# Limits in Calculus

A Tutorial Introduction

Santo D'Agostino

October 14, 2024

This book is dedicated to my parents, Francesco and Lucia D'Agostino, and to my wife's parents, Matteo and Concetta Cirocco. They were all wonderful learners, and I learned a lot from each of them.

Copyright ©2024 Santo D'Agostino

# Preface

This book introduces students to the most important fundamental ideas of calculus, starting with the idea of a limit, with thorough, tutorial-style explanations, detailed step-by-step examples, and then continues to develop other important concepts. The book is directed to high-school students to help them prepare for university calculus courses. University courses are very fast-paced compared to high-school courses, and many university students have difficulty understanding the limit concept in the short time available and at the fast pace of university courses. Understanding limits before arriving at university will make your experience much better if you will be taking a university calculus course.

This book may also be helpful to students who are already attending a university calculus course, as an additional detailed, foundational resource. No previous knowledge of calculus is assumed in this book, although it is expected that students have a good, high-school level understanding of functions. Besides containing many basic exercises, the book also contains numerous challenging exercises and opportunities for digging more deeply into the subject, so although it is intended to explain the basic concepts clearly and thoroughly, it can also serve as a source of enrichment material.

The idea of a limit is subtle, and it took the top mathematicians in the world over 200 years to fully understand the idea. It's no wonder, then, that students have difficulty absorbing the idea of a limit in calculus in a short time! One of the features of this book is that some of the history of the development of the limit concept is briefly discussed. I hope this will help readers understand the limit concept more deeply. I also hope that reading a bit about the history of the development of the limit concept will encourage students to persist in working at understanding limits, while highlighting the difficulties that our great ancestors had when grappling with the foundations of calculus.

This book helps students to understand limits by presenting a fundamental problem — how can one calculate the slope of a curve — and then presenting solutions to this problem that start at a very basic, concrete, numerical level, and gradually increase in generality and sophistication, culminating in the precise modern definition of a limit. At all stages of this development, the text proceeds step-by-step, with numerous, detailed examples integrating numerical, graphical, and symbolic approaches. The text is written in plain language, and is tutorial-style, in the sense that it anticipates questions students commonly ask and discusses the questions in a conversational style. Other aspects of limits are also discussed in detail.

A discussion of the concept of continuity is included, as well as some important tools based on continuity, such as the intermediate value theorem. The book also discusses the basic ideas of differential calculus, starting with the definition of the derivative, and including the basic connections between rate of change, derivative, and slope. Numerous graphs are included, to help make clear the interplay between graphical and symbolic (algebraic) perspectives on calculus, and also incorporating some physical ideas.

A recurring theme in this book is infinity, and this concept is thoroughly discussed in several

places in the book, particularly in Section 8.3. Misconceptions about infinity abound, and this book aims to provide students with healthy and accurate ways of thinking about infinity, which are applicable in calculus and elsewhere in mathematics.

The conversational style of the book, and the informality of the presentation helps students to enter the subject and gives them a hand-rail to help them climb the staircase of understanding. But this book goes beyond the elementary, gradually building up to the precise definition of the limit, which is first presented concretely, with diagrams explained in detail, and followed up with numerous fully worked out examples of calculating limits using the precise definition. These step-by-step examples are accompanied by descriptions of the thinking process that are meant to demystify what is typically a very challenging process for first-year university students. A sufficient number of proofs of limits using the precise definition of the limit are presented so that students who work through all of them will develop a good understanding of the mechanics and spirit of such proofs.

Various key theorems about limits and derivatives are also proved, again in a step-by-step way, to make them as accessible as possible. One of the main features of this book is that proofs of theorems are left to the last two chapters of the book. Students are presented with numerous examples and discussion so that they can assimilate the core ideas, and only then is their attention directed towards formal proofs of theorems.

The introduction to the book (Chapter 1) provides a concise summary of the basic ideas of calculus, and places them in historical and physical context. Chapter 2 provides a rapid review of the most immediately needed prerequisite skills. In Chapter 3 the concept of a limit is introduced in the context of the fundamental problem of calculating the slope of a curve at a point. Numerous examples are given, culminating in the definition of the derivative in Chapter 4. Limits are applied more generally than just to determining the slope of a curve, and this more general use of limits is discussed in detail starting in Chapter 5. Subsequent chapters deal with one-sided limits, continuity, and asymptotes. Chapter 9 discusses rates of change in the context of position-time graphs and velocity-time graphs for one-dimensional motions.

Chapter 10 discusses infinite sequences and series, and the discussion of the limit concept in this context ties together the discussion of the limit of a function and the limit of a sequence. Chapter 11 introduces the precise definition of a limit, motivates the need for a precise definition, and then discusses many examples of proving that supposed limits are correct using the precise definition of the limit. Chapter 12 states and proves a number of fundamental theorems in calculus. As usual for this book, the proofs are done step-by-step, with everything thoroughly explained, making it an ideal self-study primer for learning how to prove theorems in calculus. Together, the theoretical material in the final two chapters amounts to a quarter of the book.

Starting with Chapter 3, each chapter begins with a brief overview and ends with a brief summary. This book contains many exercises to help readers to understand the concepts by practicing their use. Answers for all exercises are included, so that this book can be most useful for self-study.

Various kinds of colour-coded feature boxes are included to draw readers' attention to important aspects of calculus. Key concepts, definitions, and theorems are highlighted in yellow boxes. Common misconceptions are highlighted in pink boxes labelled "Careful." Historical information is provided throughout the text, and especially in brown boxes labelled "History." A full list of feature boxes is included at the end of this preface.

## How to Use this Book

One of the difficulties many students have throughout their university programs is poor prerequisite algebra skills. For this reason, there are numerous examples and exercises in this book that help

students practice their algebra skills. Working on your algebra skills while you are still in high school, and bringing them up to an excellent level, will benefit you enormously when you reach university.

Reading actively leads to the best learning, and reading actively includes asking questions as you read. Many questions are highlighted in this text, both to anticipate questions that (in my experience) most students have when encountering this subject for the first time, and also to stimulate students to ask their own questions. To this end, keeping your own notebook with all of your questions, and consulting it frequently, will be very helpful for you as you work through this book.

The best way to deeply learn the material in this book is to read through the text carefully and thoughtfully, and work through all of the examples. Note any questions you have or points that you don't understand. Then, on another day, work through the examples again from scratch; that is, read the statement of the problem, and then try to reproduce the full solution by yourself, without looking at the worked-out solution in the text. If you get stuck, peek at the solution just enough to get you unstuck, and then go back to trying on your own. On another day, try again, starting from scratch as before. With enough repetitions you will be able to solve all of the examples by yourself, which will be a good sign that you understand the material, and which will also give you tremendous confidence.[1] If you get really stuck in working out an example, in the sense that you just don't understand something about the reasoning, then go back and read the reasoning before the example. Don't feel bad about anything. Like many things, learning calculus is hard, even for mathematicians, and the key is to allow yourself enough time, work at it repeatedly, and be gentle with yourself. Remember that repetition and review are essential; even the greatest mathematicians and the greatest learners don't understand everything the first time through; everyone uses repetition and review. This is an important point, and is worth repeating. The fact that you need repetition and review is not a sign that you are dumb, because absolutely *everyone* requires repetition and review to learn difficult subjects. And mathematics is a difficult subject.

Each time you see a highlighted question in this book is an invitation for you to slow down, pause, give the question some thought, and only after sufficient thought (and perhaps some work with pencil and paper) move on to continue reading. Pausing and thinking at these times will greatly improve your learning experience, both by helping you to remember what you are reading, and also by practicing a behaviour (active reading by frequently asking questions) that will help you to dramatically improve your learning ability. This will be a tremendous benefit for you when you begin university studies.

Following the guidance of three slogans will help you to become an excellent learner: Daily work, review, and repetition.

- Daily work: Just like athletic sports training, practicing a little bit every day leads to placing your learning in long-term memory, where it will be with you for a lifetime. Cramming, or studying a subject only once per week, is grossly inefficient, as leaving study sessions spaced by a week means we forget everything we learned the previous week. Instead, studying a little bit every day keeps the subject fresh in your mind. This takes some practice to organize yourself and be efficient, but it will accelerate your learning and decrease your stress significantly. Don't neglect rest; daily work means something like working five or six days per week. Make sure to take at least one day off per week to rest and rejuvenate, or you may lose enthusiasm and even suffer "burnout."

[1] Remember, competence breeds confidence.

- Review: Regular reviews of previously learned material will dramatically improve your long-term retention. In other words, don't just move on to new material every day; start work every day by reviewing what you did yesterday, and every week devote a little time to reviewing what you did last week and last month. You can optimize this using what is called spaced repetition, but at the start keep things simple and just briefly review as outlined above.

- Repetition: Finally, repeating the solutions of exercises and problems, starting from scratch every time, and only peeking at the solution if you get stuck, will ensure that you master their solution. This is especially important with difficult exercises, dealing with concepts that you find difficult to understand. Once you can solve these difficult exercises from scratch without looking at the solution, your understanding (and therefore, your confidence) will soar. Don't neglect what you find difficult! Stay with it! Keep at it, and you will be amazed at how soon what was previously difficult is now routine for you.

A very important skill for you to develop to ensure your success in university studies of mathematics is to take responsibility for your own learning. In high school teachers tell you exactly what to do, and students can succeed by simply following instructions. This no longer works in university, because you are expected to be an expert learner and are expected to know how to learn and what to do to succeed. Start working on becoming an independent learner now, before you arrive at university. You can do this by practicing every aspect of active learning. One way to do this is to construct your own examples and play with them. Don't just rely on exercises in the textbook (do them, of course, but don't restrict yourself to just them), but think of your own examples. Work on them with friends, and discover means to test and check your own examples. More specific advice along these lines is given later in this book.

If this book will be revised, you can always find the latest version posted at `https://fomap.org/`.

Best wishes in learning this subject! Remember that the most effective way to learn is to study a little bit every day, most days of the week, and to review what you have learned frequently. Practice all of the key processes and ideas frequently (solving the same examples repeatedly from scratch), and you will soon transfer them into your long-term memory, where you will have them for life.

## A Few Words of Encouragement

I first learned calculus from the book *Calculus Made Easy*, by Silvanus P. Thompson, the first edition of which was written in 1914. I borrowed a copy of the book from the Welland Public Library in the summer of 1974, and it was sufficiently well-written that I was able to learn quite a bit from it on my own. That fall I enrolled in a calculus course at my high school, and the teacher (coincidentally also named Thompson) was a good one, and so I was able to learn more about calculus. At university I had a number of really excellent teachers, most notably Peter Taylor, who was quite inspiring. I was lucky to have such good teachers, and then I was lucky again to have the opportunity to teach calculus for many years, which allowed me to continue learning more about the subject. So my first words of encouragement are to seek out good teachers, and to make good use of them. They are only too happy to pass on their learning and help you understand.

The epigraph of Thompson's friendly book is what he calls an ancient Simian proverb:

> What one fool can do, another can.

This is meant to be encouraging as well. When we first begin to learn a subject, it's difficult, and mathematics in particular is hard for everyone, including mathematicians! It's just that they know very well how hard the subject is, and they are willing to put the time in to learn it and learn it well. It doesn't help that many of us, for whatever reason, think that we are so much dumber than everyone else. So that's the second piece of encouragement: Don't worry about being stupid, as everyone else is more or less the same as you. What matters is putting in the work, and we can all learn a subject as long as we put in the right work for long enough. So hang in there, do the work, and trust that you too can learn what so many others have learned.

Richard Feynman, one of the great physicists of the 20th century, reiterated Thompson's advice in an interview with a magazine:[2]

> I don't believe in the idea that there are a few peculiar people capable of understanding math, and the rest of the world is normal. Math is a human discovery, and it's no more complicated than humans can understand. I had a calculus book once that said, "What one fool can do, another can." What we've been able to work out about nature may look abstract and threatening to someone who hasn't studied it, but it was fools who did it, and in the next generation, all the fools will understand it.

Now Feynman was a great man, but he was a physicist, not a mathematician. Perhaps we should hear from a mathematician; OK, here is Thomas W. Körner:[3]

> Mathematicians find mathematics hard and are not surprised or dismayed if it takes them a long time and a lot of hard work to understand a piece of mathematics. On the other hand, most of them would agree that the only reason we find mathematics hard is that we are stupid.
>
> The basic ideas of the calculus, like the basic ideas of the rest of mathematics, are easy (how else would a bunch of apes fresh out of the trees be able to find them?), but calculus requires a lot of work to master (after all, we are just a bunch of apes fresh out of the trees).

So, please, hang in there, give yourself lots of time, start learning calculus before you get to university, and do the work. Calculus is hard for everyone, even professional mathematicians, even top scientists. So let's not get down on ourselves if we find it hard too. Let's be gentle on ourselves, find friendly study partners who are willing to travel this road with us,[4] find good teachers (in person, in books, or both), and let us take the time to work long and consistently, a little bit every day, and enjoy the growth in our understanding.

Remember the slogans *daily work*, *review*, and *repetition*, and you will be amazed at how much you can learn over time, growing a little bit every day. Looking back after your consistent work brings you a little bit more understanding each day, you will be deeply satisfied, and motivated to keep at it so that you can learn a little bit more! This will help keep you going through the inevitable periods of frustration that we all go through when we are engaged in the process of learning something new, significant, and therefore challenging.

---

[2] Reprinted on Page 194 in the book *The Pleasures of Finding Things Out*, by Richard P. Feynman, Perseus Publishing, 1999.

[3] See Page 1 of his *Calculus for the Ambitious*, Cambridge University Press, 2014.

[4] If you can't find any study partners locally, you may find them between the pages of a book written recently or long ago.

# List of Feature Boxes

## Careful!



## Challenge Problem



## Definition

### Digging Deeper



### Excursion



### Good Question



### Good Thinking Habit



### History

### Key Concept



### Levity



### Making Connections



### Play!



### Theorems

### Tricks of the Trade

# Contents

# Chapter 1

# Introduction

What is calculus and what is it used for? Calculus includes an enormous number of ideas, methods, and applications, and this section is an attempt to provide a brief overview.

Most of the interesting phenomena that are analyzed scientifically involve change. The flow of wind and water, the orbits of the planets, the path of a baseball, the movement of a shark, the decay of a pile of leaves, the digestion of food, the growth of a child — all are situations involving change. In any change situation, one of the most important questions is: What is the rate of change? That is, how fast is the change occurring? One aspect of calculus (differential calculus) addresses such questions.

Most quantities of interest in science are modelled by continuous functions. Differential calculus can be considered to be a tool for analyzing continuous functions. Besides the class of applications mentioned in the previous paragraph, calculus is therefore also a fundamental tool in pure mathematics.

The graph of a function provides a very useful visual representation of the function. For a straight-line graph of a function of time, the slope of the graph represents the rate of change of the quantity modelled by the function. Thus, the slope of a graph is of key importance in applications. However, how does one determine the slope of the graph of a function that is not a straight line? This is the key question addressed by differential calculus.

Thus, we have a connected set of concepts: Rate of change is a scientifically useful quantity, which is related to the slope of the graph of the function that models the quantity, which can be calculated by an algebraic (i.e., symbolic) procedure. The algebraic procedure results in another function, related to the original function, called the **derivative** of the original function, which contains all the information about the rate of change of the original function. The process by which one obtains the derivative from the original function is called **differentiation**, and is based on the concept of the **limit** of a function.

Typical first-year university calculus textbooks contain over a thousand dense pages, meant to be worked on for a year or two. As I said, the subject of calculus includes an enormous number of ideas, techniques, and applications, and it takes a long time to explain all of them. The aim of this book is considerably more modest, and concentrates on the concept of limit, which is absolutely fundamental in calculus. If you understand the limit concept very well, this should give you a great foundation for learning calculus when you study it at university.

Calculus is a Latin word that means small stone. In ancient times small stones were used as an aid to counting. The word calculate derives from this usage of the word. In medicine, a calculus is a mineral deposit in the body, such as a kidney stone. As mathematics developed, several different systems for calculating various quantities were developed, and these are all called some kind of

calculus, as you will learn if you progress far enough in your mathematics studies. What we now call calculus used to be called *infinitesimal calculus* as a way to distinguish it from other systems of calculation. However, over time, laziness has resulted in "infinitesimal" being dropped, and so now this subject is universally known as calculus. As you learn about limits at the very beginning of your calculus studies, you will begin to understand what the "infinitesimal" has to do with the subject. We'll discuss this point very soon in this book.

Differential calculus is almost universally learned first, but the more challenging *integral calculus* is even more important and practical in applications. The purpose of differential calculus is to analyze functions to determine their properties, including their rates of change. An important basic purpose of integral calculus is to determine the total accumulated amount of a quantity that gradually changes.

For example, if you have money in your bank account or in an investment that earns interest every month, there is no need to use calculus to determine the total value of your investment at the end of a year. You simply take the initial value of the investment and add the twelve interest quantities that were earned at the end of each month. If every second two drops of water fall from a water tap, after 100 seconds a total of $2 \times 100 = 200$ drops of water will have fallen. No calculus is needed.

However, consider a ball that you drop from the top of a tall building. The ball will *gradually* pick up speed at a certain rate. What is the ball's speed after 2.4 seconds? Well, that's a more challenging problem than the investment and droplet problems in the previous paragraph, which involved *discrete* changes. The falling ball involves *continuous* changes; the ball's speed increases gradually and continuously. This is a problem for integral calculus, although in the simple situation of no air resistance, one can calculate the result readily without using the full power of integral calculus.

The acceleration function of the ball is the derivative of the ball's velocity function. We can rely on various experiments to help us determine the acceleration function of the ball. What we seek is the value of the velocity function of the ball after 2.4 seconds. In this context, we can refer to the velocity function as the **anti-derivative** of the acceleration function. The process of determining the velocity function from the given acceleration function is called **anti-differentiation**, or, equivalently, **integration**.

So, to summarize the differences between differential and integral calculus:

Differential calculus: You have a function of time that models a quantity. The derivative of the function tells you the rate of change of the quantity.

Integral calculus: You have a function that models the rate of change of a quantity in time. The anti-derivative of the function tells you the accumulated amount of the quantity after some time has passed.

Phrased in this way, you can see that the basic problems of differential calculus and integral calculus are sort of inverses of each other.[1] In differential calculus you determine the derivative of a function to help you analyze the function. In integral calculus problems, you know the derivative function, and you seek the original function, or at least a certain value of the original function.

Just as the basic problem of differential calculus has a geometric exemplar (calculating the slope of a curve), so does the basic problem of integral calculus, which will now be described. Each human brain has an enormous visual cortex, and so visual representations are a great aid to learning. Therefore many concepts in mathematics are first presented as geometric situations, or other situations that can be readily visualized, so that they will be as memorable as possible, even

[1] In your future university calculus course you will learn about what is called the fundamental theorem of calculus, which makes this relationship between differentiation and integration both explicit and precise.

though their most general applications typically go far beyond the basic geometric situations we first use to explain the concepts.

Back to the exemplar problem of integral calculus: Suppose that you have the graph of a function and you wish to determine the area enclosed by the graph, the horizontal axis, and two vertical line. For example, consider the shaded region in Figure 1.1.

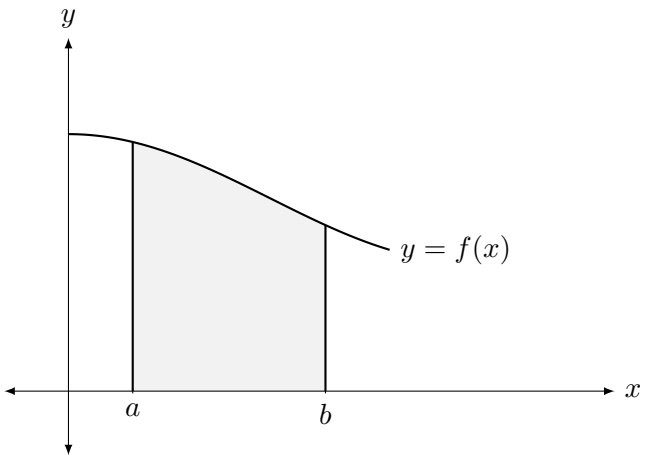



Figure 1.1: Calculating the area of the shaded region is a fundamental kind of problem in integral calculus.

Suppose that the graph represents the acceleration of a falling object plotted against time. It is a fact that the area of the shaded region then represents the change in the object's velocity between the times $a$ and $b$. Thus, the problem of determining the change in the object's velocity has been reduced to a geometric problem, and this kind of graph is useful in developing the basic concepts and techniques of integral calculus. Once the techniques of integral calculus have been understood, they are applicable to other situations beyond the original geometric situation in Figure 1.1.

For example, suppose that you are an engineer designing a propulsion system for sending a space craft to the Moon. You will first have to determine the position of the Moon at all future times. You can do this using Newtonian mechanics, one of the best and most useful physical theories available. In particular, you would use Newton's second law of motion,

$$\mathbf{F} = m\mathbf{a}$$

which can also be written as

$$\mathbf{a} = \frac{\mathbf{F}}{m}$$

In applying Newton's law of motion, you focus your attention on the Moon. You then add up all of the forces acting on the Moon at this moment as vectors (this is the "$\mathbf{F}$" part of the formula), and then divide by the mass of the Moon (this is the "$m$" part of the formula). (The forces acting on the Moon are gravitational forces exerted by the Sun, the Earth, other planets, and so on.) The result is a vector that tells you the magnitude and direction of the acceleration of the Moon right now (this is the "$\mathbf{a}$" part of the formula). Knowing the position and the velocity of the Moon right now, you can then predict what the position of the Moon will be a fraction of a second from now. But by then the positions of the Moon, the Sun, the Earth, and the rest of the planets will have changed, so you will have to re-do the calculation to determine where the Moon will be a fraction of a second later. This is the kind of calculation that you can automate using an electronic computer, but you will have to understand the situation thoroughly so that you can program the computer correctly and then confidently analyze the results to make sure no errors have been made.

In summary, careful measurements can give you the current position and velocity of the Moon, and the positions of all of the other planets. Then you use Newton's law of gravity to determine the magnitudes and directions of all the forces acting on the Moon. Then you add up all of the forces as vectors and then apply Newton's second law of motion to determine the Moon's acceleration. Then you use integral calculus to determine the Moon's subsequent velocity and subsequent position. The process here is like the process of anti-differentiation (integration) as described earlier.

Once you know the Moon's location at all future times (i.e., you know the Moon's position function), you will be able to plan how to propel your space craft so that it will reach the Moon gently and safely at the right time and place.

The same sort of problem-solving procedure occurs in many different fields of science. Newton's second law of motion can be considered to be a *differential equation*, because it is a relation involving a quantity of interest (position), its derivatives (velocity and acceleration), and perhaps other quantities. In the Moon example, it is the position function of the Moon that is of main interest, but Newton's second law of motion involves the second derivative of the position function. One has to differentiate the position function to obtain the velocity function, then differentiate the velocity function to obtain the acceleration function. Two differentiations are needed to go from position function to acceleration function. For this reason, Newton's second law of motion is said to be a second-order differential equation for the position function.

Many of the most important quantities in physics, engineering, and other scientific fields satisfy differential equations, and many of them are second-order differential equations. Thus, gaining a deep understanding of physics, engineering, and many other fields of science, requires an understanding of differential equations and how to solve them. Once you understand integral calculus you will be able to build on this to start tackling differential equations. And understanding differential calculus (which is the subject of this book) is the foundation for understanding integral calculus.

The subject of differential equations can therefore be considered to be an advanced branch of calculus. The usual sequence of learning is that you first learn about "single-variable" calculus; that is, you learn how to apply calculus tools to functions of one variable, which are the kind you can plot on a sheet of paper. Then you can learn about differential equations, and at the same time learn about applying calculus to functions of several variables; the latter is called multi-variable calculus. Multi-variable calculus is applicable to many more complex models of our three-dimensional world than single-variable calculus, but as always in mathematics we start with the more basic subject, understand the more basic subject well, and then build on that understanding by tackling more complex subjects.

In the overall scheme of mathematics, calculus, vector calculus, and differential equations are part of a branch of mathematics called real analysis. There are other branches of analysis, such as complex analysis, functional analysis, numerical analysis, and you might also consider probability and statistics to be in this category as well. Taking a wider view to include other branches of mathematics, the three[2] pillars of mathematics are analysis, topology, algebra, and combinatorics.[3] All of the other numerous branches of mathematics are built upon these pillars. Combinatorics is the systematic study of counting techniques; counting a small number of things can be easy, but counting all possible ways that various things can occur can be very difficult, and ingenious techniques have been dreamed up to cope with these difficulties. Topology is concerned with the shapes of various kinds of geometric objects, and the shapes of various kinds of mathematical spaces, and in particular which kinds of properties of geometrical objects are invariant with respect to continuous deformations. Algebra, broadly speaking, deals with structural matters in mathematics, such as identifying interesting mathematical systems to study, and then abstracting the essential

[2]There are three kinds of people in the world: Those who can count and those who can't count.

[3]A little joke, as combinatorics still doesn't get the respect it deserves from some sectors of the mathematics community, according to some prominent combinatorics practitioners.

properties of the systems so that theorems can be proved about all possible examples that share the same properties. Thus, in advanced algebra, one specifies various systems using axioms (definitions used as starting points), and then one uses logic to prove what one can about all such systems. Algebra is therefore harder for most students to deal with, because of its abstraction, but the rewards are many, for this kind of abstract, structural approach yields many insights that one would not have obtained by studying only concrete examples. However, beginners should learn by careful study of numerous well-chosen examples as a start, and only get into abstractions later.

The same spirit of abstraction obtains in the study of mathematical logic and set theory, which lie at the foundations of mathematics. There are other ways to categorize mathematics (pure mathematics vs. applied mathematics, for example), but thinking in terms of the pillars of analysis, algebra, topology, and combinatorics may be helpful for you as you navigate the vast landscape of mathematics.

Historically, it's interesting to note that one of the fundamental ideas of integral calculus originated with Archimedes. His method of exhaustion, a systematic procedure for approximating the area of a circle using polygons, and improving the accuracy by using polygons with an increasing number of sides in a step-by-step way, was devised over 2200 years ago! Unfortunately, the algebraic and numerical tools had not been developed yet, and the next major advances had to wait until the 1600s, with the work of Fermat, Barrow, and others. The development of analytic geometry (the idea of using coordinate systems and algebra to study geometric figures) by Fermat, Descartes, and others was vital. All of these researchers paved the way for Newton and Leibniz to unify the diverse results of many others into a coherent system in the late 1600s, which was then further developed over the following decades and centuries by numerous other workers.

It's worth noting that Newton developed integral calculus because he desired to solve a specific problem. If you are a budding researcher, it's a good idea to keep a journal with problems that you think of, and jot down ideas for possible solutions. By looking at your problem journal regularly, you can keep them in the forefront of your thinking, and increase the probability of solving them. Every time you learn something new, you can go to your problem journal with fresh eyes, and explore whether your new knowledge can help you to solve your problems. This is what all good researchers do. Newton was busy working out his theory of gravity, and applying it to the Earth, Moon, and the rest of the solar system. In doing so, he initially made the assumption that the gravitational force due to a spherical planet with a density that depends only on the distance from the centre of the planet could be calculated as if all of the mass of the planet were concentrated at its centre. This worked out well, but the assumption displeased him. Could he prove this fact, so that he did not have to assume it? Yes, he certainly did, but he had to invent integral calculus to do so!

The development of calculus did not stop with the work of Newton and Leibniz. As is typical with most mathematical discoveries of this scale, the geniuses who invented the field did not fully understand it. They were able to get their tools to work because of their great thinking power (which includes good intuition), but they did not fully dot every "i" and cross every "t." This was a time of conflict between the rationality of the age of enlightenment and the dogmatism of the church. Many mathematicians and scientists of that time were devoutly religious, but some were openly derisive of religious extremism. Religious leaders were naturally sensitive to the criticism they were receiving from some scientists, and they fought back. Bishop George Berkeley levelled quite a few pointed criticisms at calculus in 1734, to show that these supposedly rational scientists were not reasoning very well at all. The criticisms were quite valid; there were unsolved problems at the foundations of calculus. Newton, Leibniz, and other mathematicians worked with what they called "infinitesimals," but not all of their manoeuvers were well-justified. But as I said, this is the way it always goes; early researchers discover wonderful ideas, and use them to develop powerful methods for solving problems. Sometimes it takes many years before the foundations are tidied up.

Diderot founded his encyclopedia in 1751, a long project attempting to "encircle" all knowledge between the covers of its volumes. His co-founder, d'Alembert wrote many articles for this project, and in his article on calculus he stated that the foundations of calculus had not been clarified yet, but in his opinion the idea of a limit would be fundamental. Cauchy spearheaded a movement to strengthen the proofs of mathematical results in the 1800s, particularly in calculus. His great work *Cours d'Analyse* was published in 1821, and encouraged the spread of a more rigorous approach to mathematical analysis throughout Europe. But even with all of this attention from so many workers for so long, it was not until 1872 that the currently accepted definition of the limit of a function was introduced by Weierstrass, and published by his student Heine. We shall discuss this most precise definition of a limit towards the end of this book. At this point it is sufficient to say that the concept of limit has proven to be fundamental in calculus, and it did indeed tidy up the loose ends that were left dangling by Newton, Leibniz, and their immediate followers.[4]

The discovery of mathematics is not entirely a logical process; it is a creative one. Eventually, the foundations of the subject are strengthened by formulating the newly discovered field as an axiomatic system, with clear definitions and then theorems clearly stated and proved logically. Unfortunately, textbooks are often written in an axiomatic style (that is, emphasizing the internal logic of the subject without providing enough examples and discussion), which is not the best way for most people to learn. The way virtually everyone learns mathematics is by first carefully studying a sufficient number of examples of a new concept, including its applications, and only later tying the examples all together with theorems and their proofs. Keep this in mind when you read mathematics textbooks.

This concludes a brief overview of calculus. The standard way to understand calculus is by beginning with an understanding of the limit concept, and becoming familiar with calculations involving limits. For this reason, we devote this book primarily to the subject of limits in calculus, going step-by-step and with lots of examples and explanations, and a few further developments. This book will give you a strong foundation for learning calculus in university courses, which will go much further than in this book.

[4]The history of the development of calculus is fascinating, and if you would like to read more about it, a good starting point is the Wikipedia page "History of calculus." To dive even deeper, a great source is the book *The Historical Development of the Calculus*, by C.H. Edwards, Jr., Springer, 1994.

# Chapter 2

# Review of Prerequisite Skills

Provided that you can easily work through the following problems, you are well-prepared to tackle this book. If you stumble on any of the problems, it is wise to practice these skills with suitable practice material before beginning to study this book, or at least in the early stages while you study this book. (Answers follow after the questions.)

Nowadays many students who enter university programs that require a bit of mathematics struggle mightily because they lack good algebra skills. The ability to quickly and accurately perform standard, high-school algebraic manipulations is essential for success in university mathematics. If your algebra skills are not top-notch[1], then take the time **now** to bring your algebra skills up to an excellent level, *before* you begin university. You will be so happy that you have done so! University studies are very fast-paced, and there is little free time; trying to patch up any holes in your preparation after you begin university studies will be stressful, and is unlikely to be successful.

Use the questions below as a *starter*[2] in helping you to diagnose your own level, and then work hard now to sharpen your essential prerequisite skills before you arrive at university. See the Suggestions for Further Reading section at the end of this book for suggestions on good sources for practice material to strengthen your essential prerequisite skills to optimally prepare yourself for university mathematics study.

1. Determine the slope of each line.

   (a) The line that passes through the points $(1, 2)$ and $(4, 7)$.
   (b) The line that passes through the points $(-1, -2)$ and $(-4, -7)$.
   (c) The line that passes through the points $(1, -2)$ and $(4, -7)$.
   (d) The line that passes through the points $(-1, 2)$ and $(-4, 7)$.

2. Sketch all four lines from Exercise 1 on the same axes. Notice how the slopes of the lines are related.

3. Determine an equation for each line in Exercise 1.

4. Determine an equation for the line that passes through the point $(-1, 3)$ and

   (a) has slope 2. (b) has slope $-2$. (c) is vertical. (d) is horizontal.

5. Determine which of the following lines has a slope that is

   (a) positive. (b) negative. (c) zero. (d) undefined.

[1] top-top-top-notch

[2] You will need to have a good grounding in all of high-school mathematics, and all of high-school mathematics cannot be reviewed in a few pages. This chapter is the briefest of starters only.

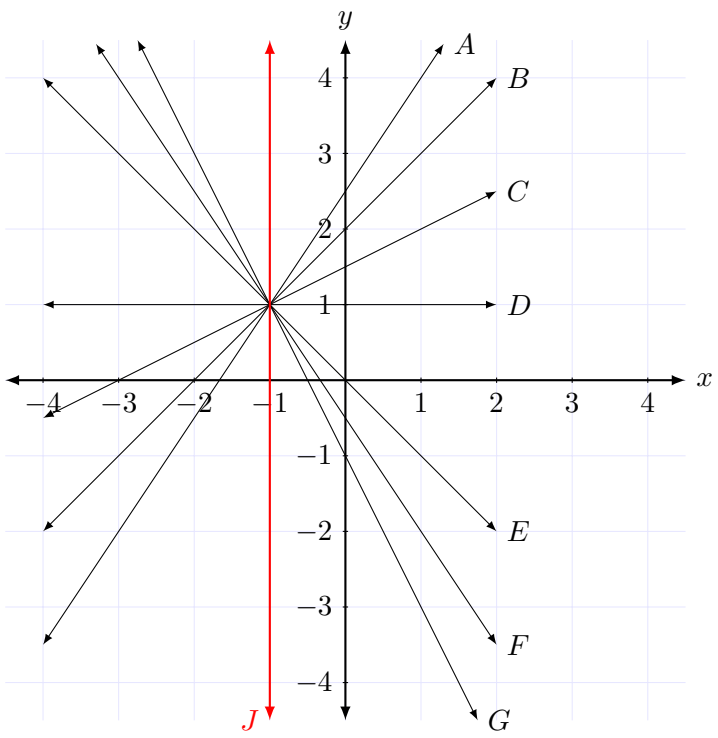



Figure 2.1:

6. Determine the average speed in each case.

   (a) A car travels on a straight road for a distance of 150 km in a time of 2 h.

   (b) A car travels on a winding road for a distance of 80 km in a time of 2 h.

   (c) A car makes a round trip from home to a nearby city 30 km away, and back again, in a total time of 1.5 h.

7. For the function $f(x) = x^2 - 2x + 3$, determine

   (a) $f(4)$ (b) $f(-2)$ (c) $f(a)$ (d) $f(a+h)$ (e) $f(x+1)$

8. Repeat Exercise 7 for each function.

   (i) $f(x) = \dfrac{3x}{x+2}$ (ii) $f(x) = \sqrt{2x+5}$

9. Expand and simplify each algebraic expression. State any restriction on the variables.

   (a) $(x-2)(x+3)$ (b) $(2x-1)(3x+4)$ (c) $(2x+1)(x^2-2x+3)$

   (d) $\dfrac{x^2-x-2}{x^2+x-6}$ (e) $\dfrac{(4+h)^2-4^2}{h}$ (f) $\dfrac{(a+h)^2-a^2}{h}$

10. Factor each expression as completely as possible.

    (a) $2xy + 6xz$ (b) $x^2 - 3x + 2$ (c) $4x^2 - 9$ (d) $x^2 + 1$

    (e) $4x^2 + 4x - 3$ (f) $x^3 - 4x^2 + x + 6$ (g) $8x^3 - 27$ (h) $8x^3 + 27$

11. Determine the domain and range of each function.

    (a) $y = x^2 - 3$ (b) $y = \dfrac{x+2}{x-3}$ (c) $y = \dfrac{x^2-5x+6}{x^2-2x-3}$ (d) $y = \sqrt{2x-5}$

12. Rationalize the denominator of each expression.

(a) $\dfrac{1}{\sqrt{2}}$ (b) $\dfrac{1}{\sqrt{x}-\sqrt{3}}$ (c) $\dfrac{4x}{\sqrt{x+1}+\sqrt{x-2}}$

13. Rationalize the numerator of each expression.

(a) $\dfrac{\sqrt{2x}-\sqrt{5}}{x+1}$ (b) $\dfrac{\sqrt{x+4}+\sqrt{3x-2}}{\sqrt{x+4}-\sqrt{3x-2}}$ (c) $\dfrac{\sqrt{a+h}-\sqrt{a}}{h}$

## ANSWERS

1. (a) $\frac{5}{3}$ (b) $\frac{5}{3}$ (c) $-\frac{5}{3}$ (d) $-\frac{5}{3}$

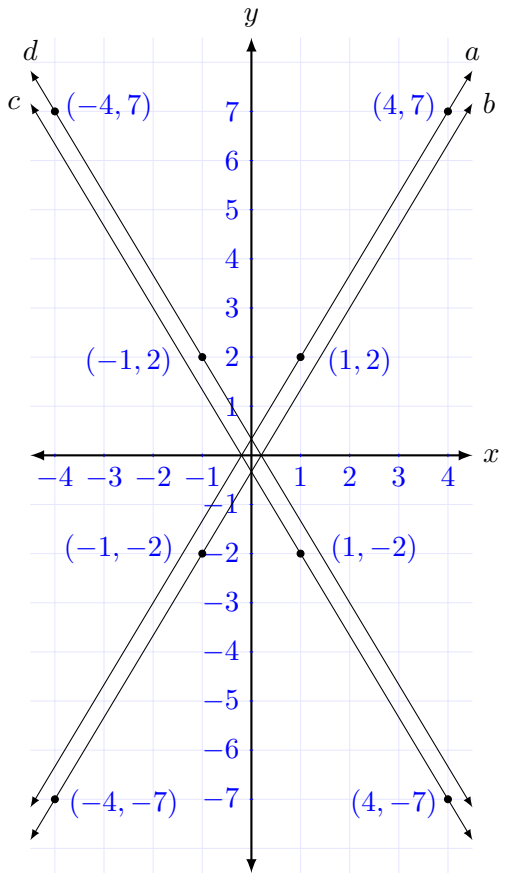



Figure 2.2:

2.

3. (a) $y = \frac{5}{3}x + \frac{1}{3}$ (b) $y = \frac{5}{3}x - \frac{1}{3}$ (c) $y = -\frac{5}{3} - \frac{1}{3}$ (d) $y = -\frac{5}{3} + \frac{1}{3}$

4. (a) $y = 2x + 5$ (b) $y = -2x + 1$ (c) $x = -1$ (d) $y = 3$

5. (a) $A$, $B$, and $C$ (b) $E$, $F$, and $G$ (c) $D$ (d) $J$

6. (a) 75 km/h (b) 40 km/h (c) 40 km/h

7. (a) 11 (b) 11 (c) $a^2 - 2a + 3$ (d) $a^2 + 2ah + h^2 - 2a - 2h + 3$ (e) $x^2 + 2$

8. (i) (a) 2 (b) does not exist (c) $\dfrac{3a}{a+2}$ (d) $\dfrac{3(a+h)}{a+h+2}$ (e) $\dfrac{3(x+1)}{x+3}$ (ii) (a) $\sqrt{13}$ (b) 1 (c) $\sqrt{2a+5}$ (d) $\sqrt{2a+2h+5}$ (e) $\sqrt{2x+7}$

9. (a) $x^2 + x - 6$ (b) $6x^2 + 2x - 4$ (c) $2x^3 - 3x^2 + 4x + 3$ (d) $\dfrac{x+1}{x+3}$; $x \neq 2$ (e) $8 + h$; $h \neq 0$ (f) $2a + h$; $h \neq 0$

10. (a) $2x(y+3z)$ (b) $(x-1)(x-2)$ (c) $(2x-3)(2x+3)$ (d) cannot be factored using only real numbers (e) $(2x+3)(2x-1)$ (f) $(x+1)(x-2)(x-3)$ (g) $(2x-3)(4x^2+6x+9)$ (h) $(2x+3)(4x^2-6x+9)$

11. (a) Domain: $\{x \in \mathbb{R}\}$; Range: $\{y \in \mathbb{R} \,|\, y \geq -3\}$ (b) Domain: $\{x \in \mathbb{R} \,|\, x \neq 3\}$; Range: $\{y \in \mathbb{R} \,|\, y \neq 1\}$ (c) Domain: $\{x \in \mathbb{R} \,|\, x \neq 3\,,\, x \neq -1\}$; Range: $\{y \in \mathbb{R} \,|\, y \neq 1\}$ (d) Domain: $\{x \in \mathbb{R} \,|\, x \geq \frac{5}{2}\}$; Range: $\{y \in \mathbb{R} \,|\, y \geq 0\}$

12. (a) $\dfrac{\sqrt{2}}{2}$ (b) $\dfrac{\sqrt{x}+\sqrt{3}}{x-3}$ (c) $\dfrac{4x\left(\sqrt{x+1}-\sqrt{x-2}\right)}{3}$

13. (a) $\dfrac{2x-5}{(x+1)\left(\sqrt{2x}+\sqrt{5}\right)}$ (b) $\dfrac{-x+3}{2x+1-\sqrt{x+4}\sqrt{3x-2}}$ (c) $\dfrac{1}{\sqrt{a+h}+\sqrt{a}}$

## HISTORY

### Pierre de Fermat (1607–1665) and analytic geometry

It is possible to study geometric figures without using formulas. This approach is known as synthetic geometry, and was pioneered by ancient Greek mathematicians over 2000 years ago, with research and developments still ongoing today. With the help of advances in algebra in the 1500s and 1600s (namely, the beginning of the use of letters to stand for unknown quantities), Pierre Fermat founded the field of analytic geometry. He introduced coördinate axes so that each point on a plane could be described by an ordered pair of real numbers. We now call such a plane a Cartesian plane, in honour of Descartes, or $xy$-plane, because the coördinates are almost universally denoted by $x$ and $y$.

Introducing coördinate axes allows us to describe geometric figures using formulas involving the coördinates $x$ and $y$, and then allows us to determine properties of the geometric figures by analysing the formulas; hence the term analytic geometry. Fermat used his new invention to determine formulas for straight lines, circles, parabolas, ellipses, hyperbolas, and other curves. He also showed that all first-degree and second-degree formulas describe lines, circles, parabolas, ellipses, and hyperbolas.

Fermat made is living as a lawyer and judge, and lived in Toulouse, far from the intellectual centre of France in Paris. He was a shy and modest man, and his passion was mathematics, to which he devoted a lot of time. His mathematical discoveries were communicated to Parisian mathematicians by letter, and his work on analytic geometry, which was done in 1629, was circulated in Paris in 1636, with Fermat's letters passing from one excited mathematician to another, facilitated by Marin Mersenne. René Descartes read Fermat's work, and included it in his book on geometry in 1637, with improved notation. This is perhaps part of the reason why Descartes is often given credit for Fermat's work.

The ancient Greek Mathematician Apollonius of Perga had inklings of Fermat's idea almost 1800 years previously! The Persian mathematician Omar Khayyam made progress in treating geometric figures from an analytic approach in his book *Treatise on Demonstrations of Problems of Algebra*, published in the year 1070. Neither of these greats went as far as Fermat, but both deserve mention, because research is a collective activity, and the researchers of one era stimulate each other and build on the work of their predecessors. Remember this when you think of your own future contributions to the world! Many people do very important and useful work, not just the few greats that you may read about in textbooks.

Fermat produced many other mathematical researches in his lifetime, including the statement of his famous "last" theorem, that the equation $a^n + b^n = c^n$ has natural number solutions only if $n = 1$ or $n = 2$. Fermat claims to have proved this theorem, which he stated in the margin of one of the books in his library, but he claimed that there was not enough space in the margin of the book to write out the proof. Fermat did a lot of his work in the margins of his personal library books, and perhaps he did have a proof. However, the theorem was finally proved publicly in 1994 by Andrew Wiles, after more than 350 years of labour by numerous mathematicians.

Despite the fact that Fermat produced a lot of significant mathematics, for our purposes in this book, the salient facts about Fermat's work is that analytic geometry provided an essential arena for calculus, and that Fermat himself made many developments in both differential and integral calculus, before Newton and Leibniz were born. Newton gave credit to Fermat, saying that his early ideas about calculus came from reading Fermat's works.

To begin reading more about this aspect of Fermat's work, a good source is Section A.13 of *Calculus Gems*, by George F. Simmons, McGraw-Hill, 1992.

# Chapter 3

# Slope and Rate of Change

**OVERVIEW**

The slope of a graph indicates the rate of change of the graphed quantity. This is an important conceptual foundation for understanding and applying calculus.

In this chapter we explore the connection between slope and rate of change.

**WARMUP**

Before you tackle this chapter, make sure you can solve the following exercises. If you have difficulties, please review the appropriate prerequisites.

(Answers at end.)

1. Determine the slope of the line joining the points $(1, -2)$ and $(3, 4)$.
2. Rank the lines in Figure 3.1 in order of increasing slope.
3. Rank the lines in Figure 3.2 in order of increasing slope.

---

Answers: 1. slope $= \dfrac{4-(-2)}{3-1} = \dfrac{6}{2} = 3$; 2. C, B, A; 3. F, E, D.

In Figure 3.1, you can calculate the slope of each line by drawing triangles that have the same base, as in Figure 3.3, and using the "rise-over-run" definition of the slope of a line.

From Figure 3.3, the slope of line $A$ can be calculated using the rise-over-run definition applied to the points $(-1, 1)$ and $(1, 4)$:

$$\text{slope of line } A = \frac{\text{rise}}{\text{run}} = \frac{4-1}{1-(-1)} = \frac{3}{2} = 1.5$$

Thus, the slope of line $A$ is $\dfrac{3}{2}$, which is the same as 1.5. This means that if you imagine the line to be the side of a hill, and you always walk from left to right, then every time you move over to the right by 2 units, you rise 3 units. Equivalently, every time you move over to the right by 1 unit, you rise 1.5 units.

Similarly, the slopes of the other two lines are:

$$\begin{aligned} \text{slope of line } B &= \frac{\text{rise}}{\text{run}} = \frac{3-1}{1-(-1)} = \frac{2}{2} = 1 \\ \text{slope of line } C &= \frac{\text{rise}}{\text{run}} = \frac{2-1}{1-(-1)} = \frac{1}{2} = 0.5 \end{aligned}$$

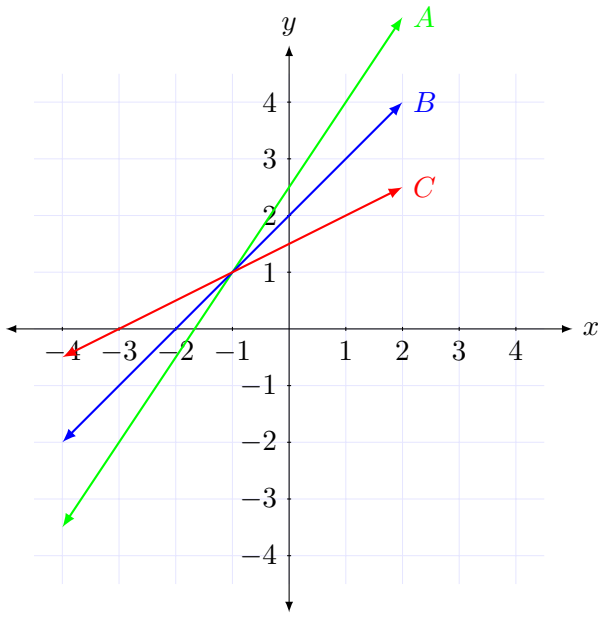



Figure 3.1: Rank the lines in order of increasing slope.

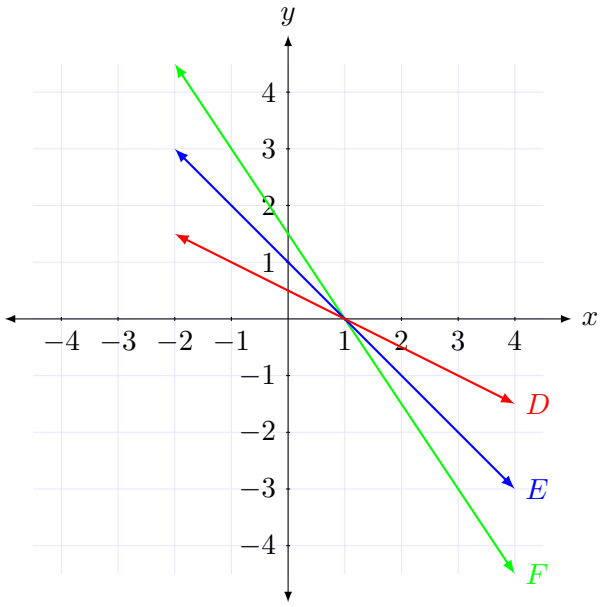



Figure 3.2: Rank the lines in order of increasing slope.

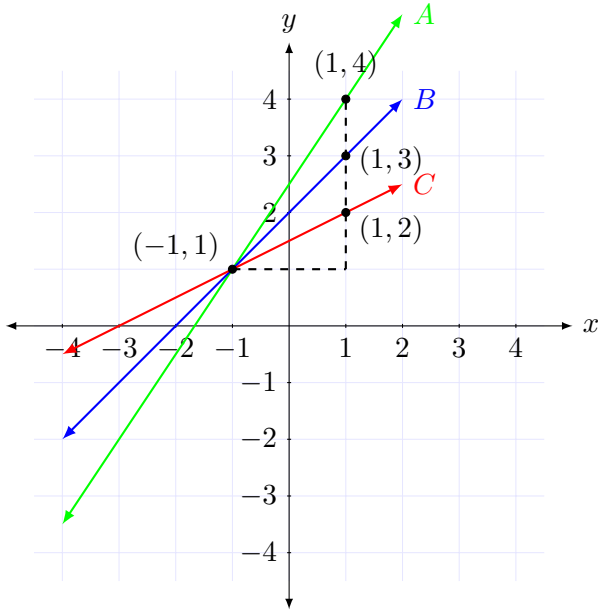



Figure 3.3: Calculate the slope using "rise-over-run."

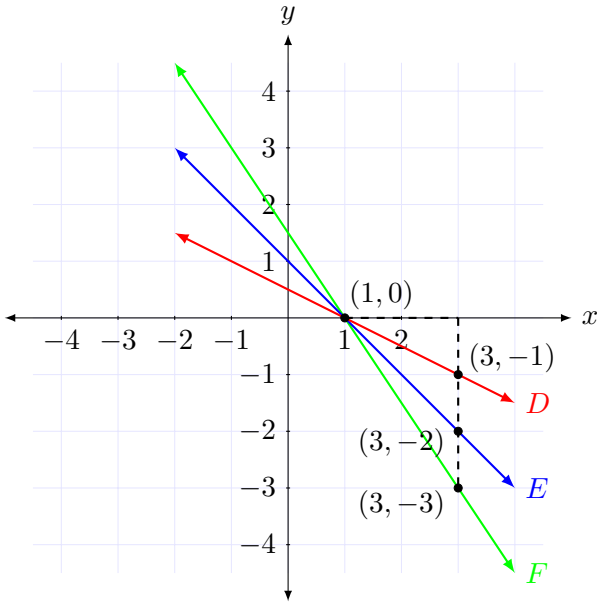



Figure 3.4: Calculate the slope using "rise-over-run."

Thus, in order of increasing slopes, line $C$ has the least slope, followed by line $B$, followed by line $A$ with the greatest slope.

Note that all three of the lines in Figure 3.3 have a positive slope and all the lines in Figure 3.4 have a negative slope. You can see this at a glance by using the hill-climbing analogy for slope: Imagine that each line represents the side of a hill, and you always climb from left to right (because the positive $x$-axis points towards the right). If you climb uphill, the line has a positive slope, and if you climb downhill, the line has a negative slope. Thus all of the lines in Figure 3.3 have a positive slope and all of the lines in Figure 3.4 have a negative slope; check this again by looking at

the graphs. By drawing triangles all with the same base, as we just did for the lines with positive slope, you can calculate the values of the slopes of the lines in Figure 3.4. The result is that $F$ has the smallest slope, $E$ is next, and $D$ has the largest slope.

**Question:** Comparing two lines, does the steeper one necessarily have the greater slope?

Note that for lines with positive slope, the steeper the hill, the larger the slope, which matches our every-day sense of steepness. For lines with negative slope, the opposite is true. This means that more care is needed when ranking the slopes of lines with negative slopes. It may help you to think about temperatures. A temperature of $-2$ is higher than $-3$, which is higher than $-5$. Thus, a slope of $-2$ is greater than a slope of $-3$, which is greater than a slope of $-5$, even though the line with a slope of $-5$ is the steepest of the three.

Also remember that horizontal lines have slope equal to zero, and that it's not possible to represent the slope of a vertical line using a number. We can describe the slope of a vertical line with words (sheer rise, cliff face, etc.) but a numerical value for the slope of a vertical line does not exist. If you apply the rise-over-run definition of the slope of a line to try to calculate the slope of a vertical line, you will end up with an expression that includes division by zero, which is undefined (i.e., makes no sense). Try it yourself to see in detail why the slope of a vertical line cannot be specified by a number.

**Question:** Is infinity a number?

**CAREFUL!**

**Infinity is NOT a number**

One of the common errors made by calculus learners is to consider infinity as a number. True, there are number systems such as the extended real number system in which infinity is successfully treated as a number, and you can ponder on them if you wish, but for our purposes at this level of learning calculus, infinity is decidedly not a number.

Numbers satisfy various properties, and infinity does not satisfy these properties. For example, it might seem reasonable to state that $\infty + 1 = \infty$ (how could this be any different?), but then by the usual properties of numbers, we should be able to subtract $\infty$ from both sides of the equation to obtain $1 = 0$, which is nonsense. This is enough to rule out infinity as a number, but you can have fun deriving all kinds of contradictions based on the assumption that it is a number. (It won't take you long to prove that all real numbers are equal, for example, which is further reinforcement that the assumption that infinity is a number is not valid.)

Be alert to the use of the symbol $\infty$ in various arguments in calculus textbooks for various purposes. Recognize that it is a time-saving symbol that represents various facts or processes, and facilitates stating results in a compact form. While you are striving to understand the facts and processes it represents, remind yourself regularly that infinity, while a very useful concept, is not a number.

### Slope and Rate of Change

The following story illustrates why slope is one of the most important concepts in calculus.

Alice and Basil graduate from university and visit a financial advisor who can accurately predict the near future.[1] The advisor foresees yearly salaries for them from the year 2021 until the year 2025 as shown in Figure 3.5. Who will be better off?

[1] Like all calculus textbooks, this book also has a few improbable situations and unlikely characters.

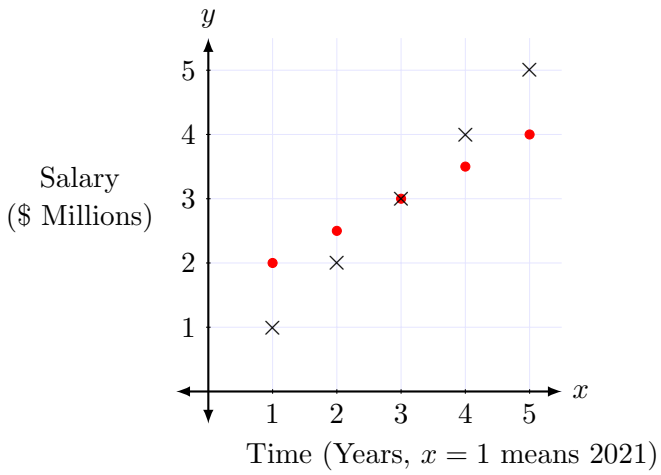



Figure 3.5: Salaries for Alice (×) and Basil (•).

If we continue to use the analogy of a hill to represent a graph, then it is the *height* of the graph (the $y$-coördinate) at a particular $x$-value that represents the salary for that year. Thus, in the year 2021, Basil's salary is \$2 million and Alice's salary is \$1 million. Some might argue that Basil is better off, since in the early years his salary is larger, so he can invest more money sooner and therefore profit more. Others might argue that they are equally well off, since each makes a total of \$15 million over the five years. It's dangerous to *extrapolate* (to guess what happens after the year 2025), but if the trends in Figure 3.5 continue, is it clear that in the long run Alice is better off? Although both salaries are increasing, Basil's is increasing at a rate of \$0.5 million per year, whereas Alice's salary is increasing at a rate of \$1 million per year. Thus, the *rate of change* of Alice's salary is greater than the *rate of change* of Basil's salary.

Note that the slope of Basil's salary graph is 0.5 and the slope of Alice's salary graph is 1. The same conclusion is true of all graphs, and provides the definition of slope:

**KEY CONCEPT**

**Slope**

The slope of a graph is the rate of change of its height.

What is special about the graphs in Figure 3.5? First, the graphs are sets of *discrete* points (a series of separated dots)—this is the kind of graph that results from plotting experimental measurements, and which therefore one encounters frequently in science. Most of the graphs that we'll study in this book are continuous—lines and curves that don't have any breaks in them—because those graphs are more frequently encountered in applications. (For example, in the case of experimental measurements, it is almost always assumed that the quantities being measured are continuous, even though only a few measurements are made. Therefore, the separated dots on the graph are usually joined by a smooth line or curve, and it is the formula for the smooth line or curve that is analyzed.)

The second thing about the graphs in Figure 3.5 that makes them special is that the dots lie on straight lines, so the graphs are *linear*. For a straight line, the slope is the same everywhere on the line, so a single number is enough to specify the slope of a line. That's not the case for a curve, as you can see for example in Figure 3.6. At some points the curve is more steep, at some points less steep, at some points the slope is positive (going uphill if you move from left to right), and at

some points the slope is negative (downhill). So it is clear that it will be impossible to describe the slope of an entire curve with just one number—it's going to be more complicated than that.

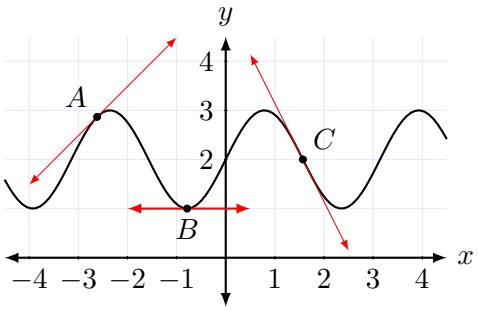



Figure 3.6: The slope of a curved graph is not the same at each point.

**Question:** How can we determine the slope of a curved graph?

But surely we could use a number to describe the slope of a curve at one particular point, wouldn't you say? For example, for the curve in Figure 3.6, if we wanted to know the slope of the curve at the point $A$, we could perhaps sketch a straight line that is the best approximation to the curve at $A$ (this is called the *tangent line* to the curve at the point $A$), and then find the slope of the line. So it seems that the slope of the curve at $A$ is about 1, the slope at $B$ is about 0, and the slope at $C$ is about $-2$. Doesn't this seem reasonable if you are climbing on a curved hillside and wish to describe how steep it is at your particular location?

This method is good for helping us to understand the concept of slope, and it is useful for obtaining a quick estimate of the slope of a curve at a certain point. However, this method is not very satisfying because it depends too much on our drafting skills, because it is not very accurate, and because it doesn't help us to analyze formulas. Our goal is to be able to calculate the slope of a curve *precisely* and efficiently by applying some kind of algebraic procedure to the formula for the curve.

Next we'll begin work towards this goal by describing a method for calculating the slope of a curve at a point. Later in this book we develop more powerful and practical methods of determining the slope of a curve, but to develop such practical methods, and to deeply understand them, we have to start here with the basic method we will now begin to study. This basic method incorporates some of the most important fundamental concepts in calculus.

## 3.1 Calculating the Slope of a Graph at a Point Using a Limit: Numerical and Visual Approach

We'll start by applying an important principle for solving mathematics problems: If you can't calculate a quantity of interest exactly, first approximate it. Then improve the approximation. This idea works best when you can find a *systematic* process to improve the approximation to any desired accuracy. In this way, one can repeatedly apply the process with as many repetitions (called *iterations*) as needed to obtain the desired accuracy.

## KEY CONCEPT

**Systematic iterative approximations**

A key idea in mathematical analysis is to come up with an iterative (i.e., repeated) procedure that systematically approximates a quantity that you are trying to calculate, in such a way that the approximation is improved with each subsequent iteration.

This important idea is at the heart of calculus, and is the essence of the limit concept. It is also a key idea for any professional in mathematics, science, and computer science who needs to compute any quantity. Thus, this basic idea is vital for an enormous number of workers, and is worth your time and attention.

To start off, let's apply this key idea in an attempt to determine the slope of the graph of $f(x) = \dfrac{x^2}{4} + 1$ at the point where $x = 1$. Since we have no idea how to determine the slope of a curve, we'll approximate the curve by a straight line, and use the straight line to estimate the slope of the curve. Then we'll see if we can systematically improve the estimate.

## GOOD THINKING HABIT

**Relating new concepts to ones you already know**

Another good thinking habit that mathematicians use repeatedly is to relate new concepts to ones already understood. In this case, we're trying to understand how to calculate the slope of a curve, so we begin by approximating the curve by a straight line, because we already know how to calculate the slope of a line.

So far we only know how to find the slope of a straight line. So let's approximate the curve near $x = 1$ with a straight line, then calculate the slope of the straight line. We'll choose two points on the curve and use the line joining them as the approximation—such a line is called a *secant line*. (We use two points on the curve because we have a formula for the curve, so we'll be able to find the $y$-coördinates of the points given their $x$-coördinates.) To start off, let's choose the points on the curve that have $x$-coördinates 1 and 3; that is, the points $A(1, 1.25)$ and $B(3, 3.25)$ (see Figure 3.7). The slope of secant line $AB$ is

$$m = \frac{3.25 - 1.25}{3 - 1} = 1$$

Study Figure 3.7 and note that near the point $A(1, 1.25)$ the secant line $AB$ is steeper than the curve.

**Question:** Is this clear? If not, you may have to think about this and play with this situation for a while. For example, imagine that you are walking (from left to right, as always) up the curve, then walking up the secant line; which is steeper near the point $A$?

Thus the slope of the curve at $A$ is less than the slope of the secant line, which is 1. Another way to say this is that the slope of the secant line $AB$ is an *overestimate* for the slope of the curve at $A$.

**Question:** How accurate is this estimate?

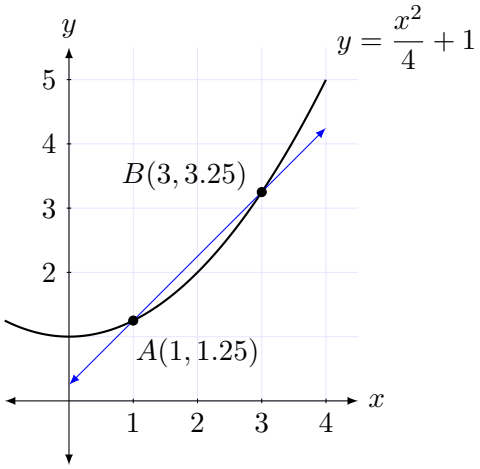



Figure 3.7: The secant line $AB$ gives an overestimate for the slope of the curve at $A$.

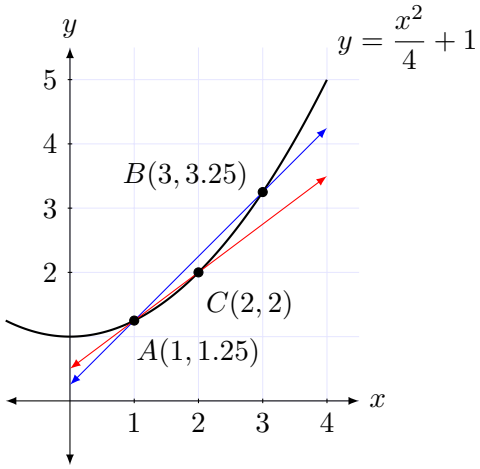



Figure 3.8: The secant line $AC$ also gives an overestimate for the slope of the curve at $A$, but it gives a better estimate than $AB$.

At this point in our thinking process, we have no idea how good this estimate is.[2] An estimate without some sense of its accuracy is of no value.

**Question:** How can we both get a sense for the accuracy of the estimate, and also systematically improve its accuracy?

Suppose we find the slope of the secant line joining the points $A(1, 1.25)$ and $C(2, 2)$. By examining Figure 3.8 and thinking about steepness, note that the slope of secant line $AC$ is less than the slope of secant line $AB$, but still greater than the slope of the curve at $A$.[3] Thus, the slope of secant line $AC$ is a better estimate to the slope of the curve at $A(1, 1.25)$ than $AB$ is. The slope of secant line $AC$ is

$$m = \frac{2 - 1.25}{2 - 1} = 0.75$$

We can continue this game[4] by picking a point $P$ on the curve that is between $A$ and $C$ and finding the slope of the secant line $AP$. By repeating this process over and over, all the while moving the point $P$ closer and closer to $A$, we would get better and better estimates to the slope of the curve at the point $A(1, 1.25)$. The following table summarizes the results of a few such calculations; I encourage you to verify the figures in the table. You could just as well choose other values of $x_2$, as long as they decrease towards 1.

**Question:** Is there anything special about the points chosen in the previous calculations to estimate the slope of the graph at $x = 1$?

Note that each calculated slope in Table 3.1 is an overestimate of the true slope of the curve at the point $A$ (i.e., larger than the true value). There was nothing very special about the points chosen in the calculation; other nearby points would have served just as well. The points I chose make calculations relatively simple and (I hope) make the conclusion clear.

[2] Perhaps this comment is unduly harsh. We know that in this case the slope in question is positive, so at least we know that the actual slope is greater than zero and less than 1, which is some sense of accuracy. However, in more complicated situations, such as if a graph "wiggles" wildly, we might not get a sense from a graph whether the slope is positive or negative at a certain point.

[3] Once again, is this clear? If not, you will have to think about this and do some play, sketch some lines, do some calculations; whatever you need to do to convince yourself of this conclusion, take the time to convince yourself that this is correct.

[4] I should have said "systematic procedure" instead of game, but it's enough fun to be called a game, right?

Table 3.1: Calculations for overestimates to the slope of the graph of $f(x) = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$.

| $x_1$ | $x_2$ | $y_1$ | $y_2 = \frac{x_2^2}{4} + 1$ | $h = x_2 - x_1$ | $y_2 - y_1$ | $m = \frac{y_2 - y_1}{x_2 - x_1}$ |
|---|---|---|---|---|---|---|
| 1 | 3 | 1.25 | 3.25 | 2 | 2 | 1 |
| 1 | 2.5 | 1.25 | 2.5625 | 1.5 | 1.3125 | 0.875 |
| 1 | 2 | 1.25 | 2 | 1 | 0.75 | 0.75 |
| 1 | 1.5 | 1.25 | 1.5625 | 0.5 | 0.3125 | 0.625 |
| 1 | 1.1 | 1.25 | 1.3025 | 0.1 | 0.0525 | 0.525 |
| 1 | 1.01 | 1.25 | 1.255025 | 0.01 | 0.005025 | 0.5025 |
| 1 | 1.001 | 1.25 | 1.25050025 | 0.001 | 0.00050025 | 0.50025 |
| 1 | 1.0001 | 1.25 | 1.2500500025 | 0.0001 | 0.0000500025 | 0.500025 |

So what is the slope of the graph of $f(x) = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$? It's hard to say, isn't it? From the table of overestimates, it's clear that the slope of the curve at the point $A$ is less than 0.500025, but we can't be sure how much less the true value is.

It's a bit like being told to estimate the distance from Toronto to Vancouver, and saying, "It's about 10 km." That's a very poor estimate, partly because there is no statement of its accuracy. In normal every-day discourse, saying a distance is "about 10 km" contains a certain unspoken understanding about its degree of accuracy. If the true distance were 85 km, we wouldn't consider it very accurate to say the distance were about 10 km. However, in every-day life we might consider an estimate of 100 km reasonably accurate.

In mathematical discussions, there is no such unspoken understanding. Saying the slope of the curve at the point $A$ is about 0.500025 is useless, because it says nothing about how accurate the estimate is. To make the estimate worthwhile, we have to come up with some measure of its accuracy.

A good way to do this is to determine an *underestimate* for the slope of the curve at the point $A$. This would be like saying that the distance from Toronto to Vancouver is between 3000 km and 3500 km.[5] By giving both the underestimate (3000 km) and the overestimate (3500 km), there is a built-in statement of the accuracy of the estimate. An improved estimate, because it has greater accuracy, is to say that the distance is between 3300 km and 3400 km.

How do we obtain underestimates for the slope of the graph of $f(x) = \frac{x^2}{4} + 1$ at the point $A$? Consider Figure 3.9. Study the graph to see that the slope of the secant line $AD$ is less than the slope of the curve at $A$. (As before, imagine you are walking from left to right, and ask yourself which is steeper, the secant line $AD$ or the curve near $A$.) Thus, the slope of the secant line $AD$ is an underestimate for the slope of the curve at $A$. The slope of the secant line $AD$ is

$$m = \frac{1.25 - 1.0625}{1 - (-0.5)} = 0.125$$

So far, we know that the true slope of the curve at $A$ is between 0.125 and 0.500025.

**Question:** How can we make this approximation better?

[5] This is the distance by air, not by road.

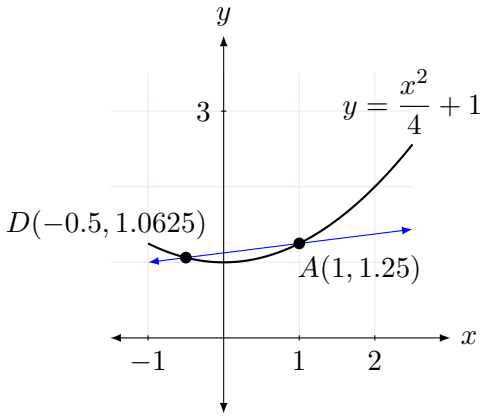



Figure 3.9: The secant line $AD$ gives an underestimate for the slope of the curve at $A$.

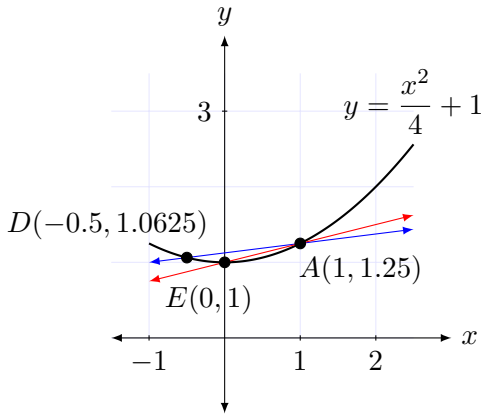



Figure 3.10: The secant line $AE$ also gives an underestimate for the slope of the curve at $A$, but it is a better estimate than $AD$.

Suppose we find the slope of the secant line joining the points $A(1, 1.25)$ and $E(0, 1)$. By examining Figure 3.10 and thinking about steepness, note that the slope of secant line $AE$ is greater than the slope of secant line $AD$, but still less than the slope of the curve at $A$. Thus, the slope of secant line $AE$ is a better approximation to the slope of the curve at $A(1, 1.25)$ than $AD$ is. The slope of secant line $AE$ is

$$m = \frac{1.25 - 1}{1 - 0} = 0.25$$

Thus, our improved estimate is that the true slope of the curve at $A$ is between 0.25 and 0.500025.

We can continue this fun game by picking a point $P$ on the curve that is between $A$ and $E$ and finding the slope of the secant line $AP$. By repeating this process over and over, all the while moving the point $P$ closer and closer to $A$, we would get better and better approximations to the slope of the curve at the point $A(1, 1.25)$. The following table summarizes the results of a few such calculations; I encourage you to verify the figures in the table. You could just as well choose other values of $x_2$, as long as they increase towards 1.

Table 3.2: Calculations for underestimates to the slope of the graph of $f(x) = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$.

| $x_1$ | $x_2$ | $y_1$ | $y_2 = \frac{x_2^2}{4} + 1$ | $h = x_2 - x_1$ | $y_2 - y_1$ | $m = \frac{y_2 - y_1}{x_2 - x_1}$ |
|---|---|---|---|---|---|---|
| 1 | −0.5 | 1.25 | 1.0625 | −1.5 | −0.1875 | 0.125 |
| 1 | 0 | 1.25 | 1 | −1 | −0.25 | 0.25 |
| 1 | 0.5 | 1.25 | 1.0625 | −0.5 | −0.1875 | 0.375 |
| 1 | 0.9 | 1.25 | 1.2025 | −0.1 | −0.0475 | 0.475 |
| 1 | 0.99 | 1.25 | 1.245025 | −0.01 | −0.004975 | 0.4975 |
| 1 | 0.999 | 1.25 | 1.24950025 | −0.001 | −0.00049975 | 0.49975 |
| 1 | 0.9999 | 1.25 | 1.2499500025 | −0.0001 | −0.0000499975 | 0.499975 |

Note that each calculated slope in Table 3.2 is an underestimate of the true slope of the curve at the point $A$ (i.e., smaller than the true value).

So what is the true value of the slope of the curve at the point $A$? It's still not clear. What seems clear is that the slope is smaller than any of the overestimate slopes in the final column of

Table 3.1, but also larger than any of the underestimate slopes in Table 3.2. Thus, the true slope of the curve at the point $A$ seems to be between 0.499975 and 0.500025. That is the best we can do at the moment.

Of course, we can always take the calculations further if we wish a more accurate estimate. As the second point approaches the point $A$, it seems that the slope of the secant line will become a better approximation to the slope of the curve at $A$.[6] But it also seems apparent that this method will never give us a definite value for the slope of the curve at the point $A$.

Let's conclude this discussion with an assessment of the advantages and disadvantages of this numerical method for estimating the slope of a curve at a point:

Table 3.3: Advantages and disadvantages of the numerical method for estimating the slope of a curve at a point.

| **Advantages** | **Disadvantages** |
|---|---|
| • the procedure can be visualized graphically<br>• the calculations are straightforward (rise-over-run)<br>• using overestimates and underestimates makes the estimate meaningful, and it seems that the accuracy can be improved by taking the calculations further<br>• the idea behind the calculations is used repeatedly in calculus, so it's worthwhile taking the time to understand it in this concrete setting | • the calculations are time-consuming<br>• it seems that the procedure will never be able to conclusively determine the precise slope of a curve at a point<br>• it's unclear whether it will be quite so easy to produce overestimates and underestimates in all cases, and it's unclear whether it will be possible to improve the accuracy indefinitely in all cases<br>• although the calculations are time-consuming, they still only help us estimate the slope at a single point on a single curve; if we desire an estimate of the slope at another point, even on the same curve, we have to repeat similar lengthy calculations all over again |

Overall, it seems that the numerical method for estimating the slope of a curve is a good start, but it would be nice if something better were available. Something better is available! We'll discuss improvements on this basic numerical method in the following pages. Before we do so, take this opportunity to practice the procedure we just illustrated by working out the following exercises.

[6] At least it seems to be true in this case; we'll see later that for curves with very wild wiggles, this is not necessarily true.

**EXERCISES**

(Answers at end.)

Use the numerical procedure outlined in this section (tables of overestimates and underestimates) to estimate the slope of the graph of each function at the indicated point.

1. Estimate the slope of the graph of $y = x^2$ at the point $A(2, 4)$.
2. Estimate the slope of the graph of $y = x^2 + 3$ at the point $A(2, 7)$.
3. Estimate the slope of the graph of $y = x^2 + 3x$ at the point $A(2, 10)$.
4. Explain graphically why the results in Exercises 1 and 2 are the same.
5. Explain graphically why the results in Exercises 1 and 3 are NOT the same.
6. Choose several of your own curved graphs, select a point on the graph, and estimate the slope of the graph at your chosen point. Check your estimate by using graphing software, zooming in on the graph near the chosen point until the graph looks nearly straight, and using the graph scale and rise-over-run.
7. Review the reasoning in this section and make sure that you understand the key points. Summarize what you have learned, write any questions or uncertainties in your journal, and discuss all of this with a study partner.

---

Answers: **1**. about 4; **2**. about 4; **3**. about 7; **4**. The graph in Exercise 2 is obtained from the graph in Exercise 1 by a vertical translation, which does NOT change the slope at a particular $x$-value. **5**. The graph in Exercise 3 is obtained from the graph in Exercise 1 by combining both a vertical translation and a horizontal translation; the horizontal translation DOES change the slope at a particular $x$-value.

## 3.2 Calculating the Slope of a Graph at a Point Using a Limit: Algebraic Approach

In the previous section we used a numerical method to estimate the slope of the graph of $f(x) = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$. Based on the calculations we've done so far, our best estimate is that the slope is between 0.499975 and 0.500025.

Now we'll repeat the same calculations, using the same process, using algebra. The algebraic approach will summarize the tables of numerical calculations in a much more streamlined way, but the basic idea behind the calculations is the same. Using the algebraic approach, we'll be able to come to a definite conclusion about the actual slope of the curve at the point $A$, as opposed to just an estimate.

To estimate the slope of the curve at the point $A$, select a nearby point $P$ on the curve, and construct the secant line $AP$; see either Figure 3.11 or Figure 3.12. The slope of the secant line $AP$ is calculated as follows. Note that after setting up the basic rise-over-run formula for the slope of the secant line $AP$, we then simplify the formula as much as possible. The reason for doing this is to compare the result to the ones obtained previously in the tables of overestimates and underestimates, Table 3.1 and Table 3.2.

According to custom, we'll use $h$ to represent the "run" in the "rise-over-run" slope calculation;

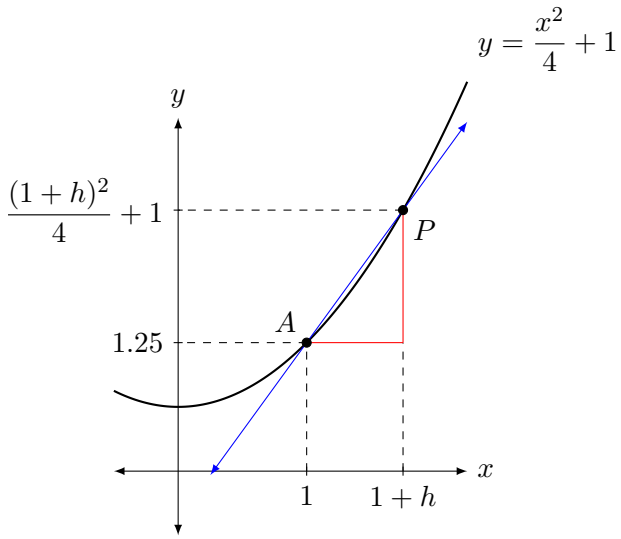



Figure 3.11: The secant line $AP$ is used in an algebraic calculation of the slope of the curve at $A$. Because the point $P$ is to the right of point $A$, the value of $h$ is positive.

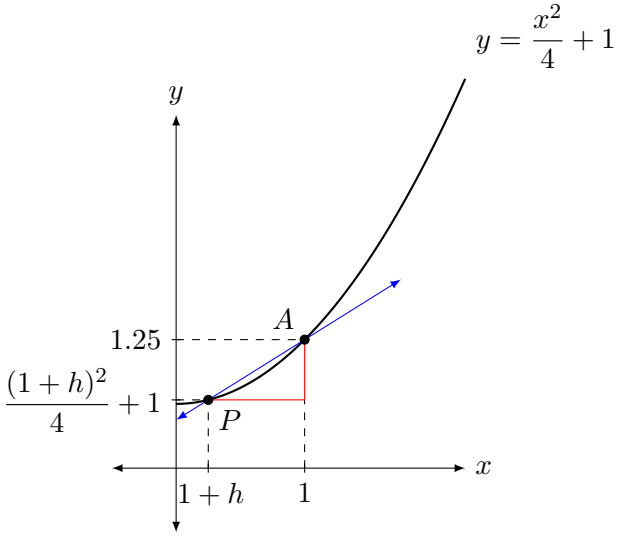



Figure 3.12: The secant line $AP$ is used in an algebraic calculation of the slope of the curve at $A$. Because the point $P$ is to the left of point $A$, the value of $h$ is negative.

that is, $h = x_2 - x_1$.

$$
\begin{aligned}
m &= \text{slope of secant line } AP \\
m &= \frac{\text{rise}}{\text{run}} \\
m &= \frac{y_2 - y_1}{x_2 - x_1} \\
m &= \frac{f(1+h) - f(1)}{(1+h) - 1} \\
m &= \frac{\left[\frac{(1+h)^2}{4} + 1\right] - \left[\frac{1^2}{4} + 1\right]}{1 + h - 1} \\
m &= \frac{\left[\frac{1+2h+h^2}{4} + 1\right] - \left[\frac{1}{4} + 1\right]}{h} \\
m &= \frac{\frac{1+2h+h^2}{4} + 1 - \frac{1}{4} - 1}{h} \\
m &= \frac{\frac{1+2h+h^2}{4} - \frac{1}{4}}{h} \\
m &= \frac{\frac{1+2h+h^2-1}{4}}{h} \times \frac{4}{4} \\
m &= \frac{1 + 2h + h^2 - 1}{4h} \\
m &= \frac{2h + h^2}{4h} \\
m &= \frac{h(2+h)}{4h} \\
m &= \frac{2+h}{4} \qquad \text{(provided that } h \neq 0\text{)} \\
m &= \frac{2}{4} + \frac{h}{4} \\
m &= 0.5 + \frac{h}{4}
\end{aligned}
$$

The formula $m = 0.5 + h/4$ tells us the slope of the secant line $AP$ for any position of the point $P$; the value of $h$ determines the position of the point $P$. Recall that $h = x_2 - x_1$, and compare this latest formula for $m$, the slope of the secant line $AP$, with the values calculated in Table 3.1 (which contains positive values of $h$) and Table 3.2 (which contains negative values of $h$).

By checking for yourself, you can verify that the formula for $m$ reproduces all of the values in the right-most column of each table. (Do this!) Thus, the formula $m = 0.5 + h/4$ is a concise summary of both tables. The algebraic approach and the numerical approach yield the same results.

So what is the slope of the curve at the point $A$? Consider the formula for the slope of a secant line that approximates the curve at $A$; that is, $m = 0.5 + h/4$. For positive values of $h$, the formula yields overestimates for the slope of the curve at $A$, and all of these values are greater than 0.5. Of course, as $h$ gets closer and closer to 0, the estimate gets closer and closer to 0.5, just as in Table 3.1. Similarly, for negative values of $h$, the formula yields underestimates for the slope of the curve at $A$, and all of these values are less than 0.5. As $h$ gets closer and closer to 0, the estimate gets closer and closer to 0.5, just as in Table 3.2.

If you believe the reasoning of the previous paragraph, you must conclude that the slope of the graph of $f(x) = \dfrac{x^2}{4} + 1$ at the point $A(1, 1.25)$ is 0.5, right? What else could it be? For example, could the slope at $A$ be 0.500007? No, and it is important to understand why. We argued that for positive values of $h$, the true value of the slope of the graph at $A$ is less than $m = 0.5 + h/4$. But if we choose the point $P$ so close to $A$ that $h = 0.000004$, then the true slope of the curve must be less than $0.5 + 0.000004/4$, which means that the true slope must be less than 0.500001. Thus, the true slope could not possibly be equal to 0.500007.

The important point is that the same kind of argument could be made to show that no matter which number $r$ we choose, however slightly greater than 0.5 the number $r$ is, that number could not possibly be the true slope of the graph at $A$, because by carefully choosing $h$ to be a small enough positive number, we can show that the true slope of the graph is actually less than $r$.

We can construct a similar argument using negative values of $h$ to show that no matter which value $s$ we choose, no matter how slightly less than 0.5 the number $s$ is, that number could not possibly be the true slope of the graph at $A$, because by carefully choosing $h$ to be a negative number whose absolute value is small enough, we can show that the true slope of the graph is actually greater than $s$.

**Question:** Does this reasoning convince you that the slope of the graph of $f(x) = \dfrac{x^2}{4} + 1$ at the point $A(1, 1.25)$ is 0.5?

The reasoning is valid for this particular function, but it depends crucially on the fact that for positive values of $h$, the formula for $m$ always gives an overestimate for the true slope, and for negative values of $h$ the formula for $m$ always gives an underestimate for the true slope. This is not true for all functions;[7] in particular, functions whose graphs have lots of "wiggles" near the chosen point $A$ will be problematic.[8] (Sketch some graphs and see if you can understand why this is so.)

Let's discuss a slightly different argument, one that does not depend on the fact that positive values of $h$ lead to overestimates in the slope formula, and negative values of $h$ lead to underestimates in the slope formula. Remember that in our initial numerical approach, we started with a secant line

[7] It is true for functions whose graphs are "concave up;" that is, if you constructed a wire model of the graph starting with a straight piece of wire, you would only have to bend the wire upwards. A more precise definition of concave-up graphs is studied in first-year university calculus courses.

[8] These more problematic situations motivate the need for a better definition of limit than the one we are introducing in this chapter. The better definition of limit is discussed in the last two chapters of this book.

$AP$ that approximated the curve, then moved the point $P$ closer to $A$ to get a better approximation.[9] What happens to $h$ as the point $P$ gets closer and closer to $A$? The value of $h$ gets closer and closer to 0.

Thus, the physical action of moving $P$ closer and closer to $A$ corresponds algebraically to evaluating the slope formula $m = 0.5 + h/4$ for values of $h$ that are closer and closer to zero, as was done in the tables. As $h$ gets closer and closer to zero, $h/4$ also gets closer and closer to zero, and therefore the slope of the secant line $AP$ gets closer and closer to 0.5. Another way to say this is:

The limit of the quantity $m = 0.5 + h/4$ as $h$ approaches zero is 0.5. In symbols,

$$\lim_{h \to 0} m = \lim_{h \to 0} \left( 0.5 + \frac{h}{4} \right) = 0.5$$

This introduces the concept of a limit in calculus, and also introduces the notation used for limits.

**Question:** Why don't we just substitute $h = 0$ in the slope formula $m = 0.5 + h/4$? Doing so leads to the same result, so why did we have to provide such lengthy reasoning?

There are two reasons for this. First, the reasoning is necessary to explain this fundamental concept of calculus. Second, there is a technical reason: Notice that in the process of simplifying the slope formula, we divided the numerator and denominator by $h$. (Equivalently, you could say we cancelled a factor of $h$ from the numerator and denominator.) This step in the procedure is valid only if $h \neq 0$, so it would be inconsistent if we said $h \neq 0$ at one point in the calculation and then turned around later and said that we're now letting $h = 0$.

**CAREFUL!**

**A tricky point that makes learning about limits so difficult**

The point discussed in the previous paragraph is really tricky, isn't it? This point was not explained very well by Newton, and led to severe criticism (which was valid) of calculus by Bishop Berkeley, as discussed elsewhere in this book. This tricky point is one of the reasons that learning about limits is difficult, and we devote a lot of time in this chapter to going over the calculations in great detail to help you get through this difficult concept.

Be careful about this tricky point. Go over it several times. Pay close attention to the arguments in this chapter, as they are foundational. Understanding these arguments will help you understand why the limit is defined as it is.

As for any difficult concepts in mathematics, the remedy for understanding them is to go over specific examples in a lot of detail, several times. This chapter will help you do this in the case of limits.

If we plot the slope formula as a function of $h$, we get the graph in Figure 3.14. Each non-zero value of $h$ corresponds to a different secant line, and the corresponding height on the graph in Figure 3.14 (above on the right) is the value of the slope of the corresponding secant line in Figure 3.13 (above on the left). Note that there is an open circle in the graph at $h = 0$ in Figure 3.14, which represents the fact that the value $h = 0$ is not allowed in this series of slope

[9] The point $A$ stays fixed, because we're interested in calculating the slope of the curve at $A$.

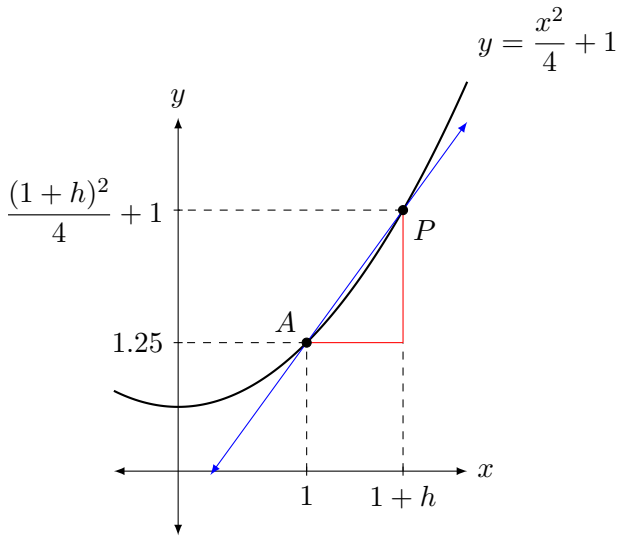



Figure 3.13: The secant line $AP$ is used in an algebraic calculation of the slope of the curve at $A$. The absolute value of $h$ is the distance between the $x$-coordinates of $A$ and $P$.

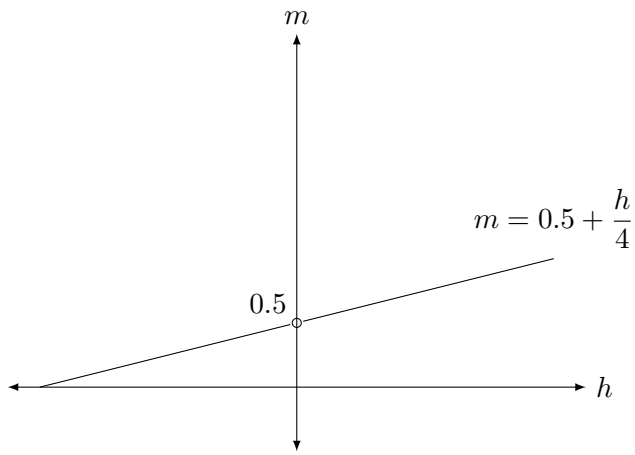



Figure 3.14: The graph shows the values of the slopes of secant lines $AP$ to the graph of $y = \frac{x^2}{4} + 1$ as a function of $h$. The value of $h$ indicates the position of the point $P$.

calculations. Also note that for positive values of $h$, the slope of the corresponding secant line is greater than 0.5. For negative values of $h$, the slope of the corresponding secant line is less than 0.5. The true value of the slope of the graph of $y = \dfrac{x^2}{4} + 1$ at the point $A$ is 0.5, the number that is less than all of the overestimates, but greater than all of the underestimates. In other words, the true value of the slope of the graph of $y = \dfrac{x^2}{4} + 1$ at the point $A$ is the value on the $m$ vs. $h$ graph that corresponds to the open circle.

**Question:** Now are you convinced that the true value for the slope of the graph of $y = \dfrac{x^2}{4} + 1$ at the point $A$ is 0.5?

Although the argument is sound, it may not be so easy to apply the argument for more complicated functions.[10] Part of the problem is that although we have named the process used to determine the slope of a curve (calculating a limit), we have not given a good definition of the limit concept, just referring to a "trend in values," which is much too sloppy to be mathematically acceptable. And we won't give a precise definition for a while, preferring to encourage you to play with the method first to get some experience. For a precise definition of the limit concept, and a logically air-tight version of the arguments given here, see Chapter 11.

Although we have not given a precise definition of a limit yet, we have described a procedure for calculating the slope of the graph of a function at a point. Next, let's look at some additional examples of applying the method to calculate the slope of the graph of a function at a point. We'll start with simple examples, and then move on to more complex examples. Once you have gained some experience with the calculations, you'll be well prepared for further discussion about the method, and the need for a precise definition will become apparent.

[10] For example, for functions with lots of "wiggles," the slopes of secant lines may fluctuate wildly as $h$ approaches 0, which might make it difficult to discern a trend in the values.

## GOOD THINKING HABIT

**Test new concepts in situations where you already know the result**

When encountering new concepts, it's a good idea to apply the new concept in a situation where you already know the result. This will give you some confidence that you are using the new concept correctly.

We have just learned a method for calculating the slope of a curve. In the next example, the method is applied to determine the slope of a straight line. Try the method on other straight lines as well.

This fits in with the usual way to learn mathematics: Start with simple applications of a new concept, or simple examples of the concept, and then gradually move to more complex examples and applications as your understanding grows.

## EXAMPLE 1

**Using a limit to determine the slope of a graph at a point**

Use the limit procedure to determine the slope of the graph of $f(x) = 2x + 3$ at the point for which $x = 5$.

## SOLUTION

Because the graph of $f$ is a straight line, and its formula is in the form $mx + b$, we can read the slope from the formula: The slope of the line, at each of its points, is 2. Thus, we know that the result of the following calculation must be 2.

Now let's actually do the limit calculation to check that the procedure produces the correct result of 2:

$$
\begin{aligned}
m &= \text{slope of secant line } AP \\
m &= \frac{\text{rise}}{\text{run}} \\
m &= \frac{y_2 - y_1}{x_2 - x_1} \\
m &= \frac{f(5+h) - f(5)}{(5+h) - 5} \\
m &= \frac{[2(5+h)+3] - [2(5)+3]}{5+h-5} \\
m &= \frac{[10+2h+3] - [10+3]}{h} \\
m &= \frac{[13+2h] - [13]}{h} \\
m &= \frac{2h}{h} \\
m &= 2 \quad \text{(provided that } h \neq 0\text{)}
\end{aligned}
$$

Thus, the slope of the graph of $f(x) = 2x + 3$ is 2, as expected.

Notice that $h$ does not appear in the final expression; this makes sense, because for a straight line, it does not matter how far away the points $A$ and $P$ are—the slope will always be the same.

**Question:** Will the same limit procedure give the correct result for the slope of every function that has a straight-line graph? If so, can you convince yourself of this using some calculations? If not, can you come up with a counterexample (i.e., an example of a function that has a straight-line graph for which the limit procedure does not give the correct result for the slope)?

Now let's calculate the slope of the graph of a quadratic function.

## EXAMPLE 2

**Using a limit to determine the slope of a graph at a point**

Use the limit procedure to determine the slope of the graph of $g(x) = x^2$ at the point for which $x = 3$.

## SOLUTION

$$
\begin{aligned}
m &= \text{slope of secant line } AP \\
m &= \frac{\text{rise}}{\text{run}} \\
m &= \frac{y_2 - y_1}{x_2 - x_1} \\
m &= \frac{g(3+h) - g(3)}{(3+h) - 3} \\
m &= \frac{\left[(3+h)^2\right] - \left[3^2\right]}{3 + h - 3} \\
m &= \frac{\left[9 + 6h + h^2\right] - [9]}{h} \\
m &= \frac{6h + h^2}{h} \\
m &= \frac{h(6 + 2h)}{h} \\
m &= 6 + 2h \quad \text{(provided that } h \neq 0\text{)}
\end{aligned}
$$

For positive values of $h$, the slopes of the secant lines $AP$ are greater than 6; for negative values of $h$, the slopes of the secant lines $AP$ are less than 6. Also, as $h$ gets closer and closer to 0, the values of the slopes of the secant lines get closer and closer to 6. That is,

$$\lim_{h \to 0} m = \lim_{h \to 0} (6 + 2h) = 6$$

Thus, the slope of the graph of $g(x) = x^2$ at the point for which $x = 3$ is 6.

**Question:** Does this seem reasonable? It would be a good idea to plot the graph and sketch some secant lines to check for yourself whether the result is reasonable.

## EXAMPLE 3

**Using a limit to determine the slope of a graph at a point**

Use the limit procedure to determine the slope of the graph of $p(x) = 4x^2 + 3x - 7$ at the point for which $x = 2$.

## SOLUTION

$$
\begin{aligned}
m &= \text{slope of secant line } AP \\
m &= \frac{\text{rise}}{\text{run}} \\
m &= \frac{y_2 - y_1}{x_2 - x_1} \\
m &= \frac{p(2+h) - p(2)}{(2+h) - 2} \\
m &= \frac{\left[4(2+h)^2 + 3(2+h) - 7\right] - \left[4(2^2) + 3(2) - 7\right]}{2 + h - 2} \\
m &= \frac{\left[4(4 + 4h + h^2) + 6 + 3h - 7\right] - \left[4(4) + 6 - 7\right]}{h} \\
m &= \frac{\left[4(4) + 16h + 4h^2 + 6 + 3h - 7\right] - 15}{h} \\
m &= \frac{\left[15 + 19h + 4h^2\right] - 15}{h} \\
m &= \frac{19h + 4h^2}{h} \\
m &= \frac{h(19 + 4h)}{h} \\
m &= 19 + 4h \quad \text{(provided that } h \neq 0\text{)}
\end{aligned}
$$

Taking the limit as $h$ approaches 0 of the expression for the slope of the secant line gives us the slope of the graph:

$$\lim_{h \to 0} m = \lim_{h \to 0} (19 + 4h) = 19$$

Thus, the slope of the graph of $p(x) = 4x^2 + 3x - 7$ at the point for which $x = 2$ is 19.

**Question:** Does this seem reasonable? It would be a good idea to plot the graph and sketch some secant lines to check for yourself whether the result is reasonable.

**EXERCISES**

(Answers at end.)

Use the algebraic procedure outlined in this section (limit of the slopes of secant lines as $h$ approaches 0) to calculate the slope of the graph of each function at the indicated point. Then draw a rough graph and sketch some secant lines to check for yourself whether the result is reasonable.

1. Calculate the slope of the graph of $y = x^2$ at the point $A(2, 4)$.
2. Calculate the slope of the graph of $y = x^2 + 3$ at the point $A(2, 7)$.
3. Calculate the slope of the graph of $y = x^2 + 3x$ at the point $A(2, 10)$.
4. Calculate the slope of the graph of $y = 2x^2 - 5x + 7$ at the point $A(-1, -10)$.
5. Calculate the slope of the graph of $y = 2x^2 - x - 5$ at the point $A(1, -4)$.

---

Answers: **1**. 4; **2**. 4; **3**. 7; **4**. −9; **5**. 3

**GOOD QUESTION**

**Does the limit procedure for determining slope work at every point on the graph of every function?**

Will the algebraic or numerical procedure for determining the slope of a curve work for all points of all graphs? NO. So which points on which graphs does it work for? How can we be sure that it works? What do the graphs look like at points for which the procedure works, and at points for which the procedure does not work? These are very good questions, and they will be discussed later in this book. It will be good for you to think about these questions now, and make an attempt at answering them. Sketching the graphs of various functions and playing with them may help!

Suppose you go through the process for determining the slope of the graph of a function $f$ at a point $(a, f(a))$. If the process is successful, we call the function $f$ *differentiable* at $x = a$. If the process is **not** successful, we say the function $f$ is **not** differentiable at $x = a$. If the process is successful for all values in the domain of $f$, then we simply say that $f$ is differentiable. This will be discussed at greater length, and with precise definitions, in Chapter 4.

## 3.3 Tangent Lines

Consider the process we have been using to calculate the slope of the graph of a function $f$ at a point $A$. It would be interesting to sketch the line that passes through $A$ that has the same slope as the graph of $f$ at $A$. This line is called the *tangent line* to the graph at $A$.

Recall our calculation of the slope of the graph of $y = \dfrac{x^2}{4} + 1$ at the point $A(1, 1.25)$; see Figure 3.7 to Figure 3.10 and Table 3.1 and Table 3.2 for the numerical approach, and see Figure 3.13 and Figure 3.14 for the algebraic approach. Recall that the result of the calculation is that the slope of the graph at $A$ is 0.5.

To determine the equation of the tangent line to the graph at $A$ (that is, the line that passes through $A$ and has slope 0.5), you can use any of the methods you learned in high school. For

example, you can let $(x, y)$ represent an arbitrary point on the line other than $A$, write an expression for the slope of the line joining $(x, y)$ and $A(1, 1.25)$, then equate the expression to 0.5, and finally solve for $y$, as follows:

$$\begin{aligned} \frac{y - 1.25}{x - 1} &= 0.5 \\ y - 1.25 &= 0.5(x - 1) \\ y - 1.25 &= 0.5x - 0.5 \\ y &= 0.5x + 0.75 \end{aligned}$$

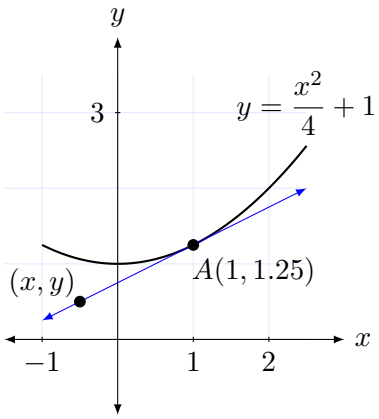



Figure 3.15: The tangent line at the point $A$ passes through the point $A$ and has the same slope as the curve at the point $A$.

Figure 3.15 shows the graph of the function together with the tangent line at $A(1, 1.25)$. If you think of the curve and tangent line as hill sides, and walk along the curve from left to right as always, then the steepness of the curved hill side **at the point** $A$ is exactly the same as the steepness of the straight hillside. The curve is less steep than the straight line before you reach $A$, and more steep after, but at the point $A$ the curve and the straight line have the same slope.

One way to summarize the previous paragraph is to say that the tangent line to the curve at $A$ is the best straight-line approximation to the curve near $A$. That's a good way to understand the tangent line geometrically.

## KEY CONCEPT

### Geometric and algebraic perspectives on tangent lines

The tangent line to a curve at a point $A$ is the best linear approximation to the curve at $A$. The slope of the tangent line is determined by calculating a certain limit.

This "best linear approximation" perspective on tangent lines is fundamental in calculus, and this concept will be used over and over as you learn about calculus. It is worth returning to this basic concept as your understanding of calculus grows.

## EXAMPLE 4

**Determining the equation of a tangent line to a graph at a point**

Determine an equation for the tangent line to the graph of $f(x) = x^2 - 3x + 1$ at the point $A(2, -1)$.

## SOLUTION

Strategy: First determine the slope $m$ of the tangent line, using an appropriate limit. Then use the slope $m$ and the point $A(2, -1)$ to determine the equation of the tangent line.

Step 1: Consider a point $P$ on the graph of the function $f$, where the $x$-coordinate of $P$ is a distance $h$ from the $x$-coordinate of $A$. The slope of the secant line $AP$ is

$$\begin{aligned}
\text{slope of secant line } AP &= \frac{\text{rise}}{\text{run}} \\
&= \frac{y_2 - y_1}{x_2 - x_1} \\
&= \frac{f(2+h) - f(2)}{(2+h) - 2} \\
\text{slope of secant line } AP &= \frac{\left[(2+h)^2 - 3(2+h) + 1\right] - \left[(2)^2 - 3(2) + 1\right]}{h} \\
&= \frac{\left[4 + 4h + h^2 - 6 - 3h + 1\right] - [4 - 6 + 1]}{h} \\
&= \frac{\left[h + h^2 - 1\right] - [-1]}{h} \\
&= \frac{h + h^2 - 1 + 1}{h} \\
&= \frac{h + h^2}{h} \\
&= \frac{h(1+h)}{h} \\
&= 1 + h \qquad \text{(provided that } h \neq 0)
\end{aligned}$$

Thus, the slope of a secant line $AP$ is $1 + h$, where $h$ represents the horizontal distance between $A$ and $P$. The slope of the secant line $AP$ is an approximation to the slope of the curve at $A$. As $P$ approaches $A$, the approximation gets better and better. The precise value $m$ of the slope of the curve at $A$, which is also the slope of the tangent line to the curve at $A$, is obtained by taking the limit of the slope of the secant line $AP$ as $h$ approaches zero:

$$\begin{aligned}
m &= \lim_{h \to 0} (1 + h) \\
m &= 1
\end{aligned}$$

As $h$ gets closer and closer to 0, the slope gets closer and closer to 1. Thus, the slope of the tangent line to the curve at $A$ is 1.

Step 2: Use the slope $m = 1$ of the tangent line and the point $A(2, -1)$ (which lies on the tangent line) to determine an equation for the tangent line. There are a number of ways to do this; we know an equation of the tangent line can be written the form $y = mx + b$, which is $y = x + b$ for this example, because we know the slope is $m = 1$. Then we can use the fact that the point $A(2, -1)$ lies on the tangent line to determine the value of $b$:

$$\begin{aligned} y &= x + b \\ -1 &= 2 + b \\ -1 - 2 &= b \\ -3 &= b \end{aligned}$$

Therefore, the equation of the tangent line is $y = x - 3$. The graph of this line is plotted in Figure 3.17.

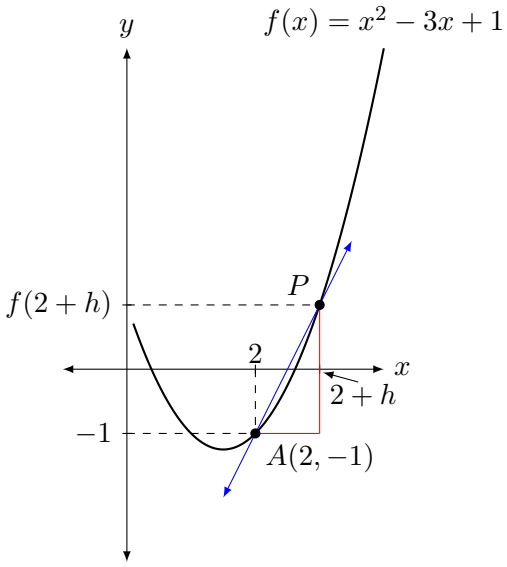



Figure 3.16: Diagram for calculating the slope of a secant line $AP$ in Example 4.

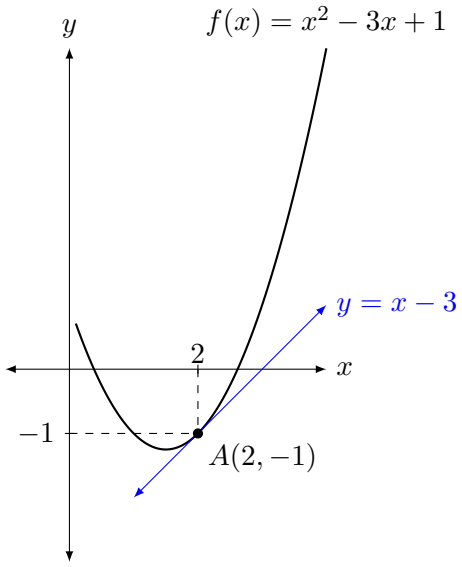



Figure 3.17: The tangent line to the graph of $f(x) = x^2 - 3x + 1$ at the point $A$ for the function in Example 4.

Now test your understanding of the process for determining the equation of a tangent line by completing the following exercises.

**EXERCISES**

(Answers at end.)

Determine an equation for the line that is tangent to the graph of the given function at the given point.

1. $y = -x^2 + 3$ at the point $A(-1, 2)$
2. $y = -x^2 + 3$ at the point $A(1, 2)$
3. $y = 3x^2 - 2x$ at the point $A(1, 1)$
4. $y = 3x^2 - 2x$ at the point $A(0, 0)$
5. $y = 2x^2 - 4x + 3$ at the point $A(1, 1)$
6. $y = 2x^2 - 4x + 3$ at the point $A(0, 3)$

---

Answers: **1.** $y = 2x + 4$; **2.** $y = -2x + 4$; **3.** $y = 4x - 3$; **4.** $y = -2x$; **5.** $y = 1$; **6.** $y = -4x + 3$

**CAREFUL!**

**Misconceptions about tangent lines**

So we now have a good algebraic perspective of tangent line: The tangent line to the graph of a function $f$ at a point $A$ on the graph of $f$ is the line through $A$ with slope determined by the limit process discussed earlier. We also have a good geometric perspective on the tangent line concept: It is the best linear approximation to the graph of $f$ near $A$. Many books try to give simpler geometric definitions of tangent line, but this invariably fails. The only way to characterize tangent lines to curves in full generality is the way described earlier in this section.

As usual, misconceptions abound on the internet (just as well as in books), and one must be careful when reading at random sites. For example, some sources attempt to characterize tangent lines by saying that they intersect a curve only once. This is clearly insufficient to capture the true nature of a tangent line, and also is incorrect. For example, the line $x = 0$ intersects the graph of $y = \cos x$ only once, but it is clearly not a tangent line to the graph. The tangent line to the graph of $y = \cos x$ at $x = 0$ is $y = 1$, as you can observe by sketching a graph of the function and the line. Note that the tangent line intersects the graph of the function at an infinite number of points!

An even more extreme example is any linear function. The tangent line to the graph of any linear function is the same line, which intersects the function at all points of the graph; that is, at all the infinite number of points on the graph.

In only certain extremely special cases can a tangent line be characterized more simply than in the way we have described in this section. One case we just discussed: Linear functions. Another simple situation is a circle, where at each point on a circle, the tangent line is the unique line through that point that is perpendicular to the radius that connects the centre of the circle to that point.

For a general graph of a function, the only way to characterize a tangent line is using the concept of limit that we are beginning to learn about in this chapter. Don't be mislead by incorrect characterizations of tangent lines!

## HISTORY

### Archimedes of Syracuse (c. 287 BCE – c. 212 BCE)

Archimedes was one of the greatest thinkers of ancient times. His main interest was mathematics, but he also made numerous advances in science and engineering. For example, he made fundamental advances in our understanding of fluids (including what we now call Archimedes's principle of buoyancy) and mechanics (including the law of the lever). He also invented and constructed ingenious mechanical devices, such as a screw pump (which are still widely used for many purposes), various devices involving pulleys, levers, catapults, and mirrors, and an ingenious astronomical device that modelled the apparent motions of the Sun and planets around the Earth.

In mathematics, Archimedes made many discoveries, especially in geometry, for which he is famous. The basic idea discussed in this chapter of using an iterative procedure with overestimates and underestimates to approximate a quantity predates Archimedes, but he used the idea (the "method of exhaustion") to great effect to approximate $\pi$. He started by inscribing a regular hexagon in a circle, calculating its area in terms of the radius of the circle, and then doing the same for a circumscribed regular hexagon. Then he successively doubled the number of sides of the inscribed and circumscribed polygons until he was working with 96-sided regular polygons. From this he was able to say that the value of $\pi$ was bounded by

$$\frac{223}{71} < \pi < \frac{22}{7}$$

Archimedes showed that the area enclosed by a parabola and a straight line is $\frac{4}{3}$ of the area of a certain inscribed triangle, again using the method of exhaustion. You might like to look up how Archimedes did this! To complete this demonstration he determined the sum of the following convergent infinite geometric series (see Chapter 10 for more details!):

$$1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots$$

In another *tour de force* of reasoning, Archimedes demonstrated that the volume of a sphere is $\frac{2}{3}$ of the volume of a minimal circumscribed cylinder, and also that the surface of a sphere is again $\frac{2}{3}$ of the surface area of a minimal circumscribed cylinder. Archimedes asked his friends to ensure that a representation of a sphere and cylinder be placed on his tomb, and this was done to honour the great mathematician.

It is absolutely amazing that Archimedes accomplished so much tremendous mathematical work without having the notation and concepts that we take for granted today, such as analytic geometry. Also, in those days they used Roman numerals, so he did not even have a decent number system for calculations! With his mastery of geometry, and his use of mechanical and geometrical concepts in his reasoning, it is not difficult to suppose that he may have invented calculus nearly 2000 years before Newton and Leibniz, if he had only had modern mathematical notation and algebra available.

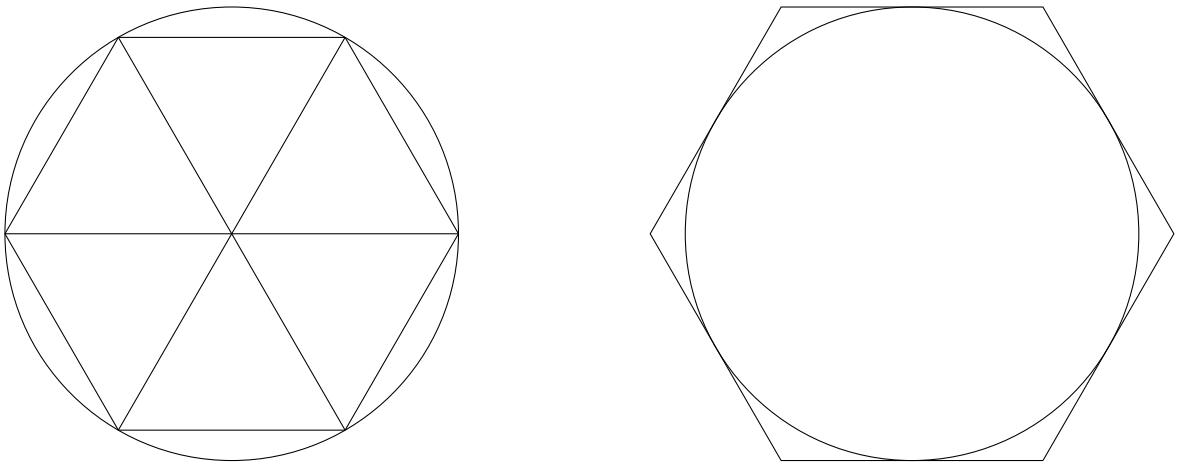


Figure 3.18: The diagram on the left shows a circle with an inscribed regular hexagon; the hexagon is separated into triangles to facilitate calculating its area. The diagram on the right shows a circle with a circumscribed regular hexagon.

## CHALLENGE PROBLEM

**In the footsteps of Archimedes**

Consider Figure 3.18. Calculate the area of the inscribed hexagon in terms of the radius of the circle. The inscribed hexagon has been separated into six congruent triangles, which may help you. Once you have completed this task, do the same for the circumscribed hexagon. You should arrive (after cancelling common factors of $r^2$) at the result that

$$\frac{3\sqrt{3}}{2} < \pi < 2\sqrt{3}r^2$$

Using a calculator, the approximate bounds for $\pi$ are therefore

$$2.598 < \pi < 3.464$$

Now double the number of sides of the approximating polygons and repeat your calculations. You should arrive at improved upper and lower bounds for $\pi$.

How far can you go? How much fun can you have here? If your calculations are systematic, then you should be able to write a short algorithm and implement it on your favourite platform to automate the calculations. It's always good to do a few rounds by hand, which helps you to really understand the process, and therefore helps to eliminate programming errors. Going slow at first usually makes things go faster in the end. How does the accuracy of the approximations improve as a function of the number of iterations? Is the convergence slow or fast? Is this a practical method for estimating *pi*? Were you able to approximate $\pi$ more accurately than Archimedes?

Have fun!

**EXCURSION**

This feature introduces a topic that is outside of the usual course of studies, but is fun to explore.

**Archimedes's argument about a sphere and a minimal circumscribed cylinder**

How did Archimedes demonstrate that the volume of a sphere is $\frac{2}{3}$ of the volume of a circumscribed cylinder?

How did Archimedes demonstrate that the surface area of a sphere is $\frac{2}{3}$ of the surface area of a circumscribed cylinder?

It may be fun for you to consider solving this problem on your own for a while, then comparing your thoughts to the method of Archimedes. You might also have fun looking up more modern methods for doing this calculation. Have fun!

**EXCURSION**

This feature introduces a topic that is outside of the usual course of studies, but is fun to explore.

**Archimedes's principle of buoyancy**

If you haven't already studied it you might find it fun to look up Archimedes's principle of buoyancy. After exploring this concept for a while, you might like to recall your own experiences of buoyancy in swimming pools, lakes, rivers, or oceans. After this, it will be a good idea to write some notes on your understanding.

After you have devoted some time to understanding Archimedes's principle of buoyancy, try solving the following problem, which is not very difficult if you have understood buoyancy, and yet it proves to be a little bit challenging for many first-year university students.

A boat floats in a bath tub that contains water. A rock is removed from the inside of the boat and is gently placed into the bath water. Does the water level of the bath tub (relative to the bath tub) go up, go down, or stay the same?

Is there a way for you to try the experiment so that you can see the result with your own eyes? (Maybe use a sink, or a small container to simulate a bath tub?)

Have fun!

## HISTORY

**The diligent workers in the shadows of the greats**

In our school mathematics courses, we typically learn very little about the people who developed the mathematics we study. This is not surprising, because there is a lot to learn in a limited amount of time. One thing that may be surprising is just how much progress unheralded researchers made over the centuries, and how essential their work was to breakthroughs by the greats. "If I have seen further than others, it is because I stood on the shoulders of giants," said Newton, and others before him as well.

Once Fermat developed analytic geometry (see the history section on Fermat and analytic geometry in the previous chapter), it became possible to produce an unlimited number of curves, just by choosing a formula and then plotting the resulting curve. Studying such curves became popular, and one of the popular things to do was to determine a formula for the tangent line to such a curve at an arbitrary point on the curve.

In those days (the early to mid 1600s), a considerable number of workers who made contributions to mathematics were not professional mathematicians; they were mathematics lovers who had other careers by which they made a living, and then they indulged their passion for mathematics in their spare time. And there were an enormous number of these enthusiasts.

Besides working on the problem of tangents, there were many other developments in calculus made by these immediate predecessors of Newton and Leibniz. The idea of integration, which originated with Archimedes, was developed by Kepler, Fermat, Cavalieri, and Torricelli, among others. Desargues, Pascal, Huygens (who was the teacher of Leibniz), Wallis, Barrow, and many others made important contributions to mathematics in general, and calculus in particular.

I mention all of these names to give you some entry points into the history of mathematics if you would like to learn more, but also to emphasize the fact that not just researchers who become famous make useful contributions. If you would like to make contributions to mathematics research, then don't worry that you are not great enough. Just ride your enthusiasm, work steadily, find a mathematical community so you have people to discuss mathematics with, and see where your journey takes you!

## SUMMARY

In this chapter, we learned a numerical version and an algebraic version of a process (calculating a limit) used for calculating the slope of the graph of a function at a point. This process is fundamental in calculus; the slope of a graph is connected to the rate of change of the quantity modelled by the graph, so knowing how to calculate the slope of a graph gives us a way of calculating rates of change.

In the next chapter we'll extend this process so that we can calculate the slope of a graph at an arbitrary point; that is, we'll treat the entire graph in one calculation, rather than having to do a separate calculation at each point.

**Make sure to regularly review the key concepts of this chapter, and also to regularly review the examples that you have worked through and the exercises that you have done. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 4

# Definition of Derivative

**OVERVIEW**

The derivative of a function is a new function that contains all of the information about the slope of the graph of the original function at each of its points. Thinking in terms of the derivative function provides a more streamlined and powerful way of calculating the slope of the graph of a function at any of its points. This is particularly useful when you need to analyze the slope of an entire graph, not just at a single point.

**WARMUP**

Before you tackle this chapter, make sure you can solve the following exercises. If you have difficulties, please review the appropriate prerequisites.

(Answers at end.)

Determine an equation for the tangent line to the graph of the function $y = x^2$ at each of the points $A(1, 1)$, $B(2, 4)$, and $C(-1.5, 2.25)$.

Answers: **1**. $y = 2x - 1$; **2**. $y = 4x - 4$; **3**. $y = -3x - 2.25$.

In the previous chapter we learned numerical and algebraic approaches for calculating the slope of the graph of a function at a point $A$. Both approaches used the same basic idea: First approximate the curve by drawing a secant line joining the point $A$ to a nearby point on the curve $P$. Then calculate the slope of the secant line $AP$. Finally, take the limit of the slope of the secant line $AP$ as the point $P$ approaches $A$ while $P$ stays on the curve. The result is the precise slope of the curve at the point $A$.

The numerical approach has some advantages and disadvantages; the algebraic approach improves on some of the disadvantages of the numerical approach. However, as you noticed in the warmup to this chapter, one of the disadvantages of the algebraic approach as we have used it so far is that if you wish to calculate the slope of a curve at several different points, you have to apply the process separately for each point. That is a lot of work; it would be nice if we could do the calculation "once and for all" for a curve instead of having to repeat the same sort of work over and over again for each point.

Let's see how we can improve the algebraic approach to calculating the slope of a curve. Rather than specify a particular point $A$, let's instead calculate the slope at an arbitrary point $A(a, f(a))$. The hope is that the result of the calculation will be a formula for the slope in terms of $a$; then, if we want to know the slope at several points, it might be relatively easy to substitute each value separately into the resulting formula. That is the hope—let's see if it works.

Let's apply this idea to the function $f(x) = x^2$. Consider the arbitrary point $A(a, a^2)$ on the graph, and a nearby point $P(a + h, (a + h)^2)$, also on the graph. The slope of the secant line $AP$

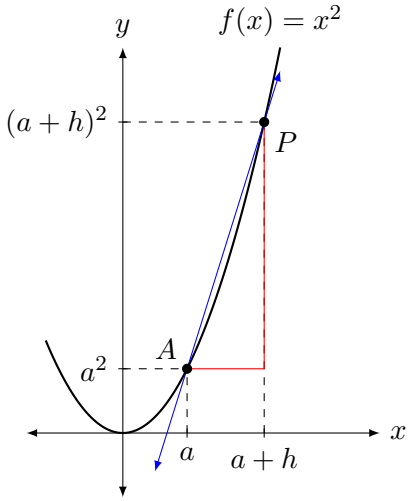



Figure 4.1: To calculate the slope of the curve at $A$, start with an expression for the slope of the secant line $AP$, then take the limit as $P$ approaches $A$ along the curve.

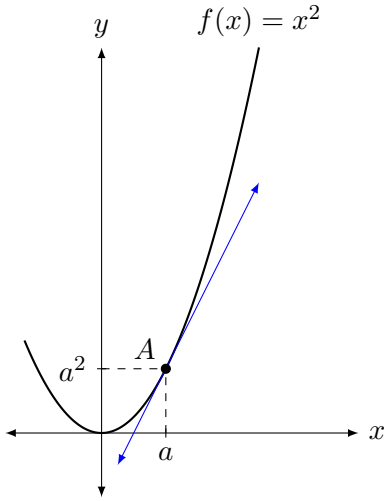



Figure 4.2: The result of taking the limit of the slopes of secant lines is the slope of the curve at $A$, which is the same as the slope of the tangent line at $A$.

is calculated as follows:

$$\begin{aligned}
\text{slope of secant line } AP &= \frac{\text{rise}}{\text{run}} \\
&= \frac{y_2 - y_1}{x_2 - x_1} \\
&= \frac{f(a+h) - f(a)}{(a+h) - a} \\
&= \frac{(a+h)^2 - a^2}{h} \\
&= \frac{a^2 + 2ah + h^2 - a^2}{h} \\
&= \frac{2ah + h^2}{h} \\
&= \frac{h(2a+h)}{h} \\
&= 2a + h \qquad \text{(provided that } h \neq 0\text{)}
\end{aligned}$$

The formula $2a + h$ represents the slope of a secant line $AP$ for various points $A$ and $P$ on the graph. If we now let the point $P$ approach the point $A$ along the curve (which amounts to taking the limit of the expression as $h$ approaches 0), then we will obtain an expression for the slope of the tangent line at the point $A$:

$$\begin{aligned}
\text{slope of tangent line at } A &= \lim_{h \to 0} (2a + h) \\
&= 2a
\end{aligned}$$

This beautiful little formula tells us that, for the graph of the function $f(x) = x^2$, the slope of the graph at $A(a, a^2)$ is $2a$.

Now, let's use this new formula to check the slopes you calculated in the Warmup questions at the beginning of this section:

| Point | Slope (using formula $2a$) | Slope (from Warmup) |
|---|---|---|
| $(1, 1)$ | 2 | 2 |
| $(2, 4)$ | 4 | 4 |
| $(-1.5, 2.25)$ | $-3$ | $-3$ |

The new formula indeed reproduces the calculations you did in the Warmup. You can see the advantage of the new approach: The limit procedure was carried out just once for the entire function, instead of three times (once for each point).

In the slope formula $2a$, consider Mr. Shakespeare's words in *Romeo and Juliet*:

> What's in a name? That which we call a rose
> By any other name would smell as sweet.

If we had labelled the point $A$ by $(x, x^2)$ instead of $(a, a^2)$, then the slope formula would have come out as $2x$. This is not saying anything new, just saying the same thing in different symbols. We could have chosen to say it in words just as well: The slope of the graph of $f(x) = x^2$ at any point is twice the value of the $x$-coordinate at that point. The point is that it doesn't matter which symbol we use, $a$, $x$, or some other symbol; the fact is the same, regardless of the symbol used.

So why would we wish to use the symbol $x$ instead of the symbol $a$ to represent the $x$-coordinate of the arbitrary point $A$? Well, because we are used to thinking of functions in terms of the symbols $x$ and $y$, and therefore it will be easier for us to recognize the slope formula $2x$ as a *new function*. We call this new function the *derivative* of the original function $f(x) = x^2$, and we symbolize the derivative function as $f'(x)$.

So, we have just learned our first derivative formula: For the function $f(x) = x^2$, the derivative function is $f'(x) = 2x$.

**CAREFUL!**

**Watch out for the same symbol used to mean two different things**

Mathematics textbooks occasionally use the same symbol to mean two different things, which can be confusing unless you are aware of it.

To avoid confusion we have used $a$ in the slope calculation we just completed, rather than $x$. In the slope calculation, the point $A$ remains fixed, and therefore the corresponding $x$-value $a$ also remains fixed. So it's useful to use two different symbols here; $x$ represents the independent variable, and $a$ represents a particular fixed value of the variable $x$.

However, many books use $x$ to stand for both quantities in this type of calculation, so one must be on guard to avoid confusion. Perhaps the best approach is to remember that, in the limit calculation, $a$ (or $x$, if you wish to call it $x$) remains fixed but arbitrary. Once the formula (slope $= 2a$, or slope $= 2x$ if you prefer) is obtained, then one understands that the value of $a$ (or the value of $x$, if you prefer) is arbitrary, and so any value in the domain of the function $f$ can be substituted for $a$ (or $x$) in the slope formula.

The formal definition of the derivative can be stated in several ways; here's one way:

**DEFINITION 1**

**Definition of the derivative function**

The derivative of the function $f$ at the point $A(a, f(a))$ is defined to be

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

provided that the limit exists. If the limit exists, then we say $f$ is differentiable at $x = a$. If $f$ is differentiable at all values of $x$ in its domain, then we simply say that $f$ is differentiable.

An equivalent definition involves the expression

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

One can obtain this expression from the one above by replacing $h$ by $(x - a)$. The two equivalent versions of the definition of derivative are illustrated in Figure 4.3 and Figure 4.4.

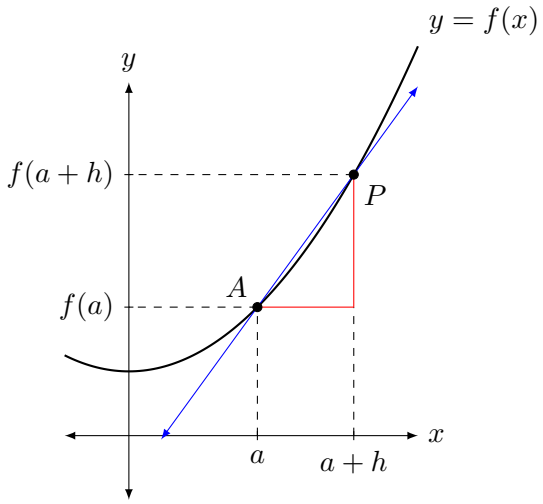



Figure 4.3: The limit of the slope of the secant line $AP$ as $P$ approaches $A$ along the curve (that is, as $h$ approaches zero) is the derivative of $f$ at the point $A$.

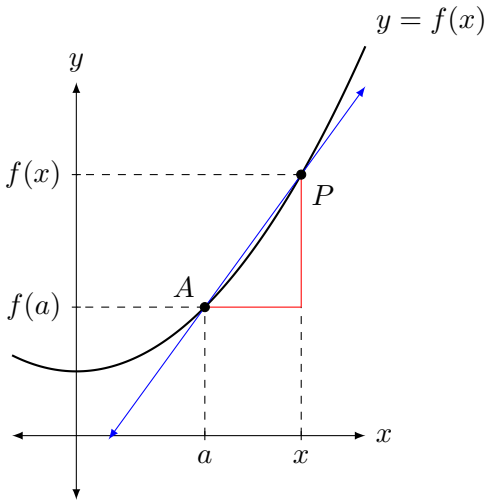



Figure 4.4: The limit of the slope of the secant line $AP$ as $P$ approaches $A$ along the curve (that is, as $x$ approaches $a$) is the derivative of $f$ at the point $A$.

Note that the definition of derivative summarizes the procedure that we have used quite a number of times already. On the right side of the definition is the limit of a quotient. The quotient itself represents the slope of a secant line $AP$ (the usual "rise-over-run"), and taking the limit as $h$ approaches 0 means to allow the point $P$ to approach the point $A$ along the curve to improve the estimate until it becomes precise.

**Question:** For which kinds of functions does the limit in the definition of the derivative not exist? What do the graphs of such functions look like?

Also note the phrase "provided that the limit exists" in the definition of the derivative. This implies that there might be some function for which the limit does not exist at some point, and therefore the function does not have a derivative at that point. We will explore this in the following sections, but you might give some thought now to whether such a function exists, and if so, what its graph might look like.

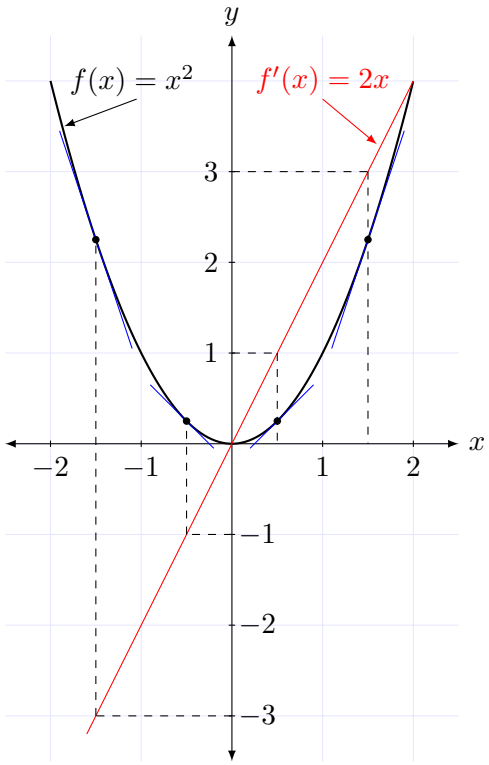



Figure 4.5: The *height* of the derivative function $f'$ (red) is equal to the *slope* of the function $f$ at the corresponding value of $x$.

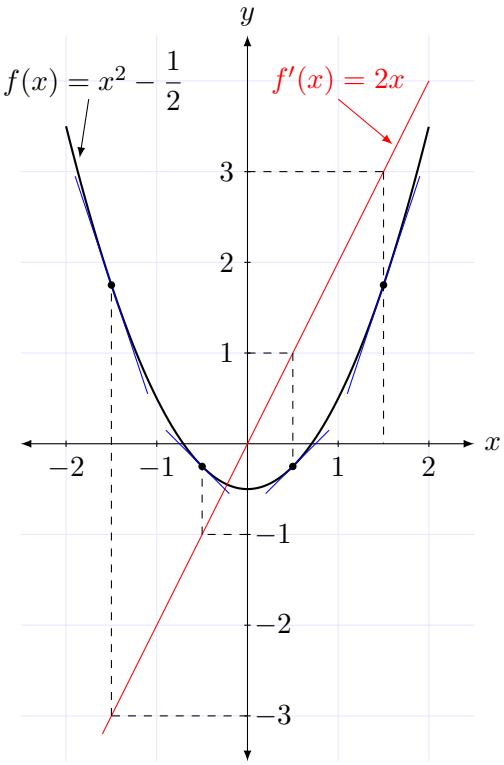



Figure 4.6: Notice that shifting the graph vertically down by $\frac{1}{2}$ unit does not change the slope of the graph at each value of $x$.

Is the derivative formula $f'(x) = 2x$ for the function $f(x) = x^2$ reasonable geometrically? Consider Figure 4.5. Note that the *height* of the derivative function $f'$ (in red on the graph) tells us the slope of the function $f$ at the same value of $x$. Follow the dashed vertical lines on the graph, and see if you can understand this point. Some representative tangent lines are also drawn to help you.

For example, at $x = 1$, the height of the derivative graph is $2(1) = 2$, and that is also the slope of the graph of $f$. Does this make sense from looking at the graph? Now make a related observation for the graph at $x = -1$: Note that the height of the derivative graph $f'$ is $-2$; does that appear to be the slope of the graph of $f$ at $x = -1$? Continue this comparison for the other indicated $x$-values on the graph (i.e., the ones with the vertical dashed lines).

Note that translating the graph vertically up or down does not change the slope of the graph at any particular $x$-value. In Figure 4.6, the graph of the function in Figure 4.5 is translated vertically down by $1/2$ unit. Check the indicated $x$-values on the graph, and compare the graph with the one in Figure 4.5 to see that the slopes of the graph of the function are not changed by this vertical translation by $1/2$ unit.

**Question:** Have you understood the points in the previous paragraphs? They are important points, and it is worth your time to copy the graph, sketch tangent lines with a ruler at various points, and sketch vertical dashed lines to verify the points we are making here. Don't rush through this; slow down and make sure you understand these key points.

The same reasoning tells us that a vertical translation by any amount, positive or negative, will not change the slope of a graph. Let's now convert this geometric statement into an equivalent algebraic one: The derivative of a function will not change if a constant value is added to or

subtracted from it.

Let's prove this fact for the function that we have been working with, $f(x) = x^2$, by finding the derivative formula for the function $g(x) = x^2 + c$, where $c$ is a constant that could be positive or negative. Using the definition of the derivative, we get:

$$\begin{aligned}
\text{slope of secant line } AP &= \frac{\text{rise}}{\text{run}} \\
&= \frac{y_2 - y_1}{x_2 - x_1} \\
&= \frac{g(a+h) - g(a)}{(a+h) - a} \\
&= \frac{\big[(a+h)^2 + c\big] - \big[a^2 + c\big]}{h} \\
&= \frac{\big[a^2 + 2ah + h^2 + c\big] - \big[a^2 + c\big]}{h} \\
&= \frac{a^2 + 2ah + h^2 + c - a^2 - c}{h} \\
&= \frac{2ah + h^2}{h} \\
&= \frac{h(2a+h)}{h} \\
&= 2a + h \qquad \text{(provided that } h \neq 0)
\end{aligned}$$

The formula $2a + h$ represents the slope of a secant line $AP$ for various points $A$ and $P$ on the graph. If we now let the point $P$ approach the point $A$ along the curve (which amounts to taking the limit of the expression as $h$ approaches 0), then we will obtain an expression for the slope of the tangent line at the point $A$:

$$\begin{aligned}
\text{slope of tangent line at } A &= \lim_{h \to 0} (2a + h) \\
&= 2a
\end{aligned}$$

This formula tells us that, for the graph of the function $g(x) = x^2 + c$, the slope of the graph at $A(a, a^2 + c)$ is $2a$. This is exactly the same formula as the slope formula for the function $f(x) = x^2$ at the point $(a, a^2)$. This shows that vertically translating the graph of the function $f$ up or down by a distance $|c|$ does not change the slope of the graph.

The proof for an arbitrary function is similar; see the theory chapters at the end of this book.

Let's try another example of calculating the slope of a curve, shall we? But this time we shall be a bit more formal (but don't worry, it's the same calculation we've been doing over and over) and use the definition of the derivative.

## EXAMPLE 5

**Determining a derivative formula**

**(a)** Determine the slope of the graph of the function $f(x) = \dfrac{1}{x}$ at the point $A(a, 1/a)$.

**(b)** Use the result of Part (a) to determine the slope of the graph of $f$ at the points $(-2, -0.5)$ and $(1, 1)$.

## SOLUTION

**(a)** The instruction tells us to determine a formula for the derivative function, $f'(x)$. Let's apply the definition of the derivative.

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} \\ f'(a) &= \lim_{h\to 0} \frac{\frac{1}{a+h} - \frac{1}{a}}{h} \end{aligned}$$

Now we need to simplify a fraction within a fraction. A smooth way to proceed is to observe that dividing by $h$ is the same as multiplying by $1/h$, and rewrite the expression as follows:

$$f'(a) = \lim_{h\to 0} \left[\frac{1}{h}\right] \left[\frac{1}{a+h} - \frac{1}{a}\right]$$

Now let's get a common denominator to simplify the difference of fractions in the right-hand bracket. To do this, multiply the numerator and denominator of the first term by $a$, and then multiply the numerator and denominator of the second term by $(a+h)$, then simplify:

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \left[\frac{1}{h}\right] \left[\frac{a}{a(a+h)} - \frac{a+h}{a(a+h)}\right] \\ f'(a) &= \lim_{h\to 0} \left[\frac{1}{h}\right] \left[\frac{a-(a+h)}{a(a+h)}\right] \\ f'(a) &= \lim_{h\to 0} \left[\frac{1}{h}\right] \left[\frac{a-a-h}{a(a+h)}\right] \\ f'(a) &= \lim_{h\to 0} \left[\frac{1}{h}\right] \left[\frac{-h}{a(a+h)}\right] \end{aligned}$$

Now we can divide the numerator and denominator both by $h$, as usual, provided that $h \neq 0$:[a]

$$f'(a) = \lim_{h\to 0} \left[\frac{-1}{a(a+h)}\right]$$

Finally we are in position to evaluate the limit. As usual, we ask ourselves what happens to the expression as $h$ gets closer and closer to 0. In this case, the result is

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \left[\frac{-1}{a(a+h)}\right] \\ f'(a) &= \frac{-1}{a(a)} \end{aligned}$$

[a] Notice that this division of numerator and denominator by $h$ has been a key step in all of our calculations of slopes of curves so far. This is typical; the point of simplifying the rise-over-run expressions is to achieve this cancellation of the $h$-factors in numerator and denominator so that the limit may be easily calculated.

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

If we wish to express the result in terms of $x$, we could just as well write the derivative formula as

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

This result is illustrated in Figure 4.7.

**(b)** When $x = -2$, the slope of the graph of $f$ is

$$\begin{aligned} f'(-2) &= -\frac{1}{(-2)^2} \\ &= -\frac{1}{4} \end{aligned}$$

When $x = 1$, the slope of the graph of $f$ is

$$\begin{aligned} f'(1) &= -\frac{1}{(1)^2} \\ &= -1 \end{aligned}$$

Thus, at the point $(-2, -1/2)$ the slope of the graph of $f$ is $-1/4$, and at the point $(1, 1)$ the slope of the graph of $f$ is $-1$. These results are illustrated in Figure 4.8.

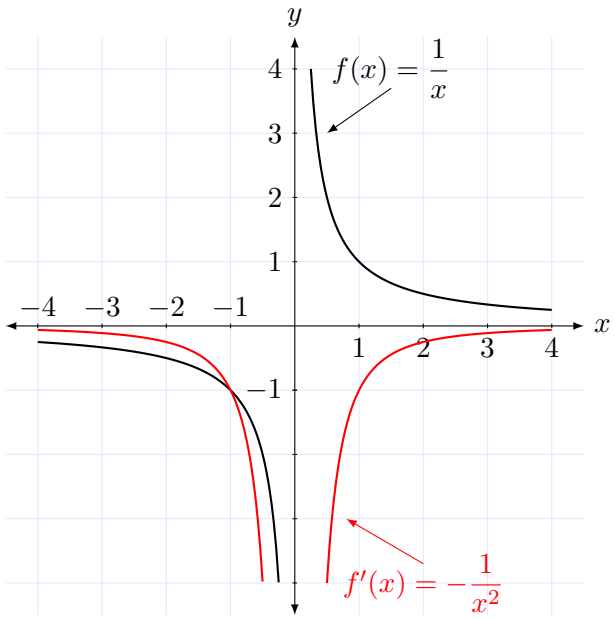



Figure 4.7: The graph of $f(x) = \frac{1}{x}$ is plotted in black, and its derivative $f'(x) = -\frac{1}{x^2}$ is plotted in red.

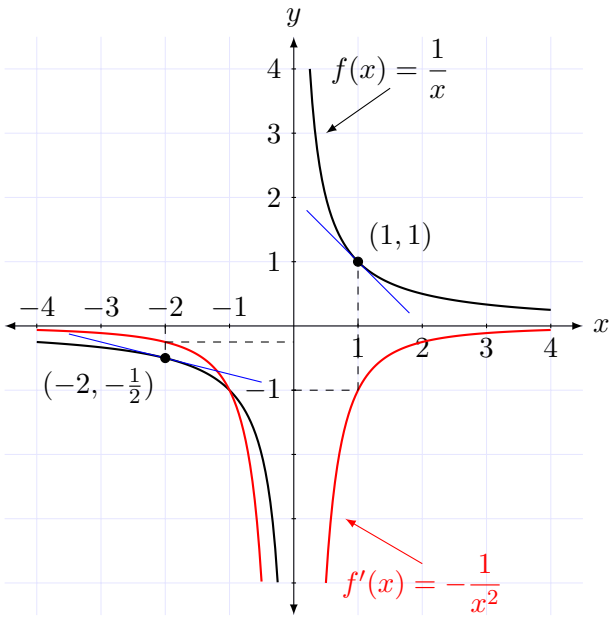



Figure 4.8: Tangent lines at $(-2, -1/2)$ and $(1, 1)$ are plotted in blue.

The results of the previous example are displayed in Figures 4.7 and 4.8. Remember that the height of the graph of $f'$ (in red) is equal to the slope of the graph of $f$ at each $x$-value. Study the

graph carefully, and make use of the dashed lines. Do the various values (height of $f'$ and slope of $f$) seem to match up? The tangent lines at the two given points are sketched to help you read off the slopes of the graph of $f$ at the given points.

Notice that the calculation of the derivative in the previous example followed exactly the same procedure introduced earlier in the chapter for determining the slope of a curve. The language we are now using is a little different, and the procedure is a bit more formal, but exactly the same ideas are being used.

Now let's have one more example.

## EXAMPLE 6

**Determining a derivative formula**

**(a)** Determine the slope of the graph of the function $f(x) = \sqrt{x}$ at the point $A(a, \sqrt{a})$.

**(b)** Use the result of Part (a) to determine the slope of the graph of $f$ at the points $(1, 1)$ and $(4, 2)$.

## SOLUTION

**(a)** The instruction tells us to determine a formula for the derivative function, $f'(x)$. Let's apply the definition of the derivative.

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} \\ f'(a) &= \lim_{h\to 0} \frac{\sqrt{a+h} - \sqrt{a}}{h} \end{aligned}$$

As usual, the next step is to cancel a factor of $h$ in numerator and denominator. However, there is no factor of $h$ in the numerator. The standard trick in this situation (when dealing with a difference of square-root expressions) is to rationalize the numerator. This means to multiply numerator and denominator by the conjugate of the square root expression, and then simplify:

$$\begin{aligned} f'(a) &= \lim_{h\to 0} \left[ \frac{\sqrt{a+h} - \sqrt{a}}{h} \cdot \frac{\sqrt{a+h} + \sqrt{a}}{\sqrt{a+h} + \sqrt{a}} \right] \\ f'(a) &= \lim_{h\to 0} \frac{\left[\sqrt{a+h} - \sqrt{a}\right]\left[\sqrt{a+h} + \sqrt{a}\right]}{h\left[\sqrt{a+h} + \sqrt{a}\right]} \\ f'(a) &= \lim_{h\to 0} \frac{a+h-\sqrt{a}\sqrt{a+h}+\sqrt{a}\sqrt{a+h}-a}{h\left[\sqrt{a+h} + \sqrt{a}\right]} \\ f'(a) &= \lim_{h\to 0} \frac{a+h-a}{h\left[\sqrt{a+h} + \sqrt{a}\right]} \\ f'(a) &= \lim_{h\to 0} \frac{h}{h\left[\sqrt{a+h} + \sqrt{a}\right]} \end{aligned}$$

Now we can divide the numerator and denominator both by $h$, as usual, provided that $h \neq 0$:

$$f'(a) = \lim_{h \to 0} \frac{1}{\sqrt{a+h} + \sqrt{a}}$$

Finally we are in position to evaluate the limit. As usual, we ask ourselves what happens to the expression as $h$ gets closer and closer to 0. In this case, the result is

$$\begin{aligned} f'(a) &= \lim_{h \to 0} \frac{1}{\sqrt{a+h} + \sqrt{a}} \\ f'(a) &= \frac{1}{\sqrt{a} + \sqrt{a}} \\ f'(a) &= \frac{1}{2\sqrt{a}} \end{aligned}$$

If we wish to express the result in terms of $x$, we could just as well write the derivative formula as

$$f'(x) = \frac{1}{2\sqrt{x}}$$

**(b)** When $x = 1$, the slope of the graph of $f$ is

$$\begin{aligned} f'(1) &= \frac{1}{2\sqrt{1}} \\ f'(1) &= \frac{1}{2(1)} \\ f'(1) &= \frac{1}{2} \end{aligned}$$

When $x = 4$, the slope of the graph of $f$ is

$$\begin{aligned} f'(4) &= \frac{1}{2\sqrt{4}} \\ f'(4) &= \frac{1}{2(2)} \\ f'(4) &= \frac{1}{4} \end{aligned}$$

Thus, at the point $(1, 1)$ the slope of the graph of $f$ is $1/2$, and at the point $(4, 2)$ the slope of the graph of $f$ is $1/4$.

---

The results of the previous example are displayed in Figure 4.10. Remember that the height of the graph of $f'$ (in red) is equal to the slope of the graph of $f$ at each $x$-value. Study the graph carefully, and make use of the dashed lines. Do the various values (height of $f'$ and slope of $f$) seem to match up? The tangent lines at the two given points are sketched to help you read off the slopes of the graph of $f$ at the given points.

In the next chapter, we'll develop our skills in calculating limits in general.

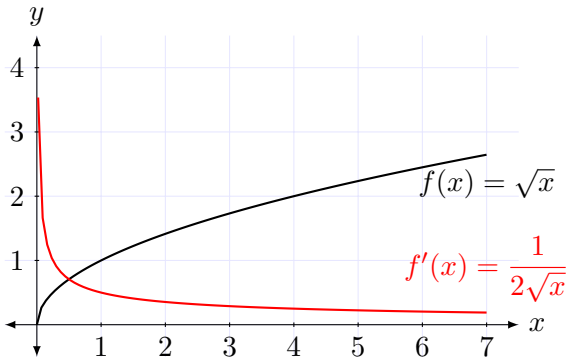



Figure 4.9: The graph of $f(x) = \sqrt{x}$ is plotted in black, and its derivative $f'(x) = \dfrac{1}{2\sqrt{x}}$ is plotted in red.

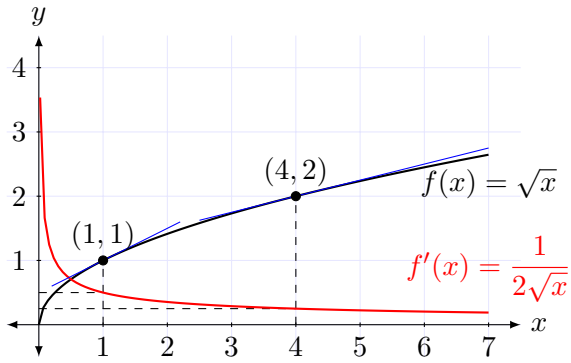



Figure 4.10: Tangent lines at $(1, 1)$ and $(4, 2)$ are plotted in blue.

## EXERCISES

(Answers at end.)

Use the definition of derivative to determine the derivative formulas for a variety of power functions, polynomial functions, rational functions, and functions involving square roots. Tabulate your results. Do you notice any interesting patterns?

Once you have determined a derivative formula for each function, evaluate the derivative formula at various values of the domain. Then plot both the original function and its derivative on the same set of axes. Sketch small segments of tangent lines at your selected values of the domain, and check for yourself that the height of the derivative function matches with the slope of the original function at these points.

One purpose of this exercise is to practice using the definition of the derivative to calculate derivatives of some simple functions. That is, this is practice in algebraic manipulations in the context of limits. Another purpose of this exercise is to strengthen your understanding of the geometric connection between a function and its derivative.

---

Answers: There may be a resistance on the part of some students to tackle such an open-ended task. However, it is very worth taking the time (which might be many hours spread over many days) to do this seriously. Do whatever is needed to encourage yourself to do this exercise seriously and compile the results. Check your results by using your favourite graphing software, such as `https://www.desmos.com/calculator`. When you get to university you will be expected to do a lot of learning on your own. Practice this skill of learning on your own now! The more systematic and detailed your exploration, the more you will learn! Have fun!

## CHALLENGE PROBLEM

### Derivative rules

Once you have calculated the derivatives of a large number of functions in the previous exercise, compile your results in a table. How will you organize the table? Try to do this in a reasonable way. Once you have organized the results, study the results and see if you can notice any patterns. Doing this may encourage you to organize the results in a different way, or perhaps use several different tables.

If you don't notice any patterns in your results, you may be encouraged to determine the derivatives of additional functions, to increase your database of functions for which you have the derivative formulas. Then study the larger database to search for patterns.

Once you notice patterns, you may wish to make some conjectures about rules for the derivative formulas of certain classes of functions. Then you might like to think about how you might prove your conjecture, or whether you can think of examples that might disprove your conjecture. This is fun!

The process described above is one aspect of mathematical thinking, and if you plan to become a mathematician you should definitely engage with this kind of process as much as you can. If you are planning to become a scientist or engineer, this process of working through examples that you construct for yourself, then compiling the examples and looking for patterns, is a very valuable skill to practice for you too. Take your time, and have fun!

As you do the work outlined in this challenge problem, you might consider how much of your work should be done "algebraically" (that is, working with the definition of the derivative and determining formulas), and how much of the work can be supplemented by thinking geometrically (that is, working with the graphs of the functions and their derivatives).

As usual, recording your thoughts, your difficulties, your results, your currently unanswered questions, and your reflections on what you have learned, will help you to become a more effective learner, and to place what you have learned in your long-term memory.

## CHALLENGE PROBLEM

**Anti-derivatives**

After you have built up an organized database of examples, as suggested by the previous challenge problem, another challenge that you can consider is to look at your table and ask yourself if you could "work backwards." That is, given the derivative of a function, could you figure out what the original function is? This process is called anti-differentiation (which is related to the process called integration), and it is an essential skill that you will practice extensively in a future university calculus course. This skill is essential for students who will study mathematics, physics, engineering, and other fields.

Practicing this process of "working backwards," i.e, anti-differentiation, may lead you to a deeper understanding of the derivative rules that you attempted to formulate in the previous challenge problem. You might also be led to conjecture some anti-differentiation rules that you can then explore and test with further examples. As in the previous challenge problem, you might consider how much of your work should be with formulas and how much of your thinking should be supplemented by work with the graphs of the functions you are studying. Geometry and algebra work hand-in-hand in calculus!

As usual, recording your thoughts, your difficulties, your results, your currently unanswered questions, and your reflections on what you have learned, will help you to become a more effective learner, and to place what you have learned in your long-term memory.

## GOOD THINKING HABIT

**How to cope with abstract mathematics textbooks**

A special message to all of you readers who are planning on becoming mathematics majors in university. Most mathematics textbooks are written in a very abstract way, and are short on examples. These books typically start with definitions, move directly to the statement of theorems and their proofs, and only then offer a few examples. This abstract approach is hard for young students to cope with. To counter this, and help yourself learn, it is of utmost importance that you learn to construct your own examples and counterexamples. Every time you encounter a new concept or a new definition, you should immediately construct for yourself quite a few examples of the new concept. For example, your textbook defines what a group is. Immediately start thinking of examples, and play with them. Is the set of all real numbers with the operation of addition a group? Can I prove this? What if the operation is multiplication? OK, after playing with these examples, move on to other examples. What about integers with addition, integers with multiplication, natural numbers with addition, natural numbers with multiplication, etc. What about other operations. You get the idea. Before you continue reading about all the wonderful theorems that can be proved about groups, you have played with quite a number of examples that you have constructed for yourself, and you are starting to get a good feel for which kinds of examples are groups and which are not groups (these are the counterexamples). The theorems and their proofs will be more meaningful to you because you will have a database of examples in your mind that will help you place the abstract arguments in a concrete context.

If you are a planning to major in mathematics, get used to constructing your own examples and counterexamples now. Take the previous exercise and the two previous challenge problems seriously. Practice these vital skills now! The work you do now, every day, will help lay a solid foundation for success in your university studies of mathematics.

Even if you are not planning on being a mathematics major, the habit outlined here is a good one in general. As a student of physics, engineering, or some other science, actively playing with examples and counterexamples is an excellent way to learn, to deepen your understanding, and to remember what you have learned for a lifetime. Practice now and have fun doing so!

## HISTORY

**Isaac Newton (1642–1727) and Gottfried Wilhelm Leibniz (1646–1716)**

Gottfried Wilhelm Lebniz was a legendary and highly influential genius, who learned so much about so many diverse fields, and wrote so much about them, that he left a mountain of unpublished manuscripts upon his death. He trained as a lawyer, and worked as a diplomat and for many years as a librarian and family historian for three successive dukes of Brunswick. He is perhaps best remembered today as a philosopher, where he made lasting contributions, but he also did important work in science, logic, philology, linguistics and the law.

His father died when Leibniz was six years old, and he was subsequently raised by his mother. He had an enormous ability to learn on his own, and because he was given free access to his father's library, he learned extensively. Leibniz visited Paris in 1672, met the Dutch physicist and mathematician Christiaan Huygens (1629–1695), and began to learn mathematics with his help. Within *three* short years he had reached a high enough level that he had invented calculus!

Isaac Newton is one of the greatest mathematicians in history, and separately one of the greatest physicists in history. (Doing the previous sentence justice would take many pages, but please read about him and his specific achievements!) Additionally he built the first reflecting telescope,

and found time to do extensive research in alchemy and theology. He was the second Lucasian Professor of Mathematics at Cambridge University, was warden and then master of the Royal Mint for the last 30 years of his life, and was president of the Royal Society (one of the earliest scientific academies in the world) for the last 24 years of his life. Newton's father died before he was born, and he was raised by his maternal grandmother.

What did Newton and Leibniz actually do that makes them deserving of being considered independently as the inventors of calculus? After all, many of their predecessors and contemporaries were "doing" calculus, so why do they get so much credit. There are two reasons for this. First, they took a collection of scattered results and methods and systematized them; that is, they organized the methods of calculus and made a system of them. Secondly, they both identified and proved the fundamental theorem of calculus, which is foundational and fundamental to further developments. (Preliminary versions of the fundamental theorem of calculus were stated and proved by James Gregory and Isaac Barrow.) One of Leibniz's legacies was that his notation was better than Newton's, and so it was easier to advance calculus to higher dimensions. (The importance of good notation for mathematical research has become better-appreciated, and so mathematicians who have developed good notation are becoming more appreciated.)

Human nature being what it is, an unfortunate dispute over priority of discovery arose among the friends and followers of Newton and Leibniz, and they were drawn into the dispute. As a result, English mathematicians stubbornly rejected Leibniz's formulation of calculus in favour of Newton's, and largely isolated themselves from continental European mathematicians, which set back English mathematics for decades, and perhaps a century.

Neither Newton nor Leibniz married or had children. Leibniz taught the Bernoulli brothers, who became influential, productive, and notable mathematicians, and they in turn continued this chain of inspiration that extended through some of the greatest mathematicians of their times: Euler, Lagrange, Gauss, Riemann, and so on. Newton had a personality quite different from Leibniz's, and left no such chain of illustrious students when he died. Leibniz loved social life, liked people and enjoyed their company, and strove to learn from everyone he met. As for Newton, according to Simmons in Section A.18 of his book *Calculus Gems*: "As an original thinker in science and mathematics he was a stupendous genius whose impact on the world can be seen by everyone; but as a man he was so strange in every way that normal people can scarcely begin to understand him. It is perhaps most accurate to think of him in medieval terms—as a consecrated, solitary, intuitive mystic for whom science and mathematics were means of reading the riddle of the universe."

## SUMMARY

In this chapter we used the definition of the derivative to calculate the slope formula for a number of graphs. In effect, we calculated formulas for the derivative functions for a number of functions. We also explored the geometric connection between a function and its derivative.

Calculating a derivative is one of the most important processes in calculus. The derivative function represents the rate of change of the function.

Having seen the importance of limits in derivative calculations, in the following sections we'll devote attention to developing our skills in calculating limits in general. Limits are used for other purposes besides calculating derivatives, as we shall see.

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 5

# Limits in General

**OVERVIEW**

In this chapter you'll improve your skill in evaluating limits. At this point in our studies, there are two main purposes for calculating a limit:

- to determine the slope of a curve (this is using the definition of the derivative)
- to understand the behaviour of a function in certain situations

In the previous chapters, our primary aim was to understand the idea behind how to calculate the slope of a curve. This required us to introduce the concept of a limit. Since limits form the current foundation of calculus (many important concepts, including the derivative, and the integral, which you will learn about in your future calculus studies, are defined in terms of limits), we'll now focus on improving your skill in evaluating limits.

Besides their use in calculating slopes of curves, limits are also useful for understanding the behaviour of a function near certain key points on its graph. For example, we might wish to understand how a function behaves near an asymptote, or near a value for which it is not defined, or near a value for which the function's behaviour has a sudden change, or its behaviour for values that are very large, either positively or negatively.

Notice that none of the purposes for limits mentioned in the previous sentence have anything to do with slope. Nevertheless, calculating the slope of a curve is one of the primary purposes of limits. This puts us in a bit of a difficult position. Recall that in using a limit of the following form to calculate a slope,

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

it makes no sense to substitute 0 for $h$ in the expression, because then we would be dividing by 0, which is undefined. This means that however we formally define a limit (which we will do later in this book), we will not be able to count on the definition to include substituting a value to determine the limit.

However, in all of the limit calculations we've done so far, we have simplified the rise-over-run quotient until there is no $h$ in the denominator, and then *in effect* we have substituted 0 for $h$ in order to evaluate the limit. This idea is universally used as part of a practical approach to evaluating limits, even though the formal definition of the limit cannot make reference to the value of the expression at the point in question, because in applying limits to the calculation of the slope of a curve, there will be no function value there.

These considerations make understanding limits difficult for many newcomers to calculus. Particularly when dealing with continuous functions, many newcomers can't understand why we have

to go through such contortions to calculate a limit—why don't we just substitute a value into the expression for the function? I hope the discussion of the previous several paragraphs has begun to clarify the reason. Let's look at some examples, in hopes that they will further clarify the issue.

Recall the very first slope calculation we did, earlier in this chapter. We calculated the slope of the graph of $y = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$. The relevant diagrams are reproduced in Figure 5.1 and Figure 5.2.

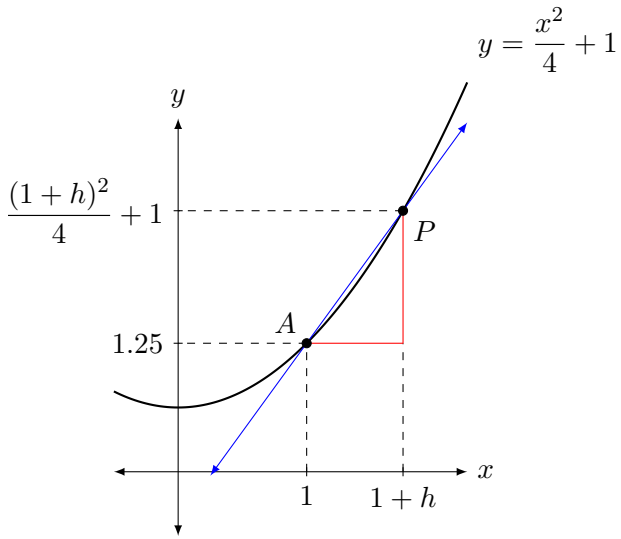



Figure 5.1: The secant line $AP$ is used in an algebraic calculation of the slope of the curve at $A$. The absolute value of $h$ is the distance between the $x$-coordinates of $A$ and $P$.

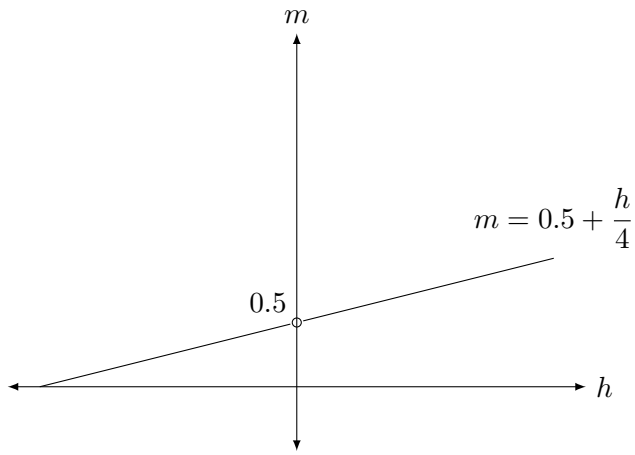



Figure 5.2: The graph shows the values of the slopes of secant lines $AP$ to the graph of $y = \frac{x^2}{4} + 1$ as a function of $h$, which indicates the position of the point $P$.

Recall our argument about the slope of the graph of $y = \frac{x^2}{4} + 1$ at the point $A(1, 1.25)$. Figure 5.2 displays all of the estimates of the slope of the curve at $A$, for various positions of the approximating secant line $AP$. We argued that all of the estimates for $h > 0$, shown in Figure 5.2 as values that are greater than 0.5, are overestimates. Also, all of the estimates for $h < 0$, shown in Figure 5.2 as values that are less than 0.5, are underestimates. Thus, the true value of the slope of the curve at $A$ can only be the $y$-value of the hole in the graph in Figure 5.2; that is, the true value of the slope of the curve at $A$ is 0.5.

If you understand the argument of the previous paragraph, you can see that although $h = 0$ is not in the domain of the formula $m = 0.5 + h/4$ (there is a hole in the graph at $h = 0$), nevertheless substituting $h = 0$ into the formula produces the correct value for the slope of the curve at $A$!

Because substituting a value into a formula is so much easier than going through intricate reasoning, it would be nice if we could come up with criteria for when substituting a value gives the correct result in a limit calculation. It is possible; first we'll state the ideas, then we'll illustrate them with examples. (Purists will note that typical developments of limits start with proper definitions of limits, then define continuity in terms of limits, then develop rules for working with limits. Our approach here is opposite, where we performed some concrete limit calculations after discussing the concept of a limit, now we will write down a few practical rules, and we will save most of the logical development of the subject for Chapters 11 and 12 of this book.)

**A practical approach to calculating limits**

1. If the function $f$ is continuous at $x = a$ (that is, there are no breaks, holes, or jumps in the graph of $f$ at $x = a$), then the limit of $f$ as $x$ approaches $a$ can be obtained by simply substituting the value $a$ for $x$ in the formula for $f$. That is,

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

2. If $f$ has a "hole" discontinuity at $x = a$, then the limit of $f$ as $x$ approaches $a$ can be obtained by "filling in the hole." That is, algebraically manipulate the expression for $f$ (if possible) so that it becomes acceptable to substitute $a$ for $x$; then evaluate the resulting formula for $x = a$ to obtain the limit.

3. If $f$ has a "jump" discontinuity at $x = a$, then

$$\lim_{x \to a} f(x) \quad \text{DOES NOT EXIST}$$

4. In more complex situations, one may have to use more powerful means: Limit laws (discussed in theory chapters at the end of this book), the squeeze theorem, or other theorems.

5. If the situation is still unclear, the last resort is to use the precise definition of the limit, discussed in the last two chapters of this book. This is the gold standard, and the fail-safe method.

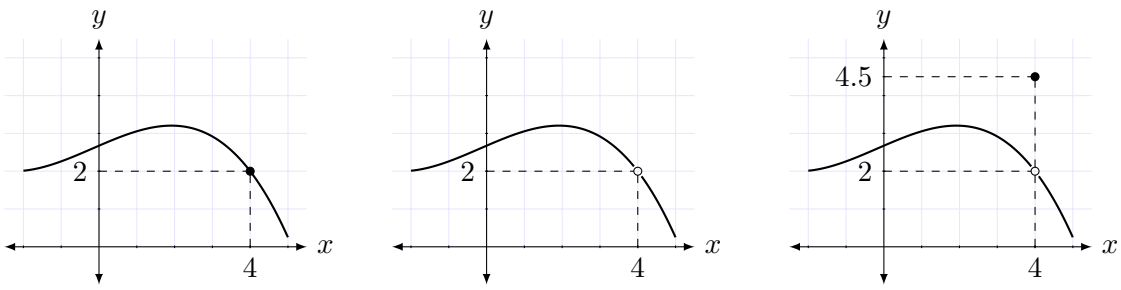



Figure 5.3: The three functions are slightly different, but for each function, the limit of the function as $x$ approaches 4 is 2. The point is that whether the function has a value at $x = 4$, and what the value is if the function does have a value, has no bearing on the existence and value of the limit of the function as $x \to 4$.

Consider the three functions graphed in Figure 5.3. In the first frame, the function is continuous at $x = 4$, and so the value of the limit of the function as $x$ approaches 4 is 2. In the second frame, the function has a hole discontinuity at $x = 4$; nevertheless, the limit of the function as $x$ approaches 4 is also 2. If you treat the function as a hillside, and imagine walking along it, then as you approach $x = 4$, either from the right or from the left, your height along the hillside is getting closer and closer to 2. Exactly the same argument results in the same conclusion about the function in the third frame. Even though the value of the function in the third frame at $x = 4$ is 4.5, the limit of the function as $x$ approaches 4 is 2.

The discussion of the functions in Figure 5.3 emphasizes that the limit of a function is not necessarily the value of the function; indeed, the function may not have a value at the point of interest, but it may indeed have a limit there.

Now let's look at a few examples of using our practical approach to calculating limits.

## EXAMPLE 7

**Calculating the limit of a function at a point of continuity**

For the function $f(x) = x^2 + 1$, calculate

$$\lim_{x \to 1} f(x)$$

## SOLUTION

Recall from your study of quadratic functions in high school that the function $f$ is continuous for all real values of $x$. Thus, we can use Step 1 in the practical approach to evaluating limits: Just substitute the given value of $x$ into the formula for the function:

$$\begin{aligned} \lim_{x \to 1} f(x) &= f(1) && \text{(because } f \text{ is continuous at } x = 1\text{)} \\ &= 1^2 + 1 \\ &= 2 \end{aligned}$$

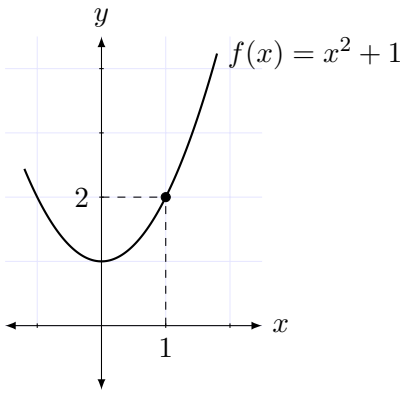



Figure 5.4: For a continuous function, the limit of the function as $x \to a$ is the value of the function at $x = a$; that is, $\lim_{x \to a} f(x) = f(a)$. For the function illustrated here, $\lim_{x \to 1} f(x) = f(1) = 2$.

Let's discuss the calculation in the previous example; refer to Figure 5.4. Remember, this limit has nothing to do with slopes; the limit in this example represents the trend of the *heights* (i.e., the function values) of the function graph as $x$ gets closer and closer to 1, either from the left or the right. This "trend" interpretation of limit can be illustrated by a table of values; see Table 5.1.

Table 5.1:

| $x$ | $f(x) = x^2 + 1$ | | $x$ | $f(x) = x^2 + 1$ |
|---|---|---|---|---|
| 0.1 | 1.01 | | 1.9 | 3.801 |
| 0.5 | 1.25 | | 1.5 | 3.25 |
| 0.9 | 1.81 | | 1.1 | 2.21 |
| 0.99 | 1.9801 | | 1.01 | 2.0201 |
| 0.999 | 1.998001 | | 1.001 | 2.002001 |
| 0.9999 | 1.99980001 | | 1.0001 | 2.00020001 |

The two columns on the left of Table 5.1 show the trend of the function values as $x$ approaches 1 from the left. The two columns on the right of Table 5.1 show the trend of the function values

as $x$ approaches 1 from the right. It appears as if the function values get closer and closer to 2 as $x$ approaches 1 from both left and right. This supports the calculation of the previous example.

## EXAMPLE 8

**Calculating the limit of a function at a point of continuity**

For the function $f(x) = \dfrac{1}{x}$, calculate

$$\lim_{x \to 2} f(x)$$

## SOLUTION

This function is continuous at $x = 2$, as you can see from the graph in Figure 5.5. Thus, we can use Step 1 evaluate the limit by substituting the given value of $x$ into the formula for the function:

$$\begin{aligned} \lim_{x \to 2} f(x) &= f(2) && \text{(because } f \text{ is continuous at } x = 2) \\ &= \frac{1}{2} \end{aligned}$$

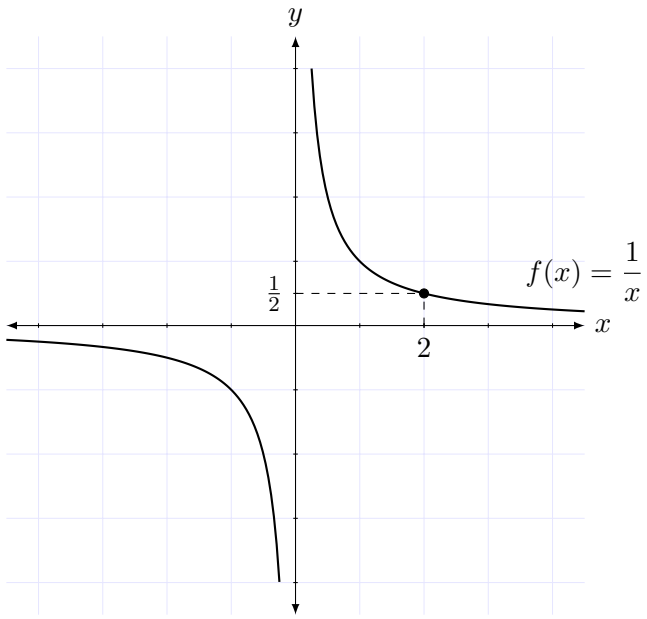



Figure 5.5: Because $f$ is continuous at $x = 2$, the limit of $f$ as $x \to 2$ is the value of the function at $x = 2$; i.e., $\lim_{x \to 2} f(x) = f(2) = \frac{1}{2}$. The fact that $f$ is not continuous at $x = 0$ is irrelevant, because at the point of interest (i.e., $x = 2$) $f$ is continuous.

To repeat, the fact that the function $f(x) = \dfrac{1}{x}$ is not continuous at $x = 0$ is irrelevant when calculating the limit of the function as $x$ approaches 2, because the function $f$ is continuous at $x = 2$.

How can one know for certain whether a function is continuous or not? Sketching a graph is not always trustworthy, because we may miss key points if we sketch the graph by hand or using computer software or a graphing calculator. The following theorem will be helpful. (For further discussion of the theorem, see the logical development of limits in the last two chapters.)

**THEOREM 1**

**List of types of continuous functions**

- The following types of functions are continuous for all $x$-values for which they are defined: Polynomial, rational, power (where the exponent may be any real number), trigonometric, inverse trigonometric, exponential, logarithmic, and hyperbolic functions.
- Algebraic combinations of continuous functions (addition, subtraction, scalar multiples, multiplication, division, and composition) are also continuous wherever they are defined.

Let's discuss the previous theorem. What the first part says is that functions such as $y = 3x^4 - 2x^2 + 0.7$, $y = \dfrac{x^2 - 2x + 5}{x^8 - 17}$, $y = x^{-3.2}$, $y = \sin x$, $y = \tan^{-1} x$, $y = 5^x$, $y = \log_3 x$, and $y = \cosh x$ are continuous wherever they are defined. The second part of the theorem says that if you combine functions such as these using any of the algebraic operations listed, then the resulting function is also continuous wherever it is defined. This means that the following functions, for example, are continuous wherever they are defined: $y = x + 2\sin x$, $y = \dfrac{\sin x}{x^2 + 1}$, $y = 3\ln x - 2\cos x$, and so on.

Some functions are continuous for all values of $x \in \mathbb{R}$. Examples are all polynomials, $y = \sin x$, $y = \cos x$, $y = k^x$ (where $k$ is any positive real number), and many others.

**EXAMPLE 9**

**Calculating the limit of a function at a point of continuity**

Calculate each limit.

$$\textbf{(a)} \lim_{x \to 0} \frac{x^2 - 3x + 4}{\cos x} \qquad \textbf{(b)} \lim_{x \to 3} \frac{2x^2 + x - 6}{x + 2}$$

**SOLUTION**

**(a)** The function $\dfrac{x^2 - 3x + 4}{\cos x}$ is an algebraic combination of continuous functions, so the function is continuous wherever it is defined. The function is defined at $x = 0$, so the limit can be calculated by substitution.

$$\begin{aligned} \lim_{x \to 0} \frac{x^2 - 3x + 4}{\cos x} &= \frac{0^2 - 3(0) + 4}{\cos 0} \\ &= \frac{4}{1} \\ &= 4 \end{aligned}$$

**(b)** The function $\dfrac{2x^2 + x - 6}{x + 2}$ is an algebraic combination of continuous functions, so the function is continuous wherever it is defined. The function is defined at $x = 3$, so the limit can be calculated by substitution.

$$\begin{aligned}\lim_{x\to 3}\frac{2x^2+x-6}{x+2} &= \frac{2(3)^2+3-6}{3+2}\\ &= \frac{15}{5}\\ &= 3\end{aligned}$$

Next, let's look at limit calculations at points where a function is not continuous.

## EXAMPLE 10

**Calculating the limit of a function at a point of discontinuity**

For the function $f(x)=\dfrac{x^2+3x+2}{x+1}$, calculate

$$\lim_{x\to -1} f(x)$$

## SOLUTION

The function $f$ is not continuous at $x=-1$, so we can't evaluate this limit by substitution. (We know the function is not continuous at $x=-1$ because it's not even defined there.)

However, notice that both the numerator and denominator equal 0 when $x=-1$; this is reminiscent of the derivative calculations we have done. There, we simplified the expression until we could cancel a factor of $h$ from both numerator and denominator; then we could determine the limit. Perhaps a similar strategy will work here.

Notice that $(x+1)$ must be a factor of the numerator, because substituting $x=-1$ into the numerator results in 0 (this is the factor theorem from high school). Therefore,

$$\begin{aligned}\lim_{x\to -1} f(x) &= \lim_{x\to -1}\frac{x^2+3x+2}{x+1}\\ \lim_{x\to -1} f(x) &= \lim_{x\to -1}\frac{(x+1)(x+2)}{x+1}\\ \lim_{x\to -1} f(x) &= \lim_{x\to -1}(x+2) \qquad \text{(cancelling the } (x+1) \text{ factors)}\\ \lim_{x\to -1} f(x) &= -1+2\\ \lim_{x\to -1} f(x) &= 1\end{aligned}$$

Continuing the discussion of the previous example, consider the graph of $f(x)=\dfrac{x^2+3x+2}{x+1}$ in Figure 5.6. Notice that the function has a hole discontinuity at $x=-1$, and yet the limit of the function is just the value that the function would have at $x=-1$ if it were continuous. In effect, you fill in the hole to calculate the limit.

Also notice that the graph of $f$ is almost identical to the graph of $y=x+2$, an expression that appears towards the end of the limit calculation. The only difference is that $f$ is not defined at $x=-1$ (the graph has a hole discontinuity there), and $y=x+2$ is continuous at $x=-1$; in other words, for the function $y=x+2$, the hole has been filled in.

The last few steps in the example, where we just substituted $x=-1$ once we had cancelled the troublesome factors of $(x+1)$, requires some thought and reasoning. However, it's the same reasoning we've been through quite a few times already. As $x$ approaches $-1$ from the left, the function values get closer and closer to 1. Similarly, as $x$ approaches $-1$ from the right, the function

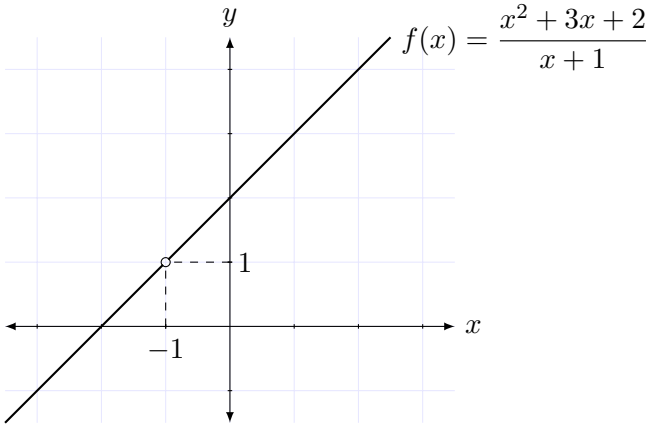



Figure 5.6: To calculate the limit of a function that has a hole discontinuity, "fill in the hole." In this case, $\lim_{x\to -1} \frac{x^2+3x+2}{x+1} = \lim_{x\to -1} \frac{(x+1)(x+2)}{x+1} = \lim_{x\to -1}(x+2) = -1+2 = 1$.

values also get closer and closer to 1. Thinking of a limit in terms of the trend in function values supports the calculation done in the example. You could also get your calculator out and construct a table of values for further support; see Table 5.2.

Table 5.2:

| $x$ | $f(x) = \frac{x^2+3x+2}{x+1}$ | | $x$ | $f(x) = \frac{x^2+3x+2}{x+1}$ |
|---|---|---|---|---|
| $-2$ | 0 | | 0 | 2 |
| $-1.5$ | 0.5 | | $-0.5$ | 1.5 |
| $-1.1$ | 0.9 | | $-0.9$ | 1.1 |
| $-1.01$ | 0.99 | | $-0.99$ | 1.01 |
| $-1.001$ | 0.999 | | $-0.999$ | 1.001 |
| $-1.0001$ | 0.9999 | | $-0.9999$ | 1.0001 |

The two columns on the left of Table 5.2 show the trend of the function values as $x$ approaches $-1$ from the left. The two columns on the right of Table 5.2 show the trend of the function values as $x$ approaches $-1$ from the right.

Tables of values may give us some support for the result in the example, but ultimately they prove nothing. It is the reasoning that provides the proof; but even the reasoning that we have presented so far in the chapter, while good, is not iron-clad. If you are still skeptical about the results, please study the formal definition of the limit (in Chapter 11), and learn how to do the proofs. That is the gold standard in the theory of limits, and should assuage any remaining doubts.[1]

[1]Maybe doubts will still remain; in that case, maybe you are ready to make the next big development in the theory of calculus! Or, like almost all of us, perhaps just a sufficient amount of time, practice, and reflection will do the trick and lead to your deep understanding, which will then remove any remaining doubts. Be patient with yourself, and give yourself the time needed to work through enough examples, reflect on them, and discuss them with fellow students and your teachers.

## EXAMPLE 11

**Calculating the limit of a rational function**

Calculate each limit:

$$\textbf{(a)}\ \lim_{x\to 4}\frac{\sqrt{x}-2}{x-3} \qquad\qquad \textbf{(b)}\ \lim_{x\to 4}\frac{\sqrt{x}-2}{x-4}$$

## SOLUTION

**(a)** Since the function $\dfrac{\sqrt{x}-2}{x-3}$ is an algebraic combination of continuous functions, it is continuous at $x=4$, provided it is defined there. So we try to substitute $x=4$ into the expression, and find that indeed the function has a value at $x=4$. Therefore, the limit can be calculated by substitution:

$$\begin{aligned}\lim_{x\to 4}\frac{\sqrt{x}-2}{x-3} &= \frac{\sqrt{4}-2}{4-3}\\ &= \frac{2-2}{1}\\ &= 0\end{aligned}$$

**(b)** We start off with the same reasoning as in part (a): Because the function $\dfrac{\sqrt{x}-2}{x-4}$ is an algebraic combination of continuous functions, it is continuous at $x=4$, provided it is defined there. So we try to substitute $x=4$ into the expression, but this time we find that the denominator is 0, so the function is not defined at $x=4$, and thus not continuous at $x=4$. Therefore, the limit **cannot** be calculated by substitution. So we move on to Step 2 in the practical approach to calculating limits: Try to algebraically simplify the expression and cancel a troublesome factor in the numerator and denominator. (We have some hope that this might work because the numerator is also 0 when $x=4$.) Because we have a square-root expression in the numerator, the usual trick here is to multiply numerator and denominator by the conjugate expression.[a]

$$\begin{aligned}\lim_{x\to 4}\frac{\sqrt{x}-2}{x-4} &= \lim_{x\to 4}\frac{\sqrt{x}-2}{x-4}\times\frac{\sqrt{x}+2}{\sqrt{x}+2}\\ &= \lim_{x\to 4}\frac{(\sqrt{x}-2)(\sqrt{x}+2)}{(x-4)(\sqrt{x}+2)}\\ &= \lim_{x\to 4}\frac{x-4}{(x-4)(\sqrt{x}+2)}\\ &= \lim_{x\to 4}\frac{1}{\sqrt{x}+2} \qquad \text{(cancelling the troublesome factors of } (x-4))\\ &= \frac{1}{\sqrt{4}+2}\\ &= \frac{1}{2+2}\\ &= \frac{1}{4}\end{aligned}$$

[a] You'll recall we have seen this trick before, when we calculated the derivative of a square root function.

Notice the key steps in the strategy of solving part (b) of the previous example:

**Strategy for evaluating the limit of a ratio of functions that has a hole discontinuity**

- Begin by trying to evaluate the limit by substitution; notice that 0 is obtained in both numerator and denominator.
- Algebraically simplify until you can cancel a troublesome factor in both numerator and denominator.
- Since the result is a function that is continuous, evaluate its limit by substitution.
- Argue that the original function must have had a hole discontinuity, so the limit of the continuous function that appears late in the solution must be equal to the limit of the original function.

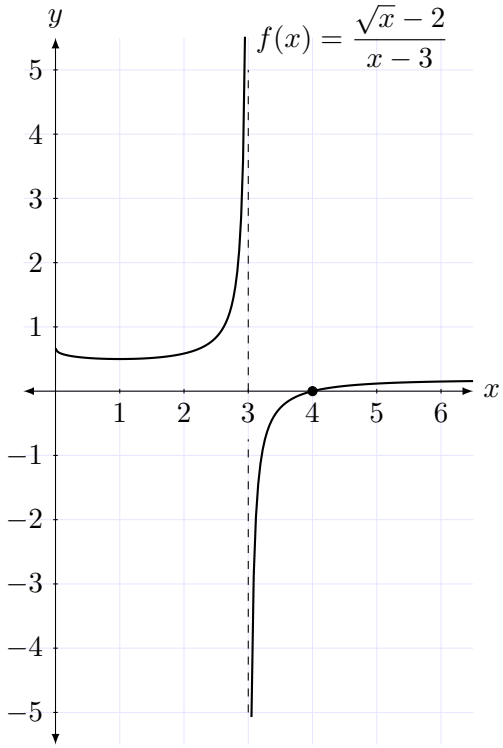



Figure 5.7: This function is continuous at $x = 4$, so its limit as $x \to 4$ can be calculated by substitution.

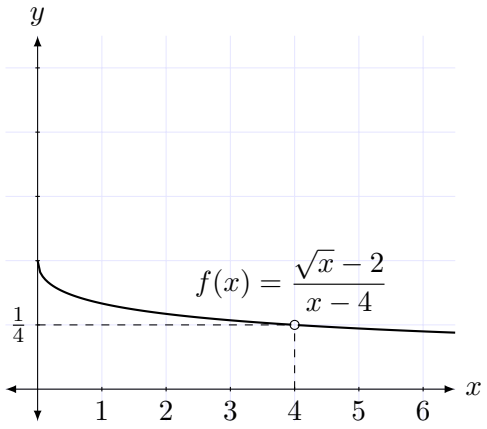



Figure 5.8: This function has a hole discontinuity at $x = 4$; one can algebraically simplify to evaluate the limit. (Note that the scale on the vertical axis is stretched compared to the horizontal axis.)

We can verify that the original function in part (b) of the previous example indeed has a hole discontinuity by graphing; see Figure 5.8. Compare its graph to the graph of the function in part (a), given in Figure 5.7. Note the vertical asymptote in the graph for part (a); even though the function is not continuous at $x = 3$, it is of no concern. The fact that the graph is continuous at the point of interest, $x = 4$, allows us to evaluate the limit as $x$ approaches 4 by substitution.

Now let's suppose we wish to determine the following limit:

$$\lim_{x \to 0} \frac{\sin x}{x}$$

We can use the same reasoning as in the previous examples. Start by recognizing that this function is an algebraic combination of continuous functions, and so is continuous wherever it is defined. This suggests substituting 0 for $x$ to attempt to evaluate the limit. However, the denominator is 0 when $x = 0$, which means that the function is not defined at $x = 0$, and therefore not continuous there. This means that we can't evaluate the limit by substitution.

However, both numerator and denominator are 0 when $x = 0$. This fits the pattern of the previous examples, where we were able to algebraically manipulate the numerator and denominator, cancel a troublesome factor, and then evaluate the limit by substitution. This gives us some hope that maybe a similar technique will work here, but unfortunately there seems to be no algebraic simplification that helps with this limit.

It's difficult to see how to proceed here. Perhaps it might help to produce a table of values. Should we set our calculator to degrees or radians? That's not clear either, so let's try both. Calculations for $x$ in degrees are in Table 5.3, and calculations for $x$ in radians are in Table 5.4.

Table 5.3: Numerical calculations for $\lim_{x \to 0} \frac{\sin x}{x}$, where $x$ is in degrees.

| $x$ | $f(x) = \frac{\sin x}{x}$ | | $x$ | $f(x) = \frac{\sin x}{x}$ |
|---|---|---|---|---|
| $-1$ | 0.034899 | | 1 | 0.034899 |
| $-0.5$ | 0.034905 | | 0.5 | 0.034905 |
| $-0.1$ | 0.0349065 | | 0.1 | 0.0349065 |
| $-0.01$ | 0.03490658433 | | 0.01 | 0.03490658433 |
| $-0.001$ | 0.03490658503 | | 0.001 | 0.03490658503 |

Table 5.4: Numerical calculations for $\lim_{x \to 0} \frac{\sin x}{x}$, where $x$ is in radians.

| $x$ | $f(x) = \frac{\sin x}{x}$ | | $x$ | $f(x) = \frac{\sin x}{x}$ |
|---|---|---|---|---|
| $-1$ | 0.841471 | | 1 | 0.841471 |
| $-0.5$ | 0.958851 | | 0.5 | 0.958851 |
| $-0.1$ | 0.998334 | | 0.1 | 0.998334 |
| $-0.01$ | 0.999983 | | 0.01 | 0.999983 |
| $-0.001$ | 0.99999983 | | 0.001 | 0.99999983 |

Notice from the tables that the values for $\frac{\sin x}{x}$ are repeated for negative and positive values of $x$; this makes sense because $\frac{\sin x}{x}$ is an even function (which follows because both $y = x$ and $y = \sin x$ are odd functions). You can verify this as follows, using $f(x) = \frac{\sin x}{x}$; to prove that $f$ is

an even function, we must show that $f(-x) = f(x)$.

$$\begin{aligned} f(-x) &= \frac{\sin(-x)}{-x} \\ f(-x) &= \frac{-\sin x}{-x} \\ f(-x) &= \frac{\sin x}{x} \\ f(-x) &= f(x) \end{aligned}$$

This proves that $\dfrac{\sin x}{x}$ is an even function. Too bad we didn't notice this before we constructed the tables, as it would have saved us half the work!

Now what can we conclude from the numerical calculations in the tables? Let's look at the table in degrees first. Notice that coming from the left, or coming from the right, it seems that we approach a similar number. However, it's not clear whether either of the sets of numbers represents overestimates or underestimates; in the absence of such an understanding, we have no way of knowing what the limit is (assuming it exists, which it might not), and we can't even say it's definitely between two numbers. The best we can do is to say that if the limit exists, it might be near 0.0349, but we can't be sure.

Similar considerations apply to the table in radians. It seems that the limit is close to 1, if it exists, but how can we be sure?

Perhaps a graph can help us understand the situation. Let's look at the graph of $y = \dfrac{\sin x}{x}$. But wait, this is not an easy graph to draw, particularly near $x = 0$. Perhaps we could make do by just analyzing the graph of $y = \sin x$, because that is a familiar graph.

**Question:** Is there a way of visualizing values of $\dfrac{\sin x}{x}$ from the graph of $y = \sin x$? If so, how?

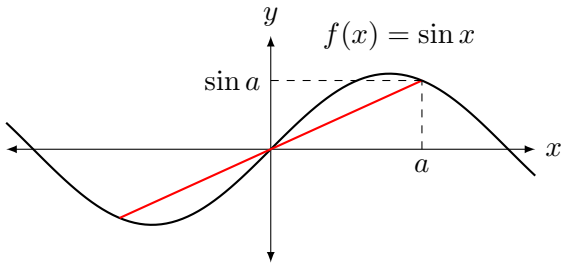



Figure 5.9: The value of $\dfrac{\sin a}{a}$ is the slope of the secant line joining $(0, 0)$ and $(a, \sin a)$.

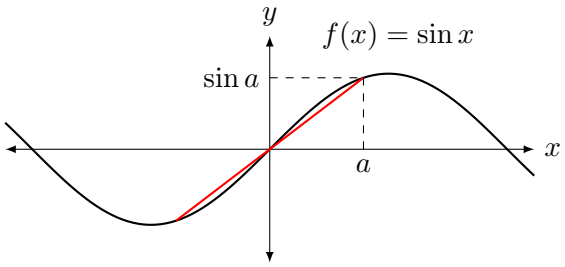



Figure 5.10: It appears that as $a \to 0$, the slope of the secant line approaches the slope of the tangent line to the graph of $f(x) = \sin x$ at $(0, 0)$.

Observe from Figure 5.9 that a value of $\dfrac{\sin x}{x}$, for $x = a$, is the slope of a tangent line joining $(0, 0)$ to $(a, \sin a)$. Notice that the same secant line is suitable for two values of $x$, one positive and one negative; this explains geometrically the matching numbers in Tables 5.3 and 5.4. Figure 5.9 shows a value of $a$ that is closer to 0.

**Question:** Can you see from the graph that as the value of $x$ gets closer and closer to 0, from either the right or the left, that the slope of the secant line seems to get closer and closer to the tangent line to the graph at $(0, 0)$? Can this be true? How can we check this?

Let's recall the definition of the derivative, and then use it to write an expression for the derivative of the function $f(x) = \sin x$ at $(0, 0)$.

$$\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{f(0+h) - f(0)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{f(h) - f(0)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{\sin h - \sin 0}{h} \\
f'(0) &= \lim_{h\to 0} \frac{\sin h - 0}{h} \\
f'(0) &= \lim_{h\to 0} \frac{\sin h}{h}
\end{aligned}$$

The previous equation confirms what appeared to be true from the graph: The limit in question,

$$\lim_{x\to 0} \frac{\sin x}{x}$$

represents the slope of the graph of $f(x) = \sin x$ at the point $(0, 0)$.[2] Based on this new insight, it certainly seems as if the limit exists; the graph is nice and smooth at $(0, 0)$, and it appears that the slope is about 1. This suggests that Table 5.4 is most relevant, and it too suggests that the limit might be about 1. However, none of this is conclusive; we can't say for sure what the limit is, or even that the limit exists.

We'll interrupt the discussion of this limit now; you'll see convincing arguments that confirm that the limit really is 1 when you continue your studies in the differential calculus of trigonometric functions. In the mean time, you might think about the mysterious numbers in Table 5.3 and what they mean. That mystery will also be cleared up in your further calculus studies.[3]

As an overall conclusion for this section, this way of thinking about a limit—as a sort of "trend" in the values of a function as the $x$-values change—is good enough for relatively simple situations. However, it can only take us so far; more complex situations require more subtle and powerful means, as we saw in the discussion of the previous example. We'll continue to develop more powerful means for evaluating limits in the rest of this book.

Also note that most books take a logical approach, which is not necessarily the best approach for learning, nor is it necessarily the practical approach that is most commonly used by practitioners for actually calculating limits. For instance, consider our practical recommendation to evaluate a limit by substitution if the function is continuous. Most books take that as the *definition* of continuity, as we shall also do, once we get around to it. The approach in this book is to present practical, concrete calculations first, and leave logical development and theory for later in the book.

---

[2] Remember Mr. Shakespeare's poetry about a rose by any other name smelling as sweetly. Whether we say $\lim_{x\to 0} \frac{\sin x}{x}$, or $\lim_{a\to 0} \frac{\sin a}{a}$, or $\lim_{h\to 0} \frac{\sin h}{h}$, all three expressions represent exactly the same thing.

[3] We'll also shed further light on this mystery in Chapter 10, if you can't wait for a university calculus course. But do think about it yourself first!

## EXERCISES

(Answers at end.)

Determine each limit. Sketch a graph of the function to check graphically whether your limit calculation is correct.

1. $\lim_{x\to 3} \dfrac{1}{x-2}$
2. $\lim_{x\to 3} \dfrac{x-3}{x-2}$
3. $\lim_{x\to \pi} \dfrac{\sin x}{x}$
4. $\lim_{x\to -1} \left(x^2-3x+5\right)$
5. $\lim_{x\to 0} \dfrac{x^2-4}{x+2}$
6. $\lim_{x\to -2} \dfrac{x^2-4}{x+2}$
7. $\lim_{x\to 0} \dfrac{x^2-4}{x-2}$
8. $\lim_{x\to 2} \dfrac{x^2-4}{x-2}$
9. $\lim_{x\to 4} \dfrac{x-9}{\sqrt{x}-3}$
10. $\lim_{x\to 9} \dfrac{x-9}{\sqrt{x}-3}$
11. $\lim_{x\to 3} \dfrac{x^2-9}{x+4}$
12. $\lim_{x\to 2} \dfrac{x^2-9}{x+4}$
13. $\lim_{x\to 0} \dfrac{x^2}{x}$
14. $\lim_{x\to 3} \dfrac{x^2}{x}$
15. $\lim_{x\to 4} \dfrac{\sqrt{x}-1}{x^2-1}$
16. $\lim_{x\to 1} \dfrac{\sqrt{x}-1}{x^2-1}$
17. $\lim_{x\to -1} \dfrac{2x^2-2x-4}{x^2+x-6}$
18. $\lim_{x\to 2} \dfrac{2x^2-2x-4}{x^2+x-6}$
19. $\lim_{x\to 1} \dfrac{x^3-7x+6}{x^3-2x^2-x+2}$
20. $\lim_{x\to 2} \dfrac{x^3-7x+6}{x^3-2x^2-x+2}$

---

Answers: 1. 1; 2. 0; 3. 0; 4. 9; 5. $-2$; 6. $-4$; 7. 2; 8. 4; 9. 5; 10. 6; 11. 0; 12. $-5/6$; 13. 0; 14. 3; 15. 1/15; 16. 1/4; 17. 0; 18. 6/5; 19. 2; 20. 5/3

## SUMMARY

This section presented a practical strategy for determining some limits, and provided opportunities for you to practice this skill. The concept of a continuous function was introduced, and a theorem about which functions are continuous was stated. For a function that is continuous at $x = a$, one can evaluate the limit of the function as $x \to a$ by substituting $a$ for $x$ into the formula for the function. For a function that has a hole discontinuity at $x = a$, one can evaluate the limit of the function as $x \to a$ by "filling in the hole." Not all limits can be effectively evaluated using these techniques; we'll learn about how to tackle other limits later in the book.

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

## HISTORY

### Ghosts of departed quantities

Calculus was developed by many workers, and their incremental progress was independently systematized by Newton and Leibniz in the late 1600s. At that time the concept of limit had not been devised yet. Even the concept of a function was still in development, and there was not yet a precise definition of a function. The term "function" appears to have been introduced by Leibniz in 1673. Thus, calculus was developed in its early days by discussing "variable quantities," which we would nowadays call variables, and the currently-accepted definition of a function was not formulated until the late 1800s, as was the currently-accepted definition of a limit. Place yourself in the shoes of Newton and Leibniz in the late 1600s, then, striving to make sense of their newly-created systems without having adequately precise definitions to work with. They were like searchers groping in a dark cave, possessing some very unusual night-vision goggles, and yet not quite able to see clearly. In this light, their progress appears all the more remarkable.

To make sense of calculus, Leibniz thought in terms of *infinitesimals* (Newton tried to avoid this concept, but had similar ideas). (d'Alembert was the first to think of a derivative in terms of limits in the 1700s.) An infinitesimal was conceived as a number that is smaller in magnitude than any real number, but not yet zero. It should be emphasized that there is no such real number! This point was made forcefully by George Berkeley in his 1734 book *The Analyst*, which was subtitled:

> A Discourse Addressed to an Infidel Mathematician: Wherein It Is Examined Whether the Object, Principles, and Inferences of the Modern Analysis Are More Distinctly Conceived, or More Evidently Deduced, Than Religious Mysteries and Points of Faith. "First Cast the Beam Out of Thine Own Eye; and Then Shalt Thou See Clearly to Cast Out the Mote Out of Thy Brother's Eye."

Berkeley had earlier attacked "free-thinkers" in response to their attacks on Christianity. Sir Edmund Halley, a noted free-thinker and devotee of Newton, mocked Berkeley's attacks, and apparently a sick friend of Berkeley's had refused Berkeley's "spiritual consolation, because Halley had convinced the friend of the untenable nature of Christian doctrine." (See page 470 of Boyer's *A History of Mathematics*.) It is speculated that "The Infidel Mathematician" in Berkeley's subtitle is Halley, and that the book was a response to Halley. (It is doubtful that the devoutly religious Newton was Berkeley's target.)

On the one hand, "Can't we all just get along?" and on the other hand, Berkeley's criticisms about the foundations of calculus were on point. Berkeley did not dispute that the results of calculus were valid (their applications in astronomy by Newton and others had been empirically supported), he simply, and correctly, pointed out that the reasoning that produced these valid

results was dodgy. In Berkeley's words (fluxions were Newton's alternative to infinitesimals),

> And what are these fluxions? The velocities of evanescent increments. And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them ghosts of departed quantities?

The last sentence is pretty biting, and makes Berkeley's point memorable. You can't say some quantity has been incremented (the $h$ in our limit arguments), then divide by this quantity as if it were non-zero, and then later suppose that this quantity is ignorable (i.e., zero), without some careful justification. We have tried to provide such careful argumentation, rough though it be, but Newton and his contemporaries did not quite do the job. But let's not be critical of them; it took two centuries of hard work by many very bright researchers to finally figure this out to the general satisfaction of the community. Better, more precise, argumentation (which reflects the conclusions of this two centuries of hard work) is found in the theory chapters at the end of this book.

The moral of this story is that creative mathematicians come up with all kinds of interesting ideas, many of which are practical and some of which are even revolutionary. But it is too much to ask of any one person, or even of any one generation of workers, to tidy up every loose end in these new fields of mathematics. The tidying-up process is also creative, but in a different sense; logic comes to the fore in the tidying-up process. Once a field is mature, then clear definitions and axioms are identified, and theorems are derived in a coherent, systematic way from the foundations. Calculus is by now a very mature field of mathematics, and if you dig more deeply into the subject you will be able to study its foundations to your heart's content. But when first learning a subject, it is beneficial to focus on numerous examples to internalize the main concepts, problems, and methods, and to save a deeper consideration of foundational issues for later study.

# Chapter 6

# Left Limits and Right Limits

**OVERVIEW**

The concepts of left limit and right limit are introduced and related to the previously developed concept of limit. These so-called "one-sided limits" are useful in studying the behaviour of functions, particularly near points of discontinuity and near asymptotes.

So far in this book we have performed quite a number of limit calculations. From the perspective of the "trend" aspect of a limit, we've considered what the trend in the function values is as $x$ approaches a certain number. To be more specific, we looked at the trend as $x$ approaches a certain number both from the left and from the right.

The following definition formalizes the idea of looking at the trends from the left and right separately.

**DEFINITION 2**

**Left limits and right limits (i.e., one-sided limits)**

- The *left limit* of $f$ as $x$ approaches $a$ exists and is equal to the number $L$, that is $\lim_{x \to a^-} f(x) = L$, provided that the function values of $f$ get closer and closer to $L$ as $x$ gets closer and closer to $a$ but $x < a$.
- The *right limit* of $f$ as $x$ approaches $a$ exists and is equal to the number $L$, that is $\lim_{x \to a^+} f(x) = L$, provided that the function values of $f$ get closer and closer to $L$ as $x$ gets closer and closer to $a$ but $x > a$.

Informally, the left limit of $f$ as $x$ approaches $a$ means the trend of the function values as $x$ approaches $a$ from the left (that is, for values of $x$ that are less than $a$). Similarly, the right limit of $f$ as $x$ approaches $a$ means the trend of the function values as $x$ approaches $a$ from the right (that is, for values of $x$ that are greater than $a$).

Why would we wish to define left and right limits? Well, there are some functions for which the trend in function values when you approach some $x$-value from the left is different from the trend in function values when you approach from the right. In this case, we say that the limit does not exist, but nevertheless, it is often of value to understand the trends in each direction.

**THEOREM 2**

**Characterization of a limit in terms of left and right limits**

- **Part 1**: If $\lim_{x\to a^-} f(x) = L$ and $\lim_{x\to a^+} f(x) = L$, then $\lim_{x\to a} f(x) = L$.
- **Part 2**: If either $\lim_{x\to a^-} f(x)$ or $\lim_{x\to a^+} f(x)$ do not exist, or if they both exist but $\lim_{x\to a^-} f(x) \neq \lim_{x\to a^+} f(x)$, then $\lim_{x\to a} f(x)$ does not exist.

The first part of the theorem states formally what we have been doing all along when calculating limits: We look at the trends from each direction, and if they both exist and are equal, then we say the limit exists. The second part of the theorem adds new information: if the trends from each direction are not equal, or if either does not exist, then we say that the limit does not exist.

For which sort of function would the left limit and right limit be unequal at some point? Consider most of the examples that we have seen so far. For a continuous function, searching for such an example seems pointless. After all, at a point of continuity, the limit of function values exists and is equal to the function value at the point in question, so the left and right limits at the same point must be equal. How about a function with a hole discontinuity? No, that doesn't work either, because at a hole discontinuity, the limit of the function also exists, and is equal to the $y$-value at the hole.

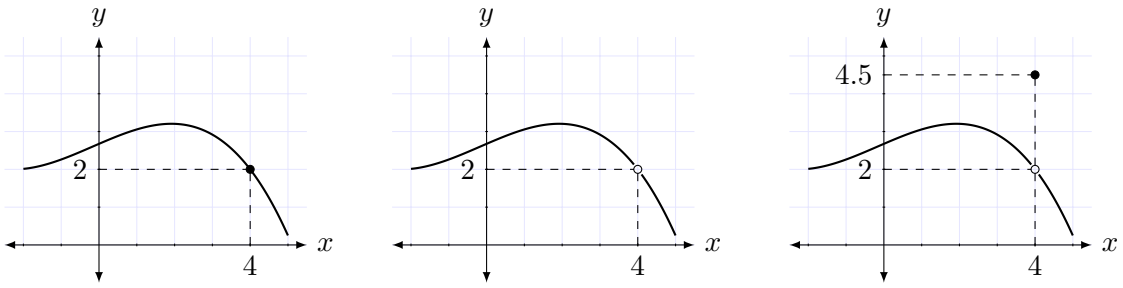



Figure 6.1: The three functions are slightly different, but for each function, the limit of the function as $x$ approaches 4 is 2. For each of these three functions, the left limit as $x \to 2$ is equal to the right limit as $x \to 2$.

Therefore, a function for which the left limit is not equal to the right limit at a point has to have a more serious discontinuity, such as a jump discontinuity at the point. Here is one example.

## EXAMPLE 12

**A function for which the left and right limits at a point are not equal**

Determine the limit

$$\lim_{x\to 0} \frac{|\,x\,|}{x}$$

## SOLUTION 1

Since the function $f(x) = \dfrac{|\,x\,|}{x}$ is not defined for $x = 0$, we won't be able to evaluate this limit by substitution. The function $f$ is similar in structure to the function $\dfrac{\sin x}{x}$, whose limit we studied in the previous section. Why don't we use the same method of analysis: Rather than try to determine $\lim_{x\to 0} \dfrac{|\,x\,|}{x}$ directly, let's consider the slope of the function $y = |\,x\,|$ at $x = 0$.

That is, consider the secant line joining the points $(a, f(a))$ and $(0, 0)$ on the graph of $f$. The slope of the secant line is

$$\text{slope of secant line} = \frac{f(a) - 0}{a - 0} = \frac{f(a)}{a}$$

which is exactly the expression of which we wish to determine the limit.

So what is the trend of the values of the slope of the secant line as $a \to 0$? Well, if $a > 0$, we can see from the graph in Figure 6.2 that the slope is 1, no matter what the value of $a$ is. Similarly, if $a < 0$, we can see from the graph that the slope of the secant line is $-1$, no matter what the value of $a$ is. This information is recorded in Figure 6.3.

Thus, because

$$\lim_{x\to 0^+} \frac{|\,x\,|}{x} = 1 \qquad \text{and} \qquad \lim_{x\to 0^-} \frac{|\,x\,|}{x} = -1$$

this means that

$$\lim_{x\to 0^+} \frac{|\,x\,|}{x} \neq \lim_{x\to 0^-} \frac{|\,x\,|}{x}$$

It follows that

$$\lim_{x\to 0} \frac{|\,x\,|}{x} \qquad \text{DOES NOT EXIST}$$

## SOLUTION 2

An alternative solution is purely algebraic, and does not rely on the graphs. (However, you can see that the essence of this solution is the same as the essence of the first solution.)

Recall that

$$|\,x\,| = \left\{ \begin{array}{rl} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{array} \right\}$$

Thus, to calculate the left and right limits of $f$, replace $|\,x\,|$ by the appropriate simpler (i.e., without absolute values signs) expression depending on whether $x > 0$ or $x < 0$. That is:

$$\begin{aligned}\lim_{x\to 0^+} \frac{|\,x\,|}{x} &= \lim_{x\to 0^+} \frac{x}{x} \\ &= \lim_{x\to 0^+} 1 \\ &= 1\end{aligned}$$

and

$$\begin{aligned}\lim_{x\to 0^-} \frac{|\,x\,|}{x} &= \lim_{x\to 0^-} \frac{-x}{x} \\ &= \lim_{x\to 0^-} -1 \\ &= -1\end{aligned}$$

Because

$$\lim_{x\to 0^+} \frac{|\,x\,|}{x} \neq \lim_{x\to 0^-} \frac{|\,x\,|}{x}$$

it follows that

$$\lim_{x\to 0} \frac{|\,x\,|}{x} \qquad \text{DOES NOT EXIST}$$

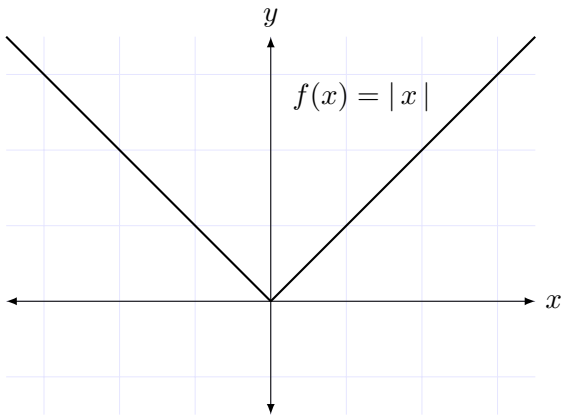



Figure 6.2: To calculate $\lim_{x\to 0} |\,x\,|/x$ think in terms of the slope of the secant line joining $(0,0)$ to another point on the graph of $y = |\,x\,|$. What happens to the slope of the secant line as $x \to 0$?

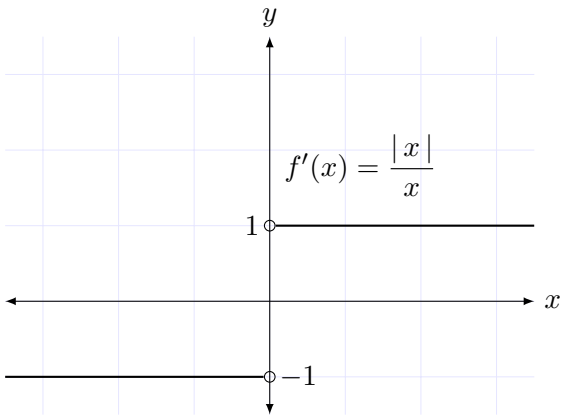



Figure 6.3: It turns out that this function is the derivative of the one in the figure on the left; see the text for details.

In the previous example, notice that the algebraic Solution 2 is faster than Solution 1, but the geometric reasoning in Solution 1 gives us insight into why the limit does not exist. Both solutions are of value, and its worthwhile studying them together until you understand that their essence is the same.

The following calculation confirms that the limit $\lim_{x\to 0} \frac{|\,x\,|}{x}$ is the derivative of the function

$f(x) = |\,x\,|$ at $x = 0$:[1]

$$\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{f(0+h) - f(0)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{f(h) - f(0)}{h} \\
f'(0) &= \lim_{h\to 0} \frac{|\,h\,| - |\,0\,|}{h} \\
f'(0) &= \lim_{h\to 0} \frac{|\,h\,| - 0}{h} \\
f'(0) &= \lim_{h\to 0} \frac{|\,h\,|}{h}
\end{aligned}$$

Since it doesn't matter what symbol we use in place of $h$, it's equally well true that

$$f'(0) = \lim_{x\to 0} \frac{|\,x\,|}{x}$$

This justifies the geometric reasoning in Solution 1 of the previous example.

One of the conclusions we can draw from the previous example is that the derivative of the function $f(x) = |\,x\,|$ does not exist at $x = 0$. This is interesting new information, as we've now experienced a function that is not differentiable at one point in its domain. We can see from the graph the geometric reason for this: The graph has a sharp corner at $(0, 0)$. Another way to say this is that it's not possible to draw a tangent line to the graph at the corner point $(0, 0)$. We infer that for a function to be differentiable at a point, its graph must be "smooth" at that point.

In general, functions for which their graphs have jump discontinuities at $x = a$ have unequal left and right limits as $x$ approaches $a$. This was one of the prime motivations for introducing the idea of left and right limits; so that we can analyze functions that have jump discontinuities, and so we have a vocabulary for describing their behaviour at the point of discontinuity.

[1] I'll let you confirm that the function $\dfrac{|\,x\,|}{x}$ is the derivative of $f$ at all other values of $x$ as well.

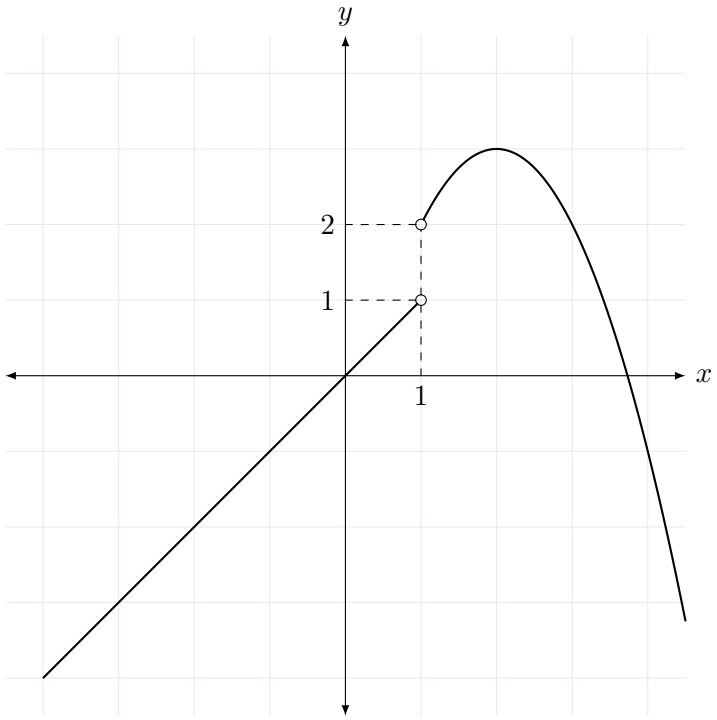



Figure 6.4: A function defined piecewise that has a jump discontinuity at $x = 1$: the function is defined by $f(x) = 3 - (x-2)^2$ for $x > 1$, and $f(x) = x$ for $x < 1$.

## EXAMPLE 13

**Left limit and right limit for a function defined piecewise**

Consider the function

$$f(x) = \left\{ \begin{array}{ll} 3-(x-2)^2 & \text{if} \quad x > 1 \\ x & \text{if} \quad x < 1 \end{array} \right\}$$

Determine $\lim\limits_{x \to 1} f(x)$.

## SOLUTION

Because the function is defined piecewise, and the pieces are separated just at the position where the limit is required, it makes sense to use left and right limits. For the right limit, we use the definition of the function towards the right of the position where the limit is required:

$$\begin{array}{rcl} \lim\limits_{x \to 1^+} f(x) & = & \lim\limits_{x \to 1^+} \left(3-(x-2)^2\right) \\ \lim\limits_{x \to 1^+} f(x) & = & 2 \end{array}$$

For the left limit, we use the definition of the function towards the left of the position where the limit is required:

$$\lim_{x\to 1^-} f(x) = \lim_{x\to 1^-} x$$
$$\lim_{x\to 1^-} f(x) = 1$$

Note that the left limit and the right limit were obtained by substitution. Does this make sense? Write a few sentences to justify this.

The left and right limits are not equal, and so it follows that

$$\lim_{x\to 1} f(x) \quad \text{DOES NOT EXIST}$$

A graph of the function $f$ is shown in Figure 6.4. Note from the graph that the function $f$ has a jump discontinuity at $x = 1$. It is possible to conclude this without looking at the graph, by comparing the left and right limits of $f$ at $x = 1$. Briefly explain.

**EXERCISES**

(Answers at end.)

1. Consider the function $f$ defined by $f(x) = \left\{ \begin{array}{ll} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0 \end{array} \right\}$

(a) Determine $\lim_{x\to 0^+} f(x)$, $\lim_{x\to 0^-} f(x)$, $\lim_{x\to 0} f(x)$

(b) Determine $\lim_{x\to 5^+} f(x)$, $\lim_{x\to 5^-} f(x)$, $\lim_{x\to 5} f(x)$

2. For the previous exercise, sketch a graph of the function to check graphically whether your limit calculations are correct. Classify each discontinuity as either a hole discontinuity or a jump discontinuity. Explain briefly how you can determine the nature of the discontinuity from the limit calculations.

3. Repeat the two previous exercises for the function

$$f(x) = \left\{ \begin{array}{ll} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{array} \right\}$$

Compare and contrast the results for the two functions.

---

Answers: 1.(a) 1, $-1$; does not exist; (b) 1, 1, 1;

2. Jump discontinuity at $x = 0$; this can be seen because the left and right limits are not equal at $x = 0$.

3. (a) 1, $-1$; does not exist; (b) 1, 1, 1; there is a jump discontinuity at $x = 0$. The point is that the actual value of the function at $x = 0$ (which is different in the two cases) does not change the values of the various limits at $x = 0$.

## HISTORY

**Leonhard Euler (1707–1783)**

Leonhard Euler was one of the greatest mathematicians in history, and certainly the most prolific. His complete works contains over 850 items, including works on mathematics, astronomy, and physics, three books on calculus, and several other subjects. (See `http://eulerarchive.maa.org/` for more information.) According to historian of mathematics Carl Boyer, Euler's two-volume introduction to calculus is the greatest mathematics textbook of modern times. Euler's writings were clear, and so very useful for students to learn from, yet they also contained research-level discoveries, often opening up entire new fields of mathematics. Simmons says (referencing C.B. Boyer in Section A.21 of *Calculus Gems*) that, "There is considerable truth in the old saying that all elementary and advanced calculus textbooks since 1748 are essentially copies of Euler or copies of copies of Euler."

Euler introduced symbols and notation that are still used today; for example, he introduced the symbol $e$ for the base of natural logarithms, and $i$ for the "imaginary unit." He discovered the remarkable formula, now known as Euler's formula

$$e^{i\theta} = \cos\theta + i\sin\theta$$

For the special case $\theta = \pi$, this becomes

$$e^{i\pi} = -1, \qquad \text{which is equivalent to} \qquad e^{i\pi} + 1 = 0$$

which is also known as Euler's formula. For some people, the second formula on the previous line is preferred to the first formula on the previous line, because the second formula contains what are considered the five most important symbols in mathematics all in one formula. If you are planning to become a mathematics or physics student at university, it is well worth digging into the more general version of Euler's formula, as understanding it and learning how to use it will help you a lot. (Plus, it is a lot of fun!)

Euler was born in Basel, Switzerland, and moved with his family to a small village shortly afterwards. When he was eight years old he was sent back to Basel to live with his maternal grandmother, and he later enrolled in the University of Basel, where he was taught by Johann Bernoulli. Most of his career was spent at the St. Petersburg Academy and the Berlin Academy.

Euler and his wife had 13 children, only five of whom survived childhood.

At the age of 28, Euler nearly died from a feverish illness, and three years later he became nearly blind in his right eye. He is reported to have taken this loss in good spirit, saying that now he would have fewer distractions. The sight in his right eye became progressively worse, and by the age of 59 he had also lost the sight in his left eye due to a cataract, so that he was completely blind. Amazingly, his productivity increased! He used his tremendous memory, powerful imagination, and the help of assistants to transcribe his dictation, and just kept working.

## SUMMARY

One-sided limits were introduced in this chapter, and they were used to study the behaviour of functions near points of discontinuity.

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 7

# Continuity

**OVERVIEW**

This chapter presents the technical definition of a continuous function in terms of limits. (Until now we have used an informal concept of continuous function: A continuous function is one for which its graph can be sketched without lifting one's pencil from the paper.) The concept of continuity is then used to develop the intermediate value theorem, an important tool in solving equations.

## 7.1 Continuous Functions

In using functions as models of realistic phenomena, by far the most useful kinds of functions are continuous ones. Among the continuous functions, the vast majority of them that are useful in applications are also differentiable. Differentiable functions are the most amenable to analysis by the most powerful tools available (calculus!), so it's a happy situation that these kinds of functions are also the most useful in applications.

Continuous functions that are not differentiable, and also discontinuous functions, are occasionally used in applications, and so if you are most interested in scientific or engineering applications of functions, you will need to learn something about non-differentiable and discontinuous functions as well.

For mathematicians, there is a spirit of seeking out all possibilities in every context, without necessarily being concerned about possible applications. This is useful, for several reasons. First, one is mindful of mathematics history, which contains numerous examples of supposedly correct statements that were later shown to be incorrect because of carefully and creatively constructed counterexamples. Thorough exploration, combined with iron-clad proofs of theorems, combine to give us certainty about what is possible and what is not possible in various contexts, and this makes for a healthy development of mathematics. Secondly, although this is a purely mathematical consideration, with no thought about applications, nevertheless one can then confidently apply mathematics in applications, as one can be sure about whether the relevant mathematics applies.

Earlier in this book we have used an intuitive sense of continuity: A function is continuous if its graph has no breaks or holes in it. One often reads that a function is continuous provided that its graph can be sketched in one piece without lifting one's pencil from the paper.

This sense of continuity is good enough for many purposes, but like almost everything we learn in mathematics, it must be sharpened (i.e., defined precisely) when moving on to advanced work. One of its deficiencies is that it relies too much on a graphical sense for its definition; how do we tell if a function is continuous just from the formula, if it is too difficult to sketch a very accurate

graph? Another deficiency is that this intuitive sense of continuity is just plain wrong if the domain of the function is not the real numbers. If you move on to advanced work, you might be a little surprised to find out that any function whatsoever whose domain is just the natural numbers is continuous, according to the precise definition of continuity. Yet such a function certainly doesn't look continuous, and can't be sketched without lifting one's pencil from the paper.

So how can we improve on the intuitive sense of continuity? As we saw earlier, if a function is continuous at a point, the limit of the function as $x$ approaches that point can be calculated by substitution. The formal definition of continuity turns this around and adopts this property, which we have already used, as the formal definition. This fits in with the general modern strategy of defining all new concepts in calculus in terms of limits, wherever possible.

Here is the formal definition of continuity:

**DEFINITION 3**

**Continuous function**

A function $f$ is continuous at $x = a$ provided that all of these conditions are satisfied:

- $\lim_{x \to a} f(x)$ exists
- $f(a)$ exists
- $\lim_{x \to a} f(x) = f(a)$

The first two conditions in the definition can be omitted if we agree that they are implicit in the third condition, but they are included for clarity. Note that the first condition eliminates the possibility of a jump discontinuity, and the second and third conditions together eliminate the possibility of a hole discontinuity. If all three conditions are satisfied, then there can't be either a jump discontinuity or a hole discontinuity, so the function must be continuous at $x = a$.

It would be interesting and instructive to come up with specific examples (in the form of graphs) for which just one of the first two conditions is violated. Try it!

If a function is continuous at each point in its domain, then we simply say that the function is continuous.

Just because a function is continuous does not guarantee that it is differentiable. A good example to keep in mind is one we encountered in the previous section; recall that the function $f(x) = |\, x \,|$ is continuous for all values of $x$, but is not differentiable at $x = 0$.

However, the opposite implication is true: If a function is differentiable at $x = a$, then it is guaranteed to be continuous at $x = a$. This is proved in the theory chapters towards the end of the book.

It sometimes happens that one wishes to create a mathematical model by piecing together two or more common types of functions. Such functions are said to be defined piecewise. Here is an example:

$$f(x) = \left\{ \begin{array}{ll} x^2 & \text{if } x \geq -1 \\ x & \text{if } x < -1 \end{array} \right\}$$

Notice from its graph in Figure 7.1 that this function is **not** continuous at $x = -1$, although it is continuous for every other value of $x$. This is reasonable, because each piece is a part of a continuous function, but the pieces don't fit together at $x = -1$. You can verify this by calculating

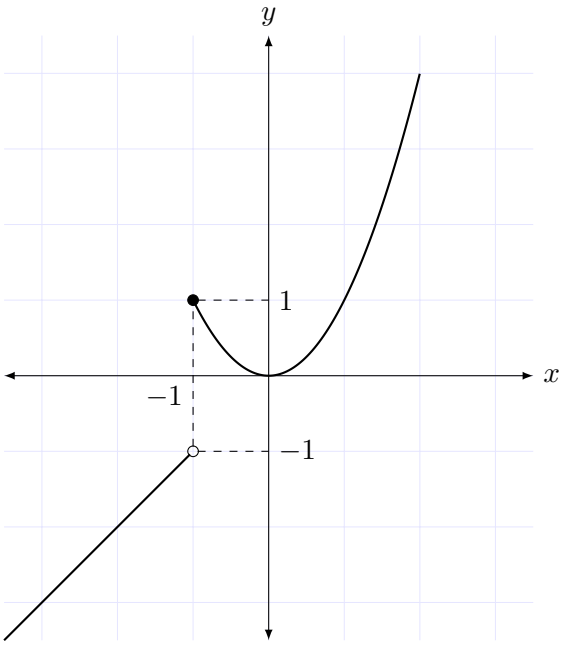



Figure 7.1: A piecewise defined function that is not continuous at $x = -1$: the function is defined by $f(x) = x^2$ for $x \geq -1$, and $f(x) = x$ for $x < -1$.

the left and right limits of the function as $x$ approaches $-1$:

$$\begin{aligned} \lim_{x \to -1^+} f(x) &= \lim_{x \to -1^+} x^2 \\ &= (-1)^2 \\ &= 1 \end{aligned}$$

and

$$\begin{aligned} \lim_{x \to -1^-} f(x) &= \lim_{x \to -1^-} x \\ &= -1 \end{aligned}$$

Since the left and right limits as $x \to -1$ are not equal, the function $f$ is not continuous at $x = -1$.

Is it possible to modify the function $x$ in a simple way to make it continuous? For instance, it seems clear from the graph that by translating the linear piece of the graph upwards by 2 units, the two pieces will fit together and the result will be a continuous function.

Let's look at a few examples of how to solve a problem such as this when the situation is not so straightforward.

## EXAMPLE 14

**Joining two functions to make a continuous functions**

Determine a value of $k$ such that the following function is continuous.

$$g_1(x) = \left\{ \begin{array}{rl} kx^2 & \text{if } x \leq 2 \\ x + k & \text{if } x > 2 \end{array} \right\}$$

## SOLUTION 1

Except possibly at the point $x = 2$, each piece of the function $g_1$ is continuous at all other values of $x$. In order that the function $g_1$ also be continuous at $x = 2$, the two pieces must fit together at $x = 2$. In other words, they must have the same value at $x = 2$. That is,

$$\begin{array}{rcll} kx^2 & = & x + k & (\text{at } x = 2) \\ k(2^2) & = & 2 + k & \\ 4k & = & 2 + k & \\ 3k & = & 2 & \\ k & = & \dfrac{2}{3} & \end{array}$$

Thus, the function $g_1$ is continuous if and only if $k = \dfrac{2}{3}$.

### DISCUSSION

Notice that in Solution 1 we substituted $x = 2$ into the expression $x + k$. However, this can be criticized because the expression is only defined for $x > 2$, so it's not valid to substitute 2 for $x$. One can argue against this criticism by saying, "OK, but we'll just modify the definition of $g_1$ as follows:

$$g_1(x) = \left\{ \begin{array}{rl} kx^2 & \text{if } x \leq 2 \\ x + k & \text{if } x \geq 2 \end{array} \right\}$$

See Figures 7.2 and 7.3, where the original and modified versions of $g_1$ are plotted for the sample value $k = 1$. For this value of $k$, the graph of $g_1$ is not continuous at $x = 2$, but the idea is to modify the value of $k$ in hopes that the two pieces of the function would join up at $x = 2$. Because each piece of the graph is defined at $x = 2$, we are now justified in using the method of Solution 1; however, this graph is no longer a function (if $k \neq 2/3$) because it has two $y$-values at $x = 2$. However, once the two pieces of graph are joined up (that is, once we determine the right value of $k$ that will join the two pieces up), then there will be only one $y$-value at $x = 2$, and so the resulting continuous graph will represent a function.

So it seems that our method ought to be acceptable. Nevertheless, some people might like a more formal procedure, and most textbooks use a solution something along the lines of the following one.

## SOLUTION 2

The function $g_1$ will certainly be continuous at all values of $x$ except $x = 2$, so we only need to

worry about $x = 2$. In order for $g_1$ to be continuous at $x = 2$, the two pieces of graph must match up. This means that the three conditions in the definition of a continuous function must be satisfied by $g_1$ at $x = 2$. Equivalently, we must show that there is a value of $k$ for which $\lim_{x\to 2^+} g_1(x)$, $\lim_{x\to 2^-} g_1(x)$, and $g_1(2)$ all exist and all have the same value.

Let's calculate each one in turn. First, $g_1(2) = 2 + k$. Next, let's work out the limits. First the right limit:

$$\begin{aligned}\lim_{x\to 2^+} g_1(x) &= \lim_{x\to 2^+} x + k \\ &= 2 + k\end{aligned}$$

So far, so good. Finally, the left limit is

$$\begin{aligned}\lim_{x\to 2^-} g_1(x) &= \lim_{x\to 2^+} kx^2 \\ &= k(2^2) \\ &= 4k\end{aligned}$$

The right limit and the value of the function at $x = 2$ are equal no matter what value of $k$ is chosen. However, in order for them also to be equal to the left limit, $k$ must satisfy:

$$\begin{aligned}4k &= 2 + k \\ 3k &= 2 \\ k &= \frac{2}{3}\end{aligned}$$

Thus, $g_1$ is continuous for all values of $x$ if and only if $k = \dfrac{2}{3}$.

In the previous example, notice that the essentials of Solutions 1 and 2 are the same, although those concerned about technicalities might prefer Solution 2.

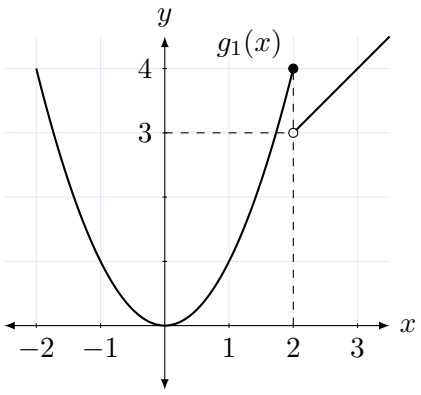



Figure 7.2: The original definition of the function $g_1$ is $g_1(x) = kx^2$ for $x \leq 2$, and $g_1(x) = x + k$ for $x > 2$. The graph is plotted using the sample value $k = 1$.

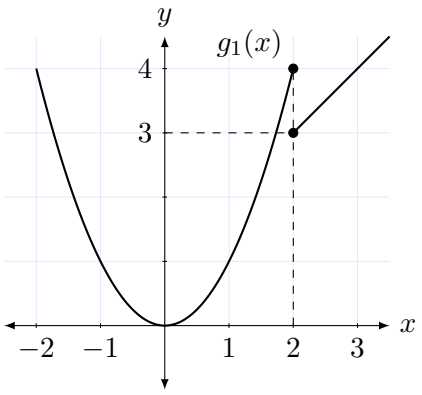



Figure 7.3: The modified definition of $g_1$ is no longer a function for the sample value $k = 1$, because there are two $y$-values at $x = 2$.

**EXAMPLE 15**

**Determining the value of a parameter to make a function continuous**

Determine a value of $k$ such that the following function is continuous.

$$g_2(x) = \left\{ \begin{array}{rl} x^2 + k & \text{if } x \leq 1 \\ kx & \text{if } x > 1 \end{array} \right\}$$

**SOLUTION**

Using the same informal procedure as in Solution 1 of the previous example, we seek a value of $k$ that makes the two pieces of the graph join. This will occur provided we can choose $k$ to satisfy

$$\begin{array}{rcl} (1)^2 + k = & = & k(1) \\ 1 + k & = & k \\ 1 & = & 0 \end{array}$$

Since the last line is an inconsistent equation, no value of $k$ can satisfy the condition and therefore no value of $k$ can be chosen to make the two pieces of graph join up. Thus, $g_2$ is **not** continuous at $x = 1$ for all values of $k$. (Of course, the function $g_2$ is continuous for all other values of $x$.)

## 7.2 The Intermediate Value Theorem

One of the primary activities in mathematics is solving equations. But what do you do if you run into a very complicated equation that you can't immediately solve by using the formulas that you have learned in high school? Well, there are various approximation methods, and some of these have been programmed into computer software, so that it's possible for a computer to chug through an iterative procedure to get a good approximation to the solution. Remember that we discussed approximations early in this book when we were developing the idea behind calculating the slope of a curve. There we made the point that an approximation method is ideal if you have a way of systematically improving the approximation in a step-by-step (i.e., iterative) procedure.

But can you even be sure that your equation has a solution? It would be annoying to waste a lot of time trying to find a solution to an equation when there actually is none. It would be good to know in advance that the equation does have a solution before you start searching for one. Even better would be to know approximately where you should search for a solution; that is, it would be good to know a small interval of the $x$-axis that is sure to contain a solution to the equation.

One of the basic ideas that is used in some approximation schemes for solving equations is as follows: Suppose that for a continuous function $f$, $f(a) < 0$ and $f(b) > 0$. Then there must be at least one value of $x$ between $a$ and $b$ for which[1] $f(x) = 0$. This means that if you can find such values $a$ and $b$, then you know for sure that there is at least one solution to the equation $f(x) = 0$ between $a$ and $b$, so you can get your computer to search for it in confidence, knowing you won't be wasting your time. And you also know where to look for a solution: A solution is sure to be found for some $x$-value between $a$ and $b$.

[1]This can be applied to the solution of any equation in this way: Take your complicated equation and rewrite it so that all of the terms in the equation are brought over to the left side, so that the right side is 0. Then use $f$ to label the complicated function on the left side of the equation.

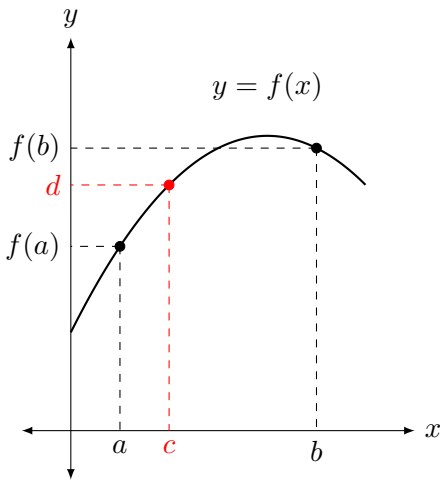



Figure 7.4: An illustration of the intermediate value theorem.

The essence of the previous paragraph is formalized and generalized as the intermediate value theorem:

**THEOREM 3**

**The intermediate value theorem**

Suppose that the function $f$ is continuous for all values of $x$ such that $a \leq x \leq b$, and suppose that $f(a) \neq f(b)$. Choose any $y$-value, call it $d$, such that $d$ is strictly between $f(a)$ and $f(b)$. (That is, either $f(a) < d < f(b)$ or $f(b) < d < f(a)$.) Then there is at least one $x$-value, call it $c$, where $a < c < b$, such that $f(c) = d$.

An informal way to state the intermediate value theorem is that if $f$ is continuous, then as you draw the curve between $(a, f(a))$ and $(b, f(b))$, your pen will cross every $y$-value between $f(a)$ and $f(b)$. This seems obvious when stated like this, doesn't it? (See Figure 7.4.) If it's so obvious why do we bother stating it? There are a couple of reasons. First, it's important to record the most important reasoning principles (theorems) for easy reference, and to draw your attention to them. More importantly, mathematicians have been burned enough times over the centuries by stating that something is obvious, only to learn later (thanks to a deep and persistent thinker) that what they thought was obvious is in fact false! So it has been learned from bitter experience that the most important tools had better be carefully stated and proved, even when they seem obvious.

And the intermediate value theorem is a good case in point. It is true if the domain of the function $f$ is an interval of real numbers, but if the domain is the natural numbers, then the theorem is false. The moral is that intuition is vital but it will only take you so far; it must work hand-in-hand with logic. A proof of the intermediate value theorem is found in the theory section towards the end of the book.

As an application of the intermediate value theorem, consider the equation

$$x \sin x = \cos(x^2)$$

Does the equation have any solutions? It's easiest to apply the intermediate value theorem if we rewrite the equation as

$$x \sin x - \cos(x^2) = 0$$

and then define the function $f$ as

$$f(x) = x \sin x - \cos(x^2)$$

The question about whether the original equation has any solutions can now be translated to, "Does the function $f$ have any zeros?" Let's use a calculator[2] to calculate a few sample values of $f$:

$$\begin{aligned} f(0) &= -1 \\ f(1) &= 0.301169 \\ f(2) &= 2.472238 \\ f(3) &= 1.334491 \\ f(4) &= -2.069551 \\ f(5) &= -5.785824 \end{aligned}$$

Notice the sign changes in the values of $f$ just calculated. Applying the intermediate value theorem[3] to the interval $[0, 1]$, we can conclude that there is a number $c$ such that $0 < c < 1$, for which $f(c) = 0$. Thus, the original equation definitely has at least one solution between 0 and 1. Applying the intermediate value theorem again to the interval $[3, 4]$, we can similarly conclude that the original equation also has at least one solution between 3 and 4.

Of course, there may be many other solutions, and with further work we might locate roughly where they are. (For starters, can you see that $f$ is an even function? If you can convince yourself of this, then you can immediately say that there are at least two more solutions to the original equation, one between $-1$ and 0, and the other between $-4$ and $-3$.) But at least we can get our computer program to approximate the solutions in the rough locations that we have identified so far. Depending on the approximation algorithm, this knowledge might save a lot of time and work.

We'll continue the discussion of an iterative method for solving equations in the last chapter of this book, where we will present the bisection method, a straightforward iterative method for solving equations. The intermediate value theorem is a very useful first step before using the bisection method, as it helps us determine where to look for solutions.

[2] Remember to put your calculator in radian mode.

[3] Do you understand why $f$ is continuous for all $x$ values?

## HISTORY

### Carl Friedrich Gauss (1777–1855)

Gauss was one of the greatest mathematicians in history. He also made important contributions to astronomy, physics, and geodesy (measurements of the Earth). Gauss was a child prodigy, and it is reported that he corrected the arithmetic in one of his father's business documents when he was three years old. His parents could not afford to pay for Gauss's education beyond the local elementary school, but fortunately the Duke of Brunswick funded Gauss's further education. Once Gauss graduated from university with a doctorate, he was appointed as a professor at the University of Göttingen and the director of the astronomical observatory there.

Gauss contributed to numerous branches of mathematics, but for our purposes here I would like to highlight the style of his writing, as compared to, say, Euler. Euler was the most prolific mathematical writer in history, and his works were clearly written, in a way that allowed readers to follow his train of thought as he made his discoveries. Euler even included some false trails or errors that he made, all for the edification of readers. Gauss was the extreme opposite in style. He published very little of his research, preferring to wait until he came up with the most "perfect" exposition of his thoughts. "Few but ripe" was Gauss's motto. In the meantime, his new ideas and discoveries continued rapidly, and so at his death he left an enormous quantity of unpublished notes. His idea of perfect exposition was to make his writing as brief as possible, without any discussion about the thought process that led him to his discoveries, and with no other words that he deemed unnecessary. As a result, his writing was typically difficult to read for other mathematicians. These two aspects of his writing, that he delayed publishing, and that he wrote opaquely, significantly slowed the proliferation of his ideas.

Gauss repeatedly defended his style over the complaints of other mathematicians, saying, for example, that architects do not leave scaffolds up once their buildings are constructed. Unfortunately, his writing style was widely emulated by other mathematicians, and became the standard for good mathematical writing in the past two centuries. As a result, even today's typical mathematics research papers and textbooks are unnecessarily hard to read.

A positive aspect of Gauss's attitudes towards mathematical thinking and writing is that he strove to return mathematical writing to the levels of logical rigour that were present in the times of the ancient Greeks, across all the branches of mathematics in which he worked.

There are other reasons for his reluctance to publish his work, one of which is that he so enjoyed discovery that he preferred to work on new research to polishing for publication work that he had already done. Mathematicians engaged Gauss in discussions about their own research, and frequently found that Gauss had anticipated their discoveries. For example, Simmons says in Section A.25 of *Calculus Gems* that Jacobi "visited Gauss on several occasions to ... tell him about his own most recent discoveries, and each time Gauss pulled 30-year-old manuscripts out of his desk and showed Jacobi what Jacobi had just shown him. ... Such was Gauss, the supreme mathematician. He surpassed the levels of achievement possible for ordinary men of genius in so many ways that one sometimes has the eerie feeling that he belonged to a higher species."

**SUMMARY**

In this section, a definition of the continuity of a function at a point is presented. Then the concept of continuity is used to develop and state the intermediate value theorem, which is a useful tool to help us solve equations. (The intermediate value theorem is an important result that helps us prove other important theorems in calculus, but we shall leave this aspect to more advanced books.)

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 8

# Asymptotes

Asymptotes help us to understand the long-term behaviour of functions — that is, the behaviour of functions far from the origin — and give us a language for describing this behaviour.

## 8.1 Vertical and Horizontal Asymptotes

**OVERVIEW**

In this section, vertical and horizontal asymptotes are defined and techniques for calculating them are developed. They are important because being able to calculate them helps us to describe long-term behaviour of functions used to model processes in time, and they also help us to understand the graphs of various types of functions.

You may recall from high school that certain functions have vertical asymptotes, and others have horizontal asymptotes. Limits give us both the language and the means for determining asymptotes. Even defining an asymptote is difficult without using the language of limits, and unfortunately some high school textbooks don't manage such definitions very well.

One of the simplest examples of a graph that has both vertical and horizontal asymptotes is the graph of the function $f(x) = \dfrac{1}{x}$; see Figure 8.1.

First let's discuss the horizontal asymptote for the graph of $f(x) = \dfrac{1}{x}$. Notice that as the $x$-values move farther and farther from the origin to the right of the graph, the $y$-values get closer and closer to 0. You might try a few values using your calculator:

| $x$ | $y$ |
|---:|---|
| 1 | 1 |
| 5 | 0.2 |
| 10 | 0.1 |
| 100 | 0.01 |
| 1 000 | 0.001 |
| 1 000 000 | 0.000 001 |

We can use limit vocabulary to summarize the behaviour of the function $f(x) = \dfrac{1}{x}$ as the $x$-values move farther and farther from the origin to the right as follows:

$$\lim_{x\to\infty} \frac{1}{x} = 0$$

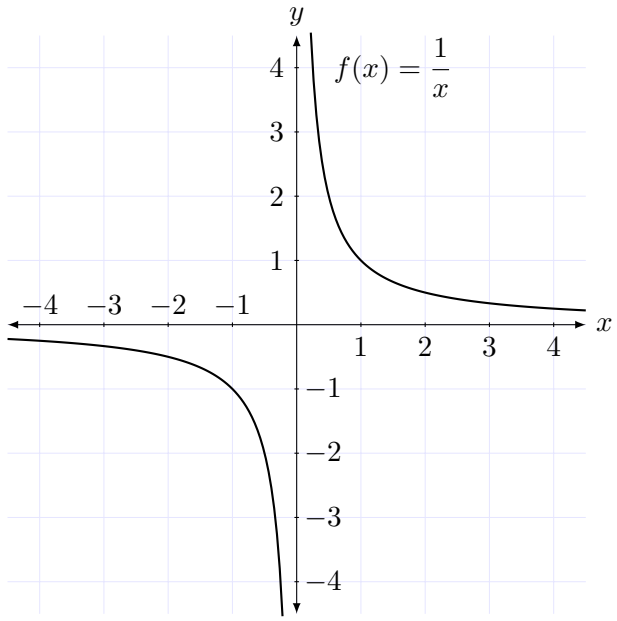



Figure 8.1: The graph of $f(x) = \dfrac{1}{x}$ has a vertical asymptote at $x = 0$ and a horizontal asymptote at $y = 0$.

Don't let the presence of the notation $x \to \infty$ mislead you into thinking that $\infty$ is a place; it is not. Nor is $\infty$ a number; there is no location on the $x$-axis (nor on the $y$-axis) that one can label and say, "Infinity is right here." Rather, $\infty$ is a concept and so one cannot operate with infinity as if it were a number. The notation $x \to \infty$ means, "keep moving to the right along the $x$-axis, indefinitely, getting farther and farther away from the origin, without boundary." Some textbooks use the phrase, "let $x$ become arbitrarily large" as equivalent to $x \to \infty$, and that is a good phrase to use if you like it and understand it.

The behaviour of the function $f(x) = \dfrac{1}{x}$ as the $x$-values move farther and farther from the origin to the left can be summarized as follows:

$$\lim_{x \to -\infty} \frac{1}{x} = 0$$

Using the function $f(x) = \dfrac{1}{x}$ as a prototype, we can define what it means for a function to have a horizontal asymptote as follows:

**DEFINITION 4**

**Horizontal asymptote**

The graph of the function $y = f(x)$ has a horizontal asymptote $y = L$ provided that either or both of the following conditions is satisfied:

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

Now let's go back to the graph of the function $f(x) = \dfrac{1}{x}$ and examine the behaviour of the graph as $x$ approaches 0 from both the left and right. Using your calculator, notice the trend in the function values as $x \to 0$ from the right:

| $x$ | $y$ |
|---|---|
| 1 | 1 |
| 0.1 | 10 |
| 0.01 | 100 |
| 0.001 | 1 000 |
| 0.000 001 | 1 000 000 |

As $x$ gets closer and closer to 0 from the right, the function values get larger and larger, without any boundary. Using limit language, this behaviour can be summarized as:

$$\lim_{x \to 0^+} \frac{1}{x} = \infty$$

Another way to say this is that this limit **does not exist**. The limit does not exist because there is no number that the function values get closer and closer to; rather, as $x \to 0^+$, the function values surpass every number that we might suggest. This means that the use of the equals sign in the previous equation is problematic, because it might mislead some readers into thinking that the limit does exist, and the value of the limit is the number $\infty$. To repeat, $\infty$ is not a number, and the limit in the previous equation does not exist. The "$= \infty$" part of the equation is a brief summary of the reason why the limit does not exist — because the function values increase indefinitely, without bound, to arbitrarily large values.

The potential for confusion means that it would be better if we did not use the equals sign in the previous equation; however, nearly every calculus text uses this notation, so we shall also use it. Just be aware of what the notation means and don't fall into the misconception.

Returning to the graph of $f(x) = \frac{1}{x}$ in Figure 8.1, the behaviour of the graph as $x \to 0$ from the left can be summarized as follows:

$$\lim_{x \to 0^-} \frac{1}{x} = -\infty$$

The "$-\infty$" in the previous equation is a brief way to explain why the limit **does not exist**: The limit does not exist because as $x$ approaches 0 from the left, the function values plunge lower and lower, decreasing indefinitely, without bound, to negative values of $y$ that are arbitrarily distant from the origin.

**DEFINITION 5**

**Vertical asymptote**

The graph of the function $y = f(x)$ has a vertical asymptote $x = a$ provided that at least one of the following conditions is satisfied:

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

Now examine the graph of the function $f(x) = \frac{1}{x^2}$ in Figure 8.2. Notice that it also has a vertical asymptote at $x = 0$, and also has a horizontal asymptote at $y = 0$, just like the function $f(x) = \frac{1}{x}$. However, the behaviour of the graph near the vertical asymptote reflects that fact that $f(x) = \frac{1}{x^2}$ is an even function, whereas $f(x) = \frac{1}{x}$ is an odd function.

Thinking about these two functions, and looking carefully at others, leads to:

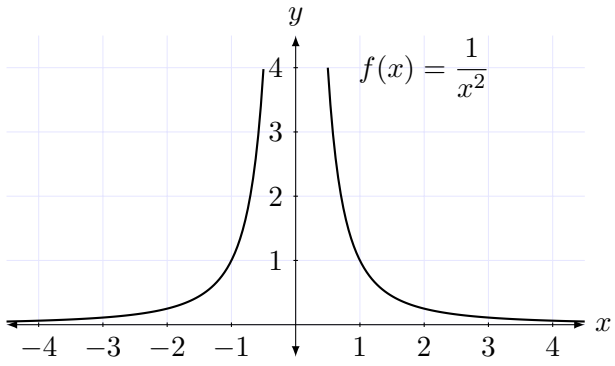



Figure 8.2: The graph of $f(x) = \dfrac{1}{x^2}$ has a vertical asymptote at $x = 0$ and a horizontal asymptote at $y = 0$.

**THEOREM 4**

**Asymptotes and limits**

(a) **Horizontal asymptotes.** If $n$ is a positive integer, then the graph of the function $y = \dfrac{1}{x^n}$ has a horizontal asymptote at $y = 0$. Furthermore,

$$\lim_{x\to\infty} \frac{1}{x^n} = 0 \qquad \text{and} \qquad \lim_{x\to-\infty} \frac{1}{x^n} = 0$$

(b) **Vertical asymptotes.** If $n$ is a positive integer, then the graph of the function $y = \dfrac{1}{x^n}$ has a vertical asymptote $x = 0$. Furthermore, if $n$ is an even positive integer, then

$$\lim_{x\to0^+} \frac{1}{x^n} = \infty \qquad \text{and} \qquad \lim_{x\to0^-} \frac{1}{x^n} = \infty$$

However, if $n$ is an odd positive integer, then

$$\lim_{x\to0^+} \frac{1}{x^n} = \infty \qquad \text{but} \qquad \lim_{x\to0^-} \frac{1}{x^n} = -\infty$$

To effectively use this theorem to determine vertical and horizontal asymptotes, we also need to make use of the following theorem:

**THEOREM 5**

**A practical approach to calculating limits (continued)**

6. (**Limit Laws**) Suppose that the function $f$ is an algebraic combination of simpler functions. Also suppose that the limit of each of the simpler functions exists. Then to evaluate the limit of $f$, just evaluate the limit of each of the simpler functions, and combine the individual limits using the same algebraic combination that forms $f$.

To be more specific, here are some fundamental instances of this idea. We also assume that $k$ is a constant, and that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both exist.

(a) $\lim\limits_{x\to a}[k\cdot f(x)] = k\left[\lim\limits_{x\to a} f(x)\right]$

(b) $\lim\limits_{x\to a}[f(x)+g(x)] = \lim\limits_{x\to a} f(x) + \lim\limits_{x\to a} g(x)$

(c) $\lim\limits_{x\to a}[f(x)-g(x)] = \lim\limits_{x\to a} f(x) - \lim\limits_{x\to a} g(x)$

(d) $\lim\limits_{x\to a}[f(x)\cdot g(x)] = \left[\lim\limits_{x\to a} f(x)\right]\cdot\left[\lim\limits_{x\to a} g(x)\right]$

(e) $\lim\limits_{x\to a}\left[\dfrac{f(x)}{g(x)}\right] = \dfrac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$, provided that $\lim\limits_{x\to a} g(x) \neq 0$

The following examples illustrate the use of the two previous theorems in determining vertical and horizontal asymptotes.

**EXAMPLE 16**

**Determining vertical and horizontal asymptotes**

Determine any (a) vertical and (b) horizontal asymptotes of the function $y = \dfrac{2x+1}{x-3}$.

**SOLUTION**

**(a)** Because we are dealing with a rational function, the possible vertical asymptotes occur where the denominator is 0. Thus, there is a possible vertical asymptote at $x = 3$.

To verify that this is indeed a vertical asymptote, let's take the limit of the function as $x \to 3$ from each side. First let's calculate the limit from the right:

$$\lim_{x\to 3^+} \frac{2x+1}{x-3}$$

As $x$ gets closer and closer to 3, the numerator gets closer and closer to $2(3)+1 = 7$, and the denominator gets closer and closer to 0. This implies that the limit does not exist, because the function values become arbitrarily large as $x$ gets closer and closer to 3. Furthermore, because $x > 3$, the denominator is positive as $x \to 3^+$, and the numerator is also positive for $x > 3$. Thus,

$$\lim_{x\to 3^+} \frac{2x+1}{x-3} = \infty$$

and so the limit does not exist. It follows that $x = 3$ is a vertical asymptote.

To determine the behaviour of the graph to the left of the asymptote, let's calculate

$$\lim_{x\to 3^-} \frac{2x+1}{x-3}$$

The same argument as above shows that this limit does not exist either. However, when $x < 0$, the numerator is positive and the denominator is negative, so the function values are negative. They become arbitrarily far from the origin as $x \to 3^-$, so

$$\lim_{x\to 3^-} \frac{2x+1}{x-3} = -\infty$$

**(b)** For horizontal asymptotes, we need to calculate the limit of the function as $x \to \infty$, and also the limit as $x \to -\infty$. The standard procedure for doing this is to divide the numerator and denominator by the highest power of $x$ present in the expression. The reason for doing this is that we can then make use of Theorem 5 to evaluate the limit.

$$\begin{aligned}
\lim_{x\to\infty} \frac{2x+1}{x-3} &= \lim_{x\to\infty} \frac{\frac{2x}{x}+\frac{1}{x}}{\frac{x}{x}-\frac{3}{x}} \\
\lim_{x\to\infty} \frac{2x+1}{x-3} &= \lim_{x\to\infty} \frac{2+\frac{1}{x}}{1-\frac{3}{x}} \\
\lim_{x\to\infty} \frac{2x+1}{x-3} &= \frac{[\lim_{x\to\infty} 2] + \left[\lim_{x\to\infty} \frac{1}{x}\right]}{[\lim_{x\to\infty} 1] - \left[\lim_{x\to\infty} \frac{3}{x}\right]} \qquad \text{(using limit laws)} \\
\lim_{x\to\infty} \frac{2x+1}{x-3} &= \frac{2+0}{1-3(0)} \\
\lim_{x\to\infty} \frac{2x+1}{x-3} &= 2
\end{aligned}$$

Because this limit exists and equals 2, the graph of the function has a horizontal asymptote $y = 2$. Note that in evaluating the limits, we made use of the fact that the limit of a constant function is the constant value. I will let you repeat the calculation for $x \to -\infty$; you'll find the same asymptote, so the only horizontal asymptote is $y = 2$. It is worthwhile sketching a graph of the function to verify your calculations. Do this!

**CAREFUL!**

**Sometimes a vertical asymptote, sometimes a hole discontinuity**

Notice that in the previous example we were careful to say that if the denominator of a rational function is 0 for a certain value of $x$ this does not guarantee that the function has a vertical asymptote at this value of $x$. Do you recall seeing any such examples? Why yes, earlier in this book we encountered many examples. When we set up limits to calculate slopes, the denominators were invariably 0, yet many of the limits existed. In such cases, the graphs of the expressions have hole discontinuities, not vertical asymptotes.

This means we can't automatically assume that a rational expression has a vertical asymptote when its denominator is 0; we must check the appropriate limits before making such a conclusion.

## EXAMPLE 17

**Determining vertical and horizontal asymptotes**

Determine any (a) vertical and (b) horizontal asymptotes of the function $y = \dfrac{x^2-1}{x^2+x-6}$.

## SOLUTION

**(a)** The possible vertical asymptotes occur where the denominator is 0. Thus, there is a possible vertical asymptote at the solutions of $x^2 + x - 6 = 0$. The quadratic expression is factorable: $(x-2)(x+3) = 0$. Thus, there are possible vertical asymptotes at $x = -3$ and $x = 2$.

Check each potential vertical asymptote:

$$\lim_{x\to -3^+} \frac{x^2-1}{x^2+x-6} = -\infty \qquad \lim_{x\to -3^-} \frac{x^2-1}{x^2+x-6} = \infty$$

$$\lim_{x\to 2^+} \frac{x^2-1}{x^2+x-6} = \infty \qquad \lim_{x\to 2^-} \frac{x^2-1}{x^2+x-6} = -\infty$$

It's clear that each of the four previous limits is either $\infty$ or $-\infty$, but which sign is correct, $+$ or $-$? Note that near each asymptote, the numerator is positive. The denominator is a quadratic expression opening up, so it is negative between the roots $-3$ and 2, and positive when $x > 2$ and when $x < -3$. This explains the signs.

Thus, there are vertical asymptotes at both $x = -3$ and $x = 2$.

**(b)** For horizontal asymptotes, we need to calculate the limit of the function as $x \to \infty$, and also the limit as $x \to -\infty$. As in the previous example, divide the numerator and denominator by the highest power of $x$ present in the expression, which is $x^2$, and then use limit laws.

$$\begin{aligned}
\lim_{x\to\infty} \frac{x^2-1}{x^2+x-6} &= \lim_{x\to\infty} \frac{\dfrac{x^2}{x^2} - \dfrac{1}{x^2}}{\dfrac{x^2}{x^2} + \dfrac{x}{x^2} - \dfrac{6}{x^2}} \\
&= \lim_{x\to\infty} \frac{1 - \dfrac{1}{x^2}}{1 + \dfrac{1}{x} - \dfrac{6}{x^2}} \\
&= \frac{[\lim_{x\to\infty} 1] - \left[\lim_{x\to\infty} \dfrac{1}{x^2}\right]}{[\lim_{x\to\infty} 1] + \left[\lim_{x\to\infty} \dfrac{1}{x}\right] - \left[\lim_{x\to\infty} \dfrac{6}{x^2}\right]} \\
&= \frac{1-(0)}{1+(0)-6(0)} \\
&= 1
\end{aligned}$$

Because this limit exists and equals 1, the graph of the function has a horizontal asymptote $y = 1$. I will let you repeat the calculation for $x \to -\infty$; you'll find the same asymptote, so the only horizontal asymptote is $y = 1$. Once again, sketching a graph of this function will be interesting. Do this!

Now modify the previous example slightly: Does the graph of the function $y = \dfrac{x^2+4x+3}{x^2+x-6}$ have vertical and horizontal asymptotes? By factoring the numerator and denominator, you will find that

$$\begin{aligned} \frac{x^2+4x+3}{x^2+x-6} &= \frac{(x+3)(x+1)}{(x+3)(x-2)} \\ &= \frac{x+1}{x-2} \qquad \text{(provided that } x \neq -3\text{)} \end{aligned}$$

By analyzing the simplified expression for this function, see if you can show that there is just one vertical asymptote, $x = 2$, and the horizontal asymptote is $y = 1$. What happens at $x = -3$? There is a hole discontinuity there (because the original expression is not defined there), but no vertical asymptote. It's worthwhile sketching a graph of this modified function and comparing it to a graph of the function in the previous example. Do this!

Now consider a polynomial function. You might recall from high school that such functions have no asymptotes. Their "end behaviour" is determined by limits as $x \to \pm\infty$. For example,

$$\lim_{x\to\infty} x^3 = \infty \qquad \text{and} \qquad \lim_{x\to-\infty} x^3 = -\infty$$

Compare these limits to the graph of the function $y = x^3$ in Figure 8.3.

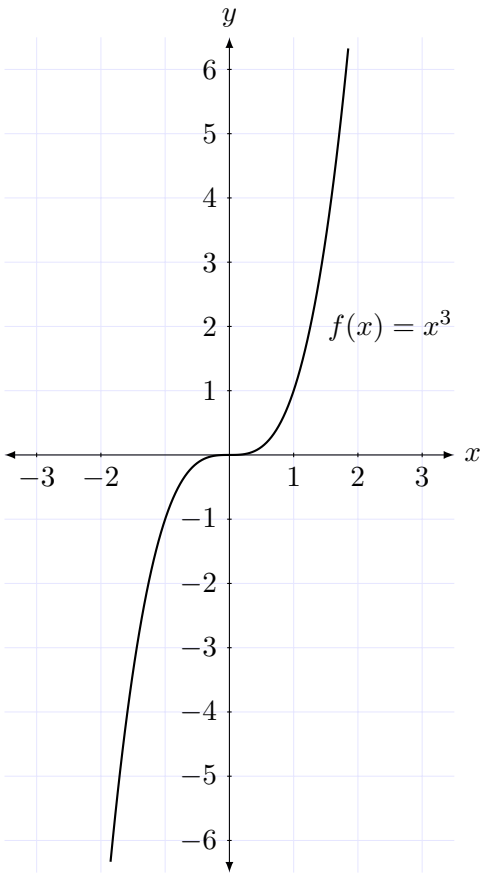



Figure 8.3: The graph of $y = x^3$, like all polynomials, has no asymptotes.

Now let's look at some additional examples of calculating vertical and horizontal asymptotes for functions that are not rational functions.

## EXAMPLE 18

**Determining asymptotes**

Determine any horizontal asymptotes for the function (a) $y = \sin x$ and (b) $y = \dfrac{\sin x}{x}$.

## SOLUTION

**(a)** The trend in the function values as $x \to \infty$ is that they oscillate endlessly without approaching a single definite value. The same is true as $x \to -\infty$. Thus,

$$\lim_{x\to\infty} \sin x \text{ DOES NOT EXIST} \qquad \text{and} \qquad \lim_{x\to-\infty} \sin x \text{ DOES NOT EXIST}$$

and therefore the sine function has no horizontal asymptote.

**(b)** Although the numerator oscillates between $-1$ and 1, the denominator gets larger and larger in absolute value as $x \to \pm\infty$. Thus, the function values get closer and closer to 0 as $x \to \pm\infty$, and so

$$\lim_{x\to\infty} \frac{\sin x}{x} = 0 \qquad \text{and} \qquad \lim_{x\to-\infty} \frac{\sin x}{x} = 0$$

Thus, the function $\dfrac{\sin x}{x}$ has a horizontal asymptote $y = 0$.

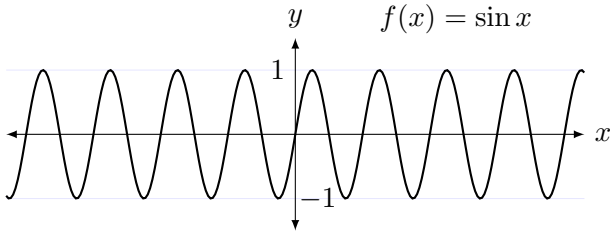



Figure 8.4: The sine function has no horizontal asymptotes. As $x \to \pm\infty$, the function values oscillate without approaching a single definite value. (The scale on the $x$-axis is compressed relative to the scale on the $y$-axis.)

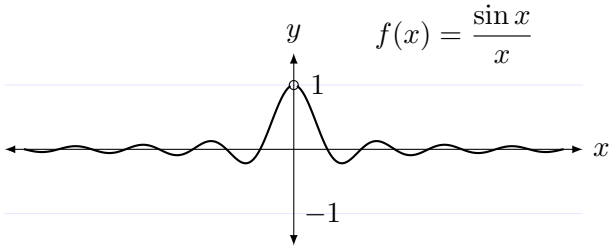



Figure 8.5: Despite its oscillations, the function $y = \dfrac{\sin x}{x}$ has the horizontal asymptote $y = 0$, because the amplitude of the oscillations approaches 0 as $x \to \pm\infty$. (The scale on the $x$-axis is compressed relative to the scale on the $y$-axis.)

## CAREFUL!

**The graph of a function may cross an asymptote**

There is nothing in the definition of a horizontal asymptote that prevents the graph of a function from crossing its horizontal asymptote. Beware of this misconception about horizontal asymptotes that you can find all over the internet. The previous example illustrates the fact that the graph of a function can indeed cross its horizontal asymptote, in this case an infinite number of times. Of course, there are plenty of functions that have horizontal asymptotes for which their graphs do not cross its asymptote, but this crossing behaviour is possible.

Part (a) of the previous example illustrates another way that a function can fail to have a limit: the function values can oscillate endlessly without approaching a single definite number. Part (b) illustrates the fact that the graph of a function can cross its asymptote.

**Question:** Is it possible for the graph of a function to cross its vertical asymptote? If so, construct such an example. If not, explain why not.

## EXAMPLE 19

**Determining asymptotes**

Determine any vertical and horizontal asymptotes for the function $y = \tan x$.

## SOLUTION

One way to analyze the tangent function is to write it in terms of sine and cosine functions:

$$\tan x = \frac{\sin x}{\cos x}$$

Possible locations of vertical asymptotes are $x$-values for which $\cos x = 0$; thinking in terms of the unit circle, these $x$-values are $x = \pm\frac{\pi}{2}$, $x = \pm\frac{3\pi}{2}$, $x = \pm\frac{5\pi}{2}$, and so on. These will indeed be vertical asymptotes provided that the numerator $\sin x$ is not equal to zero at these $x$-values. This is true, because when $\cos x = 0$, you can see from either the unit circle or the graphs of sine and cosine functions that $\sin x = \pm 1$. Thus, the graph of $y = \tan x$ has vertical asymptotes at $x = \pm\frac{\pi}{2}$, $x = \pm\frac{3\pi}{2}$, $x = \pm\frac{5\pi}{2}$, and so on.

To determine the behaviour of the graph near the vertical asymptotes, let's calculate the following limits. Note that as $x \to \frac{\pi}{2}$ from the left, the sine function is positive, and so is the cosine function. Therefore,

$$\lim_{x \to \pi/2^-} \tan x = \infty$$

However, when $x \to \frac{\pi}{2}$ from the right, the sine function is positive, but the cosine function is negative. Therefore,

$$\lim_{x \to \pi/2^+} \tan x = -\infty$$

Similar reasoning will show the behaviour of the graph near the other vertical asymptotes. However, we can equally well reason that the tangent function is periodic, with period $\pi$, so once we determine the graph for one interval of the $x$-axis of length $\pi$, we can simply repeat this piece of graph endlessly.

What about horizontal asymptotes? The tangent function is periodic, with period $\pi$. This means that as $x \to \infty$, or $x \to -\infty$, the function values do not approach a definite number, but rather repeat endlessly. This means that

$$\lim_{x \to \infty} \tan x \text{ DOES NOT EXIST} \qquad \text{and} \qquad \lim_{x \to -\infty} \tan x \text{ DOES NOT EXIST}$$

This means that the graph of $y = \tan x$ has no horizontal asymptotes, which you can see from the graph in Figure 8.6.

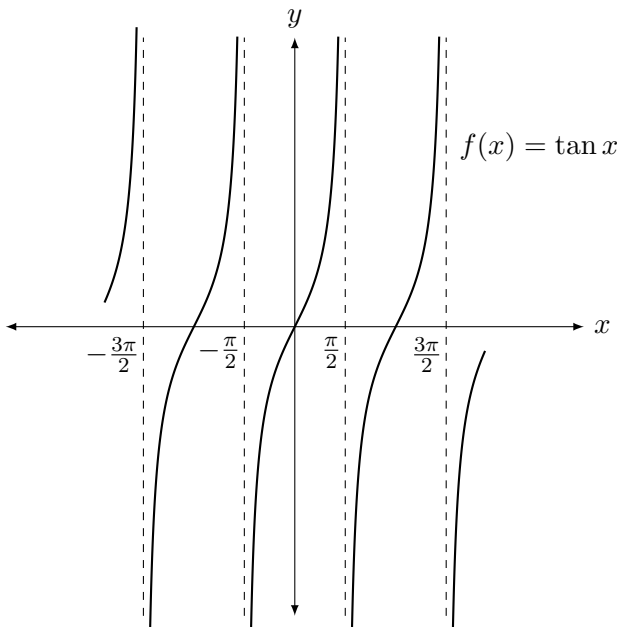



Figure 8.6: The graph of $y = \tan x$ has an infinite number of vertical asymptotes but no horizontal asymptotes.

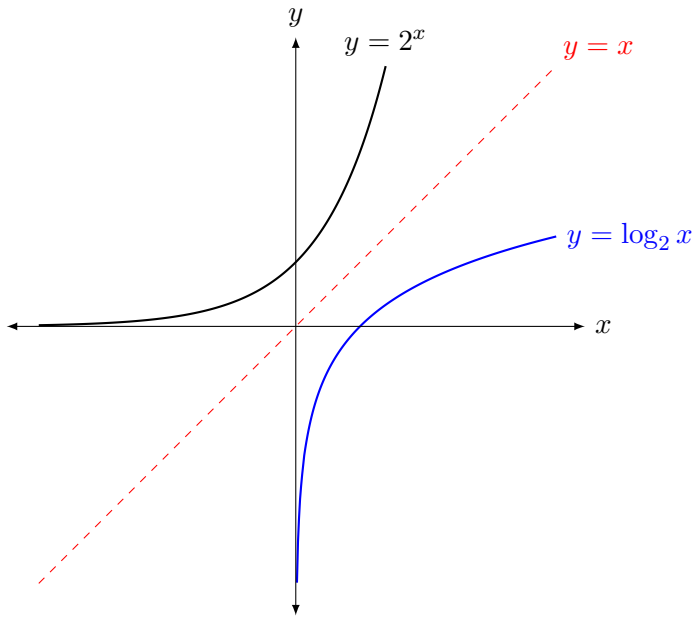



Figure 8.7: The functions $y = 2^x$ (in black) and $y = \log_2 x$ (in blue) are inverses of each other, and therefore their graphs are reflections of each other in the line $y = x$. The black graph has the horizontal asymptote $y = 0$ and the blue graph has the vertical asymptote $x = 0$.

## EXAMPLE 20

**Determining asymptotes**

Determine any vertical and horizontal asymptotes for the functions (a) $y = 2^x$ and (b) $y = \log_2 x$.

## SOLUTION

**(a)** Recall (and see Figure 8.7) that exponential functions such as $y = 2^x$ increase indefinitely as $x \to \infty$, but approach the $x$-axis asymptotically as $x \to -\infty$. Use your calculator and a table of values to get a sense for this if it is not clear. This means that

$$\lim_{x\to\infty} 2^x = \infty \qquad \text{and} \qquad \lim_{x\to-\infty} 2^x = 0$$

The graph of $y = 2^x$ has a horizontal asymptote at $y = 0$ and no vertical asymptotes.

**(b)** The function $y = \log_2 x$ is the inverse of $y = 2^x$. This means that the graph of $y = \log_2 x$ is the reflection of the graph of $y = 2^x$ in the line $y = x$. Algebraically, this is equivalent to the idea that if you interchange $x$ and $y$ in the formula for one the two functions, you will get the formula for the other. But this also means that if you interchange $x$ and $y$ in any asymptotes of one function, you will get the formula for an asymptote of the other function. Therefore, the graph of $y = \log_2 x$ has a vertical asymptote at $x = 0$ and no horizontal asymptote.

In limit language,

$$\lim_{x\to\infty} \log_2 x = \infty \qquad \text{and} \qquad \lim_{x\to 0^+} 2^x = -\infty$$

Before you study the next example, make sure you understand the following tricky point, which is a key step in the solution of the example.

Consider a specific example first. Suppose you begin with 5, then square it to get 25, and then finally take the square root. You end up back where you started, at 5. However, suppose you begin with $-5$, then square it to get 25, and then finally take the square root. You end up with 5, which is the negative of what you started with.

To summarize:

$$\sqrt{x^2} = \left\{ \begin{array}{rl} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{array} \right\}$$

Another way to express this point, which is relevant for the following example, is that if $x \geq 0$, you can replace $x$ by the equivalent expression $\sqrt{x^2}$. However, if $x < 0$, you can replace $x$ by the equivalent expression $-\sqrt{x^2}$.

It's worth going through a few more specific examples on your own to make sure that you understood this point. (Do this!) Once you do understand this point, proceed with the following example.

## EXAMPLE 21

**Determining asymptotes**

Determine any vertical and horizontal asymptotes for the function $y = \dfrac{\sqrt{3x^2+4}}{x-2}$.

## SOLUTION

The denominator is 0 when $x = 2$, yet the numerator is not 0 when $x = 2$. This means that there is a vertical asymptote at $x = 2$. The behaviour of the function near the vertical asymptote can be deduced from the following limits (the numerator is always positive, so the sign of the limit depends only on the sign of the denominator):

$$\lim_{x \to 2^+} \frac{\sqrt{3x^2+4}}{x-2} = \infty \qquad \text{and} \qquad \lim_{x \to 2^-} \frac{\sqrt{3x^2+4}}{x-2} = -\infty$$

To determine if there are horizontal asymptotes, let's begin by calculating the limit as $x \to \infty$. If we were to use the technique of dividing numerator and denominator by the highest power of $x$, we might think that we should divide numerator and denominator by $x^2$. However, because the term $3x^2$ in the numerator is under a square root sign, we'll be able to knock it out by dividing the numerator and denominator by $x$, not $x^2$. Pay careful attention to how this happens:

$$\lim_{x \to \infty} \frac{\sqrt{3x^2+4}}{x-2} = \lim_{x \to \infty} \frac{\dfrac{\sqrt{3x^2+4}}{x}}{\dfrac{x-2}{x}}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to\infty}\frac{\frac{\sqrt{3x^2+4}}{\sqrt{x^2}}}{\frac{x-2}{x}} \qquad \text{(notice how } x \text{ is replaced by } \sqrt{x^2}\text{, because } x>0\text{)}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to\infty}\frac{\sqrt{\frac{3x^2+4}{x^2}}}{\frac{x-2}{x}}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to\infty}\frac{\sqrt{\frac{3x^2}{x^2}+\frac{4}{x^2}}}{\frac{x}{x}-\frac{2}{x}}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to\infty}\frac{\sqrt{3+\frac{4}{x^2}}}{1-\frac{2}{x}}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{\sqrt{\lim_{x\to\infty}3+\lim_{x\to\infty}\left(\frac{4}{x^2}\right)}}{\lim_{x\to\infty}1-\lim_{x\to\infty}\left(\frac{2}{x}\right)}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{\sqrt{3+4\lim_{x\to\infty}\left(\frac{1}{x^2}\right)}}{1-2\lim_{x\to\infty}\left(\frac{1}{x}\right)}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{\sqrt{3+4(0)}}{1-2(0)}$$

$$\lim_{x\to\infty}\frac{\sqrt{3x^2+4}}{x-2} = \sqrt{3}$$

Because this limit exists and equals $\sqrt{3}$, therefore $y=\sqrt{3}$ is a horizontal asymptote to the graph of the function.

Now let's calculate the limit as $x\to-\infty$. The calculation is almost exactly the same as the previous limit calculation; the only difference is that when we divide the square root expression by $x$, in the immediately following step we have to replace $x$ by $-\sqrt{x^2}$, because $x<0$ as $x\to-\infty$.

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to-\infty}\frac{\frac{\sqrt{3x^2+4}}{x}}{\frac{x-2}{x}}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to-\infty}\frac{\frac{\sqrt{3x^2+4}}{-\sqrt{x^2}}}{\frac{x-2}{x}} \quad \text{(notice how } x \text{ is replaced by } -\sqrt{x^2}\text{, because } x<0)$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to-\infty}\frac{-\sqrt{\frac{3x^2+4}{x^2}}}{\frac{x-2}{x}}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to-\infty}\frac{-\sqrt{\frac{3x^2}{x^2}+\frac{4}{x^2}}}{\frac{x}{x}-\frac{2}{x}}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \lim_{x\to-\infty}\frac{-\sqrt{3+\frac{4}{x^2}}}{1-\frac{2}{x}}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{-\sqrt{\lim_{x\to-\infty}3+\lim_{x\to-\infty}\left(\frac{4}{x^2}\right)}}{\lim_{x\to-\infty}1-\lim_{x\to-\infty}\left(\frac{2}{x}\right)}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{-\sqrt{3+4\lim_{x\to-\infty}\left(\frac{1}{x^2}\right)}}{1-2\lim_{x\to-\infty}\left(\frac{1}{x}\right)}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = \frac{-\sqrt{3+4(0)}}{1-2(0)}$$

$$\lim_{x\to-\infty}\frac{\sqrt{3x^2+4}}{x-2} = -\sqrt{3}$$

Because this limit exists and equals $-\sqrt{3}$, therefore $y=-\sqrt{3}$ is a horizontal asymptote to the graph of the function. The graph therefore has two horizontal asymptotes.

The results of the limit calculations are illustrated by the graph of the function in Figure 8.8.

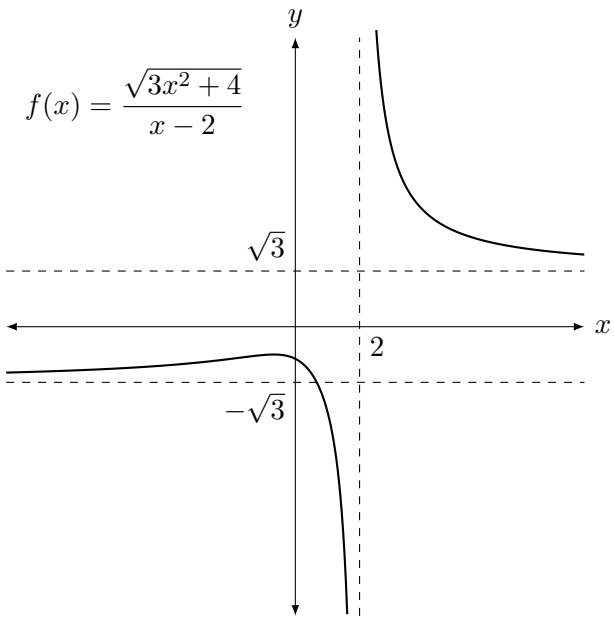



Figure 8.8: This graph has one vertical asymptote and two horizontal asymptotes.

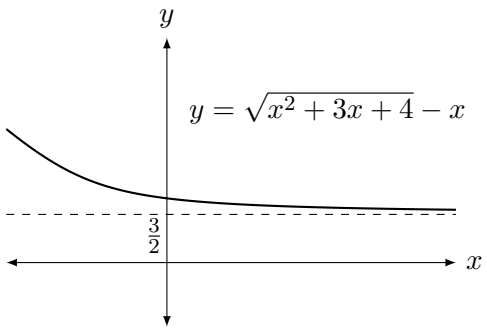



Figure 8.9: This graph has no vertical asymptote and one horizontal asymptote. The graph approaches the asymptote only as $x \to \infty$.

## EXAMPLE 22

**Determining asymptotes**

Determine any vertical and horizontal asymptotes for the graph of the function $f(x) = \sqrt{x^2+3x+4} - x$.

## SOLUTION

It is often useful to guess a limit before setting out to calculate it exactly. One way to do this is to substitute suitable numbers into a calculator. However, limits (as $x \to \infty$ or as $x \to -\infty$) of functions such as this one, which is a difference, are notably difficult to guess.[a] What should we do? Calculating this limit, or even deciding whether the limit exists, is problematic. It's not clear at first glance how to proceed.

Let's think back to the limits that we've calculated so far. The difficult ones were in the form of a quotient, but we were often able to evaluate them by cancelling a troublesome factor from numerator and denominator. If we can't think of anything better to do, we can always try to convert the formula for $f$ into a quotient, since we've got quite a bit of experience evaluating limits of this type.

OK, how do we convert the formula for $f$ into a quotient? Consider the following:

$$\sqrt{x^2+3x+4} - x = \frac{\sqrt{x^2+3x+4} - x}{1}$$

Well, sure, this is correct, but it doesn't seem very helpful, since the two expressions are virtually

[a] In fact, computers have well-known difficulties with differences of large numbers that are almost equal, so one must be careful when using software in such cases. Often one must do some reasoning and modify the expression somewhat before letting the computer do its thing. This is a strong argument for understanding what you are doing, so that you can use software wisely, not push it beyond its limitations, and be able to detect its mistakes.

identical. However, it might give us the idea of multiplying numerator and denominator by the conjugate expression, and then simplifying. This turns out to be very helpful:

$$\begin{aligned}
\sqrt{x^2+3x+4}-x &= \frac{\sqrt{x^2+3x+4}-x}{1} \\
\sqrt{x^2+3x+4}-x &= \frac{\sqrt{x^2+3x+4}-x}{1}\cdot\frac{\sqrt{x^2+3x+4}+x}{\sqrt{x^2+3x+4}+x} \\
\sqrt{x^2+3x+4}-x &= \frac{\left(\sqrt{x^2+3x+4}-x\right)\left(\sqrt{x^2+3x+4}+x\right)}{\sqrt{x^2+3x+4}+x} \\
\sqrt{x^2+3x+4}-x &= \frac{x^2+3x+4-x^2}{\sqrt{x^2+3x+4}+x} \\
\sqrt{x^2+3x+4}-x &= \frac{3x+4}{\sqrt{x^2+3x+4}+x}
\end{aligned}$$

Now you might like to get out your calculator and guess the limit of this function as $x \to \infty$ and as $x \to -\infty$; you should have an easier time with this expression than the original one, which already justifies our manoeuvers.

After you've guessed the limit, it's time to calculate it exactly by analyzing the latest formula for $f$. Before we do this, let's think about vertical asymptotes first. Are there any values of $x$ for which the function is not defined? Well, the expression contains a square root, and it could be that for certain values of $x$ the quantity under the square root sign is negative, which would make the entire expression undefined. To determine whether there are any such values of $x$, I'll use the following reasoning.

The expression under the square root sign is a quadratic expression; if it were graphed, the graph would be a parabola opening up. Thus, if it has two zeros, then the expression is negative for all $x$ values between the zeros. So a good start is to determine if the quadratic expression has zeros. It doesn't seem factorable, so I'll use the quadratic formula:

$$\begin{aligned}
0 &= x^2+3x+4 \\
x &= \frac{-b\pm\sqrt{b^2-4ac}}{2a} \\
x &= \frac{-3\pm\sqrt{3^2-4(1)(4)}}{2(1)} \\
x &= \frac{-3\pm\sqrt{9-16}}{2} \\
x &= \frac{-3\pm\sqrt{-7}}{2}
\end{aligned}$$

The square root of a negative number is not a real number, so we conclude that there are no real zeros. Thus, the expression $x^2+3x+4$ is always positive (remember, its graph is a parabola opening up), and therefore the expression $\sqrt{x^2+3x+4}$ exists for all values of $x$. Thus, the domain of $f$ is all real values of $x$. Finally, note that the square-root term in the denominator has greater magnitude than the $x$ term, so the denominator is not negative, even for negative values of $x$. Therefore, we conclude that the graph of $f$ has no vertical asymptotes.

Does the graph of $f$ have any horizontal asymptotes? Let's determine the limits as $x \to \infty$ and as $x \to -\infty$ to answer this question.

If we try to use oversimplified reasoning to calculate this limit, we may fall prey to an error that commonly occurs. It might be tempting to reason that as $x \to \infty$, $\sqrt{x^2+3x+4} \to \infty$ and also $x \to \infty$, so therefore $\sqrt{x^2+3x+4}-x \to \infty-\infty = 0$. **THIS IS NOT VALID REASONING**, because $\infty$ is not a number that can be subtracted from itself to produce another number. In situations such as this one, when we are trying to calculate the limit of a difference, and each term $\to \infty$, the limit of the entire expression may or may not exist, and if it does exist, we have no way of knowing what the limit is from this kind of reasoning. Correct reasoning involves converting the difference to a quotient and then continuing as we illustrated previously:

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{3x+4}{\sqrt{x^2+3x+4}+x} \quad \text{(see above)}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{\left(\frac{3x+4}{x}\right)}{\frac{\sqrt{x^2+3x+4}+x}{x}} \quad \text{(divide numerator and denominator by } x\text{)}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{3+\frac{4}{x}}{\frac{\sqrt{x^2+3x+4}}{x}+\frac{x}{x}}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{3+\frac{4}{x}}{\frac{\sqrt{x^2+3x+4}}{\sqrt{x^2}}+1} \quad (x \text{ is replaced by } \sqrt{x^2}\text{, because } x>0)$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{3+\frac{4}{x}}{\sqrt{\frac{x^2}{x^2}+\frac{3x}{x^2}+\frac{4}{x^2}}+1}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \lim_{x\to\infty} \frac{3+\frac{4}{x}}{\sqrt{1+\frac{3}{x}+\frac{4}{x^2}}+1}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \frac{3+0}{\sqrt{1+0+0}+1}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \frac{3}{1+1}$$

$$\lim_{x\to\infty} \sqrt{x^2+3x+4}-x = \frac{3}{2}$$

Because the limit exists and is equal to 3/2, the graph of $f$ has a horizontal asymptote at $y = \frac{3}{2}$.

Now let's calculate the limit as $x \to -\infty$. In this case, we can use simple reasoning. Rewrite the formula for $f$ as follows:

$$\sqrt{x^2+3x+4}-x = \sqrt{x^2+3x+4}+(-x)$$

Each term on the right side of the previous equation is positive as $x \to -\infty$, and each term becomes arbitrarily large as $x \to -\infty$. Thus, the limit of the **SUM** of the two terms also becomes arbitrarily large as $x \to -\infty$. That is,

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x=\infty$$

Alternatively, one can also tackle this limit using a method similar to the one we used to calculate the limit as $x\to\infty$. Here's how this would work, if you wished to go to the additional work:

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{3x+4}{\sqrt{x^2+3x+4}+x}$$

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{\frac{3x+4}{x}}{\frac{\sqrt{x^2+3x+4}+x}{x}}$$

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{3+\frac{4}{x}}{\frac{\sqrt{x^2+3x+4}}{x}+\frac{x}{x}}$$

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{3+\frac{4}{x}}{\frac{\sqrt{x^2+3x+4}}{-\sqrt{x^2}}+1} \qquad (x \text{ is replaced by } -\sqrt{x^2}, \text{ as } x<0)$$

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{3+\frac{4}{x}}{-\sqrt{\frac{x^2}{x^2}+\frac{3x}{x^2}+\frac{4}{x^2}}+1}$$

$$\lim_{x\to-\infty}\sqrt{x^2+3x+4}-x = \lim_{x\to-\infty}\frac{3+\frac{4}{x}}{-\sqrt{1+\frac{3}{x}+\frac{4}{x^2}}+1}$$

As $x\to-\infty$, the numerator of the expression $\to 3$, but the denominator $\to -1+1=0$, so the limit does not exist. A little analysis will convince you that the quantity under the square root sign is slightly less than 1 as $x\to-\infty$, which means that the denominator is positive as $x\to-\infty$. This means that the expression $\to\infty$, and so the limit does not exist.

Thus the graph of $f$ has a single horizontal asymptote, and no vertical asymptote. The result is illustrated in Figure 8.9.

## 8.2 Slant Asymptotes

The idea of an asymptote can be generalized. So far we have defined an asymptote as a vertical or horizontal line such that the graph of a function approaches the line more and more closely as you travel farther and farther away from the origin. But surely if you rotated the graph and its asymptotes, so that the asymptotes were no longer vertical or horizontal, you would still wish to call them asymptotes, wouldn't you?

In other words, we should strive to define the concept of asymptote in a more "intrinsic" way. That is, the definition should capture the geometric flavour of the concept.

One can reformulate the definition of a horizontal asymptote, in terms of vertical distance, as follows:

**DEFINITION 6**

**Horizontal asymptote**

The line $y = b$ is a horizontal asymptote for the graph of the function $y = f(x)$ provided that either one or both of the following conditions is satisfied:

$$\lim_{x\to\infty} f(x) - b = 0 \qquad \text{or} \qquad \lim_{x\to-\infty} f(x) - b = 0$$

The same conceptual structure can be used to define a slant asymptote:

**DEFINITION 7**

**Slant asymptote**

The line $y = mx + b$ is a slant asymptote for the graph of the function $y = f(x)$ provided that either one or both of the following conditions is satisfied:

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

The definition of slant asymptote may be clear geometrically (the vertical distance between the curve and the asymptote becomes smaller and smaller as $x \to \infty$ or as $x \to -\infty$), but it gives us no instructions on how to determine a slant asymptote. In other words, once you figure out what you think the slant asymptote might be, the definition gives you a way of testing it to see if it really is a slant asymptote. But how do we determine a candidate for a slant asymptote?

Let's consider two types of functions — rational functions, and other types. You may recall from high school that a rational function has a slant asymptote if the degree of the numerator is one more than the degree of the denominator. One way to determine the equation of the slant asymptote in this case is long division.

For example, consider the function $f(x) = \dfrac{x^2 - x}{x + 1}$. Because $f$ is a rational function for which the degree of the numerator is one greater than the degree of the denominator, we can conclude that the graph of $f$ has a slant asymptote. Now use long division to obtain:

$$\begin{array}{r|ll} & x & -2 \\ \hline x+1 & x^2 & -x \\ & x^2 & +x \\ \hline & & -2x \\ & & -2x \quad -2 \\ \hline & & 2 \end{array}$$

From the long division, we can see that

$$\frac{x^2 - x}{x + 1} = x - 2 + \frac{2}{x + 1}$$

If you aren't a fan of long division, here is an alternative method: Just add and subtract the correct terms in the numerator in a step-by-step fashion, and you'll achieve the same result as long

division. Here's how it works in this case (you can shorten this process with practice):

$$\begin{aligned}
\frac{x^2-x}{x+1} &= \frac{x^2+x-x-x}{x+1} \\
\frac{x^2-x}{x+1} &= \frac{x^2+x-2x}{x+1} \\
\frac{x^2-x}{x+1} &= \frac{x^2+x}{x+1}+\frac{-2x}{x+1} \\
\frac{x^2-x}{x+1} &= \frac{x(x+1)}{x+1}-2\left(\frac{x}{x+1}\right) \\
\frac{x^2-x}{x+1} &= x-2\left(\frac{x+1-1}{x+1}\right) \\
\frac{x^2-x}{x+1} &= x-2\left(\frac{(x+1)-1}{x+1}\right) \\
\frac{x^2-x}{x+1} &= x-2\left(\frac{x+1}{x+1}-\frac{1}{x+1}\right) \\
\frac{x^2-x}{x+1} &= x-2\left(1-\frac{1}{x+1}\right) \\
\frac{x^2-x}{x+1} &= x-2+\frac{2}{x+1}
\end{aligned}$$

Compare the alternative development to the long division, and you'll find that they are essentially the same process, but done slightly differently. Choose the method you like best and practice it, as the process is required frequently, and therefore it is essential to have this tool in your tool kit.

Once we have divided the polynomials in the rational function, we can read the slant asymptote from the result: The slant asymptote to the graph of $f$ is $y = x - 2$. To verify this, apply the definition of slant asymptote by calculating the following limit. We'll first look at $x \to \infty$:

$$\begin{aligned}
\lim_{x\to\infty}\left[f(x)-(x-2)\right] &= \lim_{x\to\infty}\left[\frac{x^2-x}{x+1}-(x-2)\right] \\
\lim_{x\to\infty}\left[f(x)-(x-2)\right] &= \lim_{x\to\infty}\left[\left(x-2+\frac{2}{x+1}\right)-(x-2)\right] \quad \text{(from an earlier calculation)} \\
\lim_{x\to\infty}\left[f(x)-(x-2)\right] &= \lim_{x\to\infty}\left[x-2+\frac{2}{x+1}-x+2\right] \\
\lim_{x\to\infty}\left[f(x)-(x-2)\right] &= \lim_{x\to\infty}\left[\frac{2}{x+1}\right] \\
\lim_{x\to\infty}\left[f(x)-(x-2)\right] &= 0
\end{aligned}$$

Thus, the line $y = x - 2$ is a slant asymptote to the graph of $f(x) = \dfrac{x^2-x}{x+1}$ to the right. The limit as $x \to -\infty$ follows the same pattern with the same result, so the line $y = x - 2$ is also a slant asymptote to the graph of $f(x) = \dfrac{x^2-x}{x+1}$ to the left.

The results are illustrated in Figure 8.10.

For functions that are not rational, one way to determine if they have a slant asymptote is to use guesswork and play with a calculator. Of course, after you have guessed, you will then verify that your guess is correct using the definition of slant asymptote.

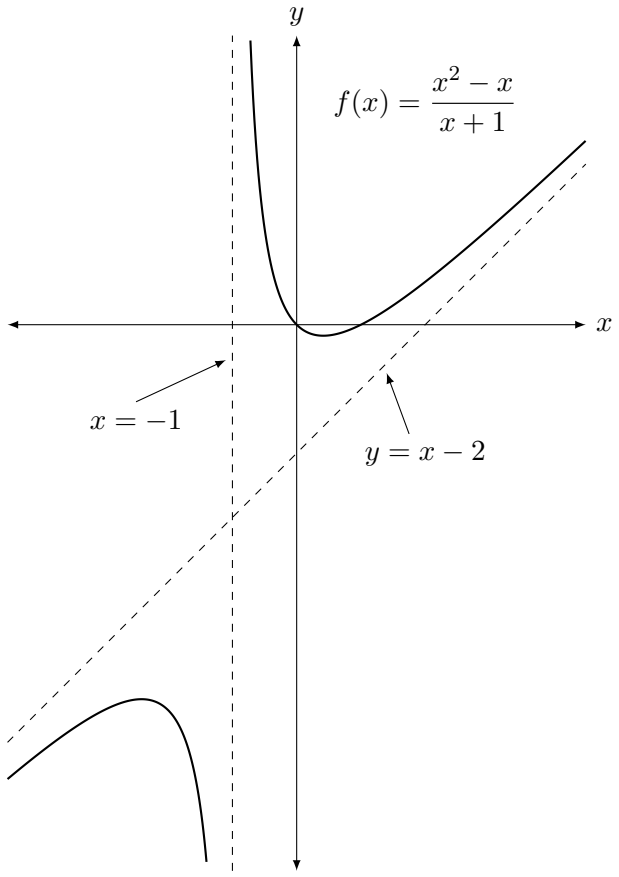



Figure 8.10: The function $f(x) = \dfrac{x^2 - x}{x + 1}$ has a vertical asymptote $x = -1$ and a slant asymptote $y = x - 2$.

Consider the function $f(x) = \sqrt{x^2 + 3x + 4} - x$, which we examined in an earlier example. We have already determined that the graph of $f$ has a horizontal asymptote to the right with equation $y = \dfrac{3}{2}$. Could the graph of $f$ also have a slant asymptote to the left?

One way to try to guess this is to use a calculator to create a table of values. Alternatively, one might reason as follows. For values of $x$ that are far to the left of the origin, $x^2$ is much greater in absolute value than $(3x + 4)$. This means that one can approximate $f$ by ignoring the terms $(3x + 4)$:

$$\begin{aligned} f(x) &= \sqrt{x^2 + 3x + 4} - x \\ f(x) &\approx \sqrt{x^2} - x \\ f(x) &\approx -x - x \quad (\sqrt{x^2} \text{ is replaced by } -x \text{ because } x < 0) \\ f(x) &\approx -2x \end{aligned}$$

Let's test this approximation using a calculator:

| $x$ | $f(x) = \sqrt{x^2 + 3x + 4} - x$ | $-2x$ |
|---|---|---|
| $-10$ | 18.60 | 20 |
| $-100$ | 198.51 | 200 |
| $-1000$ | 1998.50 | 2000 |

It appears that the approximation is fairly good, but it seems that the approximation would be better if we subtracted 1.5 from the approximating function. This suggests that there might be a slant asymptote to the left with equation $y = -2x - \frac{3}{2}$. We can test this by using the definition of slant asymptote, as follows.

$$\begin{aligned}
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\sqrt{x^2+3x+4}-x-\left(-2x-\frac{3}{2}\right)\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\sqrt{x^2+3x+4}-x+2x+\frac{3}{2}\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\sqrt{x^2+3x+4}+\left(x+\frac{3}{2}\right)\right]
\end{aligned}$$

The limit in the previous equation can't be determined by simple reasoning yet, because as $x \to -\infty$, the first term $\to \infty$ and the second term $\to -\infty$. So, we must use the typical trick of multiplying numerator and denominator by the conjugate, then simplifying, then using reasoning, as follows.

$$\begin{aligned}
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\sqrt{x^2+3x+4}+\left(x+\frac{3}{2}\right)\right]\cdot\frac{\left[\sqrt{x^2+3x+4}-\left(x+\frac{3}{2}\right)\right]}{\left[\sqrt{x^2+3x+4}-\left(x+\frac{3}{2}\right)\right]} \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\frac{x^2+3x+4-\left(x+\frac{3}{2}\right)^2}{\sqrt{x^2+3x+4}-\left(x+\frac{3}{2}\right)}\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\frac{x^2+3x+4-\left(x^2+3x+\frac{9}{4}\right)}{\sqrt{x^2+3x+4}-\left(x+\frac{3}{2}\right)}\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\frac{x^2+3x+4-x^2-3x-\frac{9}{4}}{\sqrt{x^2+3x+4}-\left(x+\frac{3}{2}\right)}\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= \lim_{x\to-\infty}\left[\frac{4-\frac{9}{4}}{\sqrt{x^2+3x+4}+\left(-x-\frac{3}{2}\right)}\right] \\
\lim_{x\to-\infty}\left[f(x)-\left(-2x-\frac{3}{2}\right)\right] &= 0
\end{aligned}$$

Note that in the second-last line of the previous calculation, the numerator is a specific number, and both terms in the denominator approach $+\infty$ as $x \to -\infty$. This explains why the limit is 0. Therefore, according to the definition of slant asymptote, the line $y = -2x - \frac{3}{2}$ is a slant asymptote. See Figure 8.11.

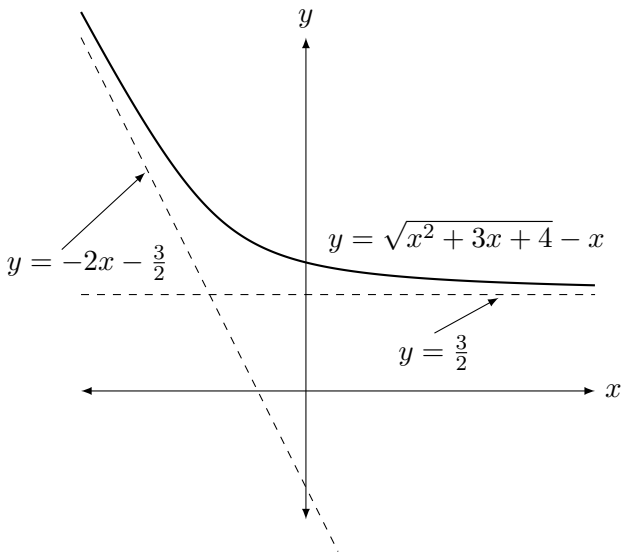



Figure 8.11: The graph of the function $f(x) = \sqrt{x^2+3x+4} - x$ has a horizontal asymptote $y = \dfrac{3}{2}$, and a slant asymptote $y = -2x - \dfrac{3}{2}$.

**EXERCISES**

(Answers below.)

For each function, determine formulas for any vertical, horizontal, or slant asymptotes by calculating appropriate limits. Use the results of limit calculations to help you sketch each graph by hand, then check your work by sketching each graph using software.

1. $y = \dfrac{1}{2x-3}$
2. $y = \dfrac{2x-1}{3x+5}$
3. $y = \dfrac{x^2-3x+2}{x^2+3x+2}$
4. $y = \dfrac{x^2-3x+2}{2x^2+5x+2}$
5. $y = \dfrac{4x^2+4x-8}{2x^2-2x-12}$
6. $y = \dfrac{3x^2+3x-6}{x^2+4x+4}$
7. $y = \cot x$
8. $y = \sec x$
9. $y = \csc x$
10. $y = \tan^2 x$
11. $y = \dfrac{x^2-1}{2x-3}$
12. $y = \dfrac{x^2-4}{x+2}$
13. $y = 3^x$
14. $y = \log_{10} x$
15. $y = \dfrac{\sqrt{2x^2+3}}{x-1}$
16. $y = \dfrac{\sqrt{3x^4+5}}{x^2+4}$
17. $y = \dfrac{x^3-x^2}{x^2+10}$
18. $y = \dfrac{x^3-x}{x+1}$
19. $y = \sqrt{9x^2+x} - 3x$
20. $y = \sqrt{x^2+2x} + x$

21. Consider the family of functions $f(x) = \dfrac{x^2-k}{x^2+1}$.

Explore this family of functions for various values of the parameter $k$. Sketch graphs of such

functions for representative values of $k$. How does the behaviour of the graph change as the value of $k$ changes? Focus on the asymptotes of the graphs and the general overall behaviour. Do your work by hand and then check your work using your favourite graphing software.

22. Repeat Exercise 21 for the family of functions $f(x) = \dfrac{x^2}{x^2 - k}$.

23. Repeat Exercise 21 for the family of functions $f(x) = \dfrac{x^2 + x}{x^2 - k}$.

24. Repeat Exercise 21 for the family of functions $f(x) = \dfrac{x^3 + x}{x^2 - k}$.

---

Answers: 1. $x = 3/2$, $y = 0$; 2. $x = -5/3$, $y = 2/3$; 3. $x = -2$, $x = -1$, $y = 1$; 4. $x = -2$, $x = -1/2$, $y = 1/2$; 5. $x = 3$, $y = 2$; 6. $x = -2$, $y = 3$; 7. $x = \pm\pi$, $x = \pm 2\pi$, etc.; 8. $x = \pm\pi/2$, $x = \pm 3\pi/2$, etc.; 9. $x = \pm\pi$, $x = \pm 2\pi$, etc.; 10. $x = \pm\pi/2$, $x = \pm 3\pi/2$, etc.; 11. $x = 3/2$, $y = x/2$; 12. none (or should we say $y = x - 2$?); 13. $y = 0$; 14. $x = 0$; 15. $x = 1$, $y = \pm\sqrt{2}$; 16. $y = \pm\sqrt{3}$; 17. $y = x - 1$; 18. no asymptotes; 19. $y = 1/6$, $y = -6x - 1/6$; 20. $y = -1$, $y = 2x + 1$

## DIGGING DEEPER

**Is it possible to define two curves being asymptotic to each other?**

Study the definition of slant asymptote. Carefully observe the structure of the definition. Can you extend the definition to define a curve being asymptotic to another curve as $x \to \infty$ or as $x \to -\infty$? If so, the next step will be to try to guess some curves that are asymptotic to each other and then use your definition to test your guesses. How will you verify your work?

If you're successful, then well done! Next, can you do the same for two curves that are asymptotic as $x$ approaches a specific number? (The asymptote would be vertical in such cases.) Once again, the next step will be to guess some curves that are asymptotic to each other in this sense, and then to use your definition to test your guesses. Once again, how will you verify your work?

This kind of work, where you have to make some small advances on your own, will deepen your understanding significantly, in a way that is quite different, and probably much more satisfying, than simply following instructions.

## SUMMARY

In this section we have defined vertical, horizontal, and slant asymptotes, and we have illustrated how to calculate them with a number of examples.

# 8.3 What is Infinity?

## OVERVIEW

Infinity is not a number, but it is something, and it is worth discussing what exactly it is, particularly as there are so many misconceptions about it. Processes in calculus involve various aspects of infinity, and so the subject of infinity is particularly relevant in calculus.

Although the symbol $\infty$, and the way it's used in some textbooks, may lead some to believe that infinity is a number, at the level of understanding of this textbook, infinity is decidedly **not** a number. So what is infinity, then?

Let's start by thinking about the natural numbers. A basic property of the natural numbers is that you can always add the number 1 to a natural number, and the result is another natural

number. For example, 7 is a natural number, and it is possible to add 1 to 7, with the result being 8, another natural number. Now this is true no matter how large the natural number you choose, because this is a property of *all* natural numbers. What, then is the largest natural number?

You will be able to understand that based on this property of natural numbers, there is no largest natural number. Suppose someone proposes to you that some natural number, no matter how large, is the largest natural number. You could counter this proposal by simply adding 1 to the proposed largest natural number to produce a natural number that is even larger. But that new natural number is not the largest either, because you can also add 1 to it to obtain an even larger one.

So there is no largest natural number. One could say that there is an unlimited number of natural numbers. Another way to say this is that there is an infinite number of natural numbers. This usage of the word infinite summarizes the fact that there is an unlimited number of natural numbers.

There are many cars on Earth, but if you had to do so, you could count all of them. You could take a super-snapshot of Earth at a particular time, and then you could carefully examine this photograph and count all of the cars on Earth. No doubt the number is very large, but it is a natural number. The number of cars on Earth is *finite*, because in principle you could count the number and the result is a natural number. Similarly, you could (in principle) count all of the atoms on Earth, at a particular time, and this too is a natural number, so we say the number of atoms on Earth is finite.[1]

Similarly, there are collections of numbers that are finite. For example, consider the collection of odd numbers that are between 1 and 100 inclusive. There are 50 such numbers, right? We could say that the set of odd natural numbers less than or equal to 100 is finite. You could easily construct any number of finite sets, such as the set of prime natural numbers that are less than 1000, the set of even numbers between 5000 and 9000 inclusive, and so on. So we have finite sets, such as the ones mentioned in this paragraph, and then we have infinite sets, such as the set of all natural numbers.

In order to be able to have some way of talking about the number of elements of a set in a unified way, whether the set is finite or infinite, mathematicians have coined the term *cardinality*. It doesn't make sense to speak about the number of elements in the set of all natural numbers, because there is no such number. It does make sense to speak about the number of elements in the set of odd numbers between 1 and 100 inclusive; this number is 50. We can say that the cardinality of the set described in the previous sentence is 50, and the cardinality of the set of all natural numbers is infinite. Thus, the concept of cardinality gives us a way of speaking about the "size" of a set, whether the number is finite or infinite.

Introducing the concept of cardinality may seem unnecessary, but let's discuss something that is potentially shocking. It certainly shocked numerous mathematicians when they learned about it from Georg Cantor about a century ago:

> There are different infinities, of different "sizes." If you prefer, there are different "levels" of infinity.

Is this not mind-boggling?? Are there really infinite sets that have different cardinalities?

To understand this amazing fact about infinities (yes, we should use the plural now), we'll first have to think about how to compare the cardinalities of two sets that are infinite. For example,

[1] It appears that some of the most serious problems we have on Earth is that we humans collectively treat some of our limited resources as if they were infinite instead of finite.

imagine the set of all natural numbers (we'll call it $A$), and then imagine the set that includes all natural numbers and that also includes the number 0 as well; we'll call this set $B$. Now it seems reasonable to say that set $B$ is larger than set $A$; after all, set $B$ includes everything in set $A$, and set $B$ also includes one number that is not in $A$. In the language of set theory, we would say that set $A$ is a proper subset of set $B$. However, this is not the way we currently understand the sizes of infinite sets, as you will see in the next few paragraphs.

Cantor came up with a criterion for comparing infinite sets that led him to his revolutionary understanding of infinities. He said that two sets have the same cardinality if you could set up a one-to-one correspondence between the two sets. That is, you have to be able to pair the elements of the two sets, so that each pairing matches an element of one set with an element of the other set, no element of either set is included in more than one pairing, and each element of each set is included in some pairing.

Think about a sports stadium with $50,000$ seats. Now imagine that the stadium is full of people, so that each seat is occupied by one person, no seats are empty, and each person in the stadium is in a seat. You don't need to count the people to determine how many of them are in the stadium; you can immediately conclude that there are $50,000$ people in the stadium, because the people are in one-to-one correspondence with the seats, and you know how many seats there are.

Now let's apply this concept of comparing cardinalities to infinite sets. In particular, consider the sets $A$ and $B$ described a few paragraphs ago. There exists a one-to-one correspondence between the sets $A$ and $B$, and so they have the same cardinality! Even though we have some sort of sense that we should be considering $B$ to be bigger than $A$, according to Cantor's definition of equal cardinalities, these two sets have the same cardinality! Can you come up with a one-to-one correspondence that confirms this?

A good way to intuitively understand this is through the story of Hilbert's Hotel. David Hilbert was one of the greatest mathematicians of about a century ago, and he constructed a series of "thought experiments" involving a hypothetical hotel that has an infinite number of hotel rooms. I imagine the hotel rooms all in a (very long!) row, rather like a motel, to model the natural numbers as we would normally plot them along a number line.

Suppose that all of the infinite number of rooms in Hilbert's Hotel are occupied at the moment, so that there are no vacant rooms. If a new prospective guest walks into the hotel's lobby desperately asking for a room for the night, is he or she out of luck? Well, not necessarily, according to the clever front desk clerk. The clerk merely asks each guest in the hotel to vacate their room and shift one room over. That is, the person in Room 1 moves to Room 2, the person in Room 2 moves to Room 3, and so on. Each existing guest is perfectly well-accommodated, but now Room 1 has been vacated, and so it is available for the new guest. Problem solved!

Once you have let this remarkable solution sink in, you might then understand why there is a temptation among some people to express this shifty business as

$$\infty + 1 = \infty \quad \textbf{(NO!!)}$$

Resist the temptation to do this! This kind of equation encourages us to treat $\infty$ as if it were a number, but we have already argued that this is not so! So avoid such nonsensical equations. Nevertheless, you can also understand the temptation to write such an equation, because it does (in a way) capture an important property of Hilbert's Hotel; even if all of the infinite number of rooms is occupied, space can always be made available for one more guest. Mind-boggling! Infinity sure is unusual.

Does this story help you to feel a bit better about the fact that the sets $A$ and $B$, described earlier, can be placed in one-to-one correspondence, and therefore have the same cardinality? Would it help more if you could find an explicit formula for such a one-to-one correspondence? Here's one,

perhaps the simplest one, that does the trick: $f(n) = n + 1$. The same formula describes the way existing guests must shift rooms: The guest in Room $n$ must shift to Room $n + 1$.

You can iterate the front desk clerk's shifty technique to accommodate two new guests, three new guests, and indeed, any *finite* number of new guests. (What is the shifting formula in such cases if there are $m$ new guests?) But this kind of shifting clearly won't work if an infinite number of new guests arrive, right? For example, let's suppose that there is a neighbouring hotel that is much like Hilbert's Hotel, in that there are an infinite number of rooms, all currently occupied by guests. There is a power outage at the other hotel; is there any way that this infinite number of guests can be squeezed into Hilbert's Hotel, with each guest having his or her own room? If there were only 5 new guests, we could just shift each existing guest five rooms over. But with an infinite number of new guests, shifting in this way doesn't work. What does it mean to shift each existing guest an infinite number of rooms over? This is meaningless! Where does the guest currently in Room 37 get moved? Because $\infty$ is not a number, saying that the guest in Room 37 should be moved to Room $37 + \infty$ has no meaning!

But the front desk clerk is very clever, and decides that if for each $n$, the guest in Room $n$ shifts to Room $2n$, then each existing guest will still be accommodated (in all the even-numbered rooms), and yet an infinite number of rooms (the odd-numbered ones) will have been vacated, allowing all of the guests displaced from the other hotel to be accommodated also! Isn't this amazing?

The previous paragraph shows that the cardinality of the even natural numbers is the same as the cardinality of the natural numbers. In some intuitive way, we would wish to say that there are half as many even numbers as natural numbers, but no, our intuition is way off when it comes to the cardinality of infinite sets. Similarly, the cardinality of the odd numbers is also the same as the cardinality of the natural numbers. What is a simple formula for a one-to-one correspondence demonstrating this latest fact?

Once again, one might be tempted to write

$$\infty + \infty = \infty \qquad \text{or} \qquad 2\infty = \infty \quad \textbf{(NO!!)}$$

to express this strange fact, but one should really avoid doing so, as $\infty$ is not a number, and therefore can't be combined in an equation like this according to the usual rules for manipulating numbers. But you can certainly see why such nonsensical equations are written in some places; they are attempts to express strange and wonderful properties of infinity in a form that is not appropriate for communicating such facts.

What if there were two or three other copies of the Hilbert Hotel, whose occupants all had to be squeezed into the Hilbert Hotel? Would you be able to do so if you were the desk clerk? Which formula proves that such redistributions of guests are possible? What if there were $m$ total copies of the Hilbert Hotel (including the Hilbert Hotel); can you do the redistribution? What is a formula that proves that such a redistribution is possible?

If you were able to complete the tasks in the previous paragraph, you will now be convinced that the cardinality of $m$ copies of the natural numbers, taken as one giant set, is the same as the cardinality of one copy of the natural numbers by itself. Remarkable!

In the previous paragraph, $m$ is a finite number. What if you had an infinite number of hotels like the Hilbert Hotel? Would you be able to fit all of the guests in all of these infinite number of hotels into just one Hilbert Hotel by redistributing all of the guests? Surely this is impossible, right? At least it's not possible using the method of the previous paragraphs for a finite number of copies of the natural numbers. It's worth pausing right now, turning away from this page, and mulling this over for some time. Return to your reading only after you have mulled things over for a while, and after having written your thoughts in your research notebook.

After mulling it over, what do you think? In fact, it is indeed possible! The cardinality of an infinite number of copies of the natural numbers is the same as the cardinality of the natural numbers! Wow! It is a little more challenging to come up with an explicit formula for a one-to-one correspondence in this case. It may help you to sketch a diagram, where each row of the diagram corresponds to a copy of the natural numbers. Then ask yourself if there is a systematic way to step your way through the entire (infinite) array of numbers, such that you are certain to eventually step on each number in each row. Doing this may help you to understand that this is possible, and provided your pathway is simple enough, you may also be able to write a formula for the correspondence. This is a challenging task, but have fun with it!

We stated earlier on that there were different levels of infinity, but so far we have only encountered one, the cardinality of the natural numbers. Each of the infinite sets we have constructed so far has the same cardinality. It turns out that the cardinality of the real numbers is greater than the cardinality of the natural numbers. The proof that this is so is due to Cantor, again, and it is based on a beautiful idea nowadays called Cantor's diagonal argument, which I'll now describe.

Consider the real numbers between 0 and 1. Cantor showed that the cardinality of this set is not equal to the cardinality of the natural numbers by proving that it is not possible to place the two sets into one-to-one correspondence. He did this by using a proof by contradiction, which is to assume that it is possible and then demonstrate a contradiction, showing that the original assumption is false. So, let's retrace Cantor's steps by assuming that it is possible to construct a one-to-one correspondence between the natural numbers and the set of real numbers between 0 and 1. In effect, this assumption is that you can place the entire set of real numbers between 0 and 1 in a list in some way. For example, here is part of one such proposed list:

$$
\begin{aligned}
&0.\textcolor{red}{3}715682\ldots\\
&0.4\textcolor{red}{9}31657\ldots\\
&0.11\textcolor{red}{5}3267\ldots\\
&0.047\textcolor{red}{4}749\ldots\\
&0.9535\textcolor{red}{3}60\ldots\\
&0.008884\textcolor{red}{1}\ldots\\
&0.558332\textcolor{red}{2}\ldots\\
&\vdots
\end{aligned}
$$

Clearly we can't display the entire list, nor can we even show the complete decimal expansion of each number in the list, but the assumption is that this can be done. Cantor then argued that this assumption is incorrect by constructing a real number between 0 and 1 that is not in the list. Do this by constructing a number that differs from the first number in the first decimal place, differs from the second number in the second decimal place, differs from the third number in the third decimal place, and so on. You can do this according to some rule to make it easier; for example, if the given digit is a 3, then make it a 5, and if the digit is not a 3, then make it a 3. Look at the list of numbers above, and apply this rule to the red digits to construct a new number:

$$0.5333533\ldots$$

The particular rule used is not essential; many other rules would work just as well. Consider the new number just constructed and note that it is not in the original list of numbers. You can tell it is not in the original list, because it is not the first number in the list (it differs in the first decimal digit), it is not the second number in the list (it differs in the second decimal digit), it is not the 47-th number in the list (it differs in the 47-th digit), and so on. Therefore, it is not in the list, and the assumption that we had a complete list of all real numbers between 0 and 1 is false.

Can you obtain a complete list of all real numbers between 0 and 1 by just including this new number at the top of the list? No. You can see that this attempt will not work by applying Cantor's diagonal argument again to the new list to construct yet another real number between 0 and 1 that is not in the new list either. No matter how many newly constructed numbers you add to the top of the list, it will never be a complete list of all real numbers between 0 and 1.

The same argument can be applied to any proposed complete list of real numbers whatsoever. Isn't this an ingenious argument? And isn't the result absolutely remarkable?

Thus, it is not possible to list all of the real numbers between 0 and 1. Another way to say this is that it is not possible to place the real numbers between 0 and 1 in one-to-one correspondence with the natural numbers, and therefore the cardinality of the real numbers between 0 and 1 is different from the cardinality of the natural numbers.

It turns out that the cardinality of the set of all real numbers is the same (!!) as the cardinality of the set of real numbers between 0 and 1. Can you argue that this must be true? Hint: If you can construct a one-to-one function that maps the entire real line into the interval of real numbers from 0 to 1, then this would be an explicit proof. A function that maps the other way would work just as well. Search your memory banks for a graph from high school that will do the trick!

Isn't it mind-boggling that the number of real numbers between 0 and 1 is the same (in the sense of one-to-one correspondence) as the number of all real numbers? These two sets have the same cardinality. Because the set of natural numbers is contained within the set of real numbers, and these two sets cannot be placed in one-to-one correspondence, we say that the cardinality of the real numbers is greater than the cardinality of the natural numbers. Thus, we have established the existence of two levels of infinity. Here is some standard terminology: Sets that either contain a finite number of elements or can be placed in one-to-one correspondence with the natural numbers (such as the even numbers, the odd numbers, the integers,[2] and so on) are called **countable** sets. Infinite sets that are countable are also called countably infinite. Infinite sets that are not countable are called **uncountable**. Thus, we have (so far) two levels of infinite sets, sets that are countably infinite (such as the natural numbers) and sets that are uncountable (such as the real numbers).

Are there any levels of infinity that are between the cardinality of the natural numbers and the cardinality of the real numbers? Cantor conjectured in 1878 that the answer to this question is no, in what is now called the continuum hypothesis. Attempts were made for many years to either prove the continuum hypothesis or to discover a counterexample, which culminated in a publication by Kurt Gödel in 1940 in which he showed that it is impossible to disprove the continuum hypothesis within standard set theory. Paul Cohen showed in 1963 that the continuum hypothesis cannot be proved within standard set theory either! This remarkable set of results shows that the continuum hypothesis is independent from standard set theory. To learn more about this very strange result, look up Gödel's incompleteness theorem.

We stated earlier that there is a whole hierarchy of infinities, but so far we have only seen examples of two levels of infinity, that of the natural numbers and that of the real numbers. How can one construct higher levels of infinity? What about the cardinality of the set of points in the plane that you used so much in high school to study functions? Surely the cardinality of the number of points in the plane is greater than the cardinality of the real line? But no, the cardinalities are the same! Proving this is more challenging, though, than Cantor's diagonal argument. (Look up the Schröder-Bernstein theorem if you are curious about this.) Similarly, the cardinality of the points in three-dimensional space is also the same as the cardinality of the real number line. Thus, simply moving to higher-dimensional spaces does not give us a greater level of infinity.

[2]Can you prove that the integers can be placed in one-to-one correspondence with the natural numbers by constructing a suitable formula?

How does one construct sets with cardinalities at higher levels of infinity? We shall leave this discussion for another time, but if you are curious you can consult a work on mathematical analysis or set theory.

Before concluding this discussion, it is worth mentioning that the ancient Greeks already distinguished between what they called actual infinity and potential infinity, and this is a useful distinctions. Actual infinity is reserved to describe an infinite set in its entirety, such as the set of natural numbers taken as a whole, or the set of real numbers taken as a whole. Potential infinity is reserved for the idea of a quantity that is increasing without bound, so that the quantity gets larger and larger with each step of the process, with no limitations on how large it gets. Our discussion of limits as $x$ "approaches infinity" fits this sense of potential infinity. In fact, we discuss limits as

$$x \to \infty \qquad \text{and} \qquad x \to -\infty$$

where we typically envision $x$ "moving" to the right indefinitely in the first case, and "moving" to the left indefinitely in the second case. It's worth emphasizing again that in both of these cases "infinity" is not a place, but rather this is a process of imagining what happens when a quantity ($x$ in this case) either "moves" to the right indefinitely or "moves" to the left indefinitely.

## HISTORY

### Augustin Louis Cauchy (1789–1857) and Niels Hendrik Abel (1802–1829)

Cauchy was one of the most prolific mathematicians in history (second to Euler), and his research spanned many branches of mathematics. He also made contributions to physics and astronomy. His greatest contributions were in the branch of mathematics called analysis, which is where calculus lies in the grand scheme of mathematics. Cauchy was the leading French mathematician of his time.

Abel was born and raised in rural Norway, and when he was 18 years old his father died, leaving the family desperately poor. Despite his poverty, and his responsibilities in trying to keep his family fed and housed, he managed to produce brilliant and original mathematics. One of his crowning achievements was to prove that there is no possible formula (analogous to the quadratic formula) for the solution of algebraic equations of fifth degree or higher, solving a long-standing problem. To do this, he made essential use of a branch of higher mathematics called group theory, which he invented for the purpose.

Abel managed to get funding to travel to Paris in 1826, which was one of the European centres for mathematics and science. There he hoped to make acquaintance with some of the leading mathematicians. He was indeed introduced to some, but his shyness prevented him from impressing. He did hit it off with August Leopold Crelle, a mathematics enthusiast who became Abel's friend. With Abel's encouragement, Crelle founded the world's first mathematics journal. In turn, Crelle did everything in his power to lobby universities to appoint Abel as a professor, a position that Abel was very worthy to hold. Crelle was not immediately successful; Abel ran out of money, and so he returned home to Norway in 1827, deep in debt, and now ill with tuberculosis. Abel continued to produce amazing works of mathematics, and Crelle continued to lobby, and eventually Abel was appointed a professor of mathematics in Berlin. Unfortunately the letter announcing the appointment was received at Abel's home two days after he had died.

The sad story of Abel may have ended differently. While he was in Paris he submitted a paper to the French Academy, hoping that it would make French mathematicians aware of him, which would help him secure a job as a professor. This could have ended his poverty. The paper itself contained, according to Jacobi, the greatest advance in integral calculus in the 1800s, and once it was read it was indeed universally hailed as brilliant. However, the paper was assigned by the French Academy to Cauchy to read and assess. Cauchy misplaced the paper, then forgot about its existence, and that was that. While Abel was waiting for a response, his money ran out, and he had to leave Paris. If Cauchy had read the paper he would certainly have realized the greatness of its author, and Abel's reputation in France would have been secured. What may have been a long and productive life for Abel came to an unfortunate and premature end.

Together with Gauss, Cauchy and Abel led a movement in mathematics to state theorems carefully and prove them rigorously. They were successful in setting high standards and in clearing up numerous confusions and errors, particularly in analysis.

**SUMMARY**

Infinity is a strange beast. Infinity is not a number, but it is a vital concept in calculus, and has a number of aspects that are worth grappling with.

Mathematicians measure the sizes of sets using the concept of cardinality, which allows them to compare infinite sets. The idea of one-to-one correspondence of sets is a key one; if two sets can be placed in one-to-one correspondence then they have the same cardinality; otherwise, the two sets do not have the same cardinality.

The set of natural numbers, the set of even natural numbers, the set of odd natural numbers, and set of integers all have the same cardinality, even though at first glance we really want to say that some of them have different sizes. However, in comparing infinite sets, we can't rely on feelings, we must use a precise concept, and the concept of cardinality serves.

Sets that have the same cardinality as the natural numbers (such as the set of even natural numbers, the set of odd natural numbers, the set of integers, and so on) are said to be countable. It's remarkable that any finite number of countable sets joined together is still countable. Even more remarkable is the fact that a countably infinite number of countable sets all joined together is still countable.

It turns out that the real numbers are not countable, a remarkable fact that can be proved using Cantor's diagonal argument. In this sense, the level of infinity of the real numbers is greater than the level of infinity of the natural numbers. Is there a level of infinity between the two? It turns out that this question is undecidable, in a sense that requires an exploration into mathematical logic; search for the continuum hypothesis if you would like to embark on such a journey.

We can also distinguish between actual infinity, which is a measure of a set as a whole (in terms of its cardinality), and potential infinity, which refers to a step-by-step process that continues indefinitely, such as the way we have conceptualized limits so far in this book.

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 9

# Rates of Change in Applications

**OVERVIEW**

In applications of calculus, it is often the rate of change of a quantity that is of primary importance. In differential calculus, the focus is on determining the rate of change for a given quantity. In integral calculus, it is the rate of change of a quantity that is somehow known (from experiment, for example, or from some other analysis), and the focus is on determining the function that describes the quantity.

Let's review the conceptual foundation of the calculus story, as we've told it so far.

A primary reason for studying mathematics is that it enables us to quantitatively describe the world and thereby better understand it. Functions are typically used as mathematical models of worldly phenomena; we hope that by mathematically analyzing the functions, we may learn about them, and then transfer what we've learned to better understand the phenomena that we modelled.[1] In this perspective, calculus is a tool for analyzing functions so that we can ultimately better understand the world. There are other perspectives on calculus, and you will learn about them if you continue your studies, but for now this perspective is reasonably practical.

A fundamental concept is the slope of a tangent line to the graph of a function. The slope indicates the rate of change of the quantity modeled by the function. And the derivative is the "algebraic" (i.e., symbolic) machinery that allows us to calculate rate of change. In summary:

$$\begin{array}{ccccc} \text{slope} & & \text{rate of change} & & \text{derivative} \\ \updownarrow & \leftrightarrow & \updownarrow & \leftrightarrow & \updownarrow \\ \text{geometric} & & \text{physical} & & \text{algebraic} \end{array}$$

So we have various perspectives (geometric, physical, and algebraic)[2] on this fundamental concept, and different phrases (slope, rate of change, derivative) to describe each perspective, but ultimately there is one concept here. Understand how these different perspectives are related, and how they all mean essentially the same thing, and you will have understood something valuable — one of the core fundamental ideas of calculus.

In this chapter we'll discuss the connections among these fundamental perspectives with a few examples.

Let's begin by discussing motion as an example. It's important to discuss motion for a number of reasons. First, motion is familiar and we can readily visualize it, and phenomena that are

[1] Of course, this is a dynamic process. One tests models by confronting them with observational or experimental data, and even the best models are typically found wanting in some way. Then one attempts to modify the models to improve them, or to create better models. Then they are tested, and the whole process repeats.

[2] We could even add "numerical" as a perspective as well, since the first process we used to estimate the slope of a curve was a numerical procedure.

concrete and easy to visualize are good ones to begin with when striving to understand a new concept. Second, almost every phenomenon in science has some motion associated with it, and so understanding motion will help us understand many scientific descriptions of the world. Finally, once we understand the descriptions of motion in terms of rates of change, we'll be able to more easily transfer the same kinds of conceptual understanding to other situations.

In learning mathematics, we typically start with the easiest situations, and then gradually increase the complexity and difficulty level. To understand motion, it is reasonable to begin with the simplest kinds of motions, such as motion in a straight line.[3]

Consider a car that moves along a straight test road. There are no other cars present, so the car can move forward or backwards without danger of collisions. If we mark the road with a scale, much like a number-line, then we can note the car's position along the road at any time. We can then plot the position for each time on a graph; because the car can't be in two positions at the same time, the result is the graph of a function. Thus, the position of the car can be thought of as a function of time.

It's customary to plot time along the horizontal axis of such a graph, and to plot the position along the vertical axis. It's also customary to identify a particular time as the "starting time," and label that as $t = 0$. In other words, it's customary to imagine a timer being used to time the car's travel, with $t = 0$ representing the instant that the timer is started.

Let's suppose that the car moves along this straight test road from position $y = +3$ m to position $x = +7$ m in 5 s, then stops, reverses, and continues to position $x = +1$ m in an additional 3 s. To make this initial example as simple as possible, let's also suppose that the speed is constant for each of the two legs of the journey we have just described.

We could represent the journey graphically as in Figure 9.1; this figure is not yet a position-time graph, but rather a simpler representation that is called a motion diagram in some textbooks.

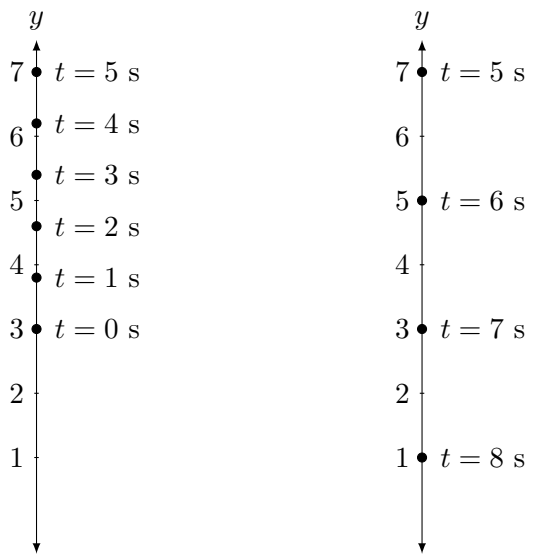



Figure 9.1: An example of a simple journey along a straight path. The motion is back-and-forth along a single straight road, but the diagram shows the first part of the journey on the left and the second part of the journey on the right for clarity. Can you visualize the motion based on following the dots in the graph?

[3]Unless you consider that stillness, which is no motion at all, is even simpler, although some would argue that stillness is a special case of straight line motion (with zero speed).

We might imagine that the road is oriented so that north is towards the upper part of the page (or screen, if you are reading this on a screen), and south is towards the bottom of the page or screen. If you follow the dots on the diagram in the order of the time labels, the car moves north at a constant speed for 5 s, stops, reverses, and then moves south at a constant speed for 3 s. A real car would pause momentarily before changing direction, and would gradually slow down before stopping, and would gradually speed up after starting to move south, but to simplify our discussion we'll pretend that the car changes direction instantly. This is somewhat typical of mathematical modelling of real physical phenomena; for the sake of simplification, the model becomes unrealistic, but one hopes that one can gain insight into the phenomenon by analyzing the simple model, and that the model can be subsequently improved by making it more realistic, at the cost of making it less simple.

**Question:** Does the diagram in Figure 9.1 help you to visualize the motion of the car, or otherwise to understand it? For example, can you tell at a glance that the car moves at a constant speed as it moves north? How can you tell? How about the second leg of the journey, when the car moves south; can you tell that the car moves at a constant speed there? Which of the two speeds is greater, the speed for the first leg of the journey or the speed for the second leg of the journey? How can you tell?

The dots in the diagram represent "snapshots" of the car at the times indicated. Because the snapshots were taken at equal time intervals (they are separated by 1 s), and because the dots are equally-spaced along the road, the car travels equal distances in equal times; this means that the car is moving at constant speed. Because the dots are spaced farther apart on the second leg of the journey, the car moves faster on the second leg of the journey; more distance is covered per second on the second leg of the journey than on the first leg.

You can calculate the speed of the car on the first leg of the journey as follows:

$$\text{speed} = \frac{\text{distance}}{\text{time interval}} = \frac{7-3}{5-0} = 0.8\,\text{m/s}$$

The car's speed on the second leg of the journey is

$$\text{speed} = \frac{\text{distance}}{\text{time interval}} = \frac{7-1}{8-5} = 2\,\text{m/s}$$

These calculations confirm that the car's speed is greater on the second leg of the journey.

It is popular (and useful) to display the same information about the motion from the diagram in Figure 9.1 in a two-dimensional plot called a position-time graph; see Figure 9.2. Each of the indicated points on the position-time graph represents the location of the car at the time of a snapshot. For example, the first indicated point at the far left of the graph represents the fact that when the timer reads $t = 0$ s, the position of the car is $y = 3$ m. The sixth indicated point represents the fact that when the timer reads $t = 5$ s, the position of the car is $y = 7$ m.

Similar interpretations apply to the other indicated points in Figure 9.2. However, the car certainly exists and has locations at the times between the snapshots, and so the actual position-time graph of the moving car should be the graph of a continuous function. We are assuming that the car moves at a constant speed during each leg of its journey, which means that the actual position-time graph of the car's motion is the graph of the continuous function shown in Figure 9.3.

The actual motion takes place along the road, which is represented by the $y$-axis; the car moves straight north, stops, and then moves straight south. The shape of the position-time graph does not represent the path of the moving car. Each point on the position-time graph represents a position of the car at a certain time.

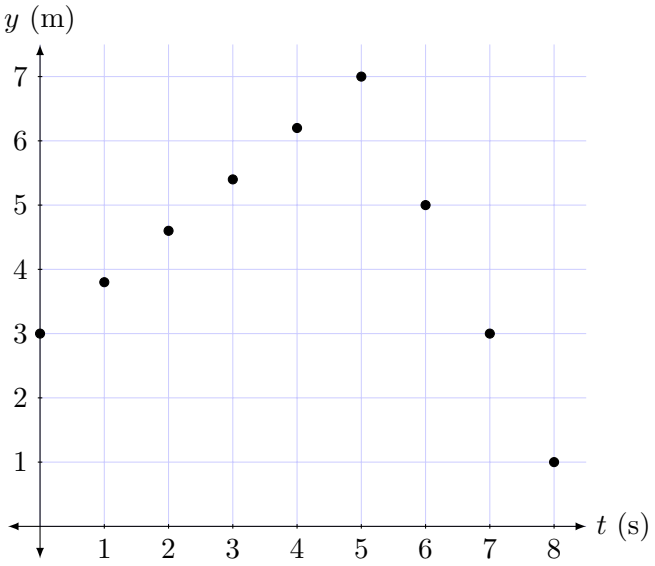



Figure 9.2: The motion represented in the previous figure is represented here in a position-time graph. Each indicated point on the graph corresponds to an indicated "snapshot" point in the previous figure.

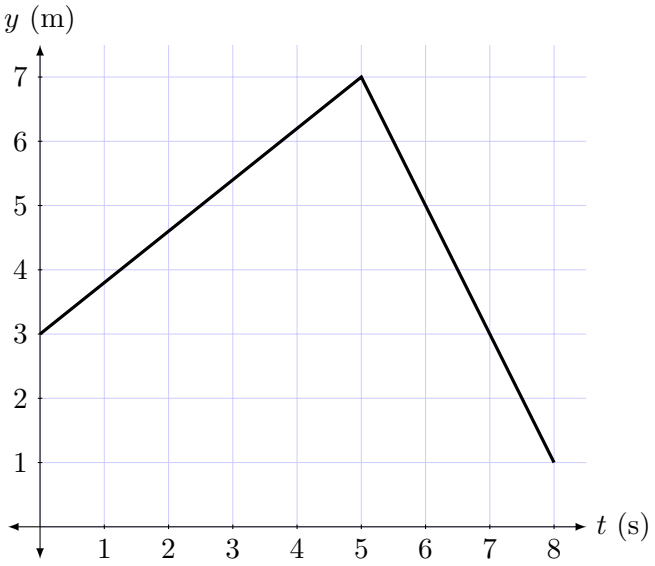



Figure 9.3: The car exists at every moment, and so the actual position-time graph of the car's motion is continuous, as in this figure. What does the slope of each segment of the graph represent?

Can you get a sense for the car's motion from the position-time graph? This is a valuable skill, which you can improve if you devote some time to it, because once you figure this out you will be able to interpret all kinds of other graphs as well.

For example, note that the slope of the position-time graph has units m/s. The value of the slope of the first segment of the graph is +0.8 m/s, and the slope of the second segment of the graph is −2 m/s. What does it mean in terms of the motion of the car that one of these slopes is positive and one is negative? Remember that the slope of the graph represents the rate of change of the quantity plotted on the vertical axis with respect to the quantity plotted on the horizontal axis. Thus, the absolute values of the two slopes represent the speeds of the car on each leg of its

journey. What do the signs of the slopes mean? On the first leg of the journey, the positive slope means that the positions are increasing as time passes, which means the car is moving along the road in the direction in which the road-marker numbers increase. (The way we set up the road, this means north; of course, in another situation, the road could be oriented differently, and the scale could be oriented in two ways along the road.) On the second leg of the journey, the negative slope means that the positions are decreasing as time passes, which means the car is moving along the road in the direction in which the road-marker numbers decrease (i.e., south).

Once you have understood the previous paragraph, interpreting a position-time graph will be straightforward. A positive slope indicates motion in the positive direction (according to the markers on the road) and a negative slope indicates motion in the negative direction.

In physics, the velocity of a moving object is a concept that includes both the object's speed and its direction of motion. Thus, the slope of the position-time graph represents the velocity of the car's motion. The magnitude of the slope represents the speed and the sign of the slope indicates the direction of motion. Figure 9.4 shows the velocity-time graph for the car's motion. The discontinuity in the velocity-time graph at $t = 5$ s corresponds to the sharp corner in the position-time graph at the same time. Both of these features indicate the unrealistic situation that the car's speed changes abruptly; in reality, changes in speed are not sudden.

The physical concept for the rate at which the car's velocity changes is called the car's acceleration. The acceleration corresponds to the slope of the velocity-time graph. As you can see, the acceleration is zero for the first leg of the journey, and also for the second leg of the journey, because the velocity does not change in either leg of the journey. However, because the velocity changes abruptly at $t = 5$ s, the acceleration makes no sense there. As described by Newton's second law of motion, the acceleration of the car is proportional to the total of all forces acting on the car, so where the acceleration is nonsensical, the force acting on the car makes no sense either; this is not physically realistic. This is yet another way of describing the fact that the motion of the car has been modelled unrealistically at $t = 5$ s.

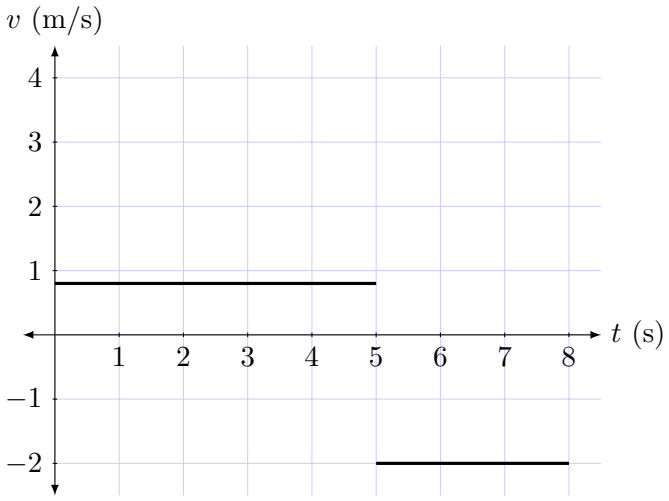



Figure 9.4: A velocity-time graph for the car's motion. Compare this to the position-time graph in the previous figure. The discontinuity in the velocity-time graph corresponds to the sharp corner in the position-time graph, both of which are signs that this model of the car's motion is unrealistic at $t = 5$ s.

It is often helpful to plot position-time graphs and velocity-time graphs together in a vertically-aligned pair, as in Figure 9.5. Compare the two graphs and note that the height of the velocity-time

graph at a particular time is the slope of the position-time graph at the same time. You might like to indicate a few vertically-aligned points on the two graphs and verify this fact.

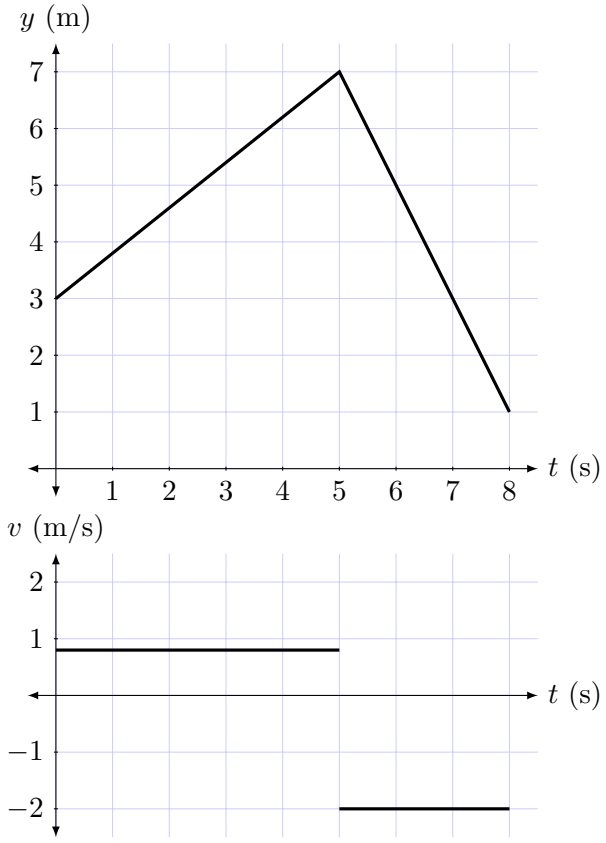



Figure 9.5: Plotting a position-time graph and the corresponding velocity-time graph together, vertically aligned, often aids understanding of the motion. These graphs are for the moving car.

A more realistic, and familiar, situation is tossing a ball vertically upwards. If we ignore air resistance (so the situation is not perfectly realistic), so that only gravity acts on the ball, then the acceleration of the ball has a constant magnitude. You are no doubt familiar with this situation, but it is always helpful to actually toss a ball upwards a few times now to remind yourself of the situation. The ball moves upwards for a while, gradually slowing down, then momentarily stops, then moves downwards and gradually speeds up. If we imagine a vertical measuring tape that we use to record the position of the ball at various times, and if the numbers on the measuring tape increase upwards, then the ball initially moves in the positive direction, then stops momentarily, then moves in the negative direction. In other words, the velocity of the ball is positive for a while (although its magnitude decreases), then the velocity is zero momentarily, then the velocity is negative.

It is a fact (confirmed by numerous experiments) that the acceleration of the ball is about $-10$ m/s$^2$ with the setup we have chosen (i.e., that the position increases in the upwards direction), and is constant if we neglect air resistance. This means that the slope of the velocity-time graph of the ball is a constant value of $-10$ m/s$^2$. Let us suppose that the initial upward speed of the ball is 25 m/s. (This value is unrealistically high unless you are a very strong athlete,[4] but it will serve to illustrate the general character of the graphs that describe such motion.) The velocity-time graph for this initial velocity is plotted in Figure 9.6.

[4]Convert the initial speed to km/h if this is not clear.

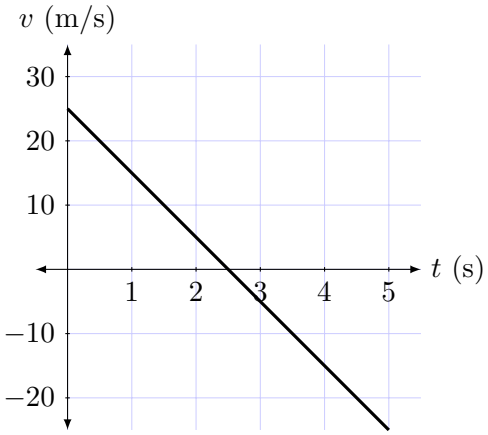



Figure 9.6: A velocity-time graph for a ball thrown vertically upwards with an initial speed of 25 m/s, with the assumption that there is no air resistance.

**CAREFUL!**

**The *sign* of the velocity indicates the direction of motion**

A common error made by beginning students is to look at a velocity-time graph such as the one in Figure 9.6 and interpret the negative slope as meaning that the object moves in the negative direction throughout the 5-s time interval plotted on the graph. **THIS IS NOT CORRECT.** The slope of the velocity-time graph is the acceleration; the fact that the acceleration is negative means the the velocity decreases. It is the **sign** of the velocity (that is, the height of the velocity-time graph, not its slope) that indicates the direction of motion. Thus, the positive velocity in the first 2.5 s corresponds to the fact that the ball moves upwards in this time interval, whereas the negative velocity from $t = 2.5$ s to $t = 5$ s indicates that the ball moves downwards in this time interval. Right at $t = 2.5$ s the velocity is zero; this is the time when the ball momentarily stops.

The absolute value of the velocity indicates the speed. You can read the speed from the velocity-time graph by noting the velocity and then ignoring its sign. Examine the graph carefully; is it clear that the ball slows down during the first 2.5 s of its motion and speeds up during the next 2.5 s?

What does the position-time graph look like for the ball thrown vertically upwards? Well, we know the ball goes up, stops, and then comes down again, so the position-time graph must do the same, because we have chosen the position scale so that it increases in the upwards direction. But can we be a little more precise? Without knowing specifically where the zero-position on the scale is located, no, we cannot be more precise about what the graph looks like. But suppose that the scale is set up so that the zero-position on the scale is located where the ball is released; then we can indeed be more precise. In fact, determining the position-time graph from the velocity-time graph is an example of a problem that belongs to *integral calculus*. The velocity function is the derivative of the position function, so in going from the velocity function to the position function, in effect we have to *anti-differentiate*.

We can explore the idea of anti-differentiation numerically in this context. Knowing that the position of the moving ball is $y = 0$ m at $t = 0$ s, and also knowing that the velocity at that time is 25 m/s, we can sketch a little part of the tangent line to the position-time graph at this initial time; see Figure 9.7. Starting at the initial time $t = 0$ s, the velocity-time graph indicates that the

initial velocity is 25 m/s. This means that the slope of the position-time graph at this time is 25 m/s; this is indicated on the position-time graph by the small piece of tangent line placed at the initial position of $y = 0$ m.

We can repeat the process for other times. For example, we can read from the velocity-time graph that the velocity at $t = 1$ s is 15, and so we can sketch a small piece of tangent line of slope 15 m/s on the position-time graph at $t = 1$ s. However, it's not clear what the vertical placement of this little piece of tangent line should be, and at this stage of our development we can't be sure about this. (This will be discussed extensively if you continue your studies in integral calculus.) For example, if we extend the tangent line sketched at $t = 0$ s out to $t = 1$ s, it will intersect the vertical line at $t = 1$ s at a position of 25 m. However, we can be certain that this is an overestimate of the position at $t = 1$ s, because this is the position that the ball would attain after 1 s if it were moving at a constant speed of 25 m/s. This is not so; we know that the ball gradually slows down in the first second. Thus, we know that the position of the ball after 1 s will be somewhat less than 25 m, but at this point in our development we can't be sure exactly what the position will be. The position has been sketched correctly on the position-time graph, but it's worth thinking a little bit about how you might be able to pin this down exactly, in preparation for developments in your future studies of integral calculus.

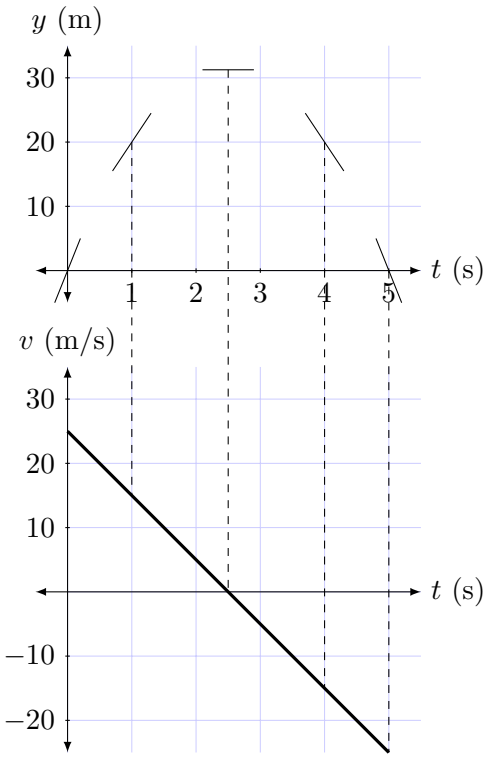



Figure 9.7: To construct a position-time graph for the moving ball from its velocity-time graph, we must anti-differentiate. The first step in an elementary version of this process is to plot a little bit of the tangent line to the graph at $t = 0$. Little bits of tangent lines at other times are also included, but it's not clear at our current level of development where these bits of tangent line should be positioned vertically on the graph. Note that the slope of the upper graph is equal to the height of the lower graph, as indicated by the dashed vertical lines.

## CHALLENGE PROBLEM

### Approximating a position-time graph from a velocity-time graph

Let's continue the discussion in the previous paragraph. How can you approximate the position-time graph given the velocity-time graph? For example, you might assume that the speed is constant during the first second and calculate the position at the end of 1 s. Then, using the new speed at $t = 1$ s (read from the velocity-time graph), you can assume that the new speed is constant for the next second, and calculate the position at $t = 2$ s, and continue similarly until you reach $t = 5$ s. This is certainly not correct, but at least you will have some approximation. How good is the approximation? It's a bit hard to say, isn't it?

But what if we separated the 5 s time interval into 0.5 s sub-intervals instead of 1 s sub-intervals? Then assuming that the initial speed is constant for the first 0.5 s is probably a better approximation than assuming that it is constant for a full second, don't you think? It seems, then, that applying this approximation scheme using 0.5 s sub-intervals will result in a better approximation for the position-time graph than using 1 s sub-intervals.

Now if you are good at programming, you can certainly construct an algorithm that will do this calculation for sub-intervals of arbitrary size. Perhaps you can also produce an animation that will trace out the resulting approximation to the position-time graph, or at least produce the graph itself without an animation.

Here's a good calculus-style question: By making the sub-intervals smaller and smaller, does the approximation to the position-time graph get better and better? It might not be clear how to even make such a judgement, considering that you might have no idea what the final position-time graph should look like.

If you have some knowledge of physics — in particular the kinematics equations for motion in a straight line with constant acceleration — then you will understand that the formula for the position function of the moving ball is

$$y = -5t^2 + 25t$$

and so you will have something to check against as you test your algorithm. For readers who don't have this previous knowledge, you can give this challenge question some thought in preparation for an in-depth discussion later in your studies of integral calculus and/or physics.

## GOOD QUESTION

### Vertically thrown ball with air resistance

How would the position-time graph of a ball thrown vertically upwards be different if air resistance were present? To explore this question you might begin by considering how the velocity-time graph would be different (qualitatively). Perhaps sketch a new velocity-time graph based on the one with no air resistance. Once you have done this, then you might follow a procedure similar to the one described in the text to approximate a new position-time graph. Does your new graph meet your expectations about what it should look like? For example, should the ball return to the thrower's hand in 5 s, or less than 5 s, or more than 5 s? Should the ball's maximum height be equal to, more than, or less than the maximum height without air resistance? Should the position-time graph with air resistance share the same symmetry properties as the position-time graph without air resistance? It will be fun to explore these questions qualitatively, and then if you can, quantitatively! What additional information would you need to explore quantitatively?

The full position-time graph (assuming no air resistance) is in Figure 9.8. Does the graph seem

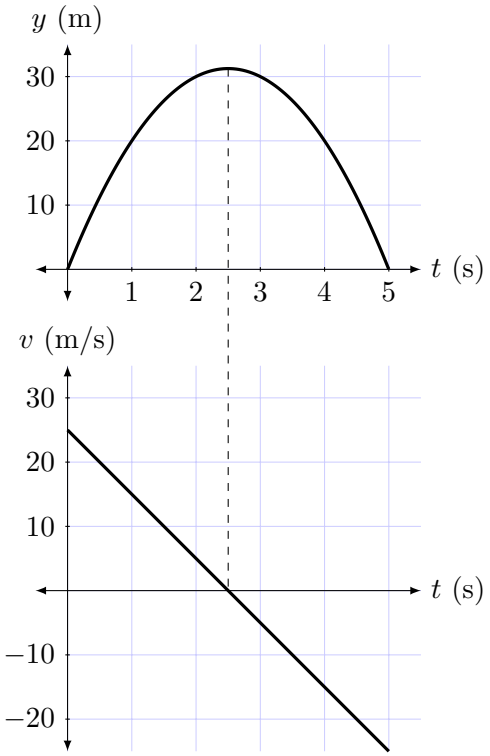



Figure 9.8: The position-time graph and the velocity-time graph for the moving ball discussed in the text.

reasonable? Does it have the general features that make sense to you based on your experiments tossing a ball vertically upwards? Are the graph's symmetry properties reasonable? Does the slope of the position-time graph at each time match with the height of the velocity-time graph at the same time? (Test this at some times to convince yourself that this is so. You can do this by copying the position-time graph into a notebook and sketching some tangent lines on it at various points, and then compare the slopes of the tangent lines with the appropriate heights of the velocity-time graph. Note the scales on the vertical axes, which are different from the scales on the horizontal axis!)

What happens at $t = 5$ s? Does the speed change abruptly to zero and remain zero? The behaviour of the ball after $t = 5$ s is beyond the scope of this discussion. What happens depends on physically what happens to the ball. Does the ball bounce from a hard surface? Does the ball land in mud and stick there without bouncing? Is the ball caught, and so brought to rest gradually? All of these situations are very complicated, and the precise position-time graph after $t = 5$ s will be correspondingly complicated.

Anti-differentiation is a fundamental process in science, and will be discussed extensively in a course on integral calculus, but we can say a few words about it now to highlight its importance. Let's think about the ball thrown vertically upwards for a moment. To analyze its motion, one could start with Newton's second law of motion. The result of such an analysis would result in the ball's acceleration function, which you could plot on an acceleration-time graph. But you are likely to be interested in other questions, such as the speed of the ball at various times, or the position of the ball at various times. To obtain the velocity function from the acceleration function, one anti-differentiates; this process is also known as integration. To obtain the position function, one then anti-differentiates (i.e., integrates) the velocity function.

Laws of physics are typically like this; they don't directly tell us about quantities of primary

interest, but rather they give us relationships involving the rates of change of quantities of primary interest. Then it is up to us to integrate to obtain the quantities of primary interest.

For simple laws of physics, integration suffices. More sophisticated laws of physics are expressed mathematically in terms of what are called differential equations, which are relationships among derivatives of various quantities. The process of solving a differential equation is similar in spirit to solving an integration problem, but may be more complicated.

All of this will be discussed later in your study of calculus and physics, but you now have a bit of a sneak preview and (one hopes) a bit of insight into how the whole process works.

There are numerous other examples of the applications of rate of change in science, engineering, finance, economics, and so on, but the focus of this book is on limits, and so we restrict ourselves to the examples already presented, and leave further examples for another time and place.

**SUMMARY**

In applications of calculus, it is often the rate of change of a quantity that is of primary importance. In differential calculus, the focus is on determining the rate of change for a given quantity. In integral calculus, it is the rate of change of a quantity that is somehow known (from experiment, for example, or from some other analysis), and the focus is on determining the function that describes the quantity.

**Make sure to regularly review the key concepts of this chapter and the previous chapters, and also to regularly review the examples that you have worked through and the exercises that you have done, both in this chapter and the previous chapters. Review and repetition is the key to placing your learning in your long-term memory.**

# Chapter 10

# Sequences, Series, and Limits

**OVERVIEW**

The rough conception of limit presented at first relies on the idea of a sequence of approximations. In this chapter we formally discuss sequences and series, discuss limits in this context, and connect the idea of the limit of a sequence with the idea of the limit of a function.

We have concentrated our attention in this book on the concept of the limit of a function, and its use in calculus, both to calculate slopes and to calculate long-term behaviour for the graphs of functions.

Recall the rough conception of limit that we used when calculating the slope of the graph of a function. We first approximated the slope, then proceeded to improve the approximation in a step-by-step procedure. In our initial numerical approach, we listed the approximations in a table, then asked ourselves if the trend in the values of the approximations approached a definite value. If so, then we call this definite value the limit. If not, then we say the limit does not exist.

Another way to express the point of the previous paragraph more briefly is to say that the derivative is defined in terms of a limit.

If we widen our perspective a little bit, and forget for a moment about the origin of the problem, which was to calculate the slope of a curve, our rough conception of limit amounts to looking at a list of numbers to see if there is a trend in their values. The other purpose for calculating a limit that we have studied so far, which is to look at the trend in function values, could be conceived of in a similar way. For example, you could imagine sampling function values at various $x$-values, and looking at the trend in the sampled function values.

Are there other applications of this basic idea? The answer is yes. For example, the integral, the other main concept in calculus, is also defined in terms of a certain limit. The continuity of a function is also defined in terms of a limit.

Taking a wider, more abstract perspective therefore motivates us to study the concept of the limit of a list of numbers; that is, the limit of a sequence of numbers.

## 10.1 Sequences

A list of numbers is called a **sequence** of numbers; each number in the list is called a *term* of the sequence. One can have a finite sequence of numbers (i.e., one with a finite number of terms), such as

$$1, 3, 8, 19, 11, 12$$

or an infinite sequence (one with an unending number of terms), such as

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots$$

In calculus, one is more interested in infinite sequences, because one is concerned with approximations that can be made arbitrarily accurate as one proceeds indefinitely. But we will briefly review finite sequences as a reminder of your high-school studies.

Technically, you can define an infinite sequence to be a function the domain of which is the natural numbers. That is, $f(1)$ is the first term of the sequence, $f(2)$ is the second term of the sequence, and so on. A finite sequence with $k$ terms can be defined as a function the domain of which is the set of natural numbers from 1 to $k$ inclusive.

A **series** is a sum of numbers, and an infinite series is a sum of numbers with an infinite number of terms. An example of an infinite series is:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

In our work with calculating the slope of a curve, our conceptual understanding involved calculating the slopes of a sequence of secant lines, and determining whether the trend in the sequence of slope values zeroed in on a limiting value. Not every example we looked at had such limiting values; not every function was differentiable at a point of interest.

In the context of infinite sequences, the words convergence and divergence are used to describe the situations where the sequences either have limits or do not have limits. If an infinite sequence of numbers has a limit, we say that the sequence **converges**. If an infinite sequence of numbers does not have a limit, we say that the sequence **diverges**, which is another way of saying that the sequence does not converge.

But what does the limit of a sequence mean? We can connect this kind of limit with one of the kinds of limits we have studied earlier; so called "limits at infinity." Because an infinite sequence can be considered to be a function the domain of which is the natural numbers — i.e., $y = f(x)$, where $x$ is a natural number — you can imagine plotting such a function as a collection of dots on our usual $xy$-axis system. Then deciding whether an infinite sequence has a limit, and what its value is (if it exists), is the same as deciding whether

$$\lim_{x \to \infty} f(x)$$

exists and if so what is its value. Thus, the idea of the limit of a sequence is not much of a new concept; it is very similar to the idea of the limit of a function "at infinity," which we have studied earlier.

**DEFINITION 8**

**Convergence of an infinite sequence**

An infinite sequence that has $n$-th term $f(n)$ converges to a limit $L$ provided that

$$\lim_{n \to \infty} f(n) = L$$

If the limit does not exist then we say that the infinite sequence diverges.

For functions the domains of which are the natural numbers, it is traditional to use $n$ for the independent variable instead of the usual $x$. The terms of a corresponding sequence are typically symbolized by $f(n)$, or $y_n$, or $x_n$, and so on. See Figure 10.1 for an example.

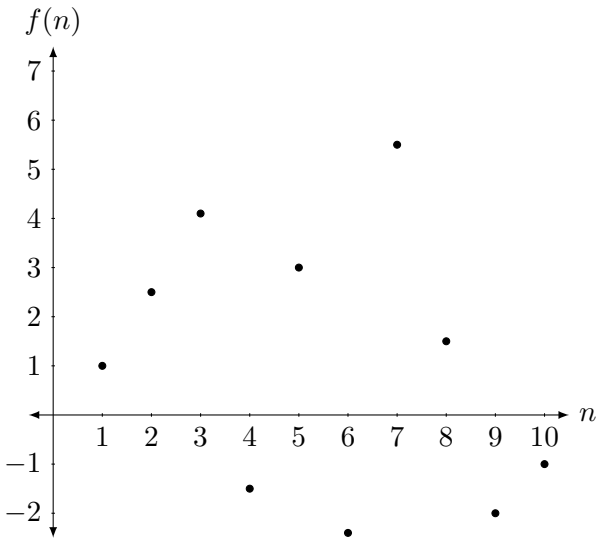



Figure 10.1: A sequence of numbers can be thought of us a function the domain of which is the natural numbers, and so the graph of a sequence is a collection of dots.

Let's look at a few examples. First consider the sequence with terms that satisfy the formula

$$f(n) = \frac{1}{n}$$

The terms of this sequence are $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$, and so on. Does this sequence converge? Well, we have previously studied the function $y = \frac{1}{x}$, and we know that

$$\lim_{x \to \infty} \frac{1}{x} = 0$$

and so it appears that the sequence converges to 0 as well. In other words, the limit of the sequence is 0. The trend in the function values $\frac{1}{x}$, as $x \to \infty$, is 0 for real values of $x$, and so the same is true for the trend in sequence values $\frac{1}{n}$, as $n \to \infty$, for natural numbers $n$. See Figure 10.2.

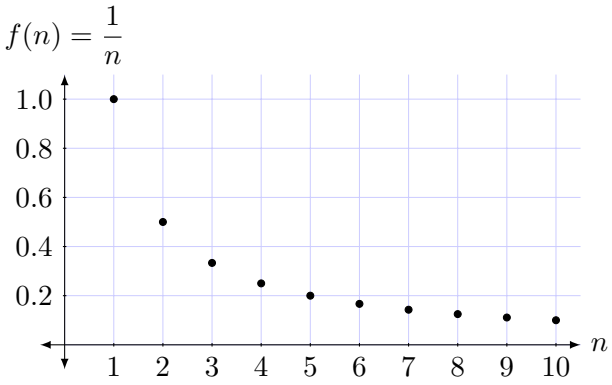



Figure 10.2: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = \frac{1}{n}$. Note that the scale on the vertical axis is not the same as the scale on the horizontal axis. The sequence has a limit of 0; that is, the sequence converges to 0.

A similar situation involves the sequence with terms that satisfy the formula

$$f(n) = 5 + \frac{1}{n}$$

Is it clear that this sequence converges to 5? Write out the first few terms and plot a graph and then carefully examine it. How can you convince yourself that this is true? Which limit will it be helpful for you calculate? See Figure 10.3.

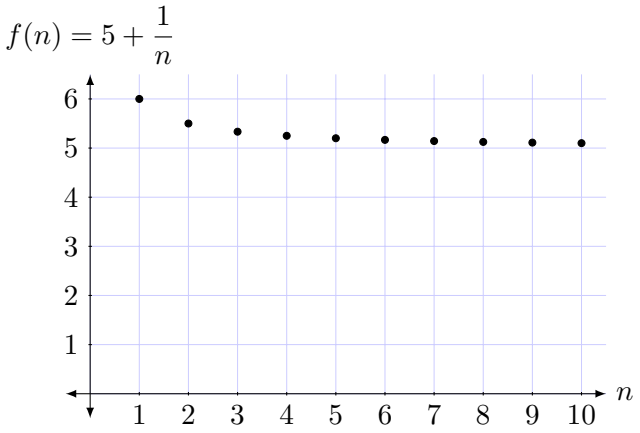



Figure 10.3: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = 5 + \frac{1}{n}$. The sequence has a limit of 5; that is, the sequence converges to 5.

Another sequence along the same lines is

$$f(n) = 3 - \frac{1}{n^2}$$

Is it clear that this sequence converges to 3? Once again, write out the first few terms and plot a graph and then carefully examine it. How can you convince yourself that this is true? Which limit will it be helpful for you calculate? See Figure 10.4.

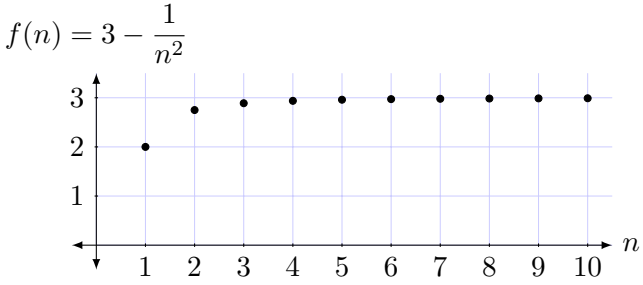



Figure 10.4: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = 3 - \frac{1}{n^2}$. The sequence has a limit of 3; that is, the sequence converges to 3.

After studying a few examples such as the ones in this section, it may become clear to you that sequences that converge correspond to functions that have a horizontal asymptote towards the right, and the $y$-value of the asymptote is the limit of the sequence. On the other hand, sequences that correspond to functions that do not have horizontal asymptotes towards the right do not converge; consider the following examples in Figures 10.5, 10.6, and 10.7.

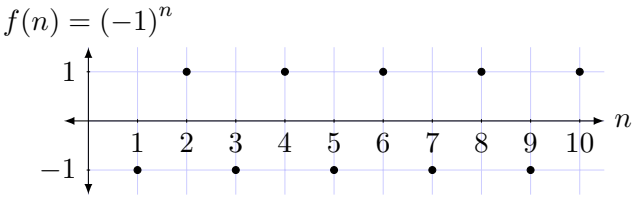



Figure 10.5: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = (-1)^n$. The sequence has no limit; that is, the sequence diverges.

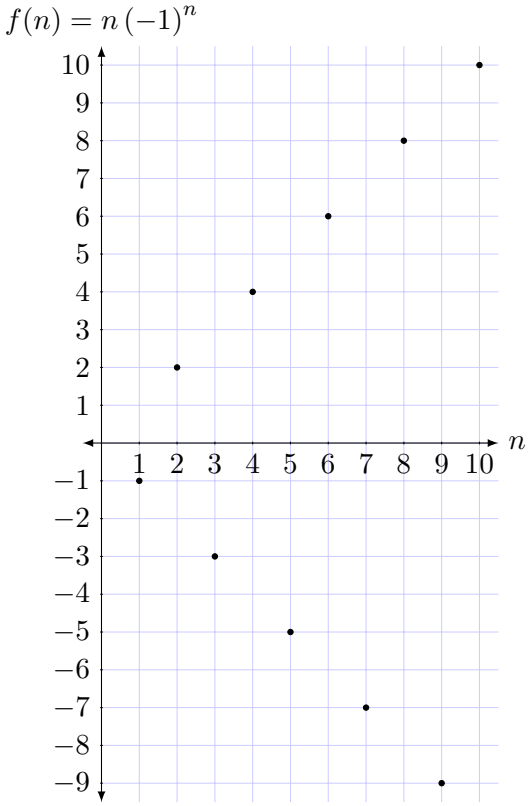



Figure 10.6: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = n\,(-1)^n$. The sequence has no limit; that is, the sequence diverges.

## EXERCISES

(Answers at end.)

Determine whether each sequence converges. For the sequences that converge, determine their limits.

1. $f(n) = 2 + \dfrac{3}{n}$
2. $f(n) = 4 - \dfrac{5}{n}$
3. $f(n) = 3 + \dfrac{(-1)^n}{n}$
4. $f(n) = 2n + 1$
5. $f(n) = 3 - n$
6. $f(n) = 6$

Answers: 1. Converges; limit is 2. 2. Converges; limit is 4. 3. Converges; limit is 3. 4. Diverges. 5. Diverges. 6. Converges; limit is 6.

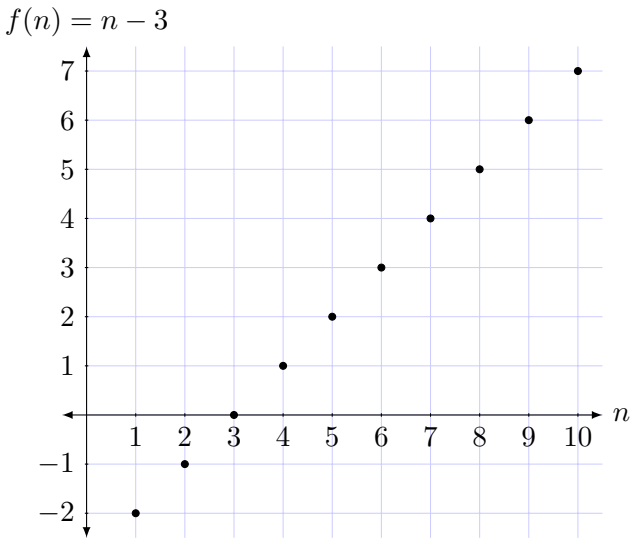



Figure 10.7: The graph illustrates the sequence of numbers that satisfies the formula $f(n) = n - 3$. The sequence has no limit; that is, the sequence diverges.

## 10.2 Series

The main problem in differential calculus is to determine the rate of change of a quantity that changes according to a known functional formula. This amounts to determining the derivative of the function. Our conceptual understanding of differentiation is founded upon the idea of increasingly good approximations; a sequence of approximations. The actual rate of change (i.e., the actual derivative) is the limit of this sequence of approximations, provided that this limit exists.

The other main branch of calculus, integral calculus, deals with a problem that is in a way inverse to the problem of determining a derivative. In integral calculus, one knows the rate of change of a quantity and one wishes to determine the accumulated amount of that quantity over some time interval. One can take the problem of determining the area of a curved plane region as an exemplar of this; the typical strategy is to slice the region up into a number of sub-regions, approximate the area of each sub-region, add the approximations to determine an estimate for the total, and then repeat the process by using more slices that are smaller than before. Once again, one encounters a sequence of approximations, and if this sequence of numbers converges, then we say that the integral exists, and the value of the limit is the sought-after accumulated amount.

It is also possible to think of this process of approximating an accumulated amount as the addition of an infinite number of terms, which is called an infinite series, and so this discussion motivates the study of infinite series.

An example of an infinite series is

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

But what does such a thing even mean? We know what it means to add two or three numbers together, or any finite number of terms, but what does it mean to add an infinite number of terms? We can make sense of this by thinking in terms of a sequence that is related to the series, which is called the **sequence of partial sums**, and which is constructed as follows. The first term of the sequence is just the first term of the series. The second term of the sequence is the sum of the first two terms of the series. The third term of the sequence is the sum of the first three terms of the

series, and so on. For our example infinite series, the sequence of partial sums is:

$$S_1 = 1$$
$$S_2 = 1 + \frac{1}{2}$$
$$S_3 = 1 + \frac{1}{2} + \frac{1}{4}$$
$$S_4 = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8}$$
$$S_5 = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16}$$

and so on. We make sense of the infinite series by saying that the infinite series converges (in other words, the infinite series has a sum) provided that the sequence of its partial sums converges. If the sequence of partial sums does converge, then we say that the sum of the infinite series is the limit of the sequence of its partial sums. On the other hand, if the sequence of partial sums for an infinite series diverges, then we say that the infinite series also diverges.

**DEFINITION 9**

**Convergence of an infinite series**

Consider an infinite series that has $n$-th term $f(n)$. The sequence of partial sums for this series has $n$-th term $S_n$ that satisfies $S_n = f(1) + f(2) + \cdots + f(n)$. The infinite series is said to converge and has sum $L$ provided that

$$\lim_{n\to\infty} S_n = L$$

If the limit does not exist then we say that the infinite series diverges.

**CAREFUL!**

**The logical structure of mathematics**

Mathematicians are fond of defining a new concept in terms of a concept already studied. In this way, a logical structure is constructed that is unambiguous and comprehensible. An example of this is defining what it means for an infinite series to converge in terms of the previously defined concept of the convergence of an infinite sequence.

One potential problem here, which happens over and over again, is the use of a single word (convergence) to mean two different (although related) concepts. Why was this done originally? Was it because of a certain confusion, where it was not noticed that there were two distinct concepts being described by the same word? Possibly. And it is also possible that the two concepts had different names originally, but that through laziness both concepts came to be referred to using the same word. We could be charitable and say that "laziness" is rather economy of language, and that it is actually helpful to label two closely related terms with the same name. This may actually be true, *once you understand both concepts and how they are related*. But for a student, who is learning these concepts for the first time, this is a potential source of confusion and frustration, and so it is worth pausing over this point, and making careful note that there are *two* concepts here, labelled with the same name, distinct though closely related, and to remind oneself frequently about exactly how they are related.

In any case, we are stuck with this situation, because it is almost impossible to alter long-held conventions, especially naming conventions. This occurs in other cases in mathematics and science, partly because there are an unlimited number of concepts, and a limited number of useful words that are therefore preferable. As you learn more, be alert for such situations, take your time over them, and be explicit about which concepts are involved and how they are related, if they are related. (An example is the term linear, which means several different things in different contexts, some of which are related, and some which are not.)

**LEVITY**

**Mathematicians and physicists**

The culture of mathematicians is significantly different from the culture of scientists, and this difference has spawned a whole genre of jokes highlighting this difference. Mathematicians tend to be deeply interested in structural matters, whereas physicists tend to be more interested in solving specific problems. These are generalizations,[a] and there are numerous exceptions, and counterexamples, but there is an element of truth to this. For example, mathematicians frequently argue for the truth of some proposition or other by showing that it is equivalent to some fact that is already known to be true. The following joke makes this point.

A mathematician and a physicist are given identical problems. They are each presented with an empty pot, a source of water, and a stove, and the problem is to boil water. Each successfully solves the problem in the same way, by filling the pot with water, then placing it on the stove, and turning the stove on.

Then they are each given a follow-up problem. They are in the same situation, with the same task, but this time the pot is already full of water. The physicist places the pot on the stove, and turns the stove on. The mathematician dumps the water out of the pot, thus reducing the problem to one that has already been solved.

There is a whole genre of such jokes, which also include comics, that are worth searching for to get a better understanding of the cultural differences between mathematicians and other scientists.

[a] And, as we know, all generalizations are wrong.

The subject of how to determine whether an infinite series converges or diverges is a large one and could fill an entire book on its own, but we will restrict ourselves to a small slice of this subject. Let's go back to the infinite series that has $n$-th term equal to $\left(\frac{1}{2}\right)^{n-1}$:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

Does this infinite series converge? It is helpful in this case to interpret the sequence of partial sums graphically. Imagine walking along a number line (maybe hopping is better) from left to right, starting at 0. The first step takes you 1 unit to the right. The second step takes you an additional $\frac{1}{2}$ unit to the right. The third step takes you an additional $\frac{1}{4}$ unit to the right, and so on. What is the trend of your journey? You clearly continue to move to the right in each step (all of the terms of the series are positive), but does your journey zero in on some location (which would mean the series converges), or do you surpass all numbers on your rightward journey (which would mean the series diverges)?

It might be helpful at this point to use your calculator to compile a table of partial sums in decimal form, to help us get a feel for whether the series might converge. Compare your results to mine:

$$S_1 = 1$$
$$S_2 = 1.5$$
$$S_3 = 1.75$$
$$S_4 = 1.875$$
$$S_5 = 1.9375$$
$$S_6 = 1.96875$$

Feel free to continue the table further if you wish. What do you think? Does the series converge? If so, what is your best guess for the sum of the series?

You might have guessed that the sum of the infinite series is 2. You might convince yourself that the series does indeed converge to 2 with the following geometric argument. You start the journey at 0 on the number line, and the first step in your journey takes you to the number 1, which is half-way to 2. The next step in your journey takes you half of the remaining distance to 2, to the position 1.5 on the number line. The next step again takes you half of the remaining distance to 2, and it's the same for each subsequent step in the journey.

Does this convince you that the sum of the series is 2? After all, each step takes you inexorably to the right, but because each step takes you half of the remaining distance to 2, it seems clear that you will never surpass 2, and so 2 deserves to be called the limit.

But wait; what if there is some smaller number, say 1.999, that you never surpass in your rightward journey? Then shouldn't this smaller number be considered the limit? Yes, certainly, if there were such a number, but there is no such number less than 2 that you never surpass in your rightward journey.

We can see that this is so with the following argument. The distance between 1.999 and 2 is 0.001. In your rightward journey, the distance between your current location and 2 after $n$ steps is

$$\frac{1}{2^{n-1}}$$

(Confirm this!) Once this distance becomes less than 0.001 you will surpass the location 1.999 on the number line on your rightward journey. This happens when $n$ satisfies

$$\begin{aligned} \frac{1}{2^{n-1}} &< 0.001 \\ 1 &< 0.001\left(2^{n-1}\right) \\ 1000 &< 2^{n-1} \\ 2^{n-1} &> 1000 \end{aligned}$$

You can use logarithms here, or play with your calculator, to determine that

$$2^{10} = 1024$$

and so we can conclude that you pass 1.999 in your rightward journey at step $n = 11$.

The same argument can be repeated for any other number smaller than 2. Thus, we can conclude that in your rightward journey, no matter which number smaller than 2 is chosen, you will surpass it with enough steps. Although you will never reach 2 in a finite number of steps, it is

reasonable to conclude that 2 is the limit of the sequence of partial sums for this series, and so the series converges to 2.

You can verify that a formula for the $n$-th partial sum is:

$$S_n = 2 - \frac{1}{2^{n-1}}$$

If you plot a graph of the function

$$y = 2 - \frac{1}{2^{x-1}}$$

you will note that the graph has a horizontal asymptote at $y = 2$, which is consistent with our assertion that the infinite series converges to 2. But maybe this is not convincing enough for you; we shall take up this question again later in this section after we have looked at a few other examples.

Let's look at another interesting infinite series:

$$1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \cdots$$

Note the alternating positive and negative signs. In graphical terms, this corresponds to a zig-zag journey along the number line, first to the right, then to the left, and continuing to alternate indefinitely. But the sizes of each step decrease, so perhaps the series converges? If so, to what number does the series converge? Again, it may be helpful to tabulate the first few partial sums for this infinite series:

$$S_1 = 1$$
$$S_2 = 0.5$$
$$S_3 = 0.75$$
$$S_4 = 0.625$$
$$S_5 = 0.6875$$
$$S_6 = 0.65625$$

The steps are shown graphically in Figure 10.8; note that the "journey" takes place along the number line, but the steps are offset for clarity.

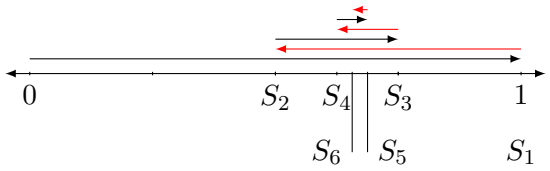



Figure 10.8: The graph illustrates the "journey" corresponding to the alternating infinite series $1 - \frac{1}{2} + \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \cdots$. The steps in the journey take place along the horizontal axis, but are offset for clarity. Does the series converge? See the text for a full explanation.

The standard argument that this series converges is as follows. The first step in your journey along the number line is 1 unit to the right. The second step is 1/2 unit to the left. The third step is to the right again, but it is shorter in size than the first step. Thus, the third step does not take you as far to the right as the first step did. The fourth step is to the left, but it is shorter than the previous steps, and the fifth step is to the right again, but it is shorter than the fourth step. Thus, the fifth step does not take you as far to the right as the third step.

The same argument can be repeated for the steps to the left to conclude that the fourth step does not take you as far to the left as the second step, the sixth step does not take you as far to the left as the fourth step, and so on.

Can you see from this argument that the sum of the series, if it converges, must be between $S_1$ and $S_2$? But if you can understand this, then you will be able to understand in the same way that the sum of the series, if it exists, must also be between $S_2$ and $S_3$. The same reasoning shows that the sum of the series is between every two consecutive terms in the sequence of partial sums. But the distance between consecutive terms in the sequence of partial sums decreases as $n$ increases, and so this is strong evidence that the series does in fact converge. From the results tabulated, we can be sure that the sum of the series is between 0.65625 and 0.6875, and we can improve this estimate to an arbitrary accuracy by simply continuing to calculate more partial sums.

This is not an ironclad proof that the infinite series converges, but this is all very suggestive. An ironclad proof involves showing that the limit of the sequence of partial sums exists, which requires a formula for the $n$-th term of this sequence. We shall construct a proof later in this section, which will allow us to calculate the exact sum of this infinite series. You might like to think about this and try to come up with the sum yourself before you reach that point in your reading.

It would be good to also study some examples of infinite series that do not converge. When mathematicians first began to study infinite series intensively, there was naturally a fair bit of sloppy reasoning, because precise definitions were only proposed later, once a sufficient amount of playing around with examples was done. Consider the following infinite series:

$$1 + (-1) + 1 + (-1) + 1 + (-1) + \cdots$$

A formula for the $n$-th term of the series is $(-1)^{n+1}$. The sequence of partial sums for this series is:

$$S_1 = 1$$
$$S_2 = 0$$
$$S_3 = 1$$
$$S_4 = 0$$
$$S_5 = 1$$
$$S_6 = 0$$

Is it clear that this sequence does not converge? In terms of your journey along the number line starting at 0, your steps take you back and forth between 0 and 1, back and forth, back and forth, without end. There is no single number to which your journey zeros in. The corresponding function (which could be a certain sine function) has no horizontal asymptote, and so the series does not converge.

Now consider the following infinite series:

$$1 + 1 + 1 + 1 + 1 + 1 + \cdots$$

This one doesn't converge either. The sequence of partial sums is $1, 2, 3, 4, 5, \ldots$. Imagining the series physically as a journey on a number line, you begin at 0, take a step of 1 unit to the right, then take another step of one unit to the right, then another, and continue indefinitely. Is it clear that your journey is unbounded? That is, no matter which positive number you state, with enough steps you will eventually surpass that number. What about 17.3? Yes, you'll pass that number with your 18th step. What about 143.29? Yes, you'll pass that number with your 144th step. Again, the corresponding function (which could be $y = x$) has no horizontal asymptote, and so the series does not converge.

**EXERCISES**

(Answers at end.)

1. Construct a series that represents a journey along the number line that starts at 3 and in each step moves 2/3 of the remaining distance to 8.
   (a) Write the first few terms of the series.
   (b) Determine a formula for the sum of the first $n$ terms of the series.
2. Construct a series that represents a journey along the number line that starts at 5 and in each step moves 3/4 of the remaining distance to 9.
   (a) Write the first few terms of the series.
   (b) Determine a formula for the sum of the first $n$ terms of the series.
3. Construct a series that represents a journey along the number line that starts at 4 and in each step moves 1/5 of the remaining distance to 6.
   (a) Write the first few terms of the series.
   (b) Determine a formula for the sum of the first $n$ terms of the series.

---

Answers: 1. (a) $3+\frac{10}{3}+\frac{10}{9}+\frac{10}{27}+\cdots$; (b) $S_n = 8-5\left(\frac{1}{3}\right)^{n-1}$

2. (a) $5+3+\frac{3}{4}+\frac{3}{16}+\cdots$; (b) $S_n = 9-4\left(\frac{1}{4}\right)^{n-1}$

3. (a) $4+\frac{2}{5}+\frac{8}{25}+\frac{32}{125}+\cdots$; (b) $S_n = 6-2\left(\frac{4}{5}\right)^{n-1}$

### 10.2.1 Finite Geometric Series

The subject of sequences and series is a large one. Although numerous techniques for determining whether a series converges have been developed, formulas for the sums of convergent infinite series are known only for a very small number of series. It will serve our purposes to discuss a few series of special type that do have known sum formulas, and leave the rest for later study in your calculus courses. One such important type of series is called a geometric series.

An example of a finite geometric series is

$$1+2+4+8+16$$

Each term after the first one can be obtained from the previous term by multiplying by the same number, which is called the common ratio of the geometric series. In the example above, the common ratio is 2. A general finite geometric series that has $n$ terms is of the form

$$a+ar+ar^2+ar^3+\cdots+ar^{n-1}$$

where $r$ is the common ratio. An infinite geometric series is of the form

$$a+ar+ar^2+ar^3+\cdots$$

**DEFINITION 10**

**Geometric series**

A geometric series having $n$ terms is of the form

$$a + ar + ar^2 + ar^3 + \cdots + ar^{n-1}$$

where $a$ is the first term and $r$ is the common ratio. An infinite geometric series with first term $a$ and common ratio $r$ is of the form

$$a + ar + ar^2 + ar^3 + \cdots$$

Another example of a finite geometric series is

$$3 + 6 + 12 + 24 + \cdots + 768 + 1536 + 3072$$

In this series, the first term is 3 and the common ratio is 2.

If a finite geometric series has only a small number of terms, then it is straightforward to add the terms one-by-one to determine the sum of the series. However, if the number of terms is large, then adding the terms one-by-one is tedious. Fortunately, there is a nice trick for determining the sum that will save a lot of time. First, label the sum of the series by a letter, such as $S$:

$$S = 3 + 6 + 12 + 24 + \cdots + 768 + 1536 + 3072$$

Next, subtract the first term from each side of the previous equation, and then factor the right side of the resulting equation by the common ratio:

$$\begin{aligned} S - 3 &= 6 + 12 + 24 + \cdots + 768 + 1536 + 3072 \\ S - 3 &= 2\,(3 + 6 + 12 + \cdots + 384 + 768 + 1536) \end{aligned}$$

Now notice that the quantity in parentheses on the right side of the previous equation is almost the original geometric series; only the final term is missing. That is, the quantity in parentheses is $S - 3072$. Therefore, we can replace the quantity in parentheses by $S - 3072$ to obtain

$$S - 3 = 2\,(S - 3072)$$

We are left with one linear equation in one unknown, which we can solve for $S$ to obtain the sum of the series:

$$\begin{aligned} S - 3 &= 2\,(S - 3072) \\ S - 3 &= 2S - 6144 \\ 6144 - 3 &= 2S - S \\ S &= 6141 \end{aligned}$$

Thus, the sum of the series $3 + 6 + 12 + 24 + \cdots + 768 + 1536 + 3072$ is 6141. You can verify this result by listing all of the terms of the series and adding them.

Using the same trick, we can obtain a formula for the sum of a general geometric series:

$$\begin{aligned}
S &= a + ar + ar^2 + ar^3 + \cdots + ar^{n-3} + ar^{n-2} + ar^{n-1} \\
S - a &= ar + ar^2 + ar^3 + \cdots + ar^{n-3} + ar^{n-2} + ar^{n-1} \\
S - a &= r\left(a + ar + ar^2 + \cdots + ar^{n-4} + ar^{n-3} + ar^{n-2}\right) \\
S - a &= r\left(S - ar^{n-1}\right) \\
S - a &= rS - ar^n \\
S - rS &= a - ar^n \\
S(1 - r) &= a(1 - r^n) \\
S &= a\left(\frac{1 - r^n}{1 - r}\right)
\end{aligned}$$

Note that the formula makes no sense if $r = 1$, as both the numerator and denominator of the formula are 0. It will be good to carefully examine the derivation of the formula to see where the derivation breaks down for $r = 1$. The line $S - a = rS - ar^n$ amounts to $S - a = S - a$ when $r = 1$, which is certainly correct, and the next step in the derivation is also a correct equation, $S - S = a - a$, but the latter amounts to $0 = 0$, which is not informative. In the case that $r = 1$, the original series is $S = a + a + a + \cdots + a$, where there are $n$ terms on the right of the equation, and so it is straightforward to conclude that if $r = 1$, then $S = na$.

The derivation is pretty attractive, isn't it? For a geometric series, one goes from each term to the next by multiplying by the same common factor $r$, so factoring $r$ in the derivation is a good thing to try. But the same idea is unlikely to work for series that are not geometric, right? But seeing this connection might encourage us to look for similar tricks that might work for other series that have some particular properties.

## KEY CONCEPT

**Sum of a finite geometric series**

For a geometric series

$$a + ar + ar^2 + ar^3 + \cdots + ar^{n-1}$$

with $n$ terms, and for which $r \neq 1$, a formula for the sum $S_n$ is

$$S_n = a\left(\frac{1 - r^n}{1 - r}\right)$$

If $r = 1$, then $S_n = na$.

## GOOD QUESTION

**Formula for the sum of a finite geometric series if $r$ is negative**

Is the formula for the sum of a geometric series also valid if $r$ is negative? Carefully go over the derivation of the formula and determine whether any of the steps are invalid if $r$ is negative. Then check the formula for some geometric series that have a small number of terms. What can you conclude?

**TRICKS OF THE TRADE**

**Should you memorize the formula for the sum of a finite geometric series?**

In the formula for the sum of a finite geometric series, we use the symbol $S_n$ for the sum instead of $S$, as a reminder that the formula applies to a series with $n$ terms, not for a series the last term of which is of the form $ar^n$. Some books quote the formula for a geometric series the last term of which is $ar^n$, in which case the formula contains $r^{n+1}$ instead of $r^n$. The appearance of both forms of the formula in various books makes it a little harder to memorize the formula, at least if you have seen both forms!

But should you memorize this formula? My strategy as a student (and I still adopt it) is to *memorize the minimum necessary formulas*. Some formulas are good to memorize, because you can derive a lot of other formulas from them. Others can be quickly derived, and so it's better to practice deriving them consistently for a while, which is a lot better way to remember them than by rote memorization.

Our human brains are a lot better at remembering meaningful things than meaningless ones, so strive to make what you remember meaningful. Our brains are also highly visual (the visual cortex of the human brain is really large), so our brains are really good at remembering visual information. I have found in my own experience that we remember information in the form of a diagram much better than information in the form of a formula, and that we remember ideas, procedures, and processes much better than formulas alone. Therefore, I have always preferred to illustrate a formula or an idea with a diagram, and I find that I remember this much better than the formula alone.

So, no, I do not have the formula for the sum of a finite geometric series memorized. I remember most of the formula, but I lack confidence about whether the exponent of $r$ in the numerator is $n$ or $n+1$. Thus, I work out the formula from scratch every time, which takes only a few seconds because I have practiced it sufficiently often. The same holds for other formulas; for example, the Newton-Raphson formula for the approximation of the root of an equation is also not in my memory, but the idea behind the method is firmly burned into my long-term memory (as is the associated diagram), so it is easy enough to derive the formula very quickly.

As your mathematical understanding grows, your need to memorize new things will decrease, because you will see more and more connections between ideas that were previously considered separate.

Let's test the formula for the sum of a finite geometric series by using it to verify that the sum of the geometric series given earlier is 6141. The first term is $a = 3$ and the common ratio is $r = 2$. We need to determine $r^n$, but we don't know the number of terms $n$. We can either list all of the terms and count them, or we can use the following trick: Note that the last term of the series is 3072, so

$$\begin{aligned} ar^{n-1} &= 3072 \\ ar^n &= 3072r \\ r^n &= \frac{3072r}{a} \\ r^n &= \frac{3072(2)}{3} \\ r^n &= 2048 \end{aligned}$$

Thus, the sum of the series is

$$S = a\left(\frac{1-r^n}{1-r}\right)$$
$$S = 3\left(\frac{1-2048}{1-2}\right)$$
$$S = 3\left(\frac{-2047}{-1}\right)$$
$$S = 3 \times 2047$$
$$S = 6141$$

The result is the same as was obtained before.

**EXERCISES**

(Answers at end.)

Determine the sum of each geometric series.

1. $2 + 8 + 32 + \cdots + 131072$
2. $20 + 2 + 0.2 + \cdots + 0.00000002$
3. $400 + 200 + 100 + \cdots + 0.390625$
4. $5 + \frac{20}{3} + \frac{80}{9} + \cdots + \frac{81920}{2187}$
5. $4 + (-8) + 16 + \cdots + 65536$
6. $4 + (-8) + 16 + \cdots - 131072$
7. $1 + (-1) + 1 + \cdots + (-1)$
8. $1 + (-1) + 1 + \cdots + 1$

Answers: 1. 174762 2. 22.22222222 3. 799.609375 4. $\frac{294875}{2187} \approx 134.8308$ 5. 43692 6. $-87380$ 7. 0 8. 1

There are common financial applications of finite geometric series involving compound interest, such as annuities, for example. You can also apply finite geometric series to any situations in which a quantity increases by a fixed ratio in each time period, or decreases by a fixed ratio in each time period. Such increases or decreases may begin modestly, but can increase significantly over time, and it is difficult for us humans to grasp the magnitude of the eventual changes over long periods of time, human nature being what it is.

You may have heard the ancient parable about an emperor who granted a wish to a subject who did him a great favour. The subject is said to have replied that he wished a very modest quantity of grain; just 1 grain of wheat for the first square of the chessboard, 2 grains of wheat for the second square, 4 grains for the third square, and so on, with double the amount of grain for each subsequent square. This may seem like a very modest request; certainly the emperor considered it very reasonable in the telling of this tale, and so he immediately granted the request. But once the emperor's accountant was called in, he quickly explained that the request is absolutely unreasonable. The total number of grains of wheat requested is

$$1 + 2 + 4 + \cdots + 2^{63}$$

This is a geometric series that has first term $a = 1$, common ratio $r = 2$, and $n = 64$ terms, so we

can apply the sum formula to obtain the total number $S$ of grains:

$$S = 1 + 2 + 4 + \cdots + 2^{63}$$
$$S = a\left(\frac{1 - r^n}{1 - r}\right)$$
$$S = (1)\left(\frac{1 - 2^{64}}{1 - 2}\right)$$
$$S = 2^{64} - 1$$
$$S \approx 1.84 \times 10^{19}$$

It may be difficult to realize just how stupendous this quantity is, so let's convert it into a more modern unit. The mass of a grain of wheat is about 40 mg, and so the total mass of wheat that the emperor committed to giving his subject is approximately

$$\text{mass} = 1.84 \times 10^{19} \times 40 \text{ mg}$$
$$\text{mass} = 1.84 \times 10^{19} \times 40 \times 10^{-6} \text{ kg}$$
$$\text{mass} = 1.84 \times 10^{19} \times 40 \times 10^{-9} \text{ tonnes}$$
$$\text{mass} = 7.36 \times 10^{11} \text{ tonnes}$$

The entire world production of wheat in the year 2020 was about 750 million tonnes, so the subject in the story was requesting nearly 1000 years of the entire world's *recent* wheat production! (The wheat production at the time of the story was, of course, drastically lower.) And yet the request seemed quite reasonable at first. This shows us the power of growth at a constant ratio.

This charming story may or may not be true, but it certainly makes an important point in a memorable way. By absorbing the point of the story we can make informed decisions about our own consumption. In particular, for quantities that grow at a constant rate per unit time (doubling per unit time is an example of this kind of growth), the eventual value of the quantity will become extremely large in a sufficient amount of time.

If we think about quantities that double at each step (where each step could be a number of years, for instance), then the geometric series from the parable indicates that the value of the quantity at each step is one greater than the sum of the entire series up to that point! That is, for the series

$$1 + 2 + 4 + 8 + 16 + \cdots$$

the value of the third term is one greater than the sum of the previous terms, the value of the fourth term is one greater than the sum of the previous terms, the value of the fifth term is one greater than the sum of all of the previous terms, and so on. Once you really start to absorb this point, it is a little terrifying, isn't it? If the consumption of any vital resource (fresh water, soil, oil, etc.) doubles per unit time period, then eventually a crisis is reached. Furthermore, if at a crisis point some discovery or technological advance results in the reserves of the quantity *doubling*, then the crisis is postponed only for a single doubling period.

Understanding growth at constant ratios, and the geometric series that model this kind of growth, will be helpful in deciding on appropriate public policy.

### 10.2.2 Infinite Geometric Series

Interestingly, the same trick used to determine the sum of a finite geometric series in the previous section also works for infinite geometric series, although justifying it requires some work. Consider

the infinite series that we studied earlier in this chapter:

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

Note that this is an example of a geometric series, and that the common ratio is $r = \frac{1}{2}$. To determine the sum of the series, start as before by labelling the sum by $S$:

$$S = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

Next, subtract the first term from both sides of the equation, and then factor the common ratio from the right side:

$$\begin{aligned} S - 1 &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots \\ S - 1 &= \frac{1}{2}\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots\right) \end{aligned}$$

Now notice that the expression in parentheses is just the original series, so we can replace the expression within the parentheses by $S$, and then solve the resulting equation for $S$:

$$\begin{aligned} S - 1 &= \frac{1}{2}(S) \\ 2S - 2 &= S \\ 2S - S &= 2 \\ S &= 2 \end{aligned}$$

The result is 2, which is the same result we obtained earlier with longer reasoning! This is very satisfying, isn't it? And the great thing is that this latest argument is absolutely rigorous, right? Well, hold on a second; no, it's not rigorous after all. The line of the argument where we factored the right hand side requires considerable justification; we are allowed to use the usual rules of algebra on finite series, but it turns out that the usual rules of algebra do not work for all infinite series! Deciding which rules work under which circumstances will require a considerable amount of work, and is beyond the scope of this book, so we shall leave such studies to your future calculus courses. But it is important that you are aware of the weak point of this argument.

An amusing illustration of a situation where the previous argument does not work is the following series:

$$1 + 2 + 4 + 8 + 16 + \cdots$$

This is a geometric series with common ratio $r = 2$. Is it clear to you that the series diverges? In this case it makes no sense for this series to have a sum. But what happens if we apply the same trick as above to try to determine the sum of the series anyway, knowing that it has no sum? Let's see:

$$\begin{aligned} S &= 1 + 2 + 4 + 8 + 16 + \cdots \\ S - 1 &= 2 + 4 + 8 + 16 + \cdots \\ S - 1 &= 2(1 + 2 + 4 + 8 + 16 + \cdots) \\ S - 1 &= 2S \\ S - 2S &= 1 \\ S &= -1 \end{aligned}$$

Now this is some high-grade nonsense! Are you telling me that you take a journey starting at 0 on the number line, take every step towards the right, with each step being bigger than the previous one, and you somehow end up to the left of your starting point? Are you telling me that you add up an infinite number of positive numbers and somehow the sum comes out negative? None of this makes any sense, and is a good illustration that this new trick of ours does not always work.

If you are left feeling a little bit uneasy about this, I don't blame you. Mathematics is supposed to be precise and its methods are supposed to be sure, so that we have no doubt when something works and when it doesn't. This is the same situation that early workers in infinite series found themselves in; they were in the process of doing a considerable amount of play, but nobody was quite certain about what worked for sure under which circumstances. It was only later, with the introduction of clearly defined concepts that further research could be done to pin all of these uncertainties down. This is the kind of work you will do in some of your future mathematics courses.

We can get a sense for when our trick for summing an infinite geometric series works by applying the trick to a general infinite geometric series:

$$\begin{aligned} S &= a + ar + ar^2 + ar^3 + \cdots \\ S - a &= ar + ar^2 + ar^3 + \cdots \\ S - a &= r\left(a + ar + ar^2 + \cdots\right) \\ S - a &= rS \\ S - rS &= a \\ S\left(1 - r\right) &= a \\ S &= \frac{a}{1-r} \end{aligned}$$

This is a useful formula, and worth remembering, but by itself it gives us no sense for when the formula is valid. However, recall that an infinite series converges if the sequence of its partial sums converges. We have earlier developed a formula for the sum of a finite geometric series, and that will be helpful now:

$$S_n = \frac{a\left(1 - r^n\right)}{1 - r}$$

Expanding this formula into two terms is revealing:

$$S_n = \frac{a}{1-r} - a\frac{r^n}{1-r}$$

The infinite series converges provided that the sequence of its partial sums has a limit:

$$\begin{aligned} \lim_{n\to\infty} S_n &= \lim_{n\to\infty}\left[\frac{a}{1-r} - a\frac{r^n}{1-r}\right] \\ \lim_{n\to\infty} S_n &= \lim_{n\to\infty}\left[\frac{a}{1-r}\right] - \lim_{n\to\infty}\left[a\frac{r^n}{1-r}\right] \\ \lim_{n\to\infty} S_n &= \frac{a}{1-r} - \frac{a}{1-r}\left[\lim_{n\to\infty} r^n\right] \end{aligned}$$

Under which circumstances does the limit on the previous line exist?[1] To answer the question requires an argument involving limits, which can be made formally, and which I'll paraphrase as

[1]Note that the first term is a constant, so does not change as we take the limit as $n \to \infty$, so we only need to carefully examine the limit in the second term.

follows. First note that if $a = 0$, the series definitely converges to 0, because the journey stays at 0. Thus, we can assume that $a \neq 0$ for the rest of the discussion. Next, notice that if $r = 1$ the series diverges, because each step in the journey on the number line has the same size, and so the journey surpasses every number (in the positive direction if $a > 0$, and in the negative direction if $a < 0$). If $r > 1$, the same conclusion applies, as now the steps are actually getting larger in size with each succeeding step. Next, if $r = -1$, the journey bounces back and forth between two numbers, and so does not "zero in" on any particular number, and so by the technical definition of convergence, the series diverges.[2] If $r < -1$, then the series also does not converge, because it bounces back and forth, but more and more wildly with each succeeding step, because the size of each step gets bigger and bigger each time.

So far, then, we conclude that the series does not converge unless $-1 < r < 1$. Next, we need to examine the interval $-1 < r < 1$, and we will be able to conclude that the series does indeed converge for values of the common ratio in this interval, according to the following argument.

We have considerable knowledge of the function $y = x^n$, a power function, and we only need recall what we learned about such functions in high school. If you need to refresh your memory, plotting this graph for various values of $n$ using your favourite software will be good, and be sure to carefully note how the graph changes as $n$ increases. You will conclude that as $n$ increases, the value of $r^n$ also increases without bound if $r > 1$ and also if $r < -1$. On the other hand, if $-1 < r < 1$, then as $n$ increases, $r^n$ decreases towards 0. If $r = 1$, the entire formula (for the sum of the series) makes no sense, because division by 0 makes no sense. If $r = -1$, then $r^n$ oscillates from $-1$ to 1 as $n$ changes, and so the limit does not exist.

Thus, if you stop summing the series at the $n$-th term, you get pretty close to the ultimate sum of the infinite series, which is $a/(1-r)$, but the difference is the final term on the right side of the following equation.

$$S_n = \frac{a}{1-r} - \left(\frac{a}{1-r}\right) r^n$$

Now you can set up a formal limit argument to show that the final term on the right side of the previous equation gets closer and closer to zero as the value of $n$ increases, provided that $-1 < r < 1$. Play with this using a calculator if this is unclear; for example, enter 0.83 and square it repeatedly and watch what happens to the numbers in the display.

In conclusion, the limit of the second term on the right side of the previous displayed equation exists and equals 0 if and only if $-1 < r < 1$. Thus, the formula developed earlier (using a trick) for the sum of an infinite geometric series is valid if and only if $-1 < r < 1$. The argument in terms of the limit of the sequence of partial sums is rigorous, or at least is a model for a rigorous argument, if you think the previous paragraph was not sufficiently detailed.

Now we have a clear guideline for when the formula for the sum of an infinite geometric series makes sense: Only when the absolute value of the common ratio is less than 1. The infinite geometric series that we studied earlier, $1 + 2 + 4 + 8 + 16 + \cdots$, where an application of the sum formula resulted in nonsense, has a common ratio of 2, and this is outside the range for which the formula is valid.

[2]Interestingly, the great Euler himself considered that the series represents the "expression" from which it is derived, which he took to be the expression for the sum that you derived earlier; that is, $a/(1-r)$. Thus, Euler figured that $1-1+1-1+1-1+\cdots = 1/2$. Nowadays we wouldn't agree with this, because according to the current definition of convergence, we recognize that this series does not converge.

## KEY CONCEPT

### Sum of an infinite geometric series

An infinite geometric series

$$a + ar + ar^2 + ar^3 + \cdots$$

converges provided that $-1 < r < 1$. A formula for the sum $S$ of a convergent infinite geometric series is

$$S = \frac{a}{1-r}$$

## EXERCISES

(Answers at end.)

Determine whether each infinite geometric series converges. Determine the sum of each series that converges.

1. $2 + \frac{4}{3} + \frac{8}{9} \cdots$
2. $20 + 2 + 0.2 + \cdots$
3. $400 + 200 + 100 + \cdots$
4. $5 + \frac{20}{3} + \frac{80}{9} + \cdots$
5. $4 - \frac{8}{5} + \frac{16}{25} - \cdots$
6. $4 + \frac{8}{5} + \frac{16}{25} + \cdots$
7. $0.3 + 0.03 + 0.003 + \cdots$
8. $0.3 - 0.03 + 0.003 - \cdots$

---

Answers: 1. Converges to 6. 2. Converges to $\frac{200}{9}$. 3. Converges to 800. 4. Diverges.
5. Converges to $\frac{20}{7}$. 6. Converges to $\frac{20}{3}$. 7. Converges to $\frac{1}{3}$. 8. Converges to $\frac{3}{11}$.

## MAKING CONNECTIONS

### Archimedes and infinite series

In a feature box on Archimedes back in Chapter 3, it was mentioned that Archimedes showed that the area enclosed by a parabola and a straight line is $\frac{4}{3}$ of the area of a certain inscribed triangle, again using the method of exhaustion. (Did you look up how Archimedes did this? If not, you might like to do this now!) To complete this demonstration Archimedes determined the sum of the following convergent infinite series:

$$1 + \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots$$

Now that we have discussed a bit about infinite series, are you able to determine the sum of the series in the previous line? What kind of series is it? Does the series converges? How do you know?

**MAKING CONNECTIONS**

**A funny thing about a divergent geometric series**

Earlier in this section we applied the formula for the sum of an infinite geometric series to the divergent series

$$1 + 2 + 4 + 8 + 16 + \cdots$$

and obtained the incorrect result that the sum is equal to $-1$. This result is patent nonsense, as the series clearly diverges and therefore has no sum. And yet, although the result is clearly incorrect, it is *meaningfully* incorrect!

To extract some meaning from this incorrect result, imagine that the series represents a journey along the number line. We have typically imagined such journeys starting at 0, so in this case the first step would be from 0 to 1, the next step would be from 1 to 3, the next step from 3 to 7, and so on. However, instead of assuming that the journey starts at 0, imagine that there were earlier steps that satisfied the same pattern. That is, the step previous to the first one had size $1/2$, and therefore went from $-1/2$ to 0, the step before that one had size $1/4$, and therefore went from $-3/4$ to $-1/2$, the step before that one had size $1/8$, and therefore went from $-7/8$ to $-3/4$, and so on. Isn't this charming? If we imagine an infinite number of steps "before" the first step starting at 0, and following the same pattern, then it seems that these "past" steps began at $-1$.

So although $-1$ is an incorrect answer to the question, "What is the sum of the series?" it is a correct answer to a different question!

This is a timely reminder not to throw away "obviously" incorrect results without reflecting on them sufficiently. You might be able to extract some insight from them.

I wonder if the same behavior occurs for other divergent geometric series? For the interested person, this will be fun territory to explore! (I learned about this charming interpretation of the nonsensical result of applying the sum formula for an infinite geometric series to a divergent geometric series from Section 45 of the delightful book *Algebra* by I.M. Gelfand and A. Shen. Check out this book for lots of additional enjoyment!)

### 10.2.3 Repeating Decimal Numbers

We have learned about repeating decimal numbers since before high school, and it's worth noting that repeating decimal numbers can be interpreted as infinite geometric series! In fact all decimal numbers, whether repeating or not, can be interpreted as infinite series. Consider the number

$$0.2785396117\ldots$$

Another way to write this number is

$$0.2 + 0.07 + 0.008 + 0.0005 + 0.00003 + 0.000009 + \cdots$$

This expresses the decimal number as an infinite series, although it is not a geometric series, because there is no common ratio. A repeating decimal can be expressed as a geometric series; for example,

$$0.4444444\ldots = 0.4 + 0.04 + 0.004 + 0.0004 + \cdots$$

which is a geometric series with common ratio $r = 0.1$. The value of the common ratio indicates that the series converges; what happens if we apply the formula for the sum of an infinite geometric

series? The result is

$$\begin{aligned}
0.4444444\ldots &= 0.4 + 0.04 + 0.004 + 0.0004 + \cdots \\
0.4444444\ldots &= \frac{a}{1-r} \\
0.4444444\ldots &= \frac{0.4}{1-0.1} \\
0.4444444\ldots &= \frac{0.4}{0.9} \\
0.4444444\ldots &= \frac{4}{9}
\end{aligned}$$

This reminds us of the fact learned in high school that every repeating decimal number is a rational number (i.e., expressible as a fraction where both terms of the fraction are integers). Besides applying the sum formula, we can adapt the trick used earlier to obtain the same result. Start by letting $S$ represent the decimal number, and then multiply both sides of the resulting equation by 10 and simplify:

$$\begin{aligned}
S &= 0.4444444\ldots \\
10S &= 4.444444\ldots
\end{aligned}$$

Subtract the first line above from the second line and notice how the decimal parts of the two decimal numbers cancel (this is the part of the argument that requires justification in a higher-level course, because the two strings are infinitely long, and how do we know that one can cancel infinitely long strings of digits?), to obtain

$$\begin{aligned}
9S &= 4 \\
S &= \frac{4}{9}
\end{aligned}$$

Cute, isn't it? As usual, this trick method suffers from the uncertainty about whether the usual rules of algebra can be applied in a situation beyond their usual domain of applicability. We'll leave such justification to more advanced courses.

More complex repeating decimal numbers can also be expressed as a fraction of two integers, where the denominator is not zero. For example, consider the decimal number

$$2.73737373\ldots$$

which is conventionally written as $2.\overline{73}$ to clarify that the group of digits "73" repeats indefinitely. Applying the same strategy as before, let $S$ represent the decimal number $2.\overline{73}$

$$S = 2.737373737373\ldots = 2.\overline{73}$$

then multiply both sides of the previous relation by 100 to obtain

$$100S = 273.737373737373\ldots = 273.\overline{73}$$

Now subtract $S$ from $100S$ to obtain

$$\begin{aligned}
100S &= 273.737373737373\ldots = 273.\overline{73} \\
S &= \quad 2.737373737373\ldots = \quad 2.\overline{73} \\
\hline
99S &= 271
\end{aligned}$$

Dividing both sides by 99, and verifying that the result is in lowest terms, we obtain

$$S = \frac{271}{99}$$

from which we conclude that the repeating decimal number $2.\overline{73}$ is equivalent to the fraction $\frac{271}{99}$.

We can summarize our study of repeating decimal numbers, and remind ourselves of some facts learned in high school, as follows:

**KEY CONCEPT**

**Characterization of rational numbers**

Each rational number can be expressed as a fraction for which the numerator and denominator are both integers, and the denominator is not zero. Conversely, each such fraction is a rational number.

Equivalently, each rational number can be expressed as either a repeating decimal number or a decimal number that terminates. Conversely, each such decimal number is a rational number.

**EXAMPLE 23**

**Converting repeating decimal numbers to fractions**

Express each repeating decimal number as a fraction.

**(a)** $0.3333\ldots = 0.\overline{3}$ **(b)** $0.142857142857\ldots = 0.\overline{142857}$

**SOLUTION**

**(a)** Let $x$ represent $0.3333\ldots$, and multiply by 10 to obtain $10x = 3.3333\ldots$. Thus,

$$\begin{aligned} 10x &= 3.3333\ldots \\ x &= 0.3333\ldots \end{aligned}$$

Subtracting the second equation from the first, we obtain

$$9x = 3$$

Solving for $x$, and reducing to lowest terms, we obtain

$$\begin{aligned} x &= \frac{3}{9} \\ &= \frac{1}{3} \end{aligned}$$

Thus, the repeating decimal number $0.3333\ldots$ is equivalent to the fraction $\frac{1}{3}$.

**(b)** Let $x = 0.\overline{142857}$ and multiply by 1 000 000 to obtain $1000000x = 142857.\overline{142857}$. Thus,

$$\begin{aligned} 1000000x &= 142857.\overline{142857} \\ x &= \quad 0.\overline{142857} \end{aligned}$$

Subtracting the second equation from the first, we obtain

$$999999x = 142857$$

Solving for $x$ and reducing to lowest terms, we obtain

$$\begin{aligned} x &= \frac{142857}{999999} \\ &= \frac{(3)(47619)}{(3)(333333)} \\ &= \frac{47619}{333333} \\ &= \frac{(3)(15873)}{(3)(111111)} \\ &= \frac{15873}{111111} \\ &= \frac{(3)(52921)}{(3)(37037)} \\ &= \frac{5291}{37037} \\ &= \frac{(11)(481)}{(11)(3367)} \\ &= \frac{481}{3367} \\ &= \frac{(13)(37)}{(13)(259)} \\ &= \frac{37}{259} \\ &= \frac{(37)(1)}{(37)(7)} \\ &= \frac{1}{7} \end{aligned}$$

Thus, we end up with the (perhaps surprising) conclusion that the repeating decimal $0.\overline{142857}$ is equivalent to the fraction $\frac{1}{7}$.

The factoring that reduced the fraction to lowest terms in Part (b) of the previous example is a considerable amount of work. I know that certain calculators will do this instantly, but I believe there is value in doing such work "by hand" and illustrating all of the steps of the thinking process.

## HISTORY

### Zeno's paradoxes

The ancient Greek philosopher Parmenides, who lived in the fifth century BCE and founded the Eleatic school, believed that change does not exist and that, in particular, motion is just an illusion. This contrasts with Heraclitus, who lived at roughly the same time, and who believed that change is continuous, summarized by his statement that nobody ever steps in the same river twice. This contrast between two world views, the static view of Parmenides vs. the dynamic view of Heraclitus, being vs. becoming, stimulated Western philosophy for centuries, including the highly influential Plato, who lived about a century later.

Zeno of Elea, one of the students of Parmenides, devised a series of arguments to support his teacher's views. The most famous of these arguments, now known as Zeno's paradoxes, is the story of Achilles and the tortoise, in which Zeno argues that the swift Achilles could never overtake a slow tortoise provided the latter were given a head start. The same argument, in simpler form, is given by Zeno in the dichotomy paradox, where he argues that it is impossible for anyone to reach any distant point by moving. The reason, says Zeno, is that before reaching the end of the journey one must first reach the midpoint of the journey, but before that one must first reach the quarter-point, and before that one must reach the eighth-point, and so on. This process requires an infinite number of steps, and is therefore impossible.

You can see that Zeno's argument contains an infinite geometric series in it, and it is interesting to think about how long ago infinite series were discussed, even without the mathematical precision that was developed more than two millennia later. There is a vast literature on Zeno's paradoxes; to learn more, you might begin at the Wikipedia page and then consult the references listed there.

**EXERCISES**

(Answers at end.)

Determine the sum of each infinite geometric series.

1. $1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+\frac{1}{81}+\cdots$
2. $1-\frac{1}{3}+\frac{1}{9}-\frac{1}{27}+\frac{1}{81}-\cdots$
3. $1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\frac{1}{256}+\cdots$
4. $1-\frac{1}{4}+\frac{1}{16}-\frac{1}{64}+\frac{1}{256}+\cdots$
5. $4+\frac{4}{5}+\frac{4}{25}+\frac{4}{125}+\frac{4}{625}+\cdots$
6. $7+\frac{7}{9}+\frac{7}{81}+\frac{7}{729}+\frac{7}{6561}+\cdots$
7. $0.5555\ldots$
8. $0.9999\ldots$
9. $0.313131\ldots$
10. $0.217217217\ldots$

---

Answers: 1. $\frac{3}{2}$; 2. $\frac{3}{4}$; 3. $\frac{4}{3}$; 4. $\frac{4}{5}$; 5. 5; 6. $\frac{63}{8}$; 7. $\frac{5}{9}$; 8. 1; 9. $\frac{31}{99}$; 10. $\frac{217}{999}$

## 10.3 The Harmonic Series

A very interesting infinite series is the harmonic series, which has $n$-th term $\frac{1}{n}$:

$$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots$$

The series is not geometric; there is no common ratio. Do you think the series converges or diverges? If you think it converges, what do you think the approximate value of its sum might be? Give this some thought and play with this before you read on.

Initially it may be difficult to decide whether the series converges or diverges. On one hand, the terms get smaller and smaller, definitely a necessary condition for convergence. But does this property guarantee convergence?

It may be interesting to play with a calculator, although the series grows so slowly that play with a calculator may not provide you with a good sense one way or the other.

There are two standard arguments that convincingly show that the harmonic series diverges. The first dates back to the 1300s and was presented then by Nicole Oresme.[3] Consider:

$$1+\frac{1}{2}+\underbrace{\frac{1}{3}+\frac{1}{4}}_{>\frac{1}{4}+\frac{1}{4}}+\underbrace{\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}}_{>\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}}+\cdots$$

Here's a paraphrase of Oresme's argument: The first term is 1. The sum of the first two terms is 1.5. If you add the next two terms, you are now beyond 2 in your journey, because the sum of the next 2 terms is certainly greater than $1/4 + 1/4$, in other words greater than $1/2$. If you add in the next 4 terms, you definitely make it past 2.5, because the sum of the next 4 terms is certainly greater than $1/2$. Similarly, if you add in the next 8 terms, you certainly make it past 3, because the sum of the next 8 terms is also certainly greater than $1/2$.

Continuing this argument, you can see that in your harmonic series journey, you definitely make it past any given positive number, and so the series does not converge. It grows very slowly indeed, but the series does not converge.

Besides the interesting fact that the harmonic series does not converge, the important take-away from this classic example is that just because the terms of a series approach zero, this is not a guarantee that the series converges. That the terms of a series approach zero is a necessary condition for convergence, but it is not a sufficient condition.

The harmonic series is so called because of the wavelengths of a vibrating string, such as you might find in a piano, a violin, a guitar, etc. A vibrating string has a longest wavelength of vibration, called the fundamental wavelength, and then shorter wavelengths of vibration, called overtones. The overtones have wavelengths that are fractions of the fundamental wavelength equal to $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and so on. For a lovely discussion about this connection between the harmonic series and musical scales, see Section 44 of *Algebra*, by I.M. Gelfand and A. Shen.

## 10.4 An Introduction to Power Series

So far we have discussed infinite series of numbers. Mathematicians have made important use of infinite series of functions for various purposes. The simplest kind of infinite series of functions is called a power series.

Power series are a fundamental tool used in many aspects of applied mathematics. For example, representing a function as a power series is one of the standard methods for solving a differential equation, and differential equations are fundamental descriptions in applied mathematics in general, and physics and engineering in particular.

Let's begin by discussing the idea of approximating a function using a polynomial, one of the fundamental ideas in calculus. One of the basic ideas of differential calculus is that the best linear approximation to a smooth curve near a point of interest on the curve is provided by the curve's tangent line through the point of interest. If the curve is the graph of a function, then the slope of the curve is the value of the function's derivative at the point of interest.

The phrase "near a point of interest" is key, because typically as you move away from the point of interest the approximation becomes worse, and the farther away from the point of interest, the worse the approximation typically becomes.

But why stop with linear approximations? Maybe we can do better with, say, quadratic approximations? Or cubic, or higher-power approximations? This leads us to the idea of a power series,

[3]The second argument requires knowledge of integral calculus, so we shall leave it for another time.

which is (typically[4]) an infinite series of powers of $x$ that represents a function. Taking only a finite number of terms of the power series amounts to a polynomial approximation. If the function is differentiable[5] at the point of interest, then the right power series will give a progressively better approximation to the function near the point of interest the more terms in the series are used. Another way to say this is that if a function has a power series representation, then each successive term in the sequence of partial sums of the power series is a progressively better approximation to the function.

When we discussed the calculation of the slope of a curve at a point, we discussed a sequence of progressively better numerical approximations to the slope, which itself is a number. Now we are taking the discussion to another level, discussing approximating entire functions by a sequence of polynomial functions. In what sense can we say that the approximations become progressively better? What, precisely, does better mean? How do you precisely conceive of and measure what "better" means in this context?

These are all interesting questions, and are typically discussed in detail towards the end of a course on integral calculus. In this section we will provide an introduction.

As a first example, consider again the sum formula for a convergent geometric series that we discussed earlier in this chapter:

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$$

In the case that the first term of the series is $a = 1$, then the series and sum formula become

$$1 + r + r^2 + r^3 + \cdots = \frac{1}{1-r}$$

So far our perspective has been to focus on the series of numbers on the left side of the previous equation, and to ask whether the series converges or not; in case of convergence, then the formula on the right side provides the sum. Let's now alter our perspective; replacing $r$ by $x$ and interchanging the sides of the equation may help to facilitate this change of perspective:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$$

The left side of the previous equation is a function, and the right side of this equation is the power series representation of this function! A graph of the function and the first two terms of the power series are plotted in Figure 10.9.

In Figure 10.9, the red line is the tangent line to the graph of the function at $x = 0$, which is the best linear approximation to the graph at $x = 0$. Observe that the red line and the black curve are almost indistinguishably near the point of tangency at $x = 0$, but farther from the point of tangency they differ more significantly.

[4] A formula for a polynomial function is its own power series, and so only has a finite number of nonzero terms. The cases of interest, where power series are useful, is for functions that are not polynomials.

[5] This is not quite correct; it is necessary for a function to be differentiable, but this condition is not sufficient to guarantee that a power series will give a good approximation to a function. See the "Digging Deeper" feature box entitled "Convergence of power series" later in this chapter for an example.

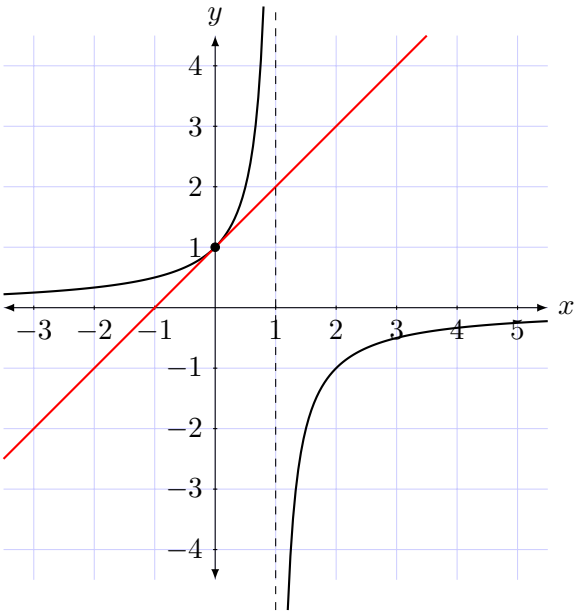



Figure 10.9: The graph illustrates the function $f(x) = \dfrac{1}{1-x}$, graphed in black, and the first two terms of its power series, $y = 1 + x$, graphed in red. Note the vertical asymptote at $x = 1$, graphed as a dashed line.

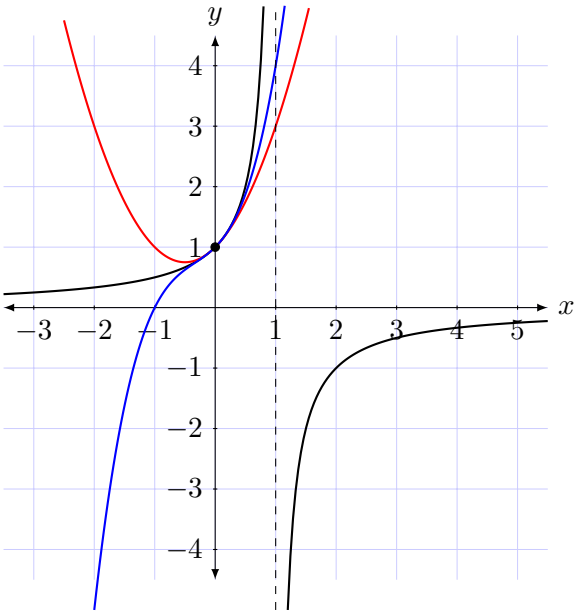



Figure 10.10: The graph illustrates the function $f(x) = \dfrac{1}{1-x}$, graphed in black, the first three terms of its power series, $y = 1 + x + x^2$, graphed in red, and the first four terms of its power series, $y = 1 + x + x^2 + x^3$, graphed in blue.

**PLAY!**

**Polynomial approximations to a function**

We can continue to plot additional terms in the power series to see what happens. In fact using online software, you can do this for a considerable number of terms and really get a sense for what happens as you increase the number of terms. In Figure 10.10, three and four terms of the power series are plotted. Carefully examine Figure 10.9 and Figure 10.10. Doesn't it seem that the quadratic approximation to the curve (in red in Figure 10.10) is a little better than the linear approximation (in red in Figure 10.9), and also doesn't it seem that the cubic approximation (in blue in Figure 10.10) is a little better yet? Perhaps it is true that the more terms in the power series you include, the better the power series approximates the actual curve. Use software to plot a sequence of partial sums of the power series; this will be interesting!

It's also worth noting that it seems that the "zone of good approximation" seems to get a little wider as we take more terms in the power series. This conjecture is also worth looking out for when you use software to examine the sequence of partial sums of the power series.

We have previously learned that the sum formula for an infinite geometric series converges if and only if $-1 < r < 1$, so it's reasonable to guess that perhaps the power series for the function $f(x) = \dfrac{1}{1-x}$ (that is, $y = 1 + x + x^2 + x^3 + \cdots$) converges if and only if $-1 < x < 1$. Watch out for this when you use software to examine the sequence of partial sums of the power series.

**PLAY!**

**Obtaining new power series from known power series**

Now that we have a power series

$$1 + x + x^2 + x^3 + \cdots$$

for the function $f(x) = \dfrac{1}{1-x}$, can we obtain power series for other functions by transforming the $x$-value in various ways? For example, replace $x$ by $-x$ in the formula for the function to obtain

$$f(x) = \frac{1}{1+x}$$

If we replace $x$ by $-x$ in the power series,

$$1 - x + x^2 - x^3 + \cdots$$

is this new power series valid for the new function? You can play with this idea by plotting sequences of partial sums for the new power series using software on the same set of axes as the new function.

If this works to your satisfaction, try transforming both the original function and its power series in various ways to explore whether the resulting power series converges to the function near the contact point. How wide does the "zone of good approximation" seem to be in each case?

This kind of exploration will prepare you well for the integral calculus course you will take, where you will state and prove many theorems about power series.

## CHALLENGE PROBLEM

**Approximating the function $f(x) = \dfrac{1}{1-x}$ at a different $x$-value**

Is it possible to determine a power series that approximates the function $f(x) = \dfrac{1}{1-x}$, but near a different $x$-value? Try it, say, for $x = 2$. Start with a linear approximation; you know how to determine the best linear approximation! Can you achieve more terms in the power series? How wide is the "zone of good approximation?"

If you can succeed at this task, the next challenge is can you do the same for other values of $x$? After you have done one or two others, perhaps you can figure out how to construct a power series approximation to this function at an arbitrary value of $x$. Once again, this kind of play will be great preparation for your further studies of power series in a future university integral calculus course.

Having a power-series representation of a function allows us to approximate the function arbitrarily accurately where the series converges. Often taking just the first few terms is good enough, and in many applications taking just a linear approximation suffices, particularly when $x$ is close to zero. The following example is an illustration.

## EXAMPLE 24

**Linear approximation for a sphere**

A sphere with radius 10 cm is to be painted with a layer of paint 0.12 cm thick. Determine the volume of paint needed.

## SOLUTION

The formula for the volume of a sphere is

$$V = \frac{4}{3}\pi r^3$$

Once the sphere is painted, its new radius (including the paint) will be 10.12 cm. The volume $\Delta V$ of paint used is the difference between the volume of the painted sphere and the "bare" sphere. Thus,

$$\begin{aligned}
\Delta V &= \frac{4}{3}\pi\left(r+\Delta r\right)^3 - \frac{4}{3}\pi r^3 \\
\Delta V &= \frac{4}{3}\pi\left[\left(r+\Delta r\right)^3 - r^3\right] \\
\Delta V &= \frac{4}{3}\pi\left[(10.12)^3 - 10^3\right] \\
\Delta V &= 152.6 \text{ cm}^3
\end{aligned}$$

We can determine a general expression for the volume of the paint by simplifying the right side of the second line above:

$$\begin{aligned}
\Delta V &= \frac{4}{3}\pi\left[\left(r+\Delta r\right)^3 - r^3\right] \\
\Delta V &= \frac{4}{3}\pi\left[\left(r^3 + 3r^2\Delta r + 3r\left(\Delta r\right)^2 + \left(\Delta r\right)^3\right) - r^3\right]
\end{aligned}$$

$$\Delta V = \frac{4}{3}\pi\left[3r^2\Delta r + 3r\left(\Delta r\right)^2 + \left(\Delta r\right)^3\right]$$

For the data given, $\Delta r = 0.012r$. For such small values of $\Delta r$ relative to $r$, we can obtain a good approximation by just retaining the first term in parentheses in the previous equation, omitting the terms involving higher powers of $\Delta r$. This is called a linear approximation, because only the first power of $\Delta r$ is retained. The results are

$$\begin{aligned}\Delta V &\approx \frac{4}{3}\pi\left[3r^2\Delta r\right]\\ \Delta V &\approx 4\pi r^2\Delta r\\ \Delta V &\approx 4\pi\left(10\right)^2\left(0.12\right)\\ \Delta V &\approx 150.8\end{aligned}$$

The approximation is within about 1% of the true value. This level of approximation might or might not be acceptable, depending on the application.

Considering the general approximation derived in the previous example, $\Delta V \approx 4\pi r^2\Delta r$, one can't help noticing that the coefficient of $\Delta r$ on the right side is equal to the derivative of the volume of the sphere with respect to its radius:

$$V = \frac{4}{3}\pi r^3 \quad \Longrightarrow \quad V'(r) = 4\pi r^2$$

This is true in general: Whenever we have a functional relationship $y = f(x)$, for small changes in $x$ we can approximate the corresponding change in $y$ as

$$\Delta y \approx f'(x)\Delta x$$

Another reason to call this a linear approximation is that it works with the tangent line to the functional relationship, as illustrated in Figure 10.11. Recalling the definition of the derivative,

$$f'(x) = \lim_{\Delta x\to 0}\frac{f(x+\Delta x) - f(x)}{\Delta x}$$

which can also be written as

$$f'(x) = \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}$$

one can deduce that the approximation $\Delta y \approx f'(x)\Delta x$ becomes exact in the limit as $\Delta x \to 0$.

In many cases an approximation based on the tangent line, which amounts to taking only the linear terms of the power series, is sufficiently accurate for scientific and engineering applications. If this is not sufficiently accurate, then one simply takes more terms in the power series until the desired accuracy is achieved.

### 10.4.1 The Taylor Methodology

We have been studying the power series for the function $f(x) = \dfrac{1}{1-x}$, but we have not yet discussed how to determine a power series for a given function. Let's now turn to this question. As a specific example, consider the sine function, and suppose that we wish to approximate it by a polynomial

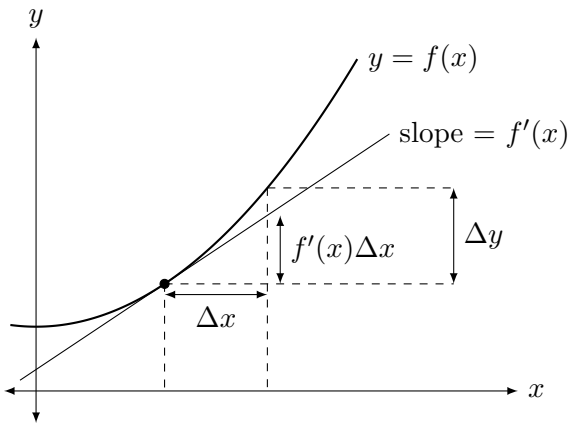



Figure 10.11: Using the tangent line as the basis for a linear approximation allows one to estimate the value of $\Delta y$ as $f'(x)\Delta x$. The approximation is generally better when $\Delta x$ is smaller, and in the limit as $\Delta x \to 0$, the approximation becomes exact.

in such a way that the approximation is good near $x = 0$. The method for determining a power series for the sine function requires that we know the derivatives of the sine and cosine functions. I'll state these derivatives, and display the graphs of the sine and cosine functions in Figure 10.12 to help you convince yourself that these formulas are at least plausible. You will derive these formulas a little more rigorously in a standard first-year university calculus course:

$$(\sin x)' = \cos x \qquad (\cos x)' = -\sin x$$

These formulas are valid provided that the angle $x$ is measured in radians. If $x$ is measured in degrees each (right side of the) derivative formula must include a factor of $\frac{\pi}{180}$.[6]

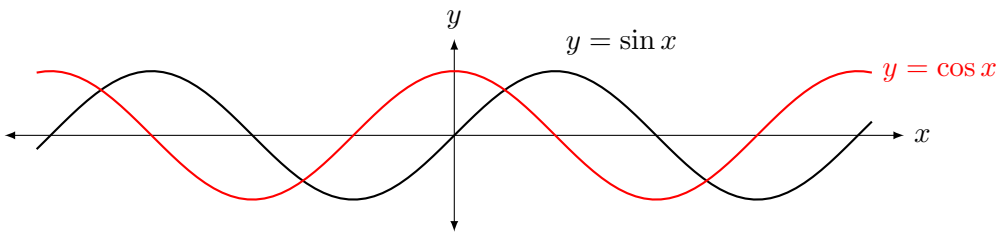



Figure 10.12: The derivative of the sine function is the cosine function and the derivative of the cosine function is the negative of the sine function.

Note that the slope of the sine graph is the value of the cosine graph at each point; test this by checking key points. Then note that the slope of the cosine graph is the negative of the value of

[6]The reason for the additional factors in degree measure are explained in a first-year university calculus course, but it is possible to understand this without detailed explanation if you think in terms of units. When $x$ is measured in degrees, the units of the derivative formula are 1/degrees. To convert degrees to radians, one must multiply by $\frac{\pi}{180}$. Therefore,

$$\frac{\pi}{180}\left(\frac{1}{\text{degrees}}\right) = \frac{\pi}{180}\left(\frac{1}{\frac{\pi}{180}\text{radians}}\right) = \frac{1}{\text{radians}}$$

Human nature being what it is, we generally prefer to use formulas with unnecessary factors in them, which explains why radian measure is used almost universally in calculus courses instead of degree measure (the formulas for the derivatives of the sine and cosine function are so much simpler using radian measure).

the sine graph at each point; again, test this by checking key points. This does not prove that the stated derivative formulas are correct, but it will make them plausible.

Let's begin with a linear approximation to the sine function; we begin to determine this by calculating the value of the derivative of the sine function at $x = 0$:

$$f(x) = \sin x \quad \Longrightarrow \quad f'(x) = \cos x \quad \Longrightarrow \quad f'(0) = 1$$

Thus, the slope of the tangent line to the graph of the sine function at $x = 0$ is 1, and so an equation for this tangent line has the form $y = mx + b$, where $m = 1$. The value of $b$ can be determined by the condition that the tangent line should touch the graph of the function at $x = 0$, and so $b = \sin 0 = 0$. Thus, an equation for the tangent line to the sine function at $x = 0$ is $y = x$, and so for values of $x$ close to 0, we can say that

$$\sin x \approx x$$

So much for the best linear approximation. Let's now seek the best quadratic approximation. That is, let's begin with a generic quadratic function, $y = ax^2 + bx + c$, and let's determine the values of the coefficients $a$, $b$, and $c$ that make this the best quadratic approximation to the sine function near $x = 0$.

**Question:** How should we do this? What are the criteria that we should use?

Certainly the value of the approximating function should be equal to the value of the given function at $x = 0$. In the case of the linear approximation, we also required the slope of the approximating linear function to be the same as the value of the given function's derivative at the contact point. So perhaps it will seem reasonable to you that we should require that all of the values of the derivatives of the approximating function should match the corresponding values of the derivatives of the given function.

Graphically, this means that we desire the graph of the approximating function to go through the same point as the graph of the given function, we desire the graphs to have the same slope, and we also desire them to bend at the same rate (i.e., the higher derivatives are equal).

These criteria I call "the Taylor methodology," because they describe how to determine what is called in textbooks the Taylor series for a function. The phrase "Taylor methodology" is not standard, but I feel it accurately represents what is a prescription for determining a power-series representation for a function; whether we call the resulting expression a Taylor series, or power series, or something else, is of no serious consequence. It's the method that is important to understand, and it's also important to practice using the method, as you will be using it a lot in your university mathematics courses.

So, let's implement the Taylor methodology to determine the best quadratic approximation to the sine function near $x = 0$. Start by determining expressions for the first few derivatives of the approximating function (the third and higher derivatives are identically zero):

$$\begin{aligned} y &= ax^2 + bx + c \\ y' &= 2ax + b \\ y'' &= 2a \end{aligned}$$

And here are the first few derivatives of the sine function:

$$\begin{aligned} f(x) &= \sin x \\ f'(x) &= \cos x \\ f''(x) &= -\sin x \end{aligned}$$

Evaluating all of the previous six expressions at $x = 0$, we obtain

$$\begin{aligned} y(0) &= a(0)^2 + b(0) + c = c \\ y'(0) &= 2a(0) + b = b \\ y''(0) &= 2a \\ f(0) &= \sin(0) = 0 \\ f'(0) &= \cos(0) = 1 \\ f''(0) &= -\sin(0) = 0 \end{aligned}$$

Matching the corresponding values, we determine the values of $a$, $b$, and $c$:

$$\begin{aligned} y(0) = f(0) &\implies c = 0 \\ y'(0) = f'(0) &\implies b = 1 \\ y'' = f''(0) &\implies 2a = 0 \end{aligned}$$

Thus, the best quadratic approximation to the sine function near $x = 0$ is $y = ax^2 + bx + c = (0)x^2 + (1)x + (0) = x$. Which is the same as the best linear approximation to the sine function near $x = 0$; that's interesting and worth thinking about.

**Question:** What are your thoughts about this?

Next, continue to determine the best cubic approximation to the sine function near $x = 0$, then the best fourth-degree approximation, and so on. It would be interesting to plot the successive approximations together with the sine function on the same set of axes; using software will make this task relatively easy. Once you have done this, perform the same task for some other commonly-used function, such as the cosine function.

**Question:** What are the results that you obtained?

After you perform these tasks a few times, you'll notice patterns, and you'll be encouraged to develop general formulas. Let's do this. Suppose your approximating power series is of the form

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots$$

Now let's use the same criteria as discussed earlier to determine the values of the coefficients $a_n$: The value of the approximating function and the values of its derivatives should match the corresponding values for the given function. First, the value of the approximating function at $x = 0$ matches the value of the given function at $x = 0$, so

$$y(0) = a_0 + a_1(0) + a_2(0)^2 + a_3(0)^3 + \cdots = a_0$$

and therefore

$$a_0 = f(0)$$

Similarly, matching first derivatives allows us to determine the value of $a_1$ in terms of the given function:

$$\begin{aligned} y' &= a_1 + 2a_2x + 3a_3x^2 + 4a_4x^3 + \cdots \\ y'(0) &= a_1 + 2a_2(0) + 3a_3(0)^2 + 4a_4(0)^3 + \cdots = a_1 \end{aligned}$$

and therefore

$$a_1 = f'(0)$$

Similarly, matching second derivatives allows us to determine the value of $a_2$ in terms of the given function:

$$y'' = 2a_2 + (3)(2)a_3x + (4)(3)a_4x^2 + (5)(4)a_5x^3 + \cdots$$
$$y''(0) = 2a_2 + (3)(2)a_3(0) + (4)(3)a_4(0)^2 + (5)(4)a_5(0)^3 + \cdots = 2a_2$$

and therefore

$$2a_2 = f''(0)$$
$$a_2 = \frac{f''(0)}{2}$$

Continuing in the same way, matching third derivatives allows us to determine the value of $a_3$ in terms of the given function:

$$y''' = (3)(2)a_3 + (4)(3)(2)a_4x + (5)(4)(3)a_5x^2 + (6)(5)(4)a_6x^3 + \cdots$$
$$y'''(0) = (3)(2)a_3 + (4)(3)(2)a_4(0) + (5)(4)(3)a_5(0)^2 + (6)(5)(4)a_6(0)^3 + \cdots = (3)(2)a_3$$

and therefore

$$(3)(2)a_3 = f'''(0)$$
$$a_3 = \frac{f'''(0)}{(3)(2)}$$

Continuing in the same way, you'll be able to conclude that the next couple of coefficients are

$$a_4 = \frac{f''''(0)}{(4)(3)(2)}$$
$$a_5 = \frac{f'''''(0)}{(5)(4)(3)(2)}$$

Continuing in the same way, you'll be able to determine that the $n$-th coefficient of the power series for the function $f(x)$ is equal to the value of the $n$-th derivative of the function (symbolized by $f^{(n)}$) at $x = 0$ divided by $n!$:

$$a_n = \frac{f^{(n)}(0)}{n!}$$

**DEFINITION 11**

**Factorial notation**

The symbol $n!$, read "n-factorial," is a compact way to write the quantity

$$n! = n(n-1)(n-2)(n-3)\cdots(3)(2)(1)$$

For example, 5!, read "five-factorial," stands for

$$5! = (5)(4)(3)(2)(1) = 120$$

By convention, zero-factorial is defined to be 1:

$$0! = 1$$

Using this compact notation often makes complicated formulas less cluttered and easier to read and understand.

Thus, the Taylor series for the function $f(x)$ expanded about $x = 0$ is

$$f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \frac{f'''(0)}{3!}x^3 + \frac{f''''(0)}{4!}x^4 + \cdots$$

If instead one wishes to determine a Taylor series for a function expanded about some other value, say $x = a$, one can show (do this!) that the Taylor series is

$$f(a) + f'(a)\,(x-a) + \frac{f''(a)}{2}\,(x-a)^2 + \frac{f'''(a)}{3!}\,(x-a)^3 + \frac{f''''(a)}{4!}\,(x-a)^4 + \cdots$$

**Question:** Were you able to show that the previous displayed equation is the correct result for the Taylor series expansion of a function about an arbitrary point $x = a$? If not, come back to this task and keep at it; it's well worth completing, as it is good practice for the kinds of calculations you will be required to do and understand once you reach university.

**EXERCISES**

(Answers at end.)

Using the Taylor methodology (i.e., using the formulas given above), determine power series for (a) $\sin x$ and (b) $\cos x$.

---

Answers:

(a) $x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \cdots$

(b) $1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + \frac{x^{2n}}{(2n)!} + \cdots$

One can't help but notice that the power series representation of the sine function contains only odd powers of $x$, whereas the power series representation of the cosine function contains only even powers. Evidently the definitions of even and odd functions are consistent with the idea of even and odd powers. That's fun! And it also sheds light on the question we were pondering a couple of pages ago, where we determined that the best quadratic approximation to the sine function is the same as the best linear approximation.

**Question:** Do the power series for the sine and cosine function converge? If so, for which values of $x$ do they converge?

The question of convergence arises; for which values of $x$ does a power series converge? This is a question that is typically studied in a first-year university integral calculus course, where you will learn various methods for determining the values of $x$ for which a power series converges. The general answer is that a power series either converges for no values of $x$, for just a single value of $x$, for an interval of values centred at the expansion point, or for all values of $x$. In case a power series converges for all values in an interval, the interval might be open, closed, or include just one of its endpoints.

If a power series converges for a particular value of $x$, does its sum for this value of $x$ necessarily equal the value of the function (at the same value of $x$) that the power series is supposed to represent? Alas the answer to this question is "No, not necessarily!"[7] There are exotic functions for which the power series does not equal the function that it is supposed to represent! This complicates matters; we have a procedure (which I have been calling the Taylor methodology) that

[7] Isn't mathematics rich and interesting! Cases such as this one emphasize a point that we have been making repeatedly about the importance of rigorous proof in mathematics.

produces a power series for a given function, but what good is the power series if it doesn't (in some sense) equal the function?

Fortunately, for the functions that commonly arise in applications, it is true that if you determine a power series for a function, then for values of $x$ for which the power series converges, the sum of the power series is equal to the value of the function for the same value of $x$, so that we can safely say that the power series does accurately represent the function where the series converges.

The power series given above for the sine and cosine functions each converge for all real values of $x$, so they are particularly nice. Such functions are called *analytic* functions.

## DIGGING DEEPER

### Convergence of power series

The Taylor methodology provides a means for determining a power series for a function, but does the power series actually converge to the function? The answer might be yes for all values of $x$, or perhaps yes for some values of $x$, or in the worst case, no for virtually all values of $x$. Convergence is an important criterion; if the power series doesn't actually converge to the function it is supposed to represent, then what good is it?

An example of a bizarre function, for which the Taylor series converges to the function at just one single value of $x$, is

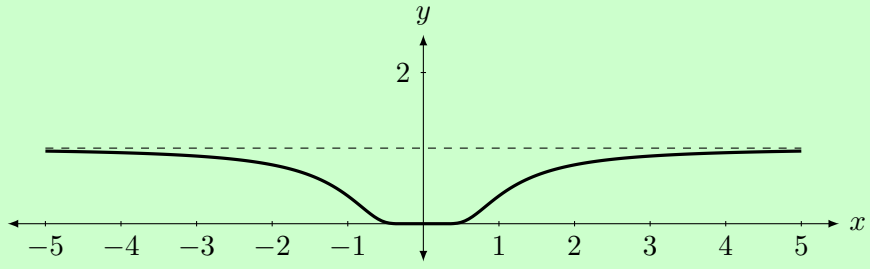
$$f(x) = \left\{ \begin{array}{ll} e^{-1/x^2} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{array} \right\}$$

If you haven't studied the special number $e$ yet, then you are in for a treat when you see it in your first-year university calculus course. This special number, sometimes called Euler's number, is an irrational number, and its value (which is approximately 2.7) can be expressed by the infinite series

$$1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots$$

The natural exponential function, $y = e^x$ has the wonderful property that it is its own derivative! Based on this, and the fact that $e^0 = 1$, can you sketch a rough graph?

To determine the Taylor series for the function $y = f(x)$, expanded about $x = 0$, you will have to apply the definition of the derivative to determine the values of all the derivatives at $x = 0$. (Do this!) Do this, and you will determine that all of the derivatives of this bizarre function have a value of 0 at $x = 0$. This means that the Taylor series for this bizarre function is identically zero, the zero function, the graph of which is a horizontal line at $y = 0$. This Taylor series is equal to the function at just one point, $x = 0$, as you can see from the following graph.



If the function represented a position function for a moving object, where $x$ represents time and $y$ represents position, it would be a very bizarre motion, wouldn't it? Can you imagine a certain motion along the $y$-axis that starts from rest at time $x = 0$, has zero initial velocity, zero initial acceleration, the initial rate of change of acceleration is zero, indeed all derivatives of the position function are zero at time $x = 0$ ... and yet, and yet ... somehow it manages to move away from the origin! And somehow it manages to make it all the way to (nearly) $y = 1$. Bizarre.

This function was discovered in 1821 by Cauchy, and provided a wake-up call. By that time, considerable work had been done on power series, and it was widely thought that every function could be represented by its Taylor series. Cauchy's discovery showed that this is not so.

The point of this bizarre example is that one must be careful, and one must prove relevant theorems so that one understands the scope and limitations of one's mathematical methods. Specifically, for our purposes at the moment, it's good to know that when you use the Taylor methodology to determine a power series, there's no guarantee that the resulting power series actually converges to the function. Therefore, there's no guarantee that the resulting power series is a good representation of the function anywhere.

In a future university calculus course, you will learn more about the conditions for which a power series really does converge to the function it is supposed to represent. For now we will restrict ourselves to the following facts. A real-valued function for which its Taylor series expanded about some point actually does converge to the function in an open interval about that point is called a *real analytic* function, or just an analytic function for short. All of the usual elementary functions that we know and love (polynomial functions, exponential functions, logarithm functions, trigonometric functions, power functions) are analytic functions.

The Taylor methodology is a fail-safe method for determining a power series, but there are other methods that are sometimes easier, and why shouldn't we use the easier methods if the results are the same?

For example, if you have a power series for a particular function, then differentiating the power series term-by-term results in a power series for the derivative of the function. The region of convergence of the differentiated power series is more-or-less the same as the original power series (the only difference might be the endpoints of the interval of convergence, if the region of convergence is an interval; once again, you will be going over the specific theorems in a first-year university calculus course). This very useful fact provides us with a nice trick for determining new power series from old ones. A similar theorem makes the analogous statement about anti-differentiating, which provides us with another trick.

For example, starting from the power series for the sine function,

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \cdots$$

you can differentiate term-by-term to obtain a power series for the cosine function:

$$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots + \frac{x^{2n}}{(2n)!} + \cdots$$

This is much easier than starting from scratch and using the Taylor methodology on the cosine function. Because the power series for the sine function converges to the sine function for all real values of $x$, the analogous statement can be made about the power series for the cosine function.

**Question:** What happens when you differentiate the power series for the cosine function term-by-term? Try it! Is the result reasonable?

### 10.4.2 Binomial Series

One of the bread-and-butter tools used in physics is to make approximations using a binomial series. Binomial series were discovered by Newton, and he put them to great use for numerical approximations; for example, he devised a very efficient approximation to the value of $\pi$ that was much better than anything done up to that time, and he used binomial approximations to calculate all kinds of square roots, cube roots, etc. Nowadays physicists typically use binomial approximations to simplify expressions rather than for numerical calculations, since everyone has numerical computers at their finger tips. For instance, a nasty differential equation might be usefully simplified by replacing some complicated expression with a more friendly one using a binomial series.

A modern formulation of a version of a binomial series is as follows:

$$(1+x)^r = 1 + rx + \frac{r(r-1)}{2}x^2 + \frac{r(r-1)(r-2)}{3!}x^3 + \frac{r(r-1)(r-2)(r-3)}{4!}x^4 + \cdots$$

where $r$ is a real number. The series converges if $-1 < x < 1$, and its typical use for approximations involves using only the first few terms, which can be a reasonable approximation when $x$ is very close to zero. In the special case that $r$ is a natural number, observe that after the first $(r+1)$ terms, each subsequent term is identically zero, so the series is finite. In this case, it is popular to write the finite series in the form

$$(a+b)^n = \binom{n}{0}a^n b^0 + \binom{n}{1}a^{n-1}b + \binom{n}{n-2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}a^0 b^n$$

where the coefficients in the terms are combinatorial factors known as *binomial coefficients*, and are defined by

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

These binomial coefficients fit into a nice geometrical pattern, known as Pascal's triangle; if you haven't seen Pascal's triangle, or don't remember its charms, you will certainly wish to check it out and refresh your memory.

First let's look at a calculation that would have been typical in Newton's time (and which was a great advance over the previous laborious methods), and was also typical as little as 40 or so years ago, which is when hand-held electronic calculators became generally available. Suppose you need to calculate the value of the square root of five. There is a complicated algorithm (a distant cousin of long division) that one could use in the old days for calculating this quantity, but a binomial series is much better used, as follows. A binomial series converges when $x$ is close to zero (specifically between $-1$ and 1), which motivates the first few steps in the following development:

$$\sqrt{5} = \sqrt{4+1}$$
$$\sqrt{5} = \sqrt{4\left(1+\frac{1}{4}\right)}$$
$$\sqrt{5} = 2\left(1+\frac{1}{4}\right)^{1/2}$$

Now approximate the second factor in the previous line using a binomial series, with $x = 1/4$. If only the first two terms are used, the approximation is

$$\sqrt{5} = 2\left(1+\frac{1}{2}\times\frac{1}{4}\right) = 2.25$$

Checking with a calculator, $2.25^2 = 5.0625$, which may or may not be a good enough approximation, depending on your needs. Using the first three terms of a binomial approximation, we get

$$\sqrt{5} = 2\left(1 + \frac{1}{2} \times \frac{1}{4} + \frac{(1/2)(-1/2)}{2!} \times (1/4)^2\right) = 2.234375$$

Checking with a calculator, $2.234375^2 \approx 4.9924$. We could continue with further approximations, and I encourage you to do so if you find it fun; otherwise, I hope you get the idea. The more terms used in the approximation, the greater the accuracy.

**Question:** How many decimal places of accuracy were you able to achieve in your approximation of $\sqrt{5}$ using the binomial series?

One might complain about this development because of the use I've made of a hand-held calculator to check the work. If I was going to use a calculator, why bother with a binomial series at all; just enter 5, press the square root key and be done with it! But remember, someone had to program this calculator; how did she do it? She used some algorithm or other, and whether it was a binomial series or some other clever trick, it helps to know about these techniques. One day you may have to write some complex computer program, and then you might make use of such methods. Having some experience playing with various approximation schemes will be very useful to you in making a wise choice in your construction of an appropriate algorithm.

The use of a binomial series for numerical calculations is ancient history, although I believe it's a bit of history worth knowing. More important is the use of binomial series in simplifying algebraic expressions, whether for solving differential equations or for some other reasons. For example, if you apply Newton's second law to the analysis of a simple pendulum (drawing a free-body diagram, of course, to guide your reasoning), the resulting second-order differential equation describing the motion of the pendulum is

$$\theta'' + \frac{g}{L}\sin\theta = 0$$

where $L$ is the length of the pendulum, $g$ is the acceleration due to gravity, and $\theta$ is the angular position of the pendulum measured from the vertical. This is a very nasty differential equation, and its solution requires advanced concepts such as elliptic integrals. However, if the amplitude is small, one can approximate the differential equation by replacing $\sin\theta$ using just the first term of the power series for $\sin\theta$; that is, $\sin\theta \approx \theta$ for small angles. The approximate differential equation is

$$\theta'' + \frac{g}{L}\theta = 0$$

which is a standard differential equation that can be solved exactly, and represents simple harmonic motion (which is a special kind of regular, back-and-forth motion that continues indefinitely in the absence of friction). Making the approximation provides us with several insights: First, that a simple pendulum oscillates approximately in simple harmonic motion if the amplitude is small, and second that the oscillation is not exactly simple harmonic motion. Your grandfather clock needs regular corrections if it is to keep time accurately in the long run, as small errors build up over time.

**EXERCISES**

(Answers at end.)

1. Use the first three terms of a binomial series to approximate $\sqrt{10}$.

2. In special relativity, the expression for the total energy of a freely moving particle that has mass $m$ is

$$E = \gamma mc^2$$

where the Lorentz factor $\gamma$ is related to the speed $v$ of the particle and the speed of light $c$ by

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Use a binomial series to expand the expression for the energy of a freely moving particle and identify the physical meaning of the first two terms.

---

Answers: 1. 3.162

2.

$$E = \gamma mc^2$$
$$E = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} mc^2$$
$$E \approx \left(1 + \frac{1}{2}\frac{v^2}{c^2} + \cdots\right) mc^2$$
$$E \approx mc^2 + \frac{1}{2}mv^2 + \cdots$$

The first term represents the "rest energy" of the particle and the second term represents the (Newtonian) kinetic energy of the particle. (The rest of the terms of the series also contribute to the kinetic energy of the particle, but this was not known before Einstein's work.) This is the kind of argument used by Einstein in the early 1900s to propose that matter can be converted to energy and vice versa!

## 10.5 An Iterative Method for Approximating the Solution of an Equation

One of the main themes of this book has been the idea of calculating a quantity by using a sequence of approximations that can be made arbitrarily good, so that one can achieve any level of desired accuracy by continuing the process of sequential approximation far enough. This idea is used frequently in the construction of computer algorithms, where the term **iteration** is used to describe this process.

An example of an iterative procedure involves the use of the Newton-Raphson method for approximating the zero of a function, which we will now describe.[8]

Suppose that we wish to approximate the square root of 2. One way to do this is to determine the positive $x$-intercept of the graph of the function $f(x) = x^2 - 2$; see Figure 10.13. The idea

[8] Compare this with the bisection method in Section 12.6.5; each method has its strengths and weaknesses.

behind the Newton-Raphson method is to first guess the desired value; the initial guess we shall use is $x_0 = 2$, although some other value might be better. The next step is to determine an equation for the tangent line to the graph of the function at $x = x_0$; it frequently happens that the $x$-intercept of this tangent line is a better approximation to the desired value than the initial guess. Once this $x$-intercept is determined, call it $x_1$. Then the process can be repeated, starting by determining an equation for the tangent line to the graph at $x = x_1$. The process can be repeated as often as needed to produce an approximation with the desired accuracy.

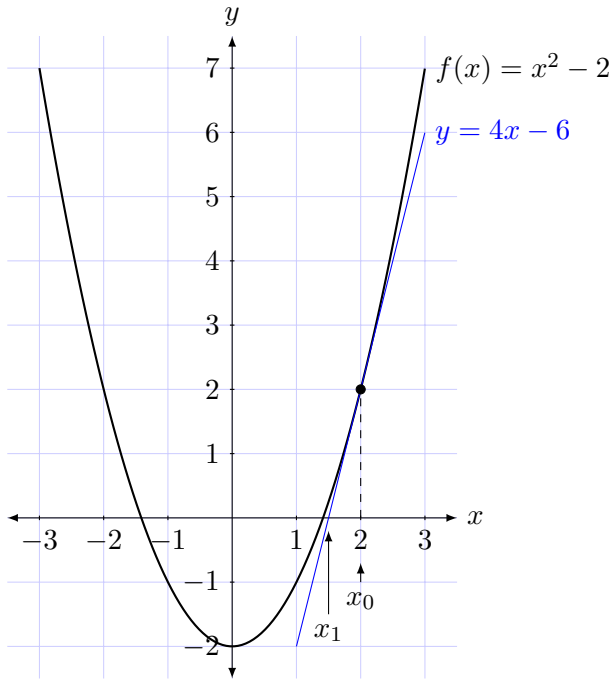



Figure 10.13: The figure illustrates the Newton-Raphson method for estimating an $x$-intercept of the graph of a function. See the text for a full explanation.

For the function $f(x) = x^2 - 2$, one can use the definition of the derivative to determine that $f'(x) = 2x$. Therefore, the slope of the tangent line to the graph of $f$ at $x = 2$ is $m = 2(2) = 4$, and so an equation for the tangent line has the form $y = 4x + b$. Because $f(2) = 2^2 - 2 = 2$, we can calculate the value of $b$ as follows:

$$\begin{aligned} y &= 4x + b \\ 2 &= 4(2) + b \\ 2 &= 8 + b \\ b &= -6 \end{aligned}$$

Thus, an equation for the tangent line to the graph of $f$ at $x = 2$ is $y = 4x - 6$. This line is plotted in blue in Figure 10.13. The $x$-intercept of this tangent line is supposed to be a better approximation to $\sqrt{2}$ than the original guess of $x_0 = 2$. The $x$-intercept of the tangent line can be calculated as follows:

$$\begin{aligned} 0 &= 4x - 6 \\ 4x &= 6 \\ x &= \frac{6}{4} \\ x &= 1.5 \end{aligned}$$

Thus, a better estimate of the value of $\sqrt{2}$ is $x_1 = 1.5$. Now we can repeat the process to determine better and better estimates of $\sqrt{2}$, stopping when we have reached the desired accuracy. Because this method does not give us upper and lower bounds on the estimate, one typically stops when the distance between successive estimates is smaller than a specified accuracy.

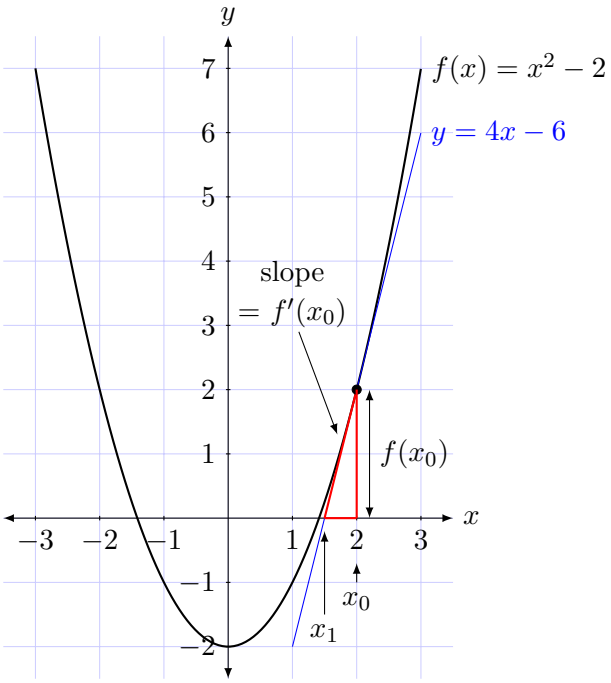



Figure 10.14: In the Newton-Raphson method, one can obtain a relation between $x_1$ and $x_0$ by applying "slope = rise over run" to the red triangle. See the text for full details.

We can make the repeated calculation of these estimates more efficient by developing a general formula for the Newton-Raphson method, which can be done by focussing attention on the red triangle in Figure 10.13 that has base between $x_1$ and $x_0$, has altitude given by the dashed line from $(x_0, 0)$ and $(x_0, f(x_0))$, and hypotenuse the part of the tangent line that extends from $(x_1, 0)$ to $(x_0, f(x_0))$. See Figure 10.14. We can obtain a relation between $x_1$ and $x_0$ by applying "slope = rise over run" to the red triangle in Figure 10.14, and then solving for $x_1$:

$$\begin{aligned} \text{slope} &= \frac{\text{rise}}{\text{run}} \\ f'(x_0) &= \frac{f(x_0)}{x_0 - x_1} \\ x_0 - x_1 &= \frac{f(x_0)}{f'(x_0)} \\ x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

Starting with a guess $x_0$ for the $x$-intercept of the graph of the function, the formula on the previous line then produces a (presumably) better estimate. But then the argument can be repeated by using $x_1$ in the previous role of $x_0$, to produce a (presumably) even better estimate:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

The process can be repeated indefinitely, so that once the $n$th estimate has been obtained, the

next, $(n+1)$-th, estimate is given by the Newton-Raphson formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

**KEY CONCEPT**

**Newton-Raphson formula**

The Newton-Raphson formula provides an iterative procedure for estimating a zero of function:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Going back to the problem of estimating the value of $\sqrt{2}$, we can insert the formulas $f(x) = x^2 - 2$ and $f'(x) = 2x$ into the Newton-Raphson formula to facilitate calculations:

$$\begin{aligned}
x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \\
x_{n+1} &= x_n - \frac{x_n^2 - 2}{2x_n} \\
x_{n+1} &= \frac{2x_n^2}{2x_n} - \frac{x_n^2 - 2}{2x_n} \\
x_{n+1} &= \frac{x_n^2 + 2}{2x_n} \\
x_{n+1} &= \frac{x_n}{2} + \frac{1}{x_n}
\end{aligned}$$

The last step in the previous derivation results in an expression that minimizes the number of operations that must be done in each calculation, and therefore results in faster computing times, which is especially important if you decide to program the calculations for automatic computing. Let's use the formula to perform a few calculations using a hand-calculator, starting with the same initial guess as before, $x_0 = 2$. The results are tabulated below. Perform the calculations yourself and compare your results to the ones below to ensure that you understand the procedure.

$$\begin{aligned}
x_0 &= 2 \\
x_1 &= 1.5 \\
x_2 &= 1.4166666666 \\
x_3 &= 1.414215686 \\
x_4 &= 1.414213562 \\
x_5 &= 1.414213562
\end{aligned}$$

That's pretty fast convergence, isn't it? With only four iterations of the formula we have obtained a result that is correct to ten significant figures. (The fifth iteration confirms that the result is stable.) You can check your result by calculating $\sqrt{2}$ directly using your hand calculator.

## EXAMPLE 25

**Using the Newton-Raphson method to estimate a zero of a function**

Estimate the solution to the equation $0 = x^3 - 3x + 3$.

## SOLUTION

Apply the Newton-Raphson method to the function $f(x) = x^3 - 3x + 3$. First use the definition of the derivative to calculate the derivative of the function. The result is

$$f'(x) = 3x^2 - 3$$

Now let's insert the formulas for $f$ and $f'$ into the Newton-Raphson formula and simplify:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$
$$x_{n+1} = x_n - \frac{x_n^3 - 3x_n + 3}{3x_n^2 - 3}$$
$$x_{n+1} = \frac{x_n\left(3x_n^2 - 3\right)}{3x_n^2 - 3} - \frac{x_n^3 - 3x_n + 3}{3x_n^2 - 3}$$
$$x_{n+1} = \frac{3x_n^3 - 3x_n}{3x_n^2 - 3} - \frac{x_n^3 - 3x_n + 3}{3x_n^2 - 3}$$
$$x_{n+1} = \frac{2x_n^3 - 3}{3x_n^2 - 3}$$

Trying a few guesses, it seems that $-2$ is close to the zero, so let's start with $x_0 = -2$. Using a hand-calculator, the results of the iteration are:

$$x_0 = -2$$
$$x_1 = -2.111111111$$
$$x_2 = -2.103835979$$
$$x_3 = -2.103803403$$
$$x_4 = -2.103803403$$

Check the result by substituting it into the original equation using a hand-calculator.

**Question:** In the previous example, what happens if we choose a different initial guess?

For example, what happens if we choose $x_0 = 0$ as an initial guess in the previous example? Let's see:

$$x_0 = 0$$
$$x_1 = -1$$
$x_2$ does not exist

Oops! That was a quick and unsuccessful try. Apparently we had bad luck and the first estimate brought us to a value of $x$ for which the tangent line is horizontal (the derivative of the function is in the denominator of the Newton-Raphson formula). Because a horizontal line does not intersect

the $x$-axis, that is the end of that attempt. What happens if we choose $x_0 = 4$?

$$\begin{aligned}
x_0 &= 4 \\
x_1 &= 2.777777778 \\
x_2 &= 1.978690087 \\
x_3 &= 1.428596107 \\
x_4 &= 0.906664675 \\
x_5 &= 2.827182697 \\
x_6 &= 2.011314029 \\
x_7 &= 1.452808047 \\
x_8 &= 0.940211801 \\
x_9 &= 3.843932367 \\
x_{10} &= 2.676053811
\end{aligned}$$

This is not looking very good, as the estimates keep bouncing around. Nevertheless, let's persevere. (I used a spreadsheet to make the calculations easier for me; you could do the same, or program your favourite software.)

$$\begin{aligned}
x_{11} &= 1.911288314 \\
x_{12} &= 1.377543069 \\
x_{13} &= 0.827413194 \\
x_{14} &= 1.973324423 \\
x_{15} &= 1.424585069 \\
x_{16} &= 0.900884698 \\
x_{17} &= 2.720525404 \\
x_{18} &= 1.940796670 \\
x_{19} &= 1.400079550 \\
x_{20} &= 0.864013262 \\
x_{21} &= 2.248682615 \\
x_{22} &= 1.622161975 \\
x_{23} &= 1.131362147 \\
x_{24} &= -0.123529471 \\
x_{25} &= 1.016772133 \\
x_{26} &= -8.846068500 \\
x_{27} &= -5.986662075 \\
x_{28} &= -4.134365526 \\
x_{29} &= -2.989652542 \\
x_{30} &= -2.370160561 \\
x_{31} &= -2.138854639 \\
x_{32} &= -2.104534491 \\
x_{33} &= -2.103803731 \\
x_{34} &= -2.103803403 \\
x_{35} &= -2.103803403
\end{aligned}$$

OK, the method finally yielded the correct result, but that took quite a number of iterations. From this attempt, where the convergence was not very efficient, with lots of bouncing around, and the previous attempt, which failed, we have learned that the starting guess in this method is important. Sketching a graph of the function from the previous example will be helpful. It turns out that if the graph of the function has a peak or a valley between the initial guess and the final result, then the process may not be very efficient. In this case, the graph has both a peak and a valley between the initial guess and the correct value.

This reinforces that applying the Newton-Raphson method effectively requires a certain amount of preliminary analysis, and a certain level of understanding of functions and their graphs. Once you learn more about calculus, which is a powerful tool for analyzing functions, you will be able to apply the method more effectively.

It turns out that for some graphs, if the guess is just bad enough, then the iteration process will not converge.

**Question:** Can you sketch a graph that has this problem? See the next set of exercises for an example, but try to find an example yourself first.

In Section 12.6.5 you will learn the bisection method, which is a fail-safe method for calculating zeros of functions. Because determining zeros of functions amounts to solving equations, any such method for determining zeros of a function is important, because there are lots of equations to be solved in mathematics and its applications to science. The bisection method typically converges very slowly, but it always works. The Newton-Raphson method typically converges rapidly, but it is sensitive to initial guesses, and sometimes doesn't converge very well or at all. The moral is that it is helpful to have many tools in your tool box!

## TRICKS OF THE TRADE

### Should you memorize the Newton-Raphson formula?

As I mentioned earlier, it is best to memorize the absolute minimum amount of material, and practice your skills systematically so that you can reproduce any formula you need. As your understanding grows, the amount of elementary material that you have to memorize by rote decreases.

Rather than memorize the Newton-Raphson formula, I highly recommend practicing sketching a figure somewhat like a simplified version of Figure 10.14, and then practice deriving the Newton-Raphson formula. Doing this a number of times, using spaced repetition, will help you to internalize the idea, and you will be able to reproduce the formula with a lot less work and a lot more understanding than if you take a lot of time to memorize the formula by rote repetition.

## CHALLENGE PROBLEM

### An iterative process for determining the square root of an arbitrary real number

Earlier we applied the Newton-Raphson method to the function $f(x) = x^2 - 2$ to determine an iterative formula for $\sqrt{2}$:

$$x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$$

Develop an iterative formula similar to this one that can be used to estimate $\sqrt{k}$, for any positive real number $k$.

## EXERCISES

(Answers at end.)

Use the Newton-Raphson method to estimate a zero of the function, estimate the value of the quantity, or solve the equation, as appropriate. Try various starting guesses and use a sufficient number of iterations to produce a valid result. Check your result!

1. $\sqrt{5}$
2. $\sqrt{15}$
3. $x^2 - x - 1 = 0$
4. $x^2 + x - 1 = 0$
5. $x^3 - 3x + 2 = 0$
6. $x^3 - 3x + 1 = 0$
7. $f(x) = x^{1/2}$
8. $f(x) = x^{1/3}$

---

Answers: 1. 2.236067977 2. 3.872983346 3. 1.618033989 and 0.618033988 4. $-1.618033989$ and 0.618033988

5. $-2$ and 1 6. $-1.879385242$, 0.347296355, and 1.532088886

7. The solution is $x = 0$, but using the Newton-Raphson method with any non-zero starting guess leads to problems, because the first estimate will be negative, which is outside the domain of the function. This explains why we used a quadratic function instead of a square-root function to estimate $\sqrt{2}$ earlier in this section.

8. The solution is $x = 0$, but using the Newton-Raphson method with any non-zero starting guess leads to problems, because the resulting sequence of approximations diverges. For example, starting with $x_0 = 1$, the method yields $x_1 = -2$, $x_2 = 4$, $x_3 = -8$, and in general, $x_{n+1} = -2x_n$. This is one of the functions mentioned earlier in this section for which the Newton-Raphson method does not work.

After reflecting on this section, including the exercises, you might conclude that the Newton-Raphson method works well when the initial guess is on a steep portion of the graph that leads directly to the zero with no intervening peaks or valleys. If the graph is not very steep at the initial guess, then problems may arise.

## HISTORY

### Who was Raphson?

Joseph Raphson was a contemporary of Newton, and published what we now call the Newton-Raphson iterative method for the approximate solution of equations in his book *Universal Analysis of Equations* in 1690. It is agreed that Newton had devised an essentially similar method in 1671, but Newton did not publish this method until 1736 in his book *Method of Fluxions*. Newton made the discovery first, but Raphson published first, so they are jointly credited.

Only scant details about Raphson's life are available, which is surprising considering the era in which he lived. He shows signs of being very well educated, but it is not clear that he attended any university. He is at least partly Jewish, and he may have had Irish background as well, although he was born and died in England, and seems to have lived there for most, if not all, of his life. Even the years of his birth and death are unclear!

Raphson was elected to the Royal Society at an unusually young age of 21, if one of the proposed years of his birth is to be believed. He communicated with the other leading mathematicians of England, including Newton.

**DIGGING DEEPER**

**The logistic map and chaos**

The logistic map can be thought of as a family of iterative procedures that satisfy the relation

$$x_{n+1} = rx_n\,(1 - x_n)$$

The logistic map has been used as a model for certain populations of living creatures, where $x_n$ represents the population in time period $n$, and the parameter $r$ is related to the ultimate population. The units used for the population could be thousands of individuals, millions of individuals, etc. One can study the mathematical properties of this family of population models without being concerned about the units.

The long-term value of the population is of interest. Does the collection of creatures eventually become extinct? Does the population reach some stable value? If so, what is the stable value?

The answers to these questions depend on the value of the parameter $r$, so each value of $r$ can be thought of as describing a different situation. Interestingly, the answers to the questions sometimes depends on the initial population, $x_0$. Sometimes the answers to these questions about long-term behaviour change significantly for very slight changes in the initial value $x_0$; this is the concept of *chaos*.

Among the most interesting values of $r$ in the family of logistic models are real numbers in the domain $0 < r \leq 4$, because values of $r$ of 4 or greater can lead to negative populations, which may not be relevant depending on the application.

You might enjoy exploring the logistic model for various values of $r$. Choose a value of $r$, then choose various starting populations (using $0 < x_0 < 1$), and then program a spreadsheet or some other favourite software to output some values of $x_n$. You might also enjoy plotting the data, with values of $n$ on the horizontal axis and values of $x_n$ on the vertical axis. Once you feel you have explored sufficiently for a specific value of $r$, try other values of $r$.

For some values of $r$, the population dies off to zero in the long term. For some other values of $r$, the population tends to $\dfrac{r-1}{r}$ in the long term. For some other values of $r$, the population eventually alternates between two values. For some other values of $r$, the population eventually alternates between four values, eight values, sixteen values, and so on. This will be very fun to explore!

Examples of chaotic systems in nature are Earth's weather and the orbits of planets. The list of applications of chaos theory is very long, and includes categories across mathematics, science, engineering, and even social sciences. To learn more about chaos, you can search on the terms chaos and dynamical systems.

## 10.6 $p$-Series

The harmonic series can be generalized to produce a family of infinite series commonly called $p$-series, where $p$ is a constant real number:

$$1 + \frac{1}{2^p} + \frac{1}{3^p} + \frac{1}{4^p} + \cdots$$

For example, for $p = 3$, this series is

$$1 + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{4^3} + \cdots$$

for $p = -2$, this series is

$$1 + 2^2 + 3^2 + 4^2 + \cdots$$

and so on. Note that none of the $p$-series, including the harmonic series (which is a $p$-series with $p = 1$), is a geometric series; they have quite different properties. In integral calculus courses it is proved that a $p$-series converges if and only if $p > 1$.

**Question:** How could you prove this? This question is well beyond the level of this book, but you might like to give it some thought. Full details to come in a future university calculus course!

This means that the harmonic series forms a sort of "boundary" series in the family of all $p$-series, in the sense that if $p \leq 1$ the $p$-series diverges, but if $p > 1$ the $p$-series diverges. The harmonic series is the limiting case; with a value of $p$ ever so slightly greater than 1, the series converges, but if $p$ is exactly 1 the series diverges.

It's typically much, much more difficult to determine an exact value for the sum of a convergent series than to determine whether or not a series converges. Even though it is known that a $p$-series converges if and only if $p > 1$, sum formulas are known only in the cases that $p$ is an even natural number.

**Question:** How is it possible to determine exact formulas for the sum of such $p$-series? (See the following paragraph.)

For odd natural numbers it's an open question, and for other real numbers it's also an open question, and presumably forbiddingly difficult to work on.

Consider a $p$-series with $p = 2$:

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots$$

Euler presented a beautiful argument to derive the sum of this convergent series. It was known in Euler's time that the sine function could be written as a power series, as we have seen earlier in this chapter:

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

Euler reasoned as follows, in a way that would not be considered rigorous nowadays, but nevertheless illustrates how creative people bend the rules to make discoveries. Euler reasoned that because the sine function has many zeros, he should be able to factor the power series given above as a *product* of linear factors, which he did (more or less) as follows:

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = (x)\left(1 - \frac{x}{\pi}\right)\left(1 + \frac{x}{\pi}\right)\left(1 - \frac{x}{2\pi}\right)\left(1 - \frac{x}{2\pi}\right)\cdots$$

because the zeros of the sine function are at $x = 0$, $x = \pm\pi$, $x = \pm 2\pi$, and so on. Factoring $x$ from the left side of the previous equation and then dividing both sides of the equation by $x$, we obtain

$$1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \cdots = \left(1 - \frac{x}{\pi}\right)\left(1 + \frac{x}{\pi}\right)\left(1 - \frac{x}{2\pi}\right)\left(1 - \frac{x}{2\pi}\right)\cdots$$

Next, we expand the product of adjacent factors on the right side of the previous line to obtain

$$1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \cdots = \left(1 - \frac{x^2}{\pi^2}\right)\left(1 - \frac{x^2}{4\pi^2}\right)\left(1 - \frac{x^2}{9\pi^2}\right)\cdots$$

The next step is to expand the entire product on the right side of the previous line. This is a bit tricky, as it's an infinite product. Nevertheless, if we proceed slowly and systematically, it can be done. First notice that there is only one constant term that results from expanding the right side of the previous line, and this term is 1. This matches the constant term on the left side of the equation, so all is well. Next, note that all of the rest of the terms in the expansion of the right side of the previous line are even powers of $x$. So let's start with the lowest even power, and collect all of the terms involving $x^2$. How are such terms produced in the expansion? Well, the only way you get an $x^2$-term is to choose a 1 from each factor except for one single factor where you choose the $x^2$ term. When you collect up all such factors, you end up with

$$-\frac{x^2}{\pi^2} - \frac{x^2}{4\pi^2} - \frac{x^2}{9\pi^2} - \frac{x^2}{16\pi^2} - \cdots$$

**Question:** Did you understand how the term in the previous line was obtained? It may take you some time, thought, and work with pencil and paper to understand it. Do take the time needed to figure this out!

But if the two sides of the original equation are to be identical, they must have equivalent terms, so the previous line must match with the $x^2$-term on the left side of the original equation:

$$-\frac{x^2}{3!} = -\frac{x^2}{\pi^2} - \frac{x^2}{4\pi^2} - \frac{x^2}{9\pi^2} - \frac{x^2}{16\pi^2} - \cdots$$

Simplifying, we obtain

$$\begin{aligned} -x^2\left(\frac{1}{3!}\right) &= -x^2\left(\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \frac{1}{16\pi^2} + \cdots\right) \\ \frac{1}{3!} &= \frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \frac{1}{16\pi^2} + \cdots \\ \frac{\pi^2}{6} &= 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots \end{aligned}$$

Because of numerical calculations, mathematicians were aware of the approximate value for the sum of the reciprocals of the squares of the natural numbers, but it was extremely surprising that this sum would have anything whatsoever to do with $\pi$. But there you have it, as a result of Euler's ingenious play.

Euler continued in the same way to collect up the $x^4$-terms on the right side of the original equation; you might like to do this for yourself, as it's great fun. (Be systematic!) The result is:

$$\frac{\pi^4}{90} = 1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots$$

Euler continued for some of the higher even powers, and you can take this further if you like this kind of fun. Here's another one of Euler's delightful manoeuvers: He took the original conclusion for the sum of the reciprocals of the squares of the natural numbers,

$$\frac{\pi^2}{6} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots$$

and deleted the odd powers on the right side to leave only the even terms:

$$\frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \cdots$$

Then Euler factored $1/2^2$ from the expression to obtain

$$\frac{1}{2^2} + \frac{1}{4^2} + \frac{1}{6^2} + \cdots = \frac{1}{4}\left(1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots\right)$$

But we know the sum of the series in parentheses on the right side of the previous line, so this gives us an expression for the sum of the reciprocals of the squares of the even natural numbers:

$$\frac{1}{2^2}+\frac{1}{4^2}+\frac{1}{6^2}+\cdots=\frac{1}{4}\left(\frac{\pi^2}{6}\right)$$
$$\frac{1}{2^2}+\frac{1}{4^2}+\frac{1}{6^2}+\cdots=\frac{\pi^2}{24}$$

Subtracting gives a result for the sum of the reciprocals of the squares of the odd natural numbers:

$$1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=\frac{\pi^2}{6}-\frac{\pi^2}{24}$$
$$1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=\frac{3\pi^2}{24}$$
$$1+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\cdots=\frac{\pi^2}{8}$$

None of these arguments is rigorous, and each requires serious justification. Nevertheless, there is a lot of room for play following the marvelous example of the great Euler. And, of course, if you're able to obtain exact expressions for the sums of the *odd* powers of the reciprocals of the natural numbers, you'll become justly famous.

## DIGGING DEEPER

### Riemann zeta function

The perspective we have taken in this section on $p$-series is that by choosing a specific value of $p$ you get a particular infinite series of numbers, and then you can study the properties of the resulting particular series. But you can also imagine defining a function $\zeta$ as the following power series

$$\zeta(x)=\frac{1}{1^x}+\frac{1}{2^x}+\frac{1}{3^x}+\cdots$$

This function is called the Riemann zeta function, and it has applications across several branches of mathematics, and also physics. Riemann discovered the function while he was studying prime numbers, and indeed the Riemann zeta function has connections to the properties of prime numbers. One of the greatest unsolved mathematical problems currently is whether the Riemann hypothesis (which is a conjecture about the zeros of the Riemann zeta function) is correct.

This subject is way beyond the scope of this book, but an interested reader could do a search on the Riemann zeta function to learn more about it and its connections with various branches of mathematics. (Traditionally, $s$ is used in place of $x$ in the formula for the Riemann zeta function, and $s$ is allowed to be a complex number, not just a real number.)

## HISTORY

### Infinite series for $\pi$

The number $\pi$ is one of the most important and widely used numbers in mathematics. It is a fact that the ratio of the circumference to the diameter is the same number for every circle, and this was the original definition of $\pi$. The number $\pi$ is irrational, which means that it cannot be expressed as a ratio of whole numbers.

Because of its importance, numerous attempts to determine a value for $\pi$ have been made over the centuries. The approximate value of $\pi$ was known in antiquity (for example, the approximate value 3 is given in the Christian Bible), but mathematicians knew that this value was not exact and strove to do better.

Better approximations from ancient times exist in an ancient Egyptian manuscript (from about 1850 BCE), which has $\pi \approx \left(\frac{16}{9}\right)^2 \approx 3.16$, and an ancient Babylonian manuscript of about the same era, which has $\pi \approx \frac{25}{8} = 3.125$. Ancient Indian astronomers from about the fourth century BCE used $\pi \approx \frac{339}{108} \approx 3.139$.

In the third century BCE, Archimedes developed an iterative approach that is astoundingly modern in spirit. He circumscribed polygons around a circle, and then inscribed polygons with the same number of sides inside the same circle, and then calculated the area of each polygon in terms of the radius of the circle. This allowed him to obtain upper and lower bounds for the value of $\pi$. By steadily increasing the number of sides of the approximating polygons, doubling the number of sides each iteration, he was able to decrease the difference between the upper and lower bounds, thereby increasing the accuracy of the approximation. His best estimate of $\pi$, using 96-sided polygons, was $\frac{223}{71} < \pi < \frac{22}{7}$, which is about $3.141 < \pi < 3.143$. This method of Archimedes was used into the 1600s (!) by mathematicians in Europe, Asia, and the Middle East to steadily improve the estimate for $\pi$.

Great improvements in estimates of $\pi$ occurred later with the use of various infinite series, especially those that derived from power series. The Indian mathematician and astronomer Madhava of Sangamagrama developed an infinite series (later rediscovered by Gregory and Leibniz in the 1600s), which can be used to determine $\pi$:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$$

Because this is an alternating series, it also provides upper and lower bounds. However, it converges very slowly, so that after about 500,000 terms one obtains only six correct digits. A series that converges much more rapidly is that of the Indian astronomer Nilakantha Somayaji, from the 1400s:

$$\pi = 3 + \frac{4}{(2)(3)(4)} - \frac{4}{(4)(5)(6)} + \frac{4}{(6)(7)(8)} - \frac{4}{(8)(9)(10)} + \cdots$$

Nowadays, other infinite series are known that converge much more rapidly, but it's interesting to note the long history of the search for infinite series (and other methods) for computing $\pi$ more accurately. You might like to play with Somayaji's series with a hand calculator to see how rapidly it converges. You might like to program your favourite software to automatically calculate $\pi$ using Somayaji's series so that you can rapidly output values for $\pi$ using any desired number of terms.

**SUMMARY**

The initial, rough conception of limit of a function that we presented earlier in this book is formalized by the idea of the limit of a sequence. The limit of a function was initially explained as a "trend" in function values as some $x$-value approached some other $x$-value, and we can understand this trend as the limit of a corresponding sequence of function values.

In Chapter 11 we present the formal definition of the limit of a function, which is the current state of the art. Although it is not formulated in terms of sequences, other important mathematics concepts are defined in terms of sequences and their limits. For example, the definite integral, which you will learn about in a future calculus course, is defined in terms of the limit of a certain sequence.

We discussed sequences and series of numbers in this chapter, and defined the sum of a series to be the limit of the sequence of partial sums of the series, if this limit exists. We then discussed power series, which are (loosely speaking) like polynomial functions, only with an infinite number of terms. Power series are widely used in mathematical and scientific applications, for instance in solving differential equations, which are fundamental in mathematics and physics.

Finally, we discussed iterative procedures, which are essentially sequences of approximations. For example, the Newton-Raphson method is an iterative method for approximating the solution of an algebraic equation.

# Chapter 11

# Theory, Part 1: The Formal Definition of a Limit

**OVERVIEW**

In this section we study the state-of-the-art, best available conception of limit. The vague definition we've used up to now is OK for starting out, but to work out limits in truly difficult cases, we need a better definition. This better, more precise, definition is also essential for proving theorems about limits, and because just about all concepts in calculus (and advanced analysis) are based on limits, the more precise definition of a limit is essential for proving all kinds of theorems in calculus (and advanced mathematical analysis) as well.

Not all first-year university courses tackle the precise definition of limit; some prefer to leave it for a second-year course. You will need it if you are interested in going on to higher levels of mathematics, and if you are interested in the logical structure of calculus.

In this chapter we provide a step-by-step, intuitive introduction to the precise definition of the limit. We also present fully worked out examples of calculating limits using the precise definition. These step-by-step examples are accompanied by descriptions of the thinking process that are meant to demystify what is typically a very challenging process for first-year university students. As usual, careful study and repetition is the key to mastery.

So far we've used an informal conception of limit to perform limit calculations. It has served us well, for the cases we've looked at, but for more complicated limits, and for proving all kinds of theorems, we need a more precise tool. Describing this more precise tool, and getting good practice in its use, is the point of this chapter.

This is an unusually long chapter (especially if combined with the following theory chapter), but deservedly so, because the formal definition of a limit is notoriously challenging to understand and apply. We provide extensive discussion in this chapter, along with examples that are worked out in a lot of step-by-step detail. Carefully going over the discussion multiple times, and carefully working through the examples and exercises will help you to understand the formal definition of a limit and apply it successfully. The spirit of argument in this chapter is at the core of higher mathematical analysis, so anyone aspiring to higher levels of learning mathematical analysis should study this chapter and the following one very carefully.

As we discussed previously in this book, the early slope calculations by Newton, Leibniz, their contemporaries, and their predecessors, involved (in our modern notation) factoring $h$ from the numerator of the expression, then dividing numerator and denominator by $h$ (which is valid provided that $h \neq 0$), and then finally setting $h = 0$ to obtain the limit. They were well aware of the contradiction inherent in the last two steps, and they tried to justify it as best they could, but their arguments were not very convincing.

Newton and Leibniz's attempt to make sense of this type of calculation in the late 1600s was to (in effect) call $h$ an "infinitesimal," a new sort of number, which was smaller than any non-zero positive number, but not quite zero either! This justified ignoring it in the last step of the calculation (setting it equal to zero, in effect), yet allowed one to divide by it. Berkeley ridiculed this by questioning the existence of these purported infinitesimals, saying they were akin to "ghosts of departed quantities." Ouch.[1]

By the mid-1700s, Denis Diderot embarked on a massive project: The construction of the first encyclopedia in history. The literal meaning of encyclopedia is "circle of knowledge": This was an attempt to enclose all important knowledge between the covers of a single set of books. The mathematician and physicist d'Alembert wrote the article on calculus for Diderot's encyclopedia, and he stated that the foundations of the subject were still not finalized, but he had a strong feeling that mathematicians would be able to sort things out properly using the newly developed concept of limit.

Around the same time, mathematicians were struggling to properly define the concept of function. Traditionally, it was thought that functions needed to be continuous, and needed to be described by a single formula that was valid for the entire domain of the function. The work of Fourier in the early 1800s called this limited view into question, and discontinuous functions and other, more exotic types of functions began to become respectable. This influenced the development of the limit concept in the following way: Although functions were often used to model the positions of moving objects, so that it made sense to speak of the $x$-value "getting closer and closer to some number" (as we continually did earlier in this book), mathematicians began to take a more abstract approach. A function just *is*, they reasoned; nothing is moving anywhere. Nothing is "getting closer and closer" to anything else; the values are just sitting there. There ought to be a way of calculating limits that respects this way of looking at a function.

For example, Dirichlet's function cannot possibly describe the motion of any real object, but according to the modern definition it is considered to be a legitimate function:

$$D(x) = \begin{cases} 1 & \text{if} \quad x \in \mathbb{Q} \\ 0 & \text{if} \quad x \notin \mathbb{Q} \end{cases}$$

In words, the value of Dirichlet's function is 1 if $x$ is a rational number, and 0 if $x$ is an irrational number. You should devote a little bit of time to trying to visualize Dirichlet's strange function, just so that you can appreciate how strange (and impossible to visualize) it is. You should also convince yourself that it does indeed satisfy the properties of a function, according to our modern definition of a function, and so therefore really is a function. Graphing Dirichlet's function is, of course, impossible.

All this led to the currently-accepted precise definition of the limit. The idea of using inequalities to make the calculation of a limit more rigorous is due (independently) to Bolzano and Cauchy in the early 1800s, and Cauchy in particular was instrumental in bringing a higher standard of rigour in mathematical argumentation to the entire community. The currently-accepted definition was formulated by Karl Weierstrass (who also introduced the current notation for the limit of a function), and published by one of his students, Heinrich Eduard Heine, in 1872. Here is a currently accepted version of this definition:

[1] It's interesting that in the 1960s, nearly 300 years after the work of Newton and Leibniz, Abraham Robinson was able to rehabilitate infinitesimals to respectability in his alternative foundation of calculus, which he called the theory of nonstandard analysis.

**DEFINITION 12**

**Limit of a function**

The function $f$ has a limit $L$ as $x$ approaches $a$, in symbols

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

provided that for each positive real number $\varepsilon$ (that is, $\varepsilon > 0$), there exists a real number $\delta$ such that

$$\text{if} \quad 0 < |x - a| < \delta \quad \text{then} \quad |f(x) - L| < \varepsilon$$

The symbols $\varepsilon$ and $\delta$ that appear in the definition of limit are virtually universal, and the style of arguments using inequalities based on it are therefore called $\varepsilon$-$\delta$ arguments. They represent the gold-standard of argumentation in mathematical analysis.

Most students find it challenging to understand the precise definition of limit, and challenging to learn how to use it. But this is perfectly normal, and nothing to feel bad about. Remember that it took the finest mathematical minds in the world two centuries to sort this out, so be patient with yourself; if you work at it, you will be able to understand the precise definition of limit with time and practice.

To explain this definition, we'll attempt to connect it with our initial conception of limit; that is, the sense that $\lim_{x \to a} f(x) = L$ means that the values of $f(x)$ get closer and closer to $L$ as $x$ gets closer and closer to $a$. From this perspective, the precise definition of the limit formalizes the notion of "getting closer and closer" with precision and without ambiguity.

Consider Figure 11.1, which is a graph of the function $f(x) = 2x + 1$.

From what we have learned previously in this book, we can conclude that

$$\lim_{x \to 3} f(x) = 7$$

In the imprecise language we have been using in this book so far, we would say that the previous limit statement means that as $x$ gets closer and closer to 3, the corresponding function values get closer and closer to 7. Let's now explain how to view this situation using the precise definition of the limit.

The precise definition of the limit states that 7 really is the limit if for each positive value of $\varepsilon$, there exists a value of $\delta$ such that a certain property is satisfied. The figure illustrates this for a particular value of $\varepsilon$, namely $\varepsilon = 2$. As we will explain in detail shortly, any value of $\delta$ that is less than 1 will work; in the figure, the value $\delta = 0.75$ is chosen.

Let's write down the precise definition of the limit for the situation illustrated in the figure: The function $f(x) = 2x + 1$ has a limit 7 as $x$ approaches 3, in symbols

$$\lim_{x \to 3} f(x) = 7$$

provided that for each positive real number $\varepsilon$ (that is, $\varepsilon > 0$), there exists a real number $\delta$ such that

$$\text{if} \quad 0 < |x - 3| < \delta \quad \text{then} \quad |f(x) - 7| < \varepsilon$$

Rewriting this in terms of the figure, the limit of the function as $x$ approaches 3 is 7 provided that for each $\varepsilon > 0$ there exists a positive value of $\delta$ such that for all the $x$-values in the blue band defined by $\delta$ (except we don't care about $x = 3$), the corresponding function values lie in the red band defined by $\varepsilon$.

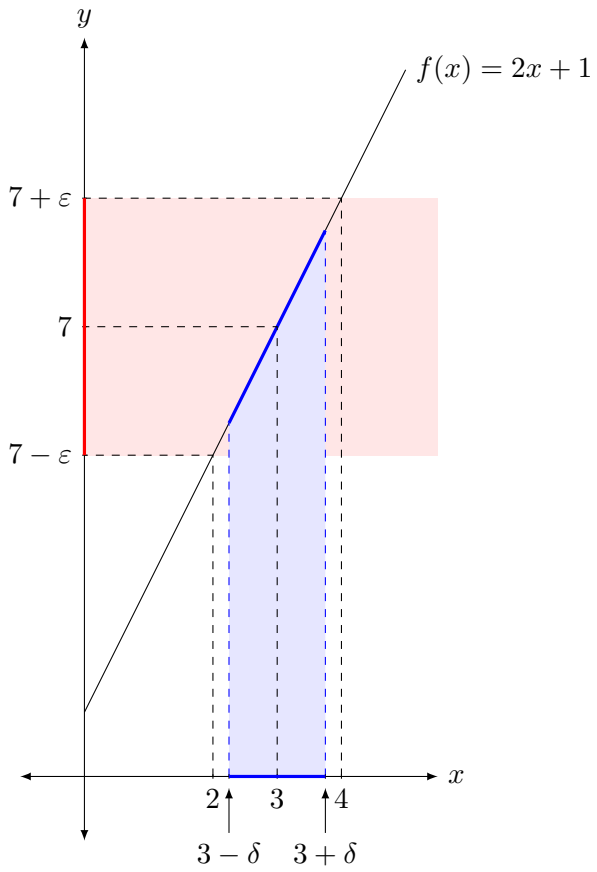



Figure 11.1: This figure illustrates some of the reasoning that is used to show that the limit of the function as $x$ approaches 3 is equal to 7, using the precise definition of the limit. See the text for the complete argument. The figure illustrates the situation for $\varepsilon = 2$. In this case, $\delta$ can be chosen to be any positive number less than 1; the figure illustrates a value of $\delta = 0.75$.

**Question:** Is this clear from the figure?

It is unlikely that this will be clear unless you go through each element of the statement, bit by bit. To begin, is it clear that $|x - 3| < \delta$ corresponds to the blue region of the $x$-axis? Because $\delta = 0.75$ in the figure, we can make this inequality more specific: $|x - 3| < 0.75$. In other words, this means all of the values along the $x$-axis that are within a distance of 0.75 units of $x = 3$. This means all the $x$-values that are in the interval between 2.25 and 3.75, not including the endpoints. Is it now clear that the condition $|x-3| < 0.75$ means all of the $x$-values in the blue band along the $x$-axis in the figure? If not, you may wish to substitute some selected $x$-values into this condition, observing that $x$-values in the blue band satisfy the condition, but $x$-values outside the blue band do not satisfy the condition. For example, for $x = 2.8$,

$$|x - 3| = |2.8 - 3| = |-0.2| = 0.2$$

and this value is indeed less than 0.75. On the other hand, for $x = 1.6$,

$$|x - 3| = |1.6 - 3| = |-1.4| = 1.4$$

and this value is **not** less than 0.75. Continue to test a few other values both inside the blue interval and outside it until it becomes clear to you that the condition $|x-3| < 0.75$ is represented graphically by the blue band along the $x$-axis.

Next, do the same with the condition $|f(x)-7|<\varepsilon$. For the value of $\varepsilon=2$ that we have chosen for the figure, this condition is $|y-7|<2$. In other words, this means all of the values along the $y$-axis that are within a distance of 2 units of $y=7$. This means all the $y$-values that are in the interval between 5 and 9, not including the endpoints. This corresponds to the red band of values along the $y$-axis in the figure. If this is not yet clear to you, then substitute some values of $y$ into the condition to become familiar with which values satisfy the conditions and which do not.

Continuing, let's now understand the difference between the condition that we studied earlier, $|x-3|<0.75$, and the slightly different condition that is actually in the definition of the limit, which is $0<|x-3|<0.75$. This latter condition means all of the $x$-values that satisfy **both** the condition $|x-3|<0.75$ that we have already studied **and** the condition $0<|x-3|$. Let's now study this; which values in the blue interval along the $x$-axis also satisfy the condition $0<|x-3|$? In other words, which of the values in the blue interval along the $x$-axis has a distance to 3 that is greater than 0? If you are standing at any point in the blue interval along the $x$-axis your distance to 3 will be greater than 0, unless of course you are standing right at the value $x=3$. Thus, the value $x=3$ does not satisfy the new condition, but every other value in the blue band along the $x$-axis does satisfy the new condition.

**Question:** Have you understood the previous paragraphs? If not, take your time, and work with pencil and paper, sketching and calculating. You will get this with time, thought, and work. If you have not understood after a reasonable time, carry on reading, but circle back to this point at a later time after you have read ahead a little.

Now that we have understood these elements, let's go back and consider the complete limit statement: 7 really is the limit of the function as $x$ approaches 3 provided that for **each** value of $\varepsilon>0$, there exists a value of $\delta>0$ such that for each value of $x$ in the blue band along the $x$-axis (excluding $x=3$), the corresponding function values lie within the red band along the $y$-axis. You can see that this is true, at least for the value $\varepsilon=2$, from the figure. Note that the part of the graphed line that is drawn in blue lies completely within the red shaded band.

The definition of the limit states that 7 really is the limit of the function as $x$ approaches 3 provided that for each value of $\varepsilon>0$, there exists a value of $\delta>0$ such that for each value of $x$ in the blue band along the $x$-axis (excluding $x=3$), the corresponding function values lie within the red band along the $y$-axis. The figure shows just one value of $\varepsilon$, so once we have fully understood the situation for this one value of $\varepsilon$ we should consider other values of $\varepsilon$. The condition states that 7 is the limit provided that for each value of $\varepsilon>0$, there exists **a** value of $\delta$ that works; for the particular value $\varepsilon=2$, we have demonstrated that the value $\delta=0.75$ "works" in the sense that it satisfies the specified condition. Is it clear that other values of $\delta$ also work? The condition only requires that **one** suitable value of $\delta$ exists, but by studying the figure we can see that there are an infinite number of $\delta$-values that work. For example, choosing a (positive) value of $\delta$ that is smaller than 0.75 also works; a smaller value just makes the blue interval along the $x$-axis smaller, which means that the stretch of the graphed line that is coloured blue is also smaller, but it still lies entirely within the red shaded band, so the limit condition is still satisfied. Making the value of $\delta$ slightly greater is also fine, provided that it is not too big. Once the value of $\delta$ reaches 1, the limit condition fails, and the limit condition will continue to fail if $\delta>1$. This is illustrated in Figure 11.2 for a value of $\delta=1.25$. Note that in this case the corresponding band of function values **does not** lie entirely within the red band, so the limit condition is not satisfied. This is the case for all values of $\delta\geq 1$.

Let's sum up what we have discussed so far. For a particular value of $\varepsilon$, namely $\varepsilon=2$, we have demonstrated that it is indeed possible to select a value of $\delta$, namely any value between 0 and 1, not inclusive, that satisfies the limit definition. In order to really be convinced that the limit of the function as $x$ approaches 3 is 7, we would have to show that the same limit condition is satisfied

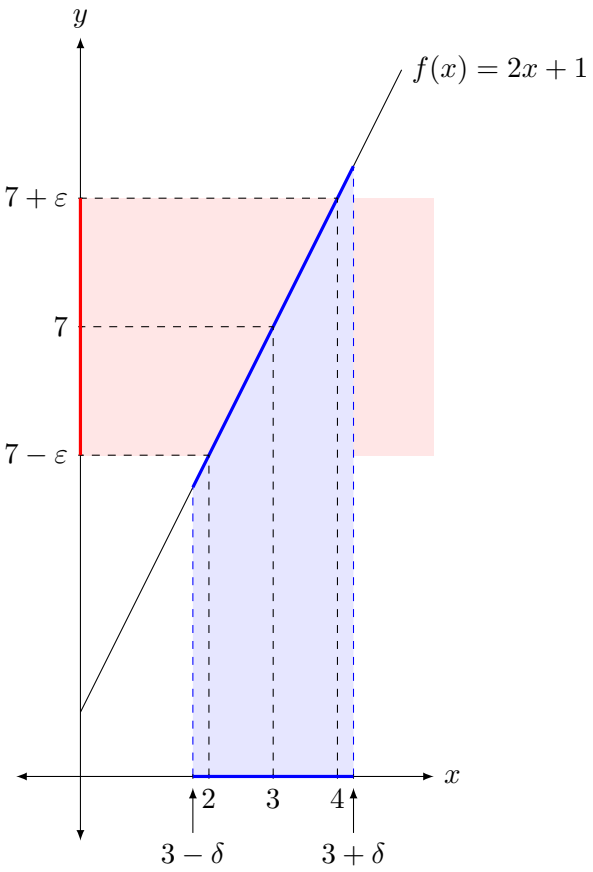



Figure 11.2: This figure continues the discussion of the reasoning that is used to show that the limit of the function as $x$ approaches 3 is equal to 7, using the precise definition of the limit. See the text for the complete argument. The figure illustrates the situation for $\varepsilon = 2$. In this case, $\delta$ can be chosen to be any positive number less than 1; the figure illustrates a value of $\delta = 1.25$. This value is **not** suitable for proving that 7 is the limit, because it does not satisfy the condition stated in the precise definition of the limit. The point is that values of $\delta$ that are less than 1 satisfy the condition, but values that are greater than or equal to 1 **do not** satisfy the condition.

for **each** positive value of $\varepsilon$. That is, we would have to show that for each positive value of $\varepsilon$, it is indeed possible to choose a value of $\delta$ that would satisfy the limit condition.

Based on the graph in Figure 11.1, does this seem possible? Yes, doesn't it? After all, if you shrink the red band vertically (that is, by using a smaller value of $\varepsilon$), we should still be able to choose a value of $\delta$ that will work, although the value of $\delta$ might have to be smaller. Consult Figure 11.3 and study the situation for $\varepsilon = 1$; is it clear that the limit condition is satisfied provided that you choose a value of $\delta$ that is less than 0.5? The figure illustrates a choice of $\delta = 0.25$; notice that for this choice, the band of function values for all $x$-values within the blue band lies entirely within the red zone. Thus, for $\varepsilon = 1$, the limit condition is satisfied for this choice of $\delta$. Is it clear that any smaller positive value of $\delta$ will also work? A smaller value of $\delta$ just means that the blue band will be smaller, and so the corresponding function values will still lie within the red zone.

For larger values of $\varepsilon$, it is "easier" to find values of $\delta$ that will satisfy the limit condition. Can you see this from the graphs in the previous three figures? Larger values of $\varepsilon$ mean a wider red zone, so even a wider blue band will result in function values that lie entirely within the red zone.

In summary, it seems possible to satisfy the limit condition no matter which value of $\varepsilon$ is given. No matter how narrow the red zone we are presented with, it seems to be possible to choose a small

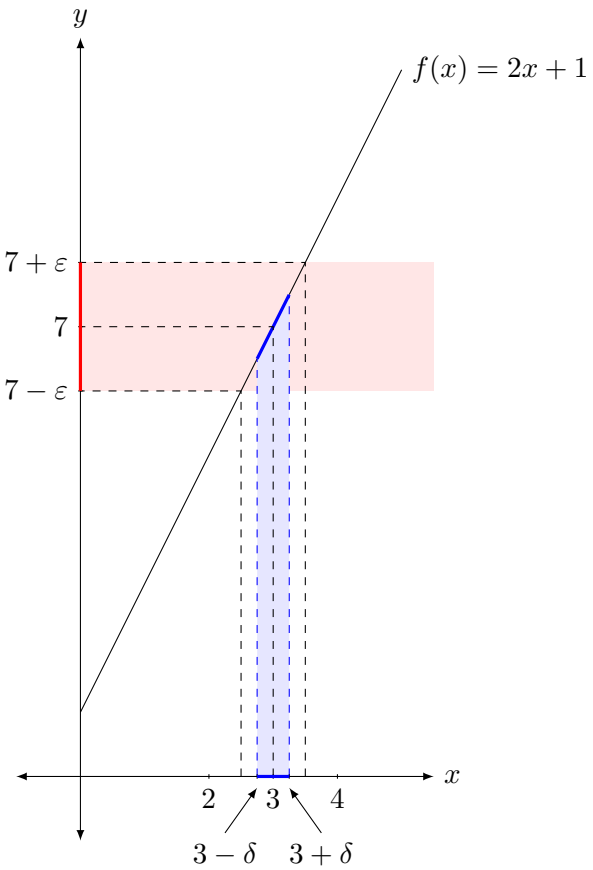



Figure 11.3: This figure continues the discussion of the reasoning that is used to show that the limit of the function as $x$ approaches 3 is equal to 7, using the precise definition of the limit. See the text for the complete argument. The figure illustrates the situation for $\varepsilon = 1$. In this case, $\delta$ can be chosen to be any positive number less than 0.5 in order to satisfy the limit condition; the figure illustrates a value of $\delta = 0.25$.

enough blue band so that all the corresponding function values lie within the red zone.

Does this convince you that the limit of the function as $x$ approaches 3 really is 7? Perhaps yes, perhaps no, but in any case, we really should provide a symbolic proof using algebra. It is very easy for humans to fool themselves, and in the history of mathematics there are many famous examples of supposed facts that were thought to be true for a while, until someone thinking more carefully showed that they were in fact not true. Thus, although geometric reasoning (including the drawing of figures) is a marvelous aid to understanding, it should not substitute for symbolic proofs, which have been the standard of rigour in mathematics for centuries. We'll provide a symbolic proof shortly, but before we do so, let's continue to study the figure to make sure that we have fully understood the situation.

Our earlier, rough sense of limit was that if

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

this means that as $x$ gets closer and closer to $a$, the function values get closer and closer to $L$. The precise definition of limit can also be interpreted in this way, but without the movement ("approaching") that is explicit in our previous treatment. The "for all $\varepsilon > 0$" part of the definition allows us simulate the "approaches $L$" part of our previous approach to limits. That is, if the limit

really is $L$, then if we make $\varepsilon$ progressively smaller, the values of $\delta$ that work also get progressively smaller, but they still exist.

You can think of the precise definition of limit as a sort of game. You propose that a certain limit exists and has value $L$. Your opponent then challenges you by giving you a positive value of $\varepsilon$, and your task is to come up with a value of $\delta$ that satisfies the limit condition. If you are able to come up with a value of $\delta$ that works, then you win. It doesn't matter that an infinite number of $\delta$-values work, all you need to do is come up with a single value that works and you win. If you always win, no matter which value of $\varepsilon$ your opponent challenges you with, then your guess about that the limit exists and is equal to $L$ is correct.

If for even one value of $\varepsilon$, it's not possible for you to come up with a value of $\delta$ that works, then you lose the game. Maybe you were wrong with your guess about the value of the limit; maybe there is no limit. In either case, the assertion that the limit exists and is equal to $L$ is incorrect.

You can even think of this game being automated, like an e-sport. The user gets to input the value of $\varepsilon$ that they choose, and then the program should output a value of $\delta$ that works, if one is possible. Imagine programming such a game; you would need an algorithm to respond to every possible user input by calculating a suitable value of $\delta$ and then outputting the result. Think about this.

**Question:** How would you program such a game? After reflecting on this question for a while, you will be prepared for the proofs that we present later in this chapter, because the thought processes behind the proofs are the same as those needed to program the computer game algorithm.

We'll get to the proofs shortly. Before we do so, it's worth emphasizing again that our previous conception of the limit, with quantities "getting closer and closer to" various values, has been replaced in the formal definition of the limit by the freedom of choice about the value of $\varepsilon$ that our "opponent" has in the e-sport game. If the opponent tries various values of $\varepsilon$, making the value smaller each try, then the allowable values of $\delta$ that we can respond with are more and more restricted each try. This allows us to interpret the process of successively playing the game by thinking in terms of values getting closer and closer to target values, if we wish, but there is no motion inherent in the definition. Nothing is moving anywhere; the opponent just selects a positive value of $\varepsilon$, and then we have to respond with a suitable value of $\delta$ if we wish to win the game.

**Question:** Have you taken the time to work through the previous pages several times in preparation for the upcoming symbolic proofs? Do this!

Once you have gone over the previous pages, ideally several times on different days, and carefully examined the previous figures, you'll be ready to absorb the formal proof that

$$\lim_{x \to 3} f(x) = 7$$

for the function $f(x) = 2x + 1$, using the formal definition of the limit. In order to complete the proof, we have to demonstrate that no matter which positive value of $\varepsilon$ is given, we are able to counter with a value of $\delta$ that satisfies the limit condition. There are an infinite number of choices for $\varepsilon$, so how will we be able to efficiently counter each given value of $\varepsilon$ with a suitable value of $\delta$? It would be a pain to have to figure it out from scratch every time; it would be a lot better if we had a formula that connected the two values, so that when we are presented with a value of $\varepsilon$, all we have to do is run it through our formula and then we would know which values of $\delta$ will work. Such a formula would allow the e-sport game to work, as mentioned earlier.

It might be helpful to tabulate the results we have obtained so far for the two given choices of $\varepsilon$:

| given value of $\varepsilon$ | values of $\delta$ that work |
|---|---|
| $\varepsilon = 2$ | $0 < \delta < 1$ |
| $\varepsilon = 1$ | $0 < \delta < 0.5$ |

Interesting, isn't it? Could it be that the values of $\delta$ that work must be less than exactly half the given value of $\varepsilon$? It would be good for you to go back to one of the three previous figures and carefully examine it. It does indeed seem so, doesn't it? It has to be connected to the fact that the slope of the graph is 2, right? Imagine if the slope of the line were different; how would that affect the relative sizes of the red and blue strips? It would be good for you to draw a few sketches and play with this idea so that you will understand it thoroughly.

**Question:** Have you taken the time to play as indicated in the previous paragraph? Do this!

Now that you have played with this sufficiently, it is time to go through the formal proof. The steps in the formal proof of a limit are always the same when using the precise definition of a limit:

## KEY CONCEPT

**Steps in a formal proof of the limit of a function**

- Guess what the limit is.
- Figure out the relation between $\delta$ and $\varepsilon$. That is, state the values of $\delta$ that work for a given value of $\varepsilon$.
- Show that your way of choosing $\delta$ for a given $\varepsilon$ really does work; i.e., show that for an arbitrary given positive value of $\varepsilon$ the limit condition is satisfied for your choice of $\delta$.

Note that the proof method we are about to present does not tell you what the limit is. You have to figure that out in some other way. Once you have figured out what the limit is, the method we are about to present will verify that your guess is correct. If you guess the value of the limit incorrectly, then your attempt at a proof will fail (as we shall see in an example later).

**Question:** What do you think will happen in the formal proof to make it fail if you guess the limit incorrectly? That is, how will you be able to tell that the proof fails? This will be discussed in an example, but it's worth keeping this in the back of your mind as you go through the successful proofs. That is, as you work through the successful proofs, ask yourself how you would recognize the proof failing if you had guessed the wrong limit.

## EXAMPLE 26

**Proving a limit using the formal definition**

Use the formal definition of limit to prove that $\lim_{x\to 3} f(x) = 7$ for the function $f(x) = 2x + 1$.

## SOLUTION

Having studied this situation for several pages now, we expect that the limit is 7, and we conjecture that choosing $\delta = \varepsilon/2$ will do the job. Let's prove this.

For each value of $\varepsilon > 0$, choose $\delta = \varepsilon/2$. Consider the values of $x$ for which

$$0 < |x - 3| < \delta$$

That is,

$$0 < |x - 3| < \frac{\varepsilon}{2}$$

Multiplying each term on the previous line by 2, it follows that the next inequality is valid for the same values of $x$:

$$0 < 2|x - 3| < \varepsilon$$

This means that the next inequality is also satisfied for the same values of $x$:

$$0 < |2x - 6| < \varepsilon$$

The next inequality is equivalent to the previous one, and so is also satisfied for the same values of $x$:

$$0 < |2x + 1 - 7| < \varepsilon$$

In other words, the next inequality is also satisfied for the same values of $x$:

$$0 < |f(x) - 7| < \varepsilon$$

And the next inequality is also satisfied for the same values of $x$:

$$|f(x) - 7| < \varepsilon$$

This completes the proof.

The reason the proof is complete is that we have shown that for each positive value of $\varepsilon$, there exists a positive value of $\delta$, namely $\delta = \varepsilon/2$, such that the values of $x$ for which $0 < |x - 3| < \delta$ are the same as the values of $x$ for which $|f(x) - 7| < \varepsilon$. In other words, for each $\varepsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - 3| < \delta$, then $|f(x) - 7| < \varepsilon$.

By the formal definition of limit, this proves that

$$\lim_{x \to 3} f(x) = 7$$

Having worked through the previous example, it's worthwhile writing out the proof for yourself, line-by-line, with $\varepsilon$ replaced by 2 and $\delta$ replaced by 1. Then compare your work to Figure 11.1. Notice that the value of $\delta$ illustrated in Figure 11.1 is less than 1. This is a reminder that any positive value of $\delta$ less than $\varepsilon/2$ will work just as well as 1 in the proof. Once you have digested this, then write the proof out a second time, but this time replace $\varepsilon$ by 1 and replace $\delta$ by 0.5. Again, note the value of $\delta$ illustrated in Figure 11.3 is less than 0.5. An infinite number of $\delta$ values will work, provided that they are all less than $\varepsilon/2$. Choosing any valid value of $\delta$ will get the proof to work.

You might like to write the proof out again a few times, each time choosing different values of $\varepsilon$, and sketching a graph labelled like Figure 11.1, with red and blue bands. Doing this will help you understand the ideas behind the formal proof.

**Question:** Have you done the work described in the previous two paragraphs? Doesn't it help enormously to understand the formal proof?

The last sentence of the previous example can be paraphrased by saying that $x$ values that are close to 3 lead to $y$-values (i.e., function values) that are close to 7. Although this is a vague statement ("close" is imprecise), it is a useful way of connecting the new precise concept of limit with the earlier, vaguer one. And "close" can be made precise; in fact it is stated very precisely in

the last line of the previous example. Along the $x$-axis, "close to 3" means "within a distance $\delta$ of 3," and along the $y$-axis, "close to 7" means "within a distance $\varepsilon$ of 7."

Recall that one of the primary purposes of limits is to calculate slope values for the graphs of functions. Recall that in such calculations, a graph of the slopes of secant lines sketched from a particular point on the graph of a function has a hole discontinuity. Therefore, to be effective for its intended purpose, limit calculations must ignore the actual function value at the point of interest. This explains why in the precise definition of the limit we use

$$0 < |x - a| < \delta$$

instead of

$$|x - a| < \delta$$

When calculating the limit of a function at a particular point, we're not allowed to care about the actual function value at that point, because for a very important case of interest there will be no function value at that point. By excluding the point of interest from the limit definition, we will be able to apply the definition effectively even where there is a hole discontinuity.

In the next example we apply the definition of limit in another simple situation. Consider the same function as before, but calculate the limit at another point.

## EXAMPLE 27

**Proving a limit using the formal definition**

Determine $\lim_{x\to 2}(2x + 1)$, and then use the formal definition of limit to prove your result.

## SOLUTION

The function $f(x) = 2x + 1$ is continuous for all values of $x$, so we know from our previous work that we can determine the limit by substitution. The result is $f(2) = 2(2) + 1 = 5$. Let's prove this using the formal definition of limit. Consider Figure 11.4 as a guide.

For each value of $\varepsilon > 0$, choose $\delta = \varepsilon/2$. We make this guess based on our earlier work with this function.

Consider the values of $x$ for which

$$0 < |x - 2| < \delta$$

That is, $0 < |x - 2| < \dfrac{\varepsilon}{2}$.

Multiplying each term on the previous line by 2, it follows that the next inequality is valid for the same values of $x$:

$$0 < 2|x - 2| < \varepsilon$$

This means that the next inequality is also satisfied for the same values of $x$:

$$0 < |2x - 4| < \varepsilon$$

The next inequality is equivalent to the previous one, and so is also satisfied for the same values of $x$:

$$0 < |2x + 1 - 5| < \varepsilon$$

In other words, the next inequality is also satisfied for the same values of $x$:

$$0 < |f(x) - 5| < \varepsilon$$

And the next inequality is also satisfied for the same values of $x$:

$$|f(x) - 5| < \varepsilon$$

This completes the proof. We have shown that for each $\varepsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - 2| < \delta$, then $|f(x) - 5| < \varepsilon$. By the formal definition of limit, this proves that

$$\lim_{x \to 2} f(x) = 5$$

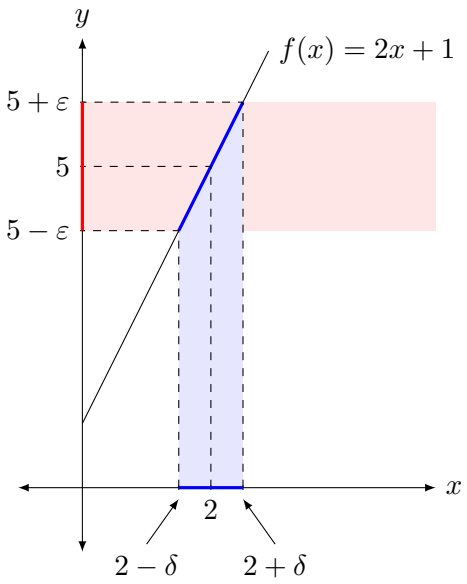



Figure 11.4: The figure illustrates the maximum value of $\delta$ that will work in a proof that the limit of the function $f(x) = 2x + 1$ as $x$ approaches 2 is 5 using the precise definition of limit. All smaller positive values of $\delta$ also work. See the example above for the complete argument.

Earlier we discussed the possibility that the slope of the graph of the function influences the range of choices for $\delta$ that work for a given $\varepsilon$. If you played with this thought by sketching some graphs you may have agreed with this. Completing the following exercises will reinforce this idea.

**EXERCISES**

(Answers at end.)

Guess each limit. Then use the precise definition of limit to prove that your guess is correct. Illustrate your work by sketching a graph in each case.

1. $\lim_{x\to 1}(3x+2)$
2. $\lim_{x\to 2}(4x+1)$
3. $\lim_{x\to 3}(0.2x-3)$
4. $\lim_{x\to 0}(0.5x+4)$
5. $\lim_{x\to -1}(2x+9)$
6. $\lim_{x\to -2}(-2x-1)$
7. $\lim_{x\to -3}(-3x-1)$
8. $\lim_{x\to 0}(6)$

---

Answers: 1. limit is 5; choose $\delta = \varepsilon/3$ for the proof, but smaller positive values also work

2. limit is 9; choose $\delta = \varepsilon/4$ for the proof, but smaller positive values also work

3. limit is $-2.4$; choose $\delta = 5\varepsilon$ for the proof, but smaller positive values also work

4. limit is 4; choose $\delta = 2\varepsilon$ for the proof, but smaller positive values also work

5. limit is 7; choose $\delta = \varepsilon/2$ for the proof, but smaller positive values also work

6. limit is 3; choose $\delta = \varepsilon/2$ for the proof, but smaller positive values also work

7. limit is 8; choose $\delta = \varepsilon/3$ for the proof, but smaller positive values also work

8. limit is 6; all positive values of $\delta$ will work in the proof

Now that you have practicing proving various limits in simple situations, to internalize the process, let's now move on to some more challenging situations. The following example is exactly like the previous few examples; in fact it generalizes them.

## EXAMPLE 28

**Proving a limit using the formal definition**

Determine $\lim_{x\to a}(mx+b)$, and then use the formal definition of limit to prove your result.

## SOLUTION

The function $f(x) = mx + b$ is continuous for all values of $x$, so we know from our previous work that we can determine the limit by substitution. The result is $f(a) = ma + b$. Let's prove this using the formal definition of limit.

For each value of $\varepsilon > 0$, choose $\delta = \varepsilon/m$. Is this guess reasonable based on the practice exercises that you completed earlier?

Consider the values of $x$ for which

$$0 < |x - a| < \delta$$

That is, $0 < |x - a| < \dfrac{\varepsilon}{m}$.

Multiplying each term on the previous line by $m$, it follows that the next inequality is valid for the same values of $x$:

$$0 < m|x - a| < \varepsilon$$

This means that the next inequality is also satisfied for the same values of $x$:

$$0 < |mx - ma| < \varepsilon$$

The next inequality is equivalent to the previous one, and so it is also satisfied for the same values of $x$:

$$0 < |mx + b - ma - b| < \varepsilon$$

In other words, the next inequality is also satisfied for the same values of $x$:

$$0 < |f(x) - (ma + b)\,| < \varepsilon$$

And the next inequality is also satisfied for the same values of $x$:

$$|f(x) - (ma + b)\,| < \varepsilon$$

This completes the proof. We have shown that for each $\varepsilon > 0$, there exists a $\delta > 0$ such that if $0 < |x - a| < \delta$, then $|f(x) - (ma + b)\,| < \varepsilon$. By the formal definition of limit, this proves that

$$\lim_{x\to a} f(x) = ma + b$$

Does the choice of $\delta$ in this example make sense compared to the choices you made in the previous exercise set? Can you sketch a graph in this case that makes the choice clear?

Now let's discuss functions that have jump discontinuities. Consider the function

$$f(x) = \begin{cases} -1 & \text{if} \quad x < 3 \\ 1 & \text{if} \quad x \geq 3 \end{cases}$$

which is illustrated in Figure 11.5.

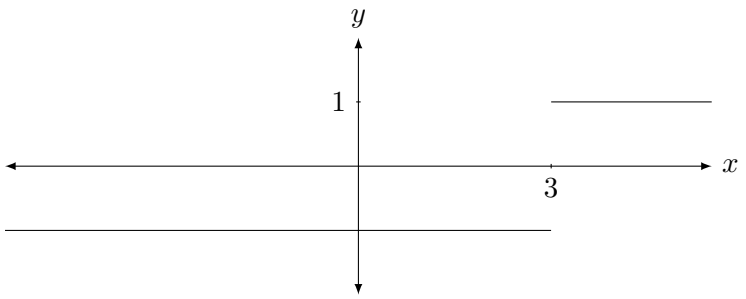



Figure 11.5: The figure illustrates a function with a jump discontinuity at $x = 3$. Because of the jump discontinuity, the limit of the function as $x \to 3$ does not exist.

Recall from our previous work earlier in this book that the limit of the function in Figure 11.5 as $x \to 3$ does not exist. We described this earlier in this book by noting that

$$\lim_{x \to 3^+} f(x) = 1 \qquad \text{and} \qquad \lim_{x \to 3^-} f(x) = -1$$

Because the left limit is not equal to the right limit, the limit does not exist. It's possible to understand that the limit does not exist from the perspective of the precise definition of the limit; study Figure 11.6.

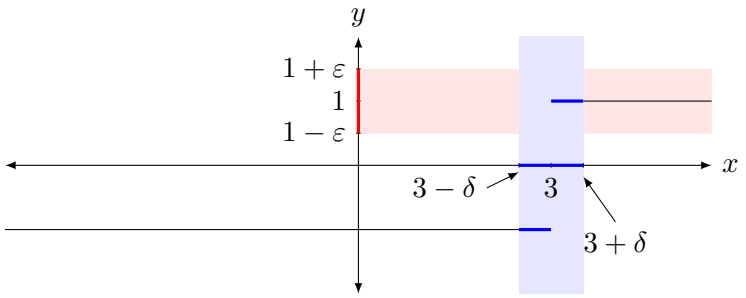



Figure 11.6: The figure illustrates a function with a jump discontinuity at $x = 3$. Because of the jump discontinuity, the limit of the function as $x \to 3$ does not exist. By studying the red and blue strips you will be able to understand from the perspective of the precise definition of the limit why this limit does not exist.

Suppose that someone tells you that the limit of the function in the previous figure as $x \to 3$ exists and is equal to 1. If you try to prove this using the precise definition of limit, you will soon see that this is not possible. Consult Figure 11.6. For the value of $\varepsilon$ shown in the figure, is it possible to choose a value of $\delta$ such that for all $x$ values that satisfy $0 < |x - 3| < \delta$, the corresponding function values satisfy $|f(x) - 1| < \varepsilon$? In other words, is it possible to choose a vertical blue strip of some width, centred at $x = 3$, such that for all $x$-values within the blue strip the corresponding function values are within the red horizontal strip?

**Question:** Is it clear from the figure that this is impossible?

The blue strip includes function values from both branches of the function, and the ones on the lower branch lie outside the red strip. No matter how thin you make the blue strip, it will always contain some function values on the lower branch of the function, which means that these function values will lie outside the red strip. Therefore, using the precise definition of the limit, it will be impossible to prove that 1 is the limit.

Sure, if you make the red strip wide enough (i.e., choose $\varepsilon$ large enough) it will include all function values, and then any value of $\delta$ will work. But remember that the precise definition of limit requires that the limit condition be satisfied **for each positive value of** $\varepsilon$. We have no control over the value of $\varepsilon$. To prove that a certain value is the limit, the limit condition must be satisfied no matter how small the value of $\varepsilon$ is.

A similar argument shows that it is not possible to prove that the limit of the function is $L$, no matter what value of $L$ is proposed. Therefore, the limit of the function as $x \to 3$ does not exist. You will be able to understand this by re-drawing Figure 11.6 and sketching various thin, horizontal "red" strips, centred at a proposed limit value. No matter what proposed limit value is chosen, if the red strip is drawn thin enough (that is, if a sufficiently small value of $\varepsilon$ is given), it will be impossible to choose a blue strip centred at $x = 3$ such that all $x$-values in the blue strip correspond to function values that all lie within the red strip.

**Question:** Try it for yourself! With a sufficient amount of play here, including sketching lots of diagrams, you will be able to convince yourself that this limit does not exist.

I trust the graphical reasoning in the previous paragraphs is convincing to most, but if you really wish to prove that the limit of this function does not exist as $x \to 3$, how should you proceed? A little bit of logic is required. The definition of limit is: $\lim_{x\to a} f(x)$ exists and is equal to $L$ provided that for each $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon$$

The double-arrow symbol on the previous line can be read as "implies," and can also be read as "if the first condition is true then so is the second condition." In logic, such implications are called "if-then" statements.

The logical structure of the definition of limit is therefore **the limit exists provided that if A is true, then B is also true**, where A represents the condition just before the implication double-arrow, and B represents the condition just after the implication double-arrow. To disprove such a statement, you would have to show that there exists at least one value of $\varepsilon$ for which there is **no** value of $\delta$ for which the implication

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon$$

is valid. (As we have been doing, let's call the previous line the limit condition.) Study Figure 11.6 and you will see how to choose a suitable value of $\varepsilon$; just make sure that $\varepsilon$ is small enough so that the horizontal red strip does not include both branches of the function. The vertical distance between the two branches is 2 units, so anything less than this will do; for example, just take $\varepsilon = 0.1$.

With this value of $\varepsilon$, let's try to prove that

$$\lim_{x\to 3} f(x) = L$$

for the function $f$ illustrated in Figure 11.6, for various values of $L$. Once we fail to do so for all possible values of $L$, we will be forced to conclude that this limit does not exist. We'll argue two cases; Case 1 is the supposition that $L \geq 0$, and Case 2 is the supposition that $L < 0$.

Case 1: Suppose that $L \geq 0$. (Actually select a value of $L$ and label it on your own hand-drawn copy of Figure 11.6, as this will help you to follow the argument. Then label each step of the following argument on your diagram.) No matter how small you select a positive value of $\delta$, for $x^* = 3 - \delta/2$, it is true that $|x^* - 3| < \delta$ (verify this!), and yet it is also true that $f(x^*) = -1$, and therefore it is also true that $|f(x^*) - L| \geq 1$ (verify this!), and so it is **not true** that $|f(x^*) - L| < \varepsilon$. Thus, for $\varepsilon = 0.1$, there is no value of $\delta$ for which the limit condition is satisfied.

Case 2: Suppose that $L < 0$. (Actually select a value of $L$ and label it on your own hand-drawn copy of Figure 11.6, as this will help you to follow the argument. Then label each step of the following argument on your diagram.) No matter how small you select a positive value of $\delta$, for $x^{**} = 3 + \delta/2$, it is true that $|x^{**} - 3| < \delta$ (verify this!), and yet it is also true that $f(x^{**}) = 1$, and therefore it is also true that $|f(x^{**}) - L| \geq 1$ (verify this!), and so it is **not true** that $|f(x^{**}) - L| < \varepsilon$. Thus, for $\varepsilon = 0.1$, there is no value of $\delta$ for which the limit condition is satisfied.

Therefore, no matter which value of $L$ we propose as the limit of the function as $x \to 3$, there is at least one value of $\varepsilon$ (namely $\varepsilon = 0.1$) for which there is no value of $\delta$ for which the limit condition

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon$$

is valid. There are always some values of $x$ for which the limit condition fails.

This completes the proof that

$$\lim_{x \to 3} f(x)$$

does not exist.

The next example illustrates that the precise definition of the limit is also effective when applied to functions with a hole discontinuity.

## EXAMPLE 29

**Using the precise definition to verify a limit**

Guess the limit and then use the precise definition of limit to verify your guess.

$$\lim_{x \to 3} \frac{2x^2 - 5x - 3}{x - 3}$$

## SOLUTION

Following our usual practical strategy for evaluating limits, we first substitute 3 for $x$ in the expression; the result is that both numerator and denominator are 0. Because both numerator and denominator are polynomials, this is a sign that they have a common factor; by the factor theorem, the common factor is $(x - 3)$. Factoring the numerator results in $2x^2 - 5x - 3 = (x - 3)(2x + 1)$.

When the common factor is cancelled, we see that the function $g(x) = \dfrac{2x^2 - 5x - 3}{x - 3}$ is nearly identical to the function $f(x) = 2x + 1$ that we have studied extensively. The only difference is that $f$ is continuous at $x = 3$, but $g$ has a hole discontinuity at $x = 3$.

You can use the formal definition of limit to prove that

$$\lim_{x \to 3} \frac{2x^2 - 5x - 3}{x - 3} = 7$$

exactly as was illustrated earlier in this section. Try it for yourself, following the same steps as before.

We have seen that for a function with a jump discontinuity, the limit of the function as $x$ approaches the point of discontinuity does not exist. But there are more complicated ways that $\lim_{x\to a} f(x)$ might not exist. For example, consider the Dirichlet function defined earlier:

$$D(x) = \begin{cases} 1 & \text{if} \quad x \in \mathbb{Q} \\ 0 & \text{if} \quad x \notin \mathbb{Q} \end{cases}$$

I don't think it's possible to form an accurate mental image of a graph of this function. (Try it and let me know if you succeed.) It's somewhat like the union of two parallel lines, one at $y = 1$ and the other at $y = 0$, but each line is riddled with holes in a strange way. Between any two points on each line, there are an infinite number of other points that lie on the graph, but also an infinite number of holes, which represent points that do not belong to the graph.

The Dirichlet function is discontinuous at every single point in its domain. This means that

$$\lim_{x\to a} D(x)$$

does not exist for each real value of $a$. Using our previous conception of limit, you can see that as $x$ approaches any particular value $a$, the corresponding function values jump around between 0 and 1, so there is no single trend in function values. Is it possible to understand that the limit does not exist using the precise definition of limit?

To be precise about why the limit discussed in the previous paragraph does not exist, consider a small value of $\varepsilon$, say $\varepsilon = 0.3$ (any value between 0 and 1, not inclusive, would also serve). No matter how small we make $\delta$, there will always be values of $x$ within a distance $\delta$ of 2 such that there are corresponding values of $f(x)$ outside the red strip. Thus, the condition specified by the definition of the limit cannot be satisfied, and so $\lim_{x\to 2} D(x) \neq 1$. Similar arguments show that $\lim_{x\to 2} D(x) \neq 0$. In fact, for any other value of $L$ that we might try, similar arguments show that $L$ is not the limit. Thus, $\lim_{x\to 2} D(x)$ does not exist for Dirichlet's function.

**Question:** Do you understand that this limit does not exist? What would you need to do here to help you understand this point?

Similar arguments show that $\lim_{x\to a} D(x)$ does not exist for Dirichlet's function, no matter which value of $a$ is chosen. Is this clear to you?

The Dirichlet function is extreme in that it is discontinuous at each point in its domain. The following function is also extreme, although it is continuous for all values of $x$ except at $x = 0$, and it is defined for all real values of $x$; what makes this function unusual is that it wiggles more and more wildly as $x \to 0$. That is, the wiggles become narrower and narrower as $x \to 0$. If the graph represented an oscillating object, with amplitude plotted against time, then the frequency of oscillation increases without bound as $x \to 0$.

$$g(x) = \begin{cases} 0 & \text{if} \quad x = 0 \\ \sin\left(\dfrac{1}{x}\right) & \text{if} \quad x \neq 0 \end{cases}$$

You might expect this graph to have "wiggles" in it, much like the graph of $y = \sin x$, and it does, but in a more complicated way. For $x > (\pi/2)^{-1}$, and for $x < -(\pi/2)^{-1}$, the graph has no wiggles, but makes a smooth approach to the asymptote $y = 0$. However, for $-(\pi/2)^{-1} \leq x \leq (\pi/2)^{-1}$, there are infinitely many wiggles, and they become squeezed more and more closely together as $x \to 0$. Once again, this is difficult to graph; see Figure 11.7 for an attempt.

Does the graph make sense? Think about the graphs of $y = \dfrac{1}{x}$ and $y = \sin x$, both of which you studied in high school. You know that the sine function is periodic, with period $2\pi$. Every

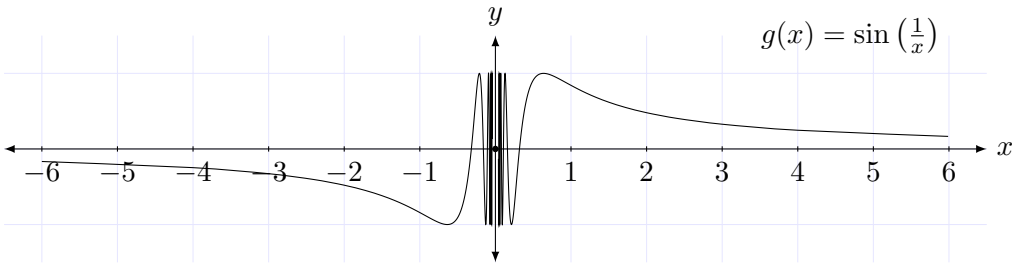



Figure 11.7: This strange function "wiggles" an infinite number of times near $x = 0$. The limit of this function as $x \to 0$ does not exist.

time the argument of the sine function changes by $2\pi$, its graph makes a complete cycle. But the argument of the sine function in the graph of $g$ in Figure 11.7 is $\frac{1}{x}$; how does this quantity cycle through periods of $2\pi$ as $x$ changes? (Sketching a rough graph of $y = \frac{1}{x}$ right now will help you follow the argument.) For very large values of $x$, the values of $\frac{1}{x}$ are quite close to zero. How far to the left do you have to move along the $x$-axis before the value of $\frac{1}{x}$ reaches $2\pi$, which would take the sine curve through its first cycle (moving from right to left)? You can determine this by setting

$$\frac{1}{x} = 2\pi$$

The result is

$$x = \frac{1}{2\pi} \approx 0.16$$

Does this seem reasonable from the graph? When is the next cycle complete, moving from left to right? Set

$$\frac{1}{x} = 4\pi$$

and solve for $x$ to obtain

$$x = \frac{1}{4\pi} \approx 0.08$$

Continuing the calculations in this way, you will understand that the wiggles in the graph of $g$ become more and more squished together as you approach the origin moving from right to left.

Alternatively, you could ask for the locations of the zeros of the function $g$. They occur every time the argument is an integer-multiple of $\pi$. Moving from right to left towards the origin, the first zero occurs at

$$x = \frac{1}{\pi} \approx 0.32$$

the next zero occurs at

$$x = \frac{1}{2\pi} \approx 0.16$$

the next zero occurs at

$$x = \frac{1}{3\pi} \approx 0.11$$

the next zero occurs at

$$x = \frac{1}{4\pi} \approx 0.08$$

and so on. You can see that the spacing between adjacent zeros decreases as you move towards the origin from right to left.

Because sine is an odd function, the graph of $g$ has similar behaviour to the left of the origin, but with opposite sign.

OK, now that we have understood the graph of $g$, let's return to our discussion of the limit of $g$ as $x \to 0$. The same sorts of arguments used for the Dirichlet function also show that $\lim_{x\to 0} g(x)$ does not exist. What could the limit be? Could it be 0? No, because consider a small value of $\varepsilon$, let's say $\varepsilon = 0.5$ (although any value between 0 and 1 not inclusive will serve). Then no matter how small $\delta$ is made, there will always be values of $x$ within a distance $\delta$ of 0 such that there are corresponding function values $g(x)$ that are outside the red $\varepsilon$-strip. The same is true no matter what value of $L$ we propose for the limit, so the limit does not exist. The problem is that the wiggles all have amplitude 1, and they get crammed together so tightly as $x \to 0$ that no matter how small you make $\delta$, there are still an infinite number of wiggles in the blue $\delta$-strip.

Note that in applying the precise definition of the limit, we must supply a guess for the limit $L$; then the definition gives us a way of confirming or denying that the supposed limit $L$ is true. The definition itself does not gives us a way of guessing the limit; that must be done independently, before we apply the definition to verify whether the guess is correct or not.

**GOOD QUESTION**

**Why doesn't the part of the precise definition of limit that reads $0 < |x - a| < \delta$, instead read as $|x - a| < \delta$? Why is the "$0 <$" included?**

Remember that when using limits to calculate slopes, we have to avoid the point $x = a$, for otherwise we would be dividing by 0. Because calculating slopes (via the definition of the derivative as a limit) is one of the major applications of limits, we have to be careful to exclude any reference to the function value at $x = a$ into the definition of limit. (We have mentioned this before, but it is such an important point that it is worth repeating.)

Now let's look at some other examples where the limit exists, and we'll see how we can use the precise definition of the limit to verify that our supposed limits are valid.[2]

[2]The practical point of having the precise definition of the limit is that it allows us to verify limits in cases where the intuitive approach is inconclusive. But, as with all new concepts, it's worthwhile practicing the precise definition on easy cases at first.

## EXAMPLE 30

**Using the precise definition of limit to prove the limit of a quadratic function**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x \to 2} \left(x^2\right)$$

## SOLUTION

The first step is to guess the limit. Quadratic functions are continuous for all real values of $x$, so we can determine the limit by substitution. Thus, we **know** that the limit is

$$\begin{aligned} \lim_{x \to 2} \left(x^2\right) &= (2)^2 \\ \lim_{x \to 2} \left(x^2\right) &= 4 \end{aligned}$$

The next step is to use the precise definition of limit to prove that 4 is indeed the limit. To do this, we must first figure out how to choose $\delta$ for a given $\varepsilon$; that is, we would like to have a simple formula for $\delta$ in terms of $\varepsilon$ that will do the job. There are various ways to do this; I'll display a method that is popular in calculus textbooks, but be aware that other methods will also work.

According to the precise definition of limit, to prove that 4 is the limit, we must show that for each positive value of $\varepsilon$, there exists a positive value of $\delta$ such that for all $x$-values that satisfy

$$0 < |x - 2| < \delta$$

the inequality

$$|x^2 - 4| < \varepsilon$$

is also satisfied. To prove this, it is helpful to have a formula for $\delta$ in terms of $\varepsilon$, and a simple formula is preferable. Recall from our previous work with using the precise definition of limit that once you find an acceptable value of $\delta$, using a smaller value of $\delta$ also works. Thus, there is no harm in restricting our attention to a small strip of values near $x = 2$. For example, we could restrict ourselves to the strip of values $1 < x < 3$. Or we could restrict ourselves to a smaller or larger strip, without any harm. We still have the task of determining a suitable value of $\delta$ for each given $\varepsilon$, but there is no harm in restricting our search to this small strip of values.

In an attempt to obtain a simple formula for $\delta$ in terms of $\varepsilon$, note that

$$x^2 - 4 = (x - 2)(x + 2)$$

Because we are restricting our attention to the values $1 < x < 3$, it follows that $|x + 2| < 5$. Thus,

$$|x^2 - 4| < 5|x - 2|$$

If we further restrict the values of $x$ so that

$$|x - 2| < \delta$$

then it follows that

$$|x^2 - 4| < 5\delta$$

Is the desired relation between $\delta$ and $\varepsilon$ now clear? We wish to ensure that

$$|x^2 - 4| < \varepsilon$$

and by comparing the previous two relations, we can arrange for the limit condition to be satisfied by choosing $5\delta = \varepsilon$ (as well as restricting $x$-values to the interval $1 < x < 3$), which is equivalent to choosing

$$\delta = \frac{\varepsilon}{5}$$

(Of course, any smaller value of $\delta$ will work just as well.) Having figured out a suitable choice for $\delta$, the last step in the proof is to verify that it works. So, for a given positive value of $\varepsilon$, choose $\delta = \varepsilon/5$, and consider the values of $x$ that satisfy

$$0 < |x - 2| < \delta$$

For these same values of $x$, the following relation is also satisfied:

$$|x - 2| < \frac{\varepsilon}{5}$$

In other words,

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |x - 2| < \frac{\varepsilon}{5}$$

Multiplying both sides of the second inequality by $|x + 2|$, we obtain

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |x - 2| \cdot |x + 2| < \frac{\varepsilon |x + 2|}{5}$$

Because $|x + 2| < 5$, if we replace $|x + 2|$ by 5 on the right side of the second inequality, the statement is still valid; after all, we are taking a valid inequality and making the larger side even larger. Therefore,

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |(x - 2)(x + 2)| < \frac{\varepsilon(5)}{5}$$

which is equivalent to

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |x^2 - 4| < \varepsilon$$

And this completes the proof. Given any positive value of $\varepsilon$, we have shown that there exists a positive value of $\delta$ such that the limit condition on the previous line is satisfied. Therefore,

$$\lim_{x \to 2} \left(x^2\right) = 4$$

**Question:** Does the proof in the previous example make sense? Does Figure 11.8 help you to make sense of the proof?

Figure 11.8 illustrates the previous example for values of $\varepsilon = 1$ and $\delta = 0.2$. Studying the dashed lines in the figure, one notices that the value of $\delta$ chosen in the example works just fine, and of course any value less than $\varepsilon/5$ would also work just fine. However, there is a bit of daylight between the vertical blue strip and the outside vertical dashed lines, which means that slightly larger values of $\delta$ would also work. An important point is that for the purposes of the proof, we

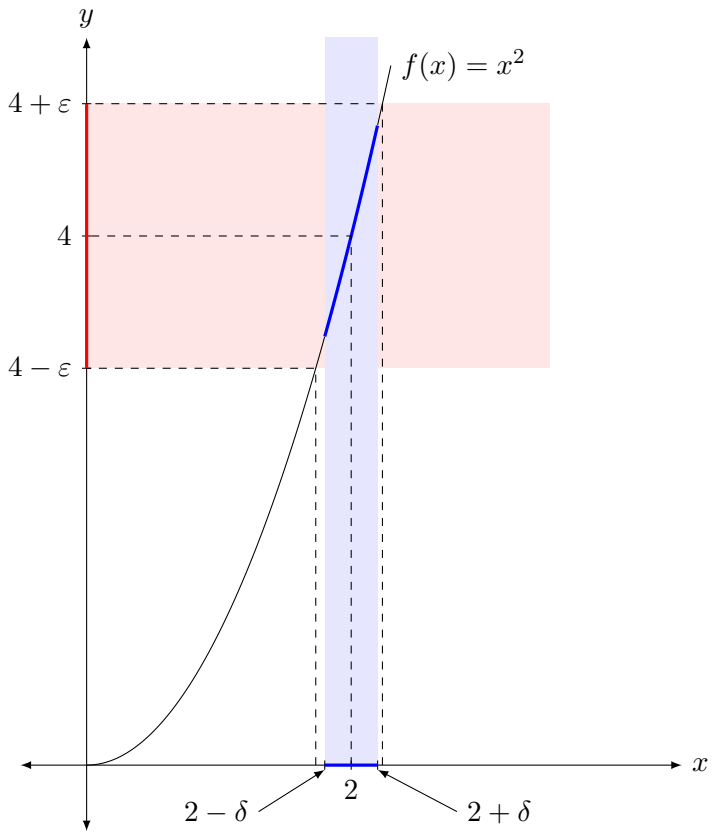



Figure 11.8: The figure may be helpful in understanding the proof (using the formal definition of limit) that the limit of the function $f(x) = x^2$ as $x \to 2$ is equal to 4. The figure illustrates values of $\varepsilon = 1$ and $\delta = 0.2$.

don't care about optimizing this; we don't care that our value of $\delta$ is the largest one possible, we just need to find a value of $\delta$ that works. Nevertheless, the curious among us will wish to explore this little gap, wouldn't we? As we studied earlier, for a linear graph with slopes $m$, good choices for $\delta$ are $\delta \leq \varepsilon/m$. What is the maximum slope of the stretch of graph within the red strip in Figure 11.8? Would choosing a value of $\delta$ equal to $\varepsilon$ divided by this maximum slope also work? Would even slightly larger values of $\delta$ work? Are these slopes even relevant? None of these questions are important for the proof just presented, but a student with a certain kind of curiosity might enjoy exploring them, and they might lend either a bit more insight or a bit more confidence. Sketch some lines on your copy of the figure and let me know how it goes!

The following example is similar to the previous one, but with a quadratic function that is a little more general.

## EXAMPLE 31

**Using the precise definition of limit to prove the limit of a quadratic function**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x\to 2}\left(3x^2+5x-7\right)$$

## SOLUTION

The first step is to guess the limit. Quadratic functions are continuous for all real values of $x$, so we can determine the limit by substitution. Thus, we **know** that the limit is

$$\begin{aligned}\lim_{x\to 2}\left(3x^2+5x-7\right) &= 3(2)^2+5(2)-7\\ \lim_{x\to 2}\left(3x^2+5x-7\right) &= 15\end{aligned}$$

The next step is to use the precise definition of limit to prove that 15 is indeed the limit. To do this, we must first figure out how to choose $\delta$ for a given $\varepsilon$; that is, we would like to have a simple formula for $\delta$ in terms of $\varepsilon$ that will do the job. As mentioned in the previous example, there are various ways to do this; I'll display one method, but variations will also work.

What we have to show is that for each given $\varepsilon > 0$, there is a $\delta > 0$ such that for all $x$ that satisfy $0 < |x-2| < \delta$, the inequality $|f(x) - 15| < \varepsilon$ is also satisfied. That is,

$$0 < |x-2| < \delta \quad \Longrightarrow \quad \left|\left(3x^2+5x-7-15\right)\right| < \varepsilon$$

which is equivalent to

$$0 < |x-2| < \delta \quad \Longrightarrow \quad \left|\left(3x^2+5x-22\right)\right| < \varepsilon$$

Based on our previous work with limits, we might guess that the quadratic expression can be factored, that $(x-2)$ is a factor, and that factoring the quadratic expression may be helpful. This is indeed correct:

$$0 < |x-2| < \delta \quad \Longrightarrow \quad |(x-2)\,[3(x-2)+17]| < \varepsilon$$

Remember, the game is that we are presented with a positive value of $\varepsilon$, and then we have to figure out how to restrict the values of $x$, if possible (i.e., choose a value of $\delta$), so that the limit condition on the previous line is satisfied. In other words, we have to figure out how to restrict the values of $x$ so that the quadratic expression is not too large — specifically it must remain less than $\varepsilon$. The $|x-2|$ factor is under control — we know it is less than $\delta$ — so we only have to worry about getting the other factor, $|3(x-2)+17|$, under control. We can do this, as we did in the previous example, by restricting the values of $x$ to be near 2; for instance, we can say that $1 < x < 3$. Remember that this is allowed, for if we ever find a value of $\delta$ that works, using a smaller value also works. With this restriction, it follows that $|x-2| < 1$, from which it follows that $|3(x-2)+17| < 20$. If this is unclear, just plot the graph of $y = 3(x-2)+17$, which is a linear function, note that the function values are all positive in the interval $|x-2| < 1$, and then observe what the maximum value of this linear function is over this interval.

It follows that for all values of $x$ that satisfy both $|x-2| < 1$ and $0 < |x-2| < \delta$,

$$|(x-2)\,[3(x-2)+17]| < \delta(20)$$

Comparing this line with the limit condition, it seems that a reasonable choice for $\delta$ in terms of $\varepsilon$ is

$$\delta = \frac{\varepsilon}{20}$$

The last step is to prove that this choice satisfies the limit condition for each value of $\varepsilon > 0$:

Given $\varepsilon > 0$, and restricting $x$-values to lie within the interval $|x - 2| < 1$, choose $\delta = \varepsilon/20$. Then,

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |x - 2| < \frac{\varepsilon}{20}$$

which is equivalent to

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad 20|x - 2| < \varepsilon$$

It follows that, for the interval of $x$-values that we are considering,

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |(x - 2)\,[3(x - 2) + 17]| < \varepsilon$$

because we have replaced 20 by a quantity that is certainly less than 20 on the interval of interest.

This completes the proof. To summarize, we have shown that given an $\varepsilon > 0$, there exists a $\delta > 0$ such that for $|x - 2| < 1$,

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |f(x) - 15| < \varepsilon$$

We can therefore conclude that

$$\lim_{x \to 2} \left(3x^2 + 5x - 7\right) = 15$$

There is nothing in the previous example that is incorrect, but there is one point that is worthy of further discussion. We stated that

$$|x - 2| < 1 \quad \Longrightarrow \quad |3(x - 2) + 17| < 20$$

which is correct, and can be understood by thinking about the graph of $y = 3(x - 2) + 17$, as we argued in the example. However, it is frequently useful to apply the triangle inequality in such situations. According to the triangle inequality,

$$|3(x - 2) + 17| \leq |3(x - 2)| + |17|$$

**Question:** Does this make sense?

If the first term on the left is negative, there will be some cancellation within the absolute value bars on the left, but there will be no cancellation on the right side of the inequality, because the terms are within their own separate absolute value bars. Thus, the right side of the inequality could be greater than the left side, but there is no way that the right side could be less than the left side. If both terms are positive, then the two sides are equal.

Continuing with the main argument, on the interval of interest, $|x - 2| < 1$, and so it is certainly true that $|3(x - 2)| < 3$, and therefore it follows that $|3(x - 2)| + |17| < 20$. Thus, we can conclude that

$$|3(x - 2) + 17| < 20$$

This argument, based on the triangle inequality, is the more commonly used one. In the previous example we preferred a more elementary argument based on the graph of a linear function.

**TRICKS OF THE TRADE**

**The triangle inequality**

If you plan on becoming a mathematics major, it will be helpful for you to practice using the triangle inequality in such arguments, as the triangle inequality is used repeatedly in higher mathematical analysis.

For a reminder about the triangle inequality, read the following feature box.

**CAREFUL!**

**The triangle inequality:** $|a+b| \leq |a| + |b|$

Reasoning with inequalities is notoriously tricky. For example, if we know that $|a+b| < 5$, does it therefore follow that $|a| + |b| < 5$? Play with this yourself by substituting some values for $a$ and $b$.

The answer to the question in the previous paragraph is "No." A counter-example is $a = 6$ and $b = -4$; in this case, $|6-4|$ is indeed less than 5, but $|6| + |-4|$ is not less than 5. In general, it is possible that $|a+b|$ is smaller than $|a| + |b|$ (this is known as the triangle inequality), so if the potentially smaller quantity is less than a certain value, it does not follow that the potentially larger value is less than that same certain value. Arguing the other way, however, is valid: If the potentially larger value is smaller than some certain value, then it is definitely true that the potentially smaller value is also less than that same certain value. This reasoning was used in the passage following the previous example.

After having played with the triangle inequality sufficiently, so that you understand it and have a good feel for it, you can prove it as follows. First note that the fact $-|s| \leq r \leq |s|$ is equivalent to the fact $|r| \leq |s|$. (Illustrate this for yourself on a diagram of the real line!) Now consider the inequalities

$$-|a| \leq a \leq |a| \quad \text{and} \quad -|b| \leq b \leq |b|$$

and add the inequalities term-by-term to obtain

$$-(|a| + |b|) \leq a + b \leq |a| + |b|$$

As you illustrated above, this is equivalent to the triangle inequality, $|a+b| \leq |a| + |b|$, which completes the proof.

The triangle inequality gets its name from a version of it involving vectors, where the double absolute-value bars mean the length of the enclosed vector:

$$||\mathbf{a} + \mathbf{b}|| \leq ||\mathbf{a}|| + ||\mathbf{b}||$$

You can understand this version of the triangle inequality intuitively by noting that the three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{a}+\mathbf{b}$ form the sides of a triangle (sketch this!). This form of the triangle inequality states that the sum of the lengths of two sides of a triangle is greater than the length of the other side. This is true no matter which two sides are chosen. Another way of intuitively understanding this version of the triangle inequality is to remember the phrase "the shortest distance between two points is a straight line."

If you know a bit about vectors and dot product, you may be able to follow this terse proof of the vector version of the triangle inequality:

$$\begin{aligned}
||\mathbf{a}+\mathbf{b}||^2 &= (\mathbf{a}+\mathbf{b})\cdot(\mathbf{a}+\mathbf{b}) \\
||\mathbf{a}+\mathbf{b}||^2 &= \mathbf{a}\cdot\mathbf{a}+2\mathbf{a}\cdot\mathbf{b}+\mathbf{b}\cdot\mathbf{b} \\
||\mathbf{a}+\mathbf{b}||^2 &= ||\mathbf{a}||^2+2||\mathbf{a}||\,||\mathbf{b}||\cos\theta+||\mathbf{b}||^2 \\
||\mathbf{a}+\mathbf{b}||^2 &\leq ||\mathbf{a}||^2+2||\mathbf{a}||\,||\mathbf{b}||+||\mathbf{b}||^2 \qquad (\text{because } \cos\theta \leq 1) \\
||\mathbf{a}+\mathbf{b}||^2 &\leq (||\mathbf{a}||+||\mathbf{b}||)^2 \\
||\mathbf{a}+\mathbf{b}|| &\leq ||\mathbf{a}||+||\mathbf{b}||
\end{aligned}$$

## GOOD QUESTION

**In the previous examples of this section, we know what the limits are. So why do we bother to prove what we already know?**

It may be worth reminding you at this point why we are bothering with these complicated arguments to justify limits that may be obvious to you. The reason for this is that you should always practice new tools in situations that you already understand, as this will help you absorb the new methods. Once you understand how to use new tools in relatively simple situations, this will give you confidence to use them in situations that are more involved, especially in situations where your old understanding (before having the new tool) is insufficient.

Remember that in complicated situations, we will not **know** what the limits are. In these situations, the precise definition of the limit is absolutely essential, so we will have to use it. It is wise to practice its use in simple situations so that we will have confidence in its use in more complicated situations.

## EXAMPLE 32

**Using the precise definition of limit to prove the limit of a rational function**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x\to 0.5}\left(\frac{3}{x}\right)$$

## SOLUTION

The first step is to guess the limit. Rational functions are continuous for all real values of $x$ at which they are defined, so we can determine the limit by substitution. Thus, we **know** that the limit is

$$\begin{aligned}\lim_{x\to 0.5}\left(\frac{3}{x}\right) &= \frac{3}{0.5}\\ \lim_{x\to 0.5}\left(\frac{3}{x}\right) &= 6\end{aligned}$$

Next we will use the precise definition of limit to prove that 6 is indeed the limit. To do this, we must first figure out how to choose $\delta$ for a given $\varepsilon$. That is, we must show that for each given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$0 < |x - 0.5| < \delta \quad \Longrightarrow \quad \left|\frac{3}{x} - 6\right| < \varepsilon$$

Observe that

$$\begin{aligned}\frac{3}{x} - 6 &= \frac{3}{x} - \frac{3}{0.5}\\ \frac{3}{x} - 6 &= 3\left(\frac{1}{x} - \frac{1}{0.5}\right)\\ \frac{3}{x} - 6 &= 3\left(\frac{0.5 - x}{0.5x}\right)\\ \frac{3}{x} - 6 &= 6\left(\frac{0.5 - x}{x}\right)\\ \frac{3}{x} - 6 &= -6\left(\frac{x - 0.5}{x}\right)\\ \frac{3}{x} - 6 &= (x - 0.5)\left(\frac{-6}{x}\right)\end{aligned}$$

Thus, the limit condition is equivalent to the condition

$$0 < |x - 0.5| < \delta \quad \Longrightarrow \quad \left|(x - 0.5)\left(\frac{-6}{x}\right)\right| < \varepsilon$$

which is equivalent to

$$0 < |x - 0.5| < \delta \quad \Longrightarrow \quad |x - 0.5|\left|\frac{6}{x}\right| < \varepsilon$$

The first factor at the far right is under control, so the task is to ensure that the second factor is also under control. We can do this by restricting the values of $x$ under consideration to a suitably small interval centred at $x = 0.5$, say $0.4 < x < 0.6$. By doing this, we can ensure that the maximum value of $6/x$ is $6/0.4 = 15$. Thus, provided that both conditions $0.4 < x < 0.6$ and $0 < |x - 0.5| < \delta$ are satisfied, it follows that

$$|(x - 0.5)| \left|\frac{6}{x}\right| < 15\delta$$

It seems that choosing $15\delta = \varepsilon$, that is choosing $\delta = \varepsilon/15$, will therefore work. This can be verified as follows:

Given $\varepsilon > 0$, choose $\delta = \varepsilon/15$. Then, restricting the values of $x$ under consideration to $0.4 < x < 0.6$, it follows that

$$\begin{aligned}
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < |x - 0.5| < \frac{\varepsilon}{15} \\
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < |x - 0.5|(15) < \varepsilon \\
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < |x - 0.5| \left|\frac{6}{x}\right| < \varepsilon \\
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < \left|\frac{6x - 3}{x}\right| < \varepsilon \\
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < \left|6 - \frac{3}{x}\right| < \varepsilon \\
0 < |x - 0.5| < \delta \quad &\Longrightarrow \quad 0 < \left|\frac{3}{x} - 6\right| < \varepsilon
\end{aligned}$$

Thus, given $\varepsilon > 0$ there exists a $\delta > 0$, specifically $\delta = \varepsilon/15$, such that (if we restrict the values of $x$ to $0.4 < x < 0.6$)

$$0 < |x - 0.5| < \delta \quad \Longrightarrow \quad \left|\frac{3}{x} - 6\right| < \varepsilon$$

We can conclude that indeed $\lim_{x \to 0.5} \left(\frac{3}{x}\right) = 6$. Sketching a graph and tracing the steps of the proof on the graph will help you understand it.

We have so far illustrated the precise definition of the limit for a few polynomial functions, one rational function, a discontinuous function, and a couple of more exotic functions. The same definition can be used on any function whatsoever, but it would take hundreds of pages to illustrate using the formal definition of a limit on all types of functions. We'll be content with just one more example, which follows.

## EXAMPLE 33

**Using the precise definition of limit to prove the limit of a radical function**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x\to 4}\left(\sqrt{x}\right)$$

## SOLUTION

The first step is to guess the limit. This function is continuous at $x = 4$, so we can determine the limit by substitution. Thus, we **know** that the limit is

$$\begin{aligned}\lim_{x\to 4}\left(\sqrt{x}\right) &= \sqrt{4}\\ \lim_{x\to 4}\left(\sqrt{x}\right) &= 2\end{aligned}$$

Next we will use the precise definition of limit to prove that 2 is indeed the limit. To do this, we must first figure out how to choose $\delta$ for a given $\varepsilon$. That is, we must show that for each given $\varepsilon > 0$, there is a $\delta > 0$ such that

$$0 < |x-4| < \delta \quad \Longrightarrow \quad \left|\sqrt{x}-2\right| < \varepsilon$$

Observe that

$$\begin{aligned}\sqrt{x}-2 &= \left(\sqrt{x}-2\right)\cdot\frac{\sqrt{x}+2}{\sqrt{x}+2}\\ \sqrt{x}-2 &= \frac{x-4}{\sqrt{x}+2}\\ \sqrt{x}-2 &= (x-4)\left(\frac{1}{\sqrt{x}+2}\right)\end{aligned}$$

Throughout the domain of the function (that is, $x \geq 0$), the value of the second factor on the right of the previous equation is no more than $1/2$. The first factor is less than $\delta$. Thus, a good choice of $\delta$ appears to be to set $\delta/2 = \varepsilon$, which means to choose $\delta = 2\varepsilon$.

To prove that this choice works, note that

$$\begin{aligned}\left|\sqrt{x}-2\right| &= |x-4|\left|\frac{1}{\sqrt{x}+2}\right|\\ \left|\sqrt{x}-2\right| &< \delta\cdot\frac{1}{2}\\ \left|\sqrt{x}-2\right| &< \varepsilon\end{aligned}$$

Thus, we have shown that for each given $\varepsilon > 0$, there exists a positive value of $\delta$, namely $\delta = 2\varepsilon$, such that

$$0 < |x-4| < \delta \quad \Longrightarrow \quad \left|\sqrt{x}-2\right| < \varepsilon$$

By the definition of limit, this proves that

$$\lim_{x\to 4}\left(\sqrt{x}\right) = 2$$

The examples in this section will serve to give you the flavour of the arguments necessary to prove that a conjectured limit is indeed correct, using the formal definition of a limit. The exercises at the end of this chapter will give you an opportunity to further practice these arguments, after you have had sufficient practice in reproducing the arguments of the previous examples without peeking at them. (It may take a few iterations before you can successfully repeat the arguments on your own; be patient and persistent.)

Before you tackle the exercises, it is worthwhile devoting a bit of time to exploring what happens when you try to use the formal definition of a limit to prove a limit that is actually incorrect. It's good to see how this goes wrong, so that you will be able to better spot your own errors should you make any in this context. For example, consider the function

$$f(x) = 3x + 5$$

and consider the limit

$$\lim_{x \to 2} f(x)$$

Because $f$ is continuous, we know that the indicated limit is $3(2) + 5 = 11$. But suppose that we make a calculation error, and that we mistakenly think that the indicated limit is actually 9. Let's try to prove this mistaken limit using the formal definition of limit and see what happens.

By the definition of limit, we would have to prove that for each given $\varepsilon > 0$, there exists a positive value of $\delta$ such that

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |(3x + 5) - 9| < \varepsilon$$

which is equivalent to

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |3x - 4| < \varepsilon$$

which is equivalent to

$$0 < |x - 2| < \delta \quad \Longrightarrow \quad |3(x - 2) + 2| < \varepsilon$$

But this is not possible. The first term on the right, $3(x - 2)$, is under control (it is no greater than $3\delta$), but there is nothing we can do about the second term, which is resolutely equal to 2. Therefore, if we are presented with a value of $\varepsilon$ that is small enough (say, $\varepsilon = 0.1$), there is no way that we can ensure that $|3(x - 2) + 2| < \varepsilon$ by choosing a small enough value of $\delta$.

Note that you can't get around this by just cherry-picking a single value of $x$ for a given $\varepsilon$; the condition has to be satisfied **for all** $x$-values in the interval defined by $0 < |x - 2| < \delta$. To see that this can't be done, it would be helpful to sketch a diagram.

Let's discuss a few final thoughts on the formal definition of a limit before moving on. Why isn't the definition the other way around? That is, why doesn't the definition say that given any positive value of $\delta$, there exists a positive value of $\epsilon$ such that the limit condition is valid. After all, our informal concept of limit is that $\lim_{x \to a} f(x) = L$ means that when $x$ is close to $a$, it follows that $f(x)$ is close to $L$. Why doesn't the formal definition parallel this phrase?

The reason for this is that such a formulation actually doesn't achieve what we wish. Consider this attempt at a definition of limit: We say that

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

provided that for each $\delta > 0$, there exists a value of $\varepsilon > 0$ such that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon$$

This doesn't work! For example, consider a function that has a jump discontinuity at $x = a$. (Sketch a graph!) For each positive value of $\delta$, it is indeed possible to choose a positive value of $\varepsilon$ such that the limit condition in the previous equation is satisfied. Just choose the value of $\varepsilon$ to be large enough.

By this new definition, the discontinuous function would have to be judged continuous, because it satisfies the condition! I hope this discussion makes clear that the proposed new definition of a limit doesn't work.

One way to think about this: We wish the formal definition of a limit to reflect our intuitive conception of limit, that as $x \to a$, $f(x) \to L$. If this were not true, how could you show it? Well, one way would be to demonstrate that there is some kind of "red zone" along the $y$-axis, centred on $y = L$, such that even as $x \to a$, there are some values of $f(x)$ that stay out of the red zone. In other words, that there exists a positive value of $\varepsilon$ such that **it is not true that**

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |f(x) - L| < \varepsilon$$

Does this make sense? If so, then maybe this gives you another perspective on why the actual formal definition of a limit is the way it is. The formal definition says that if the limit really is $L$, then there is no such red zone; for **each** positive value of $\varepsilon$, **all** values of $x$ that are sufficiently close to $a$ (i.e., within a distance $\delta$) have corresponding function values that are close to $L$ (i.e., within a distance $\varepsilon$).

**EXERCISES**

(Answers at end.)

Guess each limit. Then use the precise definition of limit to prove that your guess is correct. Illustrate your work by sketching a graph in each case.

1. $\lim\limits_{x\to 2} \left(x^2 - 3\right)$
2. $\lim\limits_{x\to 2} \left(x^3\right)$
3. $\lim\limits_{x\to 3} \left(x^2 - x + 1\right)$
4. $\lim\limits_{x\to -3} \left(x^2 + 2x + 1\right)$
5. $\lim\limits_{x\to 3} \left(\dfrac{4}{x}\right)$
6. $\lim\limits_{x\to 9} \left(\sqrt{x}\right)$

---

Answers: 1. limit is 1; restrict $x$ to $|x - 2| < 1$ and then choose $\delta = \varepsilon/5$ for the proof, but other choices also work

2. limit is 8; restrict $x$ to $|x - 2| < 1$ and then choose $\delta = \varepsilon/19$ for the proof, but other choices also work

3. limit is 7; restrict $x$ to $|x - 3| < 1$ and then choose $\delta = \varepsilon/6$ for the proof, but other choices also work

4. limit is 4; restrict $x$ to $|x + 3| < 1$ and then choose $\delta = \varepsilon/5$ for the proof, but other choices also work

5. limit is 4/3; restrict $x$ to $|x - 3| < 1$ and then choose $\delta = \varepsilon/2$ for the proof, but other choices also work

6. limit is 3; restrict $x$ to $|x - 9| < 5$ and then choose $\delta = 5\varepsilon$ for the proof, but other choices also work

## HISTORY

### Karl Weierstrass (1815–1897)

Weierstrass was born in Ostenfelde, a village in northwestern Germany. He became interested in mathematics in elementary school, and nurtured this love throughout his life. He was sent to university to prepare for a career as a government official, like his father. However, he neglected his official studies in law, economics, finance, and so on, and studied mathematics on his own instead. He also devoted a lot of time to drinking and fencing. Not surprisingly, given the objects of his attention, he did not obtain a university degree.

Having failed at his official university studies, Weierstrass then began teacher training, and after successful completion became a high-school teacher. He was an excellent and devoted teacher, and a friendly and sociable person with an active social life. However, at night he was a solitary mathematics researcher, and his companions then were the great mathematicians of the past, who communicated to Weierstrass through the books and papers they left behind. Among the great works of the past that Weierstrass studied, Abel (whom we met in a history feature in Chapter 8) was his constant companion. Weierstrass submitted a paper on Abelian functions to Crelle's journal, the same one that published so many of Abel's papers in its early days. Upon its publication in 1854 it became widely recognized as a work of genius, and a year later Weierstrass accepted an offer to become professor of mathematics in Berlin, where he remained for the rest of his life. It is quite unusual in mathematics and science for an unknown to get his big break at the age of 40, then become world famous and widely recognized as one of the leading mathematicians in the world.

Once he became a professor, Weierstrass devoted little time to writing, preferring to teach his students and help nurture their careers. His primary focus was analysis, in particular cleaning up the foundations. Abel, Cauchy, and Gauss did admirable work in this direction, but Cauchy was still relying too much on geometric feeling in his work, and definitions in use were still not precise enough. In particular, Weierstrass conceived of the precise definition of limit that is discussed in this chapter, and which is now the accepted definition of limit. His student Heine published this definition in 1872. Weierstrass worked on other foundational issues, such as the definition of real numbers.

In his time, one of the key questions was what exactly is a function, and under which conditions do power series converge to the functions that they are supposed to model. An astounding example that Weierstrass came up with, that was relevant to both questions, and also to a key theorem in calculus, is that of a function that is defined and continuous for all real numbers, and yet is not differentiable for each real number! Can you imagine sketching the graph of such an everywhere-continuous, nowhere-differentiable function? I certainly can't; it's "corners" everywhere! (When you learn a little more about power series and Fourier series you will be able to understand this strange function.)

Weierstrass was generous to many students, but notable among them was Sophie Kowalevski (Sofya Kovalevskaya). She came to Berlin as a 20-year-old student, but was denied permission to take university courses; it was the custom at that time to deny women higher education. Weierstrass recognized her abilities, and agreed to tutor Kowalevski privately, and gave her the same lessons that he was giving in his regular university lectures. With Weierstrass's instruction and support, Kowalevski became the first woman awarded a doctoral degree in mathematics, and became a role model for other young women.

Weierstrass never married (nor did any of his siblings), perhaps because of his domineering father, but he became a wonderful father figure to many students.

**SUMMARY**

In this chapter the formal definition of a limit was presented. Discussion of the concept of a limit, including its connection with our earlier, informal concept of limit, was followed by a number of examples, worked out in detail.

# Chapter 12

# Theory, Part 2

**OVERVIEW**

After introducing the formal definition of a limit in the previous section, we adapt the definition in this section to various other limit situations. Then we state and prove the limit laws and a few other important theorems.

## 12.1 Limits "at Infinity"

Limits "at infinity" are those of the form

$$\lim_{x\to\infty} f(x) \quad \text{or} \quad \lim_{x\to-\infty} f(x)$$

We'll include the usual caution that infinity is not a place, and has no location, so the phrase "limit at infinity" must be understood as a short version of a more precise phrase. Virtually all calculus textbooks use this short phrase, just don't be mislead by it when you see it.

We have used an informal understanding of limit to calculate limits at infinity earlier in this book; the following definition makes this concept precise.

**DEFINITION 13**

**Precise definition of limit "at infinity"**

The limit $\lim_{x\to\infty} f(x)$ exists and is equal to $L$ — that is, $\lim_{x\to\infty} f(x) = L$ —

provided that for each positive value of $\varepsilon$ there exists a value of $M$ such that for all $x > M$,

$$|f(x) - L| < \varepsilon$$

Similarly, the limit $\lim_{x\to-\infty} f(x)$ exists and is equal to $L$ — that is, $\lim_{x\to-\infty} f(x) = L$ —

provided that for each positive value of $\varepsilon$ there exists a value of $M$ such that for all $x < M$,

$$|f(x) - L| < \varepsilon$$

## EXAMPLE 34

**Using the precise definition of limit to prove a limit "at infinity"**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x\to-\infty}\left(\frac{2x^2}{1+x^2}\right)$$

## SOLUTION

Based on our informal understanding of limits, the limit of the function is 2, because as $x \to -\infty$

$$\frac{2x^2}{1+x^2} = \frac{\dfrac{2x^2}{x^2}}{\dfrac{1}{x^2}+\dfrac{x^2}{x^2}} = \frac{2}{\dfrac{1}{x^2}+1} \to 2$$

To prove that this value is correct, for each $\varepsilon > 0$, we must show that there exists a value of $M$ such that for all $x < M$, the limit condition

$$\left|\frac{2x^2}{1+x^2} - 2\right| < \varepsilon$$

is satisfied. It will be helpful to "solve" this inequality for $x$, because the relation between $M$ and $\varepsilon$ that we seek involves a condition on $x$. Thus:

$$\begin{aligned}
2\left|\frac{x^2}{1+x^2} - 1\right| &< \varepsilon \\
\left|\frac{x^2}{1+x^2} - \frac{1+x^2}{1+x^2}\right| &< \frac{\varepsilon}{2} \\
\left|\frac{x^2-\left(1+x^2\right)}{1+x^2}\right| &< \frac{\varepsilon}{2} \\
\left|\frac{1}{1+x^2}\right| &< \frac{\varepsilon}{2} \\
\frac{1}{1+x^2} &< \frac{\varepsilon}{2} \\
2 &< \varepsilon\left(1+x^2\right) \\
2 &< \varepsilon + \varepsilon x^2 \\
2-\varepsilon &< \varepsilon x^2 \\
x^2 &> \frac{2-\varepsilon}{\varepsilon}
\end{aligned}$$

Note that the function values all lie between the values 0 and 2. Thus, for $\varepsilon \geq 2$, the limit condition will be satisfied for all real values of $M$. For $\varepsilon < 2$, the limit condition is satisfied provided that

$$x < -\sqrt{\frac{2-\varepsilon}{\varepsilon}}$$

Therefore, given $\varepsilon \geq 2$, an arbitrary real value of $M$ will work. For a given $\varepsilon < 2$, choose

$$M = -\sqrt{\frac{2-\varepsilon}{\varepsilon}}$$

This completes the proof.

It would be wise for you to sketch a graph of the function from the previous example, select a few representative values of $\varepsilon$, calculate the corresponding values of $M$, and then illustrate the matching pairs of values on the graph. This will provide evidence for the validity of the relationship between $M$ and $\varepsilon$.

## 12.2 One-Sided Limits

The formal definition of limit can also be adapted to the situation of one-sided limits.

**DEFINITION 14**

**Precise definition of one-sided limit**

The limit

$$\lim_{x \to a^+} f(x)$$

exists and is equal to $L$ — that is,

$$\lim_{x \to a^+} f(x) = L$$

provided that for each positive value of $\varepsilon$ there exists a value of $\delta$ such that for all $x$ that satisfy $a < x < a + \delta$,

$$|f(x) - L| < \varepsilon$$

Similarly, the limit

$$\lim_{x \to a^-} f(x)$$

exists and is equal to $L$ — that is,

$$\lim_{x \to a^-} f(x) = L$$

provided that for each positive value of $\varepsilon$ there exists a value of $\delta$ such that for all $x$ that satisfy $a - \delta < x < a$,

$$|f(x) - L| < \varepsilon$$

**EXAMPLE 35**

**Using the precise definition of limit to prove a one-sided limit**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x \to 0^+} \frac{|x|}{x}$$

**SOLUTION**

For $x > 0$, $|x| = x$, so $\dfrac{|x|}{x} = 1$, so it's reasonable to guess that the limit is 1. Because the function values are constant for $x > 1$, it follows that the limit condition

$$\left| \frac{|x|}{x} - 1 \right| < \varepsilon$$

is satisfied for each positive value of $\varepsilon$ for all positive values of $x$. Thus, given $\varepsilon > 0$, one can choose a positive value of $\delta$ arbitrarily and the limit condition will be satisfied. This completes the proof. Sketch the graph to verify that $\delta$ can be chosen arbitrarily!

## 12.3 "Infinite" Limits

The formal definition of limit can also be adapted to the situation of "infinite" limits. Remember that it's not that such limits exist and are equal to infinity; rather, such limits do not exist, and using the symbol $\infty$ simply provides more information about why such limits do not exist.

**DEFINITION 15**

**Precise definition of "infinite" limit**

The limit statement $\lim_{x \to a^+} f(x) = \infty$ means that this limit does not exist because the function values increase without bound as $x \to a^+$. Formally, $\lim_{x \to a^+} f(x) = \infty$ provided that for each value of $M$ there exists a positive value of $\delta$ such that for all $x$ satisfying $a < x < a + \delta$, $f(x) > M$. Similarly, $\lim_{x \to a^+} f(x) = -\infty$ provided that for each value of $M$ there exists a positive value of $\delta$ such that for all $x$ satisfying $a < x < a + \delta$, $f(x) < M$. Similar definitions apply for left-hand limits.

It will be a good test of your understanding for you to write explicit definitions for infinite left-hand limits using the previous definition as a guide.

## EXAMPLE 36

**Using the precise definition of limit to prove an "infinite" limit**

Evaluate the limit and then use the precise definition of limit to prove your guess correct.

$$\lim_{x \to 3^-} \frac{1}{x-3}$$

## SOLUTION

Sketching a graph of the function will make the rest of this paragraph easier to understand. As $x \to 3^-$, the denominator of the formula for the function $\to 0$, and the numerator is constant, which means that the function either $\to \infty$ or $\to -\infty$. Noting that the denominator is negative for $x < 3$, it follows that

$$\lim_{x \to 3^-} \frac{1}{x-3} = -\infty$$

To prove this using the precise definition of limit, we must show that for each value of $M$ there exists a positive value of $\delta$ such that for all $x$ satisfying $3 - \delta < x < 3$,

$$\frac{1}{x-3} < M$$

Subtract 3 from each term of the inequality $3 - \delta < x < 3$ to obtain

$$-\delta < x - 3 < 0$$

from which it follows that

$$\frac{1}{x-3} < -\frac{1}{\delta}$$

This means that we can guarantee that the limit condition

$$\frac{1}{x-3} < M$$

is satisfied by choosing

$$-\frac{1}{\delta} < M$$

which is equivalent to

$$\delta < -\frac{1}{M}$$

Thus, given $M$, choose $\delta < -\dfrac{1}{M}$, which ensures that for all $x$ satisfying $3 - \delta < x < 3$,

$$\frac{1}{x-3} < M$$

This completes the proof.

It would be wise for you to sketch a graph of the function from the previous example, select

a few representative values of $M$, calculate the corresponding values of $\delta$, and then illustrate the matching pairs of values on the graph. This will provide evidence for the validity of the relationship between $\delta$ and $M$.

There are other kinds of limit statements that we can make, but the ones we have treated so far cover the most common situations, and will give you the flavour of how the formal definition of limit works in these common situations.

Next, let's recall the limit laws, stated in Chapter 8, and provide proofs of each one, based on the formal definition of limit. Again, analogous laws hold for the other types of limits, and these other limit laws can be proved in similar ways, once you have absorbed the flavour of the following proofs.

## 12.4 Limit Laws

In this section we state the limit laws, used earlier in this book, and then we prove them based on the precise definition of a limit.

**THEOREM 6**

**Limit laws**

Suppose that the function $f$ is an algebraic combination of simpler functions. Also suppose that the limit of each of the simpler functions exists. Then to evaluate the limit of $f$, just evaluate the limit of each of the simpler functions, and combine the individual limits using the same algebraic combination that forms $f$.

To be more specific, here are some fundamental instances of this idea. We also assume that $k$ is a constant, and that $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both exist.

(a) $\lim\limits_{x\to a} [k \cdot f(x)] = k \left[\lim\limits_{x\to a} f(x)\right]$

(b) $\lim\limits_{x\to a} [f(x) + g(x)] = \lim\limits_{x\to a} f(x) + \lim\limits_{x\to a} g(x)$

(c) $\lim\limits_{x\to a} [f(x) - g(x)] = \lim\limits_{x\to a} f(x) - \lim\limits_{x\to a} g(x)$

(d) $\lim\limits_{x\to a} [f(x) \cdot g(x)] = \left[\lim\limits_{x\to a} f(x)\right] \cdot \left[\lim\limits_{x\to a} g(x)\right]$

(e) $\lim\limits_{x\to a} \left[\dfrac{f(x)}{g(x)}\right] = \dfrac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$, provided that $\lim\limits_{x\to a} g(x) \neq 0$

Proof (a): Let

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

This means that given $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x$ that satisfy $0 < |x - a| < \delta$, the limit condition $|f(x) - L| < \dfrac{\varepsilon}{k}$ is also satisfied. The limit condition is equivalent to

$$k|f(x) - L| < \varepsilon$$

which is equivalent to

$$|kf(x) - kL| < \varepsilon$$

which is equivalent to the statement that

$$\lim_{x \to a} k \cdot f(x) = kL$$

which is equivalent to the statement that

$$\lim_{x \to a} k \cdot f(x) = k \cdot \left[\lim_{x \to a} k \cdot f(x)\right]$$

This completes the proof of Part (a) of the theorem.

Proof (b): Let

$$L_1 = \lim_{x \to a} f(x) \qquad \text{and} \qquad L_2 = \lim_{x \to a} g(x)$$

This means that given $\varepsilon > 0$, there exist $\delta_1 > 0$ and $\delta_2 > 0$ such that for all $x$ that satisfy $0 < |x - a| < \delta_1$, the limit condition $|f(x) - L_1| < \dfrac{\varepsilon}{2}$ is satisfied, and for all $x$ that satisfy $0 < |x - a| < \delta_2$, the limit condition $|g(x) - L_2| < \dfrac{\varepsilon}{2}$ is satisfied. For this given value of $\varepsilon$, choose $\delta$ to be less than the minimum value of $\delta_1$ and $\delta_2$. Then, for all $x$ that satisfy $0 < |x - a| < \delta$, both of the limit conditions

$$|f(x) - L_1| < \frac{\varepsilon}{2} \qquad \text{and} \qquad |g(x) - L_2| < \frac{\varepsilon}{2}$$

are satisfied. Therefore, for the same values of $x$,

$$|f(x) - L_1| + |g(x) - L_2| < \varepsilon$$

Recall the triangle inequality: $|m + n| \leq |m| + |n|$. It follows that for all $x$ that satisfy $0 < |x - a| < \delta$,

$$|f(x) - L_1 + g(x) - L_2| < \varepsilon$$

which is equivalent to

$$|[f(x) + g(x)] - [L_1 + L_2]| < \varepsilon$$

This completes the proof of Part (b) of the theorem.

Proof (c):

$$\lim_{x \to a} [f(x) - g(x)] = \lim_{x \to a} [f(x) + (-g(x))]$$

$$\lim_{x \to a} [f(x) - g(x)] = \lim_{x \to a} f(x) + \lim_{x \to a} [-g(x)] \qquad \text{(by Part (b))}$$

$$\lim_{x \to a} [f(x) - g(x)] = \lim_{x \to a} f(x) - \lim_{x \to a} g(x) \qquad \text{(by Part (a), with } k = -1)$$

Proof (d): Let

$$L_1 = \lim_{x\to a} f(x) \qquad \text{and} \qquad L_2 = \lim_{x\to a} g(x)$$

This means that given $\varepsilon > 0$, there exists $\delta_1 > 0$ and $\delta_2 > 0$ such that for all $x$ that satisfy $0 < |x-a| < \delta_1$, the limit condition $|f(x) - L_1| < \dfrac{\varepsilon}{2(1+|L_2|)}$ is satisfied, and for all $x$ that satisfy $0 < |x-a| < \delta_2$, the limit condition $|g(x) - L_2| < \dfrac{\varepsilon}{2(1+|L_1|)}$ is satisfied.

Additionally, a third limit condition is that given $\varepsilon > 0$, there exists $\delta_3 > 0$ such that for all $x$ that satisfy $0 < |x-a| < \delta_3$, $|g(x) - L_2| < 1$ is satisfied. Using the triangle inequality, it follows that for all $x$ that satisfy $0 < |x-a| < \delta_3$,

$$|g(x)| = |g(x) - L_2 + L_2| \le |g(x) - L_2| + |L_2|$$

and therefore, for the same values of $x$,

$$|g(x)| \le 1 + |L_2|$$

Now we can complete the proof. For the given value of $\varepsilon$, choose $\delta$ to be less than the minimum value of $\delta_1$, $\delta_2$, and $\delta_3$. Then, for all $x$ that satisfy $0 < |x-a| < \delta$, all three limit conditions

$$|f(x) - L_1| < \frac{\varepsilon}{2(1+|L_2|)} \qquad \text{and} \qquad |g(x) - L_2| < \frac{\varepsilon}{2(1+|L_1|)} \qquad \text{and} \qquad |g(x) - L_2| < 1$$

are satisfied. Therefore, for the same values of $x$,

$$\begin{aligned}
&|f(x)g(x) - L_1L_2| = |f(x)g(x) - L_1g(x) + L_1g(x) - L_1L_2| \\
&|f(x)g(x) - L_1L_2| \le |f(x)g(x) - L_1g(x)| + |L_1g(x) - L_1L_2| \\
&|f(x)g(x) - L_1L_2| \le |g(x)|\,|f(x) - L_1| + |L_1|\,|g(x) - L_2| \\
&|f(x)g(x) - L_1L_2| \le [1+|L_2|]\left[\frac{\varepsilon}{2(1+|L_2|)}\right] + |L_1|\left[\frac{\varepsilon}{2(1+|L_1|)}\right] \\
&|f(x)g(x) - L_1L_2| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&|f(x)g(x) - L_1L_2| \le \varepsilon
\end{aligned}$$

Moving from the third-to-last line to the second-to-last line is justified by the observation that

$$\frac{|L_1|}{1+|L_1|} < 1$$

This completes the proof of Part (d) of the theorem.

Proof (e):

We first prove the special case for which $f(x) = 1$. That is, we first prove that

$$\lim_{x\to a}\left[\frac{1}{g(x)}\right] = \frac{1}{\lim_{x\to a} g(x)}$$

provided that $\lim_{x\to a} g(x) \neq 0$. Let $L = \lim_{x\to a} g(x)$; thus, there exists $\delta_1 > 0$ such that

for all $x$ that satisfy $0 < |x - a| < \delta_1$,

$$|g(x) - L| < \frac{|L|}{2}$$

For the same values of $x$,

$$\begin{aligned}
|L| &= |L - g(x) + g(x)| \\
|L| &\leq |L - g(x)| + |g(x)| \\
|L| &< \frac{|L|}{2} + |g(x)| \\
\frac{|L|}{2} &< |g(x)| \\
\frac{1}{|g(x)|} &< \frac{2}{|L|}
\end{aligned}$$

Furthermore, because $\lim_{x \to a} g(x) = L$, given $\varepsilon > 0$, there exists $\delta_2 > 0$ such that for all $x$ that satisfy $0 < |x - a| < \delta_2$,

$$|g(x) - L| < \frac{\varepsilon |L|^2}{2}$$

Now choose $\delta$ as the minimum of $\delta_1$ and $\delta_2$. Then for $x$ that satisfy $0 < |x - a| < \delta$,

$$\begin{aligned}
\left| \frac{1}{g(x)} - \frac{1}{L} \right| &= \left| \frac{L - g(x)}{L g(x)} \right| \\
\left| \frac{1}{g(x)} - \frac{1}{L} \right| &= \frac{1}{|L|} \frac{1}{|g(x)|} |g(x) - L| \\
\left| \frac{1}{g(x)} - \frac{1}{L} \right| &< \frac{1}{|L|} \frac{2}{|L|} \frac{\varepsilon |L|^2}{2} \\
\left| \frac{1}{g(x)} - \frac{1}{L} \right| &< \varepsilon
\end{aligned}$$

Thus,

$$\lim_{x \to a} \left[ \frac{1}{g(x)} \right] = \frac{1}{L}$$

which is equivalent to

$$\lim_{x \to a} \left[ \frac{1}{g(x)} \right] = \frac{1}{\lim_{x \to a} g(x)}$$

To complete the proof, note that

$$\begin{aligned}
\lim_{x \to a} \left[ \frac{f(x)}{g(x)} \right] &= \lim_{x \to a} \left[ f(x) \cdot \frac{1}{g(x)} \right] \\
\lim_{x \to a} \left[ \frac{f(x)}{g(x)} \right] &= \left[ \lim_{x \to a} f(x) \right] \left[ \lim_{x \to a} \frac{1}{g(x)} \right]
\end{aligned}$$

$$\lim_{x\to a}\left[\frac{f(x)}{g(x)}\right] = \left[\lim_{x\to a} f(x)\right]\left[\frac{1}{\lim_{x\to a} g(x)}\right]$$
$$\lim_{x\to a}\left[\frac{f(x)}{g(x)}\right] = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$$

Some discussion of the proofs of Parts (d) and (e) of the previous theorem are warranted. They seemed very much like "rabbit-out-of-the-hat" tricks, which are very unsatisfying. Let's try to simulate the thinking that may have gone into the creation of such a proof.

In the case of the product rule for limits (Part (d) of the previous theorem), we know that the limits of $f$ and $g$ exist, and we are trying to prove that the limit of $fg$ exists. Thus, for a given $\varepsilon > 0$, we must figure out how to choose $\delta$ such that $0 < |x - a| < \delta$ implies that $|f(x)g(x) - L_1L_2| < \varepsilon$. Knowing that the limits of $f$ and $g$ exist, we already have conditions available on $L_1$ and $L_2$ separately, so we have to somehow manipulate the inequality $|f(x)g(x) - L_1L_2| < \varepsilon$ so that we can separate it into bits involving just $f$ and bits involving just $g$. Pay close attention to the trick of adding and subtracting an identical term, because a very similar trick will recur in the proof of the product rule for derivatives when you reach that point in your further calculus studies:

$$|f(x)g(x) - L_1L_2| = |f(x)g(x) - L_1g(x) + L_1g(x) - L_1L_2|$$

We can now apply the triangle inequality to the right side of the previous line to obtain

$$|f(x)g(x) - L_1L_2| \leq |f(x)g(x) - L_1g(x)| + |L_1g(x) - L_1L_2|$$

If we can make the right side of the previous line less than $\varepsilon$, this will guarantee that the right side of the line before that will also be less than $\varepsilon$. Continuing,

$$|f(x)g(x) - L_1g(x)| + |L_1g(x) - L_1L_2| = |g(x)| \cdot |f(x) - L_1| + |L_1| \cdot |g(x) - L_2|$$

We wish to ensure that the right side of the previous line is less than $\varepsilon$. Well, there are two terms; why don't we strive to ensure that each term is less than $\varepsilon/2$, so that the sum is guaranteed to be less than $\varepsilon$? OK, good. The second term is not too bad, because the first factor is constant and the second factor is "under control" — after all, the limit of $g$ is $L_2$, so we know that we can ensure that this factor is small. The first term is a little more complicated, because of the factor of $|g(x)|$; we know, however, that it is close to $L_2$ when $x$ is near $a$. We'll have to make this specific, though; "close" is way too vague for a proof. So we are encouraged to dig in there and see what we can do with this factor. Once we have bounded it, then we can work on the second factor of the first term, and then we can put it all together to construct a logically sound proof.

One might be tempted to say that the rest is details, but these details are vital! As you work your way through the details to understand the inner workings of the proof, remind yourself that the details in such proofs are not rigid; the choices made for the expressions for the various deltas has some wiggle room in them, and so other choices will also work. You might play with this to see how far you can push these choices.

We can state analogous laws for limits at infinity, one-sided limits, etc. For the right kind of person it will be a satisfying challenge to state and prove such laws. In particular, if you are planning to be a mathematics major at university, it will be a good challenge for you to try this now. Don't be discouraged if you find this difficult now! Straining against these difficulties now will make you smarter, and will make your future studies in mathematics a little more friendly.

## 12.5 The Squeeze Theorem

The squeeze theorem is a tool that is helpful for determining certain limits more easily than by using the definition of limit. Some examples follow after the statement of the theorem and its proof.

**THEOREM 7**

**Squeeze theorem: Suppose that** $g(x) \leq f(x) \leq h(x)$ **and that**

$$\lim_{x \to a} g(x) = L \qquad \textbf{and} \qquad \lim_{x \to a} h(x) = L$$

**Then**

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

Proof: Consider a given value of $\varepsilon > 0$. Because $\lim_{x \to a} g(x) = L$, there exists a $\delta_1 > 0$ such that for all $x$ that satisfy $0 < |x - a| < \delta_1$,

$$|g(x) - L| < \varepsilon$$

This condition is equivalent to

$$-\varepsilon < g(x) - L < \varepsilon$$

which is equivalent to (add $L$ to each term of the previous inequality)

$$L - \varepsilon < g(x) < L + \varepsilon$$

Using the same line of reasoning, you can show that there exists a $\delta_2 > 0$ such that for all $x$ that satisfy $0 < |x - a| < \delta_2$,

$$L - \varepsilon < h(x) < L + \varepsilon$$

Choose $\delta$ to be the minimum of $\delta_1$ and $\delta_2$. Thus, for all $x$ that satisfy $0 < |x - a| < \delta$, both of the following relations are satisfied:

$$L - \varepsilon < g(x) \qquad \text{and} \qquad h(x) < L + \varepsilon$$

Because $g(x) \leq f(x) \leq h(x)$, it follows that for the same values of $x$,

$$L - \varepsilon < f(x) \qquad \text{and} \qquad f(x) < L + \varepsilon$$

which is equivalent to

$$|f(x) - L| < \varepsilon$$

and this completes the proof. Can you sketch a graph and label it to make this proof more intuitive?

## GOOD QUESTION

**Squeeze theorem**

Could the domain over which the inequality in the squeeze theorem is satisfied be restricted and the theorem still remain valid? Play with this idea.

The following examples illustrate the utility of the squeeze theorem.

## EXAMPLE 37

**Using the squeeze theorem to determine a limit**

Determine the limit.

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right)$$

## SOLUTION

Because

$$-1 \le \sin\theta \le 1$$

it follows that

$$-1 \le \sin\left(\frac{1}{x}\right) \le 1$$

Multiplying each term of the previous inequality by $x^2$, we obtain

$$-x^2 \le x^2 \sin\left(\frac{1}{x}\right) \le x^2$$

Noting that

$$\lim_{x\to 0}\left(-x^2\right) = 0 \qquad \text{and} \qquad \lim_{x\to 0}\left(x^2\right) = 0$$

we can apply the squeeze theorem to conclude that

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right) = 0$$

Examining a graph of the function in the previous example will be helpful; see Figure 12.1, and you may wish to plot it yourself using your favourite software so that you can explore it in detail, zooming in and out. It's interesting to note that the function is not defined at $x = 0$, which means that the function has a hole discontinuity there. It's also interesting to note the asymptotic behaviour of the function, which is more apparent if you "zoom out" on the graph; see Figure 12.2. It appears that the graph is asymptotic to $y = x$. Is it clear that the sine factor approaches zero as $x \to \infty$? The quadratic factor grows without bound as $x \to \infty$, so it's interesting that the two growth rates are just right so that the product of the two factors grows approximately like $y = x$ as $x \to \infty$. Once you have a few more tools in your tool-box (to be discussed in your future calculus course) you will be able to convince yourself that $y = x$ really is an asymptote for this graph.

To understand the asymptotic behaviour of the graph as $x \to -\infty$, it's simplest to observe that sine is an odd function, and the quadratic factor is even, so overall the function in the graph is

odd. Is this enough to convince you that $y = x$ is also an asymptote when $x \to -\infty$?

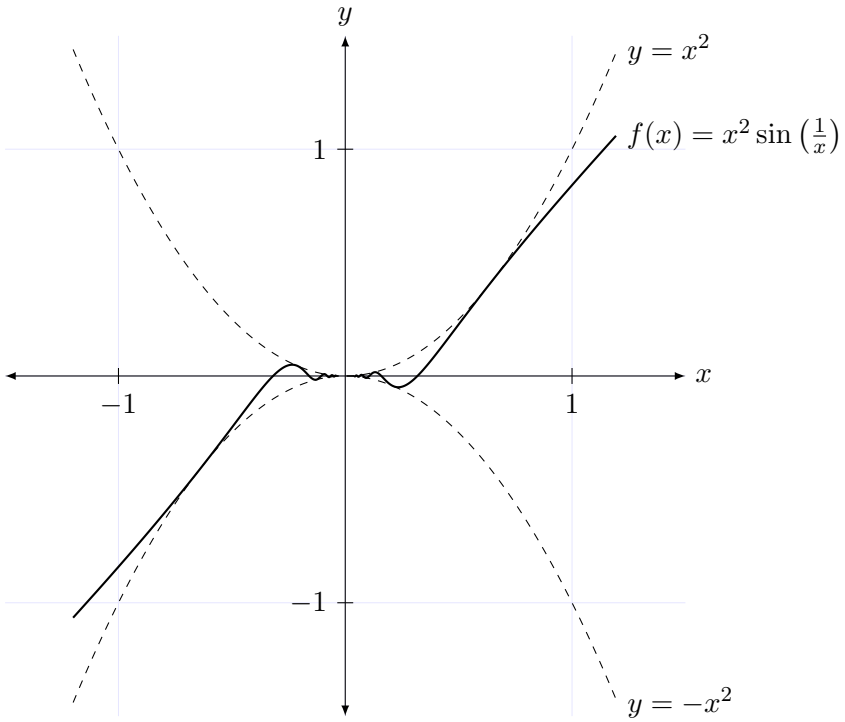



Figure 12.1: This strange function "wiggles" an infinite number of times near $x = 0$. The limit of this function as $x \to 0$ is zero.

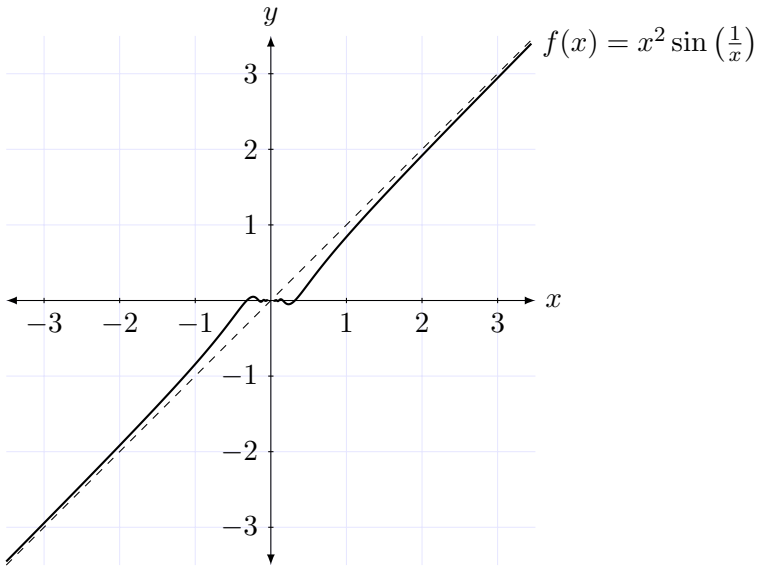



Figure 12.2: The asymptotic behaviour of this function is more apparent in this graph than in the previous one. A formula for the dashed line is $y = x$.

## DIGGING DEEPER

**Asymptotic behaviour of $f(x) = x^2 \sin\left(\frac{1}{x}\right)$**

You might use an electronic calculator to explore the following fact: For small angles, $\sin\theta$ is approximately equal to $\theta$, provided that you measure the angle in radians. (What is the approximate relationship if you measure the angle in degrees?) Furthermore, the smaller the angle, the better the approximation.

As $x \to \infty$, $1/x \to 0$, so for large values of $x$,

$$\sin\left(\frac{1}{x}\right) \approx \frac{1}{x}$$

and therefore

$$x^2 \sin\left(\frac{1}{x}\right) \approx x^2 \frac{1}{x}$$
$$x^2 \sin\left(\frac{1}{x}\right) \approx x$$

This provides further evidence that $y = x$ is indeed an asymptote for the graph of $f(x) = x^2 \sin\left(\frac{1}{x}\right)$.

A second line of argument will be apparent if you recall our introduction to power series in Chapter 10, which you will study more extensively in a future calculus course. Then you will be able to make the small-angle approximation $\sin x \approx x$ more precise:

$$\sin x \approx x - \frac{1}{6}x^3 + \frac{1}{120}x^5 - \cdots$$

If you now use the previous line to expand $f(x)$ in a power series, then its asymptotic behaviour will be evident.

A third line of argument will be convincing once you convince yourself of the interesting fact that

$$\lim_{x\to 0} \frac{\sin x}{x} = 1$$

This fact is proven with a geometric argument in virtually all first-year university calculus textbooks. Another way to write this fact is

$$\lim_{x\to 0} \left(\frac{1}{x} \cdot \sin x\right) = 1$$

Replacing $1/x$ by $y$, and observing that as $x \to 0$, $y \to \infty$, we can write this fact in an equivalent form as

$$\lim_{y\to \infty} y \sin\left(\frac{1}{y}\right) = 1$$

But what's in a name? A rose by any other name would smell as sweet, said Mr. Shakespeare, and it's the same story here. We could relabel $y$ in the previous equation by any other letter and it would still be valid. Relabeling $y$ by $x$ in the previous equation, we obtain

$$\lim_{x\to\infty} x\sin\left(\frac{1}{x}\right) = 1$$

Multiplying both sides of the previous relation by $x$, it is then quite plausible that as $x \to \infty$,

$$x^2\sin\left(\frac{1}{x}\right) \approx x$$

I wonder if it is possible to make this plausibility argument rigorously valid?

**EXAMPLE 38**

**Using the squeeze theorem to determine a limit**

Determine the limit.

$$\lim_{x\to 0} x^2\sin\left(\frac{1}{x^2}\right)$$

**SOLUTION**

Because

$$-1 \le \sin\theta \le 1$$

it follows that

$$-1 \le \sin\left(\frac{1}{x^2}\right) \le 1$$

Multiplying each term of the previous inequality by $x^2$, we obtain

$$-x^2 \le x^2\sin\left(\frac{1}{x^2}\right) \le x^2$$

Noting that

$$\lim_{x\to 0}\left(-x^2\right) = 0 \qquad \text{and} \qquad \lim_{x\to 0}\left(x^2\right) = 0$$

we can apply the squeeze theorem to conclude that

$$\lim_{x\to 0} x^2\sin\left(\frac{1}{x^2}\right) = 0$$

A graph of the function will be helpful; see Figure 12.3 and Figure 12.4. It will also be worthwhile for you to plot a graph of this function using your favourite software, so that you can explore the graph more fully.

The two previous examples illustrated functions with similar behaviour near $x = 0$, but the asymptotic behaviour of the two functions is different, and worth exploring. The function $f(x) = x^2\sin\left(\frac{1}{x^2}\right)$ appears to have a horizontal asymptote at $y = 1$, based on the graph. Using reasoning similar to the reasoning in the previous "Digging Deeper" feature box, does this seem reasonable?

You might like to generalize the previous two examples in various ways; for example, try different

powers of $x$, either in the argument of the sine function, or in the other factor. Exploring in this way will lead you to a much deeper understanding.

You might also think about how you would otherwise prove the limits in the two previous examples; this might lead you to appreciate the advantage of having a tool like the squeeze theorem.

**Question:** Can you think of some other examples of limits that are amenable to tackling by the squeeze theorem?

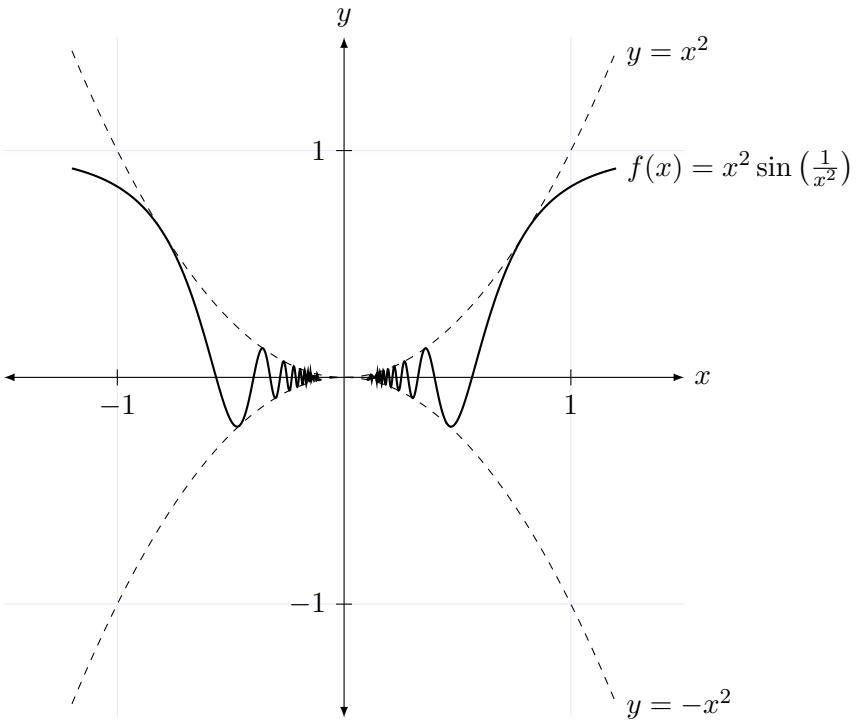



Figure 12.3: This strange function "wiggles" an infinite number of times near $x = 0$. The limit of this function as $x \to 0$ is zero.

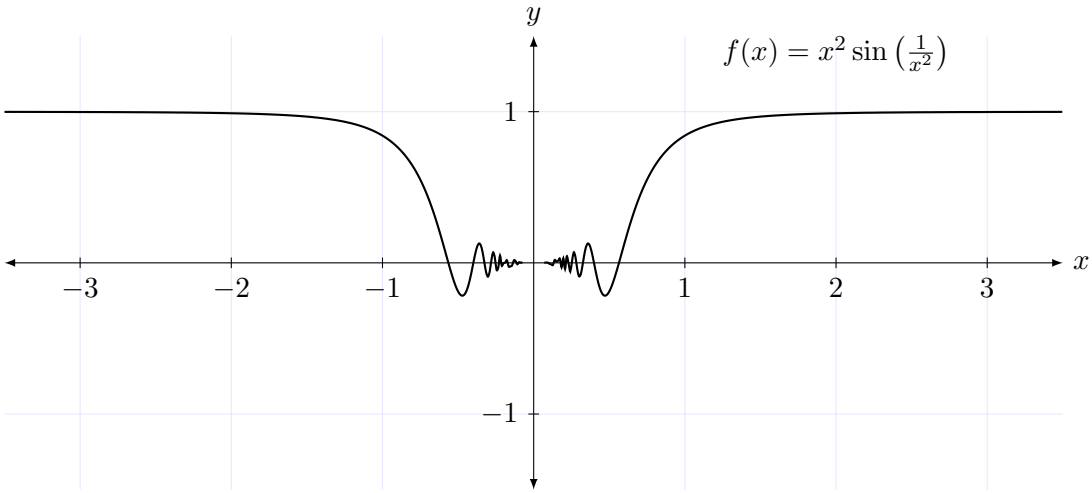



Figure 12.4: The asymptotic behaviour of this function is more apparent in this graph than in the previous one. It appears that there is a horizontal asymptote at $y = 1$.

## 12.6 Proofs of Some Theorems

In this section we provide formal proofs of some of the theorems that were quoted earlier in this book. We also state and prove some other useful theorems.

### 12.6.1 Differentiable Functions are Continuous

We discussed earlier in this book the fact that just because a function is continuous at a point **does not** mean that it is differentiable at that point; for example, the graph of the function might have a corner or cusp at that point. For example, recall the absolute value function, illustrated in Figure 12.5. The derivative of the function at $x = 0$ is defined to be

$$f'(0) = \lim_{h \to 0} \frac{f(h) - f(0)}{h - 0}$$

provided that this limit exists. We can evaluate the limit by examining the left and right limits separately. For the right limit,

$$\begin{aligned}
\lim_{h \to 0^+} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^+} \frac{f(h) - 0}{h - 0} \\
\lim_{h \to 0^+} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^+} \frac{|h|}{h} \\
\lim_{h \to 0^+} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^+} \frac{h}{h} \\
\lim_{h \to 0^+} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^+} 1 \\
\lim_{h \to 0^+} \frac{f(h) - f(0)}{h - 0} &= 1
\end{aligned}$$

For the left limit,

$$\begin{aligned}
\lim_{h \to 0^-} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^-} \frac{f(h) - 0}{h - 0} \\
\lim_{h \to 0^-} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^-} \frac{|h|}{h} \\
\lim_{h \to 0^-} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^-} \frac{-h}{h} \\
\lim_{h \to 0^-} \frac{f(h) - f(0)}{h - 0} &= \lim_{h \to 0^-} (-1) \\
\lim_{h \to 0^-} \frac{f(h) - f(0)}{h - 0} &= -1
\end{aligned}$$

Because the left and right limits are not equal, it follows that the derivative of the function at $x = 0$ does not exist, and so the function is not differentiable at $x = 0$. You can see from Figure 12.5 that the right limit and left limit that we just calculated represent the slopes of the two branches of the graph of the function. Because the two slopes are different, the two branches of the graph do not join smoothly at $x = 0$, and this sharp corner is a tell-tale sign that the function is not differentiable at $x = 0$.

To conclude this part of the discussion, just because a function is continuous at a point does not guarantee that the function is differentiable at that point. However, the converse is true; that is, if a function is differentiable at a point, then it is certainly continuous at that point. The following theorem summarizes this fact.

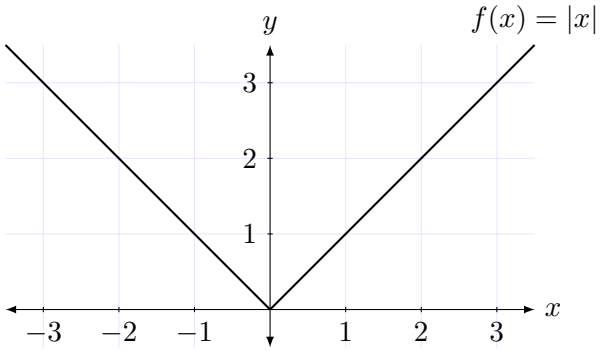



Figure 12.5: The function $f(x) = |x|$ is continuous at all values of $x$, but is not differentiable at $x = 0$, as explained in the text. Geometrically, the corner in the graph at the origin indicates that the function is not differentiable at $x = 0$.

**THEOREM 8**

**Suppose that the function $f$ is differentiable at $x = a$. Then $f$ is continuous at $x = a$.**

Proof: Consider the definition of the derivative of $f$ at $x = a$:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

Roughly speaking, asserting that $f$ is differentiable at $x = a$ is equivalent to saying the limit on the right side of the previous equation exists. But the denominator approaches zero as $x \to a$; how can the limit exist? The only way for this to happen is that the numerator must also approach zero as $x \to a$; but this is what we mean by the function $f$ being continuous at $x = a$.

So our strategy is to somehow isolate the numerator so that we can manipulate it into the condition for a function to be continuous. This involves separating the numerator and denominator, as follows:

$$f'(a) = \frac{\lim_{x \to a} (f(x) - f(a))}{\lim_{x \to a} (x - a)}$$

Expressing the limit of a quotient as the limit of the numerator divided by the limit of the denominator is justified by one of the limit laws. **Oops; no it's not in this case**. That limit law specifies that this move is valid if the limit of the denominator is not zero, and in this case the limit of the denominator *is* zero, so this argument is invalid.

OK, division failed, so let's try multiplication as a way of isolating the numerator. That is, let's multiply both sides of the starting equality by $\lim_{x \to a}(x - a)$:

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

$$\lim_{x \to a} (x - a)\, f'(a) = \left( \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \right) \lim_{x \to a} (x - a)$$

Now the two limits on the right side of the previous equation both exist, so we can use the product rule for limits to express the right side as:

$$\lim_{x\to a}(x-a)\,f'(a) = \lim_{x\to a}\left[\frac{f(x)-f(a)}{x-a}\cdot(x-a)\right]$$
$$\lim_{x\to a}(x-a)\,f'(a) = \lim_{x\to a}\left[f(x)-f(a)\right]$$

Now observe that the limit on the left side of the previous equation is zero. Therefore,

$$0 = \lim_{x\to a}\left[f(x)-f(a)\right]$$

Using a limit law on the right side of the previous equation, and then rearranging the result, we obtain

$$0 = \lim_{x\to a} f(x) - \lim_{x\to a} f(a)$$
$$\lim_{x\to a} f(x) = \lim_{x\to a} f(a)$$
$$\lim_{x\to a} f(x) = f(a)$$

The previous equation follows because $f(a)$ is a number (a constant function, if you prefer), and the limit of a number is itself. Because the previous equation is the definition of continuity, we can conclude that the function $f$ is continuous at $x = a$.

### 12.6.2 Common Functions are Continuous Where They are Defined

As we mentioned earlier in this chapter, the functions that are commonly used in mathematics and its applications at this level are all continuous where they are defined. Let's be a little more precise about this, and see which examples are relatively easy to prove.

Let's start with power functions, where the power is a whole number; that is, functions such as $y = 1$, $y = x$, $y = x^2$, and all such functions of the form $y = x^n$, where $n$ is a whole number.

**THEOREM 9**

**Continuity of power functions for whole-number exponents**

Functions of the form $f(x) = x^n$, where $n$ is a whole number, are continuous for all real values of $x$.

Proof: Suppose that $n = 0$. Then the function is $f(x) = 1$, which is continuous for all values of $x$. You can prove this by choosing a positive value of $\delta$ arbitrarily for a given positive value of $\varepsilon$.

Suppose that $n = 1$. Then the function is $f(x) = x$, which is continuous for all values of $x$. You can prove this by choosing $\delta = \varepsilon$ for a given positive value of $\varepsilon$.

For higher powers of $x$, we can repeatedly use the product rule for limits to show that the result is valid. For example, suppose that $n = 5$. Then we can write

$$\lim_{x\to a} x^5 = \lim_{x\to a} (x \cdot x \cdot x \cdot x \cdot x)$$
$$\lim_{x\to a} x^5 = \left(\lim_{x\to a} x\right)\left(\lim_{x\to a} x\right)\left(\lim_{x\to a} x\right)\left(\lim_{x\to a} x\right)\left(\lim_{x\to a} x\right)$$
$$\lim_{x\to a} x^5 = (a)\,(a)\,(a)\,(a)\,(a)$$
$$\lim_{x\to a} x^5 = a^5$$

By the definition of continuity, this shows that $f(x) = x^5$ is continuous at $x = a$.

The same reasoning can be used for any whole-number value of $n$, which completes the proof.

If you find this proof not rigorous enough for your liking, you can construct a proof using the principle of mathematical induction,[a] as follows. The cases $n = 0$ and $n = 1$ have already been proved above. Now suppose that the function $f(x) = x^k$ is continuous at $x = a$ for some natural number $k$, and we'll use this fact to show that the function $g(x) = x^{k+1}$ is also continuous at $x = a$.

Because $f$ is continuous at $x = a$, it follows that

$$\lim_{x\to a} x^k = a^k$$

Observe that

$$\lim_{x\to a} x^{k+1} = \lim_{x\to a} \left(x \cdot x^k\right)$$
$$\lim_{x\to a} x^{k+1} = \left(\lim_{x\to a} x\right)\left(\lim_{x\to a} x^k\right) \qquad \text{(by the product rule for limits)}$$
$$\lim_{x\to a} x^{k+1} = (a)\left(a^k\right) \qquad \text{(by the induction hypothesis)}$$
$$\lim_{x\to a} x^{k+1} = a^{k+1}$$

Thus, the function $g(x) = x^{k+1}$ is also continuous at $x = a$. By the principle of mathematical induction, all functions of the form $f(x) = x^n$, for all natural numbers $n$, are continuous at all real numbers.

[a] The principle of mathematical induction is typically not included in high-school mathematics instruction nowadays. If you study mathematics at university to a sufficiently high level, you will learn about mathematical induction. If you like, you may just skip the following discussion. Alternatively, you will have to learn about mathematical induction from some source before diving into the following discussion. If you are planning to become a mathematics major at university, then learning about mathematical induction now will give you a head start.

With the help of the previous theorem, and also with the help of the limit laws, it is possible to prove that all polynomial functions are continuous.

**THEOREM 10**

**Polynomial functions are continuous**

Each polynomial function is continuous at each real value.

Proof: Applying the limit laws to a polynomial function of degree $n$,

$$f(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x + b_0$$

we obtain

$$\begin{aligned}
\lim_{x\to a} f(x) &= \lim_{x\to a}\left(b_n x^n + b_{n-1}x^{n-1} + \cdots + b_2 x^2 + b_1 x + b_0\right)\\
\lim_{x\to a} f(x) &= \lim_{x\to a}\left(b_n x^n\right) + \lim_{x\to a}\left(b_{n-1}x^{n-1}\right) + \cdots + \lim_{x\to a}\left(b_2 x^2\right) + \lim_{x\to a}\left(b_1 x\right) + \lim_{x\to a}\left(b_0\right)\\
\lim_{x\to a} f(x) &= b_n \lim_{x\to a}\left(x^n\right) + b_{n-1}\lim_{x\to a}\left(x^{n-1}\right) + \cdots + b_2 \lim_{x\to a}\left(x^2\right) + b_1 \lim_{x\to a}\left(x\right) + b_0\\
\lim_{x\to a} f(x) &= b_n\left(a^n\right) + b_{n-1}\left(a^{n-1}\right) + \cdots + b_2\left(a^2\right) + b_1\left(a\right) + b_0\\
\lim_{x\to a} f(x) &= f(a)
\end{aligned}$$

Therefore, an arbitrary polynomial function of degree $n$ is continuous at an arbitrary value $x = a$. Thus all polynomial functions are continuous for all real values.

It will be good for you to work through each line of the proof of the previous theorem and identify which limit law or theorem was used at each step.

Next, it is possible to prove that all rational functions are continuous for all values of $x$ for which they are defined. (A rational function might have its denominator equal to zero (and therefore be undefined) at certain values of $x$; such a rational function might have a hole discontinuity or a vertical asymptote at such values of $x$.) Try proving this for yourself. Strive to state the theorem precisely, and then explore examples to determine whether there are any exceptions that would require you to restate the theorem more precisely. A strategy for proving this theorem is to express a rational function as a quotient of two polynomial functions, and then apply a limit law and invoke the theorem on the continuity of polynomial functions.

Next, it is a fact that algebraic combinations of continuous functions are also continuous. For example, if $f$ and $g$ are functions that are separately continuous at $x = a$, then the combinations $f + g$, $f - g$, and $fg$ are all continuous at $x = a$. Similarly, $f/g$ is continuous at $x = a$ provided that $g(a) \neq 0$. Once again, try proving these for yourself using the strategy of applying appropriate limit laws. The proofs should require no more than a few lines each.

### 12.6.3 Composition of Functions

Another very important algebraic process is the composition of functions, and theorems about how various other operations interact with the composition operation are therefore also important. For example, it is possible to interchange the order of operations when applying a limit operation with a function operation, provided that the function is continuous. The precise statement is in the following theorem.

**THEOREM 11**

**Interchanging limits and continuous functions**

If

$\lim_{x\to a} g(x) = L$ and $f$ is continuous at $L$, then

$$\lim_{x\to a} [f(g(x))] = f\left[\lim_{x\to a} g(x)\right]$$

Proof: Because $f$ is continuous at $L$, which means that

$$\lim_{y\to L} f(y) = f(L)$$

it follows that given $\varepsilon > 0$ there exists $\delta_1 > 0$ such that

$$0 < |y - L| < \delta_1 \quad \Longrightarrow \quad |f(y) - f(L)| < \varepsilon$$

Because

$$\lim_{x\to a} g(x) = L$$

it follows that given $\delta_1 > 0$ there exists $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |g(x) - L| < \delta_1$$

Identifying $y$ with $g(x)$, we can therefore conclude that given $\varepsilon > 0$ there exists $\delta > 0$ such that

$$0 < |x - a| < \delta \quad \Longrightarrow \quad |g(x) - L| < \delta_1 \quad \Longrightarrow \quad |f(g(x)) - f(L)| < \varepsilon$$

Thus, by the precise definition of a limit,

$$\lim_{x\to a} [f(g(x))] = f(L)$$
$$\lim_{x\to a} [f(g(x))] = f\left[\lim_{x\to a} g(x)\right]$$

The composition of continuous functions is continuous, as stated in the next theorem. Recall the definition of the symbol for composition of functions: $(f \circ g)(x) = f(g(x))$

**THEOREM 12**

**A composition of continuous functions is continuous**

Suppose that the function $g$ is continuous at $a$ and also that the function $f$ is continuous at $g(a)$. Then the composition $f \circ g$ is continuous at $a$.

Proof:

$$\lim_{x\to a} [f(g(x))] = f\left[\lim_{x\to a} g(x)\right] \qquad \text{(by Theorem 11, because } f \text{ is continuous at } g(a))$$
$$\lim_{x\to a} [f(g(x))] = f(g(a)) \qquad \text{(because } g \text{ is continuous at } a)$$

By the definition of continuous function, this means that $f \circ g$ is continuous at $x = a$.

### 12.6.4 Intermediate Value Theorem

The intermediate value theorem is a useful technical tool. After stating the theorem we apply the theorem in a technique (called the bisection method) for approximating the solution of a difficult equation.

You can get an intuitive sense for the theorem by imagining sketching a graph of a continuous function $f$ that satisfies these conditions: Suppose that $a < b$, that $f(a) < f(b)$, and also that $f(a) < d < f(b)$. Note that if you sketch the graph of $f$ on the interval $a \leq x \leq b$ your pencil will have to cross the horizontal line $y = d$ at some point on the interval. (Because the function is continuous, you must keep your pencil on the paper throughout the sketching process.) At this intersection point, the function value is equal to $d$, and this is the substance of the intermediate value theorem. See Figure 12.6 and try sketching a graph for yourself. You'll understand that it is not possible to sketch a continuous function graph joining the indicated points without the graph crossing the blue line.

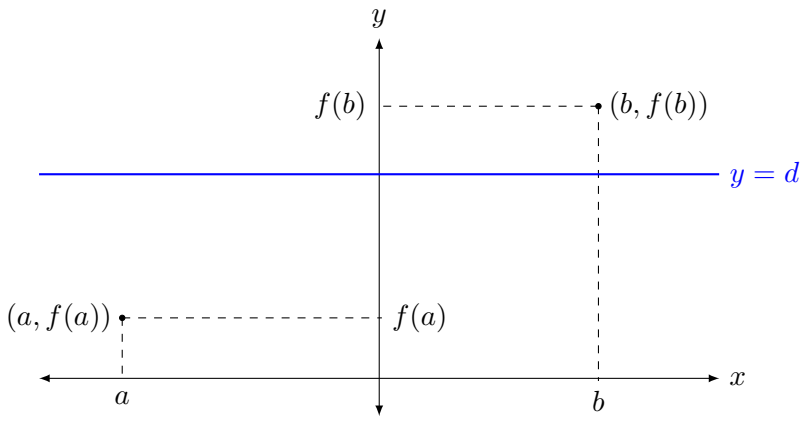



Figure 12.6: This figure provides you with intuition about the intermediate value theorem. If you sketch the graph of a continuous function connecting the two indicated points, the graph must cross the horizontal blue line.

After considering the graph for a while, I hope that you'll agree that the intermediate value theorem is indeed natural and believable. However, mathematicians have been burned over the centuries too many times by believing things that seem entirely natural, only to have a very clever colleague later construct an ingenious counter-example that showed their beliefs to be false. Therefore, one must not simply rely on feelings, although they may be a good start. In mathematics one must strive for clear and rigorous proofs of theorems.

**THEOREM 13**

**The intermediate value theorem**

Suppose that the function $f$ is continuous for all values of $x$ such that $a \leq x \leq b$, where $a \neq b$, and suppose that $f(a) \neq f(b)$. Choose any $y$-value, call it $d$, such that $d$ is strictly between $f(a)$ and $f(b)$. (That is, either $f(a) < d < f(b)$ or $f(b) < d < f(a)$.) Then there is at least one $x$-value, call it $c$, where $a < c < b$, such that $f(c) = d$.

Proof: The standard proof of this theorem depends on facts about infinite sequences that we have not studied yet, so we shall leave the proof for your further studies in calculus.

Although we have deferred the proof of the intermediate value theorem to a more advanced

course, some evidence for its validity can be obtained by its use in solving equations, as the following examples show.

### 12.6.5 The Bisection Method for Solving Equations

Almost all equations encountered at high-school level and below can be solved "analytically;" this means that there is a formula for the solution, and so in principle one can obtain an exact result. Of course, not all equations encountered in practical situations are like this; naturally what we learn in school begins with the easiest situations and then later builds towards the more challenging situations. What do scientists, engineers, mathematicians, or other practical people do when they encounter an equation that is too complicated to be solved by formula? Typically one turns to software tools, but someone had to program these software tools, and it is valuable to have some ideas about how such software is programmed, because you may need to adapt some such ideas in your own work down the road.

The branch of mathematics devoted to obtaining approximate solutions to equations that cannot be solved analytically is called "numerical methods," and entire books have been written about small slices of this sub-field. We clearly don't have room in an introductory book about limits to devote much time and space to this field, but we introduce the bisection method in this section, which is a simple idea for approximating the solution to a difficult equation. This idea is based on the intermediate value theorem.

As we have already described near the beginning of this book, the very best approximation schemes are those that can be improved by iterative (i.e., step-by-step) processes so that one can obtain as good an approximation as desired by performing as many iterations as needed. The bisection method is such a method.

We'll introduce the bisection method by considering an example. Suppose you would like to solve the equation

$$\cos x = x$$

There is no formula that provides an analytic solution; that is, of the form $x = \ldots$. Are there any solutions at all? This is an important first step: Do some preliminary analysis to convince yourself that there are indeed some solutions before you waste a lot of time searching for possible solutions that don't actually exist. The functions $y = \cos x$ and $y = x$ are familiar enough that a quick sketch will provide some guidance. After you draw a rough sketch, consult Figure 12.7.

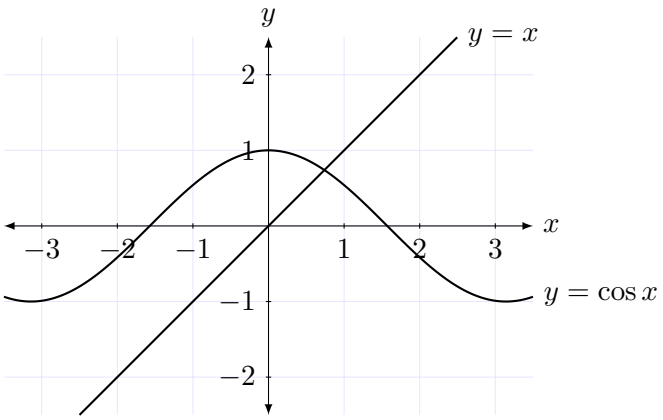



Figure 12.7: The $x$-coördinate of the intersection point of the graphs of $y = \cos x$ and $y = x$ represents the solution of the equation $\cos x = x$.

Based on our understanding of the properties of the graph of the cosine function (it is periodic), the figure makes it clear that there is indeed a solution to the equation, and that there is only one solution. It also appears from the figure that the solution is somewhere between $x = 0$ and $x = 1$.

The next step in finding the solution to the equation is to rearrange the equation into the equivalent form

$$\cos x - x = 0$$

This may seem like a pointless manipulation, but it will become apparent shortly that it does simplify our task a little bit. This is typically the case, and so it is common to rearrange all equations that need to be solved into the form (some combination of quantities) = 0. Figure 12.8 shows the graph of the function $g(x) = \cos x - x$. The solution to the equation $\cos x - x = 0$ corresponds to the $x$-coördinate of the point where the graph of the function $g(x) = \cos x - x$ intersects the $x$-axis.

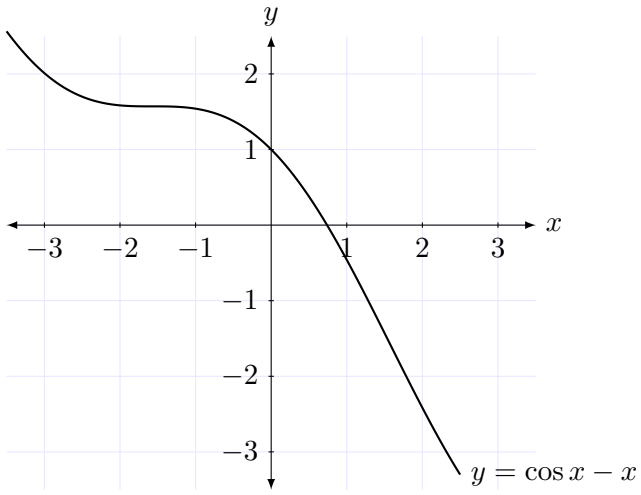



Figure 12.8: The $x$-coördinate of the point where the graph of $g(x) = \cos x - x$ intersects the $x$-axis represents the solution of the equation $\cos x = x$.

Now let's use the bisection method to approximate the solution of the equation $\cos x = x$. Remember that we are searching for a value of $x$ for which $g(x) = 0$. Using an electronic calculator (remembering to use radian mode), you can verify that (rounded to four decimal places)

$$g(0) = 1 \qquad \text{and} \qquad g(1) = -0.4597$$

Note that $g$ is a continuous function, and therefore we can apply the intermediate value theorem to $g$ on the interval between $x = 0$ and $x = 1$ to conclude that there is indeed a value of $x$ between 0 and 1 for which $g(x) = 0$. Now calculate the value of $g$ at the midpoint of the interval:

$$g(0.5) = 0.3776$$

Noting that $g(0.5)$ is positive and $g(1)$ is negative, we can apply the intermediate value theorem to the interval between $x = 0.5$ and $x = 1$ to conclude that there is a value of $x$ between 0.5 and 1 for which $g(x) = 0$. Next calculate the value of $g$ at the midpoint of this new interval:

$$g(0.75) = -0.0183$$

Because $g(0.5) > 0$ and $g(0.75) < 0$, we can apply the intermediate value theorem to the interval between $x = 0.5$ and $x = 0.75$ to conclude that there is a value of $x$ between 0.5 and 0.75 for which $g(x) = 0$. Once again we calculate the value of $g$ at the midpoint of the current interval:

$$g(0.625) = 0.1860$$

We can conclude that the solution lies between $x = 0.625$ and $x = 0.75$.

Continuing in this way for a few more steps, we obtain the following results. (Try it for yourself!)

$$\begin{aligned} g(0.6875) &= 0.0853 \\ g(0.71875) &= 0.0339 \\ g(0.734375) &= 0.0079 \\ g(0.7421875) &= -0.0052 \end{aligned}$$

Based on the steps performed so far, we can conclude that the solution to the equation $\cos x = x$ is between 0.734375 and 0.7421875. Thus, after seven bisections of the original interval, we can be certain about the first decimal place in the solution, but are uncertain about the second decimal place, which could be 3 or 4. The method appears to converge on the solution slowly, which is not great, but the method is very straightforward and easy to program.

Using software, one obtains the approximation

$$x \approx 0.739085$$

which is better than the approximation we have obtained so far with the bisection method. One wonders how many iterations of the bisection method would be needed to obtain this level of accuracy. If you are good at programming, you might pursue this question.

Although the bisection method converges only very slowly, it is fail-safe in the sense that as long as the initial interval contains only one solution to the equation, the bisection method is guaranteed to converge on the result.

Do you understand why rearranging the equation $\cos x = x$ into the form $\cos x - x = 0$ is helpful? Note that if we used the original form of the equation, at each step we would have to compare the values of the two sides of the equation to see which is greater. But one of the sides of the equation changes at each step, so one must be very attentive. Using the rearranged form of the equation, the task is simpler, because one only has to note the sign of the newly-calculated function value at the midpoint of the interval.

**EXERCISES**

(Answers at end.)

Use the bisection method to approximate the solution to each equation. Use your judgement to choose a reasonable starting interval and a reasonable number of iterations in each case. If you can program the method, then try a few "by hand" (that is, using an electronic calculator), and then try automating a few. Practicing both by hand and by writing a program will be useful to you.

1. $\sin x = x - 1$
2. $2^x = -x$
3. $x^3 + 2x = -1$
4. $x^4 + 5x = 3$
5. $3^x = x^2$
6. $2^x = x^2$

---

Answers: 1. $x \approx 1.93456$; 2. $x \approx -0.641186$; 3. $x \approx -0.453398$; 4. $x \approx -1.87572$ and $x \approx 0.577719$; 5. $x \approx -0.686027$; 6. $x = 2$, $x = 4$, and $x \approx -0.766665$

In summary, the intermediate value theorem is a technical result about continuous functions that provides useful tools for solving equations. The bisection method is conceptually simple, and also easy to implement. It is effective provided that the starting interval contains exactly one solution. If you are good at programming, you can enjoy programming this method.

It would be worthwhile to determine approximate solutions to various equations using both the bisection method and the Newton-Raphson method; the latter was studied in Section 10.5. After some practice, you will realize that each method has advantages and disadvantages. The bisection method is guaranteed to work, but it converges very slowly. The Newton-Raphson method often converges rapidly, but it doesn't work very well for some functions, and doesn't work at all for others. It is also sensitive to the initial guess.

The moral, as always, is that it is helpful to have many tools in your tool kit!

**HISTORY**

**Communication of research results over the millennia**

An important aspect of mathematics and science research is the communication and publication of discoveries.

People interested in learning have always found a way to get together with like-minded people, gathering to discuss, learn from each other, argue with each other, and do all the other things that humans do together. In ancient times there were academies where students who were able went to learn from great masters and to meet fellow students, and knowledge was passed on by word of mouth and through books. The first modern universities were founded about 1000 years ago. In medieval times in Europe, the church was a source of both academic learning and also a means of preserving and transmitting academic knowledge; one gets the image of monks devoting long, candle-lit evenings copying books by hand. The invention of the printing press in 1450 allowed many more books to be published, and this contributed to the rapidity of the dissemination of new ideas and new discoveries. In the Renaissance, scientists continued to communicate with each other by gathering together, but they also wrote letters to each other, and published their findings in books.

The first learned societies, in contrast to universities, were founded in the 1600s; for example, the Academy of the Lynx-like in Italy in 1603, the Royal Society of London for Improving Natural Knowledge (Royal Society for short) was founded in 1660, the French Academy of Sciences was founded in 1666, the Berlin Academy of Sciences in 1700, and so on. These and many other learned societies made it their point to gather top minds together, to facilitate discussion, and to provide a forum for the announcement and debate of research findings. Many of these learned societies also collected and published research papers, so that the findings could be spread more widely.

Eventually, independent journals arose that were not attached to learned societies or academies. The first that was devoted entirely to mathematics was the Journal for Pure and Applied Mathematics, founded by Leopold Crelle in 1826, but there were many others.

Nowadays, there are so many more people engaging in research, and so much research being produced, that there are an enormous number of research journals, across all the fields of mathematics and science. It is impossible to keep up with even a tiny fraction of research being produced, so one must be wisely selective. Since the 1970s, journals have used peer review as a kind of quality control to maintain high publishing standards. Each paper is sent out to one or more reviewers in the same field for their critical comments, and then the authors of the paper can respond. Journal editors then decide which papers are of sufficiently high quality that they merit publication. One of the problems with mathematics and science journals lately has been the rise of junk journals. As with everything, scams abound, and one must be careful.

There are also nowadays a number of popular magazines that are intended to explain current research advances in a simplified form suitable for a general audience. One can learn a lot from the good articles in these magazines, but quality varies, and it is sometimes difficult for a lay person to distinguish the metal from the dross.

In the recent past, researchers would send what are called pre-print papers (preprints, for short) to each other to speed up communication. These pre-print papers were submitted to journals for peer review, but then also sent to other researchers, with the understanding that they had not undergone the peer review process yet. The publishing process, including peer review, takes months or sometimes years, and it is useful to disseminate research results before the peer review process is finished, so that in-the-know researchers can learn and build on ideas without delay. Since 1991, it has become standard for researchers in mathematics, physics, and other fields to post the papers they submit to journals to an online server known as arXiv (a play on words of archive). This has replaced the concept of preprints in a more convenient form, allowing the entire community to see research immediately. As with journals, there is so much posted to arXiv every day that one must select wisely to read a very tiny subset. You can check out this scientific communications server here: `https://arxiv.org/` .

If you will be a mathematics or science researcher one day, it will be good for you to practice writing now. What you write does not have to be an original discovery, but by writing up some of the interesting things that you learn, in a way that would be understandable by your friends or family members, you will begin to learn and practice an essential aspect of mathematics and science research and scholarship: Clearly communicating something of importance.

# Chapter 13

# Review Exercises

## 13.1 Review Exercises Involving Calculation

Answers are found at the end.

1. Consider the function $f(x) = 2x^2 - 4x + 5$.
   (a) Determine the slope of the graph of $f$ at the point for which $x = 3$.
   (b) Determine an equation for the tangent line to the graph of $f$ at the point for which $x = 3$.
   (c) Determine a formula for the derivative of $f$.
2. Repeat the previous exercise for the function $f(x) = 5x^3 + 2x^2 - 3x + 7$ at the point for which $x = 2$.
3. Determine a derivative formula for each function.
   (a) $f(x) = \dfrac{5}{x}$
   (b) $f(x) = \dfrac{3}{x^2}$
   (c) $f(x) = \sqrt{x+4}$
   (d) $f(x) = \dfrac{1}{\sqrt{x}}$
   (e) $f(x) = x^2$
   (f) $f(x) = x^3$
   (g) $f(x) = x^4$
   (h) $f(x) = x^5$
   (i) $f(x) = x^3 + x^2$
   (j) $f(x) = x^5 - x^4$
   (k) $f(x) = 7x^2$
4. Determine each limit.
   (a) $\lim_{x\to 4} \left(x^2 - 5\right)$
   (b) $\lim_{x\to 3} \left(2x^3 - 1\right)$

(c) $\lim_{x\to 2}\left(\sqrt{x+1}\right)$

(d) $\lim_{x\to -1}\left(\sin(2\pi x)\right)$

(e) $\lim_{x\to -5}\left(\dfrac{1}{x+2}\right)$

5. Determine each limit.

(a) $\lim_{x\to -1}\left(\dfrac{x^2-1}{x+1}\right)$

(b) $\lim_{x\to 1}\left(\dfrac{x^2-1}{x+1}\right)$

(c) $\lim_{x\to 3}\left(\dfrac{x-3}{x^2-4x+3}\right)$

(d) $\lim_{x\to 2}\left(\dfrac{x^3-8}{x^2-x-2}\right)$

(e) $\lim_{x\to 9}\left(\dfrac{x-9}{\sqrt{x}-3}\right)$

(f) $\lim_{x\to 4}\left(\dfrac{x-4}{x-1}\right)$

(g) $\lim_{x\to 1}\left(\dfrac{x-4}{x-2}\right)$

6. Consider the function

$$f(x)=\left\{\begin{array}{lll} 4-(x-1)^2 & \text{if} & x>2 \\ kx & \text{if} & x<2 \end{array}\right\}$$

(a) Suppose that $k=5$. Determine

$$\lim_{x\to 2} f(x)$$

(b) Determine the value of $k$ for which the following limit exists.

$$\lim_{x\to 2} f(x)$$

7. Consider the function

$$f(x)=\left\{\begin{array}{lll} x^2+3x & \text{if} & x>-1 \\ 4-3x^2 & \text{if} & x<-1 \end{array}\right\}$$

Determine each limit.

(a) $\lim_{x\to -1^+} f(x)$

(b) $\lim_{x\to -1^-} f(x)$

(c) $\lim_{x\to -1} f(x)$

8. Consider the function

$$f(x) = \left\{ \begin{array}{lll} x^3 + 2x & \text{if} & x > 3 \\ 3x^2 + 6 & \text{if} & x < 3 \end{array} \right\}$$

Determine each limit.

(a) $\lim_{x \to 3^+} f(x)$

(b) $\lim_{x \to 3^-} f(x)$

(c) $\lim_{x \to 3} f(x)$

9. Determine any vertical and horizontal asymptotes for each function.

(a) $y = \dfrac{x}{x^4}$

(b) $y = \dfrac{x-1}{x^2}$

(c) $y = \dfrac{x+1}{x^2-1}$

(d) $y = \dfrac{x^2+5x+6}{x^2+3x+2}$

(e) $y = \dfrac{\cos x}{x}$

(f) $y = 5^{x-2} + 1$

(g) $y = \log_4 (x-3) + 2$

(h) $y = \dfrac{\sqrt{2x^2+3}}{x-1}$

10. Determine any vertical, horizontal, and slant asymptotes for each function.

(a) $y = \sqrt{x^2+2x+3} - x$

(b) $y = \dfrac{x^2+x}{x-1}$

(c) $y = \dfrac{x^3+2x^2}{x^2-4}$

(d) $y = \dfrac{x^3+2x^2-x-2}{x+2}$

(e) $y = \dfrac{x^3+2x^2-x-2}{x+1}$

(f) $y = \dfrac{x^3+2x^2-x-2}{x+3}$

(g) $y = \dfrac{x^3+2x^2-x-2}{x^2+3}$

11. Determine the following limit.

$$\lim_{x \to 0} \left( \frac{\sqrt{x^2+4}-2}{x^2} \right)$$

12. Determine whether each sequence converges. For the sequences that converge, determine their limits.

(a) $f(n) = 2 - \dfrac{5}{n^2}$

(b) $f(n) = 1 + \dfrac{1}{n}$

(c) $f(n) = 3 + n(-1)^n$

(d) $f(n) = -5 + \dfrac{(-1)^n}{n^2}$

(e) $f(n) = 3n - n^2$

(f) $f(n) = 5$

13. Determine the sum of each series.

(a) $25 + 125 + 625 + \cdots + 5^{13}$

(b) $3 + 1 + 3^{-1} + \cdots + 3^{-15}$

(c) $1 + 2 + 4 + \cdots + 2^{25}$

(d) $1 + (-2) + 4 + (-8) + \cdots + 2^{20}$

(e) $1 + \dfrac{1}{2} + \dfrac{1}{4} + \cdots + \dfrac{1}{2^{12}}$

14. Determine the sum of each series.

(a) $0.4 + 0.04 + 0.004 + 0.0004 + \cdots$

(b) $0.5 - 0.05 + 0.005 - 0.0005 + \cdots$

(c) $1 + \dfrac{1}{6} + \dfrac{1}{6^2} + \dfrac{1}{6^3} + \cdots$

(d) $\dfrac{1}{8} + \dfrac{1}{8^2} + \dfrac{1}{8^3} + \cdots$

(e) $1 + \dfrac{3}{2} + \dfrac{9}{4} + \dfrac{27}{8} + \cdots$

15. Express each repeating decimal number as a fraction.

(a) $0.474747\ldots$

(b) $0.131313\ldots$

(c) $12.571571571\ldots$

(d) $0.314231423142\ldots$

16. Use the Newton-Raphson method to estimate the solutions of each equation, correct to five decimal places.

(a) $x^2 + x - 5 = 0$

(b) $x^2 - x - 5 = 0$

(c) $x^3 - 5x + 2 = 0$

(d) $x - 7^{1/3} = 0$

(e) $x - \sqrt{24} = 0$

17. Guess the limit of each function, then use the formal definition of the limit of a function to prove that your guess is correct.

(a) $\lim_{x \to 5} (3x + 2)$

(b) $\lim_{x \to -2} \left(x^2 - 1\right)$

(c) $\lim_{x\to 3}\left(2x^2+4x-1\right)$

(d) $\lim_{x\to 4}\left(\dfrac{1}{x}\right)$

(e) $\lim_{x\to 5}\left(\sqrt{x}\right)$

(f) $\lim_{x\to -1}\left(\dfrac{2x+1}{x^2-1}\right)$

(g) $\lim_{x\to 2}\left(\dfrac{\sqrt{x}}{x+1}\right)$

(h) $\lim_{x\to 1}\left(mx+b\right)$

(i) $\lim_{x\to 1}\left(ax^2+bx+c\right)$

18. Use the bisection method to estimate the solution of each equation, correct to five decimal places.

(a) $\cos x = x+2$

(b) $x^5+x=1$

(c) $10^x = 5-x$

(d) $\log_{10}(x) = x-1$

(e) $x^2-2=5^x$

Answers to review exercises:

1. (a) 8 (b) $y=8x-13$ (c) $f'(x)=4x-4$

2. (a) 35 (b) $y=35x+14$ (c) $f'(x)=15x^2+4x-3$

3. (a) $f'(x)=-\dfrac{5}{x^2}$ (b) $f'(x)=-\dfrac{6}{x^3}$ (c) $f'(x)=\dfrac{1}{\sqrt{x+4}}$ (d) $f'(x)=-\dfrac{1}{x\sqrt{x}}$
(e) $f'(x)=2x$ (f) $f'(x)=3x^2$ (g) $f'(x)=4x^3$ (h) $f'(x)=5x^4$
(i) $f'(x)=3x^2+2x$ (j) $f'(x)=5x^4-4x^3$ (k) $f'(x)=14x$

4. (a) 11 (b) 53 (c) $\sqrt{3}$ (d) 0 $-\dfrac{1}{3}$

5. (a) $-2$ (b) 0 (c) $\dfrac{1}{2}$ (d) 4 (e) 6 (f) 0 (g) 3

6. (a) The limit does not exist. (b) $\dfrac{3}{2}$

7. (a) $-2$ (b) 1 (c) The limit does not exist.

8. (a) 21 (b) 33 (c) The limit does not exist.

9. (a) vertical asymptote $x=0$, horizontal asymptote $y=0$ (b) vertical asymptote $x=0$, horizontal asymptote $y=0$ (c) vertical asymptote $x=1$, horizontal asymptote $y=0$ (d) vertical asymptotes $x=-1$, horizontal asymptote $y=1$ (e) vertical asymptote $x=0$, horizontal asymptote $y=0$ (f) no vertical asymptote, horizontal asymptote $y=1$ (g) vertical asymptote $x=3$, no horizontal asymptote (h) vertical asymptote $x=1$, horizontal asymptotes $y=\pm\sqrt{2}$

10. (a) no vertical asymptote, horizontal asymptote $y = 1$, slant asymptote $y = -2x - 1$ (b) vertical asymptote $x = 1$, no horizontal asymptote, slant asymptote $y = x - 2$ (c) vertical asymptote $x = 2$, no horizontal asymptote, slant asymptote $y = x + 2$ (d) no asymptotes (e) no asymptotes (f) vertical asymptote $x = -3$, no horizontal asymptote, no slant asymptote (g) no vertical asymptote, no horizontal asymptote, slant asymptote $y = x + 2$

11. $\frac{1}{4}$

12. (a) converges, and the limit is 2 (b) converges, and the limit is 1 (c) diverges (d) converges, and the limit is $-5$ (e) diverges (f) converges, and the limit is 5

13. (a) $\frac{5^{14} - 5^2}{4} = 1\,525\,878\,900$ (b) $\frac{9 - 3^{-15}}{2}$ (c) $2^{26} - 1 = 67\,108\,863$ (d) $\frac{2^{21} + 1}{3} = 699\,051$ (e) $2 - \frac{1}{2^{12}}$

14. (a) $\frac{4}{9}$ (b) $\frac{5}{11}$ (c) $\frac{6}{5}$ (d) $\frac{1}{7}$ (e) diverges

15. (a) $\frac{47}{99}$ (b) $\frac{13}{99}$ (c) $12 + \frac{571}{999}$ (d) $\frac{3142}{9999}$

16. (a) $-2.79129$ and $1.79129$ (b) $-1.79129$ and $2.79129$ (c) $-2.41421$, $0.41421$, and $2$ (d) $1.91293$ (e) $4.89898$

17. (a) 17 (b) 3 (c) 29 (d) $\frac{1}{4}$ (e) $\sqrt{5}$ (f) the limit does not exist (g) $\frac{\sqrt{2}}{3}$ (h) $m + b$ (i) $a + b + c$

18. (a) $-2.98827$ (b) $0.75488$ (c) $0.63953$ (d) $0.137129$ (e) $-1.44818$

## 13.2 Conceptual Review Exercises: True or False

For each statement, state clearly whether the statement is true or false, then carefully explain your choice. If the statement is true, a good explanation could include a proof. If the statement is false, a good explanation could include a counter-example.

1. The "steeper" a line is, the greater its slope.
2. Infinity is a number that is larger than any imaginable number.
3. The rate of change of a quantity is a measure of the function values used to model that quantity; the greater the function values, the greater the rate of change.
4. The tangent line to the graph of a function intersects the graph at only one point.
5. Every function can be differentiated to produce its derivative function; that is, the derivative of every function exists for all values of $x$ in the domain of the function.
6. If a function is continuous at a point in its domain, then the function is differentiable at that point.
7. If the function $f$ has a discontinuity at $x = a$, then $\lim_{x\to a} f(x)$ does not exist.
8. If the function $f$ has a hole discontinuity at $x = a$, then $\lim_{x\to a} f(x)$ does not exist.
9. If the function $f$ has a jump discontinuity at $x = a$, then $\lim_{x\to a} f(x)$ does not exist.
10. If the function $f$ is continuous at $x = a$, then $\lim_{x\to a} f(x)$ exists, but the value of this limit may not be equal to $f(a)$.
11. If the function $f$ has a vertical asymptote at $x = a$, then $\lim_{x\to a} f(x)$ does not exist.
12. If the function $f$ has a horizontal asymptote at $y = b$, then $\lim_{x\to\infty} f(x)$ exists.
13. If the function $f$ has a horizontal asymptote at $y = b$, then $\lim_{x\to-\infty} f(x)$ exists.
14. If $f(a)$ exists, then $f$ is continuous at $x = a$.
15. The intermediate value theorem guarantees that, for example, if $a < b$ and $f(a) < f(b)$, then there exists $c$ such that $a < c < b$ and $f(a) < f(c) < f(b)$.
16. If a function has both a vertical asymptote and a horizontal asymptote, then it cannot also have a slant asymptote.
17. If $\lim_{x\to a} f(x)$ does not exist, then $f$ has a vertical asymptote at $x = a$.
18. A polynomial function can have a vertical asymptote.
19. A polynomial function cannot have a horizontal asymptote.
20. There are more integers than positive integers.
21. There are more real numbers than integers.
22. There are more real numbers than rational numbers.
23. There are more real numbers than the real numbers between 0 and 1.

24. The real numbers can, in principle, be placed in a list.

25. For an object moving in a straight line, the slope of the object's velocity-time graph indicates the direction of its motion.

26. If the infinite sequence $f(1), f(2), f(3), \ldots$ converges, then the corresponding infinite series $f(1) + f(2) + f(3) + \cdots$ also converges.

27. If the infinite sequence $f(1), f(2), f(3), \ldots$ converges to 0, then the corresponding infinite series $f(1) + f(2) + f(3) + \cdots$ also converges.

28. Every infinite geometric series converges.

## 13.3 Conceptual Review Exercises: Discussion Questions

Writing a few sentences about each of the following questions will remind you about key ideas of the chapter, and will test your understanding. If you have difficulty answering any of them, then review of the corresponding sections is warranted.

1. Discuss the connection between slope and rate of change.
2. Is the numerical procedure discussed in the early part of this chapter effective for determining the slope of any graph?
3. Discuss advantages and disadvantages of the numerical procedure for estimating the slope of a curve at a point.
4. What is a tangent line?
5. How is the derivative of a function defined?
6. What does it mean for a function to be differentiable at a point of its graph? Can you tell if a function is differentiable or not by examining its graph?
7. Compare and contrast jump discontinuities and hole discontinuities. Are there any other kinds?
8. In calculating a limit that arises from a calculation of a derivative, the numerator and denominator of the resulting quotient both approach zero. How does one get around this difficulty without violating the rules of algebra?
9. When calculating a limit, is it ever justified to simply substitute a value? Explain.
10. Is infinity a number? Explain.
11. If a limit is infinite, does this mean the limit exists? Explain.
12. What is a "ghost of a departed quantity?" What are their significance?
13. How are limits related to left-hand limits and right-hand limits.
14. What is the intermediate value theorem? What is one of its applications?
15. What is an asymptote? How can you determine asymptotes of various type?
16. Is it true that the graph of a function cannot intersect one of its asymptotes? Explain.
17. Are there various levels of infinity? Explain.
18. What is the formal definition of limit? Why is it needed?
19. What is the triangle inequality? How could you describe it in plain terms?
20. What is the squeeze theorem? What is its value?
21. Is every continuous function differentiable? Explain.
22. Is the product of two continuous functions continuous? Explain.
23. Is the quotient of two continuous functions continuous? Explain.
24. What is the bisection method? What is it used for?
25. What is the Newton-Raphson method? What is it used for?

# Suggestions for Further Reading

To practice and review your technical skills in preparation for university calculus, work your way through *Conquering Calculus: Post Secondary Preparation*, by Miroslav Lovric, Nelson, 2017 (also available for free at `https://archive.org/details/conqueringcalcul0000lovr/`). Daily practice, review, and repetition will ensure that your technical skills and background knowledge are super-sharp!

If you would like to continue studying university calculus, start working your way through a big university calculus book with lots of exercises. You can get such a book very inexpensively at a used book store, if you wish, or you could try a source of free online textbooks, such as this one: `https://openstax.org/details/books/calculus-volume-1`.

To learn more about the concepts of calculus, especially qualitative analysis, read *Calculus: The Analysis of Functions*, by Peter D. Taylor, Wall & Emerson, 1992 (also available for free at `https://archive.org/details/calculus00pete`). Absorbing this book will help you think like a mathematician and also help you to apply the concepts of calculus in practice.

If you are planning on majoring in mathematics at university, a good book to dive into is *Calculus for the Ambitious*, by Thomas W. Körner, Cambridge University Press, 2014. It's challenging, direct, honest, and contains quite a bit of humour. Absorbing this book will help you think like a mathematician and also help you to apply the concepts of calculus in practice.

To read further about the amazing mathematicians who developed calculus (and a lot of other mathematics as well), a great place to begin is the book *Calculus Gems: Brief Lives and Memorable Mathematics*, by George F. Simmons, McGraw-Hill, 1992 (also available for free at `https://archive.org/details/calculusgemsbrie0000simm_d7n9`).

For a much more extensive history of mathematics as a whole, a good source is *A History of Mathematics*, by Carl B. Boyer, Wiley, 1968 (or the 1991 edition, revised by Uta C. Merzbach, also available for free at `https://archive.org/details/isbn_9870471543976/`).

For a history of calculus specifically, read *The Historical Development of the Calculus*, by C.H. Edwards, Jr., Springer, 1979 (also available for free at `https://archive.org/details/historicaldevelo0000edwa`).

If you would like to read the book I first learned calculus from when I was in high school (*Calculus Made Easy*, by Sylvanus P. Thompson; I borrowed a copy from my local public library), you can find a free copy here: `https://calculusmadeeasy.org/`. An edition that has been revised and updated by Martin Gardner can be found here: `https://archive.org/details/calculusmadeeasy00thom_0`.

# Index