big| \leq A(x - x_0)^2$$

holds for all x with $|x - x_0| < d$. This is not particularly explicit. We like to get an estimate for $|f(x) - f(x_0)|$. The next theorem provides an estimate in terms of A, d, $f'(x_0)$, and $|x - x_0|$.

Theorem 2.35. [13] *Suppose that f is differentiable at x_0, and A and d are as above. Then*

$$|f(x) - f(x_0)| \leq (|f'(x_0)| + Ad)|x - x_0| \tag{2.36}$$

for all x with $|x - x_0| < d$.

Paying less attention to details, if f is differentiable at x_0, then there exist numbers C and $d > 0$ such that

$$|f(x) - f(x_0)| \leq C|x - x_0|$$

for all x with $|x - x_0| < d$.

Example 2.36. Let $f(x) = \sin x$ and $x_0 = \pi/6$. Then $f(x_0) = 1/2$ and $f'(x_0) = \cos(\pi/6) = \sqrt{3}/2$. We saw that we may use $A = 1$ and any d, see Example 2.11. Set $d = \pi/12$, then we find that $|f'(x_0)| + Ad < 1.13$. The theorem says that

$$\left|\sin x - \frac{1}{2}\right| \leq 1.13 \left|x - \frac{\pi}{2}\right|$$

as long as $x \in (\pi/12, 3\pi/12)$.

Exercise 62. Suppose f, $C = |f'(x_0)| + Ad$ and d are as in the theorem. Show that the graph of f is trapped between two lines over the interval $(x_0 - d, x_0 + d)$. These lines intersect in the point $(x_0, f(x_0))$ and have slope C, resp., $-C$.

Proof of Theorem 2.35. Differentiability of the function f at x_0 assures us of the existence of numbers A and $d > 0$, such that

$$f'(x_0)(x - x_0) - A(x - x_0)^2 \leq f(x) - f(x_0) \leq f'(x_0)(x - x_0) + A(x - x_0)^2$$

for all x with $|x - x_0| < d$. The inequality is a variation of the one in (2.7). For a moment, set $h = x - x_0$ and $x = x_0 + h$. With this substitution, our previous inequality reads

$$f'(x_0)h - Ah^2 \leq f(x_0 + h) - f(x_0) \leq f'(x_0)h + Ah^2.$$

[13] We are not showing that differentiability at a point x implies continuity at this point. The strong notion of differentiability which we are using, implies that a function is also strongly continuous at the point under consideration. Technically speaking, a Lipschitz condition of order 2 implies one of order 1. We will introduce strongly continuous functions when we discuss integration, and when we look for a class of functions for which the Fundamental Theorem of Calculus holds.

It follows that

$$-|f'(x_0)||h| - Ah^2 \leq f(x_0+h) - f(x_0) \leq |f'(x_0)||h| + Ah^2.$$

Basic properties of the absolute value, see Section 5.2, and the assumption that $|h| < d$, tell us that

$$\begin{aligned} |f(x_0+h) - f(x_0)| &\leq |f'(x_0)h| + Ah^2 \\ &= (|f'(x_0)| + A|h|)|h| \\ &\leq (|f'(x_0)| + Ad)|h|. \end{aligned}$$

Reverting to our original notation, we have shown that

$$|f(x) - f(x_0)| \leq (|f'(x_0)| + Ad)|x - x_0|,$$

and this is what we claimed. □

2.10 Being Close Versus Looking Like a Line

A minor correction or adjustment of your intuition may be advised in case you developed the misconception that being close to a line is synonymous with looking like a line. We said that a function f is differentiable at a point x_0 if the graph of f is 'close to' a line near the point $(x_0, f(x_0))$. We went on to make this expression 'being close' precise. Based on the examples so far, you may have gotten the impression that this means that the graph "looks like" a line when you zoom in on the point[14]. In fact, this was the case in previous examples. Still, this is not what the definition says.

We illustrate the difference between being close to a line and looking like a line in an example. The function is $f(x) = x^2 \sin(1/x)$. The expression makes no sense for $x = 0$, and we set $f(0) = 0$. This function is differentiable at $x = 0$. Setting $l(x) = 0$, we see that

$$|f(x) - l(x)| = |f(x)| \leq x^2$$

for all $x \in (-\infty, \infty)$, so that the estimate in the definition of differentiability holds. In particular, the tangent line to the graph at $(0,0)$ is the x-axis.

[14]We do not attempt to define what it means to look like a straight line. It may be intuitively clear, but when you try to make this mathematically precise, then you face a formidable task. For the purpose of the discussion we ignore the problems which arise when we want to really zoom in closely on a point, and when we exceed the abilities of the graphing software. This problem might render our discussion useless to begin with.

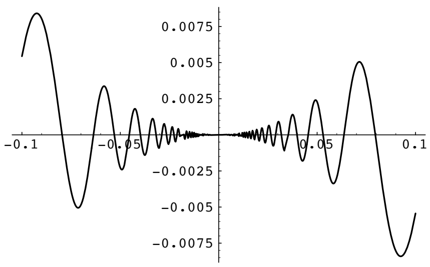

Figure 2.18: $f(x) = x^2 \sin(1/x)$

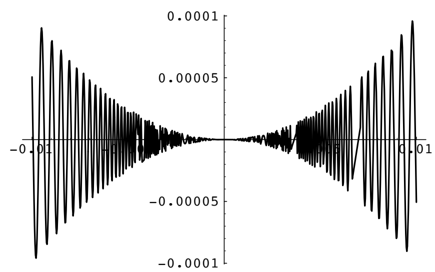

Figure 2.19: $f(x) = x^2 \sin(1/x)$

You see the graph of the function f in Figures 2.18 and 2.19 over two different intervals. Apparently, the graph is trapped between the upper parabola $p(x) = x^2$ and the lower parabola $q(x) = -x^2$. As you can almost see them, particularly in the second picture, we abstained from showing them. Whether we use the calculation from above or the picture, we are convinced of the differentiability of the function at $x = 0$. In our well specified sense, the graph of the function is close to the x-axis. On the other hand, by no stretch of imagination will you say that the graph of the function looks like a line.

You may think of this example as being esoteric. In a way it is. In this sense it is forgivable if you start out with an intuition which needs upgrading later. This is part of learning.

2.11 Rules of Differentiation

There are formulas for calculating the derivative of a composite function from the derivatives of its constituents. These formulas are the topic of this section. These formulas, together with the knowledge of the derivatives of some basic functions, turn the process of differentiation for many functions into an algorithm, a rather mechanical process. You can do it even on the computer, which means that no "understanding" is required. You are expected to learn the basic rules, be able to apply the accurately, and practice many examples. In the last section of this chapter we summarize the computational results of this section. We collect the rules established in this section and tabulate the derivatives of many of the important functions which we considered.

2.11.1 Linearity of the Derivative

The first two rules state that differentiation is compatible with addition of functions and multiplication with a constant. In a more mathematical language one says that differentiation is linear. Let f and g be functions, and assume that both of them are differentiable at x. Let c be a real number. Then $f+g$ and cf are differentiable at x and their derivatives are given by

$$(f+g)'(x) = f'(x) + g'(x) \quad \text{and} \quad (cf)'(x) = cf'(x). \tag{2.37}$$

In Leibnitz' notation this reads

$$\frac{d}{dx}(f+g)(x) = \frac{df}{dx}(x) + \frac{dg}{dx}(x) \quad \text{and} \quad \frac{d}{dx}(cf)(x) = c\frac{df}{dx}(x). \tag{2.38}$$

You may prefer to remember these rules in words. The derivative of a sum of functions is the sum of the derivatives of the function. The derivative of a scalar multiple of function is the multiple of the derivative.

Example 2.37. Differentiate

$$h(x) = x^2 + \sin x.$$

We set $f(x) = x^2$ and $g(x) = \sin x$. Then $h(x) = f(x) + g(x)$. Previously we found that $f'(x) = 2x$ and that $g'(x) = \cos x$. We conclude that

$$h'(x) = \frac{dh}{dx}(x) = 2x + \cos x. \quad \Diamond$$

Example 2.38. Differentiate

$$k(x) = 3\cos x.$$

We set $f(x) = \cos x$ and $c = 3$. Then $k(x) = cf(x)$. We found previously that $f'(x) = -\sin x$. We conclude that

$$k'(x) = \frac{dk}{dx}(x) = -3\sin x. \quad \Diamond$$

Example 2.39. Differentiate $\log_a x$, the logarithm functions for an arbitrary positive base a, $a \neq 1$.

We use the formula for $\log_a x$ from (1.23), $\log_a x = \frac{\ln x}{\ln a}$. In this sense $\log_a x = cf(x)$ where $c = 1/\ln a$ and $f(x) = \ln x$. We stated previously that $\ln' x = 1/x$ (see Theorem 2.13 on page 52). Using the linearity of the derivative, we find

$$\log_a' x = \frac{d}{dx}\left(\frac{\ln x}{\ln a}\right) = \frac{1}{\ln a}\ln' x = \frac{1}{\ln a} \times \frac{1}{x} = \frac{1}{x \ln a}.$$

Specifically we find

$$\log_3' x = \frac{1}{x\ln 3} \quad \text{and} \quad \log_{1/3}' x = \frac{1}{x\ln(1/3)} = -\frac{1}{x\ln 3}.$$

We may even be more specific, and see at what rate the logarithm functions are increasing at a specific point.

$$\log_5' 2 = \frac{1}{2\ln 5} = 0.310667.$$

The numerical value is obtained from a calculator, and exact up to 6 decimal places. The equation says that $\log_5 x$ is increasing at a rate of approximately 0.310667 when $x = 2$. $\Diamond$

Exercise 63. Find the derivatives of the following functions:

(1) $f(x) = 5 + 7\sin x$ (2) $g(x) = 3\log_2(x)$ (3) $h(x) = 3\sin x - 5\cos x$.

Suppose f and g are defined and differentiable on an interval, or a union of intervals. Thinking of f and g more as functions, and not so much as functions evaluated at a point, we may omit (x) from the notation. Then the differentiation rules are

$$(f+g)' = f' + g' \quad \text{or} \quad \frac{d}{dx}(f+g) = \frac{df}{dx} + \frac{dg}{dx} \tag{2.39}$$

and

$$(cf)' = cf' \quad \text{or} \quad \frac{d}{dx}(cf) = c\frac{df}{dx}. \tag{2.40}$$

2.11.2 Product and Quotient Rules

Next we state the product and the quotient rule. They allow us to calculate the derivatives of products and quotients of functions. Again, let f and g be functions, and assume that both of them are differentiable at x. For the quotient rule assume in addition that $g(x) \neq 0$. Then the product fg and the quotient f/g are differentiable at x and their derivatives are given by

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x) \tag{2.41}$$

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}. \tag{2.42}$$

In Leibnitz' notation these formulas become

$$\frac{d}{dx}(fg)(x) = \frac{df}{dx}(x)g(x) + f(x)\frac{dg}{dx}(x) \tag{2.43}$$

$$\frac{d}{dx}\left(\frac{f}{g}\right)(x) = \frac{\frac{df}{dx}(x)g(x) - f(x)\frac{dg}{dx}(x)}{[g(x)]^2}. \tag{2.44}$$

Example 2.40. Differentiate the function

$$h(x) = x^2 \sin x.$$

Write $h(x) = f(x)g(x)$ with $f(x) = x^2$ and $g(x) = \sin x$. In Section 2.3 we worked out that $f'(x) = 2x$ and that $g'(x) = \cos x$ (see Table 2.2). Putting this into the product formula yields

$$h'(x) = f'(x)g(x) + f(x)g'(x) = 2x \sin x + x^2 \cos x. \quad \diamondsuit$$

Exercise 64. Find the derivatives of the following functions:

(1) $f(x) = x \cos x$ (2) $g(x) = x^2 e^x$ (3) $h(x) = x \ln x$ (4) $k(x) = x\sqrt{2x+3}$.

Example 2.41. Differentiate the function

$$k(x) = 1/x.$$

The function is defined for all non-zero real numbers. To differentiate $k(x)$ we set $k(x) = f(x)/g(x)$ with $f(x) = 1$ and $g(x) = x$. Then $f'(x) = 0$ and $g'(x) = 1$, and we find that

$$k'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2} = \frac{-1}{x^2}. \quad \diamondsuit$$

Example 2.42. Show:

(2.45) If $f(x) = x^n$, then $f'(x) = nx^{n-1}$ for $n = 0$, 1, 2, 3, 4, 5, etc

Solution: We learned already the first three cases.

- If $n = 0$, then $f(x) = 1$ (by definition) and $f'(x) = 0$.
- If $n = 1$, then $f(x) = x$ and $f'(x) = 1$.
- If $n = 2$, then $f(x) = x^2$ and $f'(x) = 2x$.

Suppose $f(x) = x^3$. To calculate $f'(x)$, we set $f(x) = g(x)h(x)$ with $g(x) = x^2$ and $h(x) = x$. Previously we found that $g'(x) = 2x$ and $h'(x) = 1$. According to the product rule we find that $f'(x) = g'(x)h(x) + g(x)h'(x) = 2xx + x^2 = 3x^2$. This means:

- If $n = 3$, then $f(x) = x^3$ and $f'(x) = 3x^2$.

Let's push our calculations one n further. Suppose $f(x) = x^4$. We set $f(x) = g(x)h(x)$ with $g(x) = x^3$ and $h(x) = x$. Using the previous calculation we find $f'(x) = g'(x)h(x) + g(x)h'(x) = 3x^2x + x^3 = 4x^3$.

- If $n = 4$, then $f(x) = x^4$ and $f'(x) = 4x^3$.

Exercise 65. Show:

- If $f(x) = x^5$, then $f'(x) = 5x^4$.
- If $f(x) = x^6$, then $f'(x) = 6x^5$.

Proceeding with larger and larger values for n, and in each step using previous results (formally speaking we are doing an induction), we find the general result claimed in (2.45). $\diamondsuit$

Example 2.43. Find the derivative of an arbitrary polynomial.

Solution: A polynomial is a finite sum of multiples of non-negative powers of the variable, i.e., a function of the form

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where the a_i are constants. Using Example 2.42 and the linearity of the derivative we see right away that

$$f'(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1.$$

Here is a specific example, a special case of the formula which we just derived.

$$\text{If } f(x) = 4x^5 - 3x^2 + 4x + 5, \text{ then } \quad f'(x) = 20x^4 - 6x + 4. \quad \diamondsuit$$

Exercise 66. Find the derivatives of the following functions:

$$(1)\ f(x) = 4x^7 - 3x^5 + x^2 - 1 \quad (2)\ g(x) = 8x^5 - 7x^3 + 2x + 1.$$

Example 2.44. Find the derivative of an arbitrary rational function.

Solution: Rational functions are functions of the form

$$r(x) = \frac{p(x)}{q(x)},$$

where $p(x)$ and $q(x)$ are polynomials. We assume, as it is typically done, that $p(x)$ and $q(x)$ do not have any common zeros[15]. The quotient rule tells us now that

$$r'(x) = \frac{p'(x)q(x) - p(x)q'(x)}{[q(x)]^2}.$$

Each of the terms in this formula is known due to Example 2.43. The function $r(x)$ is defined for all x where $q(x) \neq 0$, and the expression for $r'(x)$ is valid for the same values of x.

To be specific, if $r(x) = (x^2 - 5)/(x^3 + 1)$, then

$$r'(x) = \frac{2x(x^3+1) - (x^2-5)3x^2}{(x^3+1)^2} = \frac{-x^4 + 15x^2 + 2x}{(x^3+1)^2}. \quad \diamondsuit$$

Exercise 67. Find the derivatives of the following functions:

$$(1)\ f(x) = \frac{3x+1}{x^2+1} \quad (2)\ g(x) = \frac{x^2+2x+4}{3x-7} \quad (3)\ h(x) = \frac{x^3 - x + 1}{16x^2 - 7x + 4}.$$

Example 2.45. Find the derivative of

$$f(x) = \tan x.$$

Solution: We express $f(x)$ as a quotient of two functions, $f(x) = \sin x / \cos x$, and apply the quotient rule. Use also that $\sin' x = \cos x$ (see

[15] This assumption can be forced in the following sense. Suppose $x = a$ is a common zero of $p(x)$ and $q(x)$. Then $p(x) = p_1(x)(x-a)$ and $q(x) = q_1(x)(x-a)$, where $p_1(x)$ and $q_1(x)$ are once more polynomials. Instead of our initial expression $r(x) = p(x)/q(x)$, we may cancel the common factor $(x-a)$ and replace the expression for $r(x)$ by $p_1(x)/q_1(x)$. We repeat this process of cancelling common factors until the numerator and denominator of the fraction describing $r(x)$ have no common zero anymore.

Example 2.11 on page 50) and $\cos' x = -\sin x$ (see Exercise 37 on page 51), and the identity $\cos^2 x + \sin^2 x = 1$ (see (5.18)). We find

(2.46)
$$\tan' x = \frac{\sin' x \cos x - \sin x \cos' x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x.$$

Some books and computer programs will give this result in a different form. Based on the relevant trigonometric identity, they write

$$\tan' x = 1 + \tan^2 x. \tag{2.47}$$

That draws our attention to the fact that the function $f(x) = \tan x$ satisfies the differential equation

$$f'(x) = 1 + f^2(x). \quad \diamondsuit$$

Example 2.46. Differentiate the function

$$f(x) = \sec x.$$

Solution: We write the function as a quotient: $f(x) = 1/\cos x$. The function is defined for all x for which $\cos x \neq 0$, i.e., for x not of the form $n\pi + 1/2$, where n is an integer. We apply the quotient rule, using that $\cos' x = -\sin x$ (see Exercise 37 on page 51), and that the derivative of a constant vanishes. We find

$$\sec' x = \frac{\sin x}{\cos^2 x} = \frac{\sin x}{\cos x} \cdot \frac{1}{\cos x} = \tan x \sec x. \quad \diamondsuit \tag{2.48}$$

Exercise 68. Find the derivatives of the following functions:

(1) $f(x) = x^2 \tan x$ (2) $g(x) = \cot x$ (3) $h(x) = \dfrac{\tan x}{x^2 + 4}$ (4) $k(x) = x \csc x$.

Suppose f and g are defined and differentiable on an interval, or a union of intervals. Thinking of f and g again more as functions, and not so much as functions evaluated at a point, we may once more omit (x) from the notation. Then the product rule and quotient rule become

$$(fg)' = f'g + fg' \quad \text{or} \quad \frac{d}{dx}(fg) = \frac{df}{dx}g + f\frac{dg}{dx} \tag{2.49}$$

and, wherever $g(x) \neq 0$,

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2} \quad \text{or} \quad \frac{d}{dx}\left(\frac{f}{g}\right) = \frac{\frac{df}{dx}g - f\frac{dg}{dx}}{g^2}. \tag{2.50}$$

Here g^2 is the square of the function g, given by $g^2(x) = [g(x)]^2$.

2.11.3 Chain Rule

The chain rule tells us how to calculate the derivative of a composition of functions. E.g., the function

$$h(x) = \sqrt{1 + 2\cos x}$$

may be written as a composition of two functions. In a first step, we map x to $1 + 2\cos x$ and then we take the radical of the result. Let us denote the first function by g ($g(x) = 1+2\cos x$) and the second one by f ($f(u) = \sqrt{u}$). So we are composing the functions f and g. The mathematical notation for the composition of functions, applied in this situation, is $f \circ g$. In this sense we have

$$h(x) = (f \circ g)(x) = f(g(x)).$$

For this construction to make sense, we must make sure that $f(u)$ is defined whenever $u = g(x)$ for some x in the domain of g. In our case, $f(u)$ is defined only for non-negative numbers u, so we are allowed to take only numbers x so that $1 + 2\cos x$ is non-negative. We need that $\cos x \geq -1/2$. This is the case if $x \in [-2\pi/3, 2\pi/3]$[16]. Using more mathematical terms, we need that the domain of f (the set of points to which f is applied) contains the range of g (the set in which g takes values). If this example was not enough to refresh your memory about compositions of functions, then you are encouraged to read more on this topic in Section 5.7.

The instruction (rule) for the derivative of a composition is now as follows. Let f and g be functions, and suppose that the domain of f contains the range of g, so that $f(g(x))$ is defined for all x in the domain of g. We use the name h for this composite function, so $h(x) = f(g(x))$. The *chain rule* says that whenever g is differentiable at x and f is differentiable at $g(x)$, then

$$h'(x) = (f \circ g)'(x) = f'(g(x))g'(x). \tag{2.51}$$

Here we used once more the notation $\circ$ for the composition of functions. In Leibnitz' notation the chain rule says that

$$\frac{dh}{dx}(x) = \frac{d}{dx} f(g(x)) = \frac{df}{du}(g(x))\frac{dg}{dx}(x). \tag{2.52}$$

[16] We can shift the interval by integer multiples of 2π and get more intervals on which $1 + 2\cos x$ is non-negative.

Example 2.47. Differentiate the function

$$h(x) = e^{x^2+1}.$$

Solution: We write h as a composition of two functions. Set $g(x) = x^2+1$ and $f(u) = e^u$. Then h is the composition of f and g, $h(x) = f(g(x))$. Remember that $f'(u) = f(u) = e^u$ and $g'(x) = 2x$. In particular, $f'(g(x)) = e^{x^2+1}$. The chain rule tells us that

$$h'(x) = f'(g(x))g'(x) = 2xe^{x^2+1}.$$

In the last expression we reversed the order of the factors to make the expression more readable. $\diamondsuit$

Example 2.48. Show that $g(x) = \sqrt{bx+c}$ is differentiable at all x for which $bx + c > 0$, and that the derivative is

$$g'(x) = \frac{b}{2\sqrt{bx+c}}.$$

Remark 6. This exercise carries out the generalization of the special case where $f(x) = \sqrt{x}$, promised in the proof of Proposition 2.15. It also repeats Exercise 41 using general principles instead of brute force.

Solution: Express $g(x)$ as a composition: $g(x) = f(h(x))$ where $h(x) = bx + c$ and $f(u) = \sqrt{u}$. Note that $h(x)$ is differentiable everywhere, and that $f(u)$ is differentiable for $u > 0$, as actually shown in the proof of Proposition 2.15. So $g(x)$ is differentiable at all x for which $bx + c > 0$. In the proof of Proposition 2.15 we did show that $f'(u) = 1/(2\sqrt{u})$. We also know that $h'(x) = b$. According to the chain rule we get

$$g'(x) = f'(g(x))h'(x) = \frac{b}{2\sqrt{bx+c}}. \quad \diamondsuit$$

Exercise 69. Find the derivatives of the following functions:

$$(1)\ f(x) = e^{4x-5} \quad (2)\ g(x) = e^{\cos x} \quad (3)\ h(x) = \sqrt{3x^2-5}.$$

Let us generalize Example 2.42, and not only differentiate the power of a variable, but also the power of a function.

Example 2.49. Combining Example 2.42 with the chain rule we find

$$\frac{d}{dx}u^n(x) = nu'(x)u^{n-1}(x)$$

for all natural numbers n, without any restriction on u, except for the assumption that u is differentiable at x.

Here we decompose the function $u^n(x)$ as a composition of two functions, first mapping x to $u = u(x)$ and then mapping u to its nth power $u^n(x)$. To compare our situation with the chain rule as stated, we set $g(x) = u(x)$ and $f(u) = u^n$. Then

$$h(x) = f(g(x)) = u^n(x).$$

In Example 2.42 we learned how to differentiate nth powers. In particular, $f'(u) = nu^{n-1}$. According to the chain rule:

$$h'(x) = f'(g(x))g'(x) = n(g(x))^{n-1}g'(x) = nu'(x)u^{n-1}(x).$$

We reordered the expressions so that the expression is more readable.

To be specific, here are two concrete examples:

$$\frac{d}{dx}(x^2+1)^{25} = 25(x^2+1)^{24} \cdot 2x = 50x(x^2+1)^{24}$$

and

$$\frac{d}{dx}\tan^3 x = 3\sec^2 x \tan^2 x. \quad \Diamond$$

For practice, let us do a few more examples of this kind.

Example 2.50. Differentiate

$$y(x) = (3x+2)^6.$$

In a brute force approach we could multiply $(3x+2)^6$ out and then use the formula for the derivative of a polynomial to give the answer. Here is a more elegant approach. Write $y(x)$ as a composition of functions. First we map x to $3x+2$, and then take the 6th power of the result. So we write $y(x)$ as $f(g(x))$ with $g(x) = 3x+2$ and $f(u) = u^6$. Then $g'(x) = 3$ and $f'(u) = 6u^5$. Using the chain rule we conclude

$$y'(x) = f'(g(x))g'(x) = 6(3x+2)^5 3 = 18(3x+2)^5. \quad \Diamond$$

Example 2.51. Differentiate the function

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

We may differentiate f by writing it as a composition of functions. First map x to $\cos x$, and then take the square of the result, so $f(x) = g(h(x))$

with $h(x) = \cos x$ and $g(u) = u^2$. We found previously that $h'(x) = -\sin x$ and that $g'(u) = 2u$. This yields

$$f'(x) = g'(h(x))h'(x) = -2\cos x \sin x.$$

We could also have differentiated the function using the product rule, $f(x) = \cos x \cos x$. Certainly we come up with the same answer for the derivative, and you are invited to verify this. $\diamondsuit$

Exercise 70. Find the derivatives of the following functions:

$$(1)\ f(x) = (3x^2 - 1)^{16} \quad (2)\ g(x) = \sin^7 x \quad (3)\ h(x) = \sec^3 x.$$

Example 2.52. Differentiate the function

$$f(x) = e^{\tan x}.$$

We write $f(x)$ as $g(h(x))$ with $h(x) = \tan x$ and $g(u) = e^u$. We found the derivatives of h and g before. In particular, $h'(x) = \sec^2 x$ and $g'(u) = g(u) = e^u$. Then $g'(h(x)) = e^{\tan x}$, and we may conclude that

$$\frac{d}{dx}e^{\tan x} = f'(x) = g'(h(x))g'(x) = e^{\tan x}\sec^2 x. \quad \diamondsuit$$

Generalizing two of the examples from above, we find a more general formula.

Example 2.53. Let $u(x)$ be a differentiable function.

$$\text{If} \quad f(x) = e^{u(x)} \quad \text{then} \quad f'(x) = u'(x)e^{u(x)}.$$

E.g.,

$$\text{If } f(x) = e^{\sin x} \quad \text{then} \quad f'(x) = \cos x e^{\sin x}. \quad \diamondsuit$$

Exercise 71. Find the derivatives of the following functions:

$$(1)\ f(x) = e^{\sec x} \quad (2)\ g(x) = e^{\cot x} \quad (3)\ h(x) = e^{3x^2-5x+1}.$$

Example 2.54. Differentiate the function $\ln|u|$ for $u \neq 0$.

Solution: In Theorem 2.13 on page 52 we stated that $\ln' u = 1/u$ for positive values of u. So, suppose that $u < 0$. Then $u = -|u|$ and $\ln|u| = \ln(-u)$. The chain rule tells us that, for $u < 0$,

$$\frac{d}{du}\ln|u| = \frac{1}{|u|}\frac{d}{du}(-u) = (-1)\frac{1}{-u} = \frac{1}{u}.$$

This means that for all non-zero u

$$\frac{d}{du}\ln|u| = \frac{1}{u}. \quad \diamondsuit$$

In this example we intentionally denoted the variable by u (instead of our more common name x), so that the next example is an immediate consequence of the previous one, using once more the chain rule.

Example 2.55. Let u by a function which is differentiable and nowhere zero on its domain. Then

$$\frac{d}{dx}\ln|u(x)| = \frac{u'(x)}{u(x)}.$$

To be more specific:

$$\frac{d}{dx}\ln|x^2-4| = \frac{2x}{x^2-4}$$

for all $x \neq \pm 2$. $\Diamond$

Let us apply the formula in the last example. We provide two differentiation formulas. The first one is more general, the second one may be a bit easier to comprehend. Before we give the examples, it is important to note:

Remark 7. In the following formulas we make use of the derivative of the exponential function and the logarithm function in an essential way. So far, we have not verified them, and we will have to do this later on. To avoid a circular argument, we have to make sure that we do not rely on the material in the remaining part of this section when we prove the differentiation formulas for these two function.

Example 2.56. Consider a function u which is differentiable and nowhere zero on its domain.

$$\text{If} \quad f(x) = |u(x)|^q \quad \text{then} \quad f'(x) = q\frac{u'(x)}{u(x)}|u(x)|^q. \tag{2.53}$$

Here q can be any real number.

To see this we first rewrite the function f in a different form using the exponential function and its inverse, the natural logarithm.

$$f(x) = e^{\ln f(x)} = e^{\ln(|u(x)|^q)} = e^{q\ln|u(x)|}.$$

Using Examples 2.53 and 2.55 we find

$$f'(x) = \left[\frac{d}{dx}\left(q\ln|u(x)|\right)\right]e^{q\ln|u(x)|} = q\frac{u'(x)}{u(x)}|u(x)|^q.$$

More concretely, let

$$f(x) = |x|^3.$$

As domain for this function we use the set of all non-zero real numbers, i.e., $(-\infty, 0) \cup (0, \infty)$. We set $u(x) = x$ and $q = 3$. Then $u'(x) = 1$ and

$$f'(x) = 3\frac{|x|^3}{x} = 3\frac{x^2|x|}{x} = 3x|x|.$$

Actually, the expression for $f(x)$ makes also sense for $x = 0$, and we may include this point in the domain. So we set $f(0) = 0$ and we still have $f(x) = |x|^3$, but now for all real numbers. Then, based on the definition, we can calculate that $f'(0) = 0$, and we obtain for all real numbers that:

$$\text{If } f(x) = |x|^3 \quad \text{then} \quad f'(x) = 3x|x|.$$

Here is another concrete example:

$$\frac{d}{dx}\left|\frac{1}{2} - \sin x\right|^5 = 5\frac{-\cos x}{\frac{1}{2} - \sin x}\left|\frac{1}{2} - \sin x\right|^5$$

whenever $\sin x \neq 1/2$. Specifically, we have to exclude all x of the form $\frac{\pi}{6} + 2n\pi$ and $\frac{5\pi}{6} + 2n\pi$, where n is an arbitrary integer. $\diamondsuit$

Exercise 72. Find the derivatives of the following functions:

(1) $f(x) = \ln|3x^2 - 5|$ (2) $g(x) = |\sin x - 3|^5$ (3) $h(x) = |x^2 - 4x - 1|^3$.

In each case, determine for which values of x the formula for the derivative holds.

Example 2.57. Consider a function u which is differentiable and everywhere positive on its domain, and let f be its qth power. So,

$$f(x) = u^q(x).$$

Here q can be any real number. Then

$$f'(x) = qu'(x)u^{q-1}(x).$$

The calculation is the same as the one we used to show (2.53). We can omit absolute value signs everywhere, and that allows us to cancel a power of $u(x)$ in the formula.

To be more concrete, let us differentiate $x^{1/3}$. As domain for this function we use $(0, \infty)$. Applied in this special case the formula says that

$$\frac{d}{dx}\left(x^{\frac{1}{3}}\right) = \frac{1}{3}x^{\frac{1}{3}-1} = \frac{1}{3}x^{-\frac{2}{3}}.$$

More generally, suppose $x > 0$ and q is any real number, then:

$$\frac{d}{dx}\left(x^q\right) = qx^{q-1}. \quad \diamondsuit$$

Example 2.58. Differentiate

$$h(x) = (\sin x)^{1/2} \quad \text{for} \quad x \in (0, \pi).$$

First, observe that $\sin x$ is positive on the interval, so we may apply the formula in Example 2.57:

$$f'(x) = \frac{1}{2}\cos x(\sin x)^{-1/2} = \frac{\cos x}{2\sqrt{\sin x}}. \quad \diamondsuit$$

Exercise 73. Find the derivatives of the following functions:

(1) $f(x) = (1 + 3x^2)^{3/2}$ (2) $g(x) = (\sin^2 x + 5)^{7/3}$ (3) $h(x) = (\sec^2 x + 5)^{\pi}$.

Example 2.59. We had exponential functions not only for the base e, but for any base a, where $a > 0$ and $a \neq 1$. Let us differentiate

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

Solution: According to the definition, $f(x) = e^{x \ln a}$. Using the chain rule we find

$$\frac{d}{dx}a^x = (\ln a)e^{x \ln a} = a^x \ln a.$$

To be absolutely concrete:

$$\frac{d}{dx}2^x = 2^x \ln 2$$

and

$$\frac{d}{dx}\left(\frac{1}{2}\right)^x = \left(\frac{1}{2}\right)^x \ln(1/2) = -(\ln 2)\left(\frac{1}{2}\right)^x. \quad \diamondsuit$$

Example 2.60. For $x > 0$, differentiate the function

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

Solution: This is not much harder than the problem in the previous example. We first write the function slightly differently,

$$f(x) = e^{x \ln x}.$$

Then we use the formula in Example 2.53 on page 94. We set $u(x) = x \ln x$, and differentiate this function using the product rule.

$$u'(x) = \ln x + x \ln' x = \ln x + x \frac{1}{x} = 1 + \ln x.$$

Then

$$\frac{d}{dx} x^x = u'(x) e^{u(x)} = (1 + \ln x) x^x. \quad \Diamond$$

Exercise 74. Find the derivatives of the following functions, and specify where your formula holds:

$$(1)\ f(x) = 5^x \quad (2)\ g(x) = x^{\sin x} \quad (3)\ h(x) = 3^{\cos x}.$$

Example 2.61. As an introductory example to the definition of the derivative (see Example 2) we discussed the derivative of the function

$$f(x) = \sqrt{1 - x^2}.$$

We consider this function on the open interval $(-1, 1)$. We differentiate the function using the chain rule, and for this purpose we decompose f as a composition of two functions. First we map x to $u(x) = 1 - x^2$, and then we map u to $h(u) = \sqrt{u}$. We learned that $u'(x) = -2x$, and that $h'(u) = 1/(2\sqrt{u})$. This means that

$$f'(x) = \frac{-x}{\sqrt{1 - x^2}}.$$

With this we have not only verified that the function is differentiable on the interval $(-1, 1)$, but with our calculation we have also confirmed that the slope of the tangent line to the circle is as we predicted it in Example 2 based on geometric arguments. $\Diamond$

Example 2.62. It may happen to us that a function is naturally written as a composition of more than two functions, say

$$F(x) = e^{\sqrt{x^2+1}}.$$

Here we map x to $u = x^2 + 1$ by a function which we call h, then we map u to $v = \sqrt{u}$ and call this function g, and finally we send v to e^v and call this function f. So F is the composition of the functions f, g, and h, or

$$F(x) = f(g(h(x))).$$

We can gather g and h into one function G, so $G(x) = \sqrt{x^2 + 1}$. Then we apply the chain rule twice, once to differentiate $G = g \circ h$, and once to differentiate $F = f \circ G$. We find:

$$F'(x) = f'(G(x))G'(x) \quad \text{and} \quad G'(x) = g'(h(x))h'(x).$$

This can be combined as

$$F'(x) = f'(g(h(x)))g'(h(x))h'(x).$$

Let us return to the specific example. Obviously $h'(x) = 2x$ and $g'(u) = \frac{1}{2\sqrt{u}}$. We also learned that $f'(v) = f(v)$. Putting all of this together, cancelling a factor 2, and writing the expressions in an order which makes it easy to read, we find

$$F'(x) = \frac{x}{\sqrt{x^2+1}} e^{\sqrt{x^2+1}}. \quad \Diamond$$

In the previous example we demonstrated how to calculate the derivative of a composition of three functions. The process did not depend on the specific example, and we may state our result in more generality. Let F be a function of three differentiable functions, which we call f, g, and h. So

$$F(x) = f(g(h(x))).$$

Then

$$F'(x) = f'(g(h(x)))g'(h(x))h'(x). \tag{2.54}$$

If we like to write this formula using Leibnitz' notation, then we need to give names to the variables of the functions. Denote the variable of f by v and the one of g by u. The variable of h was called x. Then the chain rule for a composition of three functions is

$$\frac{dF}{dx}(x) = \frac{df}{dv}(g(h(x)))\frac{dg}{du}(h(x))\frac{dh}{dx}(x). \tag{2.55}$$

Exercise 75. Find the derivatives of the following functions:

(1) $f(x) = e^{\sin(x^2+1)}$

(2) $g(x) = (\sin^3(x^2+7)+5)^{4/11}$

(3) $h(x) = \tan^3(5x^2 - 3x + 5)$

(4) $k(x) = (\csc^4(\cos^2 x + 3) + 3x)^{5/7}$.

Example 2.63. Differentiate the function

$$F(x) = \tan(\cos(\sqrt{x^4+2x+5})).$$

You may convince yourself that $x^4+2x+5 > 0$ for all real numbers x, so that the radical is defined for all x as well. (Use any means which come to your mind, if necessary depend on technology to graph the function.) Because $|\cos u| \le 1$ and $\tan v$ is defined if $|v| \le 1$, we find that $F(x)$ is defined for all $x \in (-\infty, \infty)$. You find the graph of $F(x)$ for $x \in [-3, 3]$ in Figure 2.20.

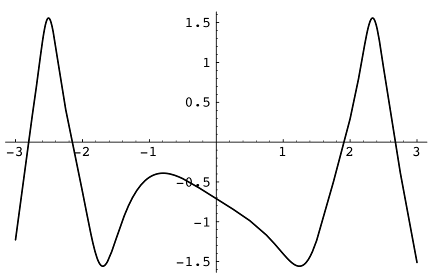

Figure 2.20: The function $F(x) = \tan(\cos(\sqrt{x^4+2x+5}))$

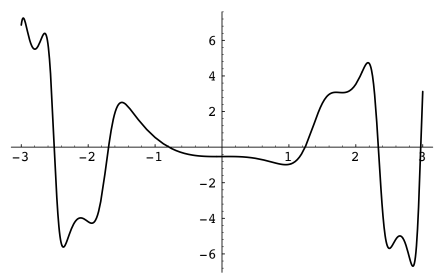

Figure 2.21: The derivative of $F(x) = \tan(\cos(\sqrt{x^4+2x+5}))$

To differentiate the function F we break it up into a composition of three functions, writing $F(x) = f(g(h(x)))$, where $h(x) = \sqrt{x^4+2x+5}$, $g(u) = \cos u$, and $f(v) = \tan v$. You learned previously that $g'(u) = -\sin u$ and $f'(v) = \sec^2 v$. Using the chain rule you also find that

$$h'(x) = \frac{4x^3+2}{2\sqrt{x^4+2x+5}} = \frac{2x^3+1}{\sqrt{x^4+2x+5}}.$$

Now we apply the chain rule and find

$$\begin{aligned} F'(x) &= f'(g(h(x)))g'(h(x))h'(x) \\ &= -\sec^2\left(\cos(\sqrt{x^4+2x+5})\right)\sin\left(\sqrt{x^4+2x+5}\right)\frac{2x^3+1}{\sqrt{x^4+2x+5}}. \end{aligned}$$

We graphed this function in Figure 2.21. $\diamondsuit$

Apparently, we could go on and on making more difficult examples. This is not our goal. You need to understand the basic tools used to compute derivatives, and that is what you were supposed to practice with the help of the examples in this section.

2.11.4 Derivatives of Inverse Functions

Intuitively, it should be clear what happens when we differentiate the inverse of a function[17]. To obtain the graph of the inverse of a function, we take the graph of the function and reflect it at the diagonal. The same applies to the tangent line to the graph of a function. This allows us to determine the derivative of the inverse function. Let us look at an example first. After having discussed the example we will determine in general where the inverse of a differentiable function is differentiable and what the derivative is.

Example 2.64. The functions $f(x) = x^3$ and $g(x) = x^{1/3}$ are inverses of each other. To see this we check that $f(g(x)) = [x^{1/3}]^3 = x$ and $g(f(x)) = x$ for all real numbers x. We restrict ourselves to the domain $(0,1)$ for both functions, f and g. We also use $(0,1)$ also as range for both of them. They are still inverses of each other because the domain of f is the range of g and vice versa. You are invited to verify this. You may rely on the graphs of the functions which are shown in Figures 2.22 and 2.23. You should observe that one figure is the reflection of the other one at the diagonal.

Let us take some point $x \in (0,1)$. We found earlier that $f'(x) = 3x^2$. Let us also take a point $y \in (0,1)$, then $g'(y) = \frac{1}{3}y^{-\frac{2}{3}}$. We used different names for the variables of f and g so that we can distinguish them. For $y = f(x)$ we find that

$$g'(f(x)) = \frac{1}{3}\left(x^3\right)^{-\frac{2}{3}} = \frac{1}{3}x^{-2} = \frac{1}{3x^2}$$

[17] You may want to review the concept of the inverse of a function, and you can do so by reading Section 5.6.

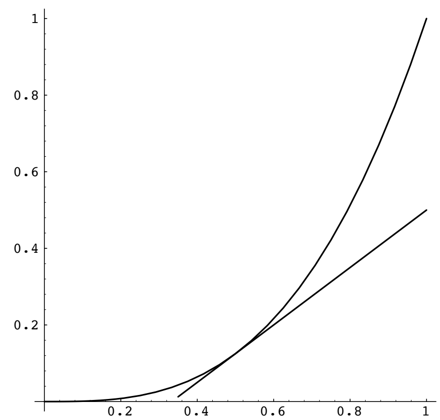

Figure 2.22: $f(x) = x^3$

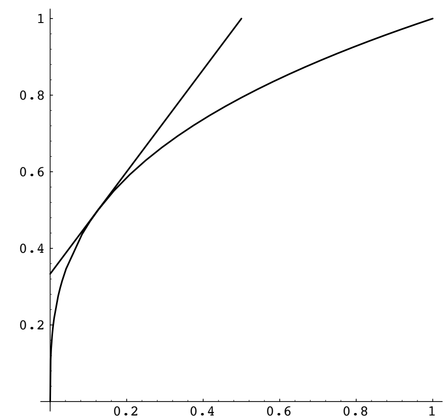

Figure 2.23: $g(x) = x^{1/3}$

so that

$$g'(f(x)) = \frac{1}{f'(x)}. \tag{2.56}$$

Let us give a numerical example. Say $x = 1/2$ and $f(x) = y = 1/8$. Then

$$f'(x) = \frac{3}{4} \quad \text{and} \quad g'(y) = \frac{1}{3}\left(\frac{1}{8}\right)^{-\frac{2}{3}} = \frac{4}{3}.$$

You see that $g'(f(x)) = 1/f'(x)$. In the figures you also see the tangent line l_1 to $f(x)$ at the point $(1/2, 1/8)$ and the tangent line l_2 to $g(y)$ at the point $(1/8, 1/2)$. The equations of the tangent lines are

$$l_1(x) = f'\left(\frac{1}{2}\right)\left(x - \frac{1}{2}\right) + \frac{1}{8} \quad \text{and} \quad l_2(y) = g'\left(\frac{1}{8}\right)\left(y - \frac{1}{8}\right) + \frac{1}{2}.$$

After putting in the values for $f'(1/2)$ and $g'(1/8)$ we have

$$l_1(x) = \frac{3}{4}\left(x - \frac{1}{2}\right) + \frac{1}{8} = \frac{3}{4}x - \frac{1}{4} \text{ and } l_2(y) = \frac{4}{3}\left(y - \frac{1}{8}\right) + \frac{1}{2} = \frac{4}{3}y + \frac{1}{3}.$$

Let us think geometrically for a moment. The tangent line to the graph at a point is a line which is close to the graph near that point. This property stays unchanged when we reflect the graph of the function and the tangent line at the diagonal of the coordinate system, in other words, if we invert the function. A line with slope m, reflected at the diagonal, will turn into a line with slope $1/m$, and this is exactly what is expressed in (2.56). You are now invited to also verify that, as we should expect, l_1 and l_2 are inverses of each other. In other words,

$$l_2(l_1(x)) = x \quad \text{and} \quad l_1(l_2(y)) = y. \quad \diamondsuit$$

Exercise 76. The inverse of $\sin x$ is called $\arcsin y$.

1. Find the equation of the tangent line to the graph of $f(x) = \sin x$ at the point $(\pi/6, 1/2)$.

2. Use geometric reasoning as in Example 2.64 to find the tangent line to the graph of $\arcsin y$ at the point $(1/2, \pi/6)$.

The following theorem is the key tool for the upcoming discussion. We will also apply it in our discussion of Newton's method.

Theorem 2.65. [**Intermediate Value Theorem**] *Let f be a differentiable (or continuous[18]) function and suppose that its domain contains the closed interval $[a, b]$. Let C be any number between $f(a)$ and $f(b)$. Then there exists a number c, where $a \le c \le b$, such that $f(c) = C$.*

This important result is typically discussed in a real analysis course. It is a consequence of the completeness of the real numbers.

It will be convenient to have a characterization of intervals. A subset J of the real line is an interval if, whenever a, $b \in J$ and $a \le c \le b$, then $c \in J$. The following two corollaries are consequences of the Intermediate Value Theorem.

Corollary 2.66. *Let f be a differentiable (or continuous) function and I an interval which in contained in the domain on f. Then the image of I is an interval.[19]*

Corollary 2.67. *Let f be a differentiable (or continuous) invertible function which is defined on an interval I. Then f is either increasing or f is decreasing.*

Using these two corollaries one can deduce:

[18] We did not, and not not wish to, define continuous functions. Every differentiable function is continuous, but not every continuous function is differentiable.

[19] In general, the image of an open interval need not be open. E.g., as you will learn later to work out, the function $f(x) = x(x-1)(x+1)$ maps the open interval $(-1, 1)$ to the closed interval $[-2\sqrt{3}/9, 2\sqrt{3}/9]$.

Theorem 2.68. *Let f be a differentiable and invertible function which is defined on an open interval (a, b). Then the image of f is an open interval (A, B).*

In this theorem a is allowed to be $-\infty$ and b to be ∞. It can happen that A is $-\infty$ and that B is ∞. You are invited to check the following examples.

Exercise 77. Verify the following:

1. If $f(x) = x^2$ and the domain is $(0, \infty)$, then the image is $(0, \infty)$.

2. If $f(x) = 1/x$ and the domain is $(0, \infty)$, then the image is $(0, \infty)$.

3. If $f(x) = \sin x$ and the domain is $(-\frac{\pi}{2}, \frac{\pi}{2})$, then the image is $(-1, 1)$.

4. If $f(x) = \cos x$ and the domain is $(0, \pi)$, then the image is $(-1, 1)$.

5. If $f(x) = e^x$ and the domain is $(-\infty, \infty)$, then the image is $(0, \infty)$.

6. If $f(x) = \ln x$ and the domain is $(0, \infty)$, then the image is $(-\infty, \infty)$.

7. If $f(x) = \tan x$ and the domain is $(-\frac{\pi}{2}, \frac{\pi}{2})$, then the image is $(-\infty, \infty)$. You see a graph of the function in Figure 2.24.

8. If $f(x) = \arctan x$ (the inverse of the tangent function) and the domain is $(-\infty, \infty)$, then the image is typically[20] taken as $(-\frac{\pi}{2}, \frac{\pi}{2})$. You see a graph of the function in Figure 2.25.

Let f be as in Theorem 2.68. Then the inverse of f is a function g which is defined on an open interval (A, B). The concept of differentiability was defined on (unions of) open intervals, so that we may ask whether, or where, g is differentiable. The answer is as follows.

Theorem 2.69. *Let f be a differentiable and invertible function which is defined on an open interval (a, b), and denote the image of f by (A, B). Denote the inverse of f by g. Then g is differentiable at all points $y \in (A, B)$ for which $f'(g(y)) \neq 0$. For these values of y and for x such that $f(x) = y$ the derivative is given by:*

$$g'(y) = \frac{1}{f'(g(y))} \quad \textit{or} \quad g'(f(x)) = \frac{1}{f'(x)}.$$

[20] Other choices are possible.

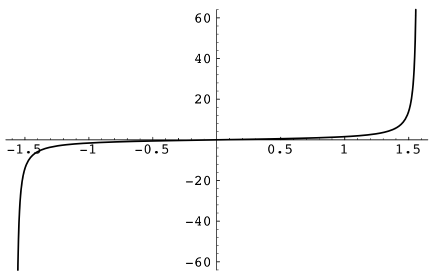

Figure 2.24: $\tan x$ on $(-\pi/2, \pi/2)$

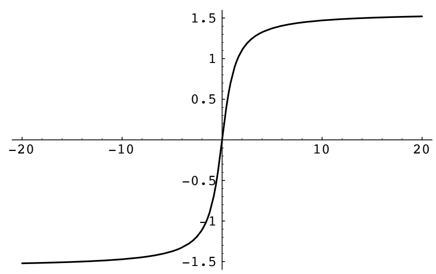

Figure 2.25: $\arctan x$ on $(-20, 20)$

Proof. We will not show the differentiability of g. Assuming it, we verify the formula for g'. By definition we have

$$f(g(y)) = y$$

for all $y \in (A, B)$. Differentiate both functions, the left hand side and right hand side of the equation. When differentiating the composition of f and g we apply the chain rule. We find

$$f'(g(y))g'(y) = 1 \quad \text{and} \quad g'(y) = \frac{1}{f'(g(y))},$$

as claimed. If $y = f(x)$, then $g(y) = g(f(x)) = x$, and we obtain the second version of the formula for the derivative of the inverse of the function:

$$g'(f(x)) = \frac{1}{f'(x)}.$$

□

We apply the theorem to find some important derivatives.

Example 2.70. Show that the exponential function is differentiable and that

$$\frac{d}{dy} e^y = e^y.$$

By definition, the exponential function is the inverse of the natural logarithm function ln. The natural logarithm function is differentiable and $\ln' x = 1/x$, see Theorem 2.13. Set $f(x) = \ln x$ and $g(y) = e^y$ in Theorem 2.69. We note that $\ln'(x) \neq 0$ for all x in $(0, \infty)$, the domain of the natural logarithm. First of all, the theorem says that the exponential function is differentiable. Secondly, the theorem provides the formula for the derivative. Specifically, we calculate that

$$\frac{d}{dy} e^y = \frac{1}{\ln'(e^y)} = \frac{1}{1/e^y} = e^y,$$

as claimed. $\diamond$

Remark 8. In the previous example we proved at least part of Theorem 2.12, assuming Theorem 2.13. Combined with the chain rule, we find that the function $f(x) = e^{ax}$ is differentiable, and that $f'(x) = ae^{ax}$. It will take a little longer before we can prove Theorem 2.13.

Example 2.71. Show that the function $g(y) = \arctan y$ (the inverse of $f(x) = \tan x$) is differentiable, and that

$$\frac{d}{dy} \arctan y = \frac{1}{1 + y^2}.$$

As domain for $\tan x$ we use the interval $(-\pi/2, \pi/2)$, and as its domain we use $(-\infty, \infty)$. Accordingly, the domain for $g(y) = \arctan y$ is the interval $(-\infty, \infty)$, and the range for this function is $(-\pi/2, \pi/2)$. You see the graph of the arctangent function in Figure 2.25 on the page before.

Solution: The function $f(x) = \tan x$ is differentiable on its entire domain, and $f'(x) = \sec^2 x$ is nowhere zero. Theorem 2.69 tells us that $g(y) = \arctan y$ is differentiable on the entire domain of this function, i.e., on the interval $(-\infty, \infty)$. The theorem also provides us with the formula for the derivative:

$$\arctan'(y) = \frac{1}{\tan'(\arctan y)} = \frac{1}{\sec^2(\arctan y)} = \cos^2(\arctan y).$$

All we need to do now is to figure out what $\cos^2(\arctan y)$ is. To do this we draw a triangle in which we identify the available data. We refer to the notation in Figure 2.26.

There you see a rectangular triangle, the right angle is at the vertex B. The angle at the vertex A is called u. The adjacent side to this angle is chosen to be of length 1, and the opposing side of length y. So, by definition,

$$\tan u = y \quad \text{and} \quad \arctan y = u.$$

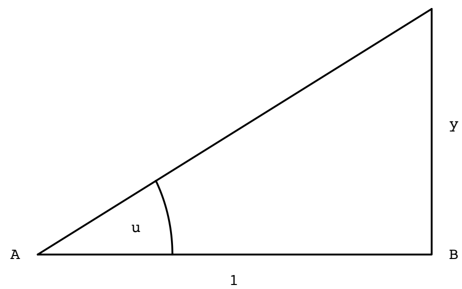

Figure 2.26: An informative triangle

By the theorem of Pythagoras, the length of the hypotenuse is $\sqrt{1+y^2}$. Then

$$\cos u = \frac{1}{\sqrt{1+y^2}} \quad \text{and} \quad \cos^2(\arctan y) = \frac{1}{1+y^2}.$$

The conclusion is that

$$\arctan'(y) = \frac{1}{1+y^2}. \tag{2.57}$$

This is exactly what we claimed.

Combined with the chain rule, we find a slightly more general formula. Suppose $u(x)$ is a differentiable function on some open interval (a, b). Then, on this interval,

$$\frac{d}{dx}\arctan(u(x)) = \frac{u'(x)}{1+u^2(x)}. \tag{2.58}$$

E.g.,

$$\text{if } f(x) = \arctan(x^2+5), \text{ then } f'(x) = \frac{2x}{1+(x^2+5)^2},$$

and

$$\text{if } f(x) = \arctan(\sin x), \text{ then } f'(x) = \frac{\cos x}{1+\sin^2 x}. \quad \Diamond$$

Exercise 78. Find the derivatives of the following functions:

(1) $f(x) = \arctan(5x - 2)$	(3) $h(x) = \operatorname{arccot} x$	(5) $j(x) = \operatorname{arccot} x^2$
(2) $g(x) = \arctan(\cos x)$	(4) $i(x) = 1/\arctan x$	(6) $k(x) = \arctan(e^x)$

In (3), $\operatorname{arccot} y$ denotes the arc-cotangent function, the inverse of the function $\cot x$. To solve the problem in (3), you may try to modify the calculation of $\arctan' y$. You may also fill in the details in the following argument. The trigonometric identities imply that $\cot x = -\tan(x - \pi/2)$. Hence $\operatorname{arccot} y$ and $-\arctan y$ differ by a constant. In particular, they have the same derivative. Then

$$\operatorname{arccot}' y = -\arctan y.$$

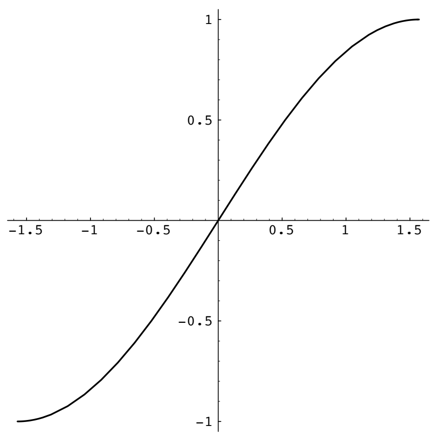

Figure 2.27: $\sin x$ on $[-\pi/2, \pi/2]$

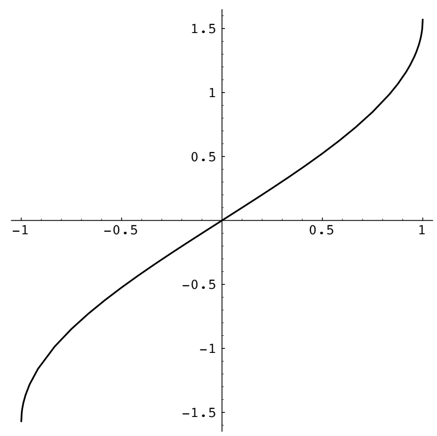

Figure 2.28: $\arcsin y$ on $[-1, 1]$

Example 2.72. Discuss the arcsine function ($\arcsin y$) and show that

$$\arcsin'(y) = \frac{1}{\sqrt{1 - y^2}}.$$

By definition, the arcsine function is the inverse of the sine function. Instead of the notation arcsin, you may also find the notation $\sin^{-1}$ for this function. In this case the superscript -1 indicates that we take the inverse of the function. You see the graphs of both functions in Figures 2.27 and

2.28. For the purpose of the differentiability discussion, we use the interval $(-\pi/2, \pi/2)$ as domain of the function $\sin x$ and as range for the function $\arcsin y$. We use the interval $(-1, 1)$ as range domain of the function $\sin x$ and as domain for the function $\arcsin y$.[21]

Solution: The cosine function, the derivative of the sine function, is nonzero on the interval $(-\pi/2, \pi/2)$, and we may conclude from Theorem 2.69 that arcsin is differentiable on $(-1, 1)$. The theorem also tells us what the derivative is:

$$\arcsin'(y) = \frac{1}{\sin'(\arcsin(y))} = \frac{1}{\cos(\arcsin(y))}.$$

This expression does not give an easy expression for $\arcsin'(y)$, and we can improve on it, using the information in a triangle similar to the one used in the previous example. We use the triangle shown in Figure 2.29.

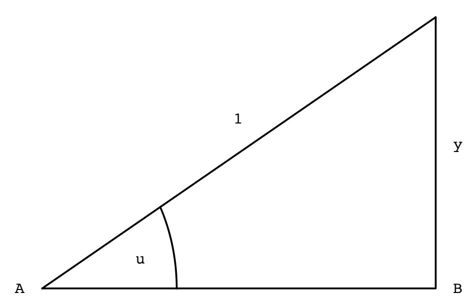

Figure 2.29: An informative triangle

According to our choices, $\sin u = y$ and $u = \arcsin y$. This means that the adjacent side to the angle u is $\cos u = \cos(\arcsin y)$. The theorem of Pythagoras tells us that $\cos u = \sqrt{1 - y^2}$, and this means that

$$\arcsin'(y) = \frac{1}{\sqrt{1 - y^2}}. \tag{2.59}$$

[21] For the purpose of definition, we could have included the end points of the intervals, but at $y = \pm 1$ $\arcsin y$ is not differentiable because $\sin'(x) = 0$ when $x = \pm\pi/2$.

In Leibniz' notation, and using $\sin^{-1}$ to denote the inverse function of sin, we get

$$\frac{d}{dx}\sin^{-1}(x) = \frac{1}{\sqrt{1-x^2}}.$$

We may once more improve on this formula. Let $u(x)$ be a differentiable function which is defined on an open interval, and suppose that $|u(x)| < 1$. Then, using the chain rule, we find that

$$\frac{d}{dx}\arcsin(u(x)) = \frac{u'(x)}{\sqrt{1-u^2(x)}}. \tag{2.60}$$

E.g., for $x \in (-1/3, 1/3)$ we have

$$\frac{d}{dx}\arcsin(3x) = \frac{3}{\sqrt{1-9x^2}},$$

and for $x \in (-1, 1)$ we have

$$\frac{d}{dx}\arcsin(x^2) = \frac{2x}{\sqrt{1-x^4}}. \quad \diamond$$

Exercise 79. Find maximal open intervals on which the following functions are defined and find their derivatives:

$$(1)\ f(x) = \arcsin(x^2-2) \quad (2)\ g(x) = \arcsin(\tan x) \quad (3)\ h(y) = \arccos y.$$

Here $\arccos y$ denotes the arccosine function, the inverse of $\cos x$. We consider it as a function with domain $(-1, 1)$ and image $(0, \pi)$. To solve the problem in (3), you may try to modify the calculation of $\arcsin'$. You may also fill in the details in the following argument. We know that $\sin x = -\cos(x + \pi/2)$. Hence $\arcsin y$ and $-\arccos y$ differ by a constant. In particular, they have the same derivative. Then

$$\arcsin' y = -\arccos y.$$

Remark 9. The formula for the derivatives of $\arcsin y$ and $\arccos y$ depends on which range or image we choose for the function. E.g., if we consider $\arcsin y$ as a function with domain $(-1, 1)$ and image $(\pi/2, 3\pi/2)$, the

$$\arcsin'(y) = \frac{-1}{\sqrt{1-y^2}}.$$

2.12 Implicit Differentiation

Until now we considered functions which were given explicitly. I.e., we were given an equation $y = f(x)$, where $f(x)$ is some instruction which assigns a value to x. The points on the graph of f are the points which satisfy the equation. Consider the equation

$$(x^2 + y^2)^2 = x^2 - y^2. \tag{2.61}$$

The solutions of this equation form a curve in the plane called a lemniscate, see Figure 2.30. Parts of this curve look like the graph of a function, such as the points for which $y \geq 0$. Without solving the equation for y, we still like to calculate the slope of curve at one of its points. This process is called *implicit differentiation.*

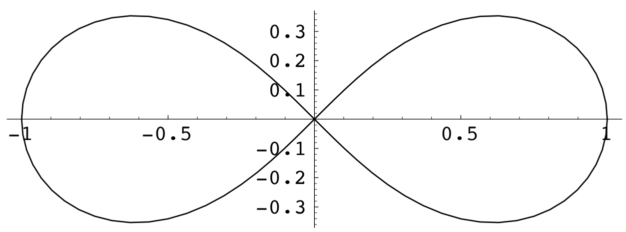

Figure 2.30: Lemniscate

Let us consider an easy situation which we have studied before.

Example 2.73. Find the slope of the tangent line to the unit circle (the curve consisting of all points which satisfy the equation $x^2 + y^2 = 1$) at the point $(1/2, \sqrt{3}/2)$.

Solution: We consider y as a function of x, and in this sense we write $y = y(x)$[22], and differentiate both sides of the equation. Apparently, $\frac{d}{dx}x^2 = 2x$. From the chain rule we deduce that $\frac{d}{dx}y^2 = 2y\frac{dy}{dx}$. That means that the derivative of the left hand side of the equation with respect to x is $2x+2y\frac{dy}{dx}$. The derivative of the right hand side is zero. The derivative of the left and right hand side of the equation have to be the same, so that we get

$$2x + 2y\frac{dy}{dx} = 0.$$

Solving the equation for $\frac{dy}{dx}$, we find

$$\frac{dy}{dx} = \frac{-x}{y}.$$

Plugging in the coordinates of the specified point, we find that

$$\left.\frac{dy}{dx}\right|_{(1/2,\sqrt{3}/2)} = \frac{-1}{\sqrt{3}}.$$

As we had to specify the x and the y coordinate of the point, we use a slightly different way to indicate at which point we evaluate the derivative. $\diamondsuit$

Example 2.74. Suppose you drop a circle of radius 1 into a parabola with the equation $y = 2x^2$. At which points will the circle touch the parabola?[23]

Solution: You see a picture of the problem in Figure 2.31. The crucial observation in this example is, that the tangent line to the parabola and the circle will be the same at the point of contact.

Suppose the coordinates of the center of the circle are $(0, a)$, then its equation is $x^2 + (y - a)^2 = 1$. Differentiating the equation of the parabola with respect to x, we find that $\frac{dy}{dx} = 4x$. Differentiating the equation of the circle with respect to x, we get

$$2x + 2(y - a)\frac{dy}{dx} = 0.$$

Assuming that $\frac{dy}{dx}$ is the same for both curves at the point of contact, we substitute $\frac{dy}{dx} = 4x$ into the second equation, cancel a factor 2, factor out an x, and find:

$$x(1 + 4(y - a)) = 0.$$

[22] We can do this only for part of the curve as y is not really a function of x. For most x there are two values of y which satisfy the equation.

[23] More sensibly, drop a ball of radius 1 into a cup whose vertical cross section is the parabola $y = 2x^2$.

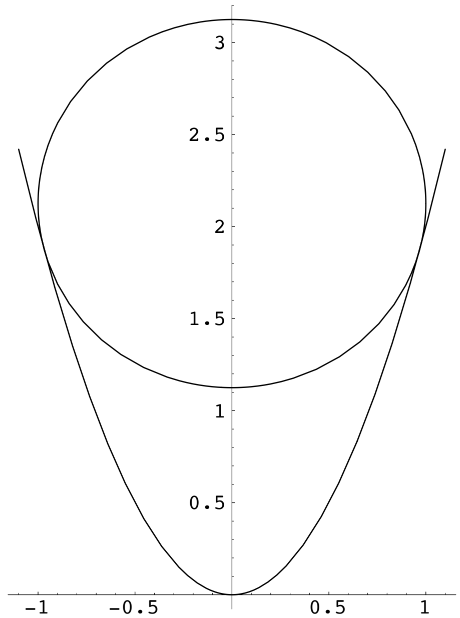

Figure 2.31: Ball in a Cup.

The ball it too large to fit into the parabola and touch at $(0, 0)$. So we may assume that $x \neq 0$. Solving the equation $1 + 4(y - a) = 0$ for y, we find that the y coordinate of the point of contact is $y = a - \frac{1}{4}$. We substitute this expression into the equation of the circle and find that the x coordinate of the point of contact is $x = \pm\frac{\sqrt{15}}{4}$. Substituting this into the equation of the parabola, we find that $y = \frac{15}{8}$ at the point of contact. In summary, the circle touches the parabola in the points

$$(x, y) = \left(\pm\frac{\sqrt{15}}{4}, \frac{15}{8}\right). \quad \Diamond$$

Example 2.75. Find the slope of the tangent line to the lemniscate

$$(x^2 + y^2)^2 = x^2 - y^2,$$

and find the coordinates of the points where the tangent line is horizontal.

Solution: You see a picture of the lemniscate in Figure 2.30. As in Example 2.73, we equate the derivatives of the left and right hand side of the equation. We consider y as a function of x. Using standard rules of differentiation, we find

$$2(x^2+y^2)(2x+2y\frac{dy}{dx}) = 2x - 2y\frac{dy}{dx}.$$

Cancelling a factor 2 and multiplying out part of the left hand side of the equation, we find

$$2x(x^2+y^2) + 2y(x^2+y^2)\frac{dy}{dx} = x - y\frac{dy}{dx}.$$

Gathering all terms with a factor $\frac{dy}{dx}$ on the left and those without on the right, we find the equation

$$(2y(x^2+y^2)+y)\frac{dy}{dx} = x(1-2(x^2+y^2)).$$

Finally we get an explicit expression for $\frac{dy}{dx}$ in terms of x and y:

$$\frac{dy}{dx} = \frac{x(1-2(x^2+y^2))}{2y(x^2+y^2)+y} = \frac{x(1-2(x^2+y^2))}{y(2(x^2+y^2)+1)}.$$

Given any point (x,y) with $y \neq 0$ on the lemniscate, we can plug it into the expression for $\frac{dy}{dx}$ and we get the slope of the curve at this point.

E.g, the point $(x,y) = (\frac{1}{2}, \frac{1}{2}\sqrt{-3+2\sqrt{3}})$ is a point on the lemniscate, and at this point the slope of the tangent line is

$$\frac{dy}{dx} = \frac{2-\sqrt{2}}{\sqrt{3}\sqrt{-3+2\sqrt{3}}}.$$

This specific calculation takes a bit of arithmetic skill and effort to carry out.

The tangent line is horizontal whenever $\frac{dy}{dx} = 0$. A quick look at Figure 2.30 tells us that we may ignore points where $x = 0$ or $y = 0$. That means that $\frac{dy}{dx} = 0$ whenever

$$1 - 2(x^2+y^2) = 0 \quad \text{or} \quad x^2+y^2 = \frac{1}{2}.$$

Substitute $x^2+y^2 = \frac{1}{2}$, and $y^2 = \frac{1}{2} - x^2$ into the equation of the curve. Then we get an equation in one variable:

$$\frac{1}{4} = x^2 - \left(\frac{1}{2} - x^2\right) \quad \text{or} \quad x^2 = \frac{3}{8} \quad \text{and} \quad y^2 = \frac{1}{8}.$$

The points at which the tangent line to the lemniscate is horizontal are

$$(x,y) = (\pm\frac{\sqrt{6}}{4}, \pm\frac{\sqrt{2}}{4}) \approx (\pm.6124, \pm.3536). \quad \Diamond$$

Exercise 80. Consider the curve given by the equation $y^2 = x^3$. Find the slope of the curve at the point $(x, y) = (4, 8)$.

Exercise 81. Consider the curve given by the equation

$$x^3 + y^3 = 1 + 3xy^2.$$

Find the slope of the curve at the point $(x, y) = (2, -1)$.

Exercise 82. Consider the curve given by the equation $x^2 = \sin y$. Find the slope of the curve at the point with coordinates $x = 1/\sqrt[4]{2}$ and $y = \pi/4$.

Exercise 83. As in Example 2.74, drop a circle into a parabola. Suppose the equation of the parabola is $y = cx^2$ for some positive constant c. Find the radius of the largest ball that will touch the bottom in the parabola.

Exercise 84. Repeat Example 2.75 with the curve given by the equation $y^2 - x^2(1 - x^2) = 0$. You find a picture of this Lissajous figure in Figure 5.7.

2.13 Related Rates

Many times you encounter situations in which you have two related variables, you know at which rate one of them changes, and you like to know at which rate the other one changes. In this section we treat such problems.

Example 2.76. Suppose the radius of a ball changes at a rate of 2 cm/min. At which rate does its volume change when $r = 20$?

Solution: Denote the volume of the ball by V and its radius by r. We use t to denote the time variable. We consider V as a function of r as well as t. The formula for the volume of a ball is $V(r) = \frac{4\pi}{3}r^3$. According the the chain rule:

$$\frac{dV}{dt} = \frac{dV}{dr}\frac{dr}{dt} = 4\pi r^2 \frac{dr}{dt}.$$

With $r = 20$ and $\frac{dr}{dt} = 2$ we get $\frac{dV}{dt} = 3200\pi$ cm^3/min. This is the rate at which the volume of the ball changes with respect to time. $\diamond$

Example 2.77. Suppose a particle moves on a circle of radius 10 cm. We think of the circle as being in the Cartesian plane. The center of the circle is at the origin $(0, 0)$. As scale we use 1 cm on both, the horizontal x-axis and the vertical y-axis. At some time the particle is at the point $(5, 5\sqrt{3})$ and moves downwards at a rate of 3 cm/min. At which rate does it move in the horizontal direction?

Solution: The equation of the circle is $x^2 + y^2 = 100$. We consider both variables, x and y, as functions of the time variable t. Implicit differentiation of the equation of the circle gives us the equation

$$2x\frac{dx}{dt} + 2y\frac{dy}{dt} = 0.$$

In the given situation $x = 5$, $y = 5\sqrt{3}$, and $\frac{dy}{dt} = -3$. We find that $\frac{dx}{dt} = 3\sqrt{3}$, so that the particle is moving to the right at a rate of $3\sqrt{3}$ cm/min. $\diamondsuit$

Example 2.78. Two ships, the Independence and Liberty, are on intersecting courses. The Independence travels straight North at a speed of 22 knots (nautical miles per hour), while the Liberty is traveling straight East at a speed of 20 knots. Currently the Independence is 12 nautical miles away from the intersection point of the courses, and the Liberty 15 nautical miles. At which rate does the distance between the ships decrease?

Solution: Draw for yourself a picture of the situation. Use the Cartesian plan as background, and place the intersection point of the courses of the ships at the origin. Use the standard convention that North is in the direction of the positive y-axis and East in the direction of the positive x-axis.

The position of both ships depends on time, which we denote by t and measure in hours. The Liberty travels along the x-axis, and we denote its position $x(t)$. The Independence travels along the y-axis, and we denote its position by $y(t)$. The distance between the ships, as a function of time, is

$$D(t) = \sqrt{x^2(t) + y^2(t)}.$$

As the rate at which $D(t)$ changes, we find

$$\frac{dD}{dt} = \frac{x(t)\frac{dx}{dt} + y(t)\frac{dy}{dt}}{\sqrt{x^2(t) + y^2(t)}}.$$

At the given instant, $x = -15$, $y = -12$, $\frac{dx}{dt} = 20$ and $\frac{dy}{dt} = 22$. We find that, at that instant, that $\frac{dD}{dt} = -29.4$. The ships are approaching each other at a speed of 29.4 knots. $\diamondsuit$

Exercise 85. Consider the situation in Example 2.78. Find the position of the ships and the distance between them 10 minutes later. Calculate the average rate at which the distance between the ships changed during these 10 minutes, and compare it with the rate of change found in the example.

Exercise 86. Two ships, the Independence and Liberty, are on intersecting courses. The Independence travels straight North at a speed of 22 knots, while the Liberty is traveling straight Northeast at a speed of 20 knots. Currently the Independence is 12 nautical miles away from the intersection point of the courses, and the Liberty 15 nautical miles. At which rate does the distance between the ships decrease?

Exercise 87. A ladder, 7 m long, is leaning against a wall. Right now the foot of the latter is 1 m away from the wall. You are pulling the foot of the ladder further away from the wall at a rate of .1 m/sec. At which rate is the top of the ladder sliding down the wall?

Exercise 88. For air at room temperature we suppose that the pressure (P) and volume (V) are related by the equation [24]

$$PV^{1.4} = C.$$

Here C is a constant.

(a) Consider P as a function of V. At which rate does $P(V)$ change with respect to V.

(b) At some instant t_0 the pressure of the gas is 25 kg/cm^2 and the volume is 200 cm^3. Find the rate of change of P if the volume increases at a rate of 10 cm^3/min.

Exercise 89. A conical cup 6 cm across and 10 cm deep is dripping. When the water is 8 cm deep, the water level is dropping at a rate of .5 cm/min. At which rate is the cup losing the water?

Exercise 90. The mass of a particle at velocity v, as perceived by an observer in resting position, is

$$\frac{m}{\sqrt{1 - v^2/c^2}},$$

where m is that mass at rest and c is the speed of light. This formula is from Einstein's special theory of relativity. At which rate is the mass changing when the particle's velocity is 90% of the speed of light, and increasing at .001c per second?

[24]Boyle-Mariotte described the relation between the pressure and volume of a gas. They derived the equation $PV^\gamma = C$. It is called the adiabatic law. The constant γ depends on the molecular structure of the gas and the temperature. For the purpose of this problem, we suppose that $\gamma = 1.4$ for air at room temperature.

2.14 Numerical Methods

In this section we introduce some methods for numerical computations. Their common feature is, that for a differentiable function we do not make a large error when we use the tangent line to the graph instead of the graph itself. This rather casual statement will become clearer when you look at the individual methods.

2.14.1 Approximation by Differentials

Suppose x_0 is an interior point of the domain of a function $f(x)$, and $f(x)$ is differentiable at x_0. Assume also that $f(x_0)$ and $f'(x_0)$ are known. The method of *approximation by differentials* provides an approximate values $f(x_1)$ if x_1 is near x_0. We use the symbol '$\approx$' to stand for 'is approximately'. One uses the formula

$$f(x_1) \approx f(x_0) + f'(x_0)(x_1 - x_0). \tag{2.62}$$

On the right hand side in (2.62) we have $l(x_1)$, the tangent line to the graph of $f(x)$ at $(x_0, f(x_0))$ evaluated at x_1. In the sense of Definition 2.2, $f(x_1)$ is close to $l(x_1)$ for x_1 near x_0.

Example 2.79. Find an approximate value for $\sqrt[3]{9}$.

Solution: We set $f(x) = \sqrt[3]{x}$, so we are supposed to find $f(9)$. Note that

$$f'(x) = \frac{1}{3}x^{-2/3}, \quad f(8) = 2, \text{ and } f'(8) = \frac{1}{12}.$$

Formula (2.62), applied with $x_1 = 9$ and $x_0 = 8$, says that

$$\sqrt[3]{9} = f(9) \approx 2 + \frac{1}{12}(9 - 8) = \frac{25}{12} \approx 2.0833.$$

Your calculator will give you $\sqrt[3]{9} \approx 2.0801$. The method gave us a pretty good answer. $\Diamond$

Example 2.80. Find an approximate value for $\tan 46°$.

Solution: We carry out the calculation in radial measure. Note that $46° = 45° + 1°$, and this corresponds to $\pi/4 + \pi/180$. Use the function $f(x) = \tan x$. Then $f'(x) = \sec^2 x$, $f(\pi/4) = 1$, and $f'(\pi/4) = 2$. Formula (2.62), applied with $x_1 = (\pi/4 + \pi/180)$ and $x_0 = \pi/4$ says

$$\tan 46° = \tan\left(\frac{\pi}{4} + \frac{\pi}{180}\right) \approx \tan\left(\frac{\pi}{4}\right) + \sec^2\left(\frac{\pi}{4}\right)\left(\frac{\pi}{180}\right) = 1 + \frac{\pi}{90} \approx 1.0349.$$

Your calculator will give you $\tan 46° \approx 1.0355$. Again we get a pretty close answer using the method. $\Diamond$

Remark 10. When you apply (2.62), then you may ask what value to take for x_0. A useful choice will be an x_0 which is close to x_1, and for which you have little difficulties finding $f(x_0)$ and $f'(x_0)$.

Exercise 91. Use approximation by differentials to find approximate values for

$$(1)\ \sqrt[5]{34} \quad (2)\ \tan 31° \quad (3)\ \ln 1.2 \quad (4)\ \arctan 1.1.$$

In each case, compare your answer with one found on your calculator.

We have been causal in (2.62) insofar as we have not estimated the error which we make using the right hand side of (2.62) instead of of the actual value of the function on the left hand side. The inequality in Definition 2.2 provides us with an estimate. Differentiability of the function $f(x)$ means that there exist numbers A and $d > 0$, such that

$$|f(x_1) - [f(x_0) + f'(x_0)(x_1 - x_0)| \leq A(x_1 - x_0)^2$$

whenever $|x_1 - x_0| < d$. Thus, if we know A and d, then we can approximate the error as long as $|x_1 - x_0| < d$.

Example 2.81. Find an approximate value for $\sin 31°$ and estimate the error.

Solution: Set $f(x) = \sin x$. The $f'(x) = \cos x$, $f(\pi/6) = 1/2$, and $f'(\pi/6) = \sqrt{3}/2$. Measuring angles in radians we set $x_0 = \pi/6$ and $x_1 = \pi/6 + \pi/180$. Applying the formula in (2.62), we find

$$\sin 31° \approx \sin\frac{\pi}{6} + \frac{\pi}{180}\cos\frac{\pi}{6} = \frac{1}{2}\left(1 + \sqrt{3}\frac{\pi}{180}\right) \approx .515115.$$

The calculator will tell that $\sin 31° \approx .515038$.

From the computation in Example 2.11 on page 50 we also know that we may use $A = 1$ and $d = \pi/4$ in the differentiability estimate. We may apply the estimate because $|x_1 - x_0| < \pi/4$. The estimate assures us that the error is at most

$$(x_1 - x_0)^2 = \left(\frac{\pi}{180}\right)^2 \leq .000305.$$

Comparison of the actual and approximate value confirm this. $\Diamond$

Example 2.82. Use approximation by differentials to find an approximate value of $\sqrt{10}$ and give an upper bound for the error.

Solution: We use $f(x) = \sqrt{x}$ and $x_0 = 9$. The $f'(x) = 1/(2\sqrt{x})$, $f(x_0) = 3$, and $f'(x_0) = 1/6$. The formula in (2.62) tells us that

$$\sqrt{10} = f(10) \approx f(9) + f'(9)(10 - 9) = 3 + \frac{1}{6} \approx 3.16666.$$

The calculator will give you $\sqrt{10} \approx 3.16228$.

For the error estimate we may use

$$A = \frac{1}{2(\sqrt{x_0})^3}$$

and any $d > 0$. This is the A which we picked in (2.18) while proving Proposition 2.15. The estimate assures us that the error is at most

$$\frac{1}{2(\sqrt{x_0})^3}(x_1 - x_0)^2 = \frac{1}{54}.$$

The actual error is again substantially less than this. $\Diamond$

Exercise 92. Use approximation by differentials to find approximate values for

$$(1)\ \cos 28^\circ \quad (2)\ \sqrt{26} \quad (3)\ \sin 47^\circ.$$

In each case, estimate also the maximal error which you may have made by using the method of approximation by differentials.

2.14.2 Newton's Method

We will encounter quite a few situations in which we have to find the zeros of a function. You have learned how to solve a quadratic equation. Assuming that $a \neq 0$, the solutions of the equation

$$ax^2 + bx + c = 0 \quad \text{are} \quad x_{1/2} = \frac{1}{2a}\left[-b \pm \sqrt{b^2 - 4c}\right].$$

There are more complicated formulas which provide algebraic expressions for the solutions of an equation of degree 3 and 4. For polynomial equations of degree 5 and larger there are no general methods which give precise answers. You are in the same predicament if the equation is not a polynomial one. In special cases you may be able to find the root, but typically you cannot.

There is a way out. We try to find approximate numerical values for the solutions. As an example let us try to find solutions of an intentionally complicated equation:

$$3\sin x + \sqrt{7 - x^2 \sin^3(\pi + \cos x)} = 0.$$

Giving a name to the expression, say calling it $f(x)$, allows us to ask instead for the zeros of a function. It is worthwhile to ask your favorite computer program to provide you with a graph of this function. You find a graph in Figure 2.32 for $x \in [-6, 7]$.

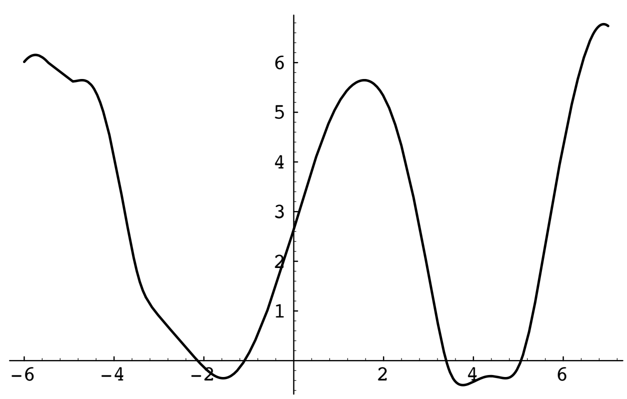

Figure 2.32: A graph

Existence of Zeros: The first question is, whether the function has any zeros. Looking at the graph, your spontaneous answer will be 'yes'. Still, you may want to justify this statement by a better argument than just saying 'it looks like this.' The Intermediate Value Theorem (see Theorem 2.65) provides us with an efficient tool:

- Suppose f is differentiable on $[a, b]$. If $f(a) > 0$ and $f(b) < 0$ (or vice versa), then there exists some $c \in (a, b)$ so that $f(c) = 0$.

Trusting the graph (or you may check this on your calculator), we note that $f(2) > 0$ and $f(4) < 0$. The function f is also differentiable. It is made up (using addition, multiplication, and composition) from functions which are differentiable. The only problem arises when the expression under the radical sign is not positive, but this does not happen for x in the interval under consideration. So f is differentiable on the interval $[2, 4]$. We conclude that f must have a zero between 2 and 4.

Finding Zeros: Suppose now that we found, by some means, a point x_0 which is close to a zero $\overline{x}$ of f, so $f(\overline{x}) = 0$. Newton's method tells us how to find a point x_1 which, under appropriate assumptions, is closer to $\overline{x}$ than x_0. The formula for x_1 is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{2.63}$$

In the hope of improving upon this result, you may iterate the process and calculate

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}, \quad x_3 = x_2 - \frac{f(x_2)}{f'(x_2)}, \quad x_4 = x_3 - \frac{f(x_3)}{f'(x_3)}, \text{ etc.}$$

Let us try this with our example. Let us pick $x_0 = 3$ as a point which is not all that far from the zero in the interval $[2, 4]$. We collect our results is Table 2.3. In the first column we keep track of the subscript n. In the second column you find the values of the corresponding x_n. In the third column we recorded the values of f for the x_n in the second column. The values are rounded off. The numerical value for $f(x)$ in the third column are quickly getting smaller. The value for $f(x)$ in the last row is so small, that it probably exceeds the accuracy with which the calculation has been carried out. So, for all practical purposes we should accept that $f(x)$ is zero for $x = 3.3930802$. We may also say that, without contemplating more about carrying out calculations to a high degree of accuracy, we have come as close to finding a zero of f as we can.

Geometry of Newton's Method: Let us give a geometric explanation for Formula (2.63). Given any x_0 at which f is defined and differentiable, we obtain the tangent line $l(x)$ to the graph of f at this point. Then x_1, as given in Formula (2.63), is the point at which $l(x)$ intersects the x-axis. Specifically,

$$l(x) = f'(x_0)(x - x_0) + f(x_0),$$

so that

$$l(x_1) = 0 \quad \text{if} \quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

n	x_n	$f(x_n)$	&	x_n	$f(x_n)$
0	3.0000000	1.74286522		4.00000000	-0.425838297
1	3.3591917	0.11578456		4.96343583	-0.191889204
2	3.3914284	0.00535801		5.08761552	0.0707101492
3	3.3930755	0.00001505		5.06137162	0.00321831863
4	3.3930802	1.2036×10^{-10}		5.06005732	8.1136276×10^{-6}

Table 2.3: Newton's Method

This means that we accept that the tangent line is close to the graph of the function, and instead of finding the zero of the function itself, we find the zero of the tangent line.

Further Reflections: Our success in the calculation depended critically on the choice of x_0. If we chose $x_0 = 4$, then the sequence of numbers turns out quite differently. As you see in the last two columns of Table 2.3, we seem to be headed for a different zero of the function.

You may also try to start Newton's method with $x = 3.8$. A first application leads you to $x = 8.09433459$, a second one to $x = 9.99399994$. At this point the expression under the radical is negative, so that the function is not even defined.

Exercise 93. Find approximate zeros of the following functions:

$$(1)\ f(x) = x^2 - 2 \quad (2)\ g(x) = x - 2\sin x \quad (3)\ h(x) = 2x - \tan x$$

Make a table as in Table 2.3, and in each example improve your original guess at least twice.

Exercise 94. Find the first positive solution of the equation:

$$x \sin x = \cos x.$$

Hint: Consider the difference of the terms in the equation as a function of x and find zeros. Then proceed as in the previous problem.

More that 4000 years ago, the Babylonians used the following algorithm for approximating radicals. Suppose you like to find $\sqrt{A}$. Pick a number x_0

close to $\sqrt{A}$. Set

$$x_1 = \frac{1}{2}\left(x_0 + \frac{A}{x_0}\right), \text{ and more generally } \quad x_{n+1} = \frac{1}{2}\left(x_n + \frac{A}{x_n}\right),$$

for $n = 0, 1, 2, 3$, etc. With each consecutive x_n you will get a better approximation of $\sqrt{A}$.

E.g., let us find a good approximation of $\sqrt{3}$. As initial guess, we use $x_0 = 2$. We apply the formula from above with $A = 3$. Then

$$x_1 = \frac{1}{2}\left(2 + \frac{3}{2}\right) = \frac{7}{4}, \quad x_2 = \frac{1}{2}\left(\frac{7}{4} + \frac{12}{7}\right) = \frac{97}{56} \quad \text{and} \quad x_3 = \frac{18817}{10864}.$$

We summarize the computation in Table 2.4. In the first column you find the subscript n. In the following two columns you find the values of x_n, once expressed as a fraction of integers, once in decimal form. In the last column you see the square of x_n. At least x_3^2 is rather close to 3. Your calculator will give you 1.73205080757 as an approximate value of $\sqrt{3}$. You see that our value for x_3 is rather precise. In fact, if you carry the calculation one step further and find x_4, then the accuracy of this approximation of $\sqrt{3}$ will exceed the accuracy of most calculators.

n	x_n	x_n	x_n^2
0	2	2.0000000000	4.0000000000
1	7/4	1.7500000000	3.0625000000
2	97/56	1.7321428571	3.0003188775
3	18817/10864	1.7320508100	3.0000000085

Table 2.4: The Babylonian Method

Exercise 95. Use the Babylonian method to find approximate values for $\sqrt{7}$, $\sqrt{35}$, and $\sqrt{19}$. Improve each initial guess at least twice. Summarize your results in a table like the one in 2.4.

Exercise 96. Show that the Babylonian method is the same as Newton's method applied to the function $f(x) = x^2 - A$.

Practical Considerations: We have relied on the graph and numerical calculations to find zeros of a function. Both methods provide results with only a certain degree of accuracy. We may make an effort to improve the accuracy. Still, we can only write down a finite number of decimal places for a real number, and in this sense we do not expect to be able to give precise answers. Our calculators only carry a finite number of significant digits, digits we can be sure about. A good computer program may give a few more significant digits. In any case, this number will decrease if the calculation involves a considerable number of steps, sometimes in ways which are difficult to predict without knowing about the mathematics involved and the technology which is used. From this point of view, we cannot achieve more than what we did above. Within the range of accuracy of the technology we found the zeros of the function as well as we could.

There is one feature of Newton's method which helps. You may say that with each iteration you make a fresh start, and in this sense previous round-off errors don't carry over.

Mathematically speaking, we can analyze under which circumstances Newton's method provides us with arbitrarily precise answers. We can also tell, how precise our answer is, or how many steps are required to achieve a desired accuracy. These are important questions, but they have little bearings on the calculations which we can carry out, unless we invest a lot more work.

2.14.3 Euler's Method

Euler's method is designed to find, by numerical means, an approximate solution of the following kind of problem:

Problem 1. *Find a function $y(t)$ which satisfies*

$$y' = \frac{dy}{dt} = F(t, y) \quad \textit{and} \quad y(t_0) = y_0. \tag{2.64}$$

Here $F(t, y)$ denotes a given function in two variables, and t_0 and y_0 are given numbers.

The first condition on y in (2.64) is a first order differential equation. The second one is called an *initial condition.* It specifies the value of the function at one point. For short, the problem in (2.64) is called an *initial value problem.*

Approach in one step: Suppose you want to find $y(T)$ for some $T \neq t_0$. Then you might try the formula

$$y(T) \approx y(t_0) + y'(t_0)(T - t_0) = y_0 + F(t_0, y_0)(T - t_0). \tag{2.65}$$

The tangent line to the graph of y at (t_0, y_0) is

$$l(t) = y(t_0) + y'(t_0)(t - t_0),$$

so that the middle term in (2.65) is just $l(T)$. The first, approximate equality in (2.65) expresses the philosophy that the graph of a differentiable function is close to its tangent line, at least as long as T is close to t_0. To get the second equality in (2.65) we use the differential equation and initial condition in (2.64), which tell us that

$$y'(t_0) = F(t_0, y(t_0)) = F(t_0, y_0).$$

The Logistic Law

The differential equation in our next example is known as the *logistic law of population growth*. In the equation, t denotes time and $y(t)$ the size of a population, which depends on t. The constants a and b are called the vital coefficients of the population. The equation was first used in population studies by the Dutch mathematician-biologist Verhulst in 1837. The equation refines the Malthusian law for population growth (see (2.26)).

In the differential equation, the term ay expresses that population growth is proportional to the size of the population. In addition, the members of the population meet and compete for food and living space. The probability of this happening is proportional to y^2, so that it is assumed that population growth is reduced by a term which is proportional to y^2.

Example 2.83. Consider the initial value problem:

$$\frac{dy}{dt} = ay - by^2 \quad \text{and} \quad y(t_0) = y_0, \tag{2.66}$$

where a and b are given constants. Find an approximate value for $y(T)$.

Remark 11. An exact solution of the initial value problem in (2.66) is given by the equation

$$y(t) = \frac{ay_0}{by_0 + (a - by_0)e^{-a(t-t_0)}} \tag{2.67}$$

This is not the time to derive this exact solution, though you are invited to verify that it satisfies (2.66). We are providing the exact solution, so that we can see how well our approximate values match it.

Solution: Setting $F(t, y) = ay - by^2$, you see that the differential equation in this example is a special case of the one in (2.64). According to the formula in (2.65) we find

$$y(T) \approx y_0 + (ay_0 - by_0^2)(T - t_0). \tag{2.68}$$

We expect a close approximation only for T close to t_0. $\diamondsuit$

Let us be even more specific and give a numerical example.

Example 2.84. Consider the initial value problem.

$$\frac{dy}{dt} = \frac{1}{10}y - \frac{1}{10000}y^2 \quad \text{and} \quad y(0) = 300. \tag{2.69}$$

Find approximate values for $y(1)$ and $y(10)$.

Solution: Substituting $a = 1/10$, $b = 1/10000$, $t_0 = 0$, and $y_0 = 300$ into the solution in (2.68), we find that

$$y(1) \approx 300 + \left(\frac{300}{10} - \frac{300^2}{10000}\right)(1 - 0) = 321.$$

According to the exact solution in (2.67), we find that

$$y(t) = \frac{3000}{3 + 7e^{-t/10}}.$$

Substituting $t = 1$, we find the exact value $y(1) = 321.4$; this number is rounded off. So, our approximate value is close.

For $T = 10$ the formula suggests that $y(10) \approx 510$. According to the exact solution for this initial value problem, $y(10) = 538.1$. For this larger value of T, the formula in (2.68) gives us a less satisfactory result. $\diamondsuit$

Multi-step approach: We like to find a remedy for the problem which we discovered in Example 2.84 for T further away from t_0. Consider again Problem 1. We want to get an approximate value for $y(T)$. For notational convenience we assume that $T > t_0$. Pick several t_i between t_0 and T:

$$t_0 < t_1 < t_2 < \cdots < t_n = T.$$

Starting out with t_0 and $y(t_0)$, we use the one step method from above to get an approximate value for $y(t_1)$. Then we pretend that $y(t_1)$ is exact, and we repeat the process. We use t_1 and $y(t_1)$ to calculate an approximate value for $y(t_2)$. Again we pretend that $y(t_2)$ is exact and use t_2 and $y(t_2)$ to calculate

$y(t_3)$. Iteratively, we calculate $[t_{i+1}, y(t_{i+1})]$ from $[t_i, y(t_i)]$ according to the formula in (2.65):

$$[t_{i+1}, y(t_{i+1})] = [t_{i+1}, y(t_i) + F(t_i, y(t_i))(t_{i+1} - t_i)] \tag{2.70}$$

We continue this process until we reach T.

For reasonably nice[25] expressions $F(t, y)$ the accuracy of the value which we get for $y(T)$ will increase with n, the number of steps we make (at least if all steps are of the same length). On the other hand, in an actual numerical computation we also make round-off errors in each step, and the more steps we make the worse the result might get. Experience will guide you in the choice of the step length.

Example 2.85. Consider the initial value problem

$$\frac{dy}{dt} = \frac{1}{10}y - \frac{1}{10000}y^2 \quad \text{and} \quad y(0) = 10. \tag{2.71}$$

1. Apply the multi-step method to find approximate values for $y(t)$ at $t = 5$, $t = 10$, $t = 15$, ... , $t = 100$. Arrange them in a table.

2. Graph the points found in the previous step together with the actual solution of the initial value problem. It is given by the equation

$$y(t) = \frac{10000}{10 + 990e^{-t/10}}. \tag{2.72}$$

3. Verify that the function $y(t)$ in (2.72) satisfies the conditions in the initial value problem in (2.71).

Solution: As points in the multi-step process we use

$$t_0 = 0,\ t_1 = 5,\ t_2 = 10,\ t_3 = 15,\ t_4 = 20,\ \ldots,\ t_{20} = 100.$$

For each t_i $(0 \le i \le 19)$ we use the formula

$$y(t_{i+1}) = y(t_i) + 5\left(\frac{y(t_i)}{10} - \frac{y_i^2(t_i)}{10000}\right)$$

and calculate $y(t_1)$, $y(t_2)$, $y(t_3)$, ... , $y(t_{20})$ consecutively. We summarize the calculation in Table 2.5.

[25] We do not want to make this term precise, but the $F(t, y)$ in Example 2.83 is of this kind.

t	$y(t)$	&	t	$y(t)$	&	t	$y(t)$
0	10.00		35	153.96		70	857.73
5	14.95		40	219.09		75	918.74
10	22.31		45	304.62		80	956.07
15	33.22		50	410.55		85	977.07
20	49.28		55	531.55		90	988.27
25	72.70		60	656.05		95	994.07
30	106.41		65	768.87		100	997.02

Table 2.5: Solution of Problem 2.85

In Figure 2.33 you see the graph of the exact solution of the initial value problem. You also see the points from Table 2.5. The points suggest a graph which does follow the actual one reasonably closely. But you see that we are definitely making errors, and they get worse as t increases[26]. You may try a shorter step length. The points will follow the curve much more closely if you use $t_1 = 1$, $t_2 = 2$, $t_3 = 3$, $\ldots$, $t_{100} = 100$ in your calculation.

We leave it to the reader to verify that the function $y(t)$ in (2.72) satisfies the conditions in (2.71). $\Diamond$

Steady States: We are not prepared to study differential equations in great depth. In particular, we are not ready to study qualitative aspects of solutions. Still, there are some note-worthy situations. Consider once more the initial value problem in (2.64):

$$y' = \frac{dy}{dt} = F(t, y) \quad \text{and} \quad y(t_0) = y_0.$$

Suppose $F(y_0, t) = 0$ for all t. Then the constant function $y(t) = y_0$ is a solution of the problem. Such a solution is called a *steady state solution.*

Example 2.86. Find the steady states of the differential equation (see (2.34) in Section 2.8)

$$f'(t) = af(t) + b. \tag{2.73}$$

[26]It is incidental that the points eventually get closer to the graph again. This is due to the specific problem, and will not occur in general.

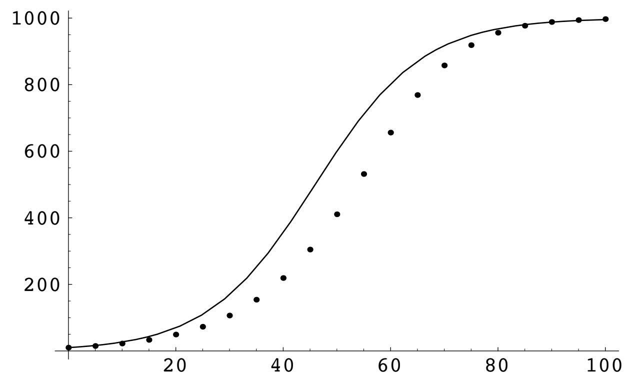

Figure 2.33: Illustration of Euler's Method

Solution: Apparently, $f'(t) = 0$ if and only if $f(t) = -b/a$. So the constant function $f(t) = -b/a$ is the only steady state of this differential equation.

In review of Example 2.33 in Section 2.8, you see that the steady state in that example is $B(t) = 40,000$. I.e., if your loan balance is \$40,000.00, the bank charges you interest at a rate of .5% per month, and you are repaying the loan at a rate of \$ 200.00 per month, then the principal balance of your account will stay unchanged. Your payments cover exactly the occuring interest charges.

For the logistic law (see Equation (2.66))

$$\frac{dy}{dt} = F(y,t) = ay - by^2 = y(a - by)$$

we find that $F(y,t) = 0$ if and only if $y = 0$ or $y = a/b$. There are two steady state solutions: $y_u(t) = 0$ and $y_s(t) = a/b$.

Let us interpret these steady state solutions for the specific numerical values of $a = 1/10$ and $b = 1/10,000$ in Example 2.85. If the initial value y_0 of the population is positive, then the population size will tend to and stabilize[27] at $y(t) = a/b = 1,000$. In this sense, $y_s(t) = a/b = 1,000$ is a stable steady state solution. It is also referred to as the *carrying capacity.* It tells you which size population of the given kind the specific habitat will support.

If the initial value y_0 is negative, then $y(t)$ will tend to $-\infty$ as time increases. If $y_0 \neq 0$, then $y(t)$ will not tend to the steady state $y(t) = 0$. In this sense, $y(t) = 0$ is an unstable steady state. ◊

Exercise 97. Consider the initial value problem

$$y'(t) = -50 + \frac{1}{2}y(t) - \frac{1}{2000}y^2(t) \quad \text{and} \quad y_0 = y(0) = 200. \tag{2.74}$$

To make the problem explicit, you should think of a population of deer in a protected wildlife preserve. There are no predators. The deer are hunted at a rate of 50 animals per year. The population has a growth rate of 50% per year. Reproduction takes place at a constant rate all year round. Finally, the last term in the differential equation accounts for the competition for space and food.

1. Use Euler's method to find the population size over the next 30 years. Proceed in 1 year steps. Tabulate and plot your results.

2. Guess at which level the population stabilizes.

3. Repeat the first two steps of the problem if hunting is stopped.

4. Repeat the first two steps of the problem if the initial population is 100 animals.

5. Find the steady states of the original equation in which hunting takes place. I.e., find for which values of y you have that $y' = 0$? You will find two values. Call the smaller one of them Y_u and the larger one Y_s. Experiment with different initial values to see which of the steady states is stable, and which one is unstable.

Orthogonal Trajectories

Let us explore a different kind of application. Suppose we are given a family $F(x, y, a) = 0$ of curves. In Figure 2.34 you see a family of ellipses

$$C_a : \; F(x, y, a) = x^2 + 3y^2 - a = 0. \tag{2.75}$$

[27] The common language meaning of these expressions suffices for the purpose of our discussion, and the mathematical definition of *'tends to'* and *'stabilizes at'* only make these terms precise.

There is one ellipse for each $a > 0$. We like to find curves D_b which intersect the curves C_a perpendicularly. (We say that D_b and C_a intersect perpendicularly in a point (x_1, y_1), if the tangent lines to the curves at this point intersect perpendicularly.) We call such a curve D_b an *orthogonal trajectory* to the family of the C_a's. You also see one orthogonal trajectory in Figure 2.34.

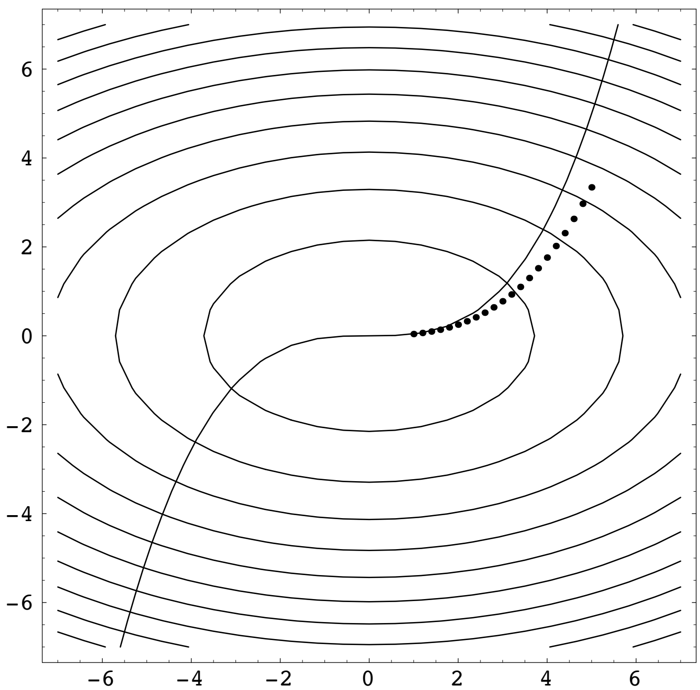

Figure 2.34: Orthogonal Trajectory to Level Curves

Let us explain where this type of situation occurs. Suppose the curves C_a are the level curves in a crater. Here a represents the elevation, so that the elevation is constant along each curve C_a. The orthogonal trajectory gives a path of steepest descent. A new lava flow which originates at some point in the crater will follow this path.

Suppose that each ellipse represents an equipotential line of an electromagnetic field. The orthogonal trajectory provides you with a path which is always in the direction of the most rapid change of the field. A charged particle will move along an orthogonal trajectory.

Suppose a stands for temperature, so that along each ellipse the temperature is constant. In this case the curves are called isothermal lines[28]. A heat seeking bug will, at any time, move in the direction in which the temperature increases most rapidly, i.e., along an orthogonal trajectory to the isothermal lines.

Suppose a stands for the concentration of a nutrient in a solution. It is constant along each curve C_a. On their search for food, bacteria will follow a path in the direction in which the concentration increases most rapidly. They will move along an orthogonal trajectory.

Example 2.87. Find orthogonal trajectories for the family of ellipses

$$C_a: \; F(x, y, a) = x^2 + 3y^2 - a = 0. \tag{2.76}$$

Solution: Differentiating the equation for the ellipses, we get

$$2x + 6y\frac{dy}{dx} = 0 \quad \text{or} \quad \frac{dy}{dx} = \frac{-x}{3y}.$$

The slope of the tangent line to a curve C_a at a point (x_1, y_1) is $\frac{-x_1}{3y_1}$. If a curve D_b intersects C_a in (x_1, y_1) perpendicularly, then we need that the slope of the tangent line to D_b at this point is $\frac{3y_1}{x_1}$. Thus, to find an orthogonal trajectory to the family of the C_a's we need to find functions which satisfy this differential equation. If we also require that the orthogonal trajectory goes through a specific point (x_0, y_0), then we end up with the initial value problem

$$\frac{dy}{dx} = \frac{3y}{x} \quad \text{and} \quad y(x_0) = y_0.$$

This is exactly the kind of problem which we solved with Euler's method. In this particular example it is not difficult to find solutions for the differential equation. They are functions of the form $y(x) = bx^3$. The orthogonal trajectory shown in Figure 2.34 has the equation $y = x^3/25$. There is one orthogonal trajectory which does not have this form, and this is the curve $x = 0$.

Let us apply Euler's method to solve the problem. Let us find approximate values for the initial value problem

$$\frac{dy}{dx} = \frac{3y}{x} \quad \text{and} \quad y(1) = \frac{1}{25}.$$

[28]The idea of isothermal lines, and with this the method in all of these applications, was pioneered by Alexander von Humbold (1769–1859).

Use $x_0 = 1$, $x_1 = 1.2$, $x_2 = 1.4$, ... , $x_{20} = 5$.

We set $(x_0, y_0) = (1, 1/25)$ and calculate (x_n, y_n) according to the formula

$$y_n = y_{n-1} + .2\frac{3y_{n-1}}{x_{n-1}} \quad \text{for } n = 1, 2, \ldots, 20.$$

Without recording the results of this calculation, we graphed the points in Figure 2.34. $\Diamond$

Exercise 98. Consider the family of hyperbolas:

$$C_a: \; x^2 - 5y^2 + a = 0.$$

There is one hyperbola for each value of a, only for $a = 0$ the hyperbola degenerates into two intersection lines.

1. Graph several of the curves C_a.
2. Find the differential equation for an orthogonal trajectory.
3. Use Euler's method to find points on the orthogonal trajectory through the point $(3, 4)$. Use the points $x_0 = 3$, $x_1 = 3.2$, $x_2 = 3.4$, ... , $x_{20} = 7$. Plot the points (x_n, y_n) in your figure.
4. Check that the graph of $y(x) = bx^{-5}$ is an orthogonal trajectory to the family of hyperbolas for every b. Determine b, so that the orthogonal trajectory passes through the point $(3, 4)$, and add this graph to your figure.

2.15 Summary

Let us collect once more all the rules of differentiation and provide a table of some of the important functions which we learned how to differentiate. We assume that f and g are real valued functions.

- Linearity of the derivative (see (2.37)): If f and g are differentiable at x and c is a real number, then

$$(f + g)'(x) = f'(x) + g'(x) \quad \text{and} \quad (cf)'(x) = cf'(x).$$

- Product rule (see (2.41)): If f and g are differentiable at x, then

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x).$$

- Quotient rule (see (2.41)): If f and g are differentiable at x and $g(x) \neq 0$, then

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}.$$

- Chain rule (see (2.51)): If $h(x) = f(g(x))$, g is differentiable at x and f is differentiable at $g(x)$, then

$$h'(x) = f'(g(x))g'(x).$$

- Generalized power rules (see Examples 2.49 and 2.57): If $f(x) = u^q(x)$, u is differentiable at x and either q is an integer or $u(x) > 0$, then

$$f'(x) = qu'(x)u^{q-1}(x).$$

- Derivative of inverse functions (see Theorem 2.69): If g is the inverse of a differentiable function f, and $f'(x) \neq 0$, resp. $f'(g(y)) \neq 0$, then

$$g'(f(x)) = \frac{1}{f'(x)} \quad \text{and} \quad g'(y) = \frac{1}{f'(g(y))}.$$

Most of the derivatives in the following table were calculated in this chapter, and the others can be obtained by the methods in this chapter, usually by an argument which is similar to one used in one of the other examples.

Previously, we have not discussed the functions arccot and arcsec. These are the inverses for the cotangent and secant function. Their domains are specified in the table. You have some freedom in choosing their range. The formula for the derivative holds with the indicated choice. There are sign changes if you alter the choice.

$f(x)$	$f'(x)$	Assumptions
x^q	qx^{q-1}	q a natural number, or $x > 0$
e^x	e^x	$x \in (-\infty, \infty)$
$\ln\lvert x\rvert$	$1/x$	$x \in (-\infty, \infty),\ x \neq 0$
$\sin x$	$\cos x$	$x \in (-\infty, \infty)$
$\cos x$	$-\sin x$	$x \in (-\infty, \infty)$
$\tan x$	$\sec^2 x$	all x for which $\tan x$ is defined
$\cot x$	$-\csc^2 x$	all x for which $\cot x$ is defined
$\sec x$	$\sec x \tan x$	all x for which $\sec x$ is defined
$\csc x$	$-\csc x \cot x$	all x for which $\csc x$ is defined
$\arctan x$	$\frac{1}{1+x^2}$	$x \in (-\infty, \infty)$
$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$	$x \in (-1, 1),\ \arcsin x \in (-\pi/2, \pi/2)$
$\arccos x$	$\frac{-1}{\sqrt{1-x^2}}$	$x \in (-1, 1),\ \arccos x \in (0, \pi)$
$\operatorname{arccot} x$	$\frac{-1}{1+x^2}$	$x \in (-\infty, \infty),\ \operatorname{arccot} x \in (0, \pi)$
$\operatorname{arcsec} x$	$\frac{1}{\lvert x\rvert\sqrt{x^2-1}}$	$x < -1$ or $x > 1,\ \operatorname{arcsec} x \in (0, \pi/2) \cup (\pi/2, \pi)$

Table 2.6: Some Derivatives

Chapter 3

Applications of the Derivative

During a debate on television in October 1984 one of the presidential candidates stated that "*the rate at which the rate of poverty is increasing is decreasing*[1]*.*" Apparently, this is a statement about the poverty rate as a function of time, but what does it really mean? The speaker was using derivatives (or rates of change) to make statements about this function. In fact, he did not only use the (first) derivative, but also the second derivative, the derivative of the derivative.

We will discuss functions on closed intervals. For this reason we extend our definition of differentiability of a function on an interval, so that it allows not only open intervals. Then we state Cauchy's Mean Value Theorem. It has consequences (corollaries) which we will use frequently. Next we will relate the first derivative to monotonicity properties of functions. We will use it to decide whether a function is increasing or decreasing, both on intervals and near a point. Next we define the second and higher derivatives. We relate the second derivative to concavity properties of the function, both on intervals and near a point. The first and second derivative are important tools for graphing functions and for finding its extrema. Finding the extrema of a function, i.e., solving optimization problems, is important in many applications of calculus. We will give some examples.

[1]We will discuss this sentence once we have developed some tools, see Remark 17. According to the transcript of the debate, which was published in the New York Times on October 8th, 1984, page B6, the precise quote is: "*Some of these facts and figures just don't add up. Yes, there has been an increase in poverty but it is a lower rate of increase than it was in the preceding years before we got here. It has begun to decline, but it is still going up.* "

3.1 Differentiability on Closed Intervals

So far we considered the idea of differentiability for functions which are defined on (a union of) open intervals, see Definition 2.8. We now consider this idea for functions which are defined on any kind of interval.

Let I be any interval, open (of the form (a, b)), closed (of the form $[a, b]$), or half open (of the form $[a, b)$ or $(a, b]$). Let J be another interval which contains I, so $I \subseteq J$. Let f be a function which is defined on I and F a function which is defined on J. We say that F *extends* f, or that F is an *extension* of f, if these functions agree on I, i.e., $F(x) = f(x)$ for all $x \in I$.

Definition 3.1. *A function is said to be differentiable on an interval I if it extends to a differentiable function on an open interval.*[2]

Remark 12. One needs to show that the derivative will be unique at all points in I. For this one needs that the interval I in this definition is neither empty nor consists of exactly one point. This will be the case whenever we consider a function on a closed interval.

Let us discuss two examples. Consider the function $f(x) = x^2$ on the interval $[0, 1]$. Is it differentiable on this interval? Yes, as extension we can use the function $F(x) = x^2$, for which we use the domain $(-\infty, \infty)$. We have seen that the function $F(x)$ is differentiable. So $f(x)$ is differentiable. In contrast, the function $g(x) = \sqrt{x}$ is not differentiable on the interval $[0, \infty)$. It is differentiable on all intervals of the form $[a, \infty)$, where $a > 0$. The only sensible candidate for the tangent line to the graph of $g(x)$ at the point $(0, 0)$ on the graph is a vertical line. The slope of this line is not a real number.

3.2 Cauchy's Mean Value Theorem

Let us start out with an

Example 3.2. For the graph pertinent to the example, see Figure 3.1. Consider the function

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

The line S through the points $(.5, f(.5)) = (.5, .25)$ and $(2.5, f(2.5)) = (2.5, 6.25)$ has slope

$$s = \frac{f(2.5) - f(.5)}{2.5 - .5} = \frac{6.25 - .25}{2.5 - .5} = 3.$$

[2] This definition is technically less painful and conceptually more sensible than one which uses one-sided derivatives.

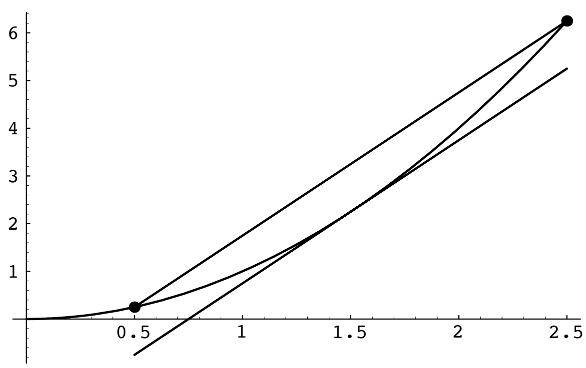

Figure 3.1: Cauchy's Theorem

We call the slope s of the line S the average slope (or average rate of change) of f over the interval $[.5, 2.5]$. Now, remember that

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

If $c = 3/2$, then $f'(c) = 3$. So, the tangent line L to the graph of f at $(3/2, f(3/2))$ has the same slope as the line S. This means, for the number $c = 3/2$, $.5 < c < 2.5$, we have that

$$f'(c) = \frac{f(2.5) - f(.5)}{2.5 - .5}.$$

In other words, there exists a number c between the endpoints of the interval, such that the slope of the graph of f at this point equals the average slope of f over the interval. In geometric terms it means that there exists a point in the interval such that the tangent line at this point is parallel to the line S, the secant line over the interval. $\diamondsuit$

The following theorem is named after Augustin-Louis Cauchy (1789–1857). It expresses the observation which we made in the example. The average slope of a differentiable function over an interval equals the slope of the graph of the function at some point in the interval.

Theorem 3.3 (Chauchy's Mean Value Theorem). *Let f be a real valued function which is defined and differentiable on the interval $[a, b]$, where $a < b$.*[3] *Then there exists a number $c \in (a, b)$ such that*

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

The following special case of the theorem, called Rolle's theorem (named after Michel Rolle (1652–1719)), is of particular interest.

Theorem 3.4 (Rolle's Theorem). *Let f be as in Theorem 3.3. If $f(a) = f(b)$, then there exists a number c between a and b (i.e., $a < c < b$) such that*

$$f'(c) = 0.$$

We are not going to say anything about the proof of these two theorems, except that Cauchy's theorem and Rolle's theorem are equivalent (each is an easy consequence of the other one), and that the proof of both of them depends heavily on the completeness[4] of the real numbers. We are also not interested in finding the points c, as they occur in the two theorems. We are interested in more general consequences.

Corollary 3.5. *Let f be a real valued function which is defined and differentiable on an interval I. If $f'(x) = 0$ for all $x \in I$, then f is constant on this interval. In other words, there exists a number d such that $f(x) = d$ for all $x \in I$.*

Proof. A different formulation of the claim is that $f(a) = f(b)$ for all a, $b \in I$. We prove this statement using Cauchy's theorem. If $f(a) \neq f(b)$, then $a \neq b$ and there exists some $c \in (a, b)$, such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \neq 0.$$

But this contradicts the assumption that $f'(c) = 0$ for all $c \in I$, and the corollary is proved. □

We are going to use the following corollary frequently.

[3]More typically it is assumed that the function is continuous on $[a, b]$ and differentiable on (a, b). Then the same conclusion is true. But, as we have not introduced continuous functions, we content ourselves with this more restrictive version of the theorem.

[4]We discussed this idea in Theorem 5.7

Corollary 3.6. *Let h and g be functions which are defined and differentiable on an interval I. If $h'(x) = g'(x)$ for all $x \in I$, then h and g differ by a constant, i.e., there exists a number d such that*

$$h(x) = g(x) + d$$

for all $x \in I$.

Proof. Apply the previous corollary to $f(x) = h(x) - g(x)$. □

Uniqueness of Solutions of Some Differential Equations

Let us apply the principles which we just discussed to finding all solutions of some differential equations. You are familiar with the fact that, for given numbers b and c, $c \neq 0$, there is exactly one number x such that

$$cx = b. \tag{3.1}$$

Considering x as the unknown, you can also express this by saying that (3.1) has a unique solution. In a differential equation the unknown is a function. We like to see to which extent some differential equations have a unique solution.

Example 3.7. Find all functions $f(x)$ which are defined and differentiable on the entire real line, and for which $f'(x) = 0$ for all x.

Solution: We know that the derivative of a constant function vanishes (is everywhere zero). Furthermore, Corollary 3.5 tells us that constants are the only functions with this property. So we found all functions which have the desired properties. The functions which we were looking for are the constant functions. ◊

Example 3.8. Find all functions $f(x)$ which are defined and differentiable on the entire real line, and whose derivative is

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

Solution: There is one obvious solution for the problem, the function $f(x) = x^2$. Corollary 3.6 says that any other solution of the problem differs from f only by a constant, so that the functions

$$f(x) = x^2 + c$$

are the only functions with the desired property. Here c is an arbitrary constant. ◊

We may formulate the ideas of the last two examples in a more general way.

Example 3.9. Suppose you are given a function $h(x)$ which is defined on an interval I. Find all functions $f(x)$ which are defined on I and for which

$$f'(x) = h(x). \tag{3.2}$$

Solution: Find[5] one function $H(x)$ which is defined on I, and for which $H'(x) = h(x)$. If there is such a function, then any solution of (3.2) is of the form

$$f(x) = H(x) + c,$$

where c is an arbitrary constant. $\diamondsuit$

Exercise 99. Find all function $f(x)$ which satisfy the equation:

$$(1)\ f'(x) = 5x^2 + 7 \quad (2)\ f'(x) = 3\sin 5x \quad (3)\ f'(x) = \sec^2 x.$$

Hint: Guess a function $H(x)$, such that $H'(x) = f'(x)$.

In the following example we verify the second claim which we made in Theorem 2.12. We like to see which functions satisfy the Malthusian Law. This law was the basis for the population and radioactive decay models discussed in Section 2.7.

Example 3.10. Find all functions $f(x)$ which are defined and differentiable on an interval and for which

$$f'(x) = af(x).$$

Solution: We know some functions $f(x)$ which satisfy the differential equation, namely all functions of the form $f(x) = ce^{ax}$ where c is a constant. We want to see once again that these are all of the solutions of the differential equation.

Let $f(x)$ be any function which satisfies the differential equation on some interval. Consider the function

$$h(x) = f(x)e^{-ax}.$$

[5] For the time being you depend on being able to guess such a function $H(x)$. By differentiating $H(x)$ you can check whether you guessed right.

As a product of differentiable functions, h is differentiable and its derivative is

$$h'(x) = f'(x)e^{-ax} - af(x)e^{-ax} = af(x)e^{-ax} - af(x)e^{-ax} = 0.$$

Corollary 3.5 tells us that $h(x)$ is a constant function. Calling the constant c we find that

$$f(x) = ce^{ax}.$$

This means that all solutions of the differential equation $f'(x) = af(x)$ are of the form $f(x) = ce^{ax}$, where c is a constant. With this we have verified the second claim in Theorem 2.12. $\diamondsuit$

Without any further information, the solutions of the differential equations are not unique. In either of the above problems, we get a unique solution if we prescribe the value of the function at one point.

Example 3.11. Find all functions $f(x)$ which are defined and differentiable on the entire real line and for which

$$f'(x) = 2f(x) \quad \text{and} \quad f(0) = 3.$$

Solution: We learned that the only functions which satisfy the differential equation $f'(x) = 2f(x)$ are of the form $f(x) = ce^{2x}$. Substituting $x = 0$ into this expression we see that $f(0) = ce^0 = c$. We conclude that $c = 3$ and that $f(x) = 3e^{2x}$. $\diamondsuit$

Remark 13. The uniqueness of the solution of an initial value problem as in the previous example is not only of theoretical importance. Imagine you study the growth rate of a strain of bacteria. Before you can publish your result, it must be certain that your experiment can be reproduced at a different time in a different location. That is a requirement which any experiment in science must satisfy. If there is more than one mathematical solution to your problem, then you have to expect that the experiment can go either way, and this would invalidate your experiment.

Exercise 100. Find the unique solutions of the problems:

1. $f'(x) = 5f(x)$ and $f(0) = 7$.
2. $f'(x) = 3f(x)$ and $f(2) = 3$.
3. $f'(x) = 2x^2 + 3$ and $f(2) = 1$.

3.3 The First Derivative and Monotonicity

One of the interesting properties of a function is whether it is increasing or decreasing. We might want to find out whether the part of a population which is infected with a disease is increasing or decreasing. We might want to know how the level of pollution in a body of water is changing. The first derivative of a function gives us information of this kind. Let us first recall the definition of the properties *increasing* and *decreasing*. Then we use the first derivative to characterize situations in which a function is monotonic and demonstrate these with some examples.

Monotonicity on Intervals

We called a function f *increasing* (resp. *decreasing*) if

$$f(b) > f(a) \text{ (resp. } f(b) < f(a))$$

whenever f is defined at a and b and $b > a$.

Theorem 3.12. *Suppose that f is a function which is defined and differentiable on an interval I.*

1. *If $f'(x) > 0$ for all $x \in I$, then f is increasing on I.*
2. *If $f'(x) < 0$ for all $x \in I$, then f is decreasing on I.*
3. *More generally, the conclusions in (1) and (2) still hold if in each finite interval $J \subset I$ there are only finitely many points at which the assumption $f'(x) > 0$, resp. $f'(x) < 0$, is not satisfied.*

Proof. We show (1). Let a and b be points in I, and suppose that $a < b$. Cauchy's theorem says that there exists a point c, $a < c < b$, such that

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

We have that $f'(c) > 0$ and $b - a > 0$, and it follows that $f(b) - f(a) > 0$. This means that $f(b) > f(a)$. The proof of the second claim is similar. We leave it and the generalization of both statements to the reader. ☐

Exercise 101. In Figures 3.2 and 3.3 you see the graphs of a function and its derivative. For each pair

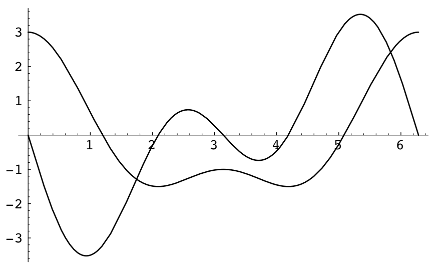

Figure 3.2: Graphs of f and f'.

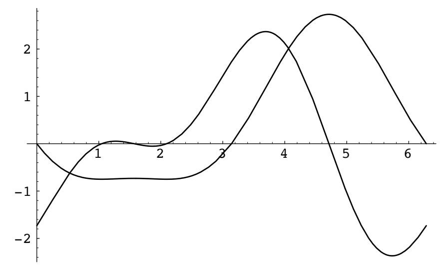

Figure 3.3: Graphs of g and g'.

1. decide which graph belongs belongs to the function and which one to its derivative.

2. determine (approximately) intervals on which the derivative is positive, resp., negative.

3. determine (approximately) intervals on which the function is increasing, resp., decreasing.

Example 3.13. Show that the natural logarithm function is increasing on the interval $(0, \infty)$.

Solution: In Theorem 2.13 on page 52 we stated that $\ln' x = 1/x$. So $\ln' x > 0$ for all $x \in (0, \infty)$. It follows from Theorem 3.12 that $\ln x$ is increasing on $x \in (0, \infty)$. You may check this result by having a look at the graph of the natural logarithm functions in Figure 1.13. $\Diamond$

Example 3.14. Show that the exponential function (for base e) is increasing on the entire real line.

Solution 1: The inverse of an increasing function is increasing (see Proposition 5.25 on page 291) and the exponential function is the inverse of the logarithm function. It follows from the previous example that the exponential function is increasing on the entire real line.

Solution 2: The exponential function is positive everywhere, see Theorem 1.12, and so is its derivative $de^x/dx = e^x$. Once again, Theorem 3.12

tells us that the function is increasing.[6]

Finally, you should confirm the result by having a look at the graph of the exponential function in Figures 1.12. $\diamond$

Example 3.15. Discuss the monotonicity properties of the function

$$f(x) = 1/x.$$

Solution: This function is defined and differentiable on the set of all nonzero real numbers. The derivative of the function is

$$f'(x) = -1/x^2,$$

and $f'(x) < 0$ for all nonzero real numbers. According to Theorem 3.12, this means that $f(x)$ is decreasing on the interval $(-\infty, 0)$, and that $f(x)$ is decreasing on the interval $(0, \infty)$. The function is not decreasing on the union of the two intervals.[7] Be sure to graph the function to confirm this finding. $\diamond$

Example 3.16. Show that the arctangent function

$$f(x) = \arctan x$$

is increasing on the entire real line.

Solution: We discussed this function in Example 2.71 on page 106. It is the inverse of the tangent function, and it is defined and differentiable on the entire real line. We found that

$$\arctan' x = \frac{1}{1+x^2}.$$

Apparently, $\arctan' x > 0$ for all real numbers, and this means that the arctangent function is increasing on $(-\infty, \infty)$. $\diamond$

Exercise 102. Discuss the monotonicity properties of the following functions:

$$(1)\ f(x) = \sqrt{x} \quad (2)\ g(x) = \frac{1}{x^2} \quad (3)\ h(x) = \frac{1}{x^3} \quad (4)\ k(x) = \operatorname{arccot} x.$$

[6] This solution stands on shaky grounds. You may say that we asserted the monotonicity of the exponential function in Theorem 1.12, so that there is nothing left to be shown.

[7] The example illustrates that it is crucial in Theorem 3.12 that we deal with functions which are defined and differentiable on an interval.

Example 3.17. You may be aware of the fact that warm blooded animals that live in cold climates are larger that their relatives of the same species that live in warm climates. Similarly, cold blooded animals that live in cold climates are smaller that their relatives of the the same species that live in a warm climate. This has been explained based on a simple mathematical observation and the theory of Darwin.

Let us first discuss the relevant mathematics. Consider a cube with side length a. Its surface area is $A(a) = 6a^2$ and its volume is $V(a) = a^3$. Let us define a function

$$E(a) = \frac{A(a)}{V(a)} = \frac{6}{a}.$$

So $E(a)$ gives the ratio between the surface area and the volume.

Similarly, as you may remember or look up in a collection of formulas, the surface area of a ball of diameter d is $A(d) = \pi d^2$, and its volume is $V(d) = \pi d^2/6$. Also for this shape we find that the ratio of A and V is

$$E(d) = \frac{A(d)}{V(d)} = \frac{6}{d}.$$

Consider any geometric shape, and suppose that you vary its size uniformly in all directions. If d denotes the length in any direction, you will find again that

$$E(d) = \frac{A(d)}{V(d)} = \frac{6}{d}.$$

It takes some work to justify this formula, but it can be done based on the example of the cube. The important fact is that

$$E'(d) = -\frac{6}{d^2} < 0 \tag{3.3}$$

for all $d > 0$. In plain English this means, as the size of an animal increases the ratio of surface area to volume decreases.

Now let us look at animals and the climate in which they live. A warm blooded animal needs energy to maintain its body temperature, particularly in cold climates. It loses heat through its surface, and the heat loss is proportional to the surface area and the temperature difference. Thus it is of advantage if, in relation to its volume, the surface area is small. This ratio improves (decreases) as the size of the animal gets bigger. Natural selection (Darwinism) should favor the larger specimens of a warm blooded species in a cold climate.

In a hot climate, the energy created by an active animal may raise the body temperature so that it exceeds its regular temperature. So the animal needs to cool down by giving off heat to its surrounding. It is of advantage if the surface area is large, in comparison to the volume. This ratio improves as the size of the animal gets smaller. So natural selection should favor smaller specimens of a warm blooded species in a warm climate.

For cold blooded animals the situation is just the other way around. Cold blooded animals have to absorb heat through their surface to reach or maintain the temperature at which they can operate (move about and find food). They have to heat up their entire body (volume) by absorbing heat through their surface. In particular, in cold climates it is important that the surface area is large, in comparison to the volume. This ratio improves as the animal gets smaller, and in this sense natural selection should favor smaller specimens of cold blooded animals in a cold climate.

Needless to say, there are other mechanisms to increase the surface area of a body than decreasing its size, and the maintenance of the body temperature is only one factor which influences the size of specimens of a species. There are many more. Larger animals need more food, are stronger but cannot hide as well, and are often less agile. All of these factors need to be taken into account to determine the optimal size of an animal. $\diamondsuit$

So far we have only discussed examples where we used (1) and (2) of Theorem 3.12. Let us show how to use the conclusion in (3). To apply it we need to determine intervals on which a function does not change signs. We recall a procedure which works well for the functions treated in these notes.

Definition 3.18. *Suppose $f(x)$ is a function. We call a point x_0 on the real line* exceptional *if either $f(x_0) = 0$ or $f(x_0)$ is not defined.*

The following result is an immediate consequence of the Intermediate Value Theorem, see Theorem 2.65 on page 103. Expressed casually it says that a differentiable function does not change signs between exceptional points.

Proposition 3.19. *Suppose $f(x)$ is a differentiable function and $f(x)$ has no exceptional points in the interval (x_0, x_1). Then $f(x) > 0$ for all points in the interval (x_0, x_1), or $f(x) < 0$ for all points in (x_0, x_1). In particular, if $f(x) > 0$ (resp., $f(x) < 0$) for one point $x \in (x_0, x_1)$, then $f(x) > 0$ (resp., $f(x) < 0$) for all points $x \in (x_0, x_1)$.*

Example 3.20. Find the intervals on which the function

$$f(x) = \frac{x^2(x^2 - 4)}{x^2 + 2x - 15}$$

is positive, resp., negative.

Solution: First, let us determine the points where the function is zero. These are the points where the numerator vanishes. The numerator of the expression for the function factors as $x^2(x-2)(x+2)$, and this expression is zero if and only if one of its factors is zero. This provides us with exceptional points $x = 0$, $x = 2$, and $x = -2$.

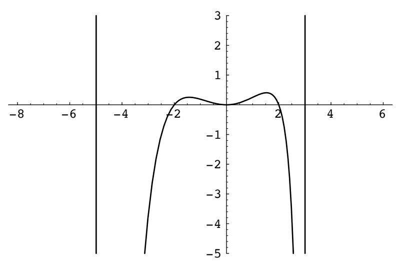

Figure 3.4: Exceptional Points

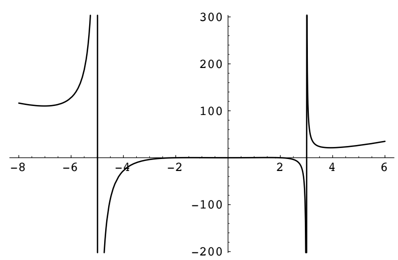

Figure 3.5: Exceptional Points

Next, let us determine the points where the function is not defined. The expression for $f(x)$ is undefined wherever the denominator is zero. It factors as $(x+5)(x-3)$. So we find two more exceptional points, $x = 3$ and $x = -5$.

The proposition tells us that on the intervals in between these exceptional points the function does not change signs. The intervals are $(-\infty, -5)$, $(-5, -2)$, $(-2, 0)$, $(0, 2)$, $(2, 3)$ and $(3, \infty)$.

Counting signs of the factors in the expression for $f(x)$, we see $f(x)$ is positive on the interval $(-\infty, -5)$, negative on $(-5, -2)$, positive on $(-2, 0)$ and on $(0, 2)$, negative on $(2, 3)$, and positive on $(3, \infty)$. You see that the sign changes at some, but not all, exceptional numbers. You see a graph of the function in Figures 3.4 and 3.5. We had to use two different y-scales to be able to display different aspects of the graph. $\diamondsuit$

Exercise 103. Find intervals on which the following functions do not change signs. Decide whether the functions are positive or negative on these intervals.

$$(1)\ f(x) = x^3 - x^2 - 5x - 3 \quad (2)\ g(x) = \frac{x}{x^3 + 5x^2 - 4x - 20}.$$

We are ready to discuss the monotonicity of functions whose derivative vanishes at some points.

Example 3.21. Find intervals on which the function

$$f(x) = 3x^2 + 5x - 4$$

is monotonic.

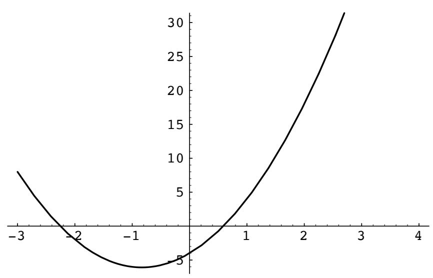

Figure 3.6: A quadratic polynomial, $f(x) = 3x^2 + 5x - 4$

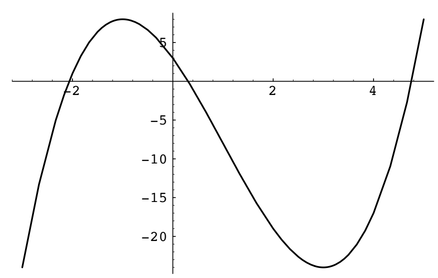

Figure 3.7: A cubic polynomial, $p(x) = x^3 - 3x^2 - 9x + 3$

Solution: We graphed the function in Figure 3.6. It is defined and differentiable on the real line. Its derivative is

$$f'(x) = 6x + 5.$$

In particular, $f'(x) > 0$ if $x > -5/6$, i.e., if $x \in (-5/6, \infty)$. So $f'(x) > 0$ for all points $x \in [-5/6, \infty)$, except at the single point $x = -5/6$. Theorem 3.12 (3) says that f is increasing on the interval $[-5/6, \infty)$. By a similar argument, f is decreasing on the interval $(-\infty, -5/6]$. $\Diamond$

Example 3.22. Find intervals on which the degree three polynomial (for a graph see Figure 3.7)

$$p(x) = x^3 - 3x^2 - 9x + 3$$

is monotonic.

Solution: The function is defined and differentiable on the real line. Its derivative is

$$p'(x) = 3x^2 - 6x - 9 = 3(x^2 - 2x - 3) = 3(x-3)(x+1).$$

We factored $p'(x)$ so that it is easy to decide where it is positive or negative. The product is positive if $(x-3)$ and $(x+1)$ are both positive $(x > 3)$ or if both are negative $(x < -1)$. We conclude that $p(x)$ is increasing on the interval $[3, \infty)$ and that it is increasing on the interval $(-\infty, -1]$. The derivative is negative on the interval $(-1, 3)$ because then $(x-3)$ is negative and $(x+1)$ is positive. The theorem implies that $p(x)$ is decreasing on the interval $[-1, 3]$. $\diamondsuit$

Example 3.23. Find intervals on which the rational function

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

is monotonic.

Solution: The simplified expression for the derivative of f is

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

The important aspect of simplifying the expression for the derivative in this form is, that numerator and denominator are expressed as products of terms, and for each of them it is apparent where it is zero. We see that the exceptional points for $f'(x)$ are $x = 1$, $x = -1$ and $x = 3$. We conclude that $f'(x)$ does not change signs on the intervals $(-\infty, -1)$, $(-1, 1)$, $(1, 3)$, and $(3, \infty)$. Counting the signs of the factors of $f'(x)$, we conclude that $f'(x) > 0$ on the intervals $(-\infty, -1)$ and $(3, \infty)$, and $f'(x) < 0$ on the intervals $(-1, 1)$ and $(1, 3)$. Observe that $f(x)$ is defined and differentiable on the entire real line with the only exception of $x = 1$. We conclude that $f(x)$ is increasing on the $(-\infty, -1]$ and $[3, \infty)$. The function is decreasing on the intervals $[-1, 1)$ and $(1, 3]$. $\diamondsuit$

Example 3.24. Find intervals on which the function

$$f(x) = \sin 2x + 2 \sin x$$

is monotonic. Restrict your discussion to the interval $[0, 2\pi]$.

Solution: We differentiate the function and rewrite the expression for the derivative so that it is easier to find its exceptional points.

$$\begin{aligned} f'(x) &= 2\cos 2x + 2\cos x \\ &= 2[2\cos^2 x + \cos x - 1] \\ &= 4(\cos x + 1)\left(\cos x - \frac{1}{2}\right). \end{aligned}$$

To see the second equality we used that $\cos 2x = \cos^2 x - \sin^2 x$ and $\sin^2 x = 1 - \cos^2 x$. To find the third equality, we solved a quadratic equation in $\cos x$. We find exceptional points where $\cos x = -1$ (i.e., $x = \pi$) and where $\cos x = \frac{1}{2}$ (i.e., $x = \frac{\pi}{3}$ and $x = \frac{5\pi}{3}$).

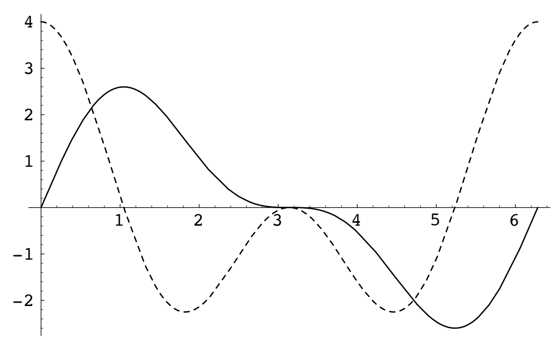

Figure 3.8: A function and its derivative.

Observe that f is differentiable on $[0, 2\pi]$, and that $f'(x) \neq 0$ at the end points of this interval. This provides us with the intervals $[0, \pi/3)$, $(\pi/3, \pi)$, $(\pi, 5\pi/3)$ and $(5\pi/3, 2\pi]$ on which f' does not change sign. Checking the sign of f' (at one point) in each of the intervals, we find that $f'(x) > 0$ for $x \in [0, \pi/3)$ and $x \in (5\pi/3, 2\pi]$, and $f'(x) < 0$ for $x \in (\pi/3, \pi)$ and $(\pi, 5\pi/3)$. We conclude that f is increasing on the interval $[0, \pi/3]$ and $[5\pi/3, 2\pi]$. The function is decreasing on the interval $[\pi/3, 5\pi/3]$, and in this interval there are three points at which $f'(x)$ is not positive.

You may confirm the calculation by having a look at Figure 3.8. There you see the graph of the function (solid line) and the graph of its derivative (dashed line). As you see, wherever $f'(x)$ is positive, there $f(x)$ is increasing. Wherever $f'(x)$ is negative, there $f(x)$ is decreasing. $\diamondsuit$

Exercise 104. Confirm the computations in Example 3.23 by graphing the function and its derivative in the same set of coordinates. Label the graphs of the functions, and indicate the intervals on which the function is increasing, resp. decreasing, and on which the derivative is positive, resp. negative.

Exercise 105. Find intervals on which the function f increases and intervals on which f decreases. In the last two problems, (g) and (h), restrict yourself to the interval $[0, 2\pi]$.

(a) $f(x) = 3x^2 + 5x + 7$	(e) $f(x) = x^3(1 + x)$
(b) $f(x) = x^3 - 3x^2 + 6$	(f) $f(x) = x/(1 + x^2)$
(c) $f(x) = (x + 3)/(x - 7)$	(g) $f(x) = \cos 2x + 2\cos x$
(d) $f(x) = x + 1/x$	(h) $f(x) = \sin^2 x - \sqrt{3}\sin x$

Monotonicity at a Point

It is quite natural to ask what it means that a function is increasing at a point, and how this concept is related to the one of being increasing on an interval. We address both questions in this subsection.

Let us say that a function is increasing at a point c if $f(x) < f(c)$ for all x in some interval to the left of c and $f(x) > f(c)$ for all x in some interval to the right of c. Expressed more formally

Definition 3.25. *Suppose f is a function and c is an interior point of its domain. We say that f is* increasing *at c if, for some $d > 0$,*

$$f(x) < f(c) \textit{ for all } x \in (c - d, c) \textit{ and } f(x) > f(c) \textit{ for all } x \in (c, c + d).$$

We say that f is decreasing *at c if this statement holds with the inequalities reversed.*

Being increasing or decreasing at a point c is a *local* property. We are making a statement about the behavior of the function on some open interval which contains c. Being increasing on an interval is a global property. For the global property the interval is given to us. For the local property we may chose the, possibly rather small, interval. The global property has to hold for any two points in the given interval. For the local property we compare $f(x)$ to $f(c)$ where c is fixed and x is any point in an open interval around c which we may chose.

Theorem 3.26. *Suppose f is a function which is defined on an open interval I. Then f is increasing (decreasing) on I if and only it it is increasing (decreasing) at each point in I.*

This theorem establishes the relation between the local and the global property. The 'only if' part is not difficult to show, but the 'if' part uses some deeper facts about finite closed intervals. Our second result gives us a valuable tool to detect monotonicity of functions at a point.

Proposition 3.27. *Let f be a function and c an interior point of its domain. If f is differentiable at c and $f'(c) > 0$, then f is increasing at c. If $f'(c) < 0$, then f is decreasing at c.*

Remark 14. A function does not have to be differentiable to be increasing. Graph the function $f(x) = 2x + |x|$ to convince yourself of this fact. A function can be differentiable and increasing at a point x, even if the assumptions of Proposition 3.27 do not hold, i.e., $f(x) = x^3$ is increasing at $x = 0$, but if $f'(0) = 0$. A function can also be increasing at a point x, but there is not open interval which contains x such that the function is increasing on this interval.

Remark 15. The ideas of of a function being increasing or decreasing at a point may be generalized to cover domains of functions which are half-closed or closed intervals, and where we like to make a statement about the behavior of a function at an endpoint. We have no specific needs for such statements, but the motivated reader is encouraged to explore them.

3.4 The Second and Higher Derivatives

Let $f(x)$ be a function which is defined on an open interval, or a union of open intervals. If the function is differentiable at each point of its domain, then $f'(x)$ is again a function with the same domain as $f(x)$. We may ask whether the function $f'(x)$ is differentiable. Its derivative, wherever it exists, is called the second derivative of f. It is denoted by $f''(x)$. This process can be iterated. The derivative of the second derivative is called the third derivative, and denoted by $f'''(x)$, etc. We will make use of the second derivative. Leibnitz's notation for the second derivative of a function $f(x)$

is d^2f/dx^2. Here is a sample computation in which you are invited to fill in the details:

$$\frac{d^2}{dx^2}e^{\sin x} = \frac{d}{dx}\cos x e^{\sin x} = (-\sin x + \cos^2 x)e^{\sin x}.$$

$f(x)$	$f'(x)$	$f''(x)$	$f'''(x)$
x^q	qx^{q-1}	$q(q-1)x^{q-2}$	$q(q-1)(q-2)x^{q-3}$
e^x	e^x	e^x	e^x
$\ln\|x\|$	$1/x$	$-1/x^2$	$2/x^3$
$\sin x$	$\cos x$	$-\sin x$	$-\cos x$
$\cos x$	$-\sin x$	$-\cos x$	$\sin x$
$\tan x$	$\sec^2 x$	$2\sec^2 x \tan x$	
$\cot x$	$-\csc^2 x$	$2\csc^2 x \cot x$	
$\sec x$	$\sec x \tan x$	$2\sec^3 x - \sec x$	
$\csc x$	$-\csc x \cot x$	$2\csc^3 x - \csc x$	
$\sinh x$	$\cosh x$	$\sinh x$	$\cosh x$
$\cosh x$	$\sinh x$	$\cosh x$	$\sinh x$
$\arctan x$	$\frac{1}{1+x^2}$	$\frac{-2x}{(1+x^2)^2}$	$\frac{6x^2-2}{(1+x^2)^3}$
$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{x}{(1-x^2)^{3/2}}$	$\frac{2x^2-1}{(1-x^2)^{5/2}}$
$\arccos x$	$\frac{-1}{\sqrt{1-x^2}}$	$\frac{-x}{(1-x^2)^{3/2}}$	$-\frac{2x^2-1}{(1-x^2)^{5/2}}$
$\operatorname{arccot} x$	$\frac{-1}{1+x^2}$	$\frac{2x}{(1+x^2)^2}$	$-\frac{6x^2-2}{(1+x^2)^3}$

Table 3.1: Some higher derivatives. We need assumptions as in Table 2.6.

We collect some examples in Table 3.1. There is nothing new to calculating higher derivatives. You just repeat what you learned before. In some calculations a few simplifications based on elementary arithmetic and trigonometric identities (as you can find them in Section 5.5 on page 276) have been employed. We don't enter all derivatives in the table. Some expressions are so large that the table will not fit on the page if we do.

In the table you see two functions which we have not introduced before. These are the hyperbolic sine and cosine functions. Their definitions are

$$\sinh x = \frac{1}{2}\left[e^x - e^{-x}\right] \quad \& \quad \cosh x = \frac{1}{2}\left[e^x + e^{-x}\right] \tag{3.4}$$

Exercise 106. Verify the formulas for the derivatives of the hyperbolic functions in Table 3.1.

Exercise 107. Verify the identity

$$\cosh^2 x - \sinh^2 x = 1.$$

The result in the previous exercise motivates the attribute 'hyperbolic'. A point (u, v) on the hyperbola

$$u^2 - v^2 = 1$$

can be expressed as $(\pm \cosh x, \sinh x)$ for some $x \in (-\infty, \infty)$.

Exercise 108. Find the second derivatives of the following functions:

(1) $f(x) = 3x^3 + 5x^2$	(6) $k(x) = \cos(x^2)$	(11) $p(x) = \ln^2(x+4)$
(2) $g(x) = \sin 5x$	(7) $l(x) = \ln 2x$	(12) $q(x) = e^{\cos x}$
(3) $h(x) = \sqrt{x^2+2}$	(8) $m(x) = \ln(x^2+3)$	(13) $r(x) = \ln(\tan x)$
(4) $i(x) = e^{5x}$	(9) $n(x) = \arctan 3x$	(14) $s(x) = e^{x^2-1}$
(5) $j(x) = \tan x$	(10) $o(x) = \sec(x^3)$	(15) $t(x) = \sin^3 x$.

3.5 The Second Derivative and Concavity

Let us start out with two examples. In Figure 3.9 you see the graph of the function

$$q(x) = x^2 - 2x + 3.$$

Consider two points on the graph, say $(-1, 6)$ and $(1, 2)$, and connect them by a line segment. As you see, the line segment lies above the graph. The same is true, if we take any two points on the graph. This property of the function will be called being concave up.

In contrast, if you consider the the graph of the function (see Figure 3.10)

$$g(x) = -x^2 + 5x - 1$$

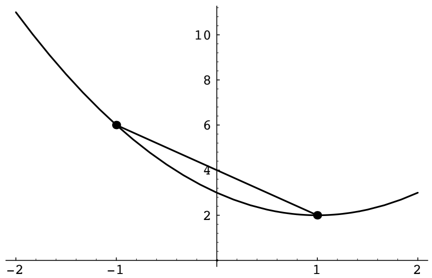

Figure 3.9: $q(x) = x^2 - 2x + 3$

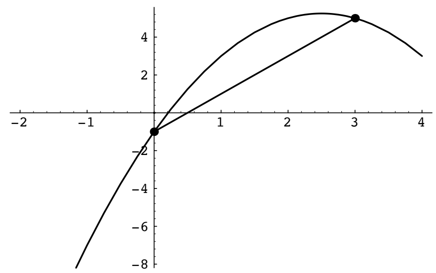

Figure 3.10: $g(x) = -x^2 + 5x - 1$

and take any two points on its graph, say $(0, -1)$ and $(3, 5)$, then the line segment which connects the two points lies below the graph. This property will be called being concave down.

We may use the monotonicity of the first derivative or information about the second derivative of a function to find criteria which tell us that a function is concave up or down on an interval. We also study the corresponding notion at a point.

Concavity on an Interval

First of all, let us define the concept of being concave up or down on an interval. Let $(a, f(a))$ and $(b, f(b))$ be two distinct points on the graph of a function f. The two point formula for a line provides us with an expression for the line through these two point. For any $x \in (-\infty, \infty)$

$$l(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a).$$

The following definition expresses in mathematical notation that a function is concave up (down) if every line segment connecting two points on its graph lies above (below) the graph.

Definition 3.28. *Let f be a function which is defined on an interval I. We say that f is* concave up *on I if*

$$f(c) < l(c)$$

for all a, b, $c \in I$ with $a < c < b$. Here $l(x)$ is the line through $(a, f(a))$ and $(b, f(b))$. We say that f is concave down *on I if*

$$f(c) > l(c)$$

for all a, b, $c \in I$ with $a < c < b$.

In Figures 3.9 and 3.10 you saw the graph of a function which is concave up and of a function which is concave down. We state a theorem which provides you with assumptions under which a function is concave up or down. We will not provide a proof of the theorem.

Theorem 3.29. *Let f be a function which is defined on an interval I.*

1. *Suppose that $f(x)$ is differentiable on I. If $f'(x)$ is increasing on I, then $f(x)$ is concave up on I. If $f'(x)$ is decreasing on I, then $f(x)$ is concave down on I.*

2. *Suppose that $f(x)$ is twice differentiable[8] on I. If $f''(x) > 0$ for all x in I, then $f(x)$ is concave up on I. If $f''(x) < 0$ for all x in I, then $f(x)$ is concave down on I.*

3. *More generally, the conclusions in (2) still hold if in each finite interval $J \subset I$ there are only finitely many points at which the assumption $f''(x) > 0$, resp. $f''(x) < 0$, is not satisfied.*

Let us apply this theorem in a few examples.

Example 3.30. Confirm the statements which we made in the prolog to this section.

Solution: The second derivative of the function (see Figure 3.9)

$$q(x) = x^2 - 2x + 3 \quad \text{is} \quad q''(x) = 2,$$

and this function is positive on the entire real line. Theorem 3.29 (2) says that q is concave up on $(-\infty, \infty)$.

In comparison, the second derivative of the function

$$g(x) = -x^2 + 5x - 1 \quad \text{is} \quad g''(x) = -2,$$

and this function is negative on the entire real line. The theorem says that g is concave down on $(-\infty, \infty)$. $\diamondsuit$

[8]Strictly speaking, so far we can consider being 'twice differentiable' only for functions which are defined on open intervals. More generally, we proceed as in Section 3.1. We say that $f(x)$ is twice differentiable on I, if $f(x)$ extends to a function $F(x)$ which is defined on an open interval J which contains I, and $F(x)$ is twice differentiable on J. The second derivative will be unique at all points in I if I does not consist of exactly one point.

Example 3.31. Show that the function

$$h(x) = \ln x$$

is concave down on the interval $(0, \infty)$.

Solution: The first derivative of $h(x)$ is $h'(x) = 1/x$, and its second derivative is (see Table 2.6)

$$h''(x) = -1/x^2.$$

Apparently $h''(x) < 0$ for all positive numbers x, so that we may conclude from Theorem 3.29 (2) that $\ln x$ is concave down on its domain, $(0, \infty)$.

Another Solution: Our calculation shows that $h'(x)$ is decreasing on the interval $(0, \infty)$, because the first derivative $h''(x)$ of $h'(x)$ is negative on this interval. Theorem 3.29 (1) implies that $h(x) = \ln x$ is concave down on its domain, $(0, \infty)$. $\Diamond$

Example 3.32. Study the concavity properties of the exponential function

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

Solution: We asserted the existence and differentiability of the exponential function in Theorems 1.12 on page 20 and 2.12 on page 52. Theorem 2.12 also says that $f(x) = f'(x)$. Applying the theorem twice we conclude that

$$f''(x) = e^x.$$

By definition, $f(x) = f''(x) > 0$ for all real numbers x, see Theorem 1.12. Theorem 3.29 (2) implies that the exponential function is concave up on its entire domain, the interval $(-\infty, \infty)$. $\Diamond$

Remark 16. Let f be a function which is defined on an interval I. Suppose that its image is an interval J, and that f has an inverse, which we call g. If f is concave up, then g is concave down, and vice versa. This follows from a geometric argument. You should convince yourself that if a secant line is above the graph, and you reflect the picture at the diagonal (that is how to get the graph of the inverse function) then, in the reflected picture, the secant line will be below the graph. We could have used this argument to exploit the statement that the logarithm function is concave down to deduce that the exponential function is concave up.

Exercise 109. In each of the Figures 3.11 and 3.12 you see the graphs of a function f and its second derivative f''. For each pair

1. decide which graph belongs belongs to the function and which one to its second derivative.
2. determine (approximately) intervals on which the second derivative is positive, resp., negative.
3. determine (approximately) intervals on which the function is concave up, resp., concave down.

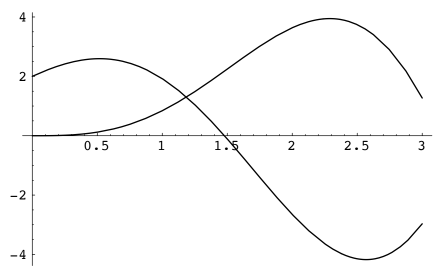

Figure 3.11: Graphs of f and f''.

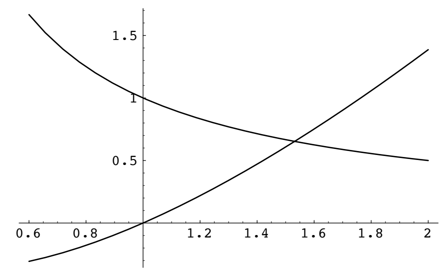

Figure 3.12: Graphs of f and f''.

Exercise 110. Discuss the concavity properties of the functions

$$(1)\ f(x) = x^2 - 5x + 8 \quad \text{and (2) } g(x) = \sqrt{3x - 1}.$$

So far we applied Theorem 3.29 (2) to obtain conclusions. Let us look at some examples where we apply condition (3).

Example 3.33. Study the concavity properties of the function

$$p(x) = x^3 - 3x^2 - 9x + 3.$$

Solution: You find the graph of this function in Figure 3.7, and we discussed its monotonicity properties in Example 3.22 on page 150. An easy calculation provides us with the second derivative of this function:

$$p''(x) = 6x - 6 = 6(x - 1).$$

We see that $p''(x) > 0$ for $x \in (1, \infty)$, and $p''(x) < 0$ for $x \in (-\infty, 1)$. This means that $p''(x) > 0$ for all $x \in [1, \infty)$ with only one exception, $x = 1$. Theorem 3.29 (3) tells us that $p(x)$ is concave up on the interval $[1, \infty)$. Similarly, $p''(x) < 0$ for $x \in [-\infty, 1)$ with only one exception, $x = 1$. One deduces that $f(x)$ is concave down on the interval $(-\infty, 1]$. $\diamondsuit$

Example 3.34. Study the monotonicity and concavity properties of the tangent function $\tan x$ on the interval $(-\pi/2, \pi/2)$.

Solution: You find the first and second derivative of the function $\tan x$ in Table 2.6 on page 136:

$$\tan' x = \sec^2 x \quad \& \quad \tan'' x = 2 \sec^2 x \tan x.$$

By definition, $\sec^2 x > 0$ on the interval $(-\pi/2, \pi/2)$. This means that $\tan x$ is increasing on $(-\pi/2, \pi/2)$, see Theorem 3.12 on page 144.

In addition, $\tan x < 0$ for $x \in (-\pi/2, 0)$ and $\tan x > 0$ for $x \in (0, \pi/2)$. This means that $\tan'' x < 0$ for $x \in (-\pi/2, 0)$ and $\tan'' x > 0$ for $x \in (0, \pi/2)$. Theorem 3.29 (3) implies that $\tan x$ is concave down on $(-\pi/2, 0]$ and concave up on $[0, \pi/2)$. Compare Figures 5.11 and 5.12 to confirm our conclusions visually. $\diamondsuit$

Example 3.35. Find intervals on which the function $f(x) = \sin x$ is concave up or concave down.

Solution: The sine function is defined and twice differentiable on the interval $(-\infty, \infty)$. Its second derivative is (see Table 3.1)

$$f''(x) = -\sin x.$$

You may use the graph shown in Figure 5.9, or the geometry of the unit circle, to conclude that $\sin x > 0$ on intervals of the form $(2n\pi, (2n+1)\pi)$ and $\sin x < 0$ on intervals of the form $((2n+1)\pi, 2n\pi)$. Here n is an arbitrary integer (whole number). We conclude that $\sin''(x) < 0$ on all intervals of the form $(2n\pi, (2n+1)\pi)$ and that $\sin x$ is concave down on the intervals $[2n\pi, (2n+1)\pi]$. Similarly, $\sin x$ is concave up on all intervals of the form $[(2n+1)\pi, 2n\pi]$. $\diamondsuit$

Remark 17. Let us discuss the statement about the poverty rate which we quoted in the beginning of the chapter: *"the rate at which the rate of poverty is increasing is decreasing."* You may view it politically. The speech writer carefully designed a sentence which was not untrue, and which ended on a positive note. Something about the poverty rate was decreasing. We have encounter functions, like ln, which are increasing and concave down (the rate at which they increase decreases). These functions are not even bounded, so arbitrarily large values are obtained as we wait long enough. That would be rather bad news, in case the function described the poverty rate. The question is whether the relevant political decisions and current social and economical conditions effect the first or the second derivative of the function $P(t)$, the poverty rate as a function of time. If $P''(t) \leq -A$ for some positive number A for a sufficiently long time, then $P'(t)$ will continue to decrease at least at rate A and eventually become negative, so that the poverty rate itself would start decreasing. Maybe a good policy will be designed to have an effect on the second derivative of a quantity which needs change. It may not bring immediate relief, but eventually lasting improvement. A change in the first derivative, without control of the second one, may bring temporary relief without solving the long term problem.

Exercise 111. Find intervals on which the following functions are concave up, resp., concave down.

1. $f(x) = x^3 - 4x^2 + 8x - 7$
2. $g(x) = x^4 + 2x^3 - 3x^2 + 5x - 2$
3. $h(x) = x + 1/x$
4. $i(x) = 2x^4 - x^2$
5. $j(x) = x/(x^2 - 1)$
6. $k(x) = 2\cos^2 x - x^2$ for $x \in [0, 2\pi]$.

Concavity at a Point

The notion of being concave up or down was defined for functions which are defined on intervals. Still, we got a picture how the function has to look like near a point, and this is the behavior which we like to capture in a definition.

Definition 3.36. *Let f be a function and c an interior point* [9] *of its domain. We say that f is* concave up*, resp.,* concave down*, at c if there exists an open interval I and a line*[10] *l such that $l(c) = f(c)$ and*

$$f(x) > l(x), \text{ resp., } f(x) < l(x),$$

[9] The idea of an interior point was defined in Definition 2.1 on page 42.

[10] A line l, as required in this definition, is called a support line. In general, there could be more than one such line, but if the function is differentiable at c, then the support line is unique.

for all $x \in I$ with $x \neq c$.

In other words, we are asking for a line $l(x)$, which agrees with f at c, and on some open interval around c the function is larger (resp., smaller) that $l(x)$. The inequality is required to be strict for $x \neq c$. You see this situation illustrated in two generic pictures in Figures 3.13 and 3.14. One shows a function which is concave up at the indicated point, one shows a function which is concave down.

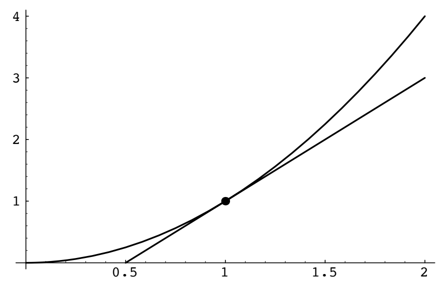

Figure 3.13: Concave up at •

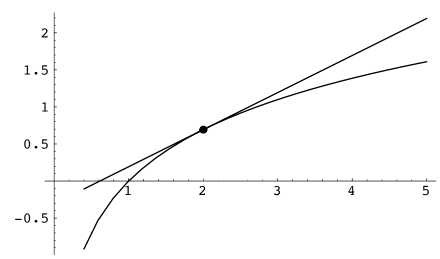

Figure 3.14: Concave down at •

There are two obvious questions. How can we detect whether a function is concave up or down at a point? What is the slope of the line l referred to in the definition? The answer to both question is given in our next theorem.

Theorem 3.37. *Let f be a function and c an interior point of its domain.*

1. *If f' is increasing at c or if $f''(c) > 0$, then f is concave up at c.*

2. *If f' is decreasing at c or if $f''(c) < 0$, then f is concave down at c.*

3. *If f is differentiable and concave up or down at c, then there is only one line which plays the role of $l(x)$ in Definition 3.36, and this line is the tangent line to the graph of f at c.*

We can use the theorem to check the concavity of a function at a point. Just calculate the second derivative of the function at the point in question, and see whether it is positive or negative. If this second derivative at the point should turn out to be 0, then the theorem is inconclusive. It does not tell us anything.

Remark 18. Item (3) in Theorem 3.37 describes the situation alluded to in Example 2 on page 40. In this case we can find the tangent line to a graph by holding a ruler against it.

Example 3.38. Check the concavity of

$$f(x) = x^5 - 7x^4 + 2x^3 + 2x^2 - 5x + 4$$

at $x = 2$.

Solution: We calculate the second derivative of f:

$$f''(x) = 20x^3 - 84x^2 + 12x + 4.$$

Evaluated at $x = 2$ we find $f''(2) = -148 < 0$. So the function is concave down at $x = 2$ $\Diamond$

Exercise 112. Decide at which points on the real line the following functions are concave up, resp., concave down:

(a) $f(x) = x^3 - 2x^2 + 5x - 3$.

(b) $f(x) = x^4 + x^3 - 3x^2 + 6x + 1$.

To relate concavity properties on an interval to those at each point in the interval we state, without proof, the following theorem.

Theorem 3.39. *Let f be a function which is defined on an open interval (a, b). Then f is concave up (resp., down) on (a, b) if and only if f is concave up (resp., down) at each point in (a, b).*

3.6 Local Extrema and Inflection Points

We are going to discuss two types of points which are particularly important in the discussion of (graphs of) functions. As we like to apply local properties of the function, we focus on interior points is the domain of the function.

Definition 3.40 (Local Extrema). *Let f be a function and c an interior point in its domain*[11]*. We say that f has a* local maximum, resp. minimum, at c *if*

$$f(c) \geq f(x), \text{ resp. } f(c) \leq f(x),$$

for all x in some open interval I around c. In this case we call $f(c)$ a local maximum, resp. minimum, of f. A local extremum is a local maximum or minimum.

[11] According to Definition 2.1 on page 42 this means that $f(x)$ is defined for all x in some open interval around c.

In other words, f has a local maximum of $f(c)$ at c if, on some open interval around c, $f(c)$ is the largest value assumed by the function. We will study tests which allow us find local extrema soon. For now, we content ourselves with an example which can be checked with bare hands.

Example 3.41. Show that the function

$$f(x) = x^2 + ax + b$$

has a local minimum at $c = -a/2$.

Solution: Completing squares, we find

$$f(x) = \left(x + \frac{a}{2}\right)^2 + \left(b - \frac{a^2}{4}\right).$$

The first expression after the equal sign in non-negative and the second one is a constant. This means that

$$f(x) \geq f(-a/2) = \left(b - \frac{a^2}{4}\right),$$

and that f has a local minimum of $(b - a^2/4)$ at $-a/2$.

E.g., the function

$$f(x) = x^2 + 2x - 1$$

has a local minimum of $f(-1) = -2$ at $x = -1$. This situation is shown in Figure 3.15. $\diamondsuit$

Exercise 113. Find the local extrema of the functions:

(a) $p(x) = x^2 + 3x - 2$.

(b) $q(x) = 3x^2 - 2x + 5$.

(c) $r(x) = -x^2 + x + 1$.

Definition 3.42 (Inflection Points). *Let f be a function and c an interior point of its domain. We call c an* inflection point *of f if the concavity of f changes at c. I.e., for some numbers a and b with $a < c < b$, we have that f is concave up on the interval $(a, c]$ and concave down on $[c, b)$, or vice versa.*

Soon we will develop tests which detect inflections points. For the moment we just give an example.

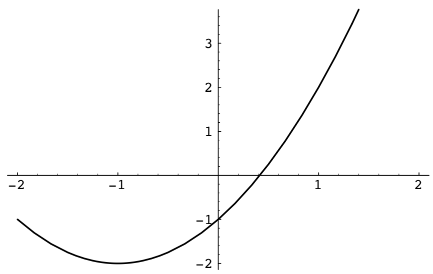

Figure 3.15: A local minimum

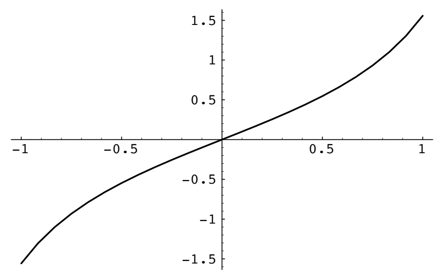

Figure 3.16: An Inflection Point

Example 3.43. Show that the function

$$f(x) = \tan x$$

has an inflection point at $c = 0$.

Solution: In Example 3.34 on page 161 we determined intervals in which the tangent function is concave up and down. Specifically, $\tan x$ is concave down on the interval $(-\pi/2, 0]$ and concave up on the interval $[0, \pi/2)$. So the concavity changes at $x = 0$ and that means that there is an inflection point at $c = 0$. You see the graph of this function in Figure 3.16. $\diamondsuit$

3.7 The First Derivative Test

In this section we discuss what is called the first derivative test. It does not detect at which points a function has local extrema, but it tells us where a function does not have a local extremum. Potentially, we would have to check every point in the domain of a function to decide whether there is a local extremum at this point, so that this could be an infinite task. Typically, the test will exclude all but a finite number of points, so that the infinite task has been reduced to a finite one.

Theorem 3.44 (First Derivative Test). *Let f be a function and c an interior point of its domain. If f is differentiable at c and $f'(c) \neq 0$, then f does not have a local extremum at c. In other words, if f has a local extremum at c, then f is either not differentiable at c or $f'(c) = 0$.*

To have an abbreviation for the points which are recognized as important in this theorem, it is customary to say:

Definition 3.45 (Critical Points). *Let f be a function and c an interior point of its domain. We say that c is a* critical point of f *if f is differentiable at c and $f'(c) = 0$, or if f is not differentiable at c.*

The first derivative test provides us with a necessary condition. If a function has a local extremum at c, then c is a critical point of the function. No local extrema can occur at points which are not critical. The test does not give a sufficient condition for a local extremum. If c is a critical point of the function, then the function need not have a local extremum at c. It makes sense to introduce one more word.

Definition 3.46 (Saddle Points). *Let f be a function and c an interior point of its domain. We say that c is a* saddle point of f *if f is differentiable at c and $f'(c) = 0$, but f does not have a local extremum at c.*

E.g., the function $f(x) = x^3$ has a saddle point at $x = 0$. This saddle point is shown in Figure 3.18. For a discussion see Example 3.50.

Proof of the First Derivative Test. Suppose that f is differentiable at c and $f'(c) > 0$. Proposition 3.27 on page 154 tells us that there exists some positive number d, such that $f(x) < f(c)$ for all $x \in (c - d, c)$, and $f(x) > f(c)$ for all $x \in (c, c + d)$. So, there are points x to the left of and arbitrarily close to c such that $f(x) < f(c)$, and there are points x to the right of and arbitrarily close to c such that $f(x) > f(c)$. This means, by definition, that f does not have a local extremum at c. If $f'(x) < 0$, then the same argument applies with inequalities reversed. If $f'(c) \neq 0$, then either $f'(c) > 0$ or $f'(c) < 0$, and in neither case we have an extremum at c. ☐

Example 3.47. Find the local extrema of the function

$$q(x) = x^2 - 2x + 3.$$

Solution: The function is differentiable for all real numbers x, and

$$q'(x) = 2x - 2 = 2(x - 1).$$

So $q'(x) \neq 0$ if $x \neq 1$. The first derivative test tells us that q does not have a local extremum at x if $x \neq 1$. The only point at which we can have a local extremum, i.e., the only critical point, is $x = 1$. If we write the function in the form

$$q(x) = (x - 1)^2 + 2,$$

then we see that q does indeed that a local minimum at $x = 1$. You should confirm this result by having a look at Figure 3.17, where this function is graphed. $\Diamond$

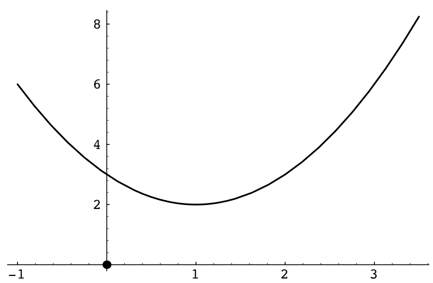

Figure 3.17: A local minimum

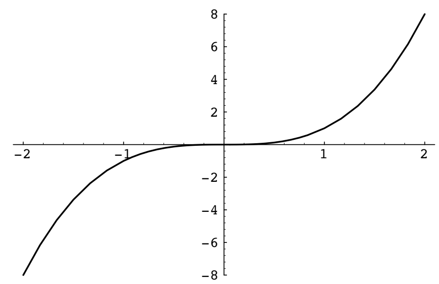

Figure 3.18: A saddle point

Example 3.48. Show that neither the exponential function nor the logarithm function have local extrema.

Solution: To verify this, remember that the derivative of $f(x) = e^x$ is $f'(x) = e^x$. This function is nowhere zero (see Theorem 1.12). Thus it has no critical point. The first derivative test tells us that the function has no local extrema.

A similar argument applies to the natural logarithm function, which is defined on the interval $(0, \infty)$. Its derivative is $\ln' x = 1/x$, and this function is nowhere zero on $(0, \infty)$. Hence the natural logarithm function has no critical points and no local extrema. $\Diamond$

Example 3.49. Find the local extrema of $f(x) = \sin x$.

Solution: As we have shown previously, $f'(x) = \cos x$, and $f'(x) = 0$ if and only if x is of the form $n\pi + \pi/2$, where n is an integer. These are the only critical points, and the only points where $\sin x$ can have a local extremum.

Observe that $\sin(n\pi + \pi/2) = 1$ if n is even, and that $\sin(n\pi + \pi/2) = -1$ if n is odd. For all x not of this form we have that $-1 < \sin x < 1$. It follows $\sin x$ as local maxima at the points of the form $n\pi + \pi/2$ for n even and local minima for n odd. $\Diamond$

Example 3.50. Show that the function

$$g(x) = x^3$$

has no local extrema, and that it has a saddle point at $x = 0$.

Solution: To see this, we differentiate g. The derivative is $g'(x) = 3x^2$, and this function is zero only when $x = 0$. The only critical point of g is at $x = 0$. The first derivative test tells us that the only point at which we can have a local extremum is $x = 0$. Our task of searching for local extrema has been substantially reduced. There is only one point left at which we have to have a closer look at the function.

Obviously, $g(x) > 0$ for all $x \in (0, \infty)$ and $g(x) < 0$ for all $x \in (-\infty, 0)$. This means that there is no local extremum at $x = 0$. As $g'(0) = 0$ and there is no local extremum at $x = 0$, the function has a saddle point at this point. $\diamondsuit$

Let us formulate a criterion which, based on first derivative information, confirms that a function has a local extremum at a point c. It gives us a sufficient condition for a local extremum to exist. Suppose c is an interior point of the domain of a function f, and suppose that for some $d > 0$ the function is increasing on $(c - d, c]$ and decreasing on $[c, c + d)$. Then f has a local maximum at c. Taking advantage of the information provided by the first derivative, we obtain the following test.

Theorem 3.51. *Suppose f is a function which is defined and differentiable on $(c-d, c+d)$ for some $d > 0$. If $f'(x) > 0$ for all $x \in (c-d, c)$ and $f'(x) < 0$ for all $x \in (c, c + d)$, then f has a local maximum at c. If $f'(x) < 0$ for all $x \in (c-d, c)$ and $f'(x) > 0$ for all $x \in (c, c+d)$, then f has a local minimum at c.*

Let us illustrate the use of the theorem with an example. You may revisit the example once we discussed the second derivative test to find a simplified argument for our conclusions.

Example 3.52. Find the local extrema of the function

$$f(x) = x^3 - 3x^2 + 2x + 2.$$

Solution: We differentiate the function, find the roots of the derivative, and factor it, so that it is easy to see where f' is positive and negative.

$$f'(x) = 3x^2 - 6x + 2 = 3\left[x - \left(1 + \frac{\sqrt{3}}{3}\right)\right]\left[x - \left(1 - \frac{\sqrt{3}}{3}\right)\right]$$

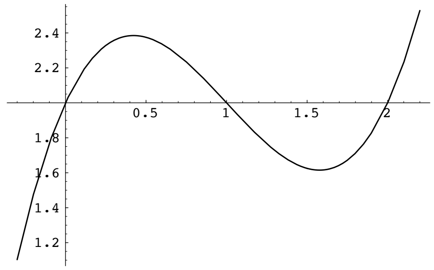

Figure 3.19: $f(x) = x^3 - 3x^2 + 2x + 2$

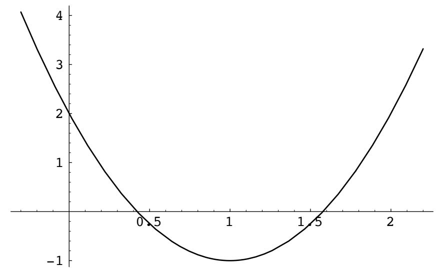

Figure 3.20: $f'(x) = 3x^2 - 6x + 2$

The expressions within the square brackets are lines. The first one of them is negative on the interval $(-\infty, 1 + \sqrt{3}/3)$ and positive on $(1 + \sqrt{3}/3, \infty)$. The second one is negative on the interval $(-\infty, 1 - \sqrt{3}/3)$ and positive on $(1 - \sqrt{3}/3, \infty)$. Taken together, $f'(x) = 0$ if $x = 1 \pm \sqrt{3}/3$, $f'(x)$ is positive on the intervals $(-\infty, 1 - \sqrt{3}/3)$ and $(1 + \sqrt{3}/3, \infty)$, and $f'(x)$ is negative on the interval $(1 - \sqrt{3}/3, 1 + \sqrt{3}/3)$. You can see graphs of f and f' in Figures 3.19 and 3.20

With this we may conclude that $x = 1 \pm \sqrt{3}/3$ are the only critical points of f, and that these are the only points where a local extremum can occur. Based on the sign of $f'(x)$ on intervals to the left and right of these two critical points we see that f has a local maximum at $x = 1 - \sqrt{3}/3$ and a local minimum at $x = 1 + \sqrt{3}/3$. $\diamondsuit$

Exercise 114. Find the local extrema of the following function:

$$(1)\ f(x) = \frac{x^2 + 3x}{x - 1} \qquad (2)\ g(x) = \sin 2x + 2 \sin x \text{ for } x \in [0, 2\pi].$$

Hint: We discussed the monotonicity properties of these functions in Examples 3.23 and 3.24.

Exercise 115. Find the local extrema of the following function:

(a) $f(x) = 3x^2 + 5x + 7$

(b) $f(x) = x^3 - 3x^2 + 6$

(c) $f(x) = (x+3)/(x-7)$

(d) $f(x) = x + 1/x$

(e) $f(x) = x^3(1+x)$

(f) $f(x) = x/(1+x^2)$

(g) $f(x) = \cos 2x + 2\cos x$ for $0 \leq x \leq 2\pi$

(h) $f(x) = \sin^2 x - \sqrt{3}\sin x$ for $0 \leq x \leq 2\pi$.

Hint: You discussed the monotonicity properties of these functions before.

3.8 The Second Derivative Test

The second derivative test provides us with a sufficient criterion for an extremum of a function at a point. When its assumptions are satisfied at a point c, then the function has a local extremum at this point. We can also tell whether it is a maximum or a minimum. Here is the test:

Theorem 3.53 (Second Derivative Test). *Let f be a function and c an interior point in its domain. Assume also that $f'(c)$ and $f''(c)$ exist and that $f'(c) = 0$. If $f''(c) > 0$, then f has a local minimum at c. If $f''(c) < 0$, then f has a local maximum at c.*

Stated differently the theorem says: If f has a critical point c, and if $f''(c)$ exists and is nonzero, then f has a local extremum at c. The sign of $f''(c)$ tells us whether the extremum is a maximum or a minimum. No statement is made in the theorem when $f''(c) = 0$. In fact, if $f'(c) = f''(c) = 0$, then there may or may not be a local extremum at c. Furthermore, the function f can have a local extremum at c, and the assumptions of the test are not satisfied. In this sense, the test provides us with a sufficient condition for the existence of a local extremum at a point. It does not provide us with a necessary condition.

The second derivative test is very easy to apply. If we only use first derivative techniques to detect local extrema, then we have to decide about the sign of the first derivative on intervals on both sides of a critical point. This can be a rather unpleasant task. In comparison, if we apply the second derivative test for this purpose, then we only have to evaluate the second derivative of a function at a critical point to find the desired information. Let us look at some examples.

Example 3.54. Find the local extrema of

$$q(x) = x^2 - 2x + 3.$$

Solution: We discussed this function in Example 3.47, see also Figure 3.17. We recognized $x = 1$ as the only critical point of this function. Apparently $q''(x) = 2$ for all x, so that the second derivative test tells us that q has a local minimum at $x = 1$. We saw this earlier in the graph and checked it by hand. $\diamondsuit$

Example 3.55. Find the local extrema of the function (for a graph, see Figure 3.7 on page 150)

$$p(x) = x^3 - 3x^2 - 9x + 3.$$

First Solution: We found previously, see Example 3.22, that p is increasing on the interval $(-\infty, -1]$ and on the interval $[3, \infty)$. The function is decreasing on the interval $[-1, 3]$. This information suffices to conclude that the function has a maximum at $x = -1$ and a minimum at $x = 3$.

Second Solution: We calculated the first derivative,

$$p'(x) = 3x^2 - 6x - 9 = 3(x + 1)(x - 3),$$

and saw that the critical points of the function are $x = -1$ and $x = 3$. Next we calculate the second derivative of the function:

$$p''(x) = 6x - 6 = 6(x - 1).$$

In particular, $f''(-1) = -12$ and $f''(3) = 12$. The second derivative test tells us that we have a local maximum at $x = -1$, because this is a critical point and $f''(-1) < 0$. We also have a local minimum at $x = 3$ because at this critical point the second derivative of the function is positive. $\diamondsuit$

Example 3.56. Use the second derivative test to find the local extrema of the function

$$f(x) = \sin x.$$

Earlier we found the critical points for this function, see Example 3.49. They are the points of the form $n\pi + \pi/2$, where n is a natural number. According to the first derivative test, they are also the only points at which we can have local extrema. The second derivative of the function is

$$f''(x) = -\sin x,$$

and as you may check, using the geometry of the circle, $f''(x) = -1$ if $x = n\pi + \pi/2$ and n is even. The second derivative test tells us that the sine function has local maxima at these points. Furthermore, $f''(x) = 1$ if $x = n\pi + \pi/2$ and n is odd. So, at these points the sine function has local minima. This conclusion should confirm what you may have suspected after inspecting the graph of the sine function in Figure 5.9. $\diamondsuit$

Proof of the Second Derivative Test. First, let us assume that $f'(c) = 0$ and $f''(c) > 0$. We would like to show that f has a minimum at c. The assumption that $f'(c) = 0$ means that the tangent line to the graph of f at $(c, f(c))$ is horizontal. Its equation is

$$l(x) = f(c).$$

The assumption that $f''(c) > 0$ means that f is concave up at c (see Theorem 3.37 (1)). Spelled out explicitly this means that

$$f(x) > l(x)$$

for some positive number d and for all $x \in (c - d, c) \cup (c, c + d)$. In other words, f has a local minimum at c.

The proof that f has a local maximum at c if $f'(c) = 0$ and $f''(c) < 0$ is similar. We leave it to the reader. $\square$

Exercise 116. Find the critical points and the local extrema.

(a) $f(x) = 4x^2 - 7x + 13$

(b) $f(x) = x^3 - 3x^2 + 6$

(c) $f(x) = x + 3/x$.

(d) $f(x) = x^2(1 - x)$

(e) $f(x) = |x^2 - 16|$

(f) $f(x) = x^2/(1 + x^2)$

3.9 Extrema of Functions

In many applications we are concerned with finding the extrema of a function. Let us start out with an example.

Example 3.57. The owner of a freighter has to decide how fast the ship should travel to deliver its cargo. The freighter is on a trip from San Francisco to Honolulu. The distance between the ports is about $2,300$ nautical miles. The maximal speed of the freighter is 25 knots (nautical miles per hour).

Let us say that the fixed expenses for operating the ship are \$10,000 per day. The income received for the trip is \$120,000, independently from the time the trip takes. In addition, the owner has to consider the fuel expenses. The ship's engine uses heavy crude, which costs \$40.00 per 1,000 liter. The problem is, that the fuel consumption increases as the ship goes faster. Let us say that at a speed of k knots the fuel consumption is

$$c(k) = 200 + 400e^{k/6} \text{ liter per hour.}$$

Solution: The first question is, which income or profit should we try to express as a function of the speed and eventually maximize. Assuming that cargo is waiting in Honolulu to be picked up by the ship for the next trip, we should try to maximize the hourly profit for the owner of the ship.

So let us denote the speed of the ship by k (knots). As a function of k, let us calculate the income and expenses per hour. The trip will take $2300/k$ hours. The income per hour (denoted by I) is

$$I(k) = \frac{120,000}{2,300/k} = \frac{1,200}{23}k.$$

The expenses per hour (denoted by E) are the sum of the fixed expenses and the fuel expenses. Expressed as a formula,

$$E(k) = \frac{10,000}{24} + .04(200 + 400e^{k/6}).$$

The resulting net profit per hour is the difference of the income and the expenses:

$$\begin{aligned} P(k) &= I(k) - E(k) \\ &= \frac{1,200}{23}k - \left[\frac{10,000}{24} + .04(200 + 400e^{k/6})\right] \end{aligned}$$

As we said, we like to maximize the hourly profit. In an attempt to find local maximum for this function, we differentiate $P(k)$.

$$P'(k) = \frac{1,200}{23} - .04\frac{400}{6}e^{k/6}.$$

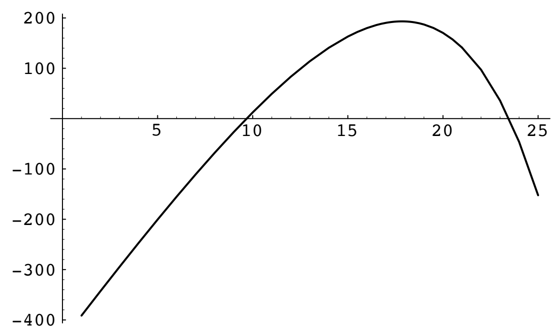

Figure 3.21: Profit as a function of speed.

After some simplifications we find that $P'(k) = 0$ if and only if

$$e^{k/6} = \frac{450}{23} \quad \text{or} \quad k = 6 \ln \left(\frac{450}{23} \right) = 17.84.$$

The numerical value is approximate. We also note that $P''(k) < 0$ for all k. So, our calculation reveals a local maximum for $P(k)$ when $k = 17.84$ knots. Plugging this value for k into the equation for P we find that for this speed the profit is about \$193.20 per hour. You see the graph of the function $P(k)$ in Figure 3.21. It confirms the calculation we have just gone through.

Have we solved the problem? Yes, but we need an additional argument to draw this conclusion. For the moment we accept that there is no other speed at which the profit exceeds the one at the local maximum. The graph supports this claim. We still need to develop the mathematical ideas which allow us to discuss such problems in general, so that we can come to the specific conclusion in the example. $\diamondsuit$

We define the notion of an absolute maximum and minimum of a function. We remind the reader that the domain of a function is the set of those points c for which $f(c)$ is defined.

Definition 3.58. *Let f be a function, and c a point in its domain. We say that f has an* absolute maximum *at c if $f(x) \leq f(c)$ for all x in the domain* *of f. Then we call $f(c)$ the* absolute maximum *of f. If $f(x) \geq f(c)$ for all x in the domain of f, then we say that f has an* absolute minimum *at c, and we call $f(c)$ the* absolute minimum *of f.*

Instead of saying that a function has an absolute extremum at c, we also say that it assumes the absolute extremum at this point. Without being specific about the point, we say that a function assumes its maximum if it does so at some point in its domain. Some functions assume their absolute extrema, others do not. Let us give an important result which tells us that, under appropriate assumtions, a function has absolute extrema, and where they occur. Then we give examples.

Theorem 3.59. *Let f be a function which is defined and differentiable*[12] *on $[a, b]$. Then f assumes its absolute maximum and minimum on $[a, b]$*[13]*. If f assumes its absolute maximum or minimum at c, then c is an endpoint of the interval or $f'(c) = 0$.*

Proof. The proof of the first claim in the theorem (for continuous functions) makes use of the completeness of the real numbers, and is typically provided in an introductory real analysis course.

Suppose f has an absolute extremum at c. If c is an interior point of the interval (i.e., $c \in (a, b)$) and $f'(c) \neq 0$, then f is either increasing or decreasing at c and cannot have an absolute extremum at this point. To have an absolute extremum at c we must have that $f'(c) = 0$. The only other possibility is that c is an endpoint of the interval. This is just what we claimed in the last sentence of the theorem. $\square$

Example 3.60 (Conclusion of Example 3.57). According to the theorem, the maximal profit for the owner of the freighter must be realized if the ship travels about 17.84 knots (the critical point for the speed which we found previously), or at the smallest or greatest speed the ship is capable of. Apparently, travelling at the smallest possible speed ($k = 0$) does not make any sense. Travelling at the maximal possible speed of 25 knots does not realize the maximal profit either. This is apparent if you look at the graph of the profit function, see Figure 3.21. You may also confirm this by plugging this value for k into the formula for the profit. You will see that

[12] We remind the reader that at function is said to be differentiable on a closed interval if the function can be extended to a function on an open interval which contains the closed one, and the function is differentiable on the open interval.

[13] To show this claim, it suffices to assume that the function is continuous on the interval. Every differentiable function is continuous. This means that the theorem holds under weaker assumptions than those we are using here.

$P(25)$ is approximately \$-152.32 per hour. So you even make a loss at this speed. Having excluded the endpoints as absolute extrema of the function, we conclude that the maximal profit is achieved if the ship travels at a speed of about 17.84 knots. $\diamondsuit$

Example 3.61. Find the absolute extrema of the function

$$f(x) = x^3 - 5x^2 + 6x + 1$$

for $x \in [0, 4]$.

Solution: We like to apply Theorem 3.59. Its assumptions are satisfied. The function is differentiable on the given interval $[0, 4]$ because it extends to a differentiable function on the entire real line.

Let us calculate the first derivative of $f(x)$, so that we can determine the critical points:

$$f'(x) = 3x^2 - 10x + 6.$$

The solutions of the quadratic equation $f'(x) = 0$ are $x = (5 \pm \sqrt{7})/3$. Given as decimal expansion, the roots of the quadratic equation are about 2.5486 and .7848. You may also check that $f''(x) = 6x - 10$, and

$$f''((5 + \sqrt{7})/3) > 0 \quad \text{and} \quad f''((5 - \sqrt{7})/3) < 0.$$

The second derivative test tells us that the function has a local minimum at $x = (5 + \sqrt{7})/3$ and a local maximum at $x = (5 - \sqrt{7})/3$. The approximate values of the function at these points are

$$f((5 + \sqrt{7})/3) = 3.1126 \quad \text{and} \quad f((5 - \sqrt{7})/3) = .3689.$$

In addition, we have to inspect the values of the function at the endpoints of the interval. We find

$$f(0) = 1 \quad \text{and} \quad f(4) = 9.$$

This means that the function assumes its absolute maximum of 9 at $x = 4$, and its absolute minimum of approximately .3689 at $x = (5 + \sqrt{7})/3$.

You may compare our calculation with the graphs of f in Figure 3.22 and the one of f' in Figure 3.23. $\diamondsuit$

Exercise 117. Find the absolute extrema of the functions on the indicated intervals.

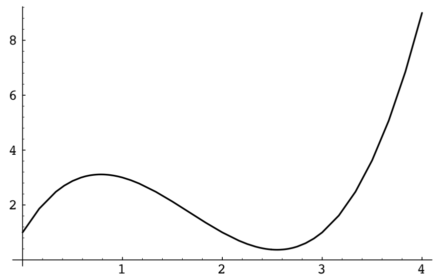

Figure 3.22: $x^3 - 5x^2 + 6x + 1$.

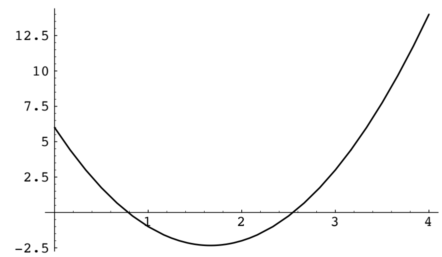

Figure 3.23: $3x^2 - 10x + 6$

(a) $f(x) = x^2 - 5x + 2$ for $x \in [0, 5]$

(b) $f(x) = x^3 + 3x^2 - 5x + 2$ for $x \in [-3, 2.5]$

(c) $f(x) = \sqrt{2 + x}/\sqrt{1 + x}$ for $x \in [0, 5]$

(d) $f(x) = \cos 2x + 2\cos x$ for $0 \le x \le 2\pi$

(e) $f(x) = \sin x + \cos x$ for $0 \le x \le 2\pi$

Example 3.62. Find the rectangle with a given perimeter so that its area is maximal.

Solution: To make this problem accessible, we introduce some notation and formulate it as a maximization problem. Let us call the perimeter of the rectangle P and its area A. To relate these two quantities we also give the sides of the rectangle names, call them w and l. Then we have

$$P = 2(w + l) \quad \text{and} \quad A = w \times l.$$

The sides of the rectangle are related, $l = P/2 - w$, and we find an expression for A as a function of w:

$$A(w) = w \times (P/2 - w).$$

There is an obvious restriction on w, it must lie in the interval $[0, P/2]$.

Mathematically formulated, our problem is now to find the absolute maximum of $A(w)$ for $w \in [0, P/2]$.

We learned how to approach such a problem. First we find the local extrema of the function A, and to do so we differentiate the function.

$$A'(w) = \frac{P}{2} - 2w.$$

We find that $A'(w) = 0$ exactly if $w = P/4$. This means that $l = P/4$, and that the rectangle is a square. It is rather apparent that the area has a local maximum when the shape of the rectangle is a square, but we may verify this using the second derivative test. The statement follows from the fact that $A''(w) = -2 < 0$.

We still have to check what happens if we look at the endpoints of the interval in which w takes its values, i.e., if $w = 0$ or $w = P/2$. In either case $A = 0$, so that the maximal value of the area of the rectangle is not obtained at an endpoint. This means, the largest area rectangle for a fixed perimeter is a square. $\diamondsuit$

There are many problems which can be analyzed via the methods presented in the previous example. The details will be different, but the steps are the same. We are presented with a 'real life' optimization problem. We introduce mathematical notation so that the problem is formulated as a mathematical optimization problem, i.e., a problem in which we are suppose to find the absolute maximum of minimum of a function. The problem may involve more than one variable. We use relations among the variables to obtain a maximization or minimization problem for a function in one variable. Then we apply the first or second derivative test to find the extrema of the function. With some additional thoughts, e.g., after checking the values at the function at the end points of its domain, we find the absolute extrema of the function which we set out to optimize. Then we draw the conclusion about the problem we started out with. Let us practice a few more examples.

Example 3.63. Find the shape of a round drum with a given surface area so that its volume is maximal.

To assure yourself that you are solving the right problem, you may ask first what type of drum should be considered. Does the drum have a bottom, a lid, or both? Let's solve the case where the drum has both, and leave the discussion of other drums to the reader.

Solution: We start our by organizing the information which we have. The surface area and the volume of a drum are functions of the radius of the drum and its height. To assure good communication with those we like to explain the solution to, we denote the height of the drum by h and its radius by r. In addition we denoted the surface area of the drum by A and its volume by V. The area of the top and bottom are each πr^2, and the area of the side is the circumference of the drum times its height, or $2\pi rh$. Adding these parts of the surface area we find that the total surface area of the drum is

$$A = 2\pi r^2 + 2\pi rh = 2\pi r(r + h).$$

This formula provides us with a relation between r and h for a given fixed surface area A. Specifically,

$$h = (A - 2\pi r^2)/2\pi r. \tag{3.5}$$

In addition, the radius of the drum has to be non-negative, and it cannot be too big, as the area of its top and bottom ($2\pi r^2$) must not exceed A. Specifically, we conclude that

$$0 \leq r \leq \sqrt{A/2\pi}.$$

We calculate the volume of the drum by multiplying the area of the base with the height, i.e.:

$$V = \pi r^2 h.$$

As it stands, V depends on two variables (r and h), and to make it accessible to our methods we have to write it as a function of one variable. We use the relation between r and h from (3.5), and find a formula for the volume of the drum as a function of its radius:

$$V(r) = \pi r^2 h = \pi r^2 (A - 2\pi r^2)/2\pi r = r(A - 2\pi r^2)/2.$$

With this, we have reformulated our practical problem into a purely mathematical one.[14] We have to find the absolute maximum of the function

$$V(r) = r(A - 2\pi r^2)/2 \quad \text{with} \quad r \in [0, \sqrt{A/2\pi}].$$

We solve it using Theorem 3.59 and the second derivative test. An easy calculation provides us with the first and second derivatives of V:

$$V'(r) = (A - 6\pi r^2)/2 \quad \& \quad V''(r) = -6\pi r.$$

[14] To be cautious, we need to remember that we divided by r in the previous calculation, and that is not permissible if $r = 0$. The formula tells us that $V(0) = 0$, but it makes little sense to talk about a drum of radius zero. As we will see, $r = 0$ will not be the viable solution for the optimization problem, and in this sense we can ignore this point at which the mathematical formulation of the problem may not agree with the real problem.

We determine the critical points of $V(r)$. Apparently,

$$V'(r) = 0 \text{ if and only if } A = 6\pi r^2, \text{ or } r = \sqrt{A/6\pi}.$$

Observe that $V''(r) < 0$ at the critical point. The second derivative test tells us that $V(r)$ has a local maximum when $r = \sqrt{A/6\pi}$.

To apply Theorem 3.59, we check the values of V at the endpoints of the interval, $r = 0$ and $r = \sqrt{A/2\pi}$. In either case we get $V(r) = 0$. For $r = \sqrt{A/6\pi}$ the volume V is certainly positive. Theorem 3.59 tells us that the absolute maximum of $V(r)$ for $0 \le r \le \sqrt{A/2\pi}$ is realized at $r = \sqrt{A/6\pi}$. We conclude that we obtain the round drum with a maximum volume if we choose $r = \sqrt{A/6\pi}$ as radius.

We now determine the height of the drum of maximal volume. Substituting $r = \sqrt{A/6\pi}$ into the formula for h gives us $h = \sqrt{2A/3\pi}$. Comparing r and h, we find that $h = 2r$. In conclusion, a drum with a given surface area will have maximal volume if its diameter is equal to its height. $\diamondsuit$

Remark 19. Why are the proportions of drums, say soda cans, different from what we would suggest as the optimal shape? First of all, you can cut the sides of the can without much loss of material to the given specifications. Before they are bent to form the sides of the can, they are rectangular. On the other hand, if you cut the round top and bottom for the can out of a sheet of metal, then some material is wasted, although it may be recycled at a cost. You may also have noticed that the top and bottom of a can are made from stronger material than its sides. To optimize the cost of the can, both arguments suggest that, in comparison, r should be smaller than suggested by our calculation, and as you see in real life, that the height of the can (or drum) should be larger than the diameter.

Example 3.64. Determine the rectangle of maximal area which can be placed between the x-axis and the graph of the function $f(x) = \sin x$.

Solution: Draw a graph of $\sin x$ so that you can follow the discussion. Because of the repeating pattern of the sine function, we restrict ourselves to the interval $[0, \pi]$, and place the rectangle between the (horizontal) x-axis and the graph of the sine function. After drawing in few rectangles, you should agree, that it will be best to place one side of the rectangle on the x-axis, and then make the rectangle as tall as possible. You will also see that it will be best to take a 'symmetric' picture. Specifically, the vertices of the rectangle should be the points $(x, 0)$, $(\pi - x, 0)$, $(x, \sin x)$ and $(\pi - x, \sin x)$ for some $x \in [0, \pi/2]$. The width of the rectangle is $\pi - 2x$ and its height is $\sin x$, so that its area is

$$A(x) = (\pi - 2x)\sin x.$$

We need to find the absolute maximum for this function for $x \in [0, \pi/2]$.

The first derivative of this function is

$$A'(x) = -2\sin x + (\pi - 2x)\cos x.$$

After a simple algebraic simplification, you find that

$$A'(x) = 0 \quad \text{if and only if} \quad \tan x = \frac{\pi - 2x}{2}.$$

We find an approximate solution of the equation. You may use Newton's method, or you may use your graphing calculator. Anyway, a fairly good approximation of the zero of $A'(x)$ is $x_0 = .710462$. You should convince yourself[15] that this is the only zero of $A'(x)$ for $x \in [0, \pi/2]$. We conclude that $A(x)$ has only one critical point, and this critical point is about at $x_0 = .710462$.

You may calculate $A''(x)$. Substituting x_0 you will see that $A''(x_0) < 0$. It follows from the second derivative test that $A(x)$ has a local maximum at x_0. Apparently $A(x) = 0$ at the endpoints $x = 0$ and $x = \pi/2$ of the interval. This tells us that $A(x)$ assumes its absolute maximum at x_0.

With this, the final answer to our problem is: The rectangle of maximal area which can be placed between the x-axis and the graph of the sine function will have a width of approximately $\pi - 2x_0 = 1.72066$ and a height of $\sin x_0 = .652183$. Its area will be about 1.12218. ◊

Exercise 118. Find the largest possible area for a rectangle with base on the x-axis and upper vertices on the curve $y = 4 - x^2$.

Exercise 119. What is the largest possible volume for a right circular cone of slant height a?

As we stated in Theorem 3.59 on page 176, a differentiable function on a closed interval of the form $I = [a, b]$ assumes its absolute extrema. If we consider functions on different kinds of intervals, then this need not be the case. E.g., the function $f(x) = x$ does not have an absolute maximum or minimum on intervals such as $(-\infty, \infty)$ or $(-1, 1)$. In many applications the following result is very useful.

Theorem 3.65. *Suppose f is defined on an interval I.*

(a) *If f is concave up on I and has a local minimum at x_0, then f assumes its absolute minimum at x_0.*

[15] One possible argument is that $\tan x$ is increasing on the interval $[0, \pi/2)$, and that $\frac{\pi-2x}{2}$ is decreasing. So these functions can intersect in only one point.

(b) *If f is concave down on I and has a local maximum at x_0, then f assumes its absolute maximum at x_0.*

(c) *Suppose f is twice differentiable on I and $f''(x) > 0$ for all $x \in I$. If $f'(x_0) = 0$, then f has a local and absolute minimum at x_0.*

(d) *Suppose f is twice differentiable on I and $f''(x) < 0$ for all $x \in I$. If $f'(x_0) = 0$, then f has a local and absolute maximum at x_0.*

Example 3.66. Find the absolute minimum of the function

$$f(x) = x + \frac{1}{x}$$

for $x \in (0, \infty)$.

Solution: We calculate the first and second derivative of $f(x)$:

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

We find that $f'(x) = 0$ if $x = 1$, and that $f''(x) > 0$ for all x in $(0, \infty)$. Part (c) of the theorem tells us that the function assumes it absolute minimum at $x = 1$. The absolute minimum of the function is $f(1) = 2$. $\diamondsuit$

Exercise 120. Consider a triangle in the plane with vertices $(0, 0)$, $(a, 0)$, and $(0, b)$. Suppose that a and b are positive, and that $(2, 5)$ lies on the line through the points $(a, 0)$, and $(0, b)$. What should the slope of the line be, so that the area of the triangle is minimal?

We will consider more optimization problems in Section 3.11.

3.10 Detection of Inflection Points

We defined inflection points as points where the concavity of a function changes. Let us start out with an example.

Example 3.67. Find the the inflection points of the function

$$g(x) = x^3 - 4x^2 + 3x - 5.$$

You see the graph of g in Figure 3.24 and the one of g' in Figure 3.25. We calculate the first and second derivative of g:

$$g'(x) = 3x^2 - 8x + 3 \quad \text{and} \quad g''(x) = 6x - 8.$$

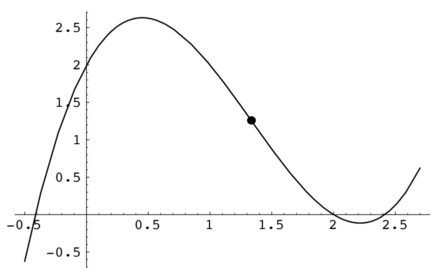

Figure 3.24: The graph of g.

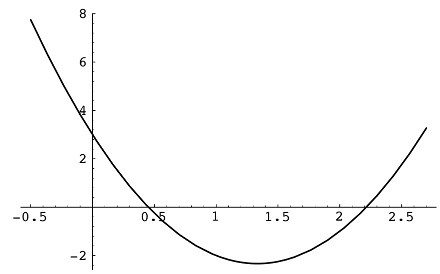

Figure 3.25: The graph of g'.

From the formula for the second derivative we conclude that

$$g''(x) < 0 \text{ if } x \in (-\infty, 4/3) \text{ and that } \quad g''(x) > 0 \text{ if } x \in (4/3, \infty).$$

This means that g is concave down on the interval $(-\infty, 4/3]$ and concave up on $[4/3, \infty)$. By definition, we have an inflection point at $x = 4/3$. You see the inflection point indicated as a dot in Figure 3.24. You see that $g'(x)$ has a local extremum at the same point.

Let us go through the same argument once more, but more geometrically. On the left of $x = 4/3$ the function $g'(x)$ is decreasing, and on the right of this value of x, $g'(x)$ is increasing. At the same time, on the left of $x = 4/3$ the derivative of $g'(x)$, i.e., $g''(x)$, is negative. On the right of this value of x, $g''(x)$ is positive. This means that g is concave down on the interval $(-\infty, 4/3]$ and concave up on $[4/3, \infty)$, and that we have an inflection point at $x = 4/3$. $\diamondsuit$

After this somewhat pure example, let us think about the relevance of inflection points. Let $f(t)$ denote the number of people who are infected with the HIV virus at time t. If $f'(t)$ is increasing during a certain period of time, i.e., if f is concave up on the interval in time, then an increasing number of people get infected, or the disease spreads at a increasing rate.

What happens as we pass an inflection point? If from some point in time on $f''(t) < 0$ or f is concave down, then f' decreases, and that means that the the rate at which the disease spreads slows down. The rate at which people get infected may still going up, but this rate does not increase as fast as previously. In mathematical terms, you passed a local maximum of f'. You will declare the passage of the inflection point a milestone in the fight against AIDS. You might say (or at least hope) that you have turned the corner in the fight. Why? You hope that the "trend" of the first derivative continues, i.e., that f' continues to decrease. You hope that f' will eventually become negative. Only this would mean that the number of HIV infected people actually declines.

So, in applications inflection points indicate turning points in the direction in which a function develops (if trends continue), and for this reason it is interesting to discuss them.

After the example and the motivation, you should appreciate a theorem which tells how to exclude points as inflection points and how to find them.

Theorem 3.68. *Let f be a function and c an interior point of its domain. Suppose that the first and second derivatives of f exist at c.*

1. *If f has an inflection point at c, then $f''(c) = 0$.*
2. *If $f''(c) = 0$, $f'''(c)$ exists and $f'''(c) \neq 0$, then f has an inflection point at c.*

Let us illustrate the use of the theorem with examples.

Example 3.69. Find the inflection points of

$$h(x) = \sin x.$$

Solution: The theorem suggests that we differentiate the function twice. We learned previously that

$$h''(x) = -\sin x.$$

The only zeros of h'' are real numbers of the form $n\pi$ where n is an integer. So, these are the only numbers where h can have a inflection point. We also know that $h'''(n\pi) = -\cos(n\pi) = \pm 1$. The theorem tells us that the inflection points of $\sin x$ are exactly the points of the form $n\pi$ where n is an integer. $\diamondsuit$

Example 3.70. Find the inflection points of

$$f(t) = 2t^4 - 6t^3 + 5t^2 - 7t + 4.$$

Solution: We calculate once more the second derivative of the function and find

$$f''(t) = 24t^2 - 36t + 10.$$

According to the theorem, we have to find the zeros of this function to determine where an inflection point can be. Solve the quadratic equation $f''(t) = 0$. The roots are

$$t = \frac{3}{4} \pm \frac{1}{12}\sqrt{21} = \frac{9 \pm \sqrt{21}}{12}.$$

Now, let us check whether there are inflection points at either of these values for t. We calculate the third derivative of f:

$$f'''(t) = 48t - 36.$$

We could plug $t = (9 \pm \sqrt{21})/12$ into the expression for f''', but this is a bit cumbersome. We see right away that $f'''(t) = 0$ exactly if $t = 3/4$, and this means that $f'''(9 \pm \sqrt{21})/12) \neq 0$. The theorem says that the inflection points of $f(t)$ are at $t = (9 \pm \sqrt{21})/12$. $\diamond$

Apparently, our ability it find inflection points of a function is limited by our ability to find the zeros of its second derivative. This may be a difficult task, particularly if you try it by hand. The task becomes a lot easier if you can rely on technology to graph the functions in question. Let us use technology for an example which would otherwise create a huge headache for you. Still, we use a function which is defined by an explicit analytic expression.

Example 3.71. Find the inflection points of the function

$$f(x) = \sqrt{1.2 + x^2 - 3(\sin x)^3}.$$

Apparently, it will take an effort to calculate the second derivative of this function, and it will be nearly impossible to find the zeros of f''. Any reasonable software has no problem with this. We asked the computer to graph f and f'' for $x \in [-3, 3]$. You see the graphs in Figures 3.26 and 3.27.

We have no problem finding approximate values of x at which f'' has a zero. Our graph shows five such values, the largest of which is about 1.8. These five zeros are only candidates for inflection points. You would still have to plug the values into the third derivative of f to verify that you actually have inflection points. But, the graph of f'' shows that f'' has a non-zero slope at those five points, so that f''', the derivative of f'', is non-zero at each of them. This means all five zeros of f'' give us inflection points of f.

A look at the graph of f barely reveals some of the inflection points, but the graph of f'' shows them clearly. Zooming in on parts of the graph will

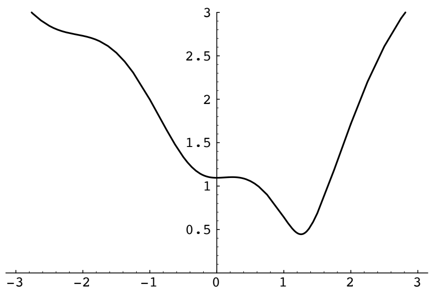

Figure 3.26: The graph of f.

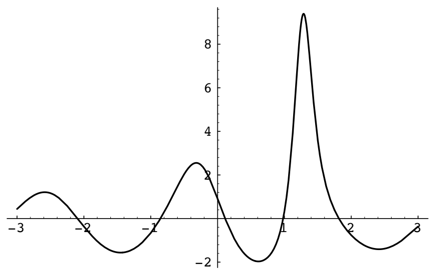

Figure 3.27: The graph of f''.

not improve this. At least in this example, the graph of f'' tells us much more about the concavity of the function f than its own graph.

The graph of f'' suggests that this function has additional zeros outside of the range over which we graphed the function. So, there may be further inflection points which we have not found yet. Use your software to study the example further. $\diamondsuit$

Remark 20. Let us have another look at the conditions in Theorem 3.68 and compare them with the ones in Theorem 3.53. We want to relate the inflection points of a function to the extrema of its derivative. So, suppose that the function f is defined and differentiable on an interval (a, b). Set $f' = g$ and let c be a point in (a, b). Apparently saying that $f''(c) = 0$ and $f'''(c) \neq 0$ is the same as saying that $g'(c) = 0$ and $g''(c) \neq 0$. The condition on f implies that f has an inflection point at c, and the condition on $g = f'$ says that f' has a local extremum at c. So, at least as long as we are considering inflection points which are detected by Theorem 3.68, the inflection points of a function correspond to local extrema of the derivative.

3.11 Optimization Problems

In this section we like to solve some more optimization problems. We solved some in Section 3.9.

Example 3.72. Suppose you want to fence off a corral. Based on the number of cattle which you like to keep in it, you decide that its area should be $6,000$ square meters. To make the job of building the corral simple, you decide that the corral should be rectangular. You also want to use as little fencing material as possible. How long should its sides be?

Solution: To make the problem accessible to a mathematical discussion, you consider the problem more abstractly. You strip away the irrelevant details and find the dimensions of a rectangle with a given area and shortest possible circumference.

To be able to discuss the problem analytically, you introduce some notation. Let us call the area of the rectangle A, the circumference C, one side length l and the other one w. There are some relations among these quantities:

$$A = w \times l \quad \text{and} \quad C = 2w + 2l.$$

Apparently, C depends on two variables, w and l, and this is more than we want to deal with. On the other hand the equation for the area allows us to relate l and w. Specifically, $l = A/w$. So C becomes a function of a single variable:

$$C(w) = 2w + \frac{2A}{w}.$$

The mathematical reformulation of the problem is: Find the absolute minimum of $C(w)$ for $w \in (0, \infty)$.

We apply the second derivative test to find the local extrema of C. First we calculate the first and second derivative of C:

$$\frac{dC}{dw} = 2 - \frac{2A}{w^2} \quad \text{and} \quad \frac{d^2C}{dw^2} = \frac{4A}{w^3}.$$

In particular, $dC/dx = 0$ if and only if $w^2 = A$, and $d^2C/dw^2 > 0$ if $w > 0$. The second derivative test tells us that $C(w)$ has a local minimum at $w = \sqrt{A}$. For the given value for A, this means that $w = 77.46$ meters. In addition, Theorem 3.65 tells us that $C(w)$ has its absolute minimum at this point.

The answer tells us more. As $l = A/w$ we also see that $l = \sqrt{A}$, so that the shape of the suggested corral will be a square. If we use a side length of 77.46 meters, then the fence will have a length of about 309.84 meters. $\Diamond$

Remark 21. In retrospect, you may say that we solved the problem previously in Example 3.62. Did we? Once we found the largest area rectangle with a given perimeter. Once we found the shortest perimeter for a rectangle with a given area. In either case we come up with a square as the optimal shape. You may find it obvious, or after some thought you may come to the conclusion, that these two problems are equivalent.

Example 3.73. An Inventory Management Problem

General Situation: Suppose you manage a drinking water reservoir, and you like to minimize the cost of providing water to the community. A pump draws clean water from a deep well and delivers it to the reservoir. There is a cost involved in keeping the water in the reservoir clean, and we assume that this cost is proportional to the amount of water in the reservoir. The pump has to be turned on and off manually, so that every time the reservoir is replenished, a worker has to drive up to the reservoir (which is located on a nearby mountain) and turn it on and off. There is a cost involved in this process as well. You need energy to operate the pump. Finally, let us assume that the consumers use the water at a constant rate.

We have to decide how often to send a worker to the reservoir, and how much water to pump into the reservoir each time.[16]

Specific Data and Notation: Let us make the problem more specific by introducing numbers. Let us say that the community uses the water at a rate of $a = 1,000$ cubic meters per day, and that the capacity of the tank is $M = 30,000$ cubic meters. Each trip to the reservoir costs $K = \$80.00$, and the fuel cost to pump one cubic meter of water is $c = \$0.70$. The cost of keeping the water clean is \$2.00 per 1,000 cubic meter and day.

Preliminary Decisions: Let us assume (justifiably) that you replenish the tank in regular intervals. What is the best strategy for one time period should be the best strategy for any time period. In this sense, let us denote the time between trips by T and the amount pumped each time by Q.

First of all, let us decide that, whatever other decisions we are going to make, we will wait until the reservoir is (about) empty before we refill it to the level of Q cubic meters. Having a safely stock would only add to the cost of keeping the water clean and increase the total cost for maintaining the water supply.

Mathematical Reformulation: To analyze the situation, we introduce a function $q(t)$, which is defined as the number of cubic meters of water in the reservoir at time t. In Figure 3.28 you see the graph of this function. You see three periods during which the water level drops to zero and the reservoir is refilled to contain Q cubic meters of water. As we fill up the reservoir to Q cubic meters and the community uses a cubic meters per day, the length of each period is $T = Q/a$. We assumed that refilling

[16] This kind of problem is considered in Operations Research under the heading of Inventory Theory. We address only the simplest problem of this kind.

the reservoir takes no time, thus the vertical line at time $t = 0$, $t = T$, etc, indicating the sharp rise in the amount of water in the reservoir.

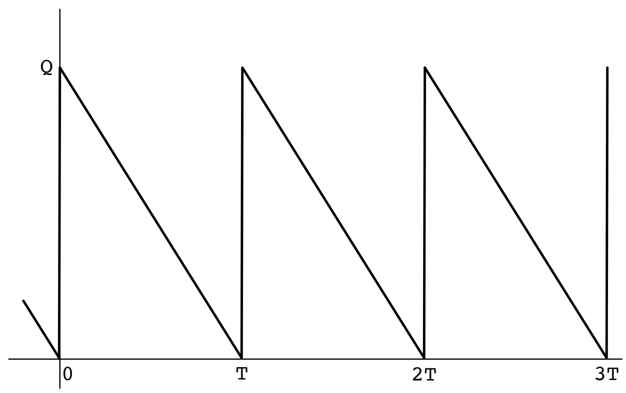

Figure 3.28: Water in Reservoir

We are now ready to calculate the cost for providing the water. First we calculate the cost for one period, i.e. from the time the tank is being replenished until it is time to replenish it again.

Storage Cost: To find the storage cost per period, we multiply the length of the period (T days), the average amount of water in the reservoir ($Q/2$ cubic meters) and the cost of keeping a cubic meter of water clean for a day ($h = 2/1000$ dollars per day):

$$\frac{ThQ}{2}.$$

Pumping Cost: We pump Q cubic meters of water into the tank, and the expense for this is cQ.

Fixed Cost per Period: Finally, we have to consider the cost for sending the worker to the reservoir, and this cost is K.

Total Cost per Period: Taken together, the expense for a single period is (in dollars)

$$e = K + cQ + \frac{ThQ}{2}.$$

Daily Cost: For a single day (or unit of time in general), and as a function of Q, the cost of providing water to the community is e/T, or

$$E(Q) = \frac{aK}{Q} + ca + \frac{Qh}{2},$$

where we used that $T = Q/a$.

Mathematical Formulation: Find the absolute minimum of $E(Q)$ for $Q \in (0, \infty)$. For the moment, we ignore the fact that reservoir is finite, and that Q should be limited. We deal with this aspect later.

We apply the second derivative test to find the local extrema for E. To do so, we calculate the first and second derivatives of E:

$$\frac{dE}{dQ} = \frac{-aK}{Q^2} + \frac{h}{2} \quad \text{and} \quad \frac{d^2E}{dQ^2} = \frac{2aK}{Q^3}.$$

The first derivative is zero if and only if

$$\frac{aK}{Q^2} = \frac{h}{2} \quad \text{and} \quad Q = \sqrt{\frac{2aK}{h}},$$

and the second derivative is positive for all positive values of Q. The second derivative test tells us, that there is a local minimum at this value of Q. Theorem 3.65 tells us, that the $E(Q)$ assumes its absolute minimum at this point.

The value of Q at which $E(Q)$ has its local and absolute minimum is called the *economical ordering quantity.* We denote it by

$$Q^* = \sqrt{\frac{2aK}{h}}.$$

For our specific values of the cost factors we find an economical ordering quantity

$$Q^* = \sqrt{\frac{2 \times 1,000 \times 80 \times 1,000}{2}} = 8,944.$$

We found the absolute minimum of $E(Q)$ for $Q \in (0,\ 30,000]$ when $Q = 8,944$ cubic meters (the value is rounded). In other words the reservoir should be replenished about every 9 days to contain $9,000$ cubic meters of water. A policy like this is almost optimal and avoids trips at night, and other inconvenient irregular schedules.

If the economical ordering quantity Q^* exceeds the capacity of the reservoir, which can happen if you increase K or decrease h substantially, then the policy will be to completely fill the reservoir each time. In this case $E(Q)$ will be decreasing on $(0, M]$. $\diamondsuit$

Remark 22. Let us contemplate what Theorem 3.65 tells us about the extrema of a function as in the previous example. Consider any function $E(Q)$ which is defined on an interval I. Assume that $E''(Q)$ exists and $E''(Q) > 0$ for all Q in the interval. The derivative of such a function is increasing and can have at most one zero and one critical point. In particular, such a function can have at most one local minimum in I. Different situations can occur.

If the function has a local minimum, then this is also the absolute minimum. This is the case if $E'(Q) = 0$ for some Q in the interior of the interval.

If $E'(Q) \neq 0$ for all interior points Q of the interval, i.e., the function does not have a local minimum, then the function is either increasing on the entire interval, or decreasing. Absolute extrema will occur at the end points of the interval as far as they belong to the interval. Where we encounter minima and maxima depends on whether the function is increasing or decreasing on the interval.

Example 3.74. Suppose you went wind surfing, and all of a sudden the wind dies down. It is absolutely calm. Sitting on your board you contemplate what to do. None of the elements of nature is going to help you, there is no wind, no current, and no waves to help you or interfere with your effort to get back to the shore. You have to paddle, and because this is a tough boring task, you want to head to the point at the shore which is closest to you. Let us suppose that your only tools are a map and something which is straight, like a piece of rope which you can make straight by stretching it.

Solution: In an attempt to solve your problem, you set up a more general, mathematically formulated problem which is supposed to give you the desired solution.

Consider the graph of a differentiable function $f(x)$ which is defined on some interval $[a, b]$ and a point (A, B) which does not lie on the graph. In Figure 3.29 you see such a situation.[17] The figure is drawn for $(A, B) =$

[17] If you use a figure as a map, then you have to make sure that the horizontal and vertical scales are the same. Otherwise, the angles are distorted.

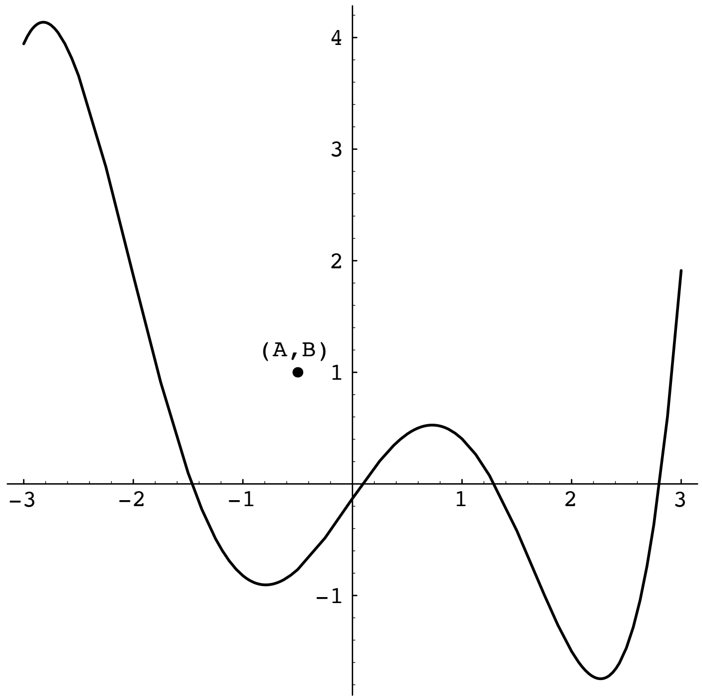

Figure 3.29: Shore Line

$(-.5, 1)$ and[18]

$$f(x) = [-.01x^7 + .7x^5 + .4x^4 - 7x^3 - x^2 + 11x - 1]/7.65.$$

If p and q are points in the plane, then we denote the distance between these two points by $D(p, q)$. Concretely, when $p = (x, y)$ and $q = (A, B)$, then, according to the theorem of Pythagoras,

$$D(p, q) = \sqrt{(x - A)^2 + (y - B)^2}.$$

As a function of the x-coordinate of the point on the graph, the distance between $(x, f(x))$ and (A, B) is

$$E(x) = D((x, f(x)), (A, B)) = \sqrt{(x - A)^2 + (f(x) - B)^2}.$$

Using the notation which we have developed so far, the problem is to find the value for x for which $E(x)$ is minimal. We like to make the following geometric observation:

[18] The somewhat unmotivated scaling factor was used so that the scale of the two axis in the graph comes out to be the same.

- If x is a critical point of the function $E(x)$, then the line which connects $(x, f(x))$ and (A, B) intersects the (tangent line to the) graph of f perpendicularly in $(x, f(x))$.

To see this assertion, differentiate $E(x)$:

$$E'(x) = \frac{2[(x-A) + (f(x)-B)f'(x)]}{2\sqrt{(x-A)^2 + (f(x)-B)^2}}.$$

So $E'(x) = 0$ if and only if

$$x - A = -(f(x) - B)f'(x).$$

As (A, B) does not lie on the graph, either $(x - A)$ or $(f(x) - B)$ has to be non-zero. So, if $E'(x) = 0$, then $f(x) - B \neq 0$. At a critical point we have that

$$f'(x) = -\frac{x-A}{f(x)-B}.$$

Let us discuss the case $f'(x) \neq 0$ first. Then $x - A \neq 0$ and the slope of the line connecting $(x, f(x))$ and (A, B) is $m = (f(x) - B)/(x - A)$. Apparently, $mf'(x) = -1$. Remember that $f'(x)$ is the slope of the tangent line to the graph of $f(x)$ at x. The relation $mf'(x) = -1$ expresses that the line connecting $(x, f(x))$ and (A, B) and the tangent line to the graph of $f(x)$ at x are perpendicular to each other.

Secondly, consider the case $f'(x) = 0$. In this case the tangent line is horizontal and $x = A$, so that the line connecting $(x, f(x))$ and (A, B) is vertical. So these two lines are still perpendicular to each other.

In either of the cases, we verified our assertion.

As a practical matter, it is not that difficult to decide the approximate location of those points on the graph where the line from this point to (A, B) intersects the shore line perpendicularly. This (hopefully) reduces your problem sufficiently, and you have only a few points on the shore line to consider as points to head for. Having only a map and a piece of string you cannot ask for much more. Try it. Take a ruler, put it on the picture, and place it so that its edge intersects the graph perpendicularly and so that (A, B) is on the edge. Two reasonable candidates $(x, f(x))$ on the shore line to head for have x-coordinate $x = .4$ and $x = -1.7$.

Let us pursue the numerical problem further, and carry out the specific calculation. We have to find solutions of the equation

$$(x - A) + (f(x) - B)f'(x) = 0,$$

or equivalently, zeros of the function

$$F(x) = (x - A) + (f(x) - B)f'(x).$$

For the given example, we have no hope of finding a simple expression for the answer. All we can do is to resort to some numerical method for finding approximate zeros. One such method is Newton's method, which we discussed previously. So let us use the values for x which we guessed based on the picture, and then improve the guesses using Newton's method. Our results are summarized in Table 3.2. In the first column we start out with $x = .38$, and improve upon this guess twice, coming up with an approximate value of $x = .3069$. The small value for $F(x)$ (in the third column) suggest that we are close to the local minimum for $E(x)$, the distance between (A, B) and the points on the graph of f. The actual distance between $(x, f(x))$ and (A, B) for the respective values of x is shown in the second column. In the last three columns you see the calculation when we start out with $x = -1.7$.

x	$E(x)$	$F(x)$	&	x	$E(x)$	$F(x)$
.38000	1.09497	+0.25241983		−1.70000	1.22918	−0.27393092
.30427	1.08643	−0.00940427		−1.67789	1.22667	−0.00731624
.30690	1.08642	−0.00002611		−1.67726	1.22667	−0.00000876

Table 3.2: Newton's Method

As it turns out, we find the relevant critical points for the function $E(x)$ which measures the distance between the point (A, B) and the points $(x, f(x))$ on the graph. These are $(.30690, .27253)$ and $(-1.67726, .65536)$, with fairly good accuracy. It is apparent that we found local minima for $E(x)$, and this may be confirmed by a calculation of $E''(x)$ for these values of x. Among the two local minima, the first one represents the point on the shore line which is closer to (A, B), as you can see from the entries in the last row of the table in column 2, resp,. column 5. Points outside the shown part of the graph don't have to be considered either (why?). This means that we should head for the point with coordinates $(.30690, .27253)$ on the shoreline. $\lozenge$

Exercise 121. Find the dimensions of a rectangle of perimeter 30 cm that has the largest area.

Exercise 122. A rectangular warehouse will have 5000 m^2 of floor space and will be separated into two rectangular rooms by an interior wall. The cost of the exterior walls is \$ 1,000.00 per linear meter and the cost of the interior wall is \$ 600.00 per linear meter. Find the dimensions of the warehouse that minimizes the construction cost.

Exercise 123. One side of a rectangular meadow is bounded by a cliff, the other three sides by straight fences. The total length of the fence is 600 meters. Determine the dimensions of the meadow so that its area is maximal.

Exercise 124. Draw a rectangle with one vertex at the origin $(0, 0)$ in the plane, one vertex on the positive x-axis, one vertex on the positive y-axis, and one vertex on the line $3x + 5y = 15$. What are the dimensions of a rectangle of this kind with maximal area?

Exercise 125. Two hallways, one 8 feet wide and one 6 feet wide, meet at a right angle. Determine the length of the longest ladder that can be carried horizontally from one hallway into the other one.

Exercise 126. A string of length 50 centimeters is to be cut into two pieces, one to form a square and one to form a circle. How should the string be cut so as to maximize the sum of the two areas? How should the string be cut so as to minimize the sum of the two areas?

Exercise 127. Inscribe a right circular cylinder into a right circular cone of height 25 cm and radius 6 cm. Find the dimensions of the cylinder if its volume is the be a maximum.

Exercise 128. A right circular cone is inscribed in a sphere of radius R. Find the dimensions of the cone if its volume is to be maximal.

Exercise 129. Find the dimensions of a right circular cone of minimal volume, so that a ball of radius 10 centimeters can be inscribed.

Exercise 130. Two ships, the Liberty and the Independence, are cruising in the waters of Hawai'i. The Liberty is travelling due West at a speed of 15 knots and the Independence due South-East at a speed of 12 knots. Their routes intersect at a point which is 40 nautical miles from the current position of the Liberty and 50 nautical miles from the current position of the Independence. How long will it take until the two ships are closest to each other, and how far from each other will they be at that time?

Exercise 131. A string of length 50 centimeters is to be cut into two pieces, one to form an equilateral triangle and one to form a circle. How should the string be cut so as to maximize (resp., minimize) the sum of the two inclosed areas?

Exercise 132. Minimize the cost of the material needed to make a round drum with a volume of 200 liter (i.e., .2 m^3) if

(a) the drum has a bottom and a top, and the same material is used for the top, bottom and sides.

(b) the drum has no top (but a bottom) and the same material is used for the bottom and sides.

(c) the drum has a bottom and a top, the same material is used for the top and bottom, and the material for the top and bottom is twice as expensive as the material for the sides.

(d) the situation is as in the previous case, but the top and the bottom are cut out of squares, and the left over material is recycled for half its value.

Exercise 133. Design a roman window with a perimeter of 4 m which admits the largest amount of light. (A roman window has the shape of a rectangle capped by a semicircle.)

Exercise 134. A rectangular banner has a red border and a white center. The width of the border at top and bottom is 15 cm, and along the sides 10 cm. The total area is 1 m^2. What should be the dimensions of the banner if the area of the white area is to be maximized?

Exercise 135. A power line is needed to connect a power station on the shore line to an island 2 km off shore. The point on the coast line closest to the island is 6 km from the power station, and, for all practical purposes, you may suppose that the shore line is straight. To lay the cable costs $40,000 per kilometer under ground and $70,000 under water. Find the minimal cost for laying the cable.

Exercise 136. Design a shaved ice container with a volume of 1 l (i.e., 1000 cm^3). The shape should be a round cone which is capped off by a round lid which has the shape of a hemisphere. To keep the ice as cold as long as possible, the surface area of the container should be minimal. What are the dimensions of the optimal cone?

3.12 Sketching Graphs

The techiques which we developed so far provide us with some valuable tools for graphing functions. Let us make a list of data which we may determine, so that we can sketch a graph rather precisely. In some cases, we will get more accurate graphs than those provided by standard software on the computer or on your graphing calculator. We gave a few examples in Section 1.4 on page 31 where we showed how technology can fail. Even if you are considering an example where technology does well, going through the following program is a good review of the material which we developed in this chapter.

Useful Information for graphing a function: For convenience, we call the function $f(x)$.

(a) Find the y-intercept of $f(x)$. Plot this point. If you consider a function on a closed interval, the you may also plot the function at its end points.

(b) Find the exceptional points[19] of $f(x)$. To find the zeros of the function we may be able to use analytical means. If they fail, we may have to use numerical means, such as Newton's method. Plot the zeros of the function.

(c) Based on the previous item, we can determine intervals on which the $f(x)$ is positive and intervals on which it is negative.

(d) Find the first derivative $f'(x)$ of $f(x)$.

(e) Repeat (b) and (c) with $f'(x)$ in place of $f(x)$. Intervals on which $f'(x)$ is positive give you intervals on which $f(x)$ is increasing, and intervals on which $f'(x)$ is negative give you intervals on which $f(x)$ is decreasing. The exceptional points of $f'(x)$ provide you with the critical points of $f(x)$. Plot the critical points (x and y value), and keep track of the intervals on which the function is increasing, resp., decreasing.

(f) Find the second derivative $f''(x)$ of $f(x)$.

(g) Repeat (b) and (c) with $f''(x)$ in place of $f(x)$. Intervals on which $f''(x)$ is positive give you intervals on which $f(x)$ is concave up, and intervals on which $f''(x)$ is negative give you intervals on which $f(x)$

[19] We defined this concept in Definition 3.18 on page 148.

is concave down. Find the inflection points of the function, i.e., the points where the concavity changes. Plot the inflection points (x and y value), and keep track of the intervals on which the function is concave up, resp., concave down.

(h) Decide at which critical points of $f(x)$ the function has a local extremum, and whether it is a minimum or a maximum.

If you now draw a graph which exhibits all of the properties which you gathered in the course of the suggested program, then your graph will look very much like the graph of $f(x)$. More importantly, the graph will have all of the essential features of the graph of $f(x)$.

Let us go through the program for a particular example.

Example 3.75. Discuss the graph of the function

$$f(x) = x^4 - 2x^3 - 3x^2 + 8x - 4 \text{ for } x \in [-3, 3].$$

Solution: To make the discussion a little easier, we note that

$$f(x) = (x-1)^2(x^2-4) = (x-1)^2(x-2)(x+2). \tag{3.6}$$

You should verify this by multiplying out the expression for $f(x)$ in (3.6).

(a): Plot the y intercept of the function and its values at the end points of the given interval: $f(-3) = 80$, $f(0) = -4$ and $f(3) = 20$.

(b): As a polynomial, the function $f(x)$ is differentiable on the given interval. The only exceptional points are its zeros. Having written $f(x)$ as in (3.6), we see right away that $f(x) = 0$ if and only if $x = -2$, $x = 1$, or $x = 2$. Plot these x-intercepts.

(c): Counting the signs of the factors of $f(x)$, we see that $f(x)$ is positive on the intervals $[-3, -2)$ and $(2, 3]$, and negative on $(-2, 1)$ and $(1, 2)$.

(d): We calculate the derivative of $f(x)$:

$$f'(x) = 2(x-1)(x^2-4) + (x-1)^2 2x = 2(x-1)(2x^2 - x - 4).$$

We based the calculation on the description of $f(x)$ in (3.6). In the first step we applied the product rule, and then we used elementary algebra.

(e): We use the quadratic formula (see (1.8)) to find the zeros of the factor $2x^2 - x - 4$ in the expression for $f'(x)$. They are $(1 \pm \sqrt{33})/4$. This allows us to factor the expression for $f'(x)$, and we find:

$$f'(x) = 4(x-1)\left(x - \frac{1}{4}[1+\sqrt{33}]\right)\left(x - \frac{1}{4}[1-\sqrt{33}]\right).$$

We conclude that:

- $f'(x)$ is negative on the interval $[-3, (1-\sqrt{33})/4)$ and $f(x)$ is decreasing on $[-3, (1-\sqrt{33})/4]$.
- $f'(x)$ is positive on the interval $((1-\sqrt{33})/4, 1)$ and $f(x)$ is increasing on $[(1-\sqrt{33})/4, 1]$.
- $f'(x)$ is negative on the interval $(1, (1+\sqrt{33})/4)$ and $f(x)$ is decreasing on $[1, (1-\sqrt{33})/4]$.
- $f'(x)$ is positive on the interval $((1+\sqrt{33})/4, 3]$ and $f(x)$ is increasing on $[(1+\sqrt{33})/4, 3]$.
- $f(x)$ has a critical point and local minimum at $(1-\sqrt{33})/4 \approx -1.19$, a critical point and local maximum at $x = 1$, and a critical point and local minimum at $(1+\sqrt{33})/4 \approx 1.69$.

The values of the function at its three critical points are approximately:

$$f(\frac{1-\sqrt{33}}{4}) \approx -12.39 \quad \& \quad f(1) = 0 \quad \& \quad f(\frac{1+\sqrt{33}}{4}) \approx -.54.$$

Plot these points.

(f): We rewrite the first derivative as $f'(x) = 4x^3 - 3x^2 - 3x + 4$, and find

$$f''(x) = 12x^2 - 12x - 6.$$

(g): We use the quadratic formula to find the zeros on $f''(x)$ and factor it:

$$f''(x) = 12\left(x - \frac{1}{2}[1+\sqrt{3}]\right)\left(x - \frac{1}{2}[1-\sqrt{3}]\right).$$

We conclude that:

- $f''(x)$ is positive on the interval $[-3, (1-\sqrt{3})/2)$ and $f(x)$ is concave up on $[-3, (1-\sqrt{3})/2]$
- $f''(x)$ is negative on the interval $((1-\sqrt{3})/2, (1+\sqrt{3})/2)$ and $f(x)$ is concave down on $[(1-\sqrt{3})/2, (1+\sqrt{3})/2]$
- $f''(x)$ is positive on the interval $((1+\sqrt{3})/2, 3]$ and $f(x)$ is concave up on $[(1+\sqrt{3})/2, 3]$
- $f(x)$ has inflection points at $x = (1-\sqrt{3})/2 \approx -.37$ and at $x = (1+\sqrt{3})/2 \approx 1.37$.

The values of the function at its inflection points is approximately:

$$f(\frac{1-\sqrt{3}}{2}) \approx -7.21 \quad \& \quad f(\frac{1-\sqrt{3}}{2}) \approx -.29.$$

Plot these points.

(h): At this point we could use the second derivative test to find at which critical points the function has local extrema, but we decided this already based on first derivative behaviour in (e).

Let us gather and organize our information. We consider the interval:

$$I_1 = [-3, -2] \qquad I_5 = \left[1, \frac{1+\sqrt{3}}{2}\right]$$
$$I_2 = \left[-2, \frac{1-\sqrt{33}}{4}\right] \qquad I_6 = \left[\frac{1+\sqrt{3}}{2}, \frac{1+\sqrt{33}}{4}\right]$$
$$I_3 = \left[\frac{1-\sqrt{33}}{4}, \frac{1-\sqrt{3}}{2}\right] \qquad I_7 = \left[\frac{1+\sqrt{33}}{4}, 2\right]$$
$$I_4 = \left[\frac{1-\sqrt{3}}{2}, 1\right] \qquad I_8 = [2, 3]\,.$$

We tabulate the which properties hold on which interval. It should be understood, that at some end points of intervals the function is zero.

Property	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
Sign	pos	neg	neg	neg	neg	neg	neg	pos
Monotonicity	dec	dec	inc	inc	dec	dec	inc	inc
Concavity	up	up	up	down	down	up	up	up

Table 3.3: Properties of the Graph

In Figure 3.30 you see the graph of the function. We have shown it on a slightly smaller interval, as the values at the endpoint a comparetively large. Showing all of the graph would show less clearly what happens near the intercept, extrema, and inflection points. The dots indicate the points which we suggests to plot.

In Figure 3.31 you see the graph of f on an even smaller interval, and parts of the graphs of f' and f''. You can use them to see that f is decreasing where f' is negative, f is concave down where f'' is negative, etc. $\Diamond$

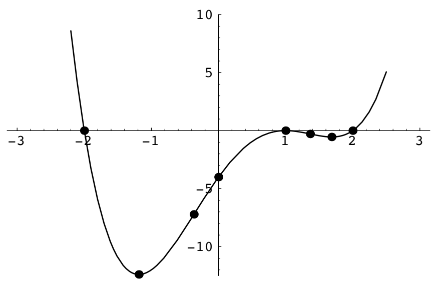

Figure 3.30: The Graph

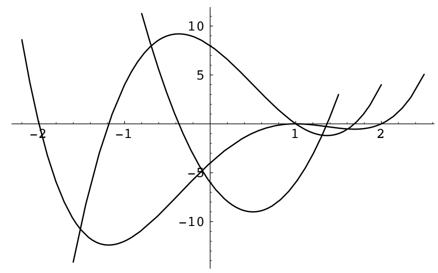

Figure 3.31: f, f', f''

Exercise 137. In analogy with the previous example, discuss the function

$$f(x) = (x-1)(x-2)(x+2) = x^3 - x^2 - 4x + 4$$

on the interval $[-3, 2.5]$. In addition, find the absolute extrema of this function.

Exercise 138. In analogy with the previous example, discuss the function

$$f(x) = x^3 - 3x + 2$$

on the interval $[-2, 2]$. In addition, find the absolute extrema of this function.

Exercise 139. In analogy with the previous example, discuss the function

$$f(x) = 2\sin x + \cos 3x$$

on the interval $[0, 2\pi]$. In addition, find the absolute extrema of this function. You may have to apply Newton's method to find zeros of f, f', and f''.

Chapter 4

Integration

We will introduce the ideas of the definite and the indefinite integral. Suppose that f is a function which is defined for all x in the closed interval $[a, b]$, and assume that f is bounded over this interval. If it exists, then the integral of f over the interval $[a, b]$ is a real number. It is denoted by

$$\int_a^b f(x)\ dx.$$

The definition is set up, so that for a non-negative function it makes sense to think of the integral as the area of the region bounded by the graph of the function, the x-axis, and the lines $x = a$ and $x = b$. The indefinite integral of a function f is the family (set) of all functions whose derivative is f. After introducing these concepts and a few examples we will state the Fundamental Theorem of Calculus. It tells us how to calculate the definite integral of a function if its indefinite integral is known.

We motivate the upcoming discussion of area, lower and upper sums, and integrability with an example.

Example 4.1. In Figure 4.1 you see the graph of a function. We would like to determine the area of the region bounded by the graph, the lines $x = 1$ and $x = 5$, and the x-axis. We denote the region by Ω.

So far, we have not even defined what the area of a region is, unless it is of a particularly nice kind, like a rectangle. Whatever concept you have in mind for the area of a region in the plane, it should have the following properties. Whenever it exists, we denote the area of a region Ω by $\text{Area}(\Omega)$.

- The area of a rectangle is the product of the lengths of its sides.

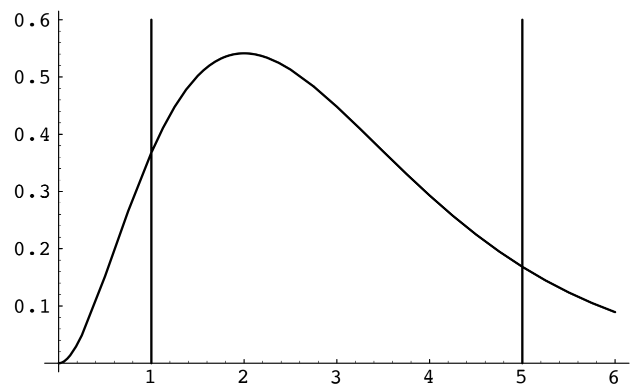

Figure 4.1: $f(x) = x^2 e^{-x}$

- Suppose that Ω_1 and Ω_2 are regions in the plane, and that the area of each of them is defined.

$$\text{If } \Omega_1 \subseteq \Omega_2, \text{ then } \operatorname{Area}(\Omega_1) \leq \operatorname{Area}(\Omega_2).$$

- Suppose that Ω_1 and Ω_2 are regions in the plane, and that the area of each of them is defined. If the regions Ω_1 and Ω_2 do not intersect, then the area of the union $\Omega_1 \cup \Omega_2$ of Ω_1 and Ω_2 is defined, and

$$\operatorname{Area}(\Omega_1 \cup \Omega_2) = \operatorname{Area}(\Omega_1) + \operatorname{Area}(\Omega_2).$$

Let us use the second principle to get a first idea about the area of our region Ω. In Figure 4.2 you see a rectangle which is contained in Ω. Its width is 4 and its height .15. This means that the area of the rectangle is .6. If Ω is to have any area, then the area should be at least .6.

In Figure 4.3 you see a rectangle which contains Ω. Its width is 4, its height .56, and its area 2.24. If Ω is to have any area, then the area should be at most 2.24.

Taking these two statements together we conclude that, if there is any way to define the area of the region Ω, then it must be between .6 and 2.24.

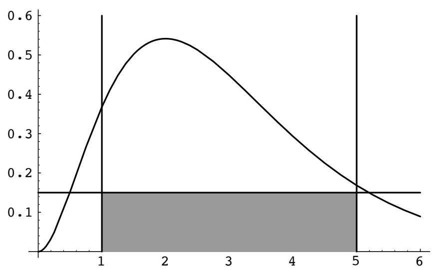

Figure 4.2: A rectangle contained in Ω

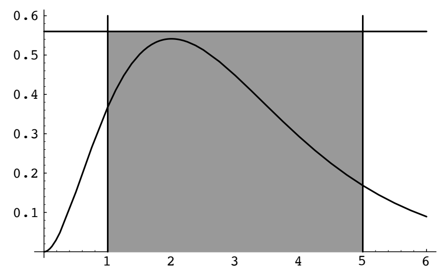

Figure 4.3: A rectangle containing Ω

The first and third principles for our concept of the area of a region can be used to derive one more principle:

- Suppose the region R in the plane is the union of a finite number of rectangles $R_1, \ldots, R_n$ and any two of them intersect at most in an edge. Then Area(R) is defined, and it is equal to the sum of the areas of the regions $R_1, \ldots, R_n$:

$$\mathrm{Area}(R) = \mathrm{Area}(R_1) + \cdots + \mathrm{Area}(R_n).$$

Using this principle, we can go one step further and determine the possible area of Ω a bit more precisely. Instead of one, we use several rectangles. In Figure 4.4 you see four rectangles, placed next to each other, such that their union is contained in Ω. The heights of the rectangles are .35, .43, .28, and .16, respectively. Each of the rectangles has width 1. The sum of the areas of the rectangles is 1.22. We conclude that the area of Ω should be at least 1.22, at least if it makes sense to talk about the area of Ω. In Figure 4.5 you see three rectangles, placed next to each other, such that their union contains Ω. The heights of the rectangles are .55, .45, and .3, respectively. Their widths are 2, 1 and 1, so that their combined area is 1.85. So, whatever idea we develop for the area of the region Ω, we expect

$$1.22 \le \mathrm{Area}(\Omega) \le 1.85.$$

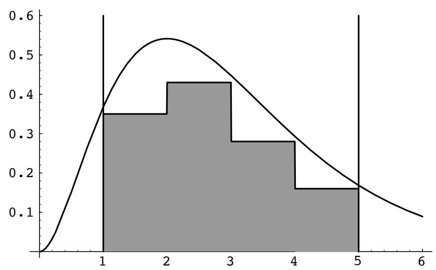

Figure 4.4: A union of rectangles contained in Ω

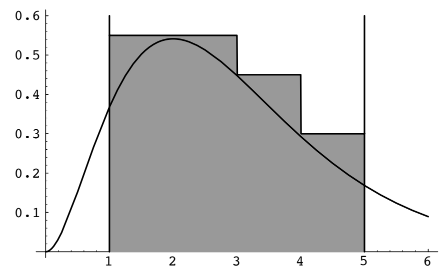

Figure 4.5: A union of rectangles containing Ω

In this second attempt we did better. We narrowed down more closely the range in which Area(Ω) will lie, if defined. We could do a lot better by using 25 rectangles (or 1,000,000 rectangles) whose union is contained, resp., contains, the region Ω. Choosing the heights of the rectangles carefully also helps.

In the following, R_l and R_u are unions of rectangles, constructed as above, so that

$$R_l \subseteq \Omega \subseteq R_u.$$

With some work, for which we do not give the details here, it is possible to show that[1]

$$\text{Area}(R_l) \leq \frac{5}{e} - \frac{37}{e^5} \leq \text{Area}(R_u). \tag{4.1}$$

This means that we could set

$$\text{Area}(\Omega) = \frac{5}{e} - \frac{37}{e^5}.$$

In addition, given any numbers L and U with $L < \frac{5}{e} - \frac{37}{e^5} < U$, we can choose the unions of the rectangles R_l and R_u, so that

$$L \leq \text{Area}(R_l) \quad \text{and} \quad \text{Area}(R_u) \leq U.$$

[1]Here e is the Euler number.

So, if we like to abide by the principles of the concept area, then there is no other choice but to set

$$\text{Area}(\Omega) = \frac{5}{e} - \frac{37}{e^5}.$$

Being left with a unique choice for the area of Ω, we make that choice, and say that

$$\text{Area}(\Omega) = \frac{5}{e} - \frac{37}{e^5}.$$

4.1 Upper and Lower Sums

The example should have prepared you for the more general definition of upper and lower sums. We say that a function f in *bounded* if there exist numbers M and m, such that $m \leq f(x) \leq M$ for all x in the domain of f, i.e., for all x for which $f(x)$ is defined.

Definition 4.2. *Consider a function f which is defined over a closed interval $[a, b]$, and suppose that f is bounded. Let us partition the interval $[a, b]$. This means that we pick numbers $x_0, x_1, \ldots, x_n$, such that*

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

Doing so, we break $[a, b]$ up into several smaller intervals $[x_0, x_1]$, $[x_1, x_2]$, $\ldots$, $[x_{n-1}, x_n]$.

Next, choose numbers m_i and M_i for $1 \leq i \leq n$, such that

$$m_i \leq f(x) \leq M_i$$

for all $x \in [x_{i-1}, x_i]$. (They exist because f is assumed to be bounded.) The lower sum S_l *and* upper sum S_u *for f with respect to the choices for the x_i, m_i and M_i are defined as*[2]

$$S_l = m_1(x_1 - x_0) + \cdots + m_n(x_n - x_{n-1})$$

and

$$S_u = M_1(x_1 - x_0) + \cdots + M_n(x_n - x_{n-1}).$$

[2]The dots in this formula represent additional terms which are supposed to be understood from context. There is a formalism, the summation notation, which makes such expressions unambiguous.

Let us make a simple observation, which you may keep in mind in the upcoming example. Although some notational effort is necessary to develop a concise proof, it is not difficult to see:

Theorem 4.3. *Let f be a function which is defined and bounded on a closed interval $[a, b]$. Let S_l be any lower sum of f and S_u any upper sum. Then*

$$S_l \leq S_u.$$

Let us repeat the statement of the theorem to emphasize its meaning. Whichever partition of the interval $[a, b]$ and whichever m_i we use in the calculation of the lower sum S_l and whichever partition of the interval and whichever M_i we use in the calculation of the upper sum S_u, the lower sum is always smaller or equal to the upper sum. Essentially, this is an immediate consequence of the fact that if we use the same partition for the interval in the calculation of the lower and upper sum, then $m_i \leq M_i$ for all i.

Example 4.4. Find upper and lower sums for $f(x) = x^2$ and $x \in [0, 1]$.

Solution: First we have to pick some partition of the interval $[0, 1]$. Let us pick

$$x_0 = 0 < x_1 = \frac{1}{5} < x_2 = \frac{2}{5} < x_3 = \frac{3}{5} < x_4 = \frac{4}{5} < x_5 = 1.$$

Secondly, we have to pick appropriate $m_1, m_2, \ldots, m_5$ and $M_1, M_2, \ldots, M_5$. We deal with the lower sum first. Let us choose

$$m_1 = 0,\ m_2 = \left(\frac{1}{5}\right)^2,\ m_3 = \left(\frac{2}{5}\right)^2,\ m_4 = \left(\frac{3}{5}\right)^2,\ \text{and } m_5 = \left(\frac{4}{5}\right)^2.$$

Apparently,

$$m_1 = 0 \leq f(x) = x^2 \text{ for all } x \in [x_0, x_1] = [0, 1/5],$$

and

$$m_2 = 1/25 \leq f(x) = x^2 \text{ for all } x \in [x_1, x_2] = [1/5, 2/5],$$

and more generally for $k = 1$, $k = 2$, ... , $k = 5$:

$$m_k = (k-1)^2/25 \leq f(x) = x^2 \text{ for all } x \in [x_{k-1}, x_k] = [(k-1)/5, k/5].$$

More geometrically speaking, each constant function m_k is smaller or equal to $f(x)$ on the interval $[x_{k-1}, x_k]$. You see this illustrated in Figure 4.6.

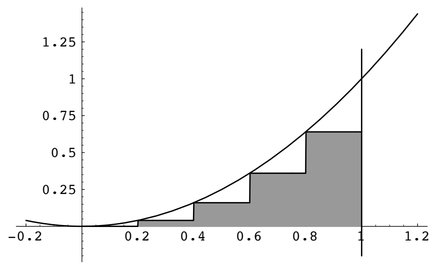

Figure 4.6: Rectangles for calculating an upper sum.

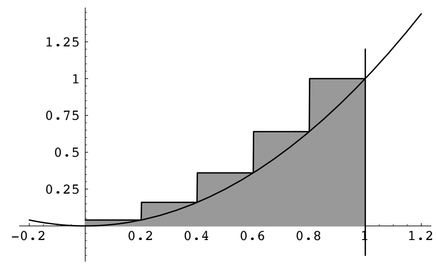

Figure 4.7: Rectangles for calculating a lower sum.

The graphs of the constant functions are horizontal lines (the top edges of the shaded rectangles), and they are below the graph.

In this example, $n = 5$ and $1 \leq k \leq 5$. We substitute the values for the x_k and the m_k into the formula for the lower sum. We also note that $x_k - x_{k-1} = 1/5$ for each k. Then

$$S_l = 0 \times \frac{1}{5} + \frac{1}{25} \times \frac{1}{5} + \frac{4}{25} \times \frac{1}{5} + \frac{9}{25} \times \frac{1}{5} + \frac{16}{25} \times \frac{1}{5} = \frac{30}{125}.$$

Next we deal with the upper sum. We choose

$$M_1 = \left(\frac{1}{5}\right)^2, \ M_2 = \left(\frac{2}{5}\right)^2, \ M_3 = \left(\frac{3}{5}\right)^2, \ M_4 = \left(\frac{4}{5}\right)^2, \text{ and } M_5 = 1.$$

Apparently,

$$f(x) = x^2 \leq M_1 = \frac{1}{25} \text{ for all } x \in [x_0, x_1] = [0, 1/5],$$

and more generally for $k = 1$, $k = 2$, ... , $k = 5$:

$$f(x) = x^2 \leq M_k = k^2/25 \text{ for all } x \in [x_{k-1}, x_k] = [(k-1)/5, k/5].$$

Speaking once more geometrically, each constant function M_k is greater or equal to $f(x)$ on the interval $[x_{k-1}, x_k]$. You see this illustrated in Figure 4.7. The graphs of the constant functions are horizontal lines (the top edges of the shaded rectangles), and they are above the graph.

Again we have $n = 5$ and $1 \le k \le 5$. We substitute the values for the x_k and the M_k into the formula for the upper sum. Once more, $x_k - x_{k-1} = 1/5$ for each k. Then

$$S_u = \frac{1}{25} \times \frac{1}{5} + \frac{4}{25} \times \frac{1}{5} + \frac{9}{25} \times \frac{1}{5} + \frac{16}{25} \times \frac{1}{5} + 1 \times \frac{1}{5} = \frac{55}{125}.$$

In conclusion, for our particular choice of the partition of the interval $[0, 1]$ (i.e., our choice of $x_1, \ldots, x_5$), and our choice for the $m_1, \ldots, m_5$ and $M_1, \ldots, M_5$ we come up with the lower and upper sums

$$S_l = \frac{30}{125} \quad \text{and} \quad S_u = \frac{55}{125}.$$

As we have been assured of in Theorem 4.3, we note that $S_l \le S_u$. $\diamondsuit$

Remark 23. Note that the lower sum in the previous example is equal to the sum of the areas of the shaded rectangles shown in Figure 4.6. The upper sum is equal to the sum of the shaded rectangles shown in Figure 4.7. In this sense, the area bounded by the graph of the function $f(x)$, the x axis, and the vertical lines $x = 0$ and $x = 1$, should be between the lower and upper sum, i.e., between .24 and .44.

Exercise 140. Repeat Example 4.4. Use again the function $f(x) = x^2$ and the interval $[0, 1]$, but partition the interval into 10 smaller intervals, each of length $1/10$. So, $x_0 = 0$, $x_1 = 1/10, \ldots$. As we did it before, use m_k as the value of the function at the left end point of the interval $[x_{k-1}, x_k]$. So $m_k = f(x_{k-1})$. Also in accordance with the example, use $M_k = f(x_k)$.

Exercise 141. Repeat Example 4.4. Use again the function $f(x) = x^2$ and the interval $[0, 1]$, but partition the interval into n smaller intervals, each of length $1/n$. Here n is any natural number. (In the example we used $n = 5$, in the previous exercise $n = 10$.) Without proof, use

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Exercise 142. Find lower and upper sums for the function $f(x) = \sin x$ over the interval $[0, \pi]$. Partition the interval into 6 smaller intervals, each of length $\pi/6$. Choose m_k as the minimum of $f(x)$ in the interval $[x_{k-1}, x_k]$ and M_k as the maximum.

Let us go through one more example. Here we illustrate what happens if the function is not non-negative on the interval under consideration.

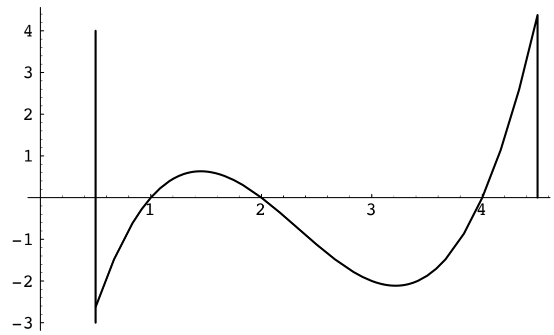

Figure 4.8: $f(x) = x^3 - 7x^2 + 14x - 8$

Example 4.5. Find upper and lower sums for the function

$$f(x) = x^3 - 7x^2 + 14x - 8$$

for $x \in [.5, 4.5]$. You see the graph of the function in Figure 4.8.

For the purpose of calculating an upper sum, we partitioned the interval $[.5, 4.5]$ using the intermediate points $x_0 = .5$, $x_1 = 1.1$, $x_2 = 2.4$, $x_3 = 3.8$, and $x_4 = 4.5$. As numbers M_i (so that $M_i \geq f(x)$ for $x \in [x_{i-1}, x_i]$) we chose $M_1 = .3$, $M_2 = .7$, $M_3 = -.9$, and $M_4 = 4.4$. With these choices, the upper sum is

$$\begin{aligned} S_u &= .3(1.1 - .5) + .7(2.4 - 1.1) + (-.9)(3.8 - 2.4) + 4.4(4.5 - 3.8) \\ &= 2.91. \end{aligned}$$

In Figure 4.9 you see four rectangles. Their areas are combined to calculate the upper sum. The areas of the ones above the x-axis are added, the ones below the axis are subtracted, in accordance with the sign of the M_i.

In the calculation of the lower sum we partitioned $[.5, 4.5]$ using $x_0 = .5$, $x_1 = .8$, $x_2 = 2.3$, $x_3 = 4.2$, and $x_4 = 4.5$. As numbers m_i (so that $m_i \leq f(x)$ for $x \in [x_{i-1}, x_i]$) we chose $m_1 = -2.7$, $m_2 = -.8$, $m_3 = -2.2$,

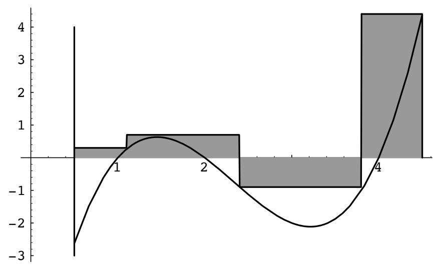

Figure 4.9: Rectangles for calculating an upper sum.

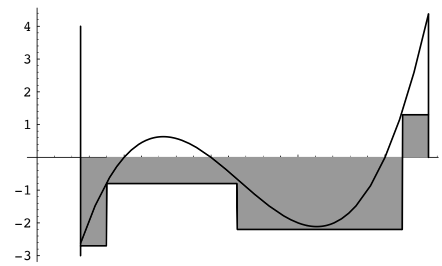

Figure 4.10: Rectangles for calculating a lower sum.

and $m_4 = 1.3$. With these choices we calculate a lower sum of

$$\begin{aligned} S_l &= -2.7(.8 - .5) + (-.8)(2.3 - .8) + (-2.2)(4.2 - 2.3) + 1.3(4.5 - 4.2) \\ &= -5.8. \end{aligned}$$

In Figure 4.10 you see four rectangles. Their areas are combined to calculate the lower sum. The areas of the ones above the x-axis are added, the ones below the axis are subtracted, in accordance with the sign of the m_i.

In summary, you see that we still combine areas of rectangles in the calculation of the upper and lower sum, only that, depending on the sign of the M_i or m_i, these rectangles are either above or below the x-axis, and depending on this, their areas are either added or subtracted.

Again you see that $S_l < S_u$, as it has to be by Theorem 4.3. The other hand, the difference between the upper and lower sum is considerable, but it can be made smaller by using finer partitions of the interval and tighter choices for the M_i and m_i. $\diamondsuit$

Exercise 143. Repeat Example 4.4 with the function $f(x) = x^2 - 1$ and the interval $[0, 2]$. Partition the interval into 8 smaller intervals, each of length 1/4. Choose m_k as the minimum of $f(x)$ in the interval $[x_{k-1}, x_k]$ and M_k as the maximum.

4.2 Integrability and Areas

As we discussed in the previous section, whatever choices we make in the calculation of lower and upper sums S_l and S_u, we always have that

$$S_l \leq S_u.$$

A crucial additional fact is stated in the next result.

Theorem 4.6. *Let f be a function which is defined and bounded on a closed interval $[a, b]$. There exists a real number Y, such that*

$$S_l \leq Y \leq S_u$$

for all lower sums S_l and upper sums S_u of f.

Idea of Proof. To deduce the theorem from the completeness of the real numbers, one observes that the set of all lower sums of f has a least upper bound. Call it Y_l. The set of all upper sums of f has a greatest lower bound. Call it Y_u. Apparently, $Y_l \leq Y_u$. Then Y is any number such that $Y_l \leq Y \leq Y_u$. □

We are now prepared to define the concept of integrability of a function.

Definition 4.7. *Let f be a function which is defined and bounded on a closed interval $[a, b]$. If there is exactly one number Y, such that*

$$S_l \leq Y \leq S_u$$

for all lower sums S_l and all upper sums S_u of f, then we say that f is integrable over the interval $[a, b]$. In this case, the number Y is called the integral of f for x between a and b. It is also denoted by

$$\int_a^b f(x)\ dx.$$

Remark 24. For completeness sake and later use, let us explain what happens when a function is not integrable. In this case there are at least two different numbers, and with this an entire interval, between all upper and lower sums. So, a function over a closed interval $[a, b]$ is not integrable if and only if the exists a positive number D such that $S_u - S_l \geq D$ for any lower sum S_u and any upper sum S_u.

On the other hand, a function is integrable if for every positive number D there is an upper sum S_u and a lower sum S_l such that $S_u - S_l < D$.

Let us illustrate the definition with an

Example 4.8. Show that $f(x) = x^2$ is integrable over the interval $[0, 1]$, and find

$$\int_0^1 x^2 \, dx.$$

Solution: In Exercise 141 you were asked to partition $[0, 1]$ into n intervals of equal length, use $m_k = f(x_{k-1})$ and $M_k = f(x_k)$, and calculate the corresponding lower and upper sums. You will come up with the solution:

$$S_l = \frac{1}{3} - \frac{1}{2n} + \frac{1}{6n^2} \quad \& \quad S_u = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}.$$

E.g., if we set $n = 1,000,000$, then

$$S_l = .3333328333335 \quad \& \quad S_u = .3333338333335.$$

Actually, using the expressions for S_u and S_l you see that $S_u - S_l = 1/n$. After some contemplation, it should be clear that $Y = 1/3$ is the only real number, so that $S_l \leq Y \leq S_u$ for all natural numbers n. According to the definition this means, that $f(x) = x^2$ is integrable over the interval $[0, 1]$ and that

$$\int_0^1 x^2 \, dx = \frac{1}{3}. \quad \diamondsuit$$

Exercise 144. Suppose that the function $f(x) = \sin x$ is integrable over the interval $[0, \pi]$. Use your solution from Exercise 142 to find numbers A and B, so that

$$A \leq \int_0^{\pi} \sin x \, dx \leq B.$$

In general it is important to explore which functions are integrable. We will explore this question soon. For the moment we content ourselves with a preliminary result. To state it, we need a definition.

Definition 4.9. *Suppose $f(x)$ is a function. We say that $f(x)$ is* non-decreasing *if $f(x_1) \leq f(x_2)$ whenever x_1 and x_2 are in the domain of $f(x)$ and $x_1 \leq x_2$. We say that $f(x)$ is* non-increasing *if $f(x_1) \geq f(x_2)$ whenever $x_1 \leq x_2$.*

Proposition 4.10. *Let $[a, b]$ be a closed interval and let f be defined and non-increasing or non-decreasing on $[a, b]$. Then f is integrable on $[a, b]$. In particular, monotonic (increasing or decreasing) functions are integrable.*

Proof. Let us assume that the function f is non-decreasing on the interval. Take any partition of the interval:

$$a = x_0 < x_1 < \cdots < x_n = b.$$

For $i = 1, \ldots, n$ we set

$$m_i = f(x_{i-1}) \qquad \& \qquad M_i = f(x_i).$$

Then, because f is non-decreasing,

$$m_i \leq f(x) \leq M_i \quad \text{for all } x \in [x_{i-1}, x_i].$$

We use the m_i and M_i to compute upper and lower sums. Let Δ be the largest value of the $x_i - x_{i-1}$. Then

$$\begin{aligned}
S_u - S_l &= [M_1(x_1 - x_0) + \cdots + M_n(x_n - x_{n-1})] \\
&\qquad -[m_1(x_1 - x_0) + \cdots + m_n(x_n - x_{n-1})] \\
&= (M_1 - m_1)(x_1 - x_0) + \cdots + (M_n - m_n)(x_n - x_{n-1}) \\
&\leq [(M_1 - m_1) + (M_2 - m_2) + \cdots + (M_n - m_n)]\,\Delta \\
&= (M_n - m_1)\Delta \\
&= [f(b) - f(a)]\Delta
\end{aligned}$$

The inequality in the computation follows from the choice of Δ. The second to last equality follows because $M_1 = m_2$, and $M_{i-1} = m_i$ more generally for all $i = 2, \ldots, n$. So, as you can see by writing out the summation in the third to last line explicitly, many terms in this sum cancel, and the only ones which remain are the ones in the next line of the calculation. Given any positive number D, we can make the partition fine enough so that $[f(b) - f(a)]\Delta < D$. Specifically, we make the partition fine enough so that the length Δ of its widest interval is less than $D/[f(b) - f(a)]$. According to our Remark 24 this means that f is integrable over the interval, as we claimed.

The proof for non-increasing functions is similar. We encourage the reader to implement the necessary modifications.

Let us explain the proof in a geometric fashion, and illustrating its steps with a concrete example, $f(x) = x^2$ over the interval $[0, 1]$. As before,

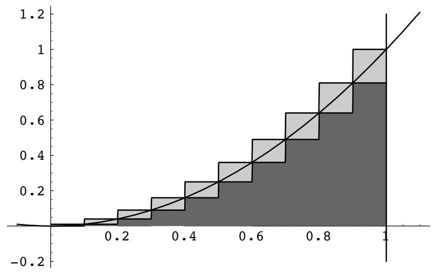

Figure 4.11: Rectangles for calculating a lower and an upper sum.

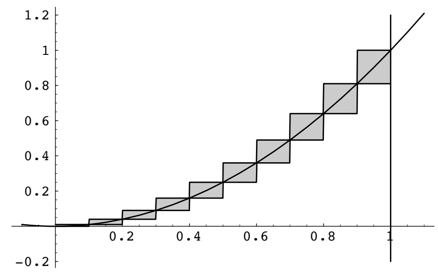

Figure 4.12: Rectangles for calculating the difference between an upper and a lower sum.

we partition the interval and choose the m_i and M_i as the value of the function at the left, resp. right, end point of the corresponding interval of the partition. The summands in the expression for the lower and upper sum represent areas of rectangles. In Figure 4.11 you see both of them. The lower sum is the sum of the areas of the darkly shaded rectangles. The upper sum is the sum of the areas of the lightly and darkly shaded rectangles. The difference between the upper and the lower sum is the sum of the lightly shaded rectangles shown in Figure 4.12. We can combine these areas by sliding the rectangles sideways so that they form one column. Its height will be $f(b) - f(a)$. Its width may vary, but in the widest place it is no wider than Δ, the width of the largest interval in the partition of $[a, b]$. That means, the difference between the upper and the lower sum is at most $[f(b) - f(a)]\Delta$. As above, we conclude that the function is integrable. □

Areas of Regions under a Graph

Let us return to our original quest. We wanted to find the area under a graph. Based on the principles which we formulated for the idea of the area of any region in the plane, we come up with the following interpretation.

Definition 4.11. *Let f be a function which is defined and non-negative on a closed interval $[a, b]$. Let Ω be the region bounded by the graph of f, the x-axis, and the lines $x = a$ and $x = b$. If f is integrable over this interval,* *then we say that the region Ω has an area and*

$$\text{Area}(\Omega) = \int_a^b f(x)\ dx.$$

Let us explain why this definition is sensible. As in the previous examples, we constructed rectangles whose union $\mathcal{R}_+$ contains the region Ω, and rectangles whose union $\mathcal{R}_-$ is contained in the region Ω. So

$$\mathcal{R}_- \subseteq \Omega \subseteq \mathcal{R}_+.$$

For the union of rectangles intersecting only in edges we defined the area. It was the sum of the areas of the individual rectangles. The collections of rectangles suggest a partition of $[a, b]$ and choices for the M_i and m_i for which

$$S_l = \text{Area}(\mathcal{R}_-) \quad \& \quad S_u = \text{Area}(\mathcal{R}_+). \tag{4.2}$$

Also, given any partition of $[a, b]$ and choices for the m_i and M_i, we can construct rectangles whose union is contained in, resp. contains, the region Ω. With the obvious choices for these rectangles we again have (4.2). That means, if there is any number $\text{Area}(\Omega)$ which qualifies to be called the area of Ω, then

$$S_l \le \text{Area}(\Omega) \le S_u.$$

These inequalities have to hold for any partition of the interval $[a, b]$ and any choices of m_i and M_i which are appropriate for calculating lower and upper sums. As $f(x)$ is assumed to be integrable, there is only one such number, and we denoted it by $\int_a^b f(x)\ dx$. I.e., There is only one possible choice for the area of Ω and that is to set

$$\text{Area}(\Omega) = \int_a^b f(x)\ dx.$$

The definition just says, that we make this one and only one possible choice.

Example 4.12. Find the area of the region Ω bounded by the graph of the function $f(x) = x^2$, the x-axis, and the lines $x = 0$ and $x = 1$.

Solution: As we have seen, $f(x)$ is integrable over the interval $[0, 1]$. So, by our definition:

$$\text{Area}(\Omega) = \int_0^1 x^2\ dx = \frac{1}{3}. \quad \diamondsuit$$

Let us give some even easier examples of integrable functions.

Example 4.13. Suppose $c \geq 0$ and $f(x) = c$ is the constant function. Then, for any closed interval $[a, b]$,

$$\int_a^b f(x)\ dx = (b-a)c.$$

Solution: This calculation is apparent from our interpretation of the integral as an area. The area enclosed by the graph, the x-axis, and the lines $x = a$ and $x = b$ is a rectangle of height c and width $b - a$. Its area is $(b - a)c$. According to the definition of the area under a graph:

$$\int_a^b f(x)\ dx = (b-a)c.$$

Strictly speaking we should have first checked that $f(x)$ is integrable over the interval. To see this, we choose $m_i = M_i = c$ in the calculation of a lower and an upper sum, whatever partition of the interval you use. With this choice, $S_u = S_l = (b - a)c$. So $Y = (b - a)c$ is the only number so that

$$S_l \leq Y \leq S_u$$

for all lower sums S_l and all upper sums S_u. This means that $f(x)$ is integrable over the interval $[a, b]$. It also implies that $\int_a^b f(x)\ dx = (b - a)c$. $\diamondsuit$

4.3 Some elementary observations

In spite of our success calculating some integrals using upper and lower sums and the definition, this is certainly not the way to go in general. To integrate "well behaved" functions we want a theory which allows us to calculate integrals more easily. We have to develop a few basic tools. These are fairly straight forward consequences of the definition of the integral.

Proposition 4.14. *If the function f is defined at a, then*

$$\int_a^a f(x)\ dx = 0 \tag{4.3}$$

In this case, the interval $[a, a]$ consists of a single point. There is only one possible partition of the interval, $a = x_0 = b$. We may still attempt to write down the formal expressions for the upper and lower sum, but they won't have any summands. In this sense the sums are zero. This is consistent with the idea of the area under the graph, which indicates the same conclusion.

Proposition 4.15. *Let $[a, b]$ be a closed interval, c a point between a and b, and f a function which is defined on the interval. Then*

$$\int_a^c f(x)\ dx + \int_c^b f(x)\ dx = \int_a^b f(x)\ dx. \tag{4.4}$$

Implicitly in the formulation of the proposition is the statement that f is integrable over $[a, b]$ if and only if it is integrable over the intervals $[a, c]$ and $[c, b]$. If one of the sides of Equation (4.4) exists, then so does the other one.

Idea of Proof. The proof is not difficult. In the calculation of the lower and upper sums over the interval $[a, b]$ one assumes (without loss of generality) that c is one of the points of the partition. Then the lower and upper sums over $[a, b]$ break up naturally into two summands, the lower and upper sums over $[a, c]$ and the one over $[c, b]$. This leads to the desired result. $\square$

Definition 4.16. *Let f be defined and integrable on the interval $[a, b]$. Then*

$$\int_a^b f(x)\ dx = -\int_b^a f(x)\ dx.$$

This definition is convenient and consistent with what we have said so far about the integral. The approach to integrals via lower and upper sums could also be generalized to include integrals $\int_a^b$ where $b < a$, leading to exactly this formula.

Using the definition of the integral it is not difficult to show:

Proposition 4.17. *Let $[a, b]$ be a closed interval and c a scalar. Suppose that f and g are integrable over the interval. Then $f+g$ and cf are integrable over $[a, b]$ and*

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

and

$$\int_a^b cf(x)\ dx = c\int_a^b f(x)\ dx.$$

We mention a few useful estimates for integrals.

Proposition 4.18. *If f is integrable over $[a,b]$, and $f(x) \geq 0$ for all $x \in [a,b]$, then*

$$\int_a^b f(x)\ dx \geq 0.$$

Proof. The proof is elementary. Just observe that under the assumptions in the proposition 0 is a lower sum, and the integral is greater or equal to any lower sum. □

Corollary 4.19. *If h and g are integrable over $[a,b]$, and $g(x) \geq h(x)$ for all $x \in [a,b]$, then*

$$\int_a^b g(x)\ dx \geq \int_a^b h(x)\ dx.$$

Proof. Set $f = g - h$. Then $f(x) \geq 0$ for all $x \in [a,b]$. The previous two propositions imply that

$$\int_a^b g(x)\ dx - \int_a^b h(x)\ dx = \int_a^b (g(x) - h(x))\ dx = \int_a^b f(x)\ dx \geq 0.$$

The claim of the proposition is an immediate consequence. □

Proposition 4.20. *Let $[a,b]$ be a closed interval and f integrable over $[a,b]$. Then the absolute value of f is integrable over $[a,b]$, and*

$$\left|\int_a^b f(x)\ dx\right| \leq \int_a^b |f(x)|\ dx. \tag{4.5}$$

The proof of this proposition is elementary, though a bit tricky.

Areas and Integrals

Let us return to the relation between areas and integrals. For a non-negative integrable function $f(x)$ over an interval $[a,b]$ we established the following relation. If Ω is the area bounded by the graph of $f(x)$, the x-axis, and the lines $x = a$ and $x = b$, then

$$\text{Area}(\Omega) = \int_a^b f(x)\ dx.$$

The question is, what happens if $f(x)$ is not non-negative?

So, let f be a function which is defined and bounded on a closed interval $[a, b]$. Let Ω be the set of points which lie between the graph of $f(x)$ and x-axis for $a \leq x \leq b$. We decompose Ω into the union of two sets. Let Ω^+ consist of those points (x, y) in the plane for which $a \leq x \leq b$ and $0 \leq y \leq f(x)$, and Ω^- of those points for which $a \leq x \leq b$ and $f(x) \leq y \leq 0$. Then Ω is the union of the sets Ω^+ and Ω^-. We decompose the area between the x-axis and the graph into the part Ω^+ above the x-axis and the part Ω^- below it. Making use of this notation, we have:

Proposition 4.21. *If f is integrable, then the areas of the regions Ω^+ and Ω^- are defined*[3] *and*

$$\int_a^b f(x)\ dx = \text{Area}(\Omega^+) - \text{Area}(\Omega^-). \tag{4.6}$$

Idea of Proof. We define two functions:

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ 0 & \text{if } f(x) \leq 0 \end{cases} \quad \text{and} \quad f^-(x) = \begin{cases} f(x) & \text{if } f(x) \leq 0 \\ 0 & \text{if } f(x) \geq 0 \end{cases}$$

One shows that the integrability of $f(x)$ implies the integrability of $f^+(x)$ and $f^-(x)$. The additivity of the integral then implies that

$$\int_a^b f(x)\ dx = \int_a^b f^+(x)\ dx + \int_a^b f^-(x)\ dx. \tag{4.7}$$

The function $f^+(x)$ is non-negative, and Ω^+ is bounded by the graph of $f^+(x)$, the x-axis, and the lines $x = a$ and $x = b$. According to Definition 4.11 we have

$$\text{Area}(\Omega^+) = \int_a^b f^+(x)\ dx. \tag{4.8}$$

Let $-\Omega^-$ be the area obtained by flipping Ω^- up, i.e., we take its mirror image along the x-axis. This process does not change areas, so $\text{Area}(\Omega^-) = \text{Area}(-\Omega^-)$. The function $-f^-(x)$ is non-negative, and $-\Omega^-$ is bounded by the graph of $-f^-(x)$, the x-axis, and the lines $x = a$ and $x = b$. According to Definition 4.11 and our elementary properties of the integral we have

$$\text{Area}(\Omega^-) = \text{Area}(-\Omega^-) = \int_a^b -f^-(x)\ dx = -\int_a^b f^-(x)\ dx. \tag{4.9}$$

[3] If you want to be really formal, then you will have to flip the region Ω^- to lie above the x-axis. Only then we can address the question of it having an area in the sense of the previous discussion.

Our claim follows now by substituting the results in (4.8) and (4.9) into (4.7). □

Example 4.22. Show

$$\int_{-\pi/2}^{\pi/2} \sin x \; dx = 0.$$

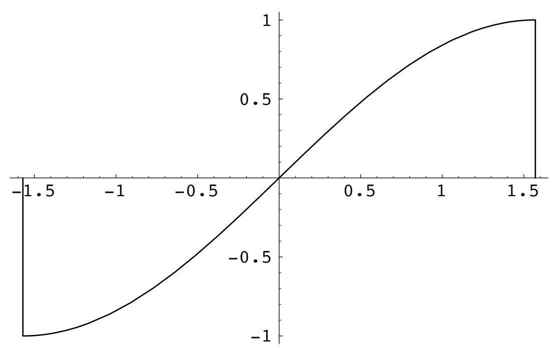

Figure 4.13: Cancelling Regions

Solution: Remember that $\sin x$ is increasing on the interval $[-\pi/2, \pi/2]$, so $\sin x$ is integrable on this interval. For reasons of symmetry, the area Ω^+ above x-axis and below the graph and Ω^- above the graph and below the x-axis have the same area. You can see these areas in Figure 4.13. That means that

$$\int_{-\pi/2}^{\pi/2} \sin x \; dx = \text{Area}(\Omega^+) - \text{Area}(\Omega^-) = 0. \quad \diamondsuit$$

Exercise 145. Show that $f(x) = \cos x$ is integrable over the interval $[0, \pi]$, and find

$$\int_0^{\pi} \cos x \; dx.$$

4.4 Integrable Functions

In general, it is important to explore which functions are integrable. As a first step we showed in Proposition 4.10 that non-increasing and non-decreasing functions over closed intervals are integrable. One may combine this result with Proposition 4.15 to show

Theorem 4.23. *Let f be defined on the interval $[a, b]$. Suppose that we can partition the interval into a finite number of intervals such that f is non-increasing or non-decreasing on each of them*[4]*. Then f is integrable on $[a, b]$. In particular, if the function is monotonic (increasing or decreasing) on each of the smaller intervals, the f is integrable on $[a, b]$.*

Remark 25. There are functions which are not integrable over any interval of the form $[a, b]$ with $a < b$.

Remark 26. Here we only discuss integrability of function over closed finite intervals, i.e., intervals of the form $[a, b]$. The discussion of integrability of functions over intervals which are not of this form, e.g., half-open intervals like $[a, b)$ or unbounded closed intervals like $[a, \infty)$, requires additional ideas and techniques which we are not in the position to discuss here.

You may now ask which functions are covered by the theorem. Probably every function which is defined on a closed interval $[a, b]$ and which you have ever seen. Nevertheless, let us make a list of function for which you may have few difficulties verifying the assumptions of the theorem and with this their integrability. In each case we assume that the function is defined at each point of the interval $[a, b]$.

- Polynomials are integrable.
- Rational functions (i.e., functions of the form $p(x)/q(x)$ where $p(x)$ and $q(x)$ are polynomials) are integrable[5].
- The trigonometric functions (sin, cos, tan, cot, sec, and csc) are integrable.

[4] In a more conventional introductory calculus course you will learn at this point that continuous functions, or more generally piecewise continuous functions, are integrable. All the functions we care about right now are covered by either theorem.

[5] The assumption that such a function is defined on the interval $[a, b]$ is equivalent to the assumption that $q(x)$ is nowhere zero in $[a, b]$.

- $f(x) = x^\alpha$ is integrable[6].
- Just making sure that the resulting functions are defined everywhere on $[a, b]$, the functions just mentioned may be added, subtracted, multiplied, divided, and composed, and one still ends up with integrable functions.

4.5 Anti-derivatives

We define the idea of an anti-derivative of a function. Having an anti-derivative of a function will (typically) make it easy to integrate it over a closed interval. We discuss and apply this idea for functions which are defined on intervals, although part of the discussion could be generalized to functions which are defined on unions of intervals.

Definition 4.24. *Let f and F be functions which are defined on the same interval I. We call F an* anti-derivative *of f if*

$$F'(x) = f(x) \quad \textit{for all } x \in I.$$

Remember that any anti-derivatives F_1 and F_2 of a function f on an interval I differ only by a constant (see Corollary 3.6). In other words, there exists a constant c, such that

$$F_1(x) = F_2(x) + c \quad \text{for all } x \in I.$$

Definition 4.25. *Let f be a function which is defined on an interval I, and suppose that f has an anti-derivative. The set of all anti-derivatives of f is called the* indefinite integral *of f. It is denoted by*

$$\int f(x)\, dx.$$

Given a function f and an anti-derivative F of it, we typically write

$$\int f(x)\, dx = F(x) + c. \tag{4.10}$$

[6]Whether $f(x) = x^\alpha$ is defined, and with this integrable, on an interval $[a, b]$ depends on α and the interval. For any real number α it suffices to assume that $a > 0$. For any real $\alpha \geq 0$, it suffices to assume $a \geq 0$. For rational numbers $\alpha = p/q$, where p and q are integers and q is odd, it suffices to assume $0 \notin [a, b]$. For non-negative integers α no assumption needs to be made on a and b.

In this expression c stands for an arbitrary constant. Different values for c result in different functions. Allowing all real numbers as possible values for c, we understand the the right hand side of (4.10) as a set of functions. The constant c in the expression is referred to as *integration constant.*

Example 4.26. Occasionally it is apparent what the indefinite integral of a function is. Given a function $f(x)$ we might know a function $F(x)$, such that $F'(x) = f(x)$. Then we can write down the indefinite integral of f in the form $F(x) + c$. Here are some examples. To check them, you may want to consult the derivatives collected in Table 2.6 on page 136.

$$\begin{aligned}
\int 1 \; dx &= x + c & \int x \; dx &= \frac{1}{2}x^2 + c \\
\int \sqrt{x} \; dx &= \frac{2}{3}x^{3/2} + c & \int x^n \; dx &= \frac{1}{n+1}x^{n+1} \; (n \neq -1) \\
\int \sin x \; dx &= -\cos x + c & \int \sec^2 x \; dx &= \tan x + c \\
\int \cos x \; dx &= \sin x + c & \int \sec x \tan x \; dx &= \sec x + c \\
\int \frac{dx}{1+x^2} &= \arctan x + c & \int \frac{dx}{\sqrt{1-x^2}} &= \arcsin x + c
\end{aligned}$$

Using the linearity of the differentiation (see the differentiation rules in (2.37)), it is easy to produce more examples. E.g.

$$\int 5x^2 - 2\cos x \; dx = \frac{5}{3}x^3 - 2\sin x + c.$$

Occasionally, an additional idea is required before we can see the anti-derivative.

Example 4.27. Find

$$\int \cos^2 x \; dx.$$

Solution: We use the trigonometric identity $\cos^2 x = (1 + \cos 2x)/2$. Then we see:

$$\int \cos^2 x \; dx = \frac{1}{2} \int (1 + \cos(2x)) \; dx = \frac{1}{2}\left[x + \frac{1}{2}\sin(2x)\right] + c. \quad \Diamond$$

Soon we will consider additional ideas for finding anti-derivatives for some standard functions.

Exercise 146. Find the following indefinite integrals:

(a) $\displaystyle\int 3\ dx$

(b) $\displaystyle\int (x+4)\ dx$

(c) $\displaystyle\int (x^2-5)\ dx$

(d) $\displaystyle\int \cos 2x\ dx$

(e) $\displaystyle\int (3+x)^3\ dx$

(f) $\displaystyle\int (3+2x)^5\ dx$

(g) $\displaystyle\int \frac{1}{x^3}\ dx$

(h) $\displaystyle\int \csc^2 x\ dx$

(i) $\displaystyle\int (1+\tan^2 x)\ dx$

(j) $\displaystyle\int \csc x \cot x\ dx$

(k) $\displaystyle\int \sin^2 x\ dx$

(l) $\displaystyle\int \sec^2(3x)\ dx$

(m) $\displaystyle\int e^{x/3}\ dx$

(n) $\displaystyle\int \frac{2x}{x^2+1}\ dx$

(o) $\displaystyle\int (4-3x)^5\ dx$

(p) $\displaystyle\int \cos(4-3x)\ dx$

(q) $\displaystyle\int \frac{2x}{(x^2+3)^2}\ dx$

(r) $\displaystyle\int x\sec^2(x^2+5)\ dx$

4.6 The Fundamental Theorem of Calculus

Right now we want a fairly general, easy to state condition which tells us when a function has an anti-derivative. For this purpose we make the following definition.

Definition 4.28. *Let f be a function which is defined on an interval I. We call f strongly continuous*[7] *if there exists a number A such that*

$$|f(x_1)-f(x_2)| \leq A|x_1-x_2|$$

for all $x_1,\ x_2 \in I$.

Geometrically speaking, a function is strongly continuous if there exists a number A such that the slope of the secant line through any two points on its graph is in between $-A$ and A. We have seen many functions like this. Polynomials over finite intervals are examples, as well as rational functions and trigonometric functions over closed intervals. More generally we have

Proposition 4.29. *If $f(x)$ is differentiable on the interval $[a,b]$, then $f(x)$ is strongly continuous on this interval.*

[7]More technically speaking, such a function f is called an L^1 function.

The proof of this proposition is not elementary, but the proposition provides us with a large class of strongly continuous functions.

Theorem 4.30. *Let f be a strongly continuous function which is defined on an interval $[a, b]$. Then f is integrable over $[a, b]$, i.e., the integral*

$$\int_a^b f(x)\ dx$$

exists.

Idea of Proof. Let A be as in Definition 4.28, and let $a = x_0 < x_1 < \cdots < x_n = b$ be a partition of the the interval $[a, b]$. Let Δ be the largest value among the $x_j - x_{j-1}$ for $1 \leq j \leq n$. Using the same notation as in the definition of upper and lower sums, we may choose m_i and M_i such that $M_i - m_i \leq A(x_i - x_{i-1})$. We calculate for the upper sum S_u and the lower sum S_l based on these choices that

$$\begin{aligned}
S_u - S_l &= (M_1 - m_1)(x_1 - x_0) + \cdots + (M_n - m_n)(x_n - x_{n-1}) \\
&\leq A(x_1 - x_0)(x_1 - x_0) + \cdots + A(x_n - x_{n-1})(x_n - x_{n-1}) \\
&\leq A\Delta\left[(x_1 - x_0) + \cdots + (x_n - x_{n-1})\right] \\
&= A\Delta(b - a).
\end{aligned}$$

Non-integrability would mean that there exists a positive number B such that $S_u - S_l \geq B$ for all partitions and all choices of M_i and m_i. Choosing the partition such that $\Delta < B/A(b-a)$ we see that $S_u - S_l < B$. This means that f is not non-integrable, or, in other words, that f is integrable. $\square$

The motivation for providing this second class of integrable functions is the following theorem.

Theorem 4.31. *Strongly continuous functions, defined over intervals, have anti-derivatives. More specifically, suppose that a function f is defined and strongly continuous over the interval I. Let $a \in I$. Then*

$$f(x) = \frac{d}{dx} \int_a^x f(t)\ dt$$

for all $x \in I$.

The major tool for calculating integrals, and the grand conclusion of our discussion of anti-derivatives is the

Theorem 4.32 (Fundamental Theorem of Calculus). *Suppose that f is a strongly continuous function over a closed interval $[a, b]$ and that F is an anti-derivative of f. Then*

$$\int_a^b f(x)\ dx = F(b) - F(a).$$

We will prove and generalize both theorems in subsections towards the end of this section, but first we want to apply the Fundamental Theorem in some calculations.

Example 4.33. Find the integral

$$\int_{-2}^{3} (x^2 - 2x + 5)\ dx.$$

Solution: The function $F(x) = x^3/3 - x^2 + 5x$ is an anti-derivative of $f(x) = x^2 - 2x + 5$. Observe also that f is strongly continuous over any finite interval. So we may apply the Fundamental Theorem, and we find

$$\begin{aligned} \int_{-2}^{3} (x^2 - 2x + 5)\ dx &= F(3) - F(-2) \\ &= \left[\frac{1}{3}3^3 - 3^2 + 5 \cdot 3\right] - \left[\frac{1}{3}(-2)^3 - (-2)^2 + 5(-2)\right] \\ &= \frac{95}{3}. \end{aligned}$$

You see the function and the area which we calculated, bounded by the graph, the x-axis, and the two indicated vertical lines, in Figure 4.14. $\diamondsuit$

Example 4.34. Find the area of the region bounded by the graph of the function of $\sin x$ for $x \in [0, \pi]$ and the x-axis. The area is shown in Figure 4.15.

In other words, we are supposed to calculate

$$\int_0^{\pi} \sin x\ dx.$$

Solution: Observe that $-\cos x$ is an anti-derivative of $\sin x$ and that we may apply the Fundamental Theorem of Calculus. We find

$$\int_0^{\pi} \sin x\ dx = -\cos \pi - (-\cos 0) = 2.$$

So the area of the region in question is 2. $\diamondsuit$

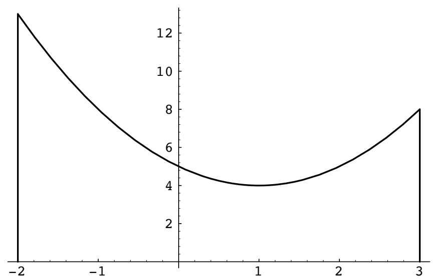

Figure 4.14: $f(x) = x^2 - 2x + 5$

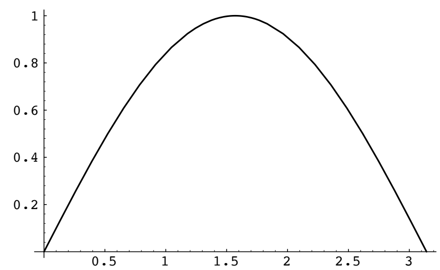

Figure 4.15: $f(x) = \sin x$

Remark 27 (Notational Convention). One commonly uses the notation

$$F(x)\Big|_a^b = F(b) - F(a).$$

This is quite convenient. E.g., we write

$$\sin x\Big|_0^\pi = \sin\pi - \sin 0.$$

If there are ambiguities due to the length of the expression to which this construction is applied, we also use the notation shown in the following example:

$$\Big[x^3 - 5x^2 + 2x - 8\Big]_3^5 = p(5) - p(3)$$

where $p(x) = x^3 - 5x^2 + 2x - 8$.

Example 4.35. Find the integral

$$\int_0^{\pi/4} \sec^2 x \; dx.$$

Solution: The function $F(x) = \tan x$ is an anti-derivative of $f(x) = \sec^2 x$, and, as a differentiable function, $\sec^2 x$ is strongly continuous on the interval $[0, \pi/4]$. Applying the Fundamental Theorem of Calculus we find

$$\int_0^{\pi/4} \sec^2 x \; dx = \tan x\Big|_0^{\pi/4} = 1. \quad \Diamond$$

Example 4.36. Find the integrals

$$\int_0^{\pi/4} \sec x \tan x \ dx \quad \text{and} \quad \int_{\pi/4}^{\pi/3} \csc x \cot x \ dx.$$

Solution: The function $F(x) = \sec x$ is an anti-derivative of $f(x) = \sec x \tan x$, and, as a differentiable function, $\sec x$ is strongly continuous on $[0, \pi/4]$. Applying the Fundamental Theorem of Calculus we find

$$\int_0^{\pi/4} \sec x \tan x \ dx = \sec x \Big|_0^{\pi/4} = \sqrt{2} - 1.$$

The function $F(x) = -\csc x$ is an anti-derivative of $f(x) = \csc x \cot x$, and, as a differentiable function, $\csc x$ is strongly continuous on $[\pi/4, \pi/3]$. Applying the Fundamental Theorem of Calculus we find

$$\int_{\pi/4}^{\pi/3} \csc x \cot x \ dx = -\csc x \Big|_{\pi/4}^{\pi/3} = \left(\frac{-2\sqrt{3}}{3} \right) - (-\sqrt{2}) = \sqrt{2} - \frac{2\sqrt{3}}{3}. \quad \diamondsuit$$

Exercise 147. Evaluate the following definite integrals:

(a) $\displaystyle\int_0^1 (3x+2) \ dx$	(g) $\displaystyle\int_0^{\pi} \frac{1}{2} \cos x \ dx$
(b) $\displaystyle\int_1^2 \frac{6-t}{t^3} \ dt$	(h) $\displaystyle\int_0^{\pi} \cos(x/2) \ dx$
(c) $\displaystyle\int_2^5 2\sqrt{x-1} \ dx$	(i) $\displaystyle\int_{-2}^{2} \lvert x^2 - 1 \rvert \ dx$
(d) $\displaystyle\int_1^0 (t^3 - t^2) \ dt$	(j) $\displaystyle\int_0^{\pi/2} \cos^2 x \ dx$
(e) $\displaystyle\int_{\pi/6}^{\pi/4} \csc x \cot x \ dx$	(k) $\displaystyle\int_0^{\pi/2} \sin^2(2x) \ dx$
(f) $\displaystyle\int_{-1}^{-1} 7x^6 \ dx$	(l) $\displaystyle\int_0^{\pi/4} \sec^2 x \ dx$

Some Proofs

Because of their importance, we like to prove Theorem 4.31 and the Fundamental Theorem of Calculus.

Proof of Theorem 4.31. Because we assumed strong continuity of f on the interval I, it follows from Theorem 4.30 that

$$F(x) = \int_a^x f(t)\ dt$$

exists. So it is our task to show that F is differentiable at x, and that $F'(x) = f(x)$.

We assume that x is not an endpoint of I. We omit (leave to the reader) the modifications of the proof which are required in the case where x is an endpoint of I.

According to the definition of differentiability (see Definition 2.10 and adjust the notation to fit the current setting) we have to show that there exist constants C and $d > 0$, so that

$$|F(x+h) - [F(x) + hf(x)]| \le Ch^2 \tag{4.11}$$

whenever $|h| < d$.

Strong continuity of f provides us with a constant A such that

$$|f(x_1) - f(x_2)| \le A|x_1 - x_2|$$

for all x_1, $x_2 \in I$.

We chose d such that $x+h \in I$ whenever $|h| < d$. In particular, $F(x+h)$ is defined. Setting $A = C$ we will show that (4.11) holds.

$$\begin{aligned}
|F(x+h) - [F(x) + hf(x)]| &= \left| \int_a^{x+h} f(t)\ dt - \int_a^x f(t)\ dt - hf(x) \right| \\
&= \left| \int_x^{x+h} f(t)\ dt - hf(x) \right| \\
&= \left| \int_x^{x+h} (f(t) - f(x))\ dt \right| \\
&\le \left| \int_x^{x+h} |f(t) - f(x)|\ dt \right| \\
&\le \int_x^{x+h} A|h|\ dt \\
&= Ah^2.
\end{aligned}$$

This means that we completed our argument (under the assumption that x is not an endpoint of I). $\square$

Proof of the Fundamental Theorem of Calculus. Essentially, the desired result is an easy consequence of Theorem 4.31. Let $F(x)$ be any anti-derivative of $f(x)$ on I, and $H(x) = \int_a^x f(t)\, dt$ the one provided by Theorem 4.31. In particular, $F'(x) = H'(x) = f(x)$. Cauchy's Theorem (see its application in Corollary 3.6) tells us that F and H differ by a constant. For some constant c and all $x \in I$:

$$H(x) = \int_a^x f(t)\, dt = F(x) + c \tag{4.12}$$

We can find out the value for c by substituting $x = a$ in this equation. In particular, we find that

$$\int_a^a f(t)\, dt = 0 = F(a) + c \qquad \text{or} \qquad c = -F(a).$$

Using this calculation of c and substituting $x = b$ in (4.12), we obtain

$$\int_a^b f(t)\, dt = F(b) - F(a),$$

as claimed. □

Extensions of the Fundamental Theorem of Calculus

Some readers may feel that our Fundamental Theorem, as stated, is somewhat restrictive. The question is: Given an anti-derivative F of the function f, when does the formula

$$\int_a^b f(x)\, dx = F(b) - F(a) \tag{4.13}$$

hold? A quick look at its proof reveals, that it is a rather formal consequence of Theorem 4.31. In this sense it is more relevant to ask in which generality Theorem 4.31 holds. There are two relevant aspects to the discussion:

- **Question:** Given a function f which is defined on an interval $[a, b]$, when does

  $$F(x) = \int_a^x f(t)\, dt$$

  exist for all $x \in [a, b]$?

- **Question:** Assuming the existence of F as in the previous question, when is $F'(x) = f(x)$ for all $x \in [a, b]$?

Let us compare our notion of differentiability with the one which is used more commonly. According to our definition, compare (2.4), a function f is *differentiable* at an interior point c of its domain, if there are number A and d and a line $l(x)$, such that $f(c) = l(c)$ and

$$\left| \frac{f(x) - l(x)}{x - c} \right| \leq A|x - c| \tag{4.14}$$

for all $x \in (c - d, c + d)$ with $x \neq c$.

In a standard treatment of calculus a weaker estimate is used. Instead of (4.14) one requires that

$$\lim_{x \to c} \left| \frac{f(x) - l(x)}{x - c} \right| = 0$$

As usual, assuming differentiability, the derivative of f at c is the slope of the line l, which is denoted by $f'(c)$.

Using the standard definition of differentiability, the beautiful result is, that the answer to both questions is affirmative if f is continuous[8]. Coming up with this coherent answer may be considered to be a major breakthrough in the development of calculus in the 19th century. Another nice feature of the standard definition of differentiability is, that it uses the weakest possible condition which still assures the uniqueness of the derivative. Through the introduction of limits and, what is now considered to be, the standard definition of differentiability, calculus became a mature, elegant mathematical theory.

Only a few non-mathematicians will appreciate this elegance. The concept of limits is complicated enough to cause major difficulties for those who see it for a first time. One important feature of these notes is, that we avoid limits and, more generally, the idea of 'approaching'. For more than a hundred years, mathematicians differentiated functions, and they developed and applied calculus, without even having introduced the idea of a limit and a continuous function. This fine point, as important as it may be for mathematicians, is of little importance for those who apply calculus, and the notes are written for them.

[8]Considering this part of the discussion to be meant for a more mature audience, we do not define the notion of continuity, or limit thereafter.

There is room to close the gap between the two approaches a bit. Let us say that a function f is *differentiable* at an interior point c of its domain, if there exist a number A, an interval $(c-p, c+p)$ around c ($p > 0$), a positive rational number α, and a line l, such that

$$|f(x) - l(x)| \leq A|x - c|^{1+\alpha} \tag{4.15}$$

for all $x \in (c - p, c + p)$. As before, $l(x)$ is called the tangent line to the graph of f at c, and its slope is $f'(c)$, the derivative of f at c. Similarly, a function is said to be *strongly continuous* on a closed interval $[a, b]$, if for some constant A and all x_1 and x_2 in $[a, b]$

$$|f(x_1) - f(x_2)| \leq A|x - c|^{\alpha}. \tag{4.16}$$

Our entire discussion of differentiation with all of its theorems and rules goes through without essential changes. The value for α may be fixed for the entire discussion, or it can be kept variable. Using this modified definition of differentiability and strong continuity, the answer to both of our questions from above is once more affirmative, and both, Theorem 4.31 and the Fundamental Theorem, will hold.

Example 4.37. Find the derivative of $f(x) = |x|^{3/2}$ at $x = 0$ and calculate the integral $\int_0^1 \sqrt{x}\, dx$.

Remark: With the original definition of differentiability, the derivative of f at $x = 0$ was not defined, and $\int_0^1 \sqrt{x}\, dx$ could not be found by a straight forward application of the Fundamental Theorem of Calculus.

Solution: Set $\alpha = 1/2$ and $l(x) = 0$. Then $|f(x) - l(x)| = |x|^{3/2}$, so that f is differentiable at $x = 0$ in the sense of the modified definition, and $f'(0) = 0$.

On the interval $[0, 1]$ the function $\sqrt{x}$ has an anti-derivative, namely $F(x) = (2x^{3/2})/3$. We can apply the Fundamental Theorem and calculate

$$\int_0^1 \sqrt{x}\, dx = \frac{2}{3}x^{3/2}\Big|_0^1 = \frac{2}{3}. \quad \diamondsuit$$

4.7 Substitution

In some cases it is not that easy to 'see' an anti-derivative of the function one likes to integrate. Substitution is a method which, when applied correctly, will simplify the expression for the function you like to integrate. You hope that you can find an anti-derivative for the simplified expression. The method is based on the chain rule for differentiation.

We explain the method. Let F and g be functions which are defined and differentiable on an interval I. Set $F' = f$. Then, according to the chain rule,

$$\frac{d}{dx}F(g(x)) = f(g(x))g'(x).$$

Assume that f and g' are strongly continuous on I. Then $f(g(x))g'(x)$ is strongly continuous as well. Knowing that $F(g(x))$ is an anti-derivative of $f(g(x))g'(x)$ we conclude that

$$\int f(g(x))g'(x)\ dx = F(g(x)) + c. \tag{4.17}$$

The left hand side of the equation denotes, by definition, the indefinite integral of $f(g(x))g'(x)$. On the right hand side of the equation you see one anti-derivative of $f(g(x))g'(x)$, to which we then added an arbitrary constant to obtain the indefinite integral of $f(g(x))g'(x)$. This means that the two expressions are the same.

Let us give a few examples to illustrate how this method can be put to use. There are no general rules what substitution must be used, rather success justifies the means. Working through the examples will teach you how to apply this method in some typical situations. It will give you at least some experience which you may then rely on in similar examples.

Example 4.38. Find the indefinite integral

$$\int (2x-3)^3\ dx.$$

Solution: Set $g(x) = 2x - 3$ and $f(u) = u^3$. Then $g'(x) = 2$. We write the integral in such a way that we see the terms from (4.17) explicitly.

$$\int (2x-3)^3\ dx = \frac{1}{2}\int (2x-3)^3 \cdot 2dx.$$

Setting $F(u) = \frac{u^4}{4}$, which is an anti-derivative of $f(u)$, we find that

$$\int (2x-3)^3\ dx = \frac{1}{8}(2x-3)^4 + c.$$

To confirm the calculation, you may verify that the derivative of the expression on the right hand side of the equation is indeed $(2x-3)^3$. $\Diamond$

There is a pattern, a way to use the notation, which can be applied to write down the steps in an integration using substitution efficiently. Setting $u = g(x)$ we write

$$du = g'(x)dx,$$

instead of $g'(x) = du/dx$[9]. Suppose also that F is an anti-derivative of f, so $F' = f$. Then the pattern for calculating an integral via substitution is

$$\int f(g(x))g'(x)\ dx = \int f(u)\ du = F(u) + c = F(g(x)) + c. \tag{4.18}$$

In the first step of this calculation we carry out the substitution, in the second one we find the anti-derivative, and in the third one we reverse the substitution. We make use of this notation in our next example.

Example 4.39. Find the indefinite integral

$$\int x\sqrt{x^2+2}\ dx.$$

Solution: We use the substitution $u = x^2 + 2$. Then $\frac{du}{dx} = 2x$, or $2xdx = du$. We find

$$\begin{aligned} \int x\sqrt{x^2+2}\ dx &= \frac{1}{2}\int \sqrt{x^2+2}\cdot 2xdx \\ &= \frac{1}{2}\int \sqrt{u}\ du \\ &= \frac{1}{3}u^{3/2} + c \\ &= \frac{1}{3}(x^2+2)^{3/2} + c. \end{aligned}$$

We suggest once more that you verify the calculation by showing that $\frac{d}{dx}(\frac{1}{3}(x^2+2)^{3/2} + c) = x\sqrt{x^2+2}$. $\diamondsuit$

Example 4.40. Find the indefinite integral

$$\int t^2(t+1)^7\ dt.$$

[9] We do not attach any particular meaning to the symbols dx and du in their own right. The equation $du = g'(x)dx$ helps us to write down what happens when we perform the substitution as in the first equality in (4.18). Thought of as infinitesimals or differentials, these symbols have a meaning, but this is beyond the scope of these notes.

Solution: We use the substitution $u = t + 1$. Then $du = dx$ and $t = u - 1$. We calculate:

$$\begin{aligned}
\int t^2(t+1)^7\ dt &= \int (u-1)^2 u^7\ du \\
&= \int (u^2 - 2u + 1)u^7\ du \\
&= \int (u^9 - 2u^8 + u^7)\ du \\
&= \frac{1}{10}u^{10} - \frac{2}{9}u^9 + \frac{1}{8}u^8 + c \\
&= \frac{1}{10}(t+1)^{10} - \frac{2}{9}(t+1)^9 + \frac{1}{8}(t+1)^8 + c. \quad \Diamond
\end{aligned}$$

Example 4.41. Find the indefinite integral

$$\int 2x\sin^2(x^2+5)\ dx.$$

Solution: We use the substitution $u = x^2+5$. Then $du = 2xdx$. We also use the trigonometric identity $\sin^2\alpha = [1-\cos 2\alpha]/2$. Then we calculate:

$$\begin{aligned}
\int 2x\sin^2(x^2+5)\ dx &= \int \sin^2 u\ du \\
&= \frac{1}{2}\int [1-\cos 2u]\ du \\
&= \frac{1}{2}\left[u - \frac{1}{2}\sin 2u\right] + c \\
&= \frac{1}{2}\left[(x^2+5) - \frac{1}{2}\sin[2(x^2+5)]\right] + c. \quad \Diamond
\end{aligned}$$

Example 4.42. Find the indefinite integral

$$\int \sec^2 x\tan x\ dx.$$

Solution: We use the substitution $u = \sec x$. Then $du = \sec x\tan x\ dx$, and we calculate that

$$\begin{aligned}
\int \sec^2 x\tan x\ dx &= \int \sec x\cdot\sec x\tan x\ dx \\
&= \int u\ du \\
&= \frac{1}{2}u^2 + c \\
&= \frac{1}{2}\sec^2 x + c. \quad \Diamond
\end{aligned}$$

Sometimes we have to apply the method of substitution twice, or more often, to work out an integral. Here is an example.

Example 4.43. Find the indefinite integral

$$\int (x^2+1)\sin^3(x^3+3x-2)\cos(x^3+3x-2)\ dx.$$

In a first step, we use the substitution $u = x^3+3x-2$. Then $du = 3(x^2+1)\ dx$. We find:

$$\int (x^2+1)\sin^3(x^3+3x-2)\cos(x^3+3x-2)\ dx \quad = \quad \frac{1}{3}\int \sin^3 u \cos u\ du.$$

In a second substitution we set $v = \sin u$. Then $dv = \cos u\ du$. Continuing the calculation we find

$$\begin{aligned}\frac{1}{3}\int \sin^3 u\cos u\ du &= \frac{1}{3}\int v^3\ dv\\ &= \frac{1}{12}v^4 + c\\ &= \frac{1}{12}\sin^4 u + c.\end{aligned}$$

Next, we reverse our first substitution. We remember that $u = x^3+3x-2$, and find that

$$\int (x^2+1)\sin^3(x^3+3x-2)\cos(x^3+3x-2)\ dx = \frac{1}{12}\sin^4(x^3+3x-2)+c$$

We suggest that you differentiate the right hand side of the last equation to verify the computation. $\diamondsuit$

Substitution and Definite Integrals

Let us now explore how substitution is used to calculate definite integrals. Assuming as before that f and g' are strongly continuous on the interval $[a,b]$, we have

$$\int_a^b f(g(x))g'(x)\ dx = \int_{g(a)}^{g(b)} f(u)\ du. \tag{4.19}$$

To see this, observe that f has an anti-derivative, which we again denote by F. Then

$$\int_a^b f(g(x))g'(x)\ dx = F(g(x))\Big|_a^b = F(u)\Big|_{g(a)}^{g(b)} = \int_{g(a)}^{g(b)} f(u)\ du.$$

The first identity is obtained as a combination of the Fundamental Theorem of Calculus and (4.17). The second one is obvious, and the third one is another application of the Fundamental Theorem of Calculus.

Let us apply this formula in a few examples.

Example 4.44. Calculate the integral

$$\int_0^1 (x^2-1)(x^3-3x+5)^3\ dx.$$

Solution: It appears promising to use the substitution $u = x^3 - 3x + 5$. Then $du = (3x^2-3)\ dx$, and $\frac{1}{3}du = (x^2-1)\ dx$. To obtain the limits for the integral we calculate $u(0) = 5$ and $u(1) = 3$. Then we find

$$\begin{aligned}\int_0^1 (x^2-1)(x^3-3x+5)^3\ dx &= \frac{1}{3}\int_5^3 u^3\ du \\ &= \frac{1}{12}u^4\Big|_5^3 \\ &= \frac{1}{12}[81-625] \\ &= -\frac{136}{3}. \quad \Diamond\end{aligned}$$

Example 4.45. Calculate the integral

$$\int_0^{\pi/4} \cos^2 x \sin x\ dx.$$

Solution: We use the substitution $u = \cos x$. Then $-du = \sin x\ dx$. If $x = 0$, then $u = 1$, and if $x = \pi/4$, then $u = \sqrt{2}/2$. Then we find

$$\begin{aligned}\int_0^{\pi/4} \cos^2 x \sin x\ dx &= -\int_1^{\sqrt{2}/2} u^2\ du \\ &= -\frac{1}{3}u^3\Big|_1^{\sqrt{2}/2} \\ &= -\frac{1}{3}\left[1-\frac{\sqrt{2}}{4}\right]. \quad \Diamond\end{aligned}$$

Example 4.46. Calculate the integral

$$\int_0^2 x(x+1)^6\ dx.$$

Solution: We use the substitution $u = x + 1$. Then $du = dx$ and $x = u - 1$. If $x = 0$, then $u = 1$, and if $x = 2$, the $u = 3$. We calculate

$$\begin{aligned}
\int_0^2 x(x+1)^6 \, dx &= \int_1^3 (u-1)u^6 \, du \\
&= \int_1^3 u^7 - u^6 \, du \\
&= \left[\frac{1}{8}u^8 - \frac{1}{7}u^7\right]_1^3 \\
&= \frac{6560}{8} - \frac{2186}{7} \\
&= \frac{3554}{7}. \quad \Diamond
\end{aligned}$$

Example 4.47. Calculate the integral

$$\int_0^{\sqrt{8}} x^3 \sqrt{x^2+1} \, dx.$$

Solution We use the substitution $u = x^2 + 1$. Then $\frac{1}{2}du = x \, dx$ and $x^2 = u - 1$. For the limits we calculate, if $x = 0$, then $u = 1$, and if $x = \sqrt{8}$, then $u = 9$. Then we find

$$\begin{aligned}
\int_0^{\sqrt{8}} x^3 \sqrt{x^2+1} \, dx &= \frac{1}{2} \int_1^9 (u-1)\sqrt{u} \, du \\
&= \frac{1}{2} \int_1^9 \left(u^{3/2} - u^{1/2}\right) \, du \\
&= \frac{1}{2} \left[\frac{2}{5}u^{5/2} - \frac{2}{3}u^{3/2}\right]_1^9 \\
&= \frac{1}{2} \left[\frac{484}{5} - \frac{52}{3}\right]
\end{aligned}$$

You may simplify the last expression. $\Diamond$

Example 4.48. Calculate the integral

$$\int_0^1 \sqrt{1-x^2} \, dx.$$

Solution: Let us carry out the calculation. We use the substitution $x = \sin u$. Then $dx = \cos u \, du$. If $x = 0$, then $u = 0$, and if $x = 1$, then

$u = \pi/2$. (For our given values of x, there are other possible values for u, but they will lead to the same results.) Then

$$\begin{aligned} \int_0^1 \sqrt{1-x^2}\,dx &= \int_0^{\pi/2} \sqrt{1-\sin^2 u}\cos u\,du \\ &= \int_0^{\pi/2} \cos^2 u\,du \\ &= \frac{\pi}{4}. \end{aligned}$$

The missing steps in the calculation are obtained from Example 4.27 $\Diamond$

Remark 28. The graph of $f(x) = \sqrt{1-x^2}$ is the northern part of a circle. Using $x \in [0,1]$ means that we calculated the area under this graph in the first quadrant, i.e., the area of one forth of the disk of radius 1. You were told long time ago in school, that the area of this unit disk is π, so that the result of the calculation is hardly surprising.

There is a more serious matter. Is the example genuine, or did we assume the answer previously? By definition, π is the ratio of the circumference of a circle by its diameter. In our calculation of the derivative of the sine and cosine functions we used the estimate that $|\sin h - h| \le h^2/2$. When we showed this, we used that $|h| \le |\tan h|$ for $h \in [-\pi/4, \pi/4]$. A typical proof of the latter inequality starts out by first showing that the area of the unit disk is π. This means, we assumed the result in the example, we did not derive it.

Exercise 148. Find the following integrals:

(a) $\displaystyle\int \frac{dx}{\sqrt{2x+1}}$

(b) $\displaystyle\int \frac{t}{(4t^2+9)^2}\,dt$

(c) $\displaystyle\int t(1+t^2)^3\,dt$

(d) $\displaystyle\int \frac{2s}{\sqrt[3]{6-5s^2}}\,ds$

(e) $\displaystyle\int \frac{b^3x^3}{\sqrt{1-a^4x^4}}\,dx$

(f) $\displaystyle\int_0^{\pi} x\cos x^2\,dx$

(g) $\displaystyle\int x^2\sqrt{x+1}\,dx$

(h) $\displaystyle\int \frac{x+3}{\sqrt{x+1}}\,dx$

(i) $\displaystyle\int \sin^2(3x)\,dx$

(j) $\displaystyle\int_0^{\pi/2} \cos^2 x\,dx$

(k) $\displaystyle\int_{-\pi/6}^{\pi/6} \sec(2x)\tan(2x)\,dx$

(l) $\displaystyle\int_0^{1/2} \frac{dx}{4+x^2}$

(m) $\displaystyle\int \frac{\sec^2 x}{\sqrt{1+\tan x}}\,dx$

(n) $\displaystyle\int \sqrt{1+\sin x}\cos x\,dx$

(o) $\displaystyle\int_0^{r} \sqrt{r^2-x^2}\,dx$

4.8 Areas between Graphs

Previously we related the integral to areas of a region under a graph. This idea can be generalized to the discussion of areas of regions between two graphs. Let us look at an example.

Example 4.49. Find the area of the region between the graphs of the functions $f(x) = x^2$ and $g(x) = \sqrt{1-x^2}$.

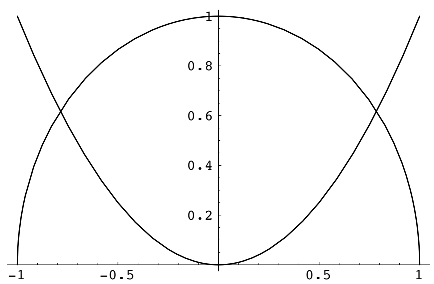

Figure 4.16: Region between two graphs

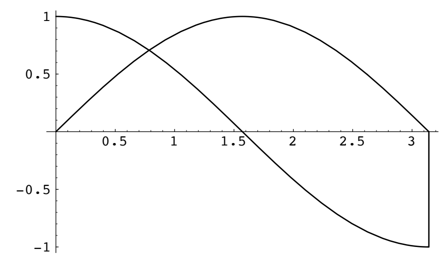

Figure 4.17: Region between two graphs

Solution To get a better picture of the problem, we draw the two graphs. They are shown in Figure 4.16. Now you see the region between the two graphs whose area we want to calculate. Let us call the region Ω.

As you see, the graphs intersect in two points. We like to find their x-coordinates. In other words, we are looking for numbers x, so that $f(x) = g(x)$. That means

$$x^2 = \sqrt{1-x^2}.$$

We square this equation and solve the resulting equation equation for x^2. According to the quadratic formula:

$$x^2 = \frac{-1 \pm \sqrt{5}}{2}.$$

Only the + sign occurs as $x^2 \geq 0$. Taking the square root, we find the

x-coordinates of the points where the curves intersect:

$$A = -\sqrt{\frac{-1+\sqrt{5}}{2}} \quad \text{and} \quad B = \sqrt{\frac{-1+\sqrt{5}}{2}}.$$

To get the area of the region under the graph of $f(x)$ and $g(x)$ over the interval $[A, B]$ we can calculate the appropriate integrals. To get the area of the region Ω between the graphs, we take the area of the region under the graph of $g(x)$ and subtract the area of the region under the graph of $f(x)$. Concretely:

$$\text{Area}(\Omega) = \int_A^B g(x)\ dx - \int_A^B f(x)\ dx = \int_A^B (g(x) - f(x))\ dx \approx 1.06651.$$

The numerical value was obtained by computer. You are invited to work out the integral with the help of the Fundamental Theorem of Calculus to verify the result. $\diamondsuit$

Some problems are a bit more subtle.

Example 4.50. Find the area of the region between the graphs of the functions $f(x) = \cos x$ and $g(x) = \sin x$ for x between 0 and π.

Solution: The region Ω between the graphs is shown in Figure 4.17. The region breaks up into two pieces, the region Ω_1 over the interval $[0, \pi/4]$ on which $f(x) \geq g(x)$, and the region Ω_2 over the interval $[\pi/4, \pi]$ where $g(x) \geq f(x)$. We calculate the areas of the regions Ω_1 and Ω_2 separately.

In each case, we proceed as in the previous example:

$$\begin{aligned}
\text{Area}(\Omega_1) &= \int_0^{\pi/4} (\cos x - \sin x)\ dx = (\sin x + \cos x)\Big|_0^{\pi/4} = \sqrt{2} - 1 \\
\text{Area}(\Omega_2) &= \int_{\pi/4}^{\pi} (\sin x - \cos x)\ dx = -(\sin x + \cos x)\Big|_{\pi/4}^{\pi} = 1 + \sqrt{2}.
\end{aligned}$$

In summary we find:

$$\text{Area}(\Omega) = \text{Area}(\Omega_1) + \text{Area}(\Omega_2) = 2\sqrt{2}.$$

An additional remark may be in place. When we compared integrals and areas, we had to take into account where the function is non-negative, resp., non-positive. Here we did not. We took care of this aspect by breaking up the interval into the part where $f(x) \geq g(x)$ and the part where $g(x) \geq f(x)$. As we are dealing with differences, it does not matter that the functions themselves are negative on part of the interval $[0, \pi]$. You should think about this aspect of the problem. $\diamondsuit$

Our general definition for the area between two graphs is as follows.

Definition 4.51. *Suppose $f(x)$ and $g(x)$ are integrable functions over an interval $[a, b]$. Let Ω be the region between the graphs of $f(x)$ and $g(x)$ for x between a and b. The area of Ω is*

$$\text{Area}(\Omega) = \int_a^b |f(x) - g(x)| \, dx.$$

This definition generalizes Definition 4.11 on page 216. The definition is also consistent with the intuitive idea of the area of a region, and it incorporates and generalizes Proposition 4.21 on page 221. Taking the absolute value of the difference of $f(x)$ and $g(x)$ allows us avoid the question where $f(x) \geq g(x)$ and where $g(x) \geq f(x)$. Typically this problem gets addressed when the integral is calculated. In some problems a and b are explicitly given, in others you have to determine them from context. In all cases it is good to graph the functions before calculating the area of the region between them. Having the correct picture in mind helps you to avoid mistakes.

Exercise 149. Sketch and find the area of the region bounded by the curves:

(a) $y = x^2$ and $y = x^3$.

(b) $y = 8 - x^2$ and $y = x^2$

(c) $y = x^2$ and $y = 3x + 5$.

(d) $y = \sin x$ and $y = \pi x - x^2$.

(e) $y = \sin x$ and $y = 2 \sin x \cos x$ for x between 0 and π.

4.9 Numerical Integration

The Fundamental Theorem of Calculus provided us with a highly efficient method for calculating definite integrals. Still, for some functions we have no good expression for its anti-derivative. In such cases we may have to rely on numerical methods for integrating. Let us take such a function, and show some methods for finding an approximate value for the integral.

We describe different ways to find, by numerical means, approximate values for the integral of a function $f(x)$ over the interval $[a, b]$:

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

In all of the different approaches we partition the interval into smaller ones:

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

Left and Right Endpoint Method: In the left endpoint method we find the value of the function at each left endpoint of the intervals of the partition. We multiply it with the length of the associated interval, and then add up the terms. Explicitly, we calculate

$$I_L = f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_1) + \cdots + f(x_{n-1})(x_n - x_{n-1}). \tag{4.20}$$

In the right endpoint method we proceed as we did on the left endpoint method, only we use the value of the function at the right endpoint instead of the left endpoint:

$$I_R = f(x_1)(x_1 - x_0) + f(x_2)(x_2 - x_1) + \cdots + f(x_n)(x_n - x_{n-1}). \tag{4.21}$$

Example 4.52. Use the left and right endpoint method to find approximate values for

$$\int_0^2 e^{-x^2}\ dx.$$

Solution: Set $f(x) = e^{-x^2}$. We carry out our calculations with 10 digits accuracy. We partition the interval $[0, 2]$ into 4 smaller intervals by choosing three intermediate points:

$$x_0 = 0 < x_1 = \frac{1}{2} < x_2 = 1 < x_3 = \frac{3}{2} < x_4 = 2.$$

Then $x_k - x_{k-1} = 1/2$ for $k = 1$, 2, 3, and 4. Formula (4.20) for I_L specializes to

$$I_L = \frac{f(0) + f(1/2) + f(1) + f(3/2)}{2}.$$

Evaluating the expressions in this formula provides us with a numerical value for I_L:

$$I_L = 1.126039724.$$

Formula (4.21) for I_R specializes to

$$I_R = \frac{f(1/2) + f(1) + f(3/2) + f(2)}{2}.$$

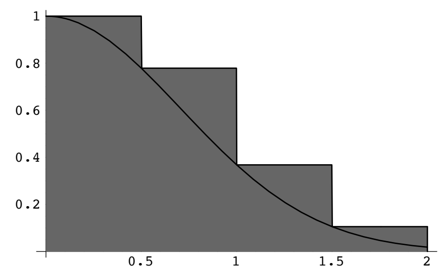

Figure 4.18: Use left endpoints

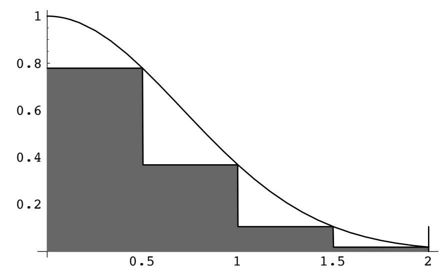

Figure 4.19: Use right endpoints

Evaluating the expressions in this formula provides us with a numerical value for I_R:

$$I_R = .6351975438.$$

Apparently, I_L and I_R are calculated by combining the areas of certain rectangles. In our case the values of $f(x)$ are all positive and all of the rectangles are above the x axis, so the areas of the rectangles are all added. Note also, that our specific function $f(x)$ is decreasing on the interval $[0, 2]$, so that I_L is an upper sum for the function $f(x)$ over the interval $[0, 2]$, and I_R is a lower sum. In this sense, we have

$$I_R = .6351975438 \leq \int_0^2 e^{-x^2}\ dx \leq I_L = 1.126039724.$$

The function and the rectangles whose areas are added to give us I_L and I_R are shown in Figure 4.18 and Figure 4.19. $\diamondsuit$

Midpoint and Trapezoid Method: We may try and improve on the endpoint methods. In the midpoint methods, we use the value of the function at the midpoints of the intervals of the partition. That should be less bias. We use the same partition and notation as above. Then the formula for the midpoint method is:

$$I_M = f\left(\frac{x_0 + x_1}{2}\right)(x_1 - x_0) + \cdots + f\left(\frac{x_n + x_{n-1}}{2}\right)(x_n - x_{n-1}). \tag{4.22}$$

In the trapezoid method we do not take the function at the average (i.e. midpoint) of the endpoints of the intervals in the partition, but we average the values of the function at the endpoints. Specifically, the formula is

$$I_T = \frac{f(x_0)+f(x_1)}{2}(x_1-x_0)+\cdots+\frac{f(x_{n-1})+f(x_n)}{2}(x_n-x_{n-1}). \tag{4.23}$$

It is quite easy to see that

$$I_T = \frac{I_L+I_R}{2}. \tag{4.24}$$

Let us explain the reference to the word trapezoid. For simplicity, suppose that $f(x)$ is non-negative on the interval $[a,b]$. Consider the trapezoid of width (x_1-x_0) which has height $f(x_0)$ at its left and $f(x_1)$ at its right edge. The area of this trapezoid is $\frac{f(x_0)+f(x_1)}{2}(x_1-x_0)$. This is the first summand in the formula for I_T, see (4.23). We have such a trapezoid over each of the intervals in the partition, and their areas are added to give I_T.

Expressed differently, we can draw a secant line through the points $(x_0, f(x_0))$ and $(x_1, f(x_1))$. This gives us the graph of a function $T(x)$ over the interval $[x_0, x_1]$. Over the interval $[x_1, x_2]$ the graph of $T(x)$ is the secant line through the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$. Proceeding in the fashion, we use appropriate secant lines above all of the intervals in the partition to define the function $T(x)$ over the entire interval $[a,b]$. Then

$$I_T = \int_a^b T(x)\ dx.$$

This integral is easily computed by the formula in (4.23).

Example 4.53. Use the midpoint and trapezoid method to find approximate values for

$$\int_0^2 e^{-x^2}\ dx.$$

Solution: We use the same partition of $[0,2]$ and accuracy as in Example 4.52. The formula for I_M (see (4.22)) specializes to

$$I_M = \frac{f(.25)+f(.75)+f(1.25)+f(1.75)}{2}.$$

Evaluating the expressions in this formula provides us with a numerical value for I_M:

$$I_M = .8827889485.$$

As for the endpoint methods, I_M is the combined area of certain rectangles. Their heights are the values $f(x_i)$ at the midpoints of the intervals of the partition. Their width are the lengths of the intervals of the partition. The areas of the rectangles are added or subtracted, depending on whether $f(x_i)$ is positive or negative. In our specific case, the areas of the rectangles are all added. You see this calculation illustrated in Figure 4.20.

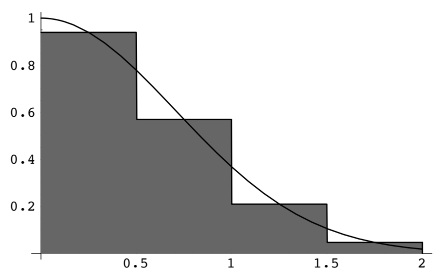

Figure 4.20: Use midpoints

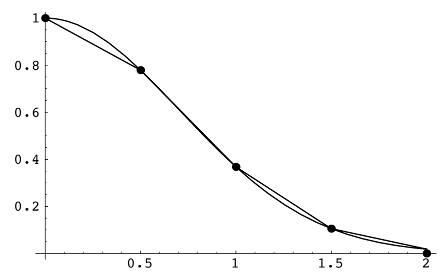

Figure 4.21: Trapezoid Method

Based on our previous calculations and Formula (4.24) we find

$$I_T = \frac{I_L + I_R}{2} = .8806186341.$$

We illustrated this calculation in Figure 4.21. There you see the function $f(x) = e^{-x^2}$ for which we like to find the integral over the interval $[0, 2]$. You also see five dots on the graph, and they are connected by straight line segments. These line segments form the graph of a function $T(x)$, and I_T is the area of the region under this graph. So

$$I_T = \int_0^2 T(x)\,dx. \quad \diamondsuit$$

Simpson's Method: In Simpson's method we combine the endpoint and midpoint methods in a weighted fashion. Again, we use the same notation for the function and the partition as above. The specific formula for an approximate value of the integral of $f(x)$ over $[a,b]$ is

$$\begin{aligned} I_S =& \frac{1}{6}\left[f(x_0)+4f\left(\frac{x_0+x_1}{2}\right)+f(x_1)\right](x_1-x_0)+\cdots \\ &+\frac{1}{6}\left[f(x_{n-1})+4f\left(\frac{x_{n-1}+x_n}{2}\right)+f(x_n)\right](x_n-x_{n-1}) \end{aligned} \tag{4.25}$$

It is quite easy to see that

$$I_S = \frac{I_L + 4I_M + I_R}{6} = \frac{I_T + 2I_M}{3}.$$

Let us explain the background to Simpson's method. We define a function $P(x)$ over the interval $[a,b]$ by defining a degree 2 polynomial on each of the intervals of the partition. The polynomial over the interval $[x_{k-1}, x_k]$ is chosen so that it agrees with $f(x)$ at the endpoints and at the midpoint of this interval. In other words, the polynomial goes through the points $(x_{k-1}, f(x_{k-1}))$, $(\frac{x_{k-1}+x_k}{2}, f(\frac{x_{k-1}+x_k}{2}))$ and $(x_k, f(x_k))$. With some work one can then show that

$$I_S = \int_a^b P(x)\ dx.$$

In this sense, Simpson's method is a refinement of the Trapezoid method. In one method we use two points on the graph and connect them by a straight line segment. In the other one we use three points on the graph and construct a parabola through them.

Example 4.54. Use Simpson's method to find an approximate value for

$$\int_0^2 e^{-x^2}\ dx.$$

Solution: We use the same partition of $[0,2]$ and accuracy as in Example 4.52. The formula for I_S (see the special case of (4.25)) specializes to

$$I_S = \frac{I_L + 4I_M + I_R}{6},$$

where I_L, I_M and I_R are as above. Using the values from Examples 4.52 and 4.53, we come up with a numerical value for I_S:

$$I_S = .88206555104.$$

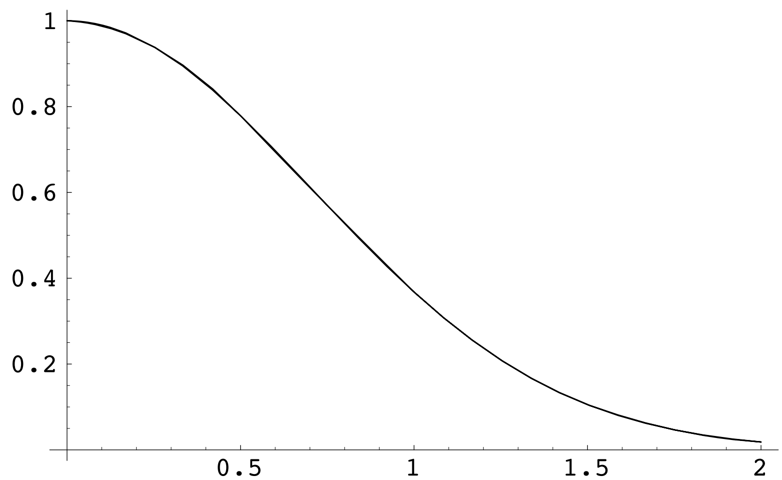

Figure 4.22: Simpson's Method

You see the method illustrated in Figure 4.22. There you see the graphs of two functions, the function $f(x) = e^{-x^2}$ and the function $P(x)$ from the discussion of Simpson's method. Only the thickness of the line suggests that there are two graphs of almost identical functions. $\diamondsuit$

Example 4.55. Compare the accuracy of the various approximate values of

$$\int_0^2 e^{-x^2}\,dx.$$

Solution: We compare the approximate values for the integral obtained by the different formulas. In addition, we vary the number n of smaller intervals into which $[0,2]$ is subdivided. We tabulate the results. They should be compared with an approximate value for the integral of

$$0.882081390762421.$$

	$n=1$	$n=10$	$n=100$	$n=1000$
I_L	2.0000000	0.9800072469	0.891895792451	0.883063050702697
I_R	0.0366313	0.7836703747	0.872262105229	0.881095681980474
I_M	0.7357589	0.8822020700	0.882082611663	0.882081402972833
I_T	1.0183156	0.8818388108	0.882078948840	0.882081366341586
I_S	0.8299445	0.8820809836	0.882081390722	0.882081390762417

Table 4.1: Approximate Values of the Integral

It should be apparent that Simpson's method is far superior to all of the other ones. E.g., Simpson's method with $n = 4$ gives a result which is better than the left and right endpoint method with $n = 1000$. Even if you use the Midpoint and trapezoid method with $n = 1000$, then the result is far less accurate that Simpson's method with $n = 100$. $\diamondsuit$

Remark 29. Why do we care about the number n of intervals into which we partition $[a, b]$ when we apply one of the methods for finding approximate numerical values for an integral. This number determines the number of algebraic operations which have to be carried out to come up with an answer. Since computers are fast, this does not seem to be particularly important. But, with every algebraic operation you have to expect round-off errors, and these can quickly add up to a substantial error in the final answer. So it is relevant to use a method which requires the least number of arithmetic operations. Simpson's method requires only about twice as many algebraic operations as the endpoint methods for the same value of n, still it gives a far superior result.

Exercise 150. Proceed as in Example 4.55 and compare the different methods applied to the calculation of

$$\int_0^{\pi/2} \sin x \, dx.$$

4.10 Applications of the Integral

In Definition 4.11 on page 216 and Proposition 4.21 on page 221 we related definite integrals to areas. Based on the context, this can have a more concrete meaning. Consider a function $f(t)$ on an interval $[a,b]$ and the integral

$$I = \int_a^b f(t)\ dt.$$

- If $f(t)$ stands for the rate at which a drug is absorbed, then I is the total amount of the drug which has been absorbed in the time interval $[a,b]$.
- If $f(t)$ stands for the rate at which people come down with the flu, then I is the number of people who contracted the sickness during the time interval $[a,b]$.
- If $f(t)$ stands for the speed with which you travel, then I stands for the total distance which you traveled during the time interval $[a,b]$.

You are invited to come up with more interpretations. In addition, the following definition expresses the common notion of the average value of a function.

Definition 4.56. *Suppose that $f(t)$ is an integrable function over the interval $[a,b]$. Then the quantity*

$$f_{av} := \frac{1}{b-a}\int_a^b f(t)\ dt$$

is called the average value *of $f(t)$ over the interval $[a,b]$.*

Let us look at an example.

Example 4.57. The river Little Brook flows into a reservoir, referred to as Beaver Pond by the locals. The amount of water carried by the river depends on the season. As a function of time, it is

$$g(t) = 2 + \sin\left(\frac{\pi t}{180}\right).$$

We measure time in days, and $t = 0$ corresponds to New Year. The units of $g(t)$ are millions of liter of water per day. Water is released from Beaver Pond at a constant rate of 2 million liters per day. At the beginning of the year, there are 200 million liters of water in the reservoir.

(a) How many liter of water will there be in Beaver Pond by the end of April?

(b) Find a function $F(t)$ which tells how much water there is in the reservoir at any day of the year.

(c) At which rate does the amount of water in the reservoir change at the beginning of September?

(d) On which days will there be 250 million liters of water in Beaver Pond?

(e) At which amount of water will the reservoir crest?

(f) On the average, by how much has the amount of water in Beaver Pond increased per day during the first three months of the year?

Solution: At the same time, water enters and leaves the pond. The net rate at which the water enters the pond is

$$f(t) = g(t) - 2 = \sin\left(\frac{\pi t}{180}\right).$$

The units are millions of liters per day. We obtain the total change of the amount of water in the reservoir by integrating $f(t)$. Set

$$A(T) = \int_0^T f(t)\ dt.$$

On the T-th day of the year, the total amount of water in Beaver Pond will be

$$F(T) = 200 + \int_0^T f(t)\ dt = 200 + \frac{180}{\pi}\left[1 - \cos\left(\frac{\pi T}{180}\right)\right].$$

The units for $F(T)$ are millions of liters. With this we have answered (b).

For our calculation we suppose that the year has 360 days and each month has 30 days. By the end of April, i.e., after 120 days, there will be

$$F(120) = 200 + \frac{180}{\pi}\left[1 - \cos\frac{2\pi}{3}\right] \approx 238.2$$

millions of liters of water in the pond. This answers (a).

Let us address (c). The rate at which the amount of water in the pond changes is

$$F'(t) = f(t).$$

At the beginning of September, i.e., after 240 days, the rate of change is

$$f(240) \approx -.866.$$

The pond is losing water at a rate of 866 thousand liters of water per day.

To answer (d), we like to know for which T we have $F(T) = 250$. We solve for T the equation:

$$250 = 200 + \frac{180}{\pi}\left[1 - \cos\left(\frac{\pi T}{180}\right)\right] \quad \text{or} \quad \cos\left(\frac{\pi T}{180}\right) = 1 - \frac{5\pi}{18}.$$

We apply the function arccos to both sides of the last equation and find

$$T = \frac{180}{\pi} \arccos\left(1 - \frac{5\pi}{18}\right) \approx 88, \text{ or } 272.$$

On the 88-th and 272-nd day of the year there will be 250 millions of liters of water in the reservoir.

To find at which amount the reservoir crests, we have to find the maximum value of $F(t)$. This occurs apparently when $\cos(\pi t/180) = -1$ or $t = 180$. The pond crests at mid-year, and then the amount of water in it is about 314.6 millions of liters of water. This answers (e).

After three months or 90 days there are about 257.3 millions of liters of water in Beaver Pond. Within this time, the amount of water has increased by 57.3 millions of liters. On the average, the amount of water in the reservoir increased by about 640,000 liters per day. $\diamond$

Exercise 151. A pain reliever has been formulated such that it is absorbed at a rate of $600\sin(\pi t)$ (mg/hr) by the body. Here t measures time in hours, $t = 0$ at the time you take the medication, and the absorption process is complete at time $t = 1$.

(a) What is the total amount of the drug which is absorbed within one hour?

(b) Find a function $F(t)$, such that $F(t)$ tells how much medication has been absorbed at time t.

(c) A total of 150 mg of the medication has to be absorbed before the drug is effective. How long does it take until this threshold is reached?

4.11 The Exponential and Logarithm Functions

In Section 1.3 we introduced the exponential function $\exp(x) = e^x$ and the natural logarithm function $\ln x$. At the time we only stated that they exist because we did not have the tools to properly define them. We will now fill in the details. Many of the routine calculations are formulated as exercises.

Observe that $f(t) = 1/t$ is a strongly continuous function on the interval $[b, \infty)$ for any $b > 0$. To see this, note that

$$|f(t_0) - f(t_1)| = \left|\frac{1}{t_0} - \frac{1}{t_1}\right| = \frac{|t_1 - t_0|}{t_0 t_1} \leq \frac{1}{b^2}|t_0 - t_1|$$

for all $t_0,\ t_1 \geq b$.

Definition 4.58. *Let $x \in (0, \infty)$. The natural logarithm of x is defined as*

$$\ln x = \int_1^x \frac{dt}{t}. \tag{4.26}$$

The integral defining $\ln x$ in (4.26) exists due to Theorem 4.30. So we defined the natural logarithm function, and its domain is $(0, \infty)$.

Theorem 4.59. *The natural logarithm function is differentiable on its entire domain $(0, \infty)$, its derivative is*

$$\ln' x = \frac{1}{x},$$

and $\ln x$ is increasing on $(0, \infty)$.

Proof. Theorem 4.31 tells us that $\ln x$ is differentiable and that $\ln' x = 1/x$. According to Theorem 3.12, the function is increasing because its derivative is positive everywhere, $\ln' x > 0$ for all $x > 0$. □

Let us also verify one of the central equations for calculating with logarithms, the third rule in Theorem 1.19.

Proposition 4.60. *For any $x,\ y > 0$,*

$$\ln(xy) = \ln x + \ln y. \tag{4.27}$$

Proof. We need a short calculation. Here x and y are fixed positive numbers. We use the substitution $u = \frac{t}{x}$, so that $du = \frac{1}{x}\ dt$. For the adjustment of the limits of integration, observe that $t/x = u = 1$ when $t = x$, and that $t/x = u = y$ when $t = xy$. Then

$$\int_x^{xy} \frac{dt}{t} = \int_x^{xy} \frac{1}{(t/x)} \frac{1}{x}\, dt = \int_1^y \frac{du}{u} = \ln y.$$

Using this calculation we deduce that

$$\ln(xy) = \int_1^{xy} \frac{dt}{t} = \int_1^x \frac{dt}{t} + \int_x^{xy} \frac{dt}{t} = \ln x + \ln y.$$

This is exactly our claim. □

Exercise 152. Show:

(1) $\ln 1 = 0$.

(2) $\ln(1/y) = -\ln y$ for all $y > 0$.

(3) $\ln(x/y) = \ln x - \ln y$ for all $x, y > 0$.

Exercise 153. Show that $\ln 4 > 1$. Hint: Using the partition

$$1 = x_0 < 2 = x_1 < 3 = x_2 < 4 = x_3,$$

find a lower sum S_l for the function $1/t$ over the interval $[1, 4]$ so that $S_l > 1$.

We can now define the Euler number:

Definition 4.61. *The number e is the unique number such that*

$$\ln e = 1 \quad \text{or, equivalently,} \quad \int_1^e \frac{dt}{t} = 1.$$

For this definition to make sense, we have to show that there is a number e which has the property used in the definition. To see this, observe that $\ln 1 = 0 < 1 < \ln 4$. Because $\ln x$ is differentiable, it follows from the Intermediate Value Theorem (see Theorem 2.65) that there is a number e for which $\ln e = 1$. It also follows that $1 < e < 4$.

Proposition 4.62. *For every real number x there exists exactly one positive number y, such that*

$$\ln y = x \tag{4.28}$$

Proof. We saw that $\ln y$ is an increasing function. This means that, for any given x, the equation

$$\ln y = x.$$

has at most one solution. We show that the equation has a solution, which then means that it has a unique solution.

Observe that $\ln(e^n) = n$ and $\ln(1/e^n) = -n$ for all natural numbers n, i.e., for all n of the form 1, 2, 3, 4, ... That means that all integers (whole numbers) are values of the natural logarithm function. Invoking once more the Intermediate Value Theorem, we conclude that for any number x which lies between two integers there is a real number y so that $\ln y = x$. But every real number lies between two integers, and so our proposition is proved. □

Exercise 154. Show that

$$\ln(a^r) = r \ln a$$

for all positive numbers a and all rational numbers r, i.e., numbers of the form $r = p/q$ where p and q are integers and $q \neq 0$.

In summary, we have seen that

Corollary 4.63. *The natural logarithm function $\ln x$ is a differentiable, increasing function with domain $(0, \infty)$ and range $(-\infty, \infty)$, and $\ln' x = 1/x$.*

We are now ready to define the exponential function.

Definition 4.64. *Given any real number x, we define $\exp(x)$ to be the unique number for which*

$$\ln(\exp(x)) = x, \tag{4.29}$$

i.e., $y = \exp(x)$ is the unique solution of the equation $\ln(y) = x$. This assignment (mapping x to $\exp(x)$) defines a function, called the exponential function, with domain $(-\infty, \infty)$ and range $(0, \infty)$.

Exercise 155. Show that the exponential function exp and the natural logarithm function ln are inverses of each other. (You may want to review the notion of two functions being inverses of each other in Section 5.6.) In addition to the equation in (4.29), you need to show that

$$\exp(\ln(y)) = y \tag{4.30}$$

for all $y \in (0, \infty)$.

Summarizing this discussion, and adding some observations which we have made elsewhere, we have:

Proposition 4.65. *The exponential function* $\exp(x)$ *is a differentiable, increasing function with domain* $(-\infty, \infty)$ *and range* $(0, \infty)$*, and the exponential function is its own derivative, i.e.,* $\exp'(x) = \exp(x)$*.*

Exercise 156. Show for all real numbers x and y that:

(1) $\exp(0) = 1$

(2) $\exp(1) = e$

(3) $\exp(x)\exp(y) = \exp(x + y)$

(4) $1/\exp(y) = \exp(-y)$

(5) $\exp(x)/\exp(y) = \exp(x - y)$.

Hint: Use the results of Exercise 152, the definition of e in Definition 4.61, and that the exponential and logarithm functions are inverses of each other.

Exercise 157. Show that $\exp(r) = e^r$ for all rational numbers r. Hint: Use Exercise 154 and that the exponential and logarithm functions are inverses of each other.

The expression e^r makes sense only if r is a rational number. If $r = p/q$ then we raise e to the r-th power and take the q-th root of the result. For an arbitrary real number we set

$$e^x = \exp(x). \tag{4.31}$$

This is consistent with the meaning of the expression for rational exponents due to Exercise 157, and it defines what we mean by raising e to any real power.

Other Bases

So far we discussed the natural logarithm function and the exponential function with base e. We now expand the discussion to other bases.

Definition 4.66. *Let* a *be a positive number,* $a \neq 1$*. Set*

$$\log_a x = \frac{\ln x}{\ln a} \quad \text{and} \quad \exp_a(x) = \exp(x \ln a). \tag{4.32}$$

We call $\log_a(x)$ *the logarithm function with base* a *and* $\exp_a(x)$ *the exponential function with base* a*. Here* $x > 0$ *if used as an argument for* $\log_a$*, so that* $(0, \infty)$ *is the domain for the function* $\log_a(x)$*. Also,* x *is any real number if used as an argument for* $\exp_a$*, so that* $(-\infty, \infty)$ *is the domain for the function* $\exp_a(x)$*.*

Exercise 158. Show

(1) $\ln a > 0$ if $a > 1$ and $\ln a < 0$ if $0 < a < 1$.

(2) $\log_a(x)$ and $\exp_a(x)$ are differentiable functions.

(3) $\log_a(x)$ and $\exp_a(x)$ are increasing functions if $a > 1$.

(4) $\log_a(x)$ and $\exp_a(x)$ are decreasing functions if $0 < a < 1$.

Exercise 159. Suppose $a > 0$ and $a \neq 1$. Show that, with the domains specified in Definition 4.66:

(1) The range of $\log_a(x)$ is $(-\infty, \infty)$.

(2) The smallest possible range of $\exp_a(x)$ is $(0, \infty)$.

Remark 30. We will always use $(0, \infty)$ as the range of $\exp_a(x)$.

Exercise 160. Suppose $a > 0$ and $a \neq 1$. Show that

(1) $\exp_a(\log_a(y)) = y$ for all $y > 0$.

(2) $\log_a(\exp_a(x)) = x$ for all real numbers x.

Taken together, the specifications for the domains and ranges for the functions $\exp_a$ and $\log_a$ and the results from Exercise 160 tell us that

Corollary 4.67. *Suppose* $a > 0$ *and* $a \neq 1$*. The functions* $\exp_a$ *and* $\log_a$ *are inverses of each other.*

Exercise 161. Suppose $a > 0$ and $a \neq 1$. Show the laws of logarithms:

(1) $\log_a 1 = 0$.

(2) $\log_a(xy) = \log_a x + \log_a y$ for all x, $y > 0$.

(3) $\log_a(1/y) = -\log_a y$ for all $y > 0$.

(4) $\log_a(x/y) = \log_a x - \log_a y$ for all x, $y > 0$.

Exercise 162. Suppose $a > 0$ and $a \neq 1$. Show the exponential laws:

(1) $\exp_a(0) = 1$

(2) $\exp_a(1) = a$

(3) $\exp_a(x)\exp_a(y) = \exp_a(x+y)$

(4) $1/\exp_a(y) = \exp_a(-y)$

(5) $\exp_a(x)/\exp_a(y) = \exp_a(x-y)$.

Exercise 163. Suppose $a > 0$, $a \neq 1$, and r is a rational number. Show

$$\log_a(a^r) = r \quad \text{and} \quad \exp_a(r) = a^r.$$

We rephrase a convention which we made previously for e. Suppose $a > 0$ and $a \neq 1$. The expression a^r makes sense if r is a rational number. If $r = p/q$ then we raise a to the r-th power and take the q-th root of the result. For an arbitrary real number we set

$$a^x = \exp_a(x). \tag{4.33}$$

This is consistent with the meaning of the expression for rational exponents due to Exercise 163, and it defines what we mean by raising a to any real power. Equation 4.33 specializes to the one in Equation 4.31 if we set $a = e$. It is also a standard convention to set

$$1^x = 1 \quad \text{and} \quad 0^x = 0$$

for any real number x. Typically 0^0 is set 1.

We can now state an equation which is typically considered to be one of the laws of logarithms:

Exercise 164. Suppose $a > 0$, $a \neq 1$, $x > 0$, and z is any real number. Then

$$\log_a(x^z) = z\log_a(x).$$

We are now ready to fill in the details for one of the major statements which we made in Section 1.3. We are ready to prove

Theorem 4.68. *Let a be a positive number, $a \neq 1$. There exists exactly one monotonic function, called the exponential function with base a and denoted by $\exp_a(x)$, which is defined for all real numbers x such that $\exp_a(x) = a^x$ whenever x is a rational number.*

Proof. In this section we constructed the function $\exp_a(x)$, and this function has all of the properties called for in the theorem. That settles the existence statement. We have to show the uniqueness statement, i.e., there in only one such function.

Suppose $f(x)$ is any monotonic function and $f(r) = a^r$ for all rational numbers r. We have to show that $f(x) = \exp_a(x)$ for all real numbers x. We assume that there exists some real number x so that $f(x) \neq \exp_a(x)$, and deduce a contradiction.

Suppose $f(x) \neq \exp_a(x)$. Then there exists some number $z \neq x$, so that $\exp_a(z) = f(x)$. We assume that f and $\exp_a$ are both increasing and $x > z$. (The cases where f and $\exp_a$ are both decreasing and $x < z$ are left to the reader.) There exists a rational number r between z and x, so $z < r < x$, and

$$\exp_a(z) < exp_a(r) = f(r) < f(x).$$

This contradicts our assertion that $\exp_a(z) = f(x)$. Our assumption that $f(x) \neq \exp_a(x)$ for some x must have been wrong. This means that $f(x) = \exp_a(x)$ for all real numbers x. $\square$

Chapter 5

Prerequisites from Precalculus

In this chapter we collect some material from precalculus which you are expected to know already. If there are some topics which you don't feel comfortable with and our discussion is too brief, then you will have to go back to the text for a more basic course and review it. Even if you know all of the material it may be good to look it over one more time to refresh your memory, and possibly learn something new. There is one important exception. In Theorem 5.7 on page 265 we state that the real numbers are complete. This property of the real numbers is typically not discussed in a precalculus course. There are also two formulas which are important for our calculation of the derivative of the sine and cosine function which you may not have seen previously, see (5.30).

5.1 The Real Numbers

After mentioning the natural numbers, integers, and rational numbers, we offer two ways to think of the real numbers, and we explain the statement that the real numbers are complete.

Let us establish some names for various kinds of numbers. The natural numbers are the numbers 1, 2, 3, 4, etc. They are useful for counting objects, and they can be added and multiplied. The integers (or whole numbers) are the numbers 0, ± 1, ± 2, ± 3, etc. These numbers can be added, subtracted, and multiplied. The rational numbers are the numbers of the form p/q where p and q are integers and $q \neq 0$. If we write out their decimal expansion, then they have a repeating block. All basic algebraic operations, addition, multiplication, subtraction, and division can be performed with them.

The real numbers are an even larger set. They include numbers such as $\sqrt{2}$ and π, numbers whose decimal expansion does not have a repeating block. A mathematically rigorous introduction of the real numbers is quite challenging. We offer two ways for thinking about them.

Real numbers are numbers which have a (not necessarily unique) decimal expansion, and any decimal expansion represents a real number. In a more geometric approach we identify the real numbers with the points on a line. Typically, we draw the line and distinguish one point which we call the origin or 0 (zero). In addition we impose a scale, which then tells us how to identify the points on the line with the decimal expansion of real numbers. Real numbers can be added, subtracted, multiplied and divided.

Real numbers are ordered. Let x and y be two real numbers, which are identified with points on the real number line. We say that x is larger than y (in mathematical notation $x > y$) if x lies to the right of y, and x is smaller than y (or $x < y$) if x lies to the left of y. If x is smaller or equal to y then this is written as $x \leq y$. If x is greater or equal to y then this is written as $x \geq y$. The statement that the real numbers are ordered means that we can compare any two of them. This is formalized by either of the following statements.

Proposition 5.1. *Suppose x and y are real numbers. Then exactly one of the following three possibilities holds:* (1) $x < y$, (2) $x > y$, or (3) $x = y$.

Proposition 5.2. *Suppose x and y are real numbers. Then either $x \geq y$ or $x \leq y$. If $x \geq y$ and $x \leq y$, then $x = y$.*

Relying on algebra instead of geometry, we could have made the following definitions.

Definition 5.3. *Suppose x and y are real numbers. We say that $x > y$ if and only if $x - y$ is positive*[1]*, and $x < y$ if and only if $x - y$ is negative.*

Most of our functions are defined on intervals of real numbers, or unions thereof. Let us characterize intervals.

Definition 5.4. *A subset I of the real numbers is an* interval *if, whenever $A \in I$, $B \in I$, and c a real number such that $A < c < B$, then $c \in I$.*

[1] To make sense out of this statement we have to specify first what a positive number is. Here we rely on your intuition.

A typical interval is $[a, b)$. It contains all x between a and b ($a < x < b$). It also contains the point a, as indicated by the use of the square bracket [. It does not contain the point b, as indicated by the use of the parenthesis).

Some concepts are preferably defined over open intervals, and some concepts are particularly meaningful if considered over closed intervals. Without going into a formal definition, we state which types of intervals are open, closed, open and closed, and neither.

Definition 5.5. *Let a and b be real numbers. Suppose that $a < b$. Intervals of the form (a, b), $(-\infty, b)$, and (a, ∞) are* open*. Intervals of the form $[a, b]$, $(-\infty, b]$, and $[a, \infty)$ are* closed*. The interval $(-\infty, \infty)$ is* open and closed*. The intervals of the form $[a, b)$ and $(a, b]$ are* neither open nor closed*, but sometimes they are called* half-open *and/or* half-closed*. In addition, $[a, a]$ is a closed interval which consists of a single point, and the empty set is an open interval.*

Still, if you like to have an explanation, then consider the following definition. A subset I of the real numbers is *open* if for all $c \in I$, there exist some positive number d, such that the interval $(c - d, c + d)$ is a subset of I. Complements of open sets are called closed. With these definitions you can deduce the characterization of the intervals as open, closed, open and closed, and neither open nor closed in Definition 5.5.

So far we have done little with the real numbers which we could not also have done with the rational ones. But, there is a fundamental difference between these two number systems. The technical term for this is that the real numbers are *complete*. We need to introduce another concept to explain completeness. Let A be a subset of the real numbers. We say that a real number M is an *upper bound* of A if $a \leq M$ for all $a \in A$. We call a real number m a *lower bound* of A if $a \geq m$ for all $a \in A$. A *least upper bound* for the set A (abbreviated as $\mathrm{lub}(A)$) is exactly what the word suggests. It is an upper bound L of A, and if M is any upper bound of A, then $L \leq M$. In the corresponding way we define the notion of a *greatest lower bound* of a set A of real numbers. Here the abbreviation is $\mathrm{glb}(A)$. A set of real numbers is *bounded* if it has an upper and a lower bound.

Example 5.6. Let A be the set of all negative real numbers. This set has no lower bound. Every non-negative number is an upper bound of A. The least upper bound of A is 0.

Theorem 5.7. (Completeness of the Real Numbers.) *Every bounded set of real numbers has a least upper bound and a greatest lower bound.*

The proof of this theorem is a subject of an introductory course in real analysis. This is well beyond the scope of a calculus course. We stated the theorem because the completeness of the real numbers is at the heart of several important theorems which we will encounter in calculus. The rational numbers are not complete, and essential aspects of calculus are false if we tried to do calculus using only these numbers.

5.2 Inequalities and Absolute Value

Using common sense and some caution, one can calculate with inequalities just as with equations. Here are the basic rules for strict inequalities. You can develop the corresponding statements for $\leq$ and $\geq$. Let x, y, and z be real numbers. Then

(1) If $x > y$ and $y > z$, then $x > z$.

(2) If $x > y$ then $x + z > y + z$.

(3) If $x > y$ and $z > 0$, then $xz > yz$.

(4) If $x > y$ and $z < 0$, then $xz < yz$.

Example 5.8. Find all solutions of the inequality:

$$2x + 6 > 10.$$

Solution: Adding -6 to both sides of the inequality turns the inequality into an equivalent one: $2x > 4$. Multiplication by $1/2$ reformulates it as $x > 2$. In the language of intervals, this means that $x \in (2, \infty)$. $\diamondsuit$

A notational tool of particular importance is the absolute value. The absolute value of x is denoted by $|x|$. E.g., $|2| = 2$ and $|-2| = 2$. In general:

$$\text{If } x \geq 0, \text{ then } |x| = x, \text{ and if } x \leq 0, \text{ then } |x| = -x. \tag{5.1}$$

On the real line you can think of $|x|$ as the distance between the point x and 0, the origin. Similarly, $|x - y|$ is the distance between the points x and y on the real line. Another way to define the absolute value is to set

$$|x| = \sqrt{x^2}, \tag{5.2}$$

where it is understood that we take the non-negative square root of x^2.

Here are some things which are conveniently expressed with the help of the absolute value. Suppose a and b are real numbers and $a > 0$.

$$\text{If}\quad |x| < a \quad\text{then}\quad -a < x < a \quad\text{or}\quad x \in (-a, a). \tag{5.3}$$

$$\text{If}\quad |x-b| < a \quad\text{then}\quad -a < x-b < a \quad\text{or}\quad x \in (b-a, b+a). \tag{5.4}$$

$$\text{If}\quad |x-b| \le a \quad\text{then}\quad -a \le x-b \le a \quad\text{or}\quad x \in [b-a, b+a]. \tag{5.5}$$

Expressed in words, $|x - b| \le a$ means that the distance between x and b is no more than a, or that x is no further away from b than a. This means that x must lie in an interval around b, no further than a to the left of b and no further than a to the right of it, in other words, $x \in [b - a, b + a]$.

E.g., $|x - 4| \le 3$ means that $x \in [1, 7]$, and $|x + 2| < 1$ means that $x \in (-3, -1)$.

Complementing (5.3), we have

$$\text{If } |x-b| \ge a, \text{ then } x \notin (b-a, b+a) \text{ or } x \in (-\infty, b-a] \cup [b+a, \infty).$$

In summary, in the last paragraph we explained how to use absolute values and inequalities to express that the difference between two numbers x and y is less (or no more) than some number a, and how to write down intervals of the form $(b - a, b + a)$ or $[b - a, b + a]$.

Properties of the absolute value:

There are some general properties of the absolute value which are important to know. Let x, y, and c be real numbers. Then

$$\begin{aligned}
&(5.6) & |x| &\ge 0\\
&(5.7) & |x| &= 0 \quad \text{if and only if } x = 0\\
&(5.8) & |cx| &= |c||x|\\
&(5.9) & |x+y| &\le |x| + |y|.\\
&(5.10) & x^2 &\le y^2 \quad \text{implies that } |x| \le |y|.
\end{aligned}$$

The statement in (5.9) is referred to as *triangle inequality.* Let us prove this inequality.

$$(x+y)^2 = x^2 + 2xy + y^2 \le |x|^2 + 2|x||y| + |y|^2 = (|x| + |y|)^2.$$

Using (5.10) and $||x| + |y|| = |x| + |y|$ we conclude

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

There is an important variant of the triangle inequality. For all real numbers x and y

$$||x| - |y|| \leq |x - y|. \tag{5.11}$$

To see this inequality, observe that

$$|x| = |(x - y) + y| \leq |x - y| + |y| \quad \text{and} \quad |y| = |x + (y - x)| \leq |x| + |x - y|.$$

This tells us that

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

and with the help of (5.3) we conclude

$$||x| - |y|| \leq |x - y|.$$

5.3 Functions, Definition and Notation

You can think of a function as a particular way of recording information. More specifically, a *function* consists of three pieces of data, a set A (called the *domain*), a set B (called the *range*), and an instruction which assigns to each element in the domain (call it x) one in the range (called the image of x). As an example from a student's life, you can consider the process of assigning grades. The domain are the students in the course, the range are the grades, say A, B, C, D, F, and the assignment is the process by which the instructor assigns a grade to each student. You come up with statements like "the grade of Curious George is an A," or "the instructor assigned to Curious George an A." You could watch the stock market and record its closing index daily. So, to each day of activity you assign a number. Assign to each item in the federal budget a number, the amount budgeted for it. You get a function from the different items to the real numbers. Consider the high and the low of the daily temperature. As a function of the date, each of them gives you a function. In a zoology experiment you might measure the concentration of a hormone in an animal at different times in its life and consider it as a function of age.

You should not have the impression that all functions depend on just one variable. If you watch a vibrating string, then its displacement from the resting position depends on the point on the string and time, so you have a function of two variables. Other processes, like the unemployment rate depend on many variables. Life is full of such functions.

Let us look at a more purely mathematical example. Use $(-\infty, \infty)$ as domain and $[0, \infty)$ as range. The assignment is to send x to x^2. Then x^2 is the image of x. Unless necessary, and that won't happen very often, we specify only the domain of the function and the instruction which provides the assignment. It is useful to introduce an efficient way to write down a function. For the example we write:

"Let $f(x) = x^2$ be defined on $(-\infty, \infty)$."

As range we could have chosen $(-\infty, \infty)$, or $[0, \infty)$, or any set which contains $[0, \infty)$. Many times we even omit the specification of the domain. So we may talk about the function $f(x) = x^2$, and we understand, without saying so, that we use the largest set as domain on which the instruction makes sense. The use of the notation $f(x) = x^2$ is efficient as it specifies the assignment by an equation relating the argument x to which the function is applied and its image $f(x)$. The careful reader will notice that we are using the letter f for the function as well as for the assignment which is part of the data of the function. There is no reason to worry about this abuse of language as it will not lead to any confusion.

In many texts on calculus you will find our statement abbreviated as: *"Consider the function $y = x^2$."* Here the reader is expected to fill in, that y is a function and $y(x)$ is given by the equation $y(x) = x^2$. Often x is referred to as the *independent variable* (the variable which can be chosen freely) and y as the *dependent variable* (the variable which depends on x and the way in which the function is defined). In Section 5.4 on page 274 we return to the discussion of this notation where it gets a more meaningful interpretation.

A related concept is the one of the *image* of a function. Given a function f with domain A and range B, the image of f consists of those $b \in B$ for which there exists an $a \in A$ such that $f(a) = b$. In other words, the image of a function f from A to B consists of the elements $f(a)$ where a varies over all the elements in the domain A.

Often we discuss functions more generally and not a specific function. In this case we might write "*Let f be a function defined on ...* " or "*Let $f(x)$ be a function defined on ...* " Which option we use is often dictated by the statement we like to make, or how we like to make it, but sometimes it is also arbitrary. You should get used to both conventions.

The names of the functions and the variables are not fixed. We might consider the area of a square as a function of its side length. If the side length is denoted by l and the area by A, then the function is expressed by the equation $A(l) = l^2$. In many applications the notation is chosen to suggest the correct interpretation, such as A for area in the example, t for time, a for acceleration, etc. We will vary the notation to avoid getting you hooked on a specific one.

Examples of Functions and Graphs

You should think of a graph of a function as information made visible. There are many ways of doing this, and which method is preferable depends on the context. Books, news papers, and TV are full of information displayed so that you can grasp it easily. Tables, bar graphs, and pie charts are simple examples. A map, as you find it in your atlas, is a very sophisticated example. Functions in two variables are often made visible with 3D graphs or by showing level curves.

How do we present a function? For the functions which we have in mind, there are several ways for doing this. First of all, we could describe them by a formula. Although this is the most efficient way in a mathematical treatment, it may not be the easiest way to convey the information to a general audience. We could give sample numerical values, say organized in a table, and many people will be happy with this. A third method is to draw the graph of the function. We will rely extensively on this method in our own treatment. We provide several examples to demonstrate these alternatives.

Let us recall the method of graphing functions in the Cartesian plane. To draw the coordinate system we start out with two perpendicular axes in the plane, one horizontal and the other one vertical. Often, but not always, they are called the x and y-axes. They intersect in the *origin* of the plane. We think of each axis as a copy of the real line, directed such that the x-values increase as we go to the right, and the y-values increase as we go up. Given any point p in the plane, we can draw a vertical line and horizontal line through it. The intersection point of the vertical line with the x-axis is a real number. It is called the x-coordinate of p. The intersection point of the horizontal line through p with the y-axis is called the y-coordinate of the point. If these intersection points are x_0 and y_0, then we also write $p = (x_0, y_0)$. Every pair (a, b) of real numbers defines a point in the plane, the point with x-coordinate a and y-coordinate b. Every point in the plane may be described by giving its x and y-coordinates. In Figure 5.1 you see this instruction applied to the point $(1.5, 2.5)$.

Graphing a function means to indicate the points $(x, f(x))$ in the plane

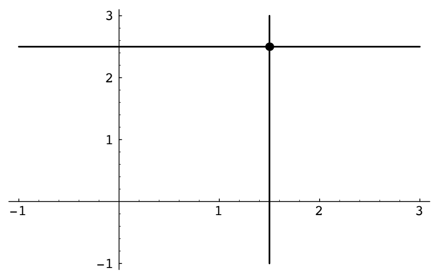

Figure 5.1: Coordinates of a point in the plane

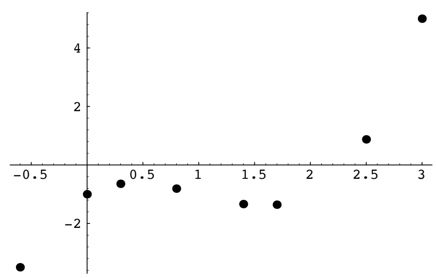

Figure 5.2: Some values of a function

for all x in the domain of f. Typically we have to restrict our attention to some interval on the x-axis (of the domain) and possibly also to some interval on the y-axis (of the range) because our piece of paper or computer screen has only finite dimensions.

In addition, we can also put down only finitely many points, we cannot find and draw infinitely many points. After drawing enough of them we'll be willing to connect the points, in a sensible way, to get a curve. The interpolation in this last step has to be done cautiously, and should be based on insight into the function. This will be possible for the functions which we are discussing in these notes.

The graphs of some functions cannot be drawn. There are graphs of bounded continuous functions over finite closed intervals which are infinitely long. The function could also be so wild, that we cannot calculate sufficiently many points which can then be connected in a sensible way, i.e., with the hope of the graph being close to the line which we drew to connect the points. There are also functions whose graph is not a curve.

Example 5.9. In Table 5.1 you find eight values for a function. We graphed these points in Figure 5.2. Suppose that the function is defined for all x in the interval $[-1, 3]$. What can you say about the function?

Solution: All you can say is that the graph goes through the indicated points. In practice, you may see these points connected by straight line segments. E.g., some papers will do it if they graph the DJI. They'll put

x	$f(x)$	&	x	$f(x)$	&	x	$f(x)$	&	x	$f(x)$
$-.6$	-3.496		$.3$	$-.643$		1.4	-1.336		2.5	$.875$
0	-1		$.8$	$-.808$		1.7	-1.357		3	5

Table 5.1: Values of a function

a point for the closing index for each day, and then connect the points by lines. If you refine the graph and include hourly updates, then the graph might look quite different.

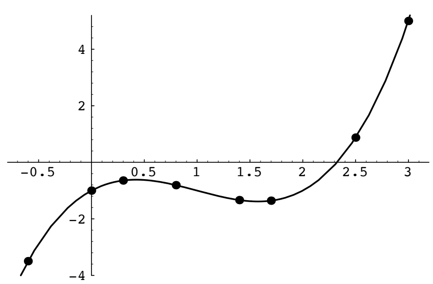

Figure 5.3: A degree three curve through the points

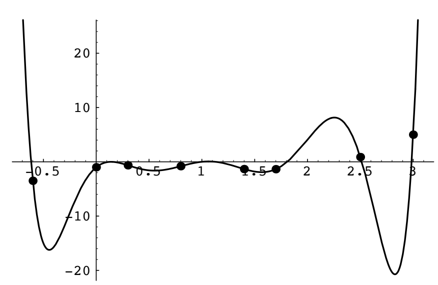

Figure 5.4: A different polynomial

With your previous experience you may also be inclined to connect the points by a smoother curve. You may remember the typical graph of a polynomial of degree three, and connect the points as shown in Figure 5.3. The actual expression for this polynomial is

$$p(x) = x^3 - 3x^2 + 2x - 1. \tag{5.12}$$

Unless you are told to try a degree three polynomial, there would not have been any reason for this solution either. You might try a different polynomial. In Figure 5.4 you see the graph of a polynomial of degree eight which goes through the same points. The graph looks quite different. We will have to stay honest and use only the information which is available to us. $\Diamond$

Example 5.10. Relate the measurement of temperature in degrees Celsius and Fahrenheit.

Solution: Depending on where you live in this world, temperature is measured on different scales. We consider two common ways for measuring temperature, the one using degrees Celsius (C) and the one using degrees Fahrenheit (F). There are further scales which we do not consider here.

A Conversion Table: You see a conversion table in Table 5.2. It allows you to go from one to the other scale for some fixed temperatures. You can guess in between values, or obtain them by interpolation. A thermometer which has both scales provides a table of this kind. Hospitals may have such a thermometer so that they can tell patients their body temperature on the scale they are used to.

deg C	deg F	&	deg F	deg C
−10	14		0	−17.78
0	32		20	−6.67
10	50		40	4.44
20	68		60	15.56
30	86		80	26.67
40	104		100	37.78
100	212		120	48.89

Table 5.2: Degrees Celsius and Fahrenheit

Functional Relation: You can provide the formula which relates the measurement of temperature in degrees Celsius and Fahrenheit. If t is the temperature in degrees Celsius and T the temperature in degrees Fahrenheit, then

$$T = \frac{9}{5}t + 32 \quad \text{and} \quad t = \frac{5}{9}(T - 32).$$

You can plug in the temperature measured on one scale, and the formula provides you the value in the other one.

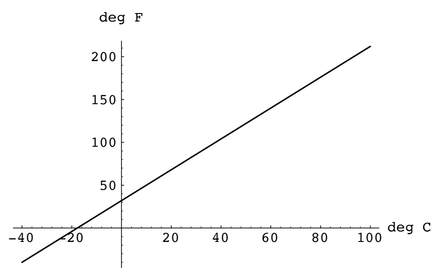

Figure 5.5: deg C to deg F

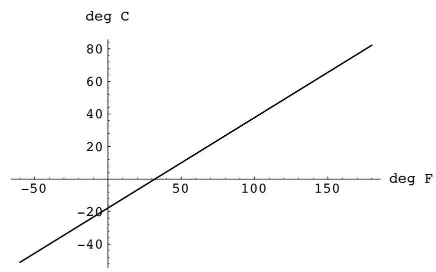

Figure 5.6: deg F to deg C

A Graphical Relation: You may graph the functions $T(t)$ and $t(T)$ and read off the conversions. You can see the graphs in Figures 5.5 and 5.6. Actually, one graph suffices. You may take the variable of either axis as the independent variable, and on the other axis you read off the value of the dependent variable. After adjusting scales, you get one graph from the other one by reflection at the diagonal. $\diamondsuit$

5.4 Graphing Equations

When we graph a function $f(x)$, then we plot the points $(x, f(x))$ in the plane. If we follow the convention to call the coordinate axes the x and y-axes, the this means that we plot the points (x, y) in the plane which are solutions of the equation

$$y = f(x) \quad \text{or} \quad y - f(x) = 0.$$

E.g., Figure 5.3 on page 272 may be considered either as the graph of the function $f(x) = x^3 - 3x^2 + 2x - 1$, or as the set of solutions of the equation $y - x^3 - 3x^2 + 2x - 1 = 0$.

More generally, we consider an equation of the form

$$F(x, y) = 0 \tag{5.13}$$

where $F(x, y)$ is some expression in x and y. Graphing the equation, or more exactly the solutions of this equation, means to plot all points (x, y) in the x-y-plane for which the equation holds, i.e., for which $F(x, y) = 0$.

Example 5.11. Graph the solution of the equation

$$F(x, y) = x^2 + y^2 - 1.$$

Solution: The graph of this equation is the circle of radius 1, centered at the origin of the coordinate system. In fact, the unit circle is defined as the set of points (x, y) in the plane which satisfy this equation. It should be emphasized that the unit circle is not the graph of a function, though, separately, the northern and the southern hemisphere are. They are graphs of the functions $f_+(x) = \sqrt{1 - x^2}$ and $f_- = -\sqrt{1 - x^2}$. $\diamondsuit$

Example 5.12. Enjoy the Lissajous figure given by the equation

$$y^2 - x^2(1 - x^2) = 0,$$

and shown in Figure 5.7.

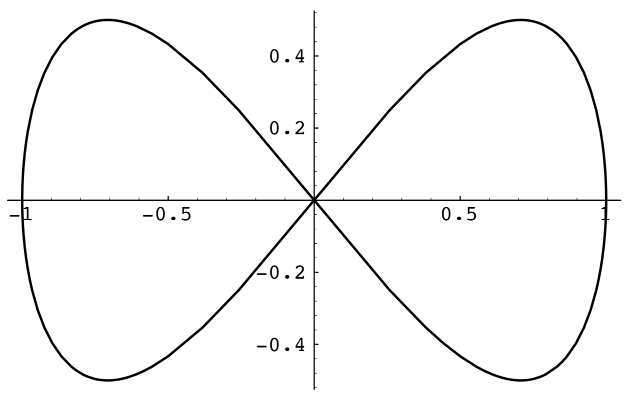

Figure 5.7: $y^2 - x^2(1 - x^2) = 0$

The point of the example is, that you get a fair amount of information about the curve by having a look at the graph. The equation will be more useful in actual calculations. $\diamondsuit$

Let us make a historical remark. The present day definition of a function goes back to the last century. In earlier centuries a function was understood as something for which one could build a machine which would draw its graph. In some older mathematics departments you still find showcases filled with these machines, which are often beautiful pieces of machinery built by craftsmen. You should keep this method in mind as one of the important ways to understand functions. It is a fourth way, aside from the three ways we discussed so far, writing down an equation, tabulating values, and graphing.

5.5 Trigonometric Functions

In this section we discuss the radian measure of angles and introduce the trigonometric functions. These are the functions sine, cosine, tangent, et. al. We collect some formulas relating these functions, and prove two formulas (see (5.30)) which are used in an essential way in the differentiation of the sine and cosine function (see Examples 2.11 on page 50 and Exercise 37 on page 51).

Arc Length and Radian Measure of Angles: Consider the unit circle (a circle with radius 1) centered at the origin in the Cartesian plane. It is shown in Figure 5.8. We take a practical approach to measuring the length of an arc on this circle. We imagine that we can straighten it out, and measure how long it is. It requires some work to introduce the idea of the length of a curve in a mathematically rigorous fashion.

Definition 5.13. *The number π is the ratio between the circumference of a circle and its diameter.*

This definition goes back to the Greeks. Stated differently it says, that the circumference of a circle of radius r is $2\pi r$. Observe that the ratio referred to in the definition does not depend on the radius of the circle.

Consider an angle α between the positive x-axis and a ray which originates at the origin of the coordinate system and intersects the unit circle in the point p. We like to find the radian measure of the angle α. Consider an arc on the unit circle which starts out at the point $(1, 0)$ and ends at p, and suppose its length is s. Then

$$\alpha = \pm s \text{ (radians)}. \tag{5.14}$$

The $+$ sign is used if the arc goes counter clockwise around the circle. The $-$ sign is used if it proceeds clockwise. We may also consider arcs which

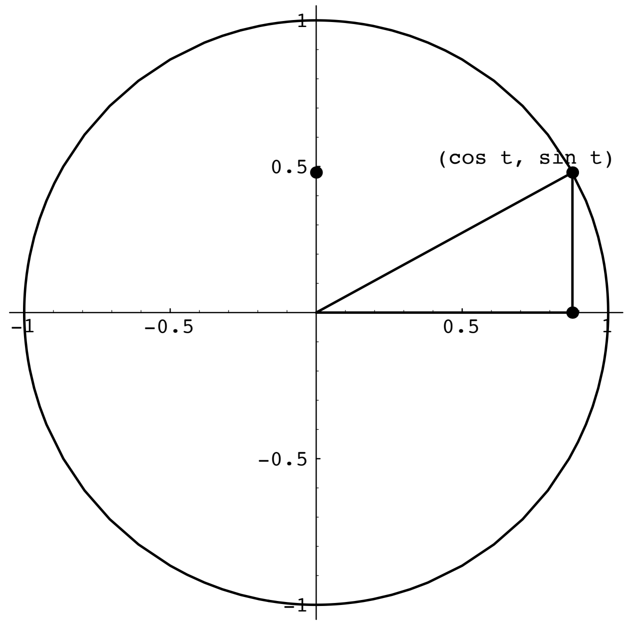

Figure 5.8: The unit circle

wrap around the circle several times before they end at p. In this sense, the radian measure of the angle α is not unique, but any two radian measures of the angle differ by an integer multiple of 2π.

Conversely, let t be any real number. We construct the angle with radian measure t. Starting at the point $(1, 0)$ we travel the distance $|t|$ along the unit circle (here $|t|$ denotes the absolute value of t). By convention, we travel counter clockwise if t is positive and clockwise if t is negative. In this way we reach a point p on the circle. Let α be the angle between the positive x-axis and the ray which starts at the origin and intersects the unit circle in p. This angle has radian measure t.

Comparison of Angles in Degrees and Radians: We suppose that you are familiar with measuring angles in degrees. The measure of half a revolution (a straight angle) comprises π radians and 180 degrees. So, one degree corresponds to $\pi/180$ radians or approximately 0.017453293 radians, and one radian corresponds to $180/\pi$ degrees, or about 57.29577951 degrees. More generally we have the conversion formula

$$x \text{ degrees} = \frac{\pi}{180}\, x \text{ radians}. \tag{5.15}$$

Trigonometric Functions: Let t be once more a real number. Starting at the point $(1, 0)$ we travel the distance $|t|$ along the unit circle (here $|t|$ denotes the absolute value of t). By convention, we travel counter clockwise if t is positive and clockwise if t is negative. In this way we reach a point $p = (x(t), y(t))$ on the circle, and we set

$$x(t) = \cos t \quad \text{and} \quad y(t) = \sin t. \tag{5.16}$$

This defines the functions $\sin t$ and $\cos t$. You see the construction implemented in Figure 5.8. We used $t = .5$. The dot on the x-axis indicates the x-coordinate $x(t) = \cos(.5)$ of the point p. The dot on the y-axis indicates its y-coordinate $y(t) = \sin(.5)$. The approximate numerical values are

$$\sin(.5) = 0.479426 \quad \text{and} \quad \cos(.5) = 0.877583.$$

You can find the graphs of the sine and cosine functions on the interval $[0, 2\pi]$ in Figures 5.9 and 5.10.

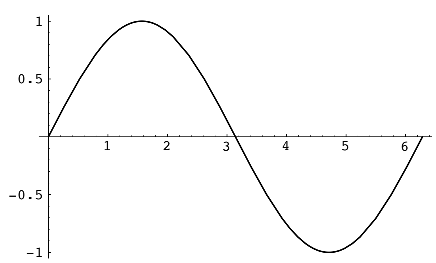

Figure 5.9: $f(x) = \sin x$

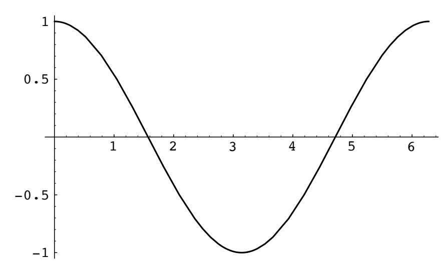

Figure 5.10: $f(x) = \cos x$

The other trigonometric functions, tangent (tan), cotangent (cot), secant (sec), and cosecant (csc) are defined as follows:

$$\tan x = \frac{\sin x}{\cos x} \quad \cot x = \frac{\cos x}{\sin x} \quad \sec x = \frac{1}{\cos x} \quad \csc x = \frac{1}{\sin x} \tag{5.17}$$

To make sure you have some idea about the behavior of the tangent and cotangent function we provided two graphs for each of them. They are drawn over different parts of the domain to show different aspects. See Figure 5.11

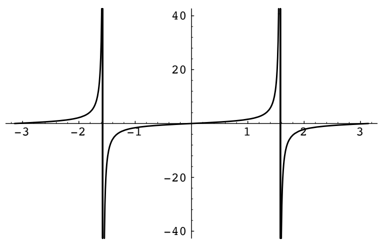

Figure 5.11: $\tan x$ on $[-\pi, \pi]$

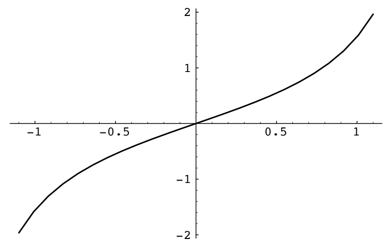

Figure 5.12: $\tan x$ on $[-1.1, 1.1]$

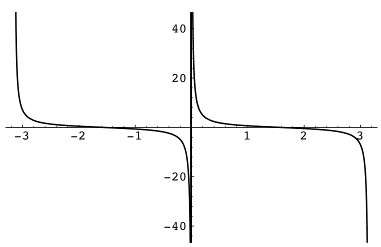

Figure 5.13: $\cot x$ on $[-\pi, \pi]$

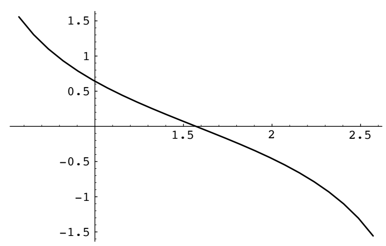

Figure 5.14: $\cot x$ on $[\frac{\pi}{2} - 1, \frac{\pi}{2} + 1]$

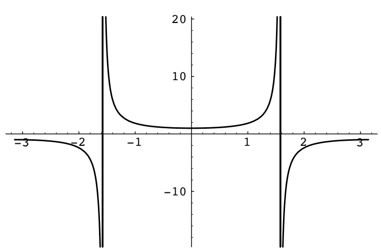

Figure 5.15: $\sec x$ on $[-\pi, \pi]$

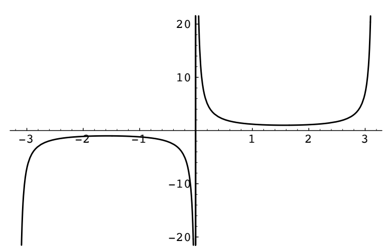

Figure 5.16: $\csc x$ on $[-\pi, \pi]$

to Figure 5.14. You can see the graphs of the secant and cosecant functions in Figure 5.15 and 5.16.

A small table with angles given in degrees and radians, as well as the associated values for the trigonometric functions is given in Table 5.3. If the functions are not defined at some point, then this is indicated by *'n/a'*. Older calculus books may still contain tables with the values of the trigonometric functions, and there are books which were published for the specific purpose of providing these tables. This is really not necessary anymore because any scientific calculator gives those values to you with rather good accuracy.

degrees	radians	$\sin x$	$\cos x$	$\tan x$	$\cot x$	$\sec x$	$\csc x$
0	0	0	1	0	n/a	1	n/a
30	$\pi/6$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{3}$	$\sqrt{3}$	$\frac{2\sqrt{3}}{3}$	2
45	$\pi/4$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	1	1	$\sqrt{2}$	$\sqrt{2}$
60	$\pi/3$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$	$\frac{\sqrt{3}}{3}$	2	$\frac{2\sqrt{3}}{3}$
90	$\pi/2$	1	0	n/a	0	1	n/a
120	$2\pi/3$	$\frac{\sqrt{3}}{2}$	$-\frac{1}{2}$	$-\sqrt{3}$	$-\frac{\sqrt{3}}{3}$	-2	$\frac{2\sqrt{3}}{3}$
135	$3\pi/4$	$\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$	-1	-1	$-\sqrt{2}$	$\sqrt{2}$
150	$5\pi/6$	$\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{3}}{3}$	$-\sqrt{3}$	$-\frac{2\sqrt{3}}{3}$	2
180	π	0	-1	0	n/a	-1	n/a

Table 5.3: Values of Trigonometric Functions

Trigonometric Functions defined at a right triangle: Occasionally it is more convenient to use a right triangle to define the trigonometric functions. To do this we return to Figure 5.8. You see a right triangle with vertices $(0,0)$, $(x,0)$ and (x,y). We may use a circle of any radius r. The right angle is at the vertex $(x,0)$ and the hypotenuse has length r. Let α be the angle at the vertex $(0,0)$. In the following the words adjacent and opposing are in relation to α. Then

$$\begin{aligned} \sin\alpha &= \frac{\text{opposing side}}{\text{hypothenuse}} & \cos\alpha &= \frac{\text{adjacent side}}{\text{hypothenuse}} \\ \tan\alpha &= \frac{\text{opposing side}}{\text{adjacent side}} & \cot\alpha &= \frac{\text{adjacent side}}{\text{opposing side}} \\ \sec\alpha &= \frac{\text{hypothenuse}}{\text{adjacent side}} & \csc\alpha &= \frac{\text{hypothenuse}}{\text{opposing side}} \end{aligned}$$

Trigonometric Identities: There are several important identities for the trigonometric functions. Some of them you should know, others you should be aware of, so that you can look them up whenever needed. From the theorem of Pythagoras and the definitions you obtain

$$\sin^2 x + \cos^2 x = 1, \quad \sec^2 x = 1 + \tan^2 x, \quad \csc^2 x = 1 + \cot^2 x. \tag{5.18}$$

The following identities are obtained from elementary geometric observations using the unit circle.

$$\begin{array}{rclclcl} \sin x &=& \sin(x+2\pi) &=& \sin(\pi - x) &=& -\sin(-x) \\ \cos x &=& \cos(x+2\pi) &=& -\cos(\pi - x) &=& \cos(-x) \\ \cos x &=& \sin(x+\frac{\pi}{2}) &=& -\cos(x+\pi) &=& -\sin(x+\frac{3\pi}{2}) \\ \sin x &=& -\cos(x+\frac{\pi}{2}) &=& -\sin(x+\pi) &=& \cos(x+\frac{3\pi}{2}) \end{array}$$

You should have seen, or even derived, the following addition formulas in precalculus.

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta \tag{5.19}$$

$$\sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta \tag{5.20}$$

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \tag{5.21}$$

$$\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta \tag{5.22}$$

$$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta} \tag{5.23}$$

$$\tan(\alpha-\beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta} \tag{5.24}$$

These formulas specialize to the double angle formulas

$$\sin 2\alpha = 2\sin\alpha\cos\alpha \quad \text{and} \quad \cos 2\alpha = \cos^2\alpha - \sin^2\alpha \tag{5.25}$$

From the addition formulas we can also obtain

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta)-\cos(\alpha+\beta)\right] \tag{5.26}$$

$$\sin\alpha\cos\beta = \frac{1}{2}\left[\sin(\alpha-\beta)+\sin(\alpha+\beta)\right] \tag{5.27}$$

$$\cos\alpha\cos\beta = \frac{1}{2}\left[\cos(\alpha-\beta)+\cos(\alpha+\beta)\right] \tag{5.28}$$

which specialize to the the half-angle formulas

$$\sin^2\alpha = \frac{1}{2}\left[1-\cos 2\alpha\right] \quad \text{and} \quad \cos^2\alpha = \frac{1}{2}\left[1+\cos 2\alpha\right] \tag{5.29}$$

Two Estimates: There are two estimates which we use when we find the derivative of the sine and cosine functions.

Theorem 5.14. *If $h \in [-\pi/4, \pi/4]$, then*[2]

$$|1-\cos h| \le \frac{h^2}{2} \quad \textit{and} \quad |h-\sin h| \le \frac{h^2}{2}. \tag{5.30}$$

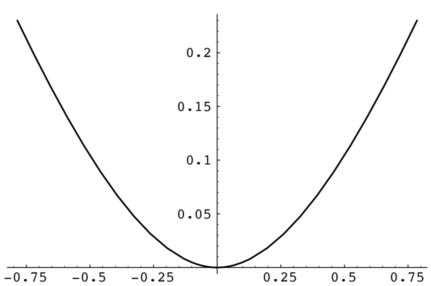

Figure 5.17: $(h^2/2) - |h-\sin h|$

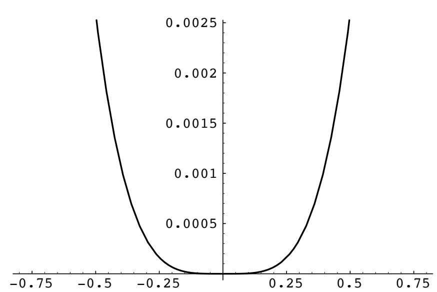

Figure 5.18: $(h^2/2) - |1-\cos h|$

As convincing evidence you may graph the functions $(h^2/2) - |1-\cos h|$ and $(h^2/2) - |h-\sin h|$ using a graphing calculator or a computer and see that both of them are everywhere non-negative. We did so in Figures 5.18

[2]The inequalities hold without the restriction on h, but we only need them on an interval around zero. Restricting ourselves to this interval simplifies the proofs somewhat.

and 5.17. You are encouraged to reproduce these graphs. A mathematician will not accept this as a proof, but ask for a logically conclusive argument. Who knows how the computer or calculator found the graph, and whether it is correct?

In the proof of Theorem 5.14 we use

Theorem 5.15. *If* $h \in [-\pi/4, \pi/4]$*, then*

$$|\sin h| \leq |h| \leq |\tan h|. \tag{5.31}$$

Proof. It suffices to show the desired inequalities for $h \in [0, \pi/4]$ because

$$|\sin(-h)| = |\sin h| \quad \& \quad |\tan(-h)| = |\tan h|.$$

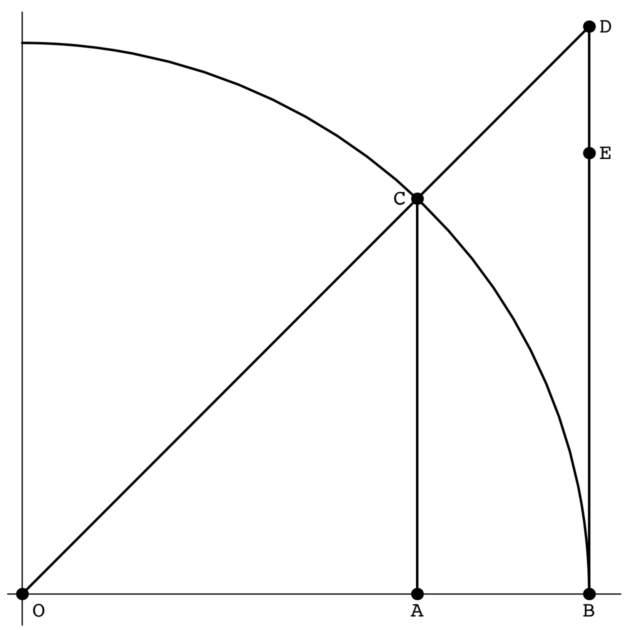

Figure 5.19: The unit circle

In Figure 5.19 you see part of the unit circle. For $h \in [0, \pi/4]$ we obtain a point $(\cos h, \sin h)$ on it. We denote the point by C. For any two points X and Y in the plane, we denote the length of the straight line segment between them by $\overline{XY}$. In addition, let $\widehat{BC}$ be the length of the arc (part of the unit circle) between B and C. Then

$$\overline{AC} = \sin h \quad \& \quad \widehat{BC} = h \quad \& \quad \overline{BD} = \tan h.$$

Evidently, $\overline{AC} \le \widehat{BC}$ because going from C straight down to the x-axis is shorter than following the circle from C to the x-axis. So we find $\sin h \le h$. Secondly, $\widehat{BC} \le \overline{BD}$. To see this, imagine that you roll the circle upwards until the point C crosses the vertical line through B in the point E. We use the process of rolling the circle along the vertical line through B to measure h. In particular, $h = \overline{BE}$. It appears to be clear[3] that $\overline{BE} \le \overline{BD}$. This verifies that $h \le \tan h$, the second inequality which we claimed in the theorem. $\square$

Proof of Theorem 5.14. To see the first estimate in (5.30) we first draw a picture to explain our notation, see Figure 5.20. There you see half of a circle of radius 1 and a triangle with vertices A, B, and C. Let $h \in [-\pi/4, \pi/4]$ be the number for which we want to show the inequality, then C is chosen as $(\cos h, \sin h)$. We drew the point C above the x-axis, which corresponds to a positive choice for h. You may place C below the axis and use a negative value for h. The following argument does not depend on it. Denote by $\overline{XY}$ the length of the straight line segment between the points X and Y. Let $\widehat{BC}$ be the length of the arc (part of the unit circle) between B and C.

From the picture we read off that

$$\overline{AB} = 2, \quad \overline{DB} = (1 - \cos h), \quad \widehat{BC} = |h|, \text{ and } \quad \overline{BC} \le \widehat{BC}.$$

Using similar triangles you see

$$\overline{AB}/\overline{BC} = \overline{BC}/\overline{DB} \quad \text{or} \quad (\overline{BC})^2 = \overline{AB} \times \overline{DB}.$$

In other words

$$2(1 - \cos h) = \overline{AB} \times \overline{DB} = (\overline{BC})^2 \le (\widehat{BC})^2 = h^2.$$

Dividing this inequality by 2 we obtain

$$|1 - \cos h| = 1 - \cos h \le \frac{h^2}{2},$$

the first estimate in (5.30).

[3] Here our argument relies on intuition, and in this sense it is not exactly rigorous. A rigorous argument requires substantial work. In particular, one needs to show that the area of a disk with radius r is πr^2, so the area of the disk with radius one is π. From this is follows by elementary geometry that the area of the slice of the disk with vertices O, B and C has area $h/2$. This slice is contained in the triangle with vertices O, B and D, and the area of the slice is $(\tan h)/2$. It follows that $h \le \tan h$.

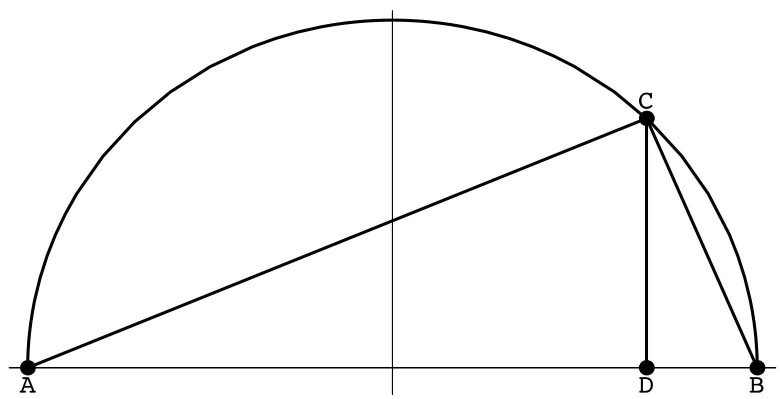

Figure 5.20: The unit circle

Next we want to show that

$$|h - \sin h| \leq \frac{h^2}{2}$$

for $h \in [-\pi/4, \pi/4]$. If $h = 0$ there is nothing to show, both sides of the inequality are zero. So we assume that $h \neq 0$. From Theorem 5.15 we deduce that

$$|\sin h| \leq |h| \leq |\tan h| = \frac{|\sin h|}{\cos h},$$

which implies (using elementary arguments for working with inequalities and that h and $\sin h$ are either both positive or negative) that

$$0 \leq \cos h \leq \frac{\sin h}{h} \leq 1.$$

Subtracting the terms in this inequality from 1 we find

$$0 \le 1 - \frac{\sin h}{h} \le 1 - \cos h \le 1.$$

Using our previous estimate for $|1 - \cos h|$ and a common denominator for one of the expressions we conclude that

$$\left|\frac{h - \sin h}{h}\right| \le |1 - \cos h| \le \frac{h^2}{2}.$$

As we assumed that $|h| \le \pi/4 < 1$, this implies that

$$|h - \sin h| \le \frac{h^2}{2},$$

which is the second inequality which we set out to prove. $\square$

5.6 Inverse Functions

As an instructive example, consider the functions $f(x) = x^2$ and $g(x) = \sqrt{x}$ on the interval $[0, \infty)$. You should be able to recognize their graphs in Figure 5.21.

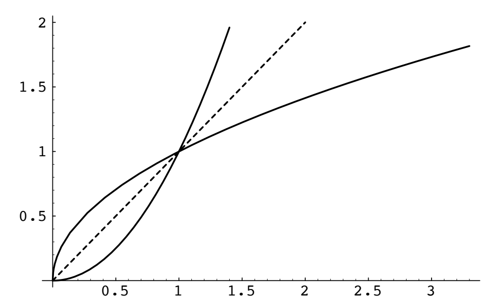

Figure 5.21: A function and its inverse

Clearly,

$$f(g(x)) = x \quad \text{and} \quad g(f(x)) = x \quad \text{for all } x \ge 0.$$

Observe that the graph of $g(x)$ is obtained from that of $f(x)$ by reflection at the diagonal. This should be clear from the picture shown in Figure 5.21.

Let us give an example which demonstrates that you are sometimes interested in the inverse of a function which is given to you.

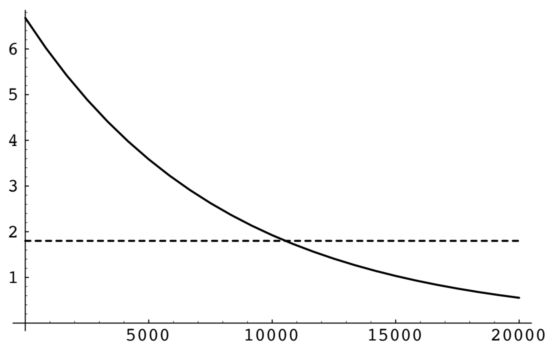

Figure 5.22: Chart for Carbon-14 Dating

Example 5.16. Suppose, you are an archaeologist and got lucky. On your recent dig you found a burial chamber, the mummy still inside, and with it a piece of wood. Well, you need to find out when the person in this grave was buried. Your lab assistant, who is scientifically trained but may not have the same intuition as you necessary to find the right place to dig, offers her help. She takes the wood to the lab and returns after an hour to report that she measured 1.8 carbon-14 disintegrations per minute and per gram of wood. You, the archaeologist, may not understand the method of carbon-14 dating. As help, the assistant offers a little chart (see Figure 5.22). As a function of time (age), it shows the number of decays (per gram and minute) expected in a sample. You are not interested in the number of decays as a function of age, but the age as a function of the number of decays. Fortunately, these two functions are inverse to each other, and you can still use the graph to figure out the age of the piece of wood. You just interchange the axes of the graph to get the graph of the function you like to read off.

To be really practical, start out at the value 1.8 on the vertical axis. Draw a horizontal line through this point (indicated as a dashed line), and find the intersection point with the graph. The x-coordinate of the intersection point tells you the ago of the piece of wood.

Well, you expect that this is also the age of the mummy. At this moment you decide to consult your thesis advisor. The find would be rather spectacular because the mummy would be a lot older than any mummy found previously and your announcement of the discovery could make you world famous or the laughing stock of the entire scientific community. $\diamondsuit$

Let us now consider the formal definition of the inverse of a function.

Definition 5.17. *Consider two functions f and g, and suppose that the range of f is equal to the domain of g, and the range of g is equal to the domain of f. We say that g is the inverse of f if*

$$\begin{aligned} g(f(x)) &= x \quad \textit{for all } x \textit{ in the domain of } f \textit{ and} \\ f(g(x)) &= x \quad \textit{for all } x \textit{ in the domain of } g. \end{aligned}$$

If g is the inverse of f, then f is also the inverse of g, and we say that f and g are inverses of each other. It is common to denote the inverse of f by f^{-1}, and we will follow this convention.

There is a possible conflict of notation. In some context it is tempting to use the symbol $f^{-1}(x)$ to denote $1/f(x)$ in analogy to the symbol $f^2(x)$ for $(f(x))^2$. We hope to express ourselves clearly enough so that no confusion arises.

Example 5.18. The function $f(x) = 1/x$ defined on $(0, \infty)$ is its own inverse, we use $(0, \infty)$ as the range. This is apparent from the calculation

$$f(f(x)) = f(1/x) = \frac{1}{1/x} = x. \quad \diamondsuit$$

It is worthwhile to study properties of a function which are related to the existence of an inverse for this function.

Definition 5.19. *A function f is said to be 1-1 if*

$$x_1 \neq x_2 \textit{ implies } f(x_1) \neq f(x_2) \textit{ for all } x_1 \textit{ and } x_2 \textit{ in the domain of } f.$$

Equivalently, we could say that $f(x_1) = f(x_2)$ implies $x_1 = x_2$.

In geometric terms being 1-1 means that a horizontal line intersects the graph of the function in at most one point.

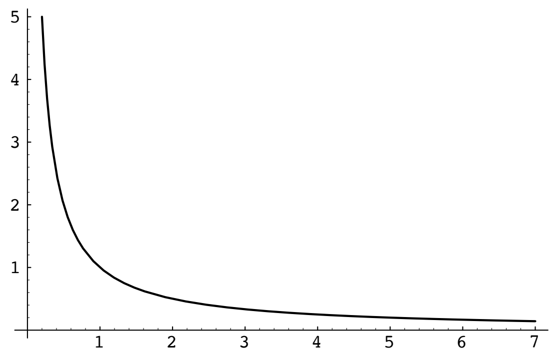

Figure 5.23: $f(x) = 1/x$ for $x \in [.2, 7]$

Example 5.20. The function $f(x) = 1/x$ (defined for all $x \neq 0$) is 1-1. Whenever $x_1 \neq x_2$, then $1/x_1 \neq 1/x_2$. You see part of the graph in Figure 5.23.

On the other hand, $g(x) = x^2$ (defined on $(-\infty, \infty)$) is not 1-1 because $g(1) = g(-1)$ although $1 \neq -1$. If we use $[0, \infty)$ as the domain for g, then the function will be 1-1. The graph of this function and its inverse are shown in Figure 5.21. $\Diamond$

Definition 5.21. *A function f with domain A and range B is said to be* onto *if for all $b \in B$ there exists an element $a \in A$ such that $f(a) = b$.*

Let us express the condition geometrically. If b is in the range of f, then the horizontal line through b must intersect the graph of the function in at least one point. Using the idea of the image of a function from Section 5.3, a function is onto if and only if its range is its image.

Remark 31. Suppose f is a function from A to B. Apparently we get a function which is onto if we use the image of f as the domain, instead of B. Typically, we even use the same name for this function with the reduced domain.

Example 5.22. The function displayed in Figure 5.22 is onto the interval $(0, 6.68]$. You would need to draw more of the graph, continuing it to the right, to see that a horizontal line through small values on the vertical axis, say $y = .1$, intersect the graph.

The function $f(x) = 1/x$ with domain $(0, \infty)$ and range $(0, \infty)$ is onto. To see this, observe that $f(1/b) = b$ for any positive number b and that $1/b$ is a positive number. See also Figure 5.23. $\diamondsuit$

Proposition 5.23. *A function has an inverse if and only if it is 1-1 and onto.*

Proof. To set up our notation, say f is a function from A to B. Assume first that f has an inverse which we call g. By definition, g is a function from B to A. We need to show that f is 1-1 and onto. Suppose $f(x) = f(y)$ for some x and y in A. Then we have $x = g(f(x)) = g(f(y)) = y$ because g is the inverse of f. So $x = y$, and this means that f is 1-1. Next, take any b in B. Then $f(g(b)) = b$ and we found an element in A, namely the element $g(b)$, which is mapped by f to b. This means that the function f is onto. This concludes the proof of one direction of our claim.

Next suppose that f is 1-1 and onto. We need to find a function g from B to A which is an inverse of f. Let b be an element in B. Because f is onto, there exists an element x in A such that $f(x) = b$. In fact, there is only one element in A with this property because f is 1-1. We obtain a well defined assignment setting $g(b) = x$. Using this approach for all elements b in B we define the function g from B to A. Combining the equations $f(x) = b$ and $g(b) = x$ we find that $g(f(x)) = g(b) = x$ and $f(g(b)) = f(x) = b$. But this just means that g is the inverse of f. This completes the proof of the other direction of our claim. $\square$

We like to have some good criteria which imply that a function is 1-1. Consider a function f which is defined on a subset of the real numbers and has values in the real numbers. Recall

Definition 5.24. *A function f is said to be* increasing *if, for all x_1 and x_2 in the domain of f,*

$$f(x_1) < f(x_2) \quad \textit{whenever } x_1 < x_2,$$

and decreasing *if*

$$f(x_1) > f(x_2) \quad \textit{whenever } x_1 < x_2.$$

A function is said to be monotonic if it is either increasing, or if it is decreasing.

Exercise 165. Show: Monotonic functions are 1-1.

E.g., the functions $f(x) = x^2$ and $f(x) = \sqrt{x}$ are increasing on the interval $[0, \infty)$. The function $f(x) = 1/x$ is decreasing on $(-\infty, 0)$ and on $(0, \infty)$, but it is not decreasing on the union of the interval $(-\infty, 0)$ and $(0, \infty)$. You are invited to graph these functions to get an assurance for these claims.

We like to make one more observation in this context. Consider again a function f which is defined on a subset of the real numbers and has values in the real numbers.

Proposition 5.25. *If f is increasing (decreasing) and f has an inverse, then the inverse is also increasing (decreasing).*

You should convince yourself that this observation is true for the examples of increasing and decreasing functions which we just gave. You are also invited to prove this proposition. Let us add another example.

Example 5.26. In Figure 5.24 you see the graph of $f(x) = \sin x$. This time we use $[-\pi/2, \pi/2]$ as domain and $[-1, 1]$ as range. A look at the graph, or a geometric argument at the unit circle, will convince you that f is 1-1, onto, and increasing. In particular, with this domain and range $\sin x$ has an inverse. In Figure 5.25 you see the graph of this inverse function. It is called arcsin. It should be apparent from the graph that arcsin with the domain $[-1, 1]$ and range $[-\pi/2, \pi/2]$ is also 1-1, onto, and increasing.

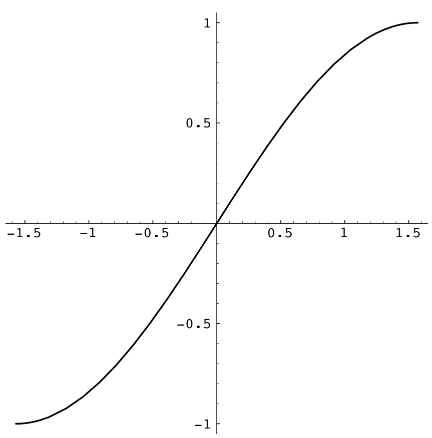

Figure 5.24: sin on $[-\pi/2, \pi/2]$

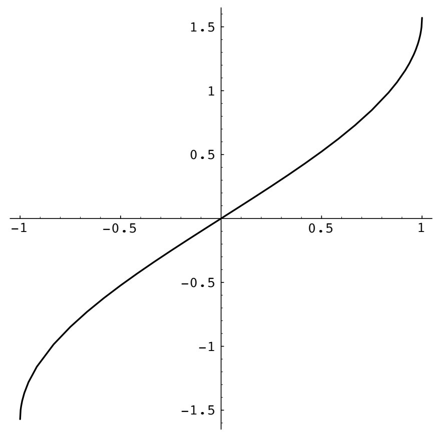

Figure 5.25: arcsin on $[-1, 1]$

5.7 New Functions From Old Ones

We describe some standard methods for constructing new functions from given ones and give examples.

Let f and g be functions with the same domain D, and let c be a constant (a real number). We define the sum $f+g$, the product fg, the quotient f/g, and the scalar product cf by the equations:

$$(f+g)(x) = f(x)+g(x) \tag{5.32}$$
$$(fg)(x) = f(x)g(x) \tag{5.33}$$
$$(f/g)(x) = f(x)/g(x) \tag{5.34}$$
$$(cf)(x) = cf(x). \tag{5.35}$$

The domain for all these functions is again D, except for f/g which is only defined for those $x \in D$ for which $g(x) \neq 0$. E.g., the first equation expresses that we are defining the function $f+g$. For every $x \in D$, we have to tell what $(f+g)(x)$ is. The instructions says that it is the sum of the values of f and g at x. Expressed casually, we defined the sum of functions by adding their values. The other definitions have similar interpretations.

Example 5.27. Let $f(x) = x^2 - 1$, $g(x) = x+1$, and $c = 3$. These function have the entire real line as domain because they are defined for all real numbers. Then we have

$$\begin{array}{ccccc}
(f+g)(x) & = & (x^2-1)+(x+1) & = & x^2+x \\
(fg)(x) & = & (x^2-1)(x+1) & = & x^3+x^2-x-1 \\
(f/g)(x) & = & (x^2-1)/(x+1) & = & x-1 \\
(cf)(x) & = & 3(x^2-1) & = & 3x^2-3
\end{array}$$

The domain for $f+g$, fg, and cf is the real line. The domain for f/g consists of the real numbers except $x=-1$, in spite of the fact that the right hand side in the equation $(f/g)(x) = x-1$ makes sense even for $x=-1$. The reason for this is, that $g(-1) = 0$ and $f(-1)/g(-1)$ is not defined.

To define the *composition* of two functions f and g we need that the domain of f contains the range of g. We use the symbol $f \circ g$ to denote the composition of f and g, and we define this function by setting

$$(f \circ g)(x) = f(g(x)). \tag{5.36}$$

This means that we first apply g to the argument x and then we apply f to $g(x)$.

E.g., if $f(u) = u^2 - 1$ and $g(x) = x + 1$, then

$$(f \circ g)(x) = f(g(x)) = f(x+1) = (x+1)^2 - 1 = x^2 + 2x.$$

If $f(x) = \sin x$ and $g(x) = x^2 + 1$, then we first map x to $x^2 + 1$, and then we take the sine of the result. We get

$$(f \circ g)(x) = \sin(x^2 + 1).$$

Remark 32. We used the symbol "$\circ$" in (5.36) for reasons of clarity, but we will usually avoid it by writing down the right hand side of this equation.

Our final method of constructing new functions from old ones is to take the inverse of a given function. This topic was discussed with several examples in Section 5.6. Another important example of this method is discussed in Section 1.3. Nevertheless, let us give one more example. In Figure 5.26 you see a graph of the cosine function, where we used $[0, \pi]$ as domain. We also specify the range as $[-1, 1]$. A look at the graph, or a geometric argument at the unit circle, will convince you that $f(x) = \cos(x)$ is 1-1, onto, and decreasing. In particular, with this domain and range $\cos x$ has an inverse. In Figure 5.27 you see the graph of this inverse function. It is called the arccosine function, and the mathematical abbreviation for it is arccos. It is the new function which we obtained from the cosine function (with the specified domain and range) by taking its inverse.

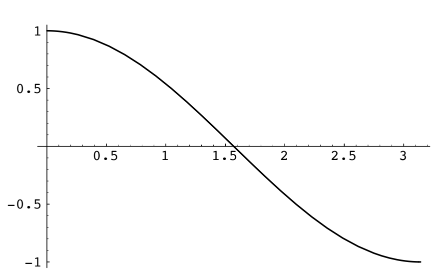

Figure 5.26: cos on $[0, \pi]$

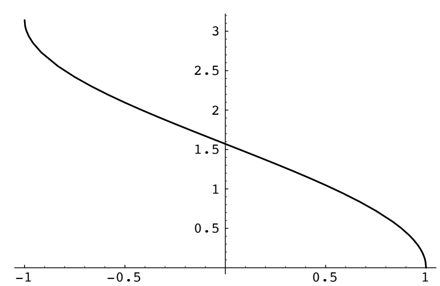

Figure 5.27: arccos on $[-1, 1]$