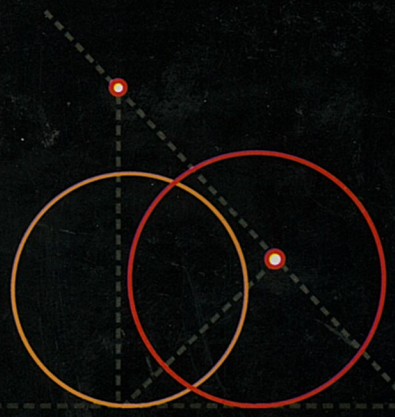
Catalog

Pathways to Real Analysis



Terrance J. Quinn • Sanjay Rai

Pathways to Real Analysis provides an introduction to several key ideas of real analysis, from Archimedes quadrature of the parabola, to the Calculus of Newton and Leibniz, power series, Cauchy's definitions of limit and integral, the inverse function theorem, the implicit function theorem, the wave equation, Fourier's heat equation and Fourier series. The book provides pathways of discovery that are mathematically natural. Examples are strategically selected in order to help the reader obtain the appropriate insights. Eventually, this initial understanding can be subsumed under a further context where one would explore and establish proofs in an axiomatic context. Prior to proving a result, however, it is helpful to first have discovered a result as a possibility. The main objective of this bookis to help promote that initial discovery, especially as relevant to the emergence of real analysis.

This is a mathematics book for college students and college teachers in science, technology, engineering and mathematics (STEM) and could serve as a supplement to a calculus sequence such as differential, integral and multi-variable calculus (Calculus I, II and III), or as a textbook for an introduction to real analysis.

Alpha Science International Ltd.
www.alphasci.com

ISBN 978-1-84265-576-4
9 781842 655764

Alpha Science

Pathways to Real Analysis

Quinn
Rai

Pathways to Real Analysis

Terrance J. Quinn
Sanjay Rai

Alpha Science

Pathways to Real Analysis

Terrance Quinn
Sanjay Rai

Alpha Science International Ltd.
Oxford, U.K.

Pathways to Real Analysis
160 pgs. | 65 figs.

Terrance Quinn
Department of Mathematical Sciences
Middle Tennessee State University
Murfreesboro, Tennessee, USA.

Sanjay Rai
Department of Mathematics
Montgomery College
Rockville, Maryland, USA.

ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

ISBN 978-1-84265-576-4

Printed in India

Terrance Quinn dedicates this book to the memory of his parents George Alphonsus Quinn and Bernice Frances Quinn and to Joan Donaghey.

Sanjay Rai dedicates this book to his wife Mamta Rai, his daughter Nandita Rai and to the memory of his father Shri Sarvadeo Rai.

Preface

This is a mathematics book for college students and college teachers in science, technology and mathematical sciences (STEM). The book could serve as a supplement to a calculus sequence such as differential, integral and multi-variable calculus (Calculus I, II and III), or as a textbook for an introduction to real analysis. We provide pathways of discovery that are mathematically natural. While we start the text with basic algebra, we lead the reader up to the inverse function theorem and the implicit function theorem for multi-variable calculus, and from there to an introduction to some of the ideas that led to the genesis of modern real analysis. Our approach is a special case of Discovery Based Learning. We invite the reader to think about clues and examples, strategically organized in order to help the reader reach the appropriate insights. This basic understanding is needed for competence in problem solving, in both pure and applied mathematics. Eventually, these initial discoveries can be subsumed into a further context where one would explore and establish proofs in an axiomatic context. Before proving a result, however, it is helpful to first have some grasp of the result as a possibility. The main objective of this book is to help promote this initial discovery, especially as relevant to the emergence of real analysis. The pace of the book is leisurely, for we have found that taking one's time is important in the learning process.

In Chapter 1 the student is led toward a basic understanding of invertible real-valued functions. Using a calculus-based approach, this is generalized to the inverse function theorem for higher dimensions. The multi-variable implicit function theorem also follows. A prerequisite for Chapter 1 is some basic calculus, such as taught in many first year college programs, or numerous high schools. Chapter 2 reviews a range of calculus results and concludes with an introduction to power series. Chapter 3 develops both the wave equation and the heat equation and ends with an introduction to Fourier series. Chapter 4, the Epilogue, includes a discussion that helps point to the rise of complex analysis.

Terrance Quinn
Sanjay Rai

Acknowledgments

Special thanks to Dr. Zine Bhoudhraa for creating the initial figures for the book.

Terrance Quinn extends thanks to Catherine Burnette, Gail Crips and Cyndi Smith, the staff of the Department of Mathematical Sciences at Middle Tennessee State University, for their constant help and support.

Sanjay Rai extends many thanks to Montgomery College for their generous support and assistance.

Rai also acknowledges US Department of Education grant number P116B060280, "Montgomery College Project Portal to Success", which helped support development of initial ideas for Chapter 1 of this book.

Contents

1

Discovering the Inverse Function Theorem

Topics: The main topic of the chapter is the Inverse Function Theorem. The culmination of the chapter invites the student to draw accumulating basic insights into a higher viewpoint known as the inverse function theorem of multi-variable calculus. The college algebra horizontal line test can then be seen to be a consequence of an intrinsic and more reaching geometry test that applies not only to graphs of real valued functions, but to general higher dimensional mappings. The Implicit Function Theorem emerges as a natural follow up result. Prerequisites for this chapter are college algebra; some trigonometry; and some familiarity with partial derivatives of real valued functions of two-variables.

1.1 ALGEBRAIC APPROACH

Example 1.1. See Figure (1.1) One mercury tube; two rulers, one scaled in °F and one scaled in °C. The temperatures of interest are from freezing to boiling, for water (at sea level say). By construction, both rulers start at 0°. Note that for °F, freezing to boiling is 32 to 212. For °C, freezing to boiling is 0 to 100.

Converting temperature °C to °F: Whatever the mercury level, one can look to either ruler to get the temperature in °F or in °C. Given °F, look to mercury, and then look to °C ruler; and vice versa. In other words for each °F there is a °C; and for each °C there is a °F.

Question 1.1. (Converting °C to °F). What is F for 1 °C? What does 1 °C mean on the scale? Note that 0 °C is freezing; 100 °C is boiling. So 1 °C is 1/100 of the way from freezing to boiling. Now look at the °F ruler. In Fahrenheit, 32 °F is freezing; 212 °F is boiling. There are 180 °F between freezing and boiling. **But °F and °C refer to same mercury tube!** So 1/100 of the way to boiling along °F ruler is 1/100 of 180. What then is °F corresponding to 1°C? Be a little bit careful here. No doubt you have the right start to the idea. But, what is the °F temperature at freezing? So, we need to add 32 °F. That is, in °F, 1 °C corresponds to [1/100 of (180)] + 32 on the °F scale.

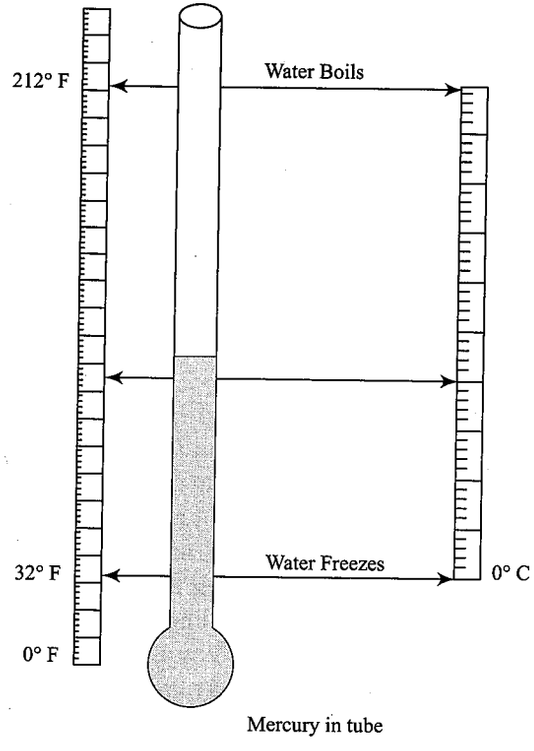


Figure 1.1

Notation. When we are discussing several conversions, it becomes cumbersome to be constantly writing the degree symbol "°". So, from now on, if we are looking for a quantity in a conversion problem, we can simplify the notation and just write F for Fahrenheit and C for Celsius.

Exercise 1.1. What is F when C = 7?; when C = 7.5?

Exercise 1.2. If C = 5, what is F?

Question 1.2. What is F for any C? That is, how do we convert any C to the corresponding F on the mercury tube?

Solution. Starting from 0 °C, note that C is a percentage of the way to boiling, namely C/100 of the way to boiling. On the °F ruler, there are 180 °F from freezing to boiling. So from where the mercury is at freezing to the mercury level at C, there are C/100[(180)] °F. But freezing starts at 32 F. Therefore the actual °F value on the °F ruler is F = C/100 (180) + 32. That is F = 1.8 C +32.

Converting Temperature °F to °C

Exercise 1.3. Using formula: If F = 100, find C. If F = 212, find C. (Use formula.)

Exercise 1.4. Given any F, what is C? That is, how do we convert any F to the corresponding C across the mercury tube?

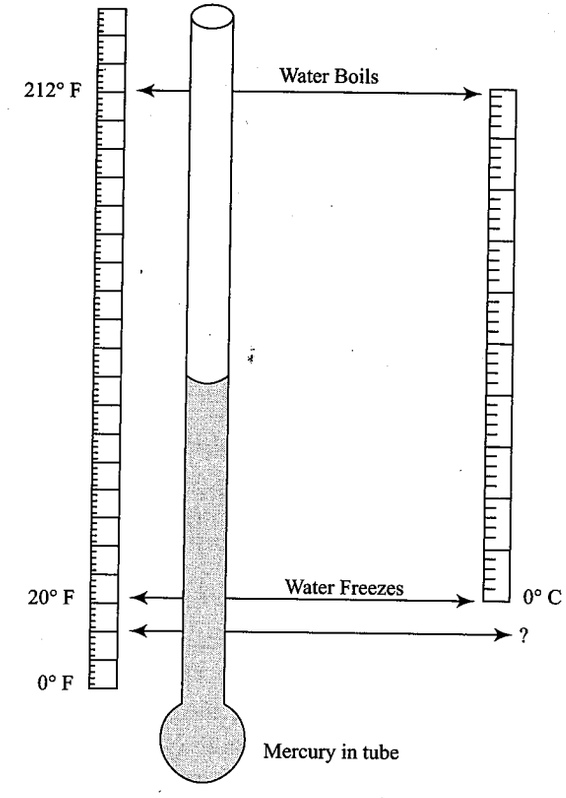


Figure 1.2

Converting Temperatures Below Zero

We have two rulers for the same mercury tube; and we have a formula that converts F to C; and vice versa, at least for temperatures from freezing to boiling. What though if the mercury drops below the freezing mark. Then it also drops below the bottom of the C ruler! How can we manage that situation?

A solution:

Turn the °C ruler around to measure the drop in mercury using the same scale. Numerically, we can track this by continuing the °C scale by naming downward distances using negative numbers.

Question 1.3. 20 °F is below freezing and so will be a negative C. Which one?

The formula we developed really only depended on what? A change of 1 °C is 1/100 of a change of 180 in °F; and then just keep track of the reference that 32 °F is freezing. In other words, the same approach can be used for temperatures below freezing.

Let's look at what we have. If $F = 20$, then $20 = 1.8C + 32$. Solving for C we get $C = -6.66$. The formula $F = 1.8C + 32$ produces F, if we know C. The formula $C = 1/1.8(F - 32)$ produces C, if we know F.

Question 1.4. Without doing any calculations, what should we get if we first convert one way, and then convert back? If we convert from $°C$ to $°F$ and then back again, what should be the result? If we convert from $°F$ to $°C$ and then back again, what should be the result?

Exercise 1.5. Now do the algebraic calculations to see whether or not the formulas produce the answer that you just gave. That is, does $C(F(C)) = C$?; and does $F(C(F)) = F$?

Exercise 1.6. Again, what does this mean on the mercury tube?

Remark 1.1. We have two formulas: Given C, we can convert C to F by $F = 1.8C + 32$.

Given F, we can convert F to C by $C = 1/1.8(F - 32)$. How can we **name** how these formulas are related to each other?

We have been talking about "converting" from one temperature scale to another.

You might know that the word "convert" comes from the Latin word ***verto***, which translates to "turn", or "turn around". The first formula $F = 1.8C + 32$ "turns the C around to F". And vice-*versa* for the other formula. So, to name how the two formulas relate to each other, we could say that each is "the turn around" of the other. Why? Because they convert, or "turn around" the degree readings from one temperature scale to the other. This even goes with what we need to do when we look at the two readings for the one mercury tube. We need to turn our attention, from one side of the tube to the other; and vice *versa*. Using the Latin again, but now for the noun rather than the action, each formula is "the turn around," the "in-verto" of the other, or in modern usage, the ***inverse*** of the other.

Exercise 1.7. In a physics lab experiment, starting at 0°F/-17.77°C, the temperature is slowly lowered to -100°F/-73.33°C. During the experiment, both scales descend below zero.

At the beginning of the experiment, the Celsius temperature is already well into the negatives, while the Fahrenheit temperature begins its descent at 0°F. Recall, though, that as the temperatures drop, the Fahrenheit readings change more rapidly than the Celsius readings, in a ratio of (1.8) : 1. So, even though the Celsius has a head start into the negatives, as both readings descend below zero, might there be a temperature where the reading on the Fahrenheit scale catches up (or rather "catches down") with the Celsius reading?

One way to answer this is to make use of our conversion formula, $F = 1.8C + 32$.

A temperature where both readings are the same would mean that we have an F and a C, with $F = C$ and $F = 1.8F + 32$ (or $C = 1.8C + 32$). Solving this equation, we get $F = C = -40$.

It is worthwhile to pause for a moment to think about what this means. Of course, degrees Fahrenheit and degrees Celsius are like apples and oranges. So, it doesn't make sense to say that – 40 of one is equal, as such, to – 40 of the other. And, not to worry, calculation does not say that. What then does our result mean?

Recall the meaning of the terms in the formula. They represent readings along two scales that were set up along a mercury tube. So, our solution $F = C = -40$ equates neither apples and oranges nor degrees Fahrenheit and degrees Celsius, but indicates that once the mercury has reached a certain point well below freezing, the readings on the two scales happen to be the same. Note also that our conversion formula also tells us how often this can occur. That is, it can happen only once, at $F = C = -40$.

Now, converting numbers from one measuring scale to another occurs in many settings besides temperature scales. So, let's look at another example.

Example 1.2. Imagine a parabolic mirror set up in a laboratory. See Figure 1.3.

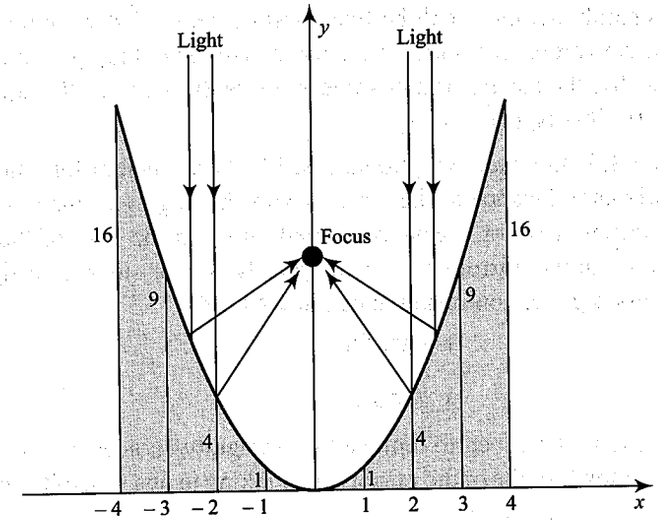


Figure 1.3

(One of the properties of a parabolic mirror is that it focuses incoming light)

Converting Distance to Height: The diagram gives a cross-section of the mirror. It has been constructed so that its height y is given by the distance x from the vertex by $y = x^2$. (By the way, why is the name "vertex" used for the lowest point on the parabola? *Clue*: That Latin word again!)

What is a convenient way to identify and distinguish points to the right of the vertex from points to the left? For example, $x = 2$ "to the right" vs. $x = -2$ "to the left"?

R: $x > 0$

Vertex: $x = 0$

L: $x < 0$

What is the height at $x = 2$? When $x = 2$, $y = (2)^2 = 4$.

What is the height $x = -2$? When $x = -2$, $y = (-2)^2 = 4$.

Converting Height to Distance: For the height $y = 4$, what is the horizontal distance from the vertex? Look at the diagram. Is there something misleading about the question? At height $y = 4$ there are two points on the mirror, and so two distances, $x = 2$ and $x = -2$. Let's see how the formula reveals this.

$$y = x^2$$
$$4 = x^2$$

so, $$x = \pm 2$$

In other words, unlike converting from one temperature scale to another, if we try to convert from height to distance from the vertex, then because the shape of the mirror is parabolic, there will be two possible vertex distances for any given height. From the vertex, one horizontal distance will be to the right and one to the left. Numerically, the square of a positive $x > 0$ is the same as the square of the negative $x < 0$. That is, $(x)^2 = (-x)^2$.

Question 1.5. Is there an exceptional case? Is there a height for which there is only one horizontal distance to the vertex? Now, having more than one possible answer to a question is not necessarily a bad thing. In the case of the mirror, it simply reflects that the mirror has two completely similar sides. The algebra also reveals this: For a given non-negative height $y \geq 0$, if

$$y = x^2, \text{ then}$$
$$x = \pm\sqrt{y}.$$

It is often useful, however, to remove the ambiguity/choice. We can do this by isolating the vertex (middle of the mirror) and looking to one side of the parabola at a time.

RHS: $y = x^2, x > 0$

Vertex/Middle: $y = 0, x = 0$

LHS: $y = x^2, x < 0$

Then to convert height to distance, we keep track of which side we are looking at:

LHS: Convert by $x = -\sqrt{y}$.

RHS:Convert by $x = +\sqrt{y}$.

Let's focus on the RHS for now. We convert distance to height by $y = x^2$; and we convert height to distance by $x = +\sqrt{y}$.

Question 1.6. Converting twice, first distance to height, and then height to distance gives what result?

Similarly, converting height to distance and then distance to height gives?

Exercise 1.8. Use the formulas to obtain the answer to the following question:

Question 1.7. Since the two formulas convert back and forth, what would be a good name for how the formulas relate to each other? (Answer: Again, the formulas are called *inverse* to each other.)

Exercise 1.9. Repeat the above exercise and questions for $y = (x - 3)^2$.

Exercise 1.10.

(i) Suppose that x converts to y by $y = x^2 - 2x + 1$.
Find how to convert y to x, that is, find the inverse formula or formulas.

(ii) Suppose that x converts to y by $y = 10x - 70$.
Find how to convert y to x, that is, find the inverse formula, or formulas.

1.2 GRAPHICAL APPROACH (GEOMETRY)

Let's look again at the example of the parabolic mirror. Suppose that we have a cross-section of the mirror, we have rulers, but we do not have a formula. Can we still in some practical way at least, convert distances to heights or heights to distances?

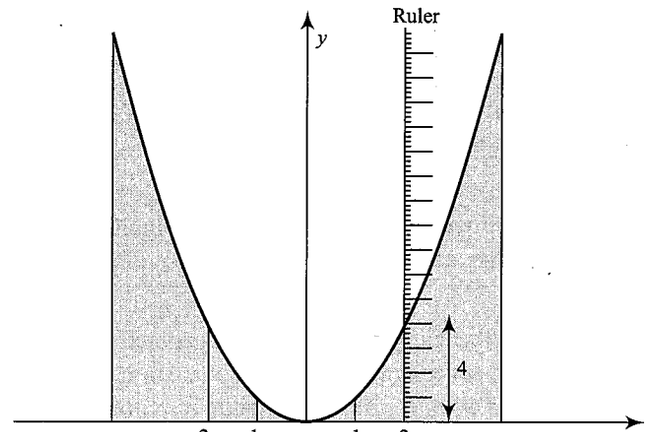


Figure 1.4

Suppose that at distance $x = 2$, the height of the mirror is 4. Then $x = 2$ converts to $y = 4$. Notice that this argument goes the other way, it can be turned around. Recalling more traditional Latin wording, we can say "conversely": The height $y = 4$ converts to $x = 2$ (at least on the RHS of the mirror!)

By the way, notice that the root of the word "conversely" is the same as for the word "inverse". In other words, if an argument is "turned around", then we can use the Latin name and say, "with - turning around", that is, "con – verto", or in modern english usage, *conversely*.

Example 1.3. Suppose that distances x and heights y are represented in the graph below.

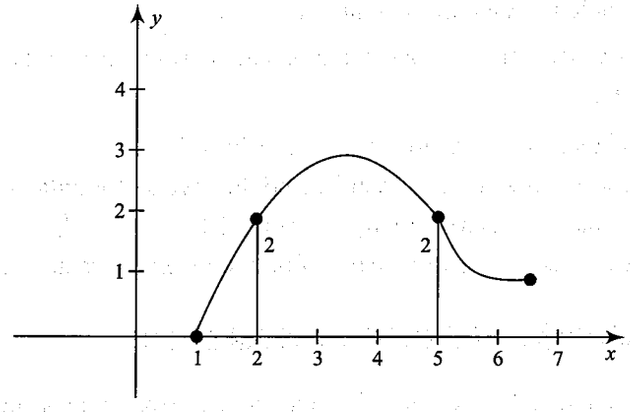


Figure 1.5

Using the rulers indicated in the diagrams, approximate answers to the following questions:

$x = 1$ converts to ? (0)
$x = 2$ converts to ? (2)
$x = 3$ converts to ? (2.75)
$y = 2$ converts to what?

There are two x's that convert to the same height $y = 2$. So again we are in a situation where if we try to convert, we get ambiguity. In that sense, it is said that there is no inverse, at least in the sense of providing one answer on the return.

Question 1.8. Consider the diagrams. Try to find a geometric feature or geometric property of the graph that reveals whether or not we can convert back and forth between x's and y's (that is without ambiguity, without multiple answers). There is ambiguity when there are two points at the same height relative to the x-axis. But, a clue: In geometry, two distinct points determine what? (Answer: A straight line.) A straight line that is parallel to the x-axis is also called horizontal.

Example 1.4. Using lines only, determine which (if any) of the following graphs could come from formulas that have inverses.

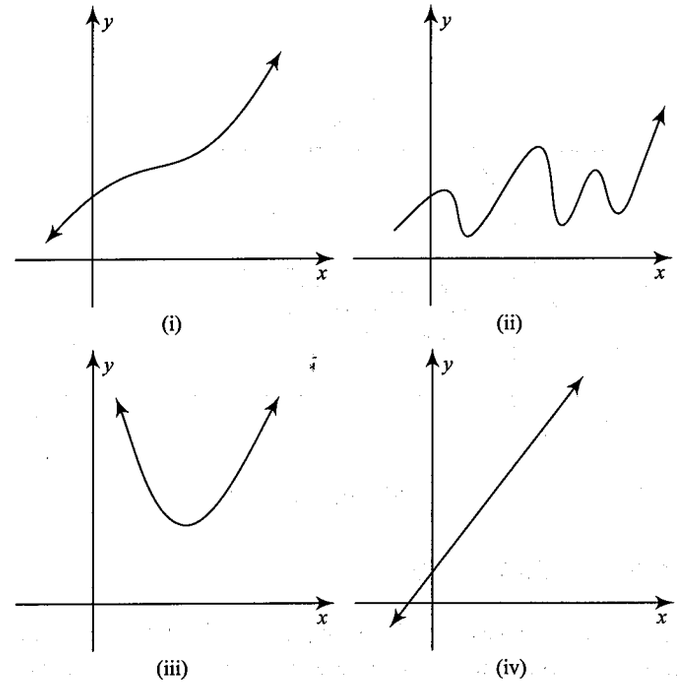
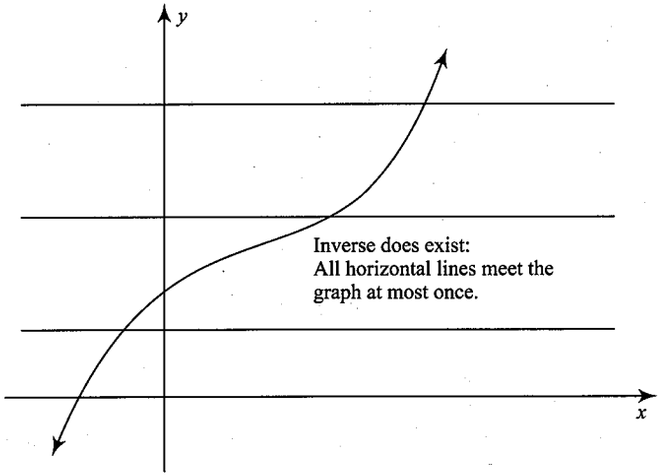


Figure 1.6

Can you put your idea into words? We come to what is often called the "horizontal line test": A graph has an inverse, or is "invertible", exactly when no horizontal line meets the graph in more than one place. Or, we can say it this way: A graph fails to have an inverse exactly when there is at least one horizontal line that meets the graph in more than one place. Notice that the horizontal line test is a graphical test for invertibility, but does not give the inverse formula itself.



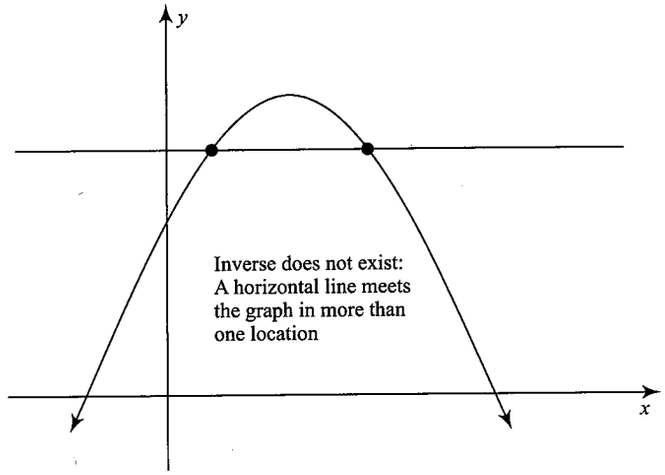


Figure 1.7

1.3 IMPLICIT FORMULAS

Exercise 1.11. Recall that we can write the equation of a line in a way that expresses in an equitable way both that x is related to y, and that y is related to x. For example, there is the line in the $x-y$ plane given by $5x + 7y = 12$. Graph this line. Find the formula that converts x to y. Find the formula that converts y to x.

Example 1.5. Recall that because of the Pythagorean formula, the equation of the circle of radius 1 centered at (0, 0) in the $x-y$ plane is $x^2 + y^2 = 1$.

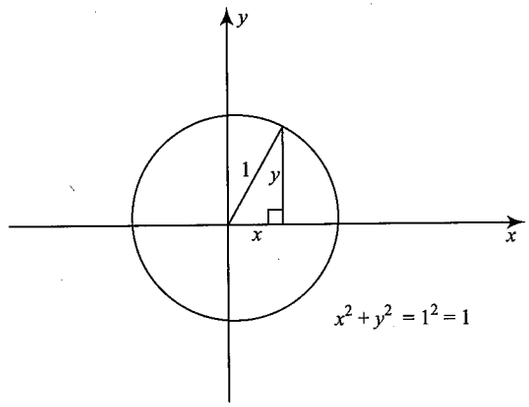


Figure 1.8

(a) Using the graphical approach (the horizontal line test) determine parts of the graph that have inverses.

(b) Using the algebraic approach, find invertible formulas and cases converting x to y.

(c) Compare and connect your answers from (a) with your answers from (b).

Exercise 1.12. Graph the ellipse given by $\frac{x^2}{36} + \frac{y^2}{4} = 1$. Answer the same questions (a), (b) and (c) as in the last example.

1.4 MORE VARIABLES

1.4.1 Rectilinear Plane Coordinates

Imagine a farm property bordered by two rivers, as in Figure (1.9).

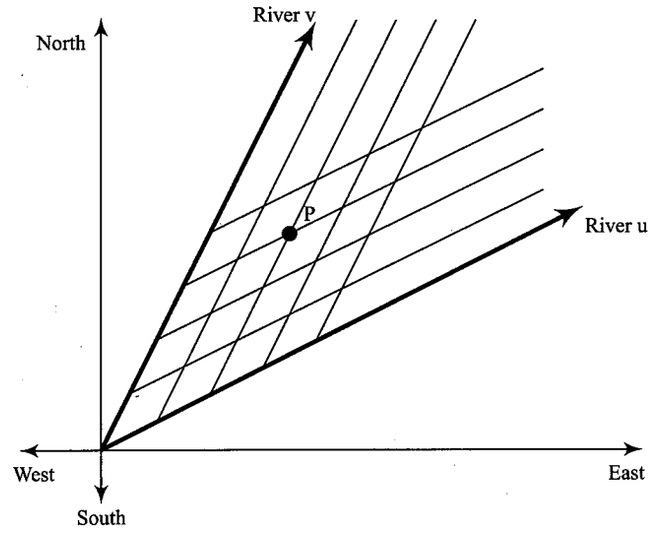


Figure 1.9

For the purposes of plowing and seeding, the farmer has for many years found it convenient to use the rivers as reference lines, for marking off the land. Each line is set apart by the distance needed for the tractor to make one pass. So, in order to account for all locations in the field, the farmer has recorded the locations in the field by a spread-sheet of pairs of numbers (u, v), or *river coordinates*. The first number u represents the distance from the river v in the direction parallel to river u; and the second number v represents the distance from the river u in the direction parallel to the river v. In the diagram, the location P, for example, is given by $(u = 2, v = 3)$.

Suppose that the state agency has required a determination of properties relative to standard geographic (x, y) north-south/east-west coordinates. The question then is how to convert the (u, v) spread sheet coordinates to geographic (x, y) coordinates.

Example 1.6.

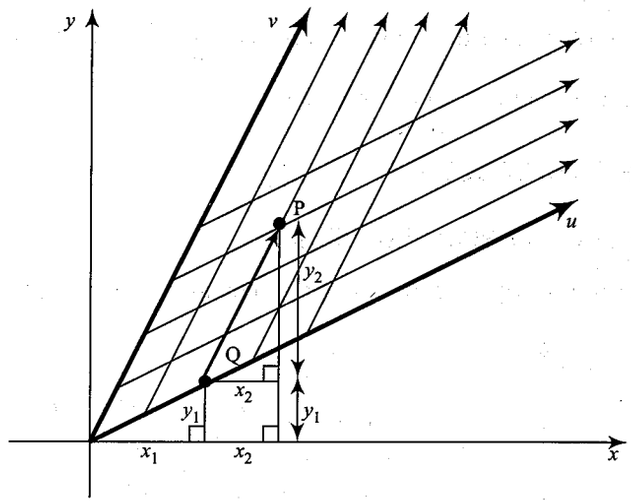


Figure 1.10

The location P is $(u, v) = (2, 3)$. We need to see what x (east-west) and y (north-south) coordinates are needed for that same location P. Let's do one geographic coordinate at a time, and so we can start with x.

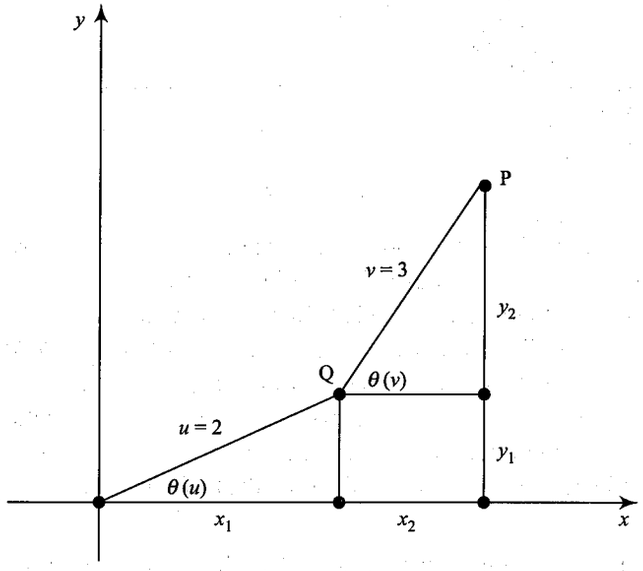


Figure 1.11

Both sets of coordinates emanate from a common vertex, or origin O -- the place where the rivers meet at the corner of the property. To reach P from O, move first in the direction of the u coordinate (parallel to the river u); and then from Q to P in

the direction of the v coordinate (parallel to the river v). The OQ segment and the QP segment each contributes to a change in the x-coordinate, x_1 and x_2 respectively. So we get that the total change in the x – coordinate is the sum $x_1 + x_2$.

Let's focus on just this special case for now. Let $\theta(u)$ be the angle between the u river and the x axis; and let $\theta(v)$ be the angle between the v river and the x axis. Looking to Figure 11, it is evident that x_1 depends on $\theta(u)$ and x_2 depends on $\theta(v)$. There are then the following questions:

$$x_1 = x_1(\theta(u)) = ? \quad \text{and } x = x_1 + x_2 = ? + ?.$$
$$x_2 = x_2(\theta(v)) = ?$$

And since similar reasoning applies to the y coordinate, we also have the questions

$$y_1 = y_1(\theta(u)) = ? \quad \text{and } y = y_1 + y_2 = ? + ?.$$
$$y_2 = y_2(\theta(v)) = ?$$

For the x coordinates, the definition of the cosine function gives

$$x_1 = 2 \cos(\theta(u))$$
$$x_2 = 3 \cos(\theta(v))$$

and so

$$x = x_1 + x_2 = 2 \cos(\theta(u)) + 3 \cos(\theta(v))$$

Exercise 1.13. $y = y_1 + y_2 = 2 \sin(\theta(u)) + 3 \sin(\theta(v))$

Exercise 1.14. For any (u, v) on the farm property, we can convert to geographic coordinates by the formulas

$$x = u \cos(\theta(u)) + v \cos(\theta(v))$$
$$y = u \cos(\theta(u)) + v \cos(\theta(v))$$

Remark 1.2. For an important special case, think of what this means if the rivers happen to be perpendicular to each other.

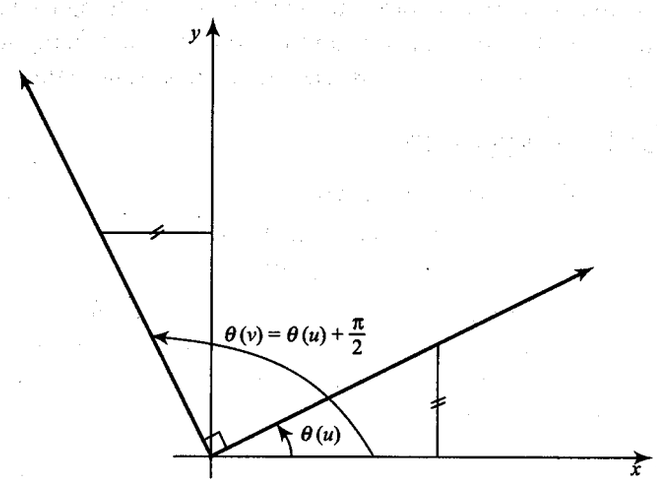


Figure 1.12

Then the angles are given by $\theta(v)=\theta(u)+\frac{\pi}{2}$. In other words, for this situation, the (u, v) coordinate system is obtained from the (x, y) coordinate system by simultaneously rotating both axes by the same angle $\theta(u)$.

Exercise 1.15. In the case $\theta(v)=\theta(u)+\frac{\pi}{2}$, and the formulas become

$$x = u\cos(\theta(u)) - v\sin(\theta(u))$$
$$y = u\cos(\theta(u)) - v\sin(\theta(u))$$

Hint: $\cos\left(\theta+\frac{\pi}{2}\right)=?$ and $\sin\left(\theta+\frac{\pi}{2}\right)=?$

This exercise gives the classical formulas for converting coordinates from one set of axes (u, v) to a set of axes (x, y) rotated clockwise by $\theta(u)$.

Exercise 1.16. Use the rotation formulas, together with similar geometric figures, to get the formulas

$$\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)$$
$$\sin(a+b) = \sin(a)\cos(b) + \cos(a)\sin(b)$$

Hint: Rotate perpendicular coordinate axes first by angle a and then by angle b. Use a similar break down as above, into geometric parts. There will be various right angle triangles stacked into the diagram. This is what accounts for there being sums.

We have solved the problem of converting the farm coordinates into geographic coordinates for the state. What about going the other way? If a state geographer gives the farmer geographic north-south/east-west coordinates (x, y), how can the farmer figure out the river coordinates for the corresponding locations on the farm? In other words, how can the farmer convert from geographic (x, y) coordinates to (u, v) coordinates? This leads to the next sub-section.

1.4.2 Algebraic Approach

We have two equations relating (x, y) and (u, v):

$$x = u\cos(\theta(u)) + v\cos(\theta(v))$$
$$y = u\cos(\theta(u)) + v\cos(\theta(v))$$

Here $\theta(u)$ and $\theta(v)$ are constants given by the angles that each of the rivers makes relative to the east-west axis. Can we solve these equations for u and v? Can this pair of equations be turned around, that is, inverted? In other words, can we find formulas for u and v given in terms of x and y, formulas of the form

$$\begin{aligned} u &= u(x, y) \\ v &= v(x, y) \end{aligned}\Bigg\}?$$

Exercise 1.17. Using elimination, show that

$$u\left[\cos(\theta(v))\sin(\theta(u)) - \cos(\theta(u))\sin(\theta(v))\right] = y\cos(\theta(v)) - x\sin(\theta(v))$$

$$v\left[\cos(\theta(u))\sin(\theta(v)) - \cos(\theta(v))\sin(\theta(u))\right] = y\cos(\theta(u)) - x\sin(\theta(u))$$

Now, recall that $\sin(a+b) = \sin(a)\cos(b)+\cos(a)\sin(b)$

So these equations can be rewritten as

$$u\left[\sin(\theta(u) - \theta(v))\right] = y\cos(\theta(v)) - x\sin(\theta(v))$$

$$v\left[\sin(\theta(v) - \theta(u))\right] = y\cos(\theta(u)) - x\sin(\theta(u))$$

This can be solved for u and v if and only if $\sin(\theta(u)-\theta(v))\neq 0$, that is, if and only if $\theta(u)-\theta(v)\neq k\pi$, for any integer k.

1.4.3 Geometric Approach

What does the condition $\theta(u) - \theta(v) \neq k\pi$ mean about the rivers? In particular, what would it mean about the rivers if $\theta(u) = \theta(v)$? Or $\theta(u) = \theta(v) + \pi$?

Let's look at what happens to the $u - v$ spread-sheet pairs (u, v) as we consider different cases.

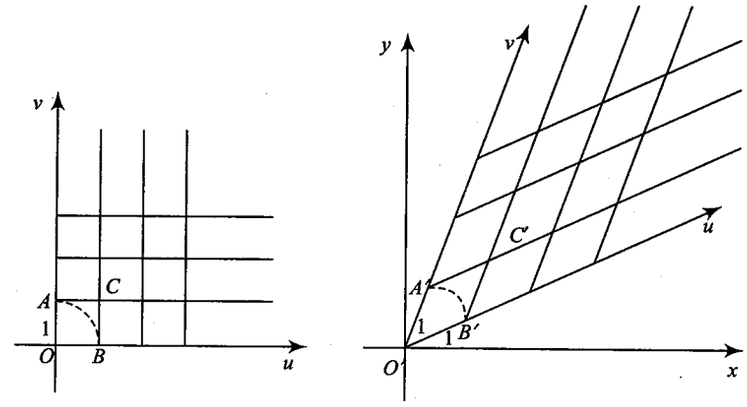


Figure 1.13

$$\theta(v) = \theta(u)+\frac{\pi}{8}$$

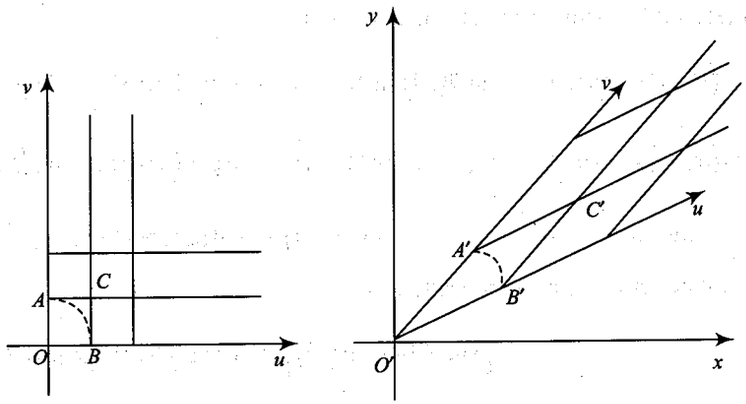


Figure 1.14

$$\theta(v) = \theta(u) + \frac{\pi}{16}$$

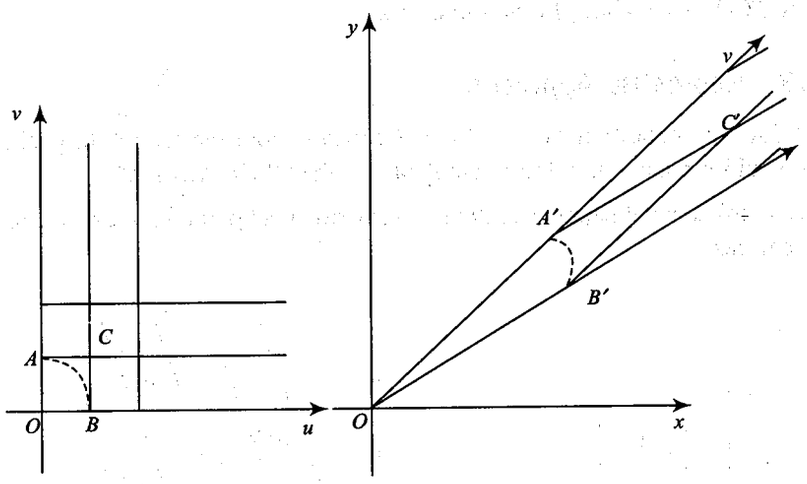


Figure 1.15

In each case the unit square OACB in the u-v spread sheet converts to a parallelogram O′A′C′B′. In Figures (1.13)-(1.15), what is happening to the parallelogram O′A′C′B′ in the x-y plane, as $\theta(v)$ gets close to $\theta(u)$? The pairs (u, v) from the *arc* (AB) convert to the pairs of (x, y) on the *arc* (A′B′). As $\theta(v)$ gets close to $\theta(u)$ what happens to the length of the *arc* (A′B′)? If $\theta(u)$ reaches $\theta(v)$, then all of the many pairs (u, v) along the *arc* (AB) convert down to the single location A′ = B′. So, in this case, if we were to try to invert from (x, y) to (u, v), we would run into a problem, at least mathematically. For given (x, y) = A′ = B′, there are then many (u, v) spread-sheet pairs which convert to (x, y) = A′ = B′. Now, arc length can be difficult to calculate. Instead of arc length, what other geometric quantity could detect the approach to "collapse" that occurs as $\theta(u)$ approaches $\theta(v)$?

Clue: Look to the grid lines.

Answer: The area of the parallelogram.

As can be seen in the diagrams, as long as the area of the new parallelogram is greater than zero, then the (x, y) locations in the parallelogram O′A′C′B′ uniquely convert back to (u, v) pairs in the square OACB. We loose uniqueness exactly when the parallelogram O′A′C′B′ collapses to a line segment O′C′, with A′ = B′.

Evidently, what happens to the unit square OACB in the spread-sheet is an indicator for what happens to the rest of the spread-sheet coordinate squares. In fact, the conversion of (u, v) to (x, y) can be inverted if and only if the area of the parallelogram O′A′C′B′ is positive.

1.4.4 Geometric Approach with Coordinates

Perhaps the discussion above has provided enough material for us to go on to a more general question. Suppose (u, v) coordinates convert to (x, y) coordinates by a (linear) equation of the form

$$x = au + bv$$
$$y = cu + dv$$

Just as for the (u, v) spread-sheet and farm field example, we can represent this in a diagram.

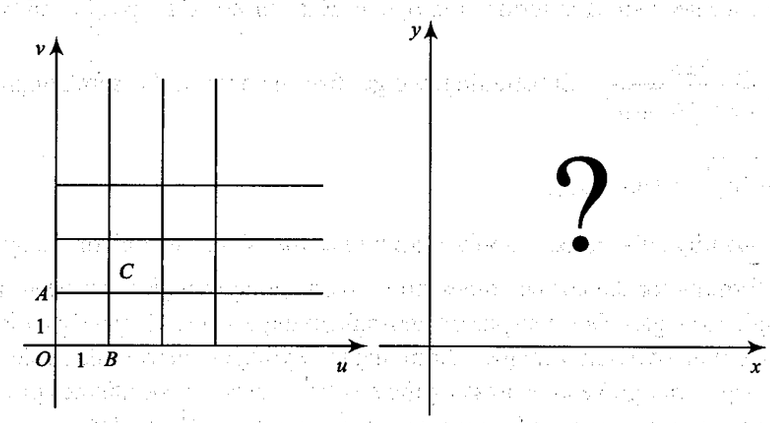


Figure 1.16

What do we draw in the (x, y) plane? How do we convert the u and v coordinate lines? How do we convert the unit square? The point A is given by $(u = 0, v = 1)$. The point B is given by $(u = 1, v = 0)$. The equation converts B to (a, c) in the (x, y) coordinates and converts A to (b, d) in the (x, y) coordinates. As the reader may recall (or calculate), since the equation is linear, all multiples of the line OB therefore convert to line segments parallel to (a, c) in the (x, y) coordinates. In the same way, the line segments along OA convert to line segments along (b, d).

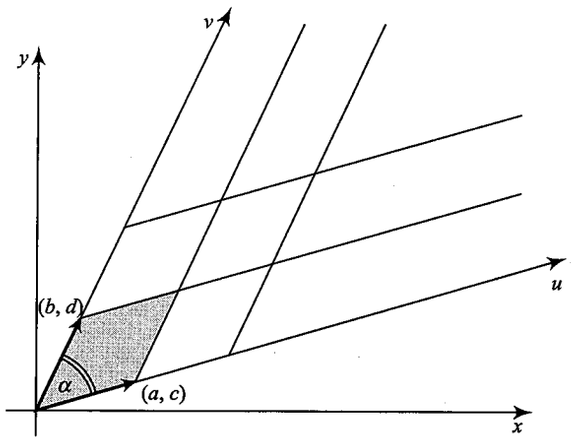


Figure 1.17

Exercise 1.18. Calculate the area of the parallelogram O′A′C′B′ determined by (a, c) and (b, c), in terms of the coefficients a, b, c, d.

Outline: Recall that the area of a parallelogram is "base times height". Then area of the parallelogram is therefore $\sqrt{a^2+c^2}\sqrt{b^2+d^2}\sin\alpha$. To complete the exercise, we need $\sin\alpha$ in terms of a, b, c, d. But, $\sin\alpha = \sqrt{1-\cos^2\alpha}$ and $\cos\alpha = \dfrac{ab+cd}{\sqrt{a^2+c^2}\sqrt{b^2+d^2}}$. Substituting, we get that the area of the parallelogram is $\sqrt{(ad-bc)^2} = |ad-bc|$.

The quantity $ad - bc$ commonly is known as the "determinant" of the system.

The formula for the cosine comes from investigating triangles that need not be right angled triangles. One may place a non-right triangle inside a larger right triangle; and then obtain two right triangles. Applying the Pythagorean formula twice (once for each right triangle) and expressing the results in terms of coordinates produces the well-known dot product formula $(a,c)\cdot(b,d) = \|(a,c)\|\,\|(b,d)\|\cos\alpha$.

Exercise 1.19. Recall the algebraic approach for the farm field problem. In the same way, use elimination to show that the system of equations

$$x = au + bv$$
$$y = cu + dv$$

can be solved for u and v if and only if $ad - bc \neq 0$. In summary, we have two equivalent formulations, that is, the geometric and the algebraic. In other words, the following are equivalent:

(i) The pair of equations $\begin{aligned} x &= au + bv \\ y &= cu + dv \end{aligned}$ can be inverted;

(ii) The area of parallelogram determined by (a, c) and (b, d) is not zero (Geometric);

(iii) $ad - bc \neq 0$ (Algebraic)

1.4.5 Three and More Variables

Suppose that we have a 3-*D* conversion formula (u, v, w) to (x, y, z) given by three equations of the form

$$x = a_{11}u + a_{12}v + a_{13}w$$
$$y = a_{21}u + a_{22}v + a_{23}w$$
$$z = a_{31}u + a_{32}v + a_{33}w$$

Exercise 1.20. Draw 3-*D* grid lines and a unit box in the (u, v, w) axes. Draw the parallelepiped to which this unit box is associated through the equations.

Clues: What are the vertices of the parallelepiped?

As you might have already figured out, the three equations convert the unit box in the (u, v, w) coordinates to a parallelepiped with vertices in the (x, y, z) coordinates are given by the edges $\begin{bmatrix} a_{11} \\ a_{21} \\ a_{21} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}, \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}$.

The volume of a parallelepiped is (area of base) times (height). Select a base. Then the height is the length of the edge rising up to the height (hypotenuse) times the sine of the angle of elevation of that length, relative to the selected base. Hence, we can use the cross-product and the dot product to calculate the volume in terms of the coefficients $a_{11}, a_{12}, \ldots, a_{33}$. It will be helpful for the student to work this out.

Can you now give both geometric and algebraic formulations for when the equations can be inverted? What is the common name given to the volume quantity when it is given in terms of the coefficients $a_{11}, a_{12}, \ldots, a_{33}$?

Answer: The *determinant*, typically written as $\det\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{21} & a_{32} & a_{33} \end{bmatrix}$.

Remark 1.3. The more advanced reader may recall that there are *determinant* formulas for two equations in two variables; three equations in three variables;

four equations in four variables; five equations in five variables; and so on. The general determinant formula is a sum of products of matrix coefficients, in certain combinations, multiplied in an alternating pattern, by + or – .

One way to obtain the general formula is to identify the key properties of "area" and "volume" from 2-*D* and 3-*D*, and to abstract those properties to define an "n-dimensional volume function". For example, in both 2-*D* and 3-*D*, doubling the length of one edge of a parallelogram or parallelepiped doubles the area and volume respectively; halving the length of any one edge results in half of the area and volume respectively; and so on. Another key property can be found by recalling that the area of a rectangle is unchanged when one side is translated along a parallel direction to produce a parallelogram. For then the base and height remain the same. This extends naturally to the 3-D case as well. That is, parallel translation of any side of a box leaves the total volume unchanged.

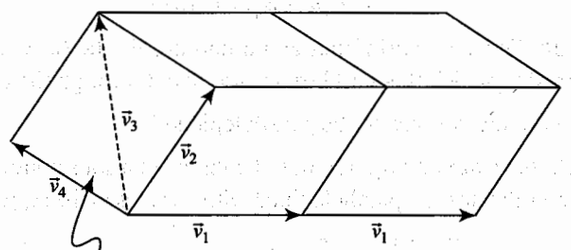


Figure 1.18

Exploring these properties further, it becomes evident that a general "*n*-dimensional volume function" should be linear in each component – called "multilinear". Note also that through both geometry and applications, the orientation of edges of a parallelogram or parallelepiped leads in a natural way to the notion of "signed area" and "signed volume" respectively. In order to generalize this property to "*n* dimensions" requires that the volume function be "alternating". That is, a switch of any two edges changes the algebraic sign of the "*n*-dimensional volume". An "*n*-dimensional volume function" will therefore be (i) "multilinear" and (ii) "alternating".

To go into more detail here would take us beyond the scope of this introductory book. For more details, please consult one of the standard linear algebra texts. If though we apply these two properties to a 4×4 matrix, then we get the formula for a 4×4 determinant; if we apply these properties to a 5×5 matrix, then we get the formula for a 5×5 determinant; and so on. In each case, we use properties (i) and (ii) to successively factor out the coefficients of the matrix, until we are left

with the traditional formula for the determinant, multiplied by the *n*-dimensional volume of the unit cube. If we define the unit cube to have unit 4-volume/5-volume/*n*-volume, then we obtain the traditional determinant formula. It can be proven by induction.

Just as in the special cases of a 2×2 or 3×3 matrix, it is possible to show algebraically that the determinant provides a criterion for when a system of equations can be inverted. The geometric formulation requires that the unit *n* box determined

by the edges $\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix},\ldots,\begin{bmatrix}0\\\vdots\\0\\1\end{bmatrix}$ does not collapse under conversion. That is, the

parallelepiped determined by*n* edges $\begin{bmatrix}a_{11}\\\vdots\\a_{n1}\end{bmatrix},\ldots,\begin{bmatrix}a_{1n}\\\vdots\\a_{nn}\end{bmatrix}$ needs to have non-zero

n-dimensional volume. But, as described above, the algebraic formulation of *n*-dimensional volume is that the determinant is not zero.

The linear conversion is given explicitly by

$$x_1 = a_{11}u_1 + \ldots + a_{1n}u_n$$
$$\vdots$$
$$x_n = a_{n1}u_1 + \ldots + a_{nn}u_n$$

Observe that this may also be written using function notation. For then the conversion is given by the function $\mathbf{F} : (u_1, \ldots, u_n) \to (x_1, \ldots, x_n)$ defined by

$$\mathbf{F}(u_1, \ldots, u_n) = \begin{pmatrix} a_{11}u_1 + \ldots + a_{1n}u_n \\ \vdots \\ a_{n1}u_1 + \ldots + a_{nn}u_n \end{pmatrix}.$$

The following are then equivalent:

(i) The equations can be inverted;

(ii) The parallelepiped $(\mathbf{F}(1, \ldots, 0), \ldots, \mathbf{F}(0, \ldots, 1))$ has non-zero n-dimensional volume;

(iii) The quantity $\det\left(\begin{bmatrix}a_{11}\\\vdots\\a_{n1}\end{bmatrix},\ldots,\begin{bmatrix}a_{1n}\\\vdots\\a_{nn}\end{bmatrix}\right) \neq 0$

Now that we have a way to determine whether or not an inverse exists for linear systems that are 2-*D*, 3-*D*, etc., the reader may wonder about the simplest case of all, the 1-*D* case. Did we miss this? Our approach at the beginning of these notes was different for the 1-*D* case. Does the determinant approach of calculating area, volume, etc. also work in the case where we have merely one linear equation of the form $x = au$?

Example 1.7. Suppose that we have a parabolic lens, and that the photo screen is situated so that the incident image I is contracted by a factor of $a = \frac{1}{2}$.

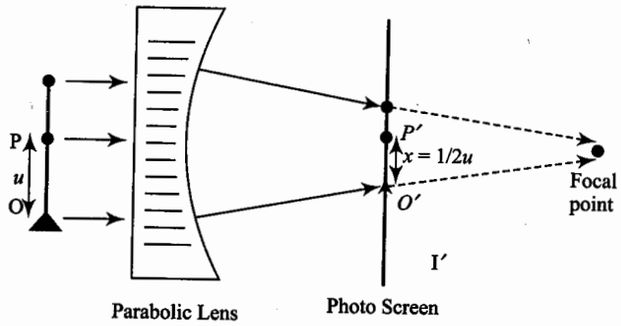


Figure 1.19

Every point P that is a distance u from the base of the image O is converted to a point P' that is a distance $x = \frac{1}{2}u$ from the image of the base O'; and vice versa.

In other words, the conversion $x = \frac{1}{2}u$ can be inverted. Evidently, the formula for the inverse can be easily calculated to be $u = 2x$.

But, what happens if the photo screen is placed at the focus of the parabolic lens?

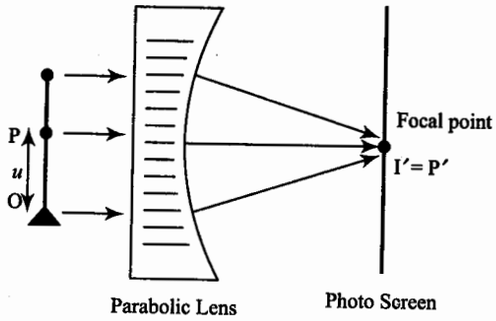


Figure 1.20

In this situation, the resulting image collapses to a single point P' at the focus. The conversion of lengths is given by $x = 0 \cdot u = 0$ for all lengths along the incident image I. Clearly, this formula cannot be inverted, for many lengths along I are converted to the zero length in I'!

While the lens was used as a model, perhaps we can now jump to the conclusion for whenever the conversion is given by $x = au$.

Conclusion: The following are equivalent:

(i) The formula $x = au$ can be inverted;
(ii) The length $u = 1$ is converted to a non-zero length. (Geometric);
(iii) The coefficient $a \neq 0$ (Algebraic; 1-*D* determinant)

Of course, if $x = au$ and $a \neq 0$ then the inverse formula is given by $u = \frac{x}{a}$.

Remark 1.4. The reader may recall that we already had a geometric test for single equations - the horizontal line test. So we now have two tests for invertibility of a formula such as $x = au$. The two tests, though, are rather different. The horizontal line test is graphical; and it goes "outside" the graph of $x = au$, by depending on whether or not horizontal lines intersect the graph in more than one location. The test just developed, however, is in terms of the intrinsic expansion/contraction factor a in the formula $x = au$. This is evidently the special 1-*D* case of our area/volume/determinant solutions for 2-*D*, 3-*D* and *n*-*D* linear systems. For, if we set $u = 1$, we get that the unit length in the u- coordinate is converted to the length $x = a \cdot 1 = a$.

1.5 THE CALCULUS APPROACH

Our work so far has been for linear systems. What about other types of formulas? For example, there is a formula $x = u^2$. How do we calculate the expansion/contraction of this formula?

Notice the following:

$$x(10) - x(9) = 100 - 81 = 19$$
$$x(3) - x(2) = 9 - 4 = 5$$

So, the expansion changes, depending on the initial values.

Or, for a 2-*D* example, using radar technology, the location of an airplane can be given by a distance r from the tower and an angle θ from a reference direction.

The numbers r and θ can be converted into geographic coordinates by

$x = r \cos \theta$ (reference direction)
$y = r \sin \theta$ (perpendicular to reference direction)

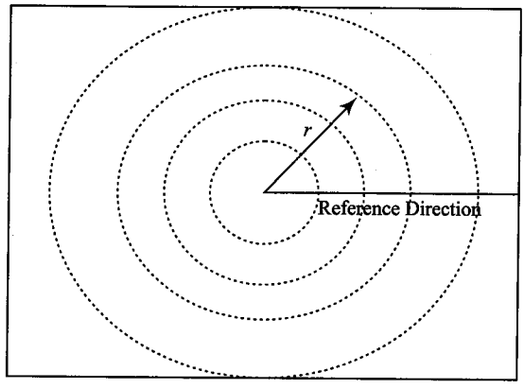


Figure 1.21

For instance, x could be miles east-west, and y could be miles north-south. The present section is devoted to the problem of determining when an inverse exists, for these and other formulas, regardless of whether or not the formulas are linear.

1.5.1 One Equation

Let's return to the formula for the height of a parabolic lens, $y = x^2$.

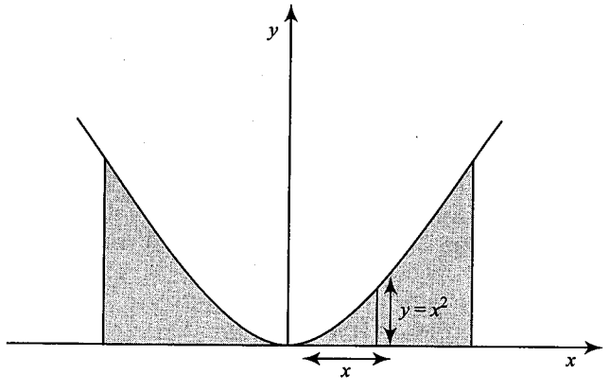


Figure 1.22

We can plot points on a graphical representation. As we discussed earlier, since $(2)^2 = (-2)^2$ and $(x)^2 = (-x)^2$ for all x, the formula in its entirety cannot be inverted.

We can invert the formula, however, if we take one side of the mirror at a time. If, for instance, we look to the RHS, then as x increases, y increases, and no heights are reached twice. So the RHS $y = x^2, x > 0$ can be inverted. That is, given y, we can solve for x by $x = \sqrt{y}$.

Our understanding of this mirror example so far depends very much on insight into the diagram for the graph. Is there, instead, a way to calculate, to precisely obtain that the RHS of the parabola formula really does continue to increase, and really does not turn back somewhere to reach some height more than once?

This is where calculus can come into play. Recall that near any $x_0 > 0$, the slope of the graph of $y = x^2$, $x > 0$ is approximated by the slope of the tangent line of slope $2x_0$. Since $x_0 > 0$, the slope at x_0 is strictly positive. It follows that for $x_0 > 0$, the graph is strictly increasing everywhere. In particular, it does not decrease anywhere.

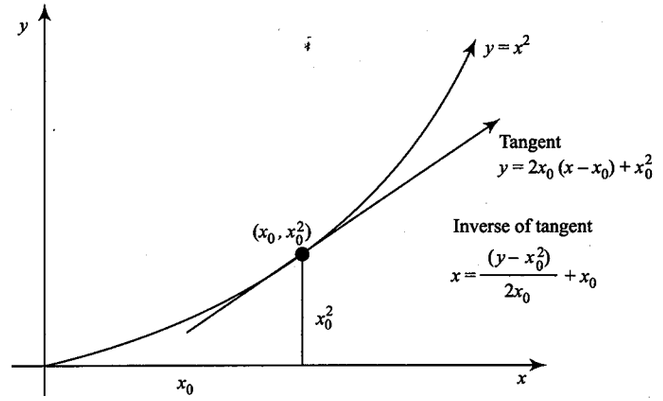


Figure 1.23

Tangent Line: $y = 2x_0(x - x_0) + x_o^2$

Inverse of tangent line: $x = \dfrac{(y - y_0)}{2x_0} + x_0^2$

Basic Theorems from calculus let us conclude that since the tangent line is invertible, so is the graph of $y = x^2$, $x > 0$, at least near $x = x_0$.

Example 1.8. Is the formula $y = x^5 - x^3 + 25$ invertible near $x = 1$? From calculus, the slope of the target line at $x = 1$ is $m = 5(1)^4 - 3(1)^2 = 2 > 0$. Therefore, at least near $x = 1$, there is an inverse. Note, however, that while we now know that an inverse exists, finding an inverse formula is another matter. Frequently, an inverse formula cannot be obtained as an explicit formula in familiar terms. Exploring when this is possible, or not possible, would take us into areas of abstract algebra and analysis beyond the scope of this introductory book.

1.5.2 Two and More Equations

Recall that radar coordinates (r, θ) (also called polar coordinates) can be converted to geographic coordinates (x, y) by

$x = r \cos \theta$ (reference direction)

$y = r \sin \theta$ (perpendicular to reference direction).

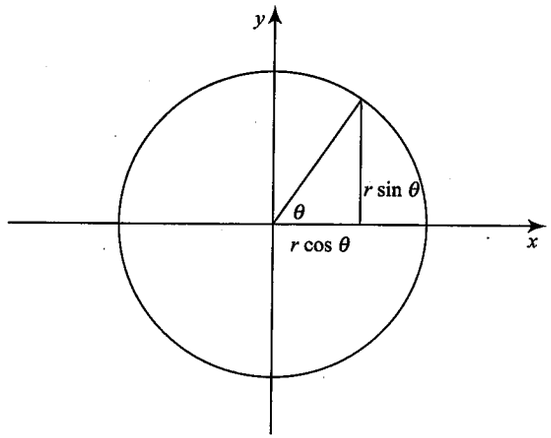


Figure 1.24

For the purposes of this example, suppose that the plane is close enough to the airport so that the curvature of the earth can be ignored. Now, recall also that in the earlier example of the two temperature scales Celsius and Fahrenheit, each referred to the same mercury tube. This is what gave a concrete-like meaning to converting temperature scales. For radar coordinates, we can think in a similar way. For, the pairs of numbers (r, θ) and (x, y) both refer to one location on the surface of the earth; and so this is what gives concrete-like meaning in this example.

Suppose a pilot uses a GPS device and reports the plane's location relative to the surface of the earth, and gives the data to the airport in geographic coordinates x (east-west) and y (north-south). The airport then finds it useful to know how far a plane is from the airport, and in what direction. How can the airport (computer) convert the geographic coordinates x and y into radar coordinates r and θ? We already have $x = r\cos\theta$ and $x = r\sin\theta$, the formulas which convert radar coordinates to geographic coordinates. We seek, therefore, the inverse formulas.

From the Pythagorean formula $r = \sqrt{x^2 + y^2} \geq 0$. We also need the angle θ. But, one way to express direction is slope, and $\frac{y}{x}$ is the slope of the radius pointing toward the location of the plane. To see what this means in terms of the angle, we can substitute to get $\frac{y}{x} = \frac{r\sin\theta}{r\cos\theta} = \tan\theta$. So, $\theta = \arctan\left(\frac{y}{x}\right)$.

What happens though if a plane as approaching from due north? In that case $x = 0$ and the inverse formula $\theta = \arctan\left(\frac{y}{x}\right)$ is not defined.

Or, what if $r = 0$? Then $x = 0$ and $y = 0$, for all angles. In this case, what could it mean to ask for the (r, θ) which converts to $(x, y) = (0, 0)$? That is, since more than one angle θ converts to $(0, 0)$, there is no unique inverse as such for $(x, y) = (0, 0)$.

Again, we return to the main question. Is there a test (algebraic, geometric or otherwise) by which to determine whether or not an inverse exists, regardless of whether or not an inverse formula can be obtained?

Our conversion formulas are

$x = r\cos\theta$ (reference direction)

$y = r\sin\theta$ (perpendicular to reference direction).

Since we had good luck with linear systems by keeping track of how initial coordinate rectangles convert, consider the following diagram:

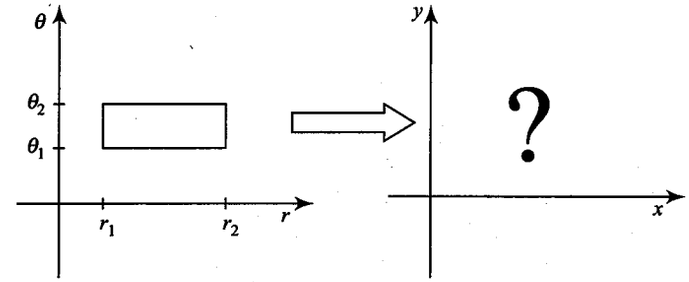


Figure 1.25

The left side of the rectangle consists of coordinates pairs where $r = r_1$ and θ varies from $\theta = \theta_1$ to $\theta = \theta_2$. The bottom edge of the rectangle is where $\theta = \theta_1$ and r varies from $r = r_1$ to $r = r_2$. This leads to the following more complete diagram:

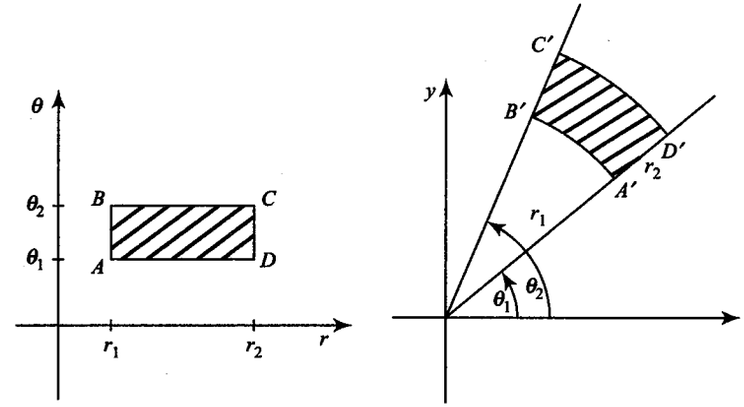


Figure 1.26

Certainly, from elementary geometry, one could argue that there is an inverse. Perhaps you have figured out that this is correct because the rectangle $ABCD$ converts to a wedge $A_2 B_2 C_2 D_2$. Horizontal lines from the (r, θ) spread-sheet of

coordinates (θ constant) convert to radii on the (x, y) map, and vertical lines from the (r, θ) spread-sheet of coordinates (r constant) convert to arcs on the (x, y) map. However, we seek more. We are looking for a test that will work for many formulas, not just radar coordinates. So, we need a reason that will work for other situations as well.

Again calculus can be of assistance. For while the conversion formulas $x = r\cos\theta$ and $x = r\sin\theta$ are certainly not linear, basic theorems of 2-D calculus tell us that near a particular (r_1, θ_1), we may approximate the original formulas by linear formulas. Using Taylor approximations, we get the explicit formulas

$$x = x(r,\theta) = x(r_1,\theta_1) + \frac{\partial x}{\partial r}(r_1,\theta_1)(r-r_1) + \frac{\partial x}{\partial \theta}(r_1,\theta_1)(\theta-\theta_1)$$

$$+ \left[\text{higher order quantities in terms of } (\theta-\theta_1)^2, (\theta-\theta_1)\bullet(r-r_1), (r-r_1)^2, \ldots\right]$$

$$y = y(r,\theta) = y(r_1,\theta_1) + \frac{\partial y}{\partial r}(r_1,\theta_1)(r-r_1) + \frac{\partial y}{\partial \theta}(r_1,\theta_1)(\theta-\theta_1)$$

$$+ \left[\text{higher order quantities in terms of } (\theta-\theta_1)^2, (\theta-\theta_1)\bullet(r-r_1), (r-r_1)^2, \ldots\right]$$

To get the linear approximation to $x(r, \theta)$ and $y(r, \theta)$, valid for r close to r_1 and θ close to θ_1, we use the first order terms from these two equations. So we get that

$$x = x(r,\theta) \approx x(r_1,\theta_1) + \frac{\partial x}{\partial r}(r_1,\theta_1)(r-r_1) + \frac{\partial x}{\partial \theta}(r_1,\theta_1)(\theta-\theta_1)$$

$$y = y(r,\theta) \approx y(r_1,\theta_1) + \frac{\partial y}{\partial r}(r_1,\theta_1)(r-r_1) + \frac{\partial y}{\partial \theta}(r_1,\theta_1)(\theta-\theta_1).$$

In matrix notation this becomes

$$\begin{pmatrix} x(r,\theta) \\ y(r,\theta) \end{pmatrix} \approx \begin{pmatrix} x(r_1,\theta_1) \\ y(r_1,\theta_1) \end{pmatrix} + \begin{pmatrix} \frac{\partial x}{\partial r}(r_1,\theta_1) & \frac{\partial x}{\partial \theta}(r_1,\theta_1) \\ \frac{\partial y}{\partial r}(r_1,\theta_1) & \frac{\partial y}{\partial \theta}(r_1,\theta_1) \end{pmatrix} \begin{pmatrix} r-r_1 \\ \theta-\theta_1 \end{pmatrix}$$

Just as for single straight line equations of the form $y \approx mx + b$, the initial values $x(r_1, \theta_1)$, $y(r_1, \theta_1)$ of the equations do not affect the invertibility of the linear equations. (We leave that as a question for the reader to explore.) To determine whether or not the equations are invertible, we can now use our earlier work and calculate the area of the parallelogram determined by the columns

$$\begin{pmatrix} \frac{\partial x}{\partial r}(r_1,\theta_1) \\ \frac{\partial y}{\partial r}(r_1,\theta_1) \end{pmatrix} \text{ and } \begin{pmatrix} \frac{\partial x}{\partial \theta}(r_1,\theta_1) \\ \frac{\partial y}{\partial \theta}(r_1,\theta_1) \end{pmatrix}.$$

In other words, we need only calculate the determinant of the linear approximation to the equations. The 2×2 matrix constructed from these column vectors is traditionally called the "Jacobian" matrix of the original conversion/equations. That is, the Jacobian matrix is the matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x}{\partial r}(r_1,\theta_1) & \frac{\partial x}{\partial \theta}(r_1,\theta_1) \\ \frac{\partial y}{\partial r}(r_1,\theta_1) & \frac{\partial y}{\partial \theta}(r_1,\theta_1) \end{pmatrix}$$ that gives the linear/first-order approximation to the original conversions/equations/transformation.

Exercise 1.21. For our radar coordinate formulas, do the calculation to obtain that the determinant of the Jacobian satisfies $\det \mathbf{J} = r_1$.

From Exercise (1.21), it follows that as long as $r_1 > 0$, that is, as long as we stay away from the center of the radar screen, the inverse exists near the reference point (r_1, θ_1). Again, note that $\det \mathbf{J} = r_1$ is a special area, for it is the area of a converted unit square from the radar coordinates (r, θ). This unit square is converted by the linear approximation $\mathbf{J}$ acting at the location (r_1, θ_1).

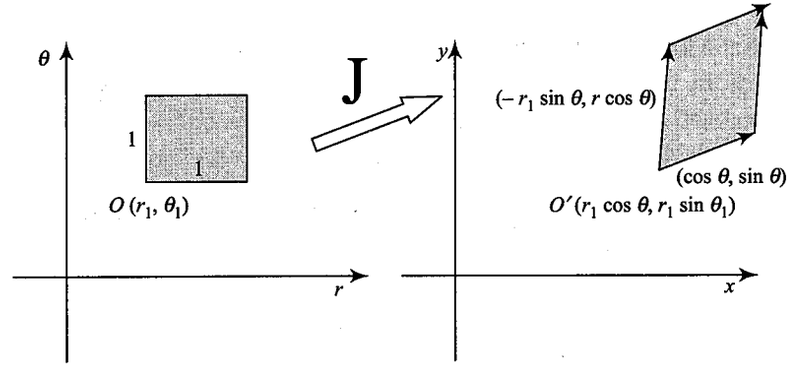


Figure 1.27

We can now summarize our results by what in multi-variable calculus is called the **Inverse Function Theorem**:

Theorem 1.1. Suppose that converts $\mathbf{F} : (u_1, \ldots, u_n) \to (x_1, \ldots, x_n)$, and that all partial derivatives exist so that $\mathbf{J}$ is the defined Jacobian. If at the point $\left(u_1^0, \ldots, u_n^0\right)$ we have $\det \mathbf{J} \neq 0$, then for a region containing the initial reference point $\left(u_1^0, \ldots, u_n^0\right)$, the conversion $\mathbf{F} : (u_1, \ldots, u_n) \to (x_1, \ldots, x_n)$ has an inverse $\mathbf{F}^{-1} : (x_1, \ldots, x_n) \to (u_1, \ldots, u_n)$ defined on a region containing the image of the reference point $F\left(u_1^0, \ldots, u_n^0\right)$.

Moreover, the derivative of the inverse is the inverse of the derivative. See Exercise 1.22.

The Idea (*Not a Proof*)

If the linear approximation is invertible, then at least near the point in question, the original formula is invertible. One way for the linear approximation to be invertible, is if under conversion to $(x_1, \ldots, x_n)$ coordinates, the unit volume in the $(u_1, \ldots, u_n)$ coordinates does not collapse to zero (geometry). But it is the determinant (algebra) of the linear approximation that gives the converted unit volume. The result follows. We note that the full statement of the Inverse Function Theorem includes a formula for the Jacobian of the inverse function. Please consult a standard multi-variable reference text for a more detailed formulation.

Exercise 1.22. Show that the Jacobian of the inverse $\mathbf{F}^{-1} : (x_1, \ldots, x_n) \to (u_1, \ldots, u_n)$ is given by $\mathbf{J}^{-1}$. *Clue:* If the inverse exists, locally we can write $[\mathbf{F}^{-1} \circ \mathbf{F}](x_1, \ldots, x_n) = (x_1, \ldots, x_n)$.

Exercise 1.23. Consider the conversion given by $\begin{matrix} x = u^2 - v^2 \\ y = u^2 + v^2 \end{matrix}$. Or, in the notation of the inverse function theorem, $\mathbf{F}(u, v) = \begin{pmatrix} u^2 - v^2 \\ u^2 + v^2 \end{pmatrix}$. Near what points (u_1, v_1) is the conversion $\mathbf{F}$ invertible? A unit rectangle in the (u, v) coordinates gets converted to what shape? What is the inverse of a vertical line in (x, y) coordinates? Draw in several vertical lines and their inverses. What is the inverse of a horizontal line in (x, y) coordinates? Draw in several horizontal lines and their inverses. New "curvi-linear" coordinates for the plane are obtained. *Clue*: See Fig. 1.28.

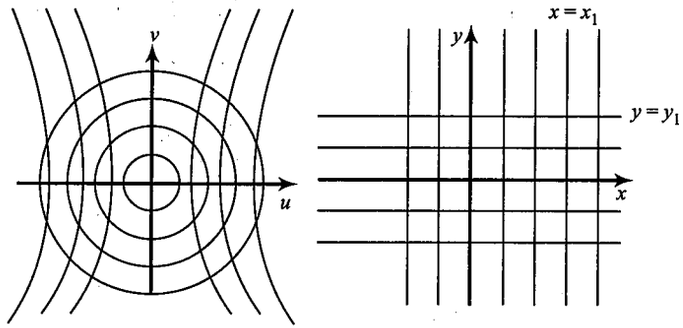


Figure 1.28

Exercise 1.24. In each of the next three problems, make a diagram; determine whether or not the conversion is invertible; use algebra to verify your solution; and determine what the Inverse Function Theorem has to say about each example. Be careful of the logic of the Inverse Function Theorem. That is, the theorem provides

sufficient conditions. Note in Figure (1.28) that along the u axis, coordinate circles are tangent to coordinate hyperbolas. Compare this to the example discussed in Section 1.4.3., the case where the two rivers are parallel.

1. Consider the function that projects (u, v) onto the line that is the u axis. Make a diagram for this situation. The function is defined by $\mathbf{F}(u, v) = (u, 0)$.
2. Consider the function defined by $\mathbf{F}(u, v) = (u^2 + 5v^2, 0)$. What about any function of the form $\mathbf{F}(u, v) = (f(u, v), 0)$.
3. Consider the function that projects (u, v) onto the line $u = v$. Construct a formula $\mathbf{F}(u, v) = (x(u, v), y(u, v))$ for this function. Your formula should only involve u's and v's. *Hint*: Recall the that the dot product of two vectors satisfies $\boldsymbol{v} \bullet \boldsymbol{w} = \|\boldsymbol{v}\| \|\boldsymbol{w}\| \cos\theta$. Another *hint*: Matrix.

Next we look to another well known and closely related result called the **Implicit Function Theorem**.

In multi-variable calculus texts, one will usually find the Inverse Function Theorem and the Implicit Function Theorem in the same chapter, or sometimes in the same section. The Inverse Function Theorem provides sufficient conditions for the existence of an inverse function in one or more variables. The Implicit Function Theorem provides sufficient conditions for when one set of variables depends as a function (perhaps only *implicitly*) on another set of variables. Although these two theorems may at first appear to be rather different, they are in fact two sides of the same coin.

In Section 3 of this chapter, we looked at the equation of a circle of unit radius, $x^2 + y^2 = 1$. Using the graph of the circle, y can be interpreted as the height of the circle off the x axis; and one can see from the graph that as long as we keep to certain quadrants, the height y is a function of horizontal distance x. In fact, for the equation $x^2 + y^2 = 1$, we can use elementary algebra to solve for y explicitly in terms of x. That is, one function is the upper semi-circle given by $y = +\sqrt{1 - x^2}$; and the other function is the lower semi-circle given by $y = -\sqrt{1 - x^2}$. Of course, if we require that a solution function go through the point $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$, then we must choose the function $y = +\sqrt{1 - x^2}$. In the same way, if we require that the solution of choice go through the point $\left(\frac{1}{2}, \frac{-\sqrt{3}}{2}\right)$, then we must choose the function $y = -\sqrt{1 - x^2}$.

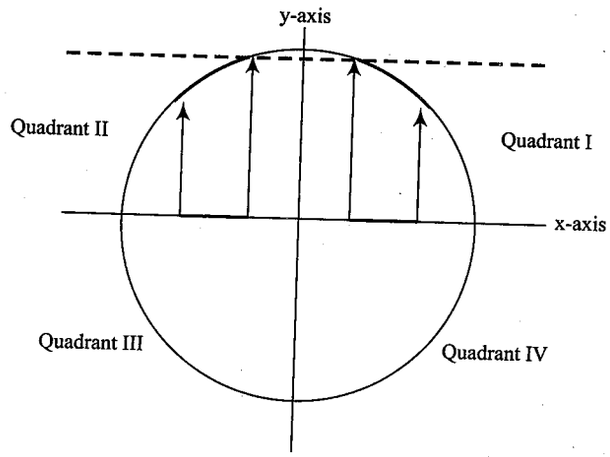


Figure 1.29

In many applications, an explicit formula $y = y(x)$ may not be necessary or even possible. It can be important, though, to at least be able to determine whether or not such a function exists. In this situation, the solution function is said to be *implicit*. Note that the name *function* is meant in the technical sense of there being a single valued relation $y = y(x)$; and again, note that an affirmative solution does not require either a unique or an explicit formula, but merely the *existence* of at least one function.

Example 1.9. Consider the relation $x^2 + \sin y + y^2 = 1$. If $x = 0$, then in order to find a value for y that goes with it to solve the equation, we need to solve $(0)^2 + \sin y + y^2 = 1$. In other words, we seek a y corresponding to $x = 0$ such that $\sin y + y^2 = 1$. If y is near zero, then $\sin y + y^2 \approx 0$ is near zero. Since the sine function is bounded, if y is very large and positive then $\sin y + y^2 >> 0$ is also large and positive. Note too that $\sin y + y^2$ is a continuous function of y. So, by the intermediate value theorem, there is at least one value for y that solves the equation $\sin y + y^2 = 1$. In other words, given $x = 0$, we obtain the existence of at least one solution y_0 to the equation $(0)^2 + \sin y_0 + y_0^2 = \sin y_0 + y_0^2 = 1$.

Let's pursue this a little more, by trying to analyze the general equation and, while doing so, remember that we have the results and techniques of calculus at our disposal. If we suppose that x is given, then we need to solve $x^2 + \sin y + y^2 = 1$. That is, we need to find y such that $\sin y + y^2 = 1 - x^2$. Note that the left hand side of this equation $\sin y + y^2$ satisfies $\sin y + y^2 \geq -1 + y^2 \geq -1$. Hence, if $|x| > \sqrt{2}$, there can be no solution, for in that case we would have $-1 > 1 - x^2$. For example, suppose that $x = 3$, then the left hand side is $\sin y + y^2 \geq -1$ but the right hand side becomes $1 - 3^2 = -8$. So, in order for there to be a solution, we will need to put

some constraints on x. This though need not be too surprising. The reader may recall that solutions of $x^2 + y^2 = 1$ also require constraints on both x and y. In the case of $x^2 + \sin y + y^2 = 1$, to help us avoid too many subtleties, let's be conservative and suppose a strict inequality $|x| < \sqrt{2}$.

The question now becomes, for x satisfying $|x| < \sqrt{2}$, does the relation $\sin y + y^2 = 1 - x^2$ have a solution $y = y(x)$ for each x that in fact produces a *function* of x? The quantity in question is $h(y) = \sin y + y^2$. We need to determine whether or not we can find y so that $\sin y + y^2 = 1 - x^2$. Now, $h(0) = \sin(0) + (0)^2 = 0$, and also $h'(y) = \cos y + 2y$. Assuming that y represents radians, we obtain $h'(y) = \cos y + 2y > 0$ for all $y \geq 0$. The function $h(y) = \sin y + y^2$ is therefore a strictly increasing function. And the key now is the familiar Inverse Function Theorem for one real variable. For, in one variable, if $h'(y) \neq 0$, then locally the function $h(y) = \sin y + y^2$ is invertible. It follows that as long as $|x| < \sqrt{2}$, the relation $\sin y + y^2 = (1 - x^2)$ does indeed imply the existence of a (differentiable) solution $y = y(x) \geq 0$.

For a moment, consider a somewhat more elaborate example such as $x^2 + \sin y + y^2 + \sin z + z^2 = 1$. While the approach taken above starts to reveal some of the usefulness of calculus, the equation $x^2 + \sin y + y^2 + \sin z + z^2 = 1$ involving three variables suggests that the somewhat ad hoc approach above may not be sufficient to deal with more complex relations. So, let's look again at the problem we have already solved, $x^2 + \sin y + y^2 = 1$. But now, let's see if we can tease out some clues that might be useful for a more general approach.

If we assume that $x^2 + \sin y + y^2$ determines $y = y(x) \geq 0$ as a function of x, then we may write $x^2 + \sin [y(x)] + [y(x)]^2 = 1$. We may not be able to solve explicitly for $y = y(x)$, but what we can do easily is obtain information on the rate of change of $y = y(x)$. For, using the Chain Rule and differentiating with respect to x we get $2x + \frac{dy}{dx}\{\cos[y] + 2[y]\} = 0$, which implies that $\frac{dy}{dx} = \frac{-2x}{\cos y + 2y}$. In other words, if there is a solution $y = y(x)$ that goes through the point (x, y), the slope is given by $\frac{dy}{dx} = \frac{-2x}{\cos y + 2y}$. (See the second graph of Figure 30.) For instance, even if we don't have a formula for the function $y = y(x)$ itself, if it exists, its slope at the solution $(x, y) = (1, 0)$ would have to be

$$\frac{dy}{dx} = \frac{-2(1)}{\cos(0) + 2(0)} = \frac{-2}{1} = -2.$$

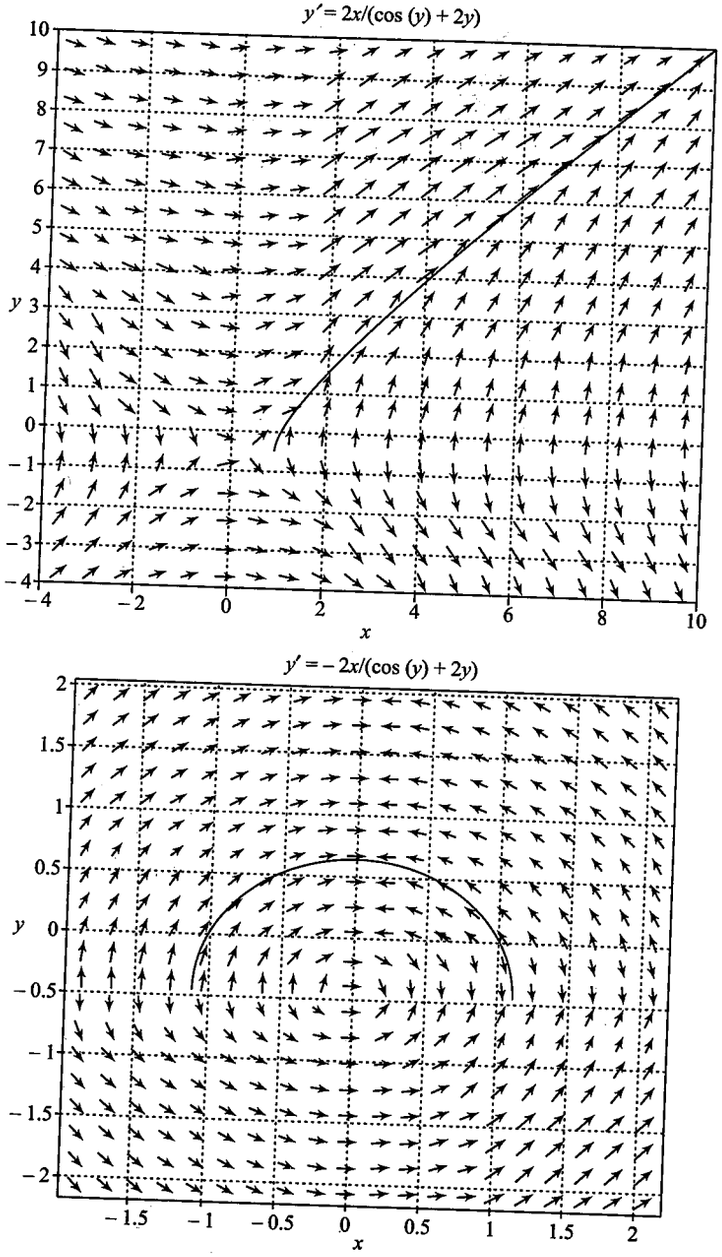


Figure 1.30

Can this argument be reversed? If there is a starting point solution (x^0, y^0), and if we can solve explicitly for $\frac{dy}{dx}$ as a well-defined function of (x, y), can we infer

the existence of an implicitly defined function $y = y(x)$ that works not only for the starting point (x^0, y^0), but so that following the slope function $\frac{dy}{dx}$, we produce nearby solutions $(x, y(x))$ as well? To make this idea precise we would need to be able to "integrate the slope field". That is, we would need to be able to find a function whose derivative (slope) coincides with the slope of the field given.

In geometric terms, this seems plausible. For, if a slope field in the (x, y) plane is given, together with some initial point, it is evident from diagrams that, at least under suitable hypotheses, this will determine a function $y = y(x)$ that fits that slope field.

Example 1.10. For now, let's skip past the subtlety just mentioned regarding possible integration of slope fields. Instead, in order to get some more clues on the *implicit* approach, let's return to our old example of the circle $x^2 + y^2 = 1$. If there is an implicit solution $y = y(x)$ solving $x^2 + [y(x)]^2 = 1$, then using the Chain Rule, $\frac{dy}{dx} = \frac{-2x}{2y} = \frac{-x}{y}$. Again, this determines an explicitly defined slope field in the (x, y) plane. Starting at (0, 1), the slopes can be seen to trace out a counter-clockwise flow along the unit circle, where the slope at the initial point (0, 1) is given by $\frac{dy}{dx} = \frac{-0}{1} = 0$. Of course, that makes sense, since the point (0, 1) is at the top of the circle. Again, finishing the solution becomes a question of integrating the slope field.

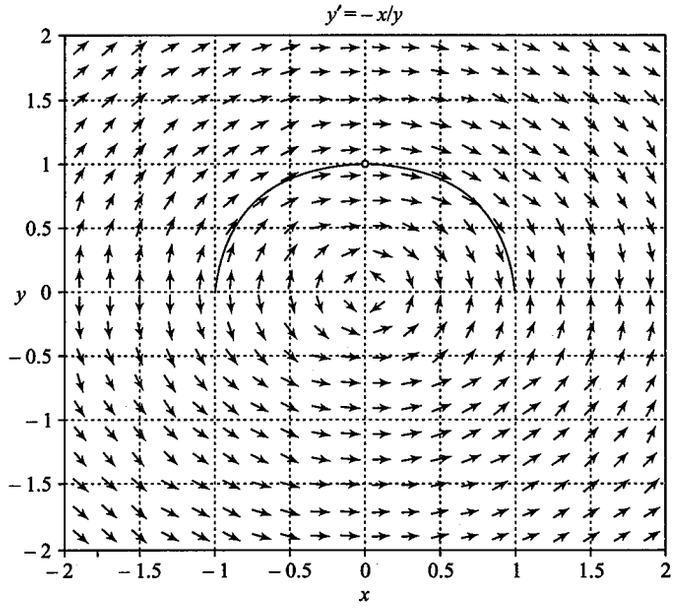


Figure 1.31

Example 1.11. The two examples we have just looked at are of the form $f(x, y) = K$, where K is constant. Following the implicit approach, if we suppose the existence of an implicitly determined function $y = y(x)$ that solves $f(x, y) = K$, then we may write $f(x, y(x)) = K$. As above, we can then differentiate with respect to x and obtain $\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{dy}{dx} = 0$. This gives $\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}$. As long as the denominator $\frac{\partial f}{\partial y} \neq 0$, we can solve for $\frac{dy}{dx}$ and so obtain a slope field in the (x, y) plane. If this can be integrated, then we obtain the existence of the implicitly determined solution function $y = y(x)$.

Example 1.12. In the notation of Example 1.11, let's suppose that $f(x, y) = g(y) = \sin y$, x arbitrary. Then $\frac{\partial f}{\partial y} \neq 0$ while $\frac{\partial f}{\partial x} = 0$. It follows that if $y(x)$ is a solution of $f(x, y) = \sin y = 1$, then $\frac{dy}{dx} \equiv -\frac{\partial f/\partial x}{\partial f/\partial y} = 0$. Of course, this implies that $y(x) =$ constant $= \frac{\pi}{2} + m2\pi$ for any integer m. In other words, not only can there be more that one solution function, but there can be infinitely many solution functions $y(x)$ solving an equation $f(x, y) = f(x, y(x)) = K$.

Example 1.13. Consider the equation $f(x, y) = x^2 + y^2 = -1$. If we use the symbolism of the calculation above, we get $\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y} = -\frac{x}{y}$. There is of course a problem, for there is no real solution function $y(x)$ to $x^2 + y(x)^2 = -1$. Remember, however, that the equation $\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}$ was derived under the hypothesis that there exists a solution. But, there may be not even one solution point for the relation, let alone be points generated by a solution function $y = y(x)$.

Example 1.14. Let's start building up the types of example we can work with. Instead of one equation in two variables, we can look at the problem of one equation in three variables. For example, $x + 2y + 3z = 5$. Of course, in this case, it is straightforward to see that the third variable z is determined as a function of (x, y). Indeed, we obtain an explicit solution $z = \frac{1}{3}\left[5 - (x + 2y)\right]$.

Example 1.15. Just as with two variables, however, it is not difficult to produce examples in three variables for which one cannot obtain an explicit solution $z = z(x, y)$. For instance, consider the equation $x^2 + \sin y + y^2 + \sin z + z^2 = 1$. Instead of

entering into ad hoc calculations for this special case, let's take a cue from our progress so far with the case of two variables. In other words, consider the general case of one equation in three variables, an equation of the form $f(x, y, z) = K$, where K is constant. If $z = z(x, y)$ exists, what can we infer? The equation can be written $f(x, y, z(x, y)) = K$. The left hand side formula depends on two variables, and so we can consider two partial rates of change, in other words, two partial derivatives. Because of the Chain Rule, this produces the following two equations:

$$\frac{\partial f}{\partial x} \cdot 1 + \frac{\partial f}{\partial y} \cdot 0 + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial x} = 0$$

$$\frac{\partial f}{\partial x} \cdot 0 + \frac{\partial f}{\partial y} \cdot 1 + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial y} = 0$$

Much as in one variable calculus, let's see if we can isolate the (partial) derivatives of the sought after unknown function $z = z(x, y)$. Rewriting, we obtain

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial x} = 0$$

$$\frac{\partial f}{\partial y} + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial y} = 0$$

Of course, as long as $\frac{\partial f}{\partial z} \neq 0$, we can solve this algebraically. This means that at each point (x, y, z) that solves $f(x, y, z) = K$, if $z = z(x, y)$ exists, then its two partial derivatives must be given by the equation $\begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix} = -\frac{1}{\left(\frac{\partial f}{\partial z}\right)} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$. And, if $f(x, y, z)$ is differentiable, then the right hand side of this last equation is a well defined pair of explicitly defined slope fields defined at points (x, y, z). So, once again, reversing the logic will depend on knowing whether or not we can integrate slope fields to produce an implicitly defined function.

Example 1.16. Will the approach work for a larger number of variables, for example, five? Let's construct an example to check this out. Consider the equation $x^2 + y^3 + z^4 + u^5 + v^6 + \sin v = 1$. The problem is to see if this implies the existence of an implicitly defined $v = v(x, y, z, u)$ satisfying (at least locally) the equation $x^2 + y^3 + z^4 + u^5 + (v(x, y, z, u))^6 + \sin(v(x, y, z, u)) = 1$. Again, we may use the Chain Rule to differentiate and so obtain information about the relative rates of change. This calculation gives a system of four equations

$$2x + 6v^5 \frac{\partial v}{\partial x} + \cos(v)\frac{\partial v}{\partial x} = 0$$

$$3y^2 + 6v^5 \frac{\partial v}{\partial y} + \cos(v)\frac{\partial v}{\partial y} = 0$$

$$4y^3 + 6v^5 \frac{\partial v}{\partial y} + \cos(v)\frac{\partial v}{\partial z} = 0$$

$$5y^4 + 6v^5 \frac{\partial v}{\partial u} + \cos(v)\frac{\partial v}{\partial u} = 0$$

We are trying to isolate the partial derivatives of the unknown function $v = (x, y, z, u)$. Basic algebra leads to

$$\left[6v^5 + \cos(v)\right]\frac{\partial v}{\partial x} = -\,[2x]$$

$$\left[6v^5 + \cos(v)\right]\frac{\partial v}{\partial y} = -\,[3y^2]$$

$$\left[6v^5 + \cos(v)\right]\frac{\partial v}{\partial z} = -\,[4y^3]$$

$$\left[6v^5 + \cos(v)\right]\frac{\partial v}{\partial u} = -\,[5y^4]$$

So, if a solution $v = (x, y, z, u)$ exists, and if $[6v^5 + \cos(v)] \neq 0$, then we obtain explicit formulas that the partial derivatives must satisfy, namely,

$$\frac{\partial v}{\partial x} = -\,[6v^5 + \cos(v)]^{-1}\,[2x]$$

$$\frac{\partial v}{\partial y} = -\,[6v^5 + \cos(v)]^{-1}\,[3y^2]$$

$$\frac{\partial v}{\partial z} = -\,[6v^5 + \cos(v)]^{-1}\,[4y^3]$$

$$\frac{\partial v}{\partial u} = -\,[6v^5 + \cos(v)]^{-1}\,[5y^4]$$

Example 1.17. It would seem that the approach will work for one equation in any number of variables. What about increasing the number of equations? Does our approach work for two equations? For three equations? For more? Let's open things up a little, and look at the case of two equations and four unknowns. As above, simple illustrations can be obtained by looking at linear equations. Take, for example, the two linear equations given by

$$x + 2y + 3u + 0 \cdot v = 4$$
$$5x + 6y + 0 \cdot u + 7v = 8.$$

Evidently, this implies that

$$3u + 0 \cdot v = 4 - (x + 2y)$$

$$0 \cdot u + 7v = 8 - (5x + 6y).$$

Elementary algebra gives the solution

$$u = \frac{1}{3}\left[4 - (x + 2y)\right]$$

$$v = \frac{1}{7}\left[8 - (5x + 6y)\right].$$

This indeed provides an explicit solution for $u = u(x, y)$ and $v = v(x, y)$. But, just as arithmetic can hide underlying algebraic pattern, in the present case it can be helpful to try to keep track of the operations involved in the calculations. The equation

$$4u + 0 \cdot v = 4 - (x + 2y)$$
$$0 \cdot u + 7v = 8 - (5x + 6y)$$

is a pair of linear equations and so one way to keep track of the operations is to use matrix notation of linear transformations. We can rewrite this last pair of equations as

$$\begin{bmatrix} 3 & 0 \\ 0 & 7 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 4 - (x + 2y) \\ 8 - (5x + 6y) \end{bmatrix}.$$

Clearly, the diagonal matrix is invertible, and we obtain

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 7 \end{bmatrix}^{-1}\begin{bmatrix} 4 - (x + 2y) \\ 8 - (5x + 6y) \end{bmatrix}$$

Example 1.18. The matrix equation of Example (1.17) is of the form

$$A\begin{bmatrix} x \\ y \end{bmatrix} + B\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} h \\ k \end{bmatrix}.$$

If B^{-1}, exists, show more generally that $\begin{bmatrix} u \\ v \end{bmatrix} = B^{-1}\left\{\begin{bmatrix} h \\ k \end{bmatrix} - A\begin{bmatrix} x \\ y \end{bmatrix}\right\}$.

Exercise 1.25. Use matrix notation to solve

$$\frac{\partial f}{\partial x} \cdot 1 + \frac{\partial f}{\partial y} \cdot 0 + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial x} = 0$$
$$\frac{\partial f}{\partial x} \cdot 0 + \frac{\partial f}{\partial y} \cdot 1 + \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial y} = 0$$

explicitly for $\begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix}$.

What is the criterion needed for the existence of the inverse matrix?

Solution: If $\det\begin{bmatrix} \frac{\partial f}{\partial z} & 0 \\ 0 & \frac{\partial f}{\partial z} \end{bmatrix} \neq 0$ (or equivalently $\frac{\partial f}{\partial z} \neq 0$) we obtain $\begin{bmatrix} \frac{\partial z}{\partial x} \\ \frac{\partial z}{\partial y} \end{bmatrix} =$

$$-\begin{bmatrix} \frac{\partial f}{\partial z} & 0 \\ 0 & \frac{\partial f}{\partial z} \end{bmatrix}^{-1}\begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}.$$

Example 1.19. Let's now look at the general case of two equations in four variables. This case has enough complexity to begin to reveal how the general result will go for several equations and several unknowns. Suppose therefore that we have two equations of the form

$$f(x, y, u, v) = K$$
$$g(x, y, u, v) = L$$

where K and L are constants.

Now, for a moment, recall what happens when these are linear equations. For instance, see Example (1.16). Gaussian elimination reveals that it is typical that two linear equations in four variables can determine two of the variables as functions of the other two variables. The situation in this Example (1.19) is similar, but the equations are not necessarily linear. As in Gaussian elimination, the objective is to determine two variables u and v as functions of the other two variables x and y. That is, we seek $u = u(x, y)$ and $v = v(x, y)$. If these functions exists, we may write

$$f(x, y, u(x, y), v(x, y)) = K$$
$$g(x, y, u(x, y), v(x, y)) = K.$$

Each equation has two partial derivatives, with respect to x and y respectively. If the examples so far are anything to go on, there will be four unknowns, the four partial derivatives $\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}$ corresponding to the unknown functions $u = u(x, y)$ and $v = v(x, y)$. Indeed, using the Chain Rule, we obtain

$$\frac{\partial f}{\partial x}\cdot 1 + \frac{\partial f}{\partial y}\cdot 0 + \frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\cdot\frac{\partial v}{\partial x} = 0$$

$$\frac{\partial f}{\partial y}\cdot 0 + \frac{\partial f}{\partial y}\cdot 1 + \frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\cdot\frac{\partial v}{\partial y} = 0$$

$$\frac{\partial g}{\partial x}\cdot 1 + \frac{\partial g}{\partial y}\cdot 0 + \frac{\partial g}{\partial u}\cdot\frac{\partial u}{\partial x} + \frac{\partial g}{\partial v}\cdot\frac{\partial v}{\partial x} = 0$$

$$\frac{\partial g}{\partial y}\cdot 0 + \frac{\partial g}{\partial y}\cdot 1 + \frac{\partial g}{\partial u}\cdot\frac{\partial u}{\partial y} + \frac{\partial g}{\partial v}\cdot\frac{\partial v}{\partial y} = 0$$

Organizing this somewhat we obtain

$$\frac{\partial f}{\partial x}\cdot 1 + \frac{\partial f}{\partial y}\cdot 0 + \frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\cdot\frac{\partial v}{\partial x} + 0\cdot\frac{\partial u}{\partial y} + 0\cdot\frac{\partial v}{\partial y} = 0$$

$$\frac{\partial f}{\partial y}\cdot 0 + \frac{\partial f}{\partial y}\cdot 1 + 0\cdot\frac{\partial u}{\partial x} + 0\cdot\frac{\partial v}{\partial x} + \frac{\partial f}{\partial u}\cdot\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\cdot\frac{\partial v}{\partial y} = 0$$

$$\frac{\partial g}{\partial x}\cdot 1 + \frac{\partial g}{\partial y}\cdot 0 + \frac{\partial g}{\partial u}\cdot\frac{\partial u}{\partial x} + \frac{\partial g}{\partial v}\cdot\frac{\partial v}{\partial x} + 0\cdot\frac{\partial u}{\partial y} + 0\cdot\frac{\partial v}{\partial y} = 0$$

$$\frac{\partial g}{\partial y}\cdot 0 + \frac{\partial g}{\partial y}\cdot 1 + 0\cdot\frac{\partial u}{\partial x} + 0\cdot\frac{\partial v}{\partial x} + \frac{\partial g}{\partial u}\cdot\frac{\partial u}{\partial y} + \frac{\partial g}{\partial v}\cdot\frac{\partial v}{\partial y} = 0$$

Our objective is to isolate the unknown partial derivatives $\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}$.

Writing this as a linear matrix equation, we obtain

$$\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}\begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial y} \end{bmatrix} = -\begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}$$

If the inverse of the matrix that multiplies on the left exists, then we can directly solve for the unknown partial derivatives. What though is our criterion for a matrix to have an inverse? From Section (1.4.5), it is the determinant that gives us the needed criterion. In the present case, that determinant is

$$\det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}.$$

Remember that the general determinant function is alternating and multilinear. You might also check a reference to find out why the determinant of a matrix is the same as the determinant of the transpose of the matrix. (This is a straightforward calculation for a 2×2 or a 3×3 matrix. The general result takes some care, and also has reaching implications.) Now, if we interchange the two middle rows, we get

$$-\det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix} = \det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix} \det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}$$

Note the sign change because of the determinant function being alternating.

Consequently, if $\det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix} \neq 0$, we can solve the system and obtain the explicit solution, not for the unknown functions $u = u(x, y)$ and $v = v(x, y)$, but for the partial derivatives $\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial y}$ of the unknown functions. That is

$$\begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial y} \end{bmatrix} = -\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}.$$

Note that at this stage, we have not yet solved the implicit function problem for two equations in four variables. We have shown that if solutions $u = u(x, y)$ and $v = v(x, y)$ exist for $\begin{matrix} f(x, y, u(x, y), v(x, y)) = K \\ g(x, y, u(x, y), v(x, y)) = L \end{matrix}$, and $\det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix} \neq 0$, then the partial derivatives must be given by

$$\begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial y} \end{bmatrix} = -\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}.$$

But, what we really need to do is be able to argue the other way. In other words, if $\det\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix} \neq 0$ and $\begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial y} \end{bmatrix} = -\begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & 0 & 0 \\ 0 & 0 & \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} \\ \frac{\partial g}{\partial y} \end{bmatrix}$, and if we have a starting point (x^0, y^0, u^0, v^0) that solves $\begin{matrix} f(x^0, y^0, u^0, v^0) = K \\ g(x^0, y^0, u^0, v^0) = L \end{matrix}$, can we somehow use the formulas for the partial derivatives to infer the existence of solution functions $u = u(x, y)$ and $v = v(x, y)$ satisfying $u^0 = u(x^0, y^0)$ and $v^0 = v(x^0, y^0)$ as well as the system of equations $\begin{matrix} f(x, y, u(x, y), v(x, y)) = K \\ g(x, y, u(x, y), v(x, y)) = L \end{matrix}$ for all (x, y) close to (x^0, y^0)?

General Case of the Implicit Function Theorem

As we have seen in each of our examples so far, one approach toward showing the possible existence of implicitly defined functions reduces, at a critical stage of the calculation, to the question of whether or not we can invert a multi-variable linear equation, and then integrate that system of explicitly given partial derivatives. So it may come as no great surprise that to complete the investigation would require a fuller discussion of the Inverse Function Theorem. That would of course go beyond the purpose of these introductory notes. However, stopping short of this more complete discussion, it may nevertheless be useful to go at least a little way toward the general Implicit Function Theorem.

In the simplest case of linear equations, you may again recall using Gaussian elimination. For example, consider the already row-reduced system

$$3x + 4y + 5z + u + 0v = 10$$
$$2x + 6y + 8z + 0u + v = 20$$

This clearly provides functions $u = u(x, y, z)$ and $v = v(c, y, z)$ given by

$$u = 10 - (3x + 4y + 5z)$$
$$v = 20 - (2x + 6y + 8z).$$

We can generalize the linear equations as follows: Let $\mathbf{A}$ be a 2×3 matrix, $\mathbf{\Lambda}$ be a 2×1 matrix of constants, and consider the form $\begin{bmatrix} u \\ v \end{bmatrix} = \mathbf{\Lambda} - \mathbf{A}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$. Evidently, this may be written in the form $-\mathbf{A}\begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} u \\ v \end{bmatrix} - \mathbf{\Lambda} = \mathbf{0}$. Of course, for the more general problem, the system of defining equations need not be linear.

What though about the number of equations? In each of our examples we find that, much as for linear systems, for one unknown function u we need one equation; for two unknown functions u, v we need two equations; and so on. In other words, as a general requirement, in order to be able to solve for m unknowns u_1, u_2, …, u_m we will need m equations.

A linear system would of course be a special case. Indeed, if you are familiar with the general results of Gaussian elimination, when the system is linear, the Implicit Function Theorem (stated below) reproduces basic existence results for Gaussian elimination.

For the set up of the general Implicit Function Theorem, we first suppose m variables $u_1, u_2, \ldots, u_m$, n variables $x_1, x_2, \ldots, x_n$, and a number of equations that equals the number m of unknown implicitly defined functions $u_1, u_2, \ldots, u_m$. In other words, we suppose a system of equations of the form

$$f_1(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0$$
$$f_2(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0$$
$$\vdots$$
$$f_m(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0.$$

We now have more variables in play, but the problem is directly analogous to the special case problem stated for the equation of the circle. [See also Figure (1.29)]. In the general case though, instead of seeking one implicitly defined function $y = y(x)$ satisfying the single defining equation $x^2 + y^2 - 1 = 0$, we now seek to determine the existence of m implicitly defined functions $u_1(x_1, x_2, \ldots, x_n)$, $u_2(x_1, x_2, \ldots, x_n)$, …, $u_m(x_1, x_2, \ldots, x_n)$ simultaneously satisfying all of the defining equations

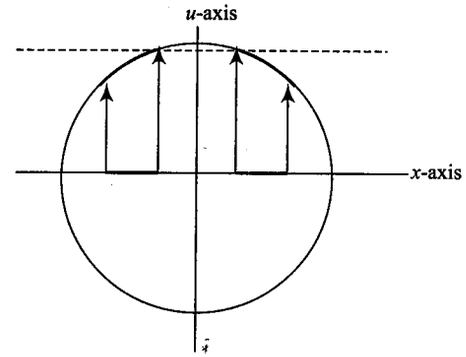


Figure 1.31

$$f_1\left(x_1, x_2, \ldots, x_n, u_1\left(x_1, x_2, \ldots, x_n\right), u_2\left(x_1, x_2, \ldots, x_n\right), \ldots, u_m\left(x_1, x_2, \ldots, x_n\right)\right) = 0$$
$$f_2\left(x_1, x_2, \ldots, x_n, u_1\left(x_1, x_2, \ldots, x_n\right), u_2\left(x_1, x_2, \ldots, x_n\right), \ldots, u_m\left(x_1, x_2, \ldots, x_n\right)\right) = 0$$
$$\vdots$$
$$f_m\left(x_1, x_2, \ldots, x_n, u_1\left(x_1, x_2, \ldots, x_n\right), u_2\left(x_1, x_2, \ldots, x_n\right), \ldots, u_m\left(x_1, x_2, \ldots, x_n\right)\right) = 0$$

If we suppose that we have solutions through some starting point (x^0, u^0), then we can differentiate each of the m equations, with respect to each of the variables $x_1, x_2, \ldots, x_n$ in turn. Using the Chain Rule, we then obtain a matrix equation of the form $\nabla_x f + \nabla_u f \bullet \nabla_x u = 0$, where $f = (f_1, f_2, \ldots, f_m)$, $u = (u_1, u_2, \ldots, u_m)$, $x = (x_1, x_2, \ldots, x_n)$ and the subscripts on the *nabla* ∇ indicate partial derivatives with respect to the variables indicated by the subscripts. This last equation becomes

$$\nabla_u f \bullet \nabla_x u = -\nabla_x f.$$

If it happens that the matrix $\nabla_u f$ is invertible, that is, if $\det \nabla_u f \neq 0$, then we obtain

$$\nabla_x u = -[\nabla_x f]^{-1} \bullet \nabla_x f.$$

In other words, we obtain an explicit and well defined system of slope fields for the partial derivatives of the unknown functions $u_1, u_2, \ldots, u_m$.

As long as knowing the slope fields is enough to infer the existence of the functions themselves, then a solution is obtained. As already discussed, geometry suggests an affirmative answer, at least as long as the quantities involved are sufficiently smooth, and as long as there is at least one solution point from which we can start following the partial derivatives to trace out the surface of a solution function. Indeed, as our examples and calculations suggest, the solution of the Implicit Function Problem can be reduced to inverting (and integrating) a special system of partial differential equations, from a given initial point.

Because of the fact that this approach relies on the existence of the inverse matrix of functions determined by $[\nabla_u f]^{-1}$, a proof of the Implicit Function Theorem

naturally depends on details that come from the general Inverse Function Theorem – the proof of which is beyond the scope of these notes. Still, perhaps it is becoming quite plausible that, at least under certain conditions, **the Inverse Function Theorem implies the Implicit Function Theorem.**

One common version of the Implicit Function Theorem is stated as follows:

Theorem 1.2. Suppose that the system of equations

$$\nabla_u f \bullet \nabla_x u = -\nabla_x f.$$

$$f_1(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0$$

$$f_2(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0$$

$$\vdots$$

$$f_m(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) = 0$$

is defined by continuously differentiable functions $f = (f_1, f_2, \ldots, f_m)$ defined for all $(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m)$ in $\mathbb{R}^n \times \mathbb{R}^m$. Suppose that (x^0, u^0) is a particular solution to $f(x^0, u^0)$. If when evaluated at (x^0, u^0), $\det \nabla_u f \neq 0$, then there exists an open neighborhood of A of x^0 and an open neighborhood of U of u^0 and a unique function $u = (u_1, u_2, \ldots, u_m) : A \to U$ satisfying $u(x^0) = u^0$ and $f(x, u(x)) = 0$ for all x in A. Furthermore, this solution function $u : A \to U$ also is continuously differentiable.

Example 1.20. Consider $f(x, u) = 4x + u^2 = 1$. This example is selected because we can first use basic algebra to solve the problem. Then we can compare to what we get if use the Implicit Function Theorem.

If $x = \frac{1}{8}$, then $u^2 = 1 - 4\left(\frac{1}{8}\right)$. So $u = \sqrt{\frac{1}{2}} = \frac{1}{\sqrt{2}}$ or $u = -\sqrt{\frac{1}{2}} = -\frac{1}{\sqrt{2}}$. In other words, for $x = \frac{1}{8}$ there are two solutions, and so two starting points that solve the equation $4x + u^2 = 1$, namely $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$ and $\left(\frac{1}{8}, -\sqrt{\frac{1}{2}}\right)$.

If we try another x close to $x = \frac{1}{8}$, say within $\frac{1}{8} - \frac{1}{32} < x < \frac{1}{8} + \frac{1}{32}$, then we can repeat the same algebra to get another choice of solutions, $u = +\sqrt{1 - 4x}$ or $u = -\sqrt{1 - 4x}$. Of course, the only way to obtain a continuous function is to be consistent in our choice, that is, either stay with the positive values $u = +\sqrt{1 - 4x}$

or stay with the negative values $u = -\sqrt{1 - 4x}$. Otherwise, if we mix these answers, taking some positive values for u and then taking some negative values for u, the heights u will be jump back and forth across x- axis, and the set of solution points we get will clearly not determine a continuous function $u = u(x)$.

So, if we happen to start with the solution $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$, then in order to have a continuous function $u = u(x)$ that solves $4x + u^2 = 1$, we will need to stay with the positive values for u. In other words, given the starting point $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$, the solution function $u = u(x)$ is unique. Of course, given the starting point $\left(\frac{1}{8}, -\sqrt{\frac{1}{2}}\right)$ we get another unique solution function. So, it is not that there is a unique solution function as such, but the function becomes unique once we specify the starting point (x^0, u^0).

If we now instead take the approach of the Implicit Function Theorem, we start with an initial point that solves the equation, for example, $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$. Now, since $f(x, u) = 4x + u^2 = 1$, we get $\det \nabla_u f = 2u$. At the initial point $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$, $u = \frac{1}{\sqrt{2}} \neq 0$. The Implicit Function Theorem now tells us that there exists a unique solution function $u = u(x)$ that both goes through the point $\left(\frac{1}{8}, \sqrt{\frac{1}{2}}\right)$ and that locally solves the equation $4x + u^2 = 1$. Note again that while the Implicit Function Theorem provides the existence of a solution, it does not provide an explicit form of the solution.

We now have some indication of how the Implicit Function Theorem depends on the Inverse Function Theorem, or rather that Inverse Function Theorem implies the Implicit Function Theorem. As you may now be aware, there are significant differences between the two theorems. Where the Inverse Function Theorem provides criteria for when an inverse function exists, the Implicit Function Theorem provides sufficient conditions for the existence of (possibly many) functions that solve an initial set of equations. Uniqueness is obtained only if we assume a given

starting point. Still, the calculations we have done so far are beginning to reveal that the two theorems are intimately related. And so the question can arise, can we go the other way?

Does the Implicit Function Theorem imply the Inverse Function Theorem?

To explore this question, let's go back to the familiar example $y=\sqrt{1-x^2}$, the function for the upper half circle in the (x, y) plane. If we restrict to $0<x<1$ this function is invertible. To see that, one may appeal to classical geometry. To be sure that we avoid circular logic however, it is better to invoke the one variable calculus version of the inverse function theorem. Taking the derivative, we get $\frac{dy}{dx}=\frac{-x}{\sqrt{1-x^2}}$. On the open unit interval this is strictly negative, from which it follows that the function is strictly decreasing and invertible on that open interval. For this one example the problem then is solved, that is, $y=\sqrt{1-x^2}$ is invertible on the open interval $0<x<1$.

What we are looking for though is a rather more general result. Our question is whether or not the general Implicit Function Theorem implies the general Inverse Function Theorem. That may be a high aim. But, if we look at this special case carefully, it is in fact possible to get some very good clues. To do that effectively, we will need to lift our view beyond the particular form of the function $y=\sqrt{1-x^2}$, and instead pose the question in terms of functions and operations. It is probably now a familiar fact to you, that the Implicit Function Theorem does not provide an explicit solution, but the existence of a solution (implicitly defined). So, here we also should not be looking for an explicit inverse, but for existence of an inverse.

Suppose then we are given a function $y=f(x)$, such as $y=\sqrt{1-x^2}$. If an inverse exists then there is a function $x=g(y)$ such that $y=f(g(y))=\sqrt{1-g(y)^2}=y$. Or, more briefly, $y=f(g(y))$. The Implicit Function Theorem regards systems of equations, and so rewriting this as an equation we can say that, if an inverse exists, then there is a function $x=g(y)$ such that $f(g(y))-y=0$. Coming closer to the form given in the Implicit Function Theorem, given the function $y=f(x)$, consider the equation $G(y,x)=f(x)-y=0$. Note the reversal of symbols – we use (y, x), not (x, y). That is, we bring our problem into alignment with the statement of the Implicit Function Theorem. In the Implicit Function Theorem as stated above, it is the function on the right that we produce as a function of the variable on the left. But, in the present problem, we are trying to find x as a function of y.

Now, observe that $(y^0, x^0)=\left(\frac{\sqrt{3}}{2}, \frac{1}{2}\right)$ is a solution to $G(y^0, x^0)=f(x^0)-y^0=0$. The Implicit Function Theorem tells us that if $\det \nabla_x G \neq 0$, then there exists a unique function $x=g(y)$ such that $\frac{1}{2}=g\left(\frac{\sqrt{3}}{2}\right)$ and such that on a neighborhood of $x^0=\frac{1}{2}$ we have $G(y,x)=y-f(g(y))=0$. But such a function $x=g(y)$ is clearly an inverse for $y=f(x)$. Note that in this special case where $y=f(x)$ is a real valued function of one real variable, $\det \nabla_x G=\frac{df}{dx}$. In other words, as one would hope, the Implicit Function Theorem reproduces the special case already known from basic one variable calculus.

Can we go further now? We are trying to get at the general case. Suppose then that $y=f(x)$ defines a function from $\mathbb{R}^m$ to $\mathbb{R}^m$. Is there a way to use the Implicit Function Theorem to determine whether or not the function $f: \mathbb{R}^n \to \mathbb{R}^m$ is invertible? Taking our cue from the discussion above, consider the equation $G(y, x)=f(x)-y$, where the symbols $f(x)$ and y now represent quantities in $\mathbb{R}^m$. Suppose that $G(y^0, x^0)=f(x^0)-y^0=0$ and that $\det \nabla_x f \neq 0$. Then, by the Implicit Function Theorem, there is a unique function $x=g(y)$ defined on an open neighborhood of x^0 and satisfying $G(y, g(y))=f(g(y))-y=0$. This of course means that $x=g(y)$ is the inverse function we are looking for.

Note 1.1. To go further would take us well beyond the introductory purpose of this book. We hope though that you now have some initial understanding of the Inverse Function Theorem, and how it is closely related to a theorem called the Implicit Function Theorem. For further study of these matters and rigorous proofs of these results, the reader may enjoy consulting one of the many excellent texts on advanced multi-variable calculus. There is, for example, the book by Jerrold E. Marsden and Michael J. Hoffman that has become a modern standard.

2

Discovering Calculus

Topics: Questions starting from antiquity, up to initial breakthroughs and advances of Newton, Leibniz. The derivative and the Fundamental "Theorem" of Calculus. Fruition of these notions in the definitions of Cauchy: convergence, derivative and the Cauchy integral. Power series.

2.1 AREAS

In the 5th century B.C.E., the Greek philosopher Zeno of Elea (ca. 490 B.C.E. – ca. 430 B.C.E.) posed what is now called "Zeno's Paradox". The apparent paradox led to much reflection, both mathematical and philosophical. The problem involves a distance to be traversed, say 2 yards. Suppose that a tortoise first travels one yard, and then 1/2 yard, and then another 1/4 yard, and then 1/16th yard, and so on. The paradox claims that the tortoise therefore cannot arrive at its destination.

Notice that the puzzle as stated does not mention anything about time. If the tortoise takes the same amount of time for each stage of its journey, then certainly it would seem reasonable to conjecture that it would not reach its destination. Is there really a paradox?

Be that as it may, the puzzle carries with it an interesting arithmetic problem. Let's look at a diagram, such as Figure 2.1.

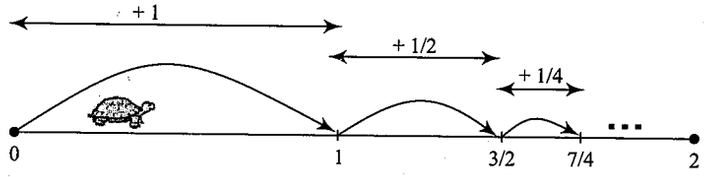


Figure 2.1

We discuss an approach to dealing with sums (of ratios) that was familiar to the great mathematician Archimedes (third century B.C.E.). As can be seen from the

diagram, at each stage of its journey, the tortoise is shy of the 2 yard mark. The distances traveled are as follows:

1

$1+\frac{1}{2} < 2$. By how much? In other words, $1+\frac{1}{2} = 2 - (\text{something})$.

Evidently, $$1+\frac{1}{2} = 2-\frac{1}{2}$$

We may continue this reasoning.

$$1+\frac{1}{2}+\frac{1}{4} < 2. \text{ By how much?}$$

$$1+\frac{1}{2}+\frac{1}{4} = 2-\frac{1}{4}$$

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8} < 2. \text{ By how much?}$$

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8} = 2-\frac{1}{8}$$

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16} < 2. \text{ By how much?}$$

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16} = 2-\frac{1}{16}$$

Conjecturing that the pattern continues, we get

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots+\frac{1}{2^n} = 2-\frac{1}{2^n}$$

We can look to verifying this general formula later. In the meantime, notice that n counts the number of stages in the journey, beyond the first yard. As n increases, we get distances traveled equal to: $2-\frac{1}{2}, 2-\frac{1}{4}, 2-\frac{1}{16}, 2-\frac{1}{32}$, and so on. In other words, as the number of stages in the journey increases (or as n increases), the closer the distance traveled is to 2 yards. There is, as it were, a *target value* of 2 [Bressoud, 11].

For another example, let's look to sums of powers of $\frac{1}{3}$. Might a similar approach be possible?

Exercise 2.1. Make a diagram similar to Fig. 2.1, but for adding powers of $\frac{1}{3}$.

Let's see how the sums work out.

$1 + \frac{1}{3} < ?$ In other words, what should we put here? In the first example, this was the target value. Evidently, we can't use the same target value. What might be a good candidate? We need something that in some way depends on the fact that this time we are using, not 2's, but 3's. Following that clue, how can the $1 + \frac{1}{2}$ and the target value 2 both be expressed in terms of 2, and perhaps fractions involving 2, the point being to hopefully see a pattern that might carry forward to the case of using 3's?

One way to do that is to observe that $1 + \frac{1}{2} < 1 + 1 = 1 + \frac{1}{(2-1)}$.

Then $1 + \frac{1}{(2-1)}$ is the target value.

Might the next case also be so concise? In other words, does a similar pattern work for the powers of $\frac{1}{3}$.

Cleary, $1 + \frac{1}{3} < 1 + \frac{1}{2}$, where the denominator "2" of "$\frac{1}{2}$" is simply 1 less than 3. If $1 + \frac{1}{2}$ is the target value, how far is $1 + \frac{1}{3}$ from that target? Consider then

$$\left(1+\frac{1}{2}\right)-\left(1+\frac{1}{3}\right) = \frac{1}{6} = \frac{1}{(3\cdot 2)}$$

That is,
$$1+\frac{1}{3} = \left(1+\frac{1}{2}\right)-\frac{1}{(3\cdot 2)}$$

So far so good. Does a pattern emerge?

Is $1+\frac{1}{3}+\frac{1}{9}<1+\frac{1}{2}$, and if so, by how much?

This time
$$1+\frac{1}{3}+\frac{1}{9} = \left(1+\frac{1}{2}\right)-\frac{1}{(9\cdot 2)}$$

For the next case,
$$1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27} = \left(1+\frac{1}{2}\right)-\frac{1}{(27\cdot 2)}$$

Conjecturing that the pattern continues, we get

$$1+\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+\ldots+\frac{1}{3^n} = \frac{3}{2}-\frac{1}{\left[\left(3^n\right)\cdot 2\right]}$$

Exercise 2.2. Using the same approach, develop a formula for sums of powers of $\frac{1}{2}$.

Clue: $1 + \frac{1}{4} < 1 + \frac{1}{3}$.

Formula: $1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\ldots+\frac{1}{4^n} = \frac{4}{3}-\frac{1}{\left[\left(3^n\right)\cdot 3\right]}$

Note that the purpose of the Exercise is not just to get the answer, but to become familiar with the details of the approach.

Exercise 2.3. Using the same approach, develop a formula for sums of powers of $\frac{1}{5}$.

Clue: $1 + \frac{1}{5} < 1 + \frac{1}{4}$.

Formula: $1+\frac{1}{5}+\frac{1}{25}+\frac{1}{125}+\ldots+\frac{1}{5^n} = \frac{5}{4}-\frac{1}{\left[\left(5^n\right)\cdot 4\right]}$

Let's now go on to a geometry problem that was solved by Archimedes, the solution of which involved rather similar patterns of sums as in the story of the tortoise. To set the stage, recall that in geometry we have area formulas for standard figures such as rectangles, parallelograms, and consequently for triangles as well. By using such standard figures to approximate, we can get some insight into other figures as well. We can even get a formula for the area of a circle. For a circle of radius r, a square of dimensions $r \times r$ is easily seen to cover one quadrant of the circle. We can ask how many of such square quadrants (of dimension $r \times r$) are needed to cover the area of the whole circle? See Figure 2.2.

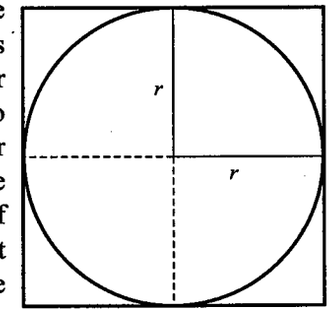


Figure 2.2

Classical geometric argument reveals that the ratio $\frac{[\text{Circle Area}]}{r^2}$ does not depend on the radius. This special ratio is therefore given its own symbol π, and for every circle of radius r its area is given by $A = \pi r^2$. From the diagram, the ratio π is

evidently less than 4. The ancients used approximating triangles to show that π is more than 3. In other words, the ratio π is some number somewhere between 3 and 4: $3 < \pi < 4$.

What though about other geometric figures?

Question 2.1. For example, what is the area under a parabola bounded by a straight line segment *AB*, as seen in Fig. 2.3?

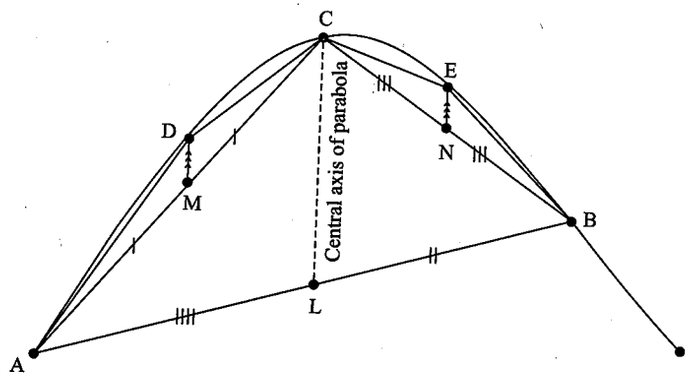


Figure 2.3

Archimedes (3rd c. B.C.E.) gave a sophisticated geometric solution to this problem that involved subtle combinations of geometric constructions and large sums of ratios [Heath, pp. 233-252].

As in Figure 2.3, Archimedes used triangle ΔABC to get a lower estimate on the area under the parabola.

Let *M* and *N* be the midpoints of *AC* and *BC*; and let *MD* and *NE* be parallel to the central axis of the parabola. Archimedes then used geometric results known in his time to show that the areas satisfy

$$\Delta CDA = \frac{1}{8}(\Delta ABC) \text{ and } (\Delta CEB) = \frac{1}{8}(\Delta ABC)$$

It follows that $$\Delta CDA + \Delta CEB = \frac{1}{4}(\Delta ABC)$$

[Burton, p. 198; Eves, p. 382]

Notice that AC and BC are again straight line segments under the parabola. So, if we again fill in areas under the parabola in the same way, we can use the same argument.

If we sum the results to the layers of triangles nestled under the parabola, we get a sum of triangular areas that, as seen from the diagram, approximates the actual area more and more closely.

From our calculations, the first approximation is $(\Delta ABC) + \frac{1}{4}(\Delta ABC) = (\Delta ABC)\left[1+\frac{1}{4}\right]$

The second approximation is

$$(\Delta ABC) + \frac{1}{4}(\Delta ABC) + \frac{1}{16}(\Delta ABC) = (\Delta ABC)\left[1+\frac{1}{4}+\frac{1}{16}\right].$$

The third approximation is

$$(\Delta ABC) + \frac{1}{4}(\Delta ABC) + \frac{1}{16}(\Delta ABC) + \frac{1}{64}(\Delta ABC) = (\Delta ABC)\left[1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}\right].$$

And so on.

After the n^{th} stage, the approximating sum of triangular areas nestled under the parabola is

$$(\Delta ABC)\left[1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\ldots+\frac{1}{4^n}\right].$$

But, what does that tell us? Evidently, part of the problem is to understand how to add powers of $\frac{1}{4}$. But, we already have a solution for that kind of sum, namely,

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

$$1+\frac{1}{4}+\frac{1}{16} = \frac{4}{3}-\left(\frac{1}{16}\right)\left(\frac{1}{3}\right)$$

$$1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64} = \frac{4}{3}-\left(\frac{1}{64}\right)\left(\frac{1}{3}\right)$$

Each of the sums is less than $\frac{4}{3}$. As we add more triangles, though, the sums get closer to 4/3.

With Archimedes, we can jump to the general case

$$1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\ldots+\frac{1}{4^n} = \left(\frac{4}{3}\right)-\left(\frac{1}{4^n}\right)\left(\frac{1}{3}\right).$$

How does this help us see what the area under the parabola is?

As we add more triangles under the parabola, the sum of powers of $\frac{1}{4}$ is always less than $\frac{4}{3}$. But, the short fall to $\frac{4}{3}$ gets smaller and smaller.

As Archimedes observed, the area under the parabola therefore cannot be less than $\frac{4}{3}$ (ΔABC). For, if K is any number less than $\frac{4}{3}$ (ΔABC), looking at our formula, we can see that by adding enough triangles, we can obtain area under the parabola that by-passes K and gets closer to $\frac{4}{3}$ (ΔABC).

Finally, from the diagram, the areas of the sum of triangles accumulate toward the area under the parabola.

We conclude that the area under the parabola is exactly $\frac{4}{3}$ (ΔABC).

Key Insight 2.1. We can use finite sums of triangles to better and better approximate (and consequently reveal) the exact area under a parabola.

Exercise 2.4. See Fig. 2.4. Consider a parabolic region bounded below by a straight line segment that is two units in length, perpendicular to the central axis of the parabola. Suppose that the vertex of the parabola is one unit in distance in height off the straight line. Use the approach of Archimedes to find the area of the parabolic region.

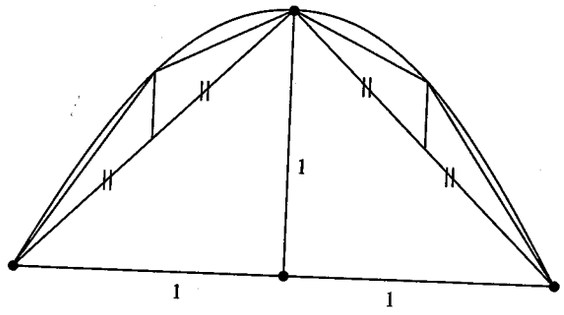


Figure 2.4

Notes 2.1. There is no claim that this result provides a rigorous definition. Indeed, while this type of approximation became a common technique in ancient and modern mathematics (e.g. Kepler [Burton, 330]), definitions for approximation did not emerge until the 18th century (Cauchy, et al), long after the initial discovery of calculus. Note also that the claim that the sum of triangular areas approximates

the parabolic area rests on an insight into the diagram, rather than on any axiom or theorem. In the early days of geometry, there are numerous results that subtly appeal to a diagram to justify a conclusion. It wasn't until the 20th century that some clarification was obtained on the rigor lacking in Euclidean geometry and, for example, the need for a "betweeness axiom". For some references on using approximation to calculate areas and volumes, see [Eves, 382], [Bressoud, 9]; and [Burton, 330].

2.2 RATES

Jon is a farmer with several acres of pasture that have lain fallow for some time. Two roads border the property, one runs north along the western side of the field, and the other runs east, along the southern side. See Figure 2.5. Jon has decided to plow a square garden in the south-west corner of the field to grow vegetables for a market garden.

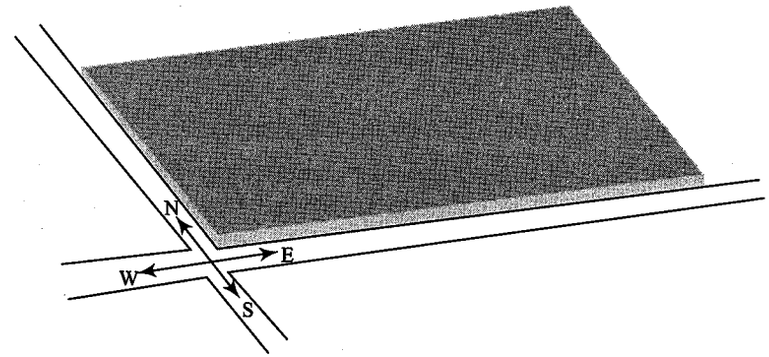


Figure 2.5

He has a roto-tiller that has different blade options. The largest blade gives a tiller cut that is 1 yard across. He tills along the outer boundary of the square, on each pass adding both length to the square and area to the garden. As the length of square increases, it becomes more work to add each additional swath across the boundary of the ever larger square. As the length of the square grows, so does the area that is added each time he makes a pass along the boundary.

Jon is a curious fellow, and so while resting on the weekend he begins to wonder about his project. He has the following question:

Question 2.2. In terms of its length, at what rate does area of a square increase?

He draws the diagram in Figure 2.6 below, and starts to calculate:

When the square is 10 yds by 10 yds, one pass adds 2(10) + 1 sq. yds. of new area;

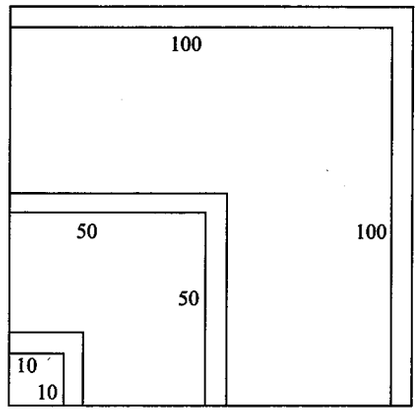


Figure 2.6

When the square is 50 yds. by 50 yds., one pass adds 2(50) + 1 sq. yds. of new area;

When the square is 100 yds. by 100 yds., one pass adds 2(100) + 1 sq. yds. of new area.

Is there a pattern to the rate at which the areas increase?

Except for the 1 sq. yd. at the north-east corner, the area increases at a rate of 2(length) sq. yds. per pass, or numerically, 2(length) sq. yds. of area per yd. of added length.

Jon has different blade options for the tiller; and a shorter blade is sometimes easier to use. He therefore wonders how the rate of change of garden area might be affected by choice of a shorter blade.

Let's suppose with Jon that the length of the square garden that we start with is 100 yds. in length, and that the tiller width is Δx. The starting area is then $(100)^2$, while the new area is $(100 + \Delta x)^2$. Looking to the diagram, the added area comes from each side plus the corner piece, that is, the added area is $2\,[(100)\,\Delta x] + (\Delta x)^2$.

More precisely, we can use algebra to calculate the ratio:

$$\frac{\text{Change in Area}}{\text{Change in Length}} = \frac{(100+\Delta x)^2-(100)^2}{\Delta x} = \frac{2(100)\Delta x+(\Delta x)^2}{\Delta x} = 2(100)+\Delta x.$$

If the blade width is $\frac{1}{2}$ yd., then the rate is $2(100)+\frac{1}{2}$ sq. yds. per yd.;

If the blade width is $\frac{1}{3}$ yd., then the rate is $2(100)+\frac{1}{3}$ sq. yds. per yd..

If the shortest $\frac{1}{4}$ yd. blade is chosen, then the rate is $2(100)+\frac{1}{4}$.

In other words, the smaller the change in length Δx, the closer the rate is to being simply the boundary length of the square, 2(100).

Exercise 2.5. (i) Let's open up the problem somewhat. Suppose that the garden is x yds. by x yds.

(ii) For a small change in length Δx, what is the main contribution to the ratio $\frac{\text{Change in Area}}{\text{Change in Length}}$?

(iii) If the square garden is 150 yds. long, and the added length is $\frac{1}{10}$ of a yd., approximately how much new area is added?

If the square garden is x yds. long, and the length is increased by a small amount Δx yds., approximately how much new area is added?

Answers: (i) For small Δx, main contribution to rate is the length of the border, $2x$ (sq. yds. per yd.); (ii) Added area is approximated by the product of the rate multiplied by the change Δx. In the present case this is $[2(150)]\left(\frac{1}{10}\right) = 30$ sq. yds.; (iii) This is the general case. Added area is approximated by the product of the rate and the change Δx, that is, $[2x]\,\Delta x$ sq. yds.

Key Insights 2.2. 1. We may use small changes Δx to better and better approximate (and consequently obtain) the exact rate of change of a square area.

2. We can turn this result around. If we have the exact rate, then for a small change Δx, the change in area is approximated by the product of the rate (in this case the boundary length $2x$) by the change Δx, that is, $[2x]\,\Delta x$.

Example 2.1. Galileo was interested in the nature of free-fall motion. By rolling balls down gently sloped planks of wood he created a controlled slower "free-fall" (or rather, a "free-roll"). He compared measurements of distances and times, and thereby discovered the Law of Falling Bodies: The distance an object falls is proportional to the square of time.

Using s for feet and t for seconds as our units for distance and time, Galileo's result is that $s = 16t^2$.

Recall that average speed is the ratio given by distance/time.

For the following questions, you may assume Galileo's Law:

How far does an object fall in the time interval $t = 1$ to $t = 2$?

What is the average speed for that second? For a small time interval $t = 1$ to $t = 1 + \Delta t$, what is the main contribution to the average speed?

How far does an object fall in the time interval $t = 2$ to $t = 3$?

What is the average speed? For a small time interval $t = 2$ to $t = 2 + \Delta t$, what is the main contribution to the average speed?

How far does an object fall in the time interval $t = t_1$ to $t = t_1 + \Delta t$?

What is the average speed? For a small time interval $t = t_1$ to $t = t_1 + \Delta t$, what is the main contribution to the average speed?

Do the computations remind you of Jon's garden? In other words, the average speed can be used to approximate the exact speed at a particular time.

What is the exact speed at a time $t = t_1$?

Answers: For the time interval $t = t_1$ to $t = t_1 + \Delta t$, the average speed is $16\{[2t_1] + \Delta t\}$. The exact speed at $t = t_1$ must therefore by $16\{[2t_1]\} = 32t_1$.

Notice that as time t_1 increases, so does the speed $32t_1$. How would you describe what happens to the speed of a stone that is dropped off of a bridge?

Example 2.2. Jan is a carpenter, and is making wooden cubic storage crates. The crates need to be strong, so Jan uses expensive hardwood along the edges of each box. Suppose that a cubic box is x inches in length along each edge. Its volume is then $V = x^3$. If l inch of the expensive hardwood is added to each length, the volume increases. When the initial cube is small, the added volume is modest. If the initial cube is large, an extra inch in length can increase the volume substantially. This leads Jan to the following question:

Question 2.3. In terms of length x, what is the rate at which volume of a cube increases?

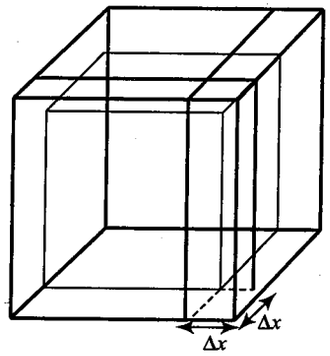


Figure 2.7

Suppose that we increase the length of the cube by Δx. Using Figure 2.7, it can be seen that while volume is added along three edges and a corner, the main added volume would seem to come from the three faces of the cube. For along each face, there is an added depth Δx. Each face has length and width x, and so the resulting added volume across each face is $x^2\Delta x$. There are though three faces. As an approximation, the main added volume would therefore seem to be $3x^2\Delta x$.

The ratio $\dfrac{\text{Added Volume}}{\text{Added Length}} = \dfrac{\Delta V}{\Delta x}$ is then approximated by $\dfrac{\Delta V}{\Delta x} = \dfrac{3x^2\Delta x}{\Delta x} = 3x^2$ cubic inches per inch of length.

How accurate is this approximation?

We can use algebra to get a more precise result.

$$\Delta V = (x + \Delta x)^3 - (x)^3 = 3x^2\Delta x + 3x(\Delta x)^2 + (\Delta x)^3.$$

The ratio is then

$$\frac{\Delta V}{\Delta x} = \frac{(x+\Delta x)^3 - (x)^3}{\Delta x} = \frac{3x^2\Delta x + 3x(\Delta x)^2 + (\Delta x)^3}{\Delta x} = 3x^2 + 3x(\Delta x) + (\Delta x)^2.$$

Remember that the length x is the fixed initial length of the cube. The question is what the ratio is for small changes Δx. In the last ratio, the two extra terms are both multiplied by Δx and $(\Delta x)^2$ respectively, while the first term $3x^2$ is unaffected by the quantity Δx. So, the smaller the change Δx, the closer the rate of change is to being $3x^2$, as anticipated from the geometry of the diagram. This can be called the *exact rate of change.*

Exercise 2.6. If the length of a cube is 40 inches, what is the exact rate at which new volume can be obtained? If the length is increased by approximately 0.5 inches, what approximately is the added volume?

Key Insights 2.3. 1. We may use small changes Δx to better and better approximate (and consequently obtain) the exact rate of change of a cubic volume.

2. We can turn this result around. If we have the exact rate, then for a small change Δx, the change in volume is approximated by the product of the rate (in this case the surface area $3x^2$) by the change Δx, that is, $[3x^2]\Delta x$.

Example 2.3. (Other integer powers of x) Suppose that x represents that number of cells in a cell population and that the number of protein molecules is given by $R(x) = x^4$. For instance, if there are $x = 10$ cells, then the number of protein molecules would be $R(10) = 10^4 = 10{,}000$. As with our other examples, we can enquire into the rate at which the number of protein molecules increases, as compared to the number x of cells present.

For Δx small,

$$\Delta R = (10+\Delta x)^4 - (10)^4 = 4(10)^3\Delta x + [\text{terms with higher powers of } \Delta x]$$

So,

$$\Delta R = (10+\Delta x)^4 - (10)^4 = 4(10)^3\Delta x + [\text{terms with } \Delta x^2, \Delta x^3, \Delta x^4]$$

and therefore

$$\frac{\Delta R}{\Delta x} = 4(10)^3 + \left[\text{terms with } \Delta x, \Delta x^2, \Delta x^3\right].$$

It now follows that for small changes in x given by Δx, the ratio of change is approximately $4(10)^3$ protein molecules per cell.

Notice that we did not need to make use of the binomial theorem. For it was enough to identify the part of the ratio $\frac{\Delta R}{\Delta x}$ essentially unaffected by the quantity Δx. So, might this approach work for other integer powers of x?

Suppose that $y(x) = x^n$. For a change in x given by Δx, $\Delta y = (x + \Delta x)^n - (x)^n$. Following the rules of algebra, the terms are produced as products, one factor from each power. We therefore get a sum of the form

$$x^n + nx^{n-1}\Delta x + (\#)(x)^{n-2}(\Delta x)^2 + (\#)(x)^{n-3}(\Delta x)^3 + \cdots + (\#)(x)(\Delta x)^{n-1} + (\Delta x)^n - (x)^n$$

$$= nx^{n-1}\Delta x + (\#)(x)^{n-2}(\Delta x)^2 + (\#)(x)^{n-3}(\Delta x)^3 + \cdots + (\#)(x)(\Delta x)^{n-1} + (\Delta x)^n$$

The ratio of change is

$$\frac{(x+\Delta x)^n - (x)^n}{\Delta x} = nx^{n-1} + (\#)(x)^{n-2}(\Delta x)^1 + (\#)(x)^{n-3}(\Delta x)^2 + \cdots (\#)(x)(\Delta x)^{n-2} + (\Delta x)^{n-1}$$

No matter what the coefficients (#) happen to be *, for a small change in x given by Δx, the main contribution to the ratio of change is nx^{n-1}. Furthermore, since x is fixed, the smaller Δx is, the closer the ratio of change is to this quantity nx^{n-1}. In other words, the exact rate of change of $y(x) = x^n$ is nx^{n-1}.

* For present purposes we do not require detailed knowledge of the coefficients (#). However, the Binomial Theorem was known in Newton's time and explicitly identifies the coefficients via the following well known expansion:

$$(a+b)^n = \binom{n}{n}a^n + \binom{n}{n-1}a^{n-1}b + \binom{n}{n-1}a^{n-1}b + \cdots + \binom{n}{2}a^2b^{n-2} + \binom{n}{1}a^1b^{n-1} + \binom{n}{0}b^n$$

where $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$.

Notation

We have been developing a notion of "exact rate of change". It is obtained from ratios of change of the form $\frac{\Delta y}{\Delta x}$. As Δx gets small, the exact rate is identified as that part of the ratio $\frac{\Delta y}{\Delta x}$ that makes the prevailing contribution.

Notation would be useful to distinguish the prevailing contribution to the ratios from the approximating ratios. The symbolism that Leibniz used is $\frac{dy}{dx}$. Newton's symbolism is $y'(x)$. Since this "exact rate" $\frac{dy}{dx} = y'(x)$ is *derived* from the approximating ratios for the given formula $y(x)$, the exact rate $\frac{dy}{dx} = y'(x)$ is traditionally called the *derivative* of $y(x)$.

Example 2.4. (Product Rule) In this example, suppose that that Jan and Jon are working together on tilling a rectangular market garden. Just as in the example above, suppose that the garden is in the south-west corner of a larger field. Jon uses a roto-tiller and cuts east and west across the northern boundary of the garden, while Jan is using a different roto-tiller, making north-south cuts along the eastern boundary of the garden area. After a day of work, the garden is 100 ft. east-west, and 50 feet north-south. See Figure 2.8. The next day, Jon and Jan work for an hour. In that time, Jan clears an additional 6 feet north, while Jan clears 7 feet east. What area of ground was tilled in that hour?

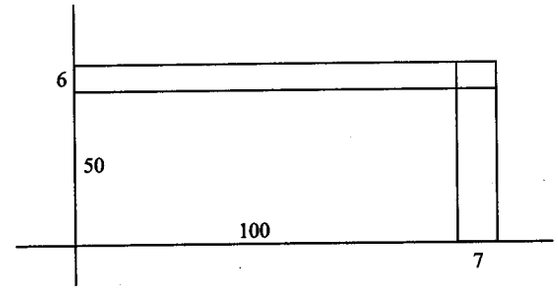


Figure 2.8

Except for the corner at the north-east corner, the area cleared is 6(100) + 7(50) sq. feet.

Exercise 2.7. Suppose that $f(x)$, $g(x)$ are two given functions, with derivatives $f'(x)$ and $g'(x)$. What is the derivative of $A(x) = f(x) \cdot g(x)$?

Clue: See Figure 2.9. Represent the product as a rectangular area, with the length of the northern boundary given by $f(x)$ and the length of the eastern boundary given by $g(x)$.

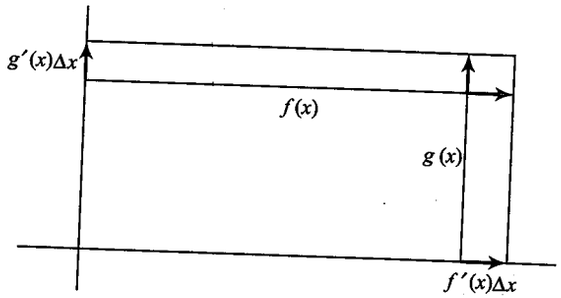


Figure 2.9

Each length has its own rate of change. That is, for small Δx, the changes in $f(x)$ and $g(x)$ are approximated by $f'(x)\,\Delta x$ and $g'(x)\,\Delta x$ respectively. Combining the change in areas coming from the north boundary and east boundary, what is the approximate total change in area? By evaluating the ratio $\dfrac{\text{change in area}}{\text{change in length}}$, what is the exact rate of change? The answer is called the "product rule": In the Leibniz notation the result is written $\dfrac{d}{dx}(f \cdot g) = \dfrac{df}{dx} \cdot g + f \cdot \dfrac{dg}{dx}$. In Newton's notation, the this becomes $(f \cdot g)' = f'g + fg'$.

Example 2.5. (Quotient Rule)

Suppose that $g(x)$ is a function, with known rate of change given by the derivative $\dfrac{dg}{dx} = g'(x)$. If $g(x) > 0$ increases rapidly, then $\dfrac{1}{g(x)}$ decreases rapidly. In a similar way, if $g(x) > 0$ decreases rapidly toward zero, then $\dfrac{1}{g(x)}$ evidently will increase. We are assuming that we know the function $g(x)$, and the exact rate $g'(x)$ at which the function changes. Is this enough to go on and determine the exact rate at which the reciprocal $\dfrac{1}{g(x)}$ changes?

In order to begin working with $\dfrac{1}{g(x)}$, it is helpful to know how this quantity is defined. Recall the algebraic definition of a reciprocal $\dfrac{1}{5}$ is as the solution of the product $x \cdot 5 = 1$. Of course, the reciprocal of a function is defined in an analogous

way, that is, for $g(x) \neq 0$, $\dfrac{1}{g(x)}$ is defined to be the solution of the product $F(x) \cdot g(x) = 1$. In more familiar notation, $\dfrac{1}{g(x)}$ is defined by the product $\dfrac{1}{g(x)} \cdot g(x) = 1$.

We are trying the find $\left[\dfrac{1}{g(x)}\right]'$, that is, the derivative of the term $\left[\dfrac{1}{g(x)}\right]$ that happens to be a factor of the *product* $\dfrac{1}{g(x)} \cdot g(x) = 1$. But, we have a product rule.

Taking the derivative of the product, we get $\left[\dfrac{1}{g(x)} \cdot g(x)\right]' = 1' = 0$, which by the product rule becomes $\left[\dfrac{1}{g(x)}\right]' g(x) + \left[\dfrac{1}{g(x)}\right] g'(x) = 0$. The only unknown term in this last equation is the quantity that we are looking for, namely, $\left[\dfrac{1}{g(x)}\right]'$. It is a worthwhile exercise to now solve for this unknown and so obtain that

$$\left[\frac{1}{g(x)}\right]' = \frac{-g'(x)}{g^2(x)}.$$

Exercise 2.8. Use the formula $\left[\dfrac{1}{g(x)}\right]' = \dfrac{-g'(x)}{g^2(x)}$ to see that the derivative of any quotient is given by $\left[\dfrac{f}{g}\right]' = \dfrac{-f'g - fg'}{g^2}$. *Hint:* $\dfrac{f}{g} = f \cdot \dfrac{1}{g}$.

Example 2.6. (Fraction powers of x) Recall that the exponential notation for the positive square root of a positive number x is $x^{1/2}$. This notation is deliberately chosen to be consistent with the addition rule for integer exponents. For then $x^{1/2} \cdot x^{1/2} = x^{(1/2)+(1/2)} = x^1 = x$. Similarly, the cube root is given by the formula $x^{1/3} \cdot x^{1/3} \cdot x^{1/3} = x^{(1/3)+(1/3)+(1/3)} = x^1 = x$.

As with other familiar functions, as x varies, so does the square root, the cube root, and so on. It follows that we may ask at what rate the roots vary, as we vary the number x. *Clue:* We seek $\frac{d\left(x^{1/2}\right)}{dx} = \left(x^{1/2}\right)'$. To find the derivative of $x^{1/2}$, we need to know what we are working with. In other words, it would be useful to have the definition of $x^{1/2}$. But this term is defined by the product $x^{1/2} \bullet x^{1/2} = x$; and we have a product rule for derivatives! Calculating, we obtain

$$\left(x^{1/2}\right)' x^{1/2} + x^{1/2}\left(x^{1/2}\right)' = x' = 1$$

$$2\left(x^{1/2}\right)' x^{1/2} = 1$$

$$\left(x^{1/2}\right)' = \frac{1}{2}x^{-1/2}$$

What about the cube root, which is defined by the multiple product $x^{1/3} \bullet x^{1/3} \bullet x^{1/3} = x^{(1/3)+(1/3)+(1/3)} = x^1 = x$?

In this case, a first calculation gives $(x^{1/3})'\,(x^{1/3} \bullet x^{1/3}) + x^{1/3}\,(x^{1/3} \bullet x^{1/3})' = 1$. This does not quite yet give us the solution. Notice though that we can apply the product rule again. This gives $3(x^{1/3})'\,x^{1/3} = 1$ from which it follows that $\left(x^{1/3}\right)' = \frac{1}{3}x^{-1/3}$

Exercise 2.9. Using the definition $x^{1/4} \bullet x^{1/4} \bullet x^{1/4} \bullet x^{1/4} = x^{(1/4)+(1/4)+(1/4)+(1/4)} = x^1 = x$, find the derivative $\frac{d\left(x^{1/4}\right)}{dx} = \left(x^{1/4}\right)'$.

Exercise 2.10. Using the definition $x^{1/n} \bullet x^{1/n} \bullet \underset{n\text{–times}}{\ldots} \bullet x^{1/n} \bullet x^{1/n} = x^{(1/n)+\cdots+(1/n)} = x^1 = x$, find the derivative $\frac{d\left(x^{1/n}\right)}{dx} = \left(x^{1/n}\right)'$.

Exercise 2.11. Using the definition $x^{3/4} \bullet x^{3/4} \bullet x^{3/4} \bullet x^{3/4} = x^{(3/4)+(3/4)+(3/4)+(3/4)} = x^4$, find the derivative $\frac{d\left(x^{3/4}\right)}{dx} = \left(x^{3/4}\right)'$.

Exercise 2.12. Using the definition $x^{m/n} \bullet x^{m/n} \bullet \underset{n\text{–times}}{\ldots} \bullet x^{m/n} \bullet x^{m/n} = x^{(m/n)+\cdots+(m/n)} = x^m$, find the derivative $\frac{d\left(x^{1/n}\right)}{dx} = \left(x^{1/n}\right)'$.

Exercise 2.13. Conjecture a rule for the derivative of any positive real power of the form x^α, where $\alpha > 0$ is real.

Exercise 2.14. What is the derivative of x^{-1}? *Clue:* One approach is of course to use the result that $\left[\frac{1}{g(x)}\right]' = \frac{-g'(x)}{g^2(x)}$. Another approach is to go back to first principles and use a defining equation for $x^{-1}x^{-1}$, namely, $\left(x^{-1}\right) \bullet x = 1$?

Exercise 2.15. What is the derivative of $x^{-2/3}$? Again, in the present context there are two natural approaches. Explore this questions for other negative fraction powers of x of the form $x^{-m/n}$ where $\frac{m}{n} > 0$.

Exercise 2.16. What is the derivative of any negative real power of the form $x^{-\alpha}$, where $\alpha > 0$ is real. Again, for one approach we may write $x^{-\alpha} = \frac{1}{x^\alpha}$ and use $\left[\frac{1}{g(x)}\right]' = \frac{-g'(x)}{g^2(x)}$. Or, going to first principles, we may write $x^{-\alpha} \bullet x^\alpha = 1$.

Exercise 2.17. Conjecture a general rule for the derivative of any real power of the form x^α.

Example 2.7. (Chain Rule): Suppose that a car gets 30 miles per gallon, and uses 2 gallons per hour. How many miles per hour is the car traveling? In other words, if we know the rate $\frac{\text{miles}}{\text{gallon}}$; and if we also know the rate $\frac{\text{gallons}}{\text{hour}}$, then we can simply multiply the results to get the rate $\frac{\text{miles}}{\text{hour}}$.

Example 2.8. (Chain Rule): Suppose that gear A is connected to gear B; and that if gear A rotates once, then gear B rotates 3 times. If gear A rotates 4 times per second, how often does gear B rotate per second? See Figure 2.10

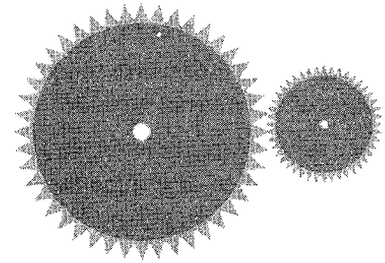


Figure 2.10

Now, consider two functions $C = f(y)$ and $y = g(x)$, connected not by teeth of a gear, but by mathematical composition. That is, let $C(x) = f(g(x))$. What is the derivative of this composition?

Clue: For a small change Δx near x, the change in $g(x)$ is approximately $\Delta g = g'(x)\,\Delta x$. But, just like when one gear is connected to another, for this small change $\Delta g = g'(x)\,\Delta x$ near $g(x)$, the change caused in the next function $f(x)$ is approximately $f'(g(x))\,\Delta g(x) = f'(g(x))\,[g'(x)\,\Delta x] = f'(g(x))\,g'(x)\,\Delta x$.

Putting this all together, we have that a small change Δx near x causes an overall change in the composition given by $\Delta[f(g(x))] \approx f'(g(x))\,g'(x)\,\Delta x$. We therefore obtain the ratio of change $\dfrac{\Delta\left[f\left(g(x)\right)\right]}{\Delta x} \approx f'\left(g(x)\right)g'(x)$; which leads to the result that $(f(g))'(x) = f'(g(x))\,g'(x)$.

Have you ever wondered why this is called the "chain" rule? There are several reasons, one of which is obtained by recalling a traditional notation for the composition of two or more functions: The composition $f(g(h(k(x))))$ is also written $(f \circ g \circ h \circ k)(x)$. Do you see links of the "chain" of dependence?

Example 2.9. (Derivative of an Inverse Function)

Solve for x in the equations $40 = 5x$; $50 = 5x$; $60 = 5x$.

More generally, given y, find the number x such that $y = 5x$.

This is called solving the inverse problem. See *Chapter 1 Discovering the Inverse Function Theorem.*

In the present example, the inverse function is of course just $f^{-1}(y) = \dfrac{1}{5}y$. In other words, to undo multiplication by 5 it is enough to simply divide by 5. For any non-horizontal line $y = f(x) = mx$, $m \neq 0$, solving for x in the same way, the inverse function is easily seen to be $f^{-1}(y) = \dfrac{1}{m}y$. Using the terminology of calculus, the derivative or slope of $f(x) = mx$, $m \neq 0$ is $f'(x) = m$ and the derivative of the inverse function is simply the slope $\left(f^{-1}\right)'(y) = \dfrac{1}{m}$.

Exercise 2.18. Use algebra to find the inverse function of the line $y = f(x) = mx + b$, $m \neq 0$. What is the slope of the inverse function?

Can we extend our work to other functions? That is, if $f(x)$ and its derivative $f'(x)$ are known functions, can we then determine the derivative $(f^{-1})'(y)$ of the inverse function $f^{-1}(y)$?

One way to approach this is to refer to the geometry of the situation. Near the reference point $(x, f(g))$ on the graph of $f(x)$, the function is approximated by the tangent line whose slope is $f'(x)$. In Example 2.19 and Exercise 2.18 above, the slope of the inverse line is simply the reciprocal of the original slope. So, in as much as the inverse tangent line approximates the inverse function, we might anticipate that the slope of the inverse function will be $\dfrac{1}{f'}$. But, where is this function to be evaluated, that is, $\dfrac{1}{f'}$ at what? Or, in symbols, $\dfrac{1}{f'} = \dfrac{1}{f'(?)}$?

The reference point for f is $(x, f(x))$, so the reference point for the inverse function is $(f(x), x)$. From using this geometric and algebraic argument, we can therefore anticipate that the derivative of the inverse function will be given by

$$\left(f^{-1}\right)'\left(f(x)\right) = \frac{1}{f'(x)}.$$

Note, however, that at this stage of our calculation, the $\left(f^{-1}\right)'\left(f(x)\right) = \dfrac{1}{f'(x)}$ is in terms of the variable x. A complete solution would require a formula in terms of y. But, under the hypothesis that the function $f(x)$ has an inverse function $f^{-1}(y)$, we can make the correspondences $y = f(x)$ and $x = f^{-1}(y)$. Substituting this into our solution we obtain $\left(f^{-1}\right)'(y) = \dfrac{1}{f'\left(f^{-1}(y)\right)}$, the traditional formula for the derivative of an inverse function.

For a more precise argument, recall how the inverse of a function is defined.

For again, having the definition gives us something to work with. Given a function $y = f(x)$, when it exists, its inverse satisfies $f^{-1}(f(x)) = x$. Observe that this is a composition, or chain, of two functions. Hence, the chain rule applies! We therefore obtain

$$\left[f^{-1}\left(f(x)\right)\right]' = x' = 1$$

$$\left(f^{-1}\right)'\left(f(x)\right) \cdot f'(x) = 1$$

Solving for the derivative of the inverse function gives

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

Since $y = f(x)$ and $x = f^{-1}(y)$, we obtain

$$(f^{-1})'(y) = \frac{1}{f'\left(f^{-1}(y)\right)}.$$

Example 2.10. Let $y = f(x) = \sin x$, the sine of an angle x, where in the present case we assume that the angle is given in radians. The inverse function on the other hand, starts with a sine (a ratio) and provides an angle (an arc length along the unit circle – remember, the angle is in radians) that would produce that ratio. In traditional notation, $x = \arcsin y = f^{-1}(y)$.

As an angle changes, the sine of the angle changes, hence there is a rate of change $f'(x)$, which as it happens is the cosine, that is $f'(x) = (\sin x) = \cos x$. (See any standard calculus text.) From our result above, we should now be able to calculate $(\arcsin y) = (f^{-1}(y))$, the derivative of the inverse to the sine function.

Our formula above is $\left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{-1}(y)\right)}$. Substituting the sine function, we get $(\arcsin y)' = \left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{-1}(y)\right)} = \frac{1}{\cos(\arcsin y)}$.

This is correct, but can we do better? Recall, for any angle θ where the inverse function is defined, we have $\sin(\arcsin \theta) = \theta$. If we could only express the denominator in terms of the sine function, we might be able to further reduce terms. Recall, however, that $(\cos x)^2 + (\sin x)^2 = 1$. For definiteness, let's take the positive square root, to see what we get, that is, take $\cos x = \sqrt{1-(\sin x)^2}$. Using this in our formula for the derivative of the inverse sine function, we get

$$(\arcsin y)' = \left(f^{-1}\right)'(y) = \frac{1}{f'\left(f^{-1}(y)\right)} = \frac{1}{\sqrt{1-\left[\sin(\arcsin y)\right]^2}}$$

$$= \frac{1}{\sqrt{1-[y]^2}} = \frac{1}{\sqrt{1-y^2}}.$$

In other words, $(\arcsin y) = \frac{1}{\sqrt{1-y^2}}$.

In the above calculation, we took a positive square root. Taking positive or negative square roots corresponds to the different quadrants of the unit circle. As the reader may verify, completely similar results are obtained in these different cases.

The Fundamental "Theorem" of Calculus

Suppose that yet another property is being tilled for market gardening. This time, the northern border follows a stream that meanders in a north-east direction.

The farmer ploughs in furrows that run north and south, on each pass extending the eastern border with a roto-tiller that gives a 1 foot wide cut. See Fig. 2.11.

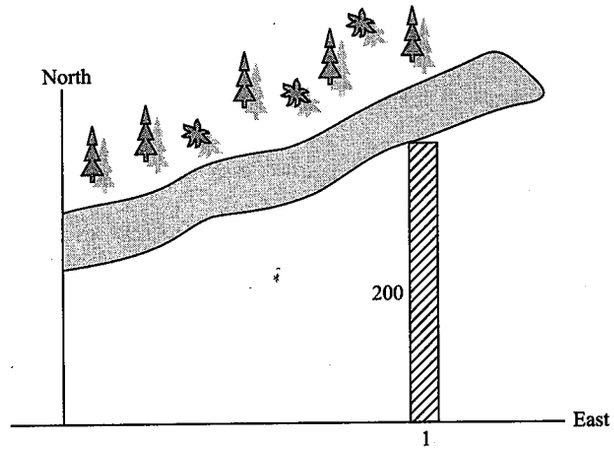


Figure 2.11

Starting at the point where the eastern border is 200 feet in length, what is the added area after one pass?

Now, suppose that $y = f(x)$, and represent its graph as in Fig. 2.12.

Let $A(x)$ be the area under the graph. What is the derivative $A'(x)$ of the area?

Clue: Starting at x where the "eastern" border is $f(x)$ feet in length, what is the added area after increasing x by Δx? In other words, the change in area $\Delta A \approx f(x)\,\Delta x$. Now look at the ratio $\frac{\Delta A}{\Delta x}$; and hence obtain that $\frac{dA}{dx} = f(x)$.

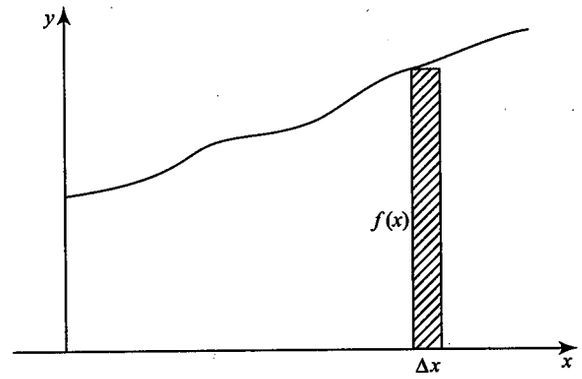


Figure 2.12

Key Insight 2.4. (The Fundamental "Theorem" of Calculus). The rate of change of area is the length of the advancing front line.

Example 2.11. Consider the function $A(x)$ that is area under the graph given by the function $f(t) = t^2$, for $1 \le t \le x$. Make a diagram. What is the rate at which area is swept out, as we move the front-line in the direction of the positive x axis? *Answer:* For each x, the length of the front-line is $y = x^2$. Therefore, by the key insight, the rate of change of the area is given by $\frac{dA}{dx} = f(x) = x^2$.

Example 2.12. Consider the function $A(x)$ that represents area under the graph of the function $f(t) = \frac{1}{t^2} = t^{-2}$, for $1 \le t \le x$. Make a diagram. What is the rate at which area is swept out, as we move the front-line in the direction of the positive x axis? *Answer:* For each x, the length of the front-line is $y = x^{-2}$. Therefore, the rate of change of the area is given by $\frac{dA}{dx} = f(x) = x^{-2}$.

Example 2.13. (Tangent Lines) For this example we will use calculus to study Galileo's formula for free-fall. Galileo's discovery is called The Law of Falling Bodies, and states that the distance an object falls is proportional to the square of time. Using the modern units of s for feet and t for seconds, the result is $s = 16t^2$.

In the discussion earlier in the chapter, we first focused on the calculation of average speed, or average rate. We then obtained the result that the exact speed at $t = t_1$ is $16\{[2t_1]\} = 32t_1$.

Free-fall can be thought of as a single process, and can be imagined as a continuum. Therefore, a natural way to represent this unity is by a single graph of $s = 16t^2$ on time and distance axes. The graph of $s = 16t^2$ is then a parabola.

Exercise 2.19. Looking back to the discussion earlier in the chapter, use geometry to represent the quantities referred to in the various questions on Galileo's Law.

For instance, for the time interval $t = 2$ to $t = 2 + \Delta t$ the average speed $\frac{16(2+\Delta t)^2 - 16(2)^2}{\Delta t}$ can be represented by the slope of the line segment joining (2, 64) to $((2 + \Delta t), 16(2 + \Delta t)^2)$.

To enhance your diagram, extend each such line segment to a line segment that extends the length of most of your diagram. Notice that the smaller Δt is, the closer the line is to being a tangent line at the point (2, 64). But, we have already determined that the exact speed at (2, 64) is obtained from the ratios $\frac{16(2+\Delta t)^2 - 16(2)^2}{\Delta t}$. See Fig. 2.13.

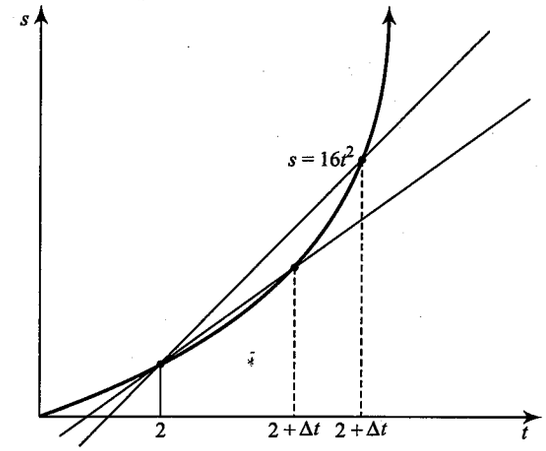


Figure 2.13

Therefore, we reach another key insight.

Key Insight 2.5. In geometry, the derivative or exact rate can be represented as the slope of the tangent line.

Exercise 2.20. Explore this idea with the graph of the formula $y = t^3$. Repeat the calculations just described for $s = 16t^2$.

Exercise 2.21. Using geometry, one might think of slope of a tangent line to a graph as being equivalent to derivative of the underlying formula. Note, however, that there are two essential terms in the phrase "slope of the tangent line". That is, 1. "slope"; and 2. "tangent line".

1. In a (t, y) coordinate plane, the slope of a line is defined as $\frac{\text{rise}}{\text{run}}$. Can you think of what kind of line might not have a slope? *Clue:* What kind of ratio is *not* defined?
2. Can you think of what kind of shape or geometric figure might not have a tangent line?

Re. 1, Ex. 2.21. A vertical line relative to (t, y) has no slope, for as defined slope would then be of the form $\frac{\text{rise}}{\text{run}} = \frac{\text{rise}}{0}$. Consider the formula of $y^2 = t$. Relative to (t, y) axes, the graph is a horizontal parabola. Looking at the graph, what is the direction of the tangent line at (0, 0)? Now, use approximations to try calculate the derivative $y = t^{1/2}$. What is happening to the ratios $\frac{(0+\Delta t)^{1/2} - 0^{1/2}}{\Delta t}$ as Δt gets small.

In other words, while the parabola $y^2 = t$ has a perfectly good tangent line at the origin, the derivative $y = t^{1/2}$ does not exist at that point. Evidently, this is not

because there is no tangent line, but simply because relative to the (t, y) axes, the tangent line at the origin is vertical.

Re. 2, Exercise 2.21. From diagrams, notice that at every point along a smooth curve we can draw in a tangent line. Consider instead a curve that has no breaks, but is not smooth. For example, look at the point (0, 0) on the graph of the function $y = |t|$. There is a corner at (0, 0). There are many lines that touch the graph at (0, 0) but these are not tangent, at least not in both positive and negative t directions at the same time. Now try to calculate the ratios $\frac{|0+\Delta t|-|0|}{\Delta t}$. For $\Delta t > 0$, the ratio is equal to 1. For $\Delta t < 0$ we get the ratio is -1. In other words, these two values are the slopes of the two lines that make up the two parts of the graph of the absolute value function.

Summary of Exercise 2.21. In some situations, the derivative can be thought of as the slope of the tangent line. For this interpretation, however, there needs to be a tangent line; and there needs to be a slope relative to the coordinate lines. Hence, the identification can break down when the tangent line exists, but is vertical relative to the given coordinate lines or when there simply is no tangent line. Examples that do not have a tangent line are easily generated by using graphs that have corners, such as the absolute value function, or any saw-tooth graph.

We will not go further into these issues, for that would take us beyond the introductory purpose of these notes.

Example 2.14. Imagine a vigorous tropical vine that grows in such a way that each day each branch produces two new branches. Draw a picture for this. Suppose that at the end of the first day there are two branches; at the end of the second day, each of these branches has produced two more, giving a new total of 4 branches; at the end of the third day each frond from the previous day has produced two branches, so there are then 8 branches at the end of the third day; and so on. In other words, after x days, the number of fronds is given by $f(x) = 2^x$. Clearly, the increase in the number of branches per day increases rapidly and, on a daily basis, is twice the number that is present at the beginning of a given day.

As we discussed earlier, to obtain the exact rate of change at a given time x, we suppose a small difference in time Δx, and look to ratios of the form $\frac{2^{x+\Delta x}-2^x}{\Delta x}$

$= \frac{2^x\left(2^{\Delta x}-1\right)}{\Delta x} = 2^x\frac{\left(2^{\Delta x}-1\right)}{\Delta x}$. Since $2^0 = 1$, the factor $\frac{\left(2^{\Delta x}-1\right)}{\Delta x} = \frac{\left(2^{\Delta x}-2^0\right)}{\Delta x}$ is the approximate rate at which the number of fronds grows near the beginning of the growth process (starting at time $x = 0$).

Thanks to the geometric interpretation of derivative as slope of a graph, the ratios $\frac{\left(2^{\Delta x}-1\right)}{\Delta x} = \frac{\left(2^{\Delta x}-2^0\right)}{\Delta x}$ can be seen to approximate the slope of the graph of the exponential function $f(x) = 2^x$ at $x = 0$. Consequently, the quantity $\frac{2^{x+\Delta x}-2^x}{\Delta x}$

$= \frac{2^x\left(2^{\Delta x}-1\right)}{\Delta x} = 2^x\frac{\left(2^{\Delta x}-1\right)}{\Delta x}$ which approximates the slope of the graph $y = 2^x$ at x, can also be seen to approximate $2^x \cdot$ (Slop of 2^x at $x = 0$). This leads to the formula $(2^x) = 2^x \cdot$ (Slop of 2^x at $x = 0$).

We may easily develop other examples. For example, if we have another type of vine that, say, triples the number of branches by the end of each day, then arguing in the same way, we would obtain $(3^x) = 3^x \cdot$ (Slop of 3^x at $x = 0$). And so on. That is, for any exponential function, $(a^x) = a^x \cdot$ (Slop of a^x at $x = 0$).

Exercise 2.22. Graph the functions $f(x) = a^x$ for $a = -3, -2, -1, 0, 1, 2, 3$. Notice (i) All of these functions go through the point (0, 1); (ii) Each has its own slope at (0, 1); and (iii) For the negative bases, the exponential function rapidly drops off toward zero, while for the positive bases, the exponential function rapidly increases.

The basic result for growth rates of exponential functions is that the rate of change of any exponential function is proportional to the function value itself. If the function is given by powers of 2 (doubling), then the exact growth rate at time x is proportional to 2^x. If the function is given by powers of 3 (tripling), then the exact growth rate at time x is proportional to 3^x. And so on. As we saw earlier, this is quite different in behavior from functions that are powers of $y = x^n$. For instance, if $y = x^3$, then $y' = 3x^2$. Since the exponent on the variable has been reduced by 1, the derivative function $y' = 3x^2$ clearly is not proportional to the original function $y = x^3$. In other words, there is no constant K say, such that $y' = Ky$. Indeed, for $x \neq 0$, $\frac{y'}{y} = \frac{3}{x}$, which is not constant in x.

Now, the "standard" exponential function is defined by the base that produces a proportionality factor that is unity. This "natural base" is denoted by the symbol e, and so we have that $(e^x)' = e^x \cdot$ (Slope of e^x at $x = 0$) $= e^x(1) = e^x$. Since this base is "natural", the inverse of this exponential function is given its own name. It is called the *natural logarithm*, and is denoted, not $\log_e(y)$, but simply $\ln(y)$, and is pronounced "lon"(y).

Exercise 2.23. Using the rule that we developed for the derivative of an inverse function, calculate $(\ln y)$. Answer: $(\ln y)' = \frac{1}{y}$.

Notes 2.2. Leibniz followed the work of Kepler and Archimedes in his initial approach to calculating areas. He introduced the elongated "S" notation, $\int f(x)\,dx$. This was to indicate that an area under a graph could be approximated by sums of narrow columns, with height given by $f(x)$ and width dx. Using the notation of Leibniz, the Fundamental "Theorem" of Calculus from the last example was written $\frac{d}{dx}\left[\int f(x)\,dx\right] = f(x)$. James Gregory (1638 – 1675) was the first to publish a (geometric) argument for the Fundamental "Theorem" of Calculus. Isaac Barrow (1630 – 1677) also gave an argument for the result, based on moving areas and tangent lines. These arguments made use of Euclidean geometry, and consequently involved certain logical oversights. Archimedes, Kepler, Leibniz, Newton and many others all had insight into how to use increasingly improved approximations to evaluate rates and areas. In modern terminology, this type of approximation is called "evaluating the limit".

Newton's interest included rates for physics and geometry, especially those involving distance and time. In physics, the derivative of position with respect to time is *velocity*, and the rate at which the *velocity* changes is called *acceleration*. If one represents a function geometrically by its graph, then the derivative is the slope of a tangent line. For the Fundamental "Theorem" of Calculus, we put the word "Theorem" in quotes because, while the initial result was known to Leibniz, and Newton (the parents of calculus!) as well as others, there were not yet definitions of either "limit" or "area". There was, therefore, not yet a "theorem" as such. Lacking rigor and proof also meant that initial insights were both incomplete and vulnerable. It was several decades before mathematicians began to reach some initial clarification on the meanings of limit, convergence and area. Maclaurin (1698 – 1746) gave an argument for the Fundamental "Theorem" of Calculus, that was in keeping with the discussion above, based on areas associated with the graph of a function. J. D'Alembert (1717 – 1783) said that "the differentiation of equations consists simply in finding the limits of the ratios of finite differences of the two variables of the equation." [Burton, Sec. 8.4]. Note, however, that D'Alembert also did not have a definition of "limit". A. L. Cauchy (1789 – 1857) was, it seems, the first to formulate a precise and usable definition of convergence [Katz, Ch. 16], a definition that in essence has survived to the present day. It is a testimony to the genius of Archimedes that in its basic structure, his argument for the quadrature of the parabola was identical to the abstract definition reached by Cauchy. In fact, there were definitions of "limit" prior to Cauchy, obtained for example by both B. Bolzano and J.A. da Cunha (1744 – 1787). These definitions were similar in concept to Cauchy's definition. Their work however was not as crystallized as Cauchy's, and did not reach the western European mathematics community for several decades [Katz, 712]. Section 2.3 below gives examples that lead up to Cauchy's definition of limit.

2.3 SERIES, POWER SERIES AND CONVERGENCE

Recall from Section (2.1) our first topic that in approximating the area under a parabola, Archimedes had a formula for the sum of powers of 1/5:

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

Archimedes argued that since for large values of n, the term $\frac{1}{3\cdot 4^n}$ gets very small, the sum must approach a target value of $\frac{4}{3}$. See also [Bressoud, 11]. He concluded with a formula for the area under a parabola.

A sum of powers of a common ratio is called a "geometric series". These have been useful throughout the development of mathematics. An approach to geometric series that would be in keeping with modern mathematics would be to use algebraic techniques. Let's keep to a numerical case for now, a ratio of 10 say. The problem, then, is to use algebra to find a formula for a sum of powers of 10.

Question 2.4. What is the sum $1 + (10) + \cdots + (10)^n = ?$

Here are a few cases:

For $n = 1$, we get $1 + (10) = 11$

For $n = 2$, we get $1 + (10) + (10)^2 = 111$

For $n = 3$, we get $1 + (10) + (10)^2 + (10)^3 = 1111$

And so on.

Evidently, these sums get large as we add higher powers of 10. It is probably not surprising that there is no target value as there was when the ratio was $\frac{1}{4}$. The pattern, though, is interesting and we can still try to find a formula for the sum.

The algebraic approach is to give the unknown a name, and then to investigate patterns of operations. Now, simply naming the unknown (e.g., calling it S) may at first glance seem like a fairly useless thing to do. But, using this algebraic approach does not merely give the unknown a name. Doing so, we identify that the unknown S is a number. Because of that, the unknown is then connected to all other numbers through the patterns of arithmetic operations. Making this explicit gives a tremendous advantage toward identifying the number in question.

In the present case, the unknown is "S" – for "sum". This gives us the equation $S = 1 + (10) + \cdots + (10)^n$. Having this equation now reveals significant arithmetic features. The sum is constructed out of powers (multiplications) of a common

ratio, so one idea is to look at how our unknown S might be affected when multiplied by that common ratio 10. This, in fact, gives us two equations to compare:

$$S = 1 + (10) + \cdots + (10)^n$$
$$10 \bullet S = (10) + \cdots + (10)^n + (10)^{n+1}$$

The unknown can now be isolated, for by combining these two equations we get

$$(10 - 1) \bullet S = -1 + (10)^{n+1}$$

That is, $(9) \bullet S = -1 + (10)^{n+1}$; and so $S = \dfrac{(10)^{n+1} - 1}{9}$.

Exercise 2.24. Was there anything in this approach that depended on the particular value of 10? Use the same approach to find a formula for the general geometric series $S = 1 + x + \cdots + x^n$. *Answer:* $S = 1 + x + \cdots + x^n = \dfrac{x^{n+1} - 1}{(x-1)}$.

This is an algebraic expression. There may need to be restrictions on the values of x for which the expression makes sense.

Typically, restrictions on the use of symbols are needed in any language. To review how that need arises in algebra, first consider the more elementary venue of arithmetic. The fraction $\dfrac{12}{3}$ poses a question: How many 3's does it take to make 12? Of course, the solution is 4, because $4 \times 3 = 12$. What about the fraction $\dfrac{12}{0}$? This fraction would also pose a question: How many 0's does it take to make a 12? Of course, you can't make a 12 from 0's!

The difficulty here is not a problem with the rules of arithmetic, but a problem with the question. The question doesn't make sense. Traditionally, division by zero is therefore called "undefined" – again, meaning that in this case the question doesn't make sense.

For an illustration of what can happen if we don't make sure that the operations are defined, consider the equality $0 \bullet 1 = 0 \bullet 0$. If we symbolically divide by 0, we get that $0 = 1$, from which it easily follows that all numbers are equal to 1, or equivalently, all numbers are equal to 0!

Returning now to algebra, note that an algebraic expression regards many instances of arithmetic. Consequently, in algebra as well, there is always the background question, "For what values of the variable(*s*) does the expression make sense?" As an example, for what values of x does the expression $\dfrac{x^3 - 1}{x - 1}$

make sense? There is only one case where there would be a problem, that is, where the denominator would become zero. Consequently, the expression is a valid algebraic expression, as long as $x \neq 1$.

Now, let's return to our geometric series $S = 1 + x + \cdots + x^n = \dfrac{x^{n+1} - 1}{(x-1)}$. For what values of x does this expression make sense? Evidently, the expression is a valid algebraic expression, as long as $x \neq 1$. This expression then is valid for any other value of x. In particular, we can retrieve the examples that we have already discussed in this chapter:

For $x = \dfrac{1}{2}$, we get $1 + \left(\dfrac{1}{2}\right) + \cdots + \left(\dfrac{1}{2}\right)^n = \dfrac{\left(\dfrac{1}{2}\right)^{n+1} - 1}{\left(\dfrac{1}{2} - 1\right)} = \dfrac{1 - \left(\dfrac{1}{2}\right)^{n+1}}{\left(1 - \dfrac{1}{2}\right)} = 2 - \dfrac{1}{2^n}$.

For $x = \dfrac{1}{4}$, we get $1 + \left(\dfrac{1}{4}\right) + \cdots + \left(\dfrac{1}{4}\right)^n = \dfrac{\left(\dfrac{1}{4}\right)^{n+1} - 1}{\left(\dfrac{1}{4} - 1\right)} = \dfrac{1 - \left(\dfrac{1}{4}\right)^{n+1}}{\left(1 - \dfrac{1}{4}\right)} = \dfrac{4}{3} - \dfrac{1}{3 \bullet 4^n}$.

For $x = 10$, we get $1 + 10 + \cdots + 10^n = \dfrac{10^{n+1} - 1}{(10-1)} = \dfrac{10^{n+1} - 1}{9}$.

For the first two examples $x = \dfrac{1}{2}; x = \dfrac{1}{4}$. As already discussed, there are target values of 2 and $\dfrac{4}{3}$ respectively. That is, the more terms we have in each sum, the closer the sum gets to its target value, To express this insight, it would be useful to have a symbolism. The symbol traditionally used is "…".

For instance, instead of writing out the entire algebraic expression

$1 + \left(\dfrac{1}{2}\right) + \cdots + \left(\dfrac{1}{2}\right)^n = \dfrac{\left(\dfrac{1}{2}\right)^{n+1} - 1}{\left(\dfrac{1}{2} - 1\right)} = \dfrac{1 - \left(\dfrac{1}{2}\right)^{n+1}}{\left(1 - \dfrac{1}{2}\right)} = 2 - \dfrac{1}{2^n}$, and then commenting on how and why the sum approaches its target value, we may abbreviate this by writing $1 + \left(\dfrac{1}{2}\right) + \left(\dfrac{1}{2}\right)^2 + \left(\dfrac{1}{2}\right)^3 + \cdots = 2$.

The notation has lead to the name "infinite sum". Notice, though, that the symbolism does not regard any "infinite sum" as such. One may think of as many terms as one pleases. But, the essential and primary meaning of the symbolism does regard an "infinity" of terms, but expresses an insight. Just as first discovered by Archimedes, the insight is that the (finite) sums approach a target value, because the additional part (the remainder) $\frac{1}{2^n}$ in the algebraic formula $2-\frac{1}{2^n}$ gets small as we increase the number of terms n.

In this context, we now have a new symbol, "⋯". Just as for fractions in arithmetic, rational expressions in algebra, and symbols in any language, we can enquire into valid use of the symbol "⋯". When does use of this symbol make sense?

For geometric series, we can answer this fairly directly, for we already have an explicit expression for the finite geometric sums: $S=1+x+\cdots+x^n=\frac{x^{n+1}-1}{(x-1)}$.

As already discussed, for the rational part on the right hand side to make sense, we need only require that $x \neq 1$. Then, to use the symbol "⋯" requires that the quantity approaches a target value. Evidently, if $|x|<1$, both criteria are met and the target value is $\frac{1}{1-x}$. To summarize these results, we can write the abbreviated expression $1+x+x^2+\cdots=\frac{1}{(1-x)}$, for $|x|<1$.

Exercise 2.25. Are there values of x for which the finite sums $S=1+(5x)+\cdots+(5x)^n=\frac{(5x)^{n+1}-1}{(5x-1)}$ approach a target value? *Clue:* Write the sum in terms of $y=5x$. Answer: We can write $S=1+(5x)+(5x)^2+\cdots$, for $|y|<1$ or equivalently for $|x|<\frac{1}{5}$.

Example 2.15. Where Archimedes enquired into the area between a parabola and a line, Gregory of St. Vincent (1584 – 1667) enquired into areas between an hyperbola such as $y=\frac{1}{x}$ and the line given by the x-axis [Katz, Ch. 12].

Before looking at Gregory's results, it may help to recall a fact about the areas of rectangles. Suppose a rectangle of height 2 and length 1, hence of area 2. If we reduce the height by a factor of $\frac{1}{5}$, but increase the length by a factor of 5, then the new rectangle will of course have the same area, namely, $\frac{2}{5}\times\frac{5}{1}$.

Now, to estimate the area under the hyperbola $y=\frac{1}{x}$, we can easily produce a lower estimate by using rectangles of width 1, and heights determined by the hyperbola itself. See Fig. 2.14.

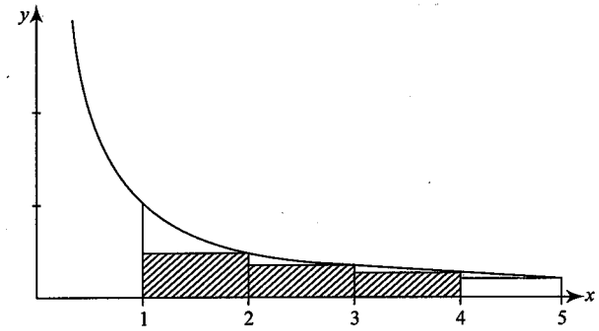


Figure 2.14

If, for example, we wish to look at the rectangular area under the hyperbola with length determined by the end points 1 and 4, then a lower estimate to the area is given by the sum of rectangular areas $\frac{1}{2}+\frac{1}{3}+\frac{1}{4}$. (Note that likewise, an upper bound is given by $1+\frac{1}{2}+\frac{1}{3}$.)

As we just observed, we can form new rectangles of the same area, as long as the height and length keep the same ratio. But, the formula for the hyperbola does just that, that is, it reduces height inversely to any increase in length. See Fig. 2.15. Start with the rectangle whose end points are $x = 1$ and $x = 2$, with height $\frac{1}{2}$ determined by the right end point $x = 2$. If we look now to a new rectangle obtained by doubling the coordinates, and using the hyperbola for getting the new height, we get the rectangle $x = 2$ and $x = 4$ and height $\frac{1}{4}$. This can be repeated. For the next case, we get a rectangle with end points $x = 4$ and $x = 8$ and height $\frac{1}{8}$. And so on.

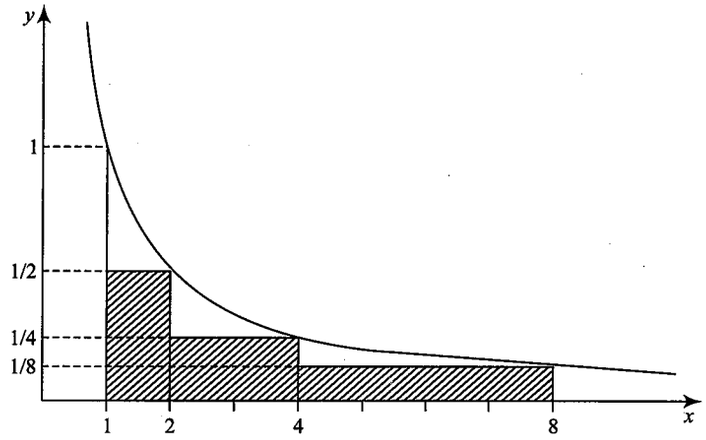


Figure 2.15

Exercise 2.26. Evaluate what happens if the initial rectangle has end points $x = 1$ and $x = 4$, with height $\frac{1}{4}$, and instead of doubling, we use a factor of 5 say. The end points are $x = 5$ and $x = 20$ and the new height is $\frac{1}{20}$. Evidently, the area is again preserved.

For the general case, suppose we start with a rectangle with end points $[a, b]$ and height $\frac{1}{b}$. The area is $(b-a)\frac{1}{b} = \frac{b-a}{b}$. Now, multiply the end points by any common factor $\alpha > 0$ to obtain the new end points $[\alpha a, \alpha b]$. Using the hyperbola over the right end point, the new height is $\frac{1}{\alpha b}$. So, we get a rectangle of length $\alpha(b-a)$ and height $\frac{1}{\alpha b}$, hence area $\frac{b-a}{b}$, which is equal to the original area.

Gregory of Vincent realized that this must also be true not only for rectangles sitting under the hyperbola, but for any area under the hyperbola [Katz, Ch. 12]. That is, he showed that given end points $[a, b]$ and positive number α, the rectangular area under the hyperbola for $[a, b]$ is equal to the rectangular area under the hyperbola for $[\alpha a, \alpha b]$. To establish that result, he used what was by then a classical approach of approximation. He partitioned the interval into many subintervals, and approximated using rectangles with height determined by the formula $y = \frac{1}{x}$. His argument then showed that since the ratio property holds for

each of the many thin rectangles, it holds for their approximating sum, and therefore the property also holds for the actual hyperbolic area.

Reading the work of Gregory of St. Vincent, A.A. de Sarasa (1618 – 1667) realized a connection to the logarithm functions [Katz, Ch. 12]. To see this, let's start with an example. Let $A(1, 15)$ denote the area under the hyperbola, between the end points $x = 1$, $x = 15$. See Fig. 2.16.

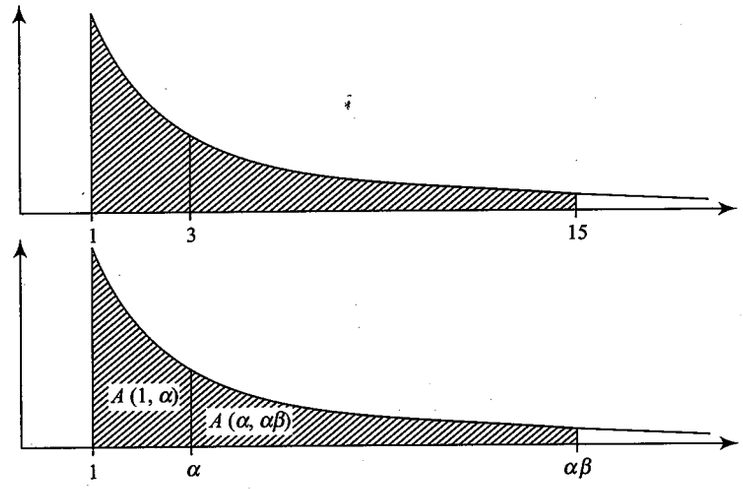


Figure 2.16

From basic geometry, $A(1, 15) = A(1, 3) + A(3, 15)$. But the end points of the second interval have a common factor of "3" and so the interval can written as $(3, 15) = (3(1), 3(5))$. We can therefore use the result of Gregory of St. Vincent. That is, the second term can be reduced to $A(3, 15) = A(1, 5)$. Putting this back into the equation, we get $A(1, 3 \cdot 5) = A(1, 3) + A(1, 5)$. In other words, for a product, we can pull back the area to a sum of two areas, both starting with the same left end point $x = 1$.

Exercise 2.27. Using the result from Gregory of St. Vincent, show that for any positive α, β, $A(1, \beta) = A(\alpha \cdot 1, \alpha \cdot \beta) = A(\alpha, \alpha\beta)$; from which it follows that $A(1, \alpha\beta) = A(1, \alpha) + A(\alpha, \alpha\beta) = A(1, \alpha) + A(1, \beta)$. Or, streamlining the notation somewhat $A(\alpha\beta) = A(1, \alpha) + A(1, \beta)$. In other words, the area under the hyperbola behaves just like a logarithm function! [Recall that logarithms (exponents) add under multiplication: For any base B, $B^x B^y = B^{x+y}$]

Having read the work of *A.A.* de Sarasa that related area under a hyperbola to a logarithm, N. Mercator (1620 – 1687) used some results from J. Wallis (1616 –

1703) on ratios of "infinite sums", and obtained a formula for a shifted logarithm $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$ [Katz, pp. 491 – 492].

In fact, Newton derived the same formula, by first obtaining a series for $\frac{1}{1+x}$.

Observe that by long division, $\frac{1}{1+x}=1-x+x^2-\frac{x^3}{1+x}$; and by repeating long division, this pattern can easily be continued. For example, in exactly the same way, we can obtain $\frac{1}{1+x}=1-x+x^2-x^3+\frac{x^4}{1+x}$. And so on. Or, one can observe that $\frac{1}{1+x}=\frac{1}{1-(-x)}$. Using the geometric series, we get that $1+(-x)+(-x)^2+\cdots+(-x)^n=\frac{1-(-x)^{n+1}}{1-(-x)}=\frac{1-(-x)^{n+1}}{1+x}$.

In other words, $1-x+x^2-x^3\cdots+(-x)^n=\frac{1-(-x)^{n+1}}{1+x}$. In fact, there is also Newton's general binomial theorem (a summation formula for $(a+b)^\alpha$, α not necessarily an integer). For $\alpha=-1$, his formula produces $\frac{1}{1+x}=(1+x)^{-1}=1-x+x^2-x^3+\cdots$.

Newton made use of the result that the area under the hyperbola $y=\frac{1}{1+x}$ is $\log(1+x)$. But, he also had a version of the Fundamental Theorem of Calculus, and so he could interpret the area under the hyperbola as an anti-derivative. He therefore integrated the summation term by term and produced the equation $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$. He used this formula "to calculate the logarithms of many small positive integers" [Katz, 508].

Example 2.16. Consider the expression $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$. If $x=1$, the left side $\log(1+1)=\log(2)$ is defined. For the right hand side, $1-\frac{1^2}{2}+\frac{1^3}{3}-\cdots=1-\frac{1}{2}+\frac{1}{3}-\cdots$. Does this series converge to some target value? See Fig. 2.17.

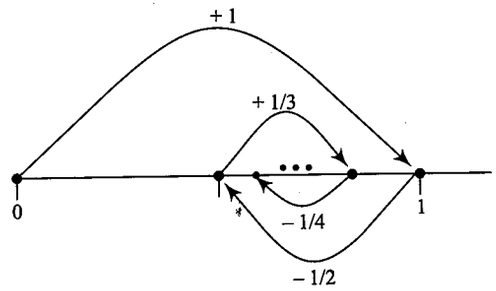


Figure 2.17

The sums $1-\frac{1}{2}+\frac{1}{3}-\cdots+(-1)^{n+1}\frac{1}{n}$ are evidently trapped inside increasingly narrow segments of the unit interval. It would seem, therefore, that there must be a target value, greater than $\frac{1}{2}$ and less than 1. This is not a proof, but it is at least a good clue. A proof would require a definition of convergence, and something equivalent to completeness of the real numbers. The calculations, though, can be directed by the (casual) geometric insight. Supposing now that for $x=1$ both sides of the equation make sense, there is still the question of whether or not they represent equal quantities: Is it true that $\log(2)=1-\frac{1^2}{2}+\frac{1^3}{3}-\cdots$? We will look to this question again later in this chapter.

Example 2.17. The previous example shows that for the formula $x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots$, setting $x=1$ produces a possible target value, something greater than $\frac{1}{2}$ and less than 1. It is also interesting to look closely at the other extreme case, $x=-1$. Of course, the left hand side of the equation is simply not defined for $x=-1$, because there is no logarithm (exponent) that produces 0. The right hand side though implicitly refers to finite sums of the form $(-1)-\frac{(-1)^2}{2}+\frac{(-1)^3}{3}-\cdots+\frac{(-1)^n}{n}=-\left[1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}\right]$. So, the existence of a target value for the right hand side of the equation will depend on the existence of a target value for sums of the form $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$.

While not a proof, one way to see that this cannot approach any target value is use the strategy of Gregory of St. Vincent. That is, assume that the area under the hyperbola $y=\frac{1}{x}$ is a logarithm for some base $B>1$. See Fig. 2.18.

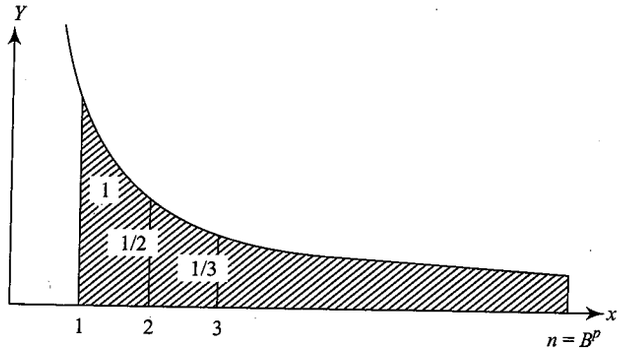


Figure 2.18

Let $A(1, n)$ be the area under the hyperbola from $x = 1$ to $x = n$ with $n = 2^p$.

Then $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n-1} > A(1,n) = \log 2^p = p\log 2$. As n increases, so does the exponent p that is needed to produce $n = 2^p$. The result follows.

Another approach uses an elementary fact already mentioned. See Figure 2.19. Consider the rectangle with end left point $x = 1$, and height determined by the right end point $x = 2$. The area of the rectangle is $\frac{1}{2}\times 1=\frac{1}{2}$. But, if we reduce the height by a factor of $\frac{1}{2}$, and increase its width by a factor of *2*, we get a new

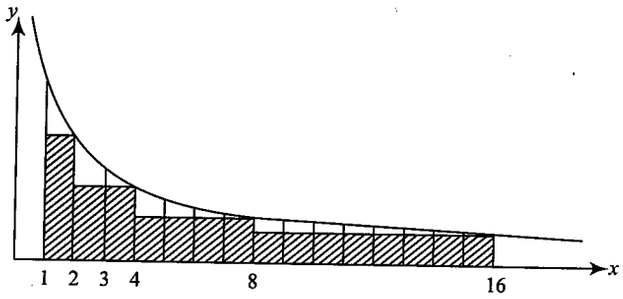


Figure 2.19

rectangle also of area $\frac{1}{2}$. This new rectangle though can be represented within the graph of the hyperbola. It will have left end point $x = 2$ and right end point $x = 4$;

and by looking also at rectangles of unit width in the diagram, it is evident that $\frac{1}{3}+\frac{1}{4}>\frac{1}{2}$. This can be repeated. Shift the rectangle with left end point $x = 2$ and right end point $x = 4$ along to a rectangle with left end point $x = 4$ and right end point $x = 8$, of height $\frac{1}{4}$. This rectangular area sits under the lower rectangles of the hyperbola, and we therefore get $\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}>\frac{1}{2}$. And so on. To be careful about this we need only formulate the result algebraically. We can partition the sum

$$\frac{1}{2^k}+\frac{1}{2^k+1}+\cdots+\frac{1}{2^{k+1}} > \frac{1}{2^{k+1}}+\frac{1}{2^{k+1}}+\cdots+\frac{1}{2^{k+1}} = \frac{2^k}{2^{k+1}} = \frac{1}{2}.$$

So, by using powers of 2 to partition the sum, we can choose large values of the form $n = 2^p$, so that the sum $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$ is larger than accumulating multiples of $\frac{1}{2}$. Multiples of $\frac{1}{2}$ grow without bound, and consequently there can be no target value for sums of the form $1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}$.

By the end of the 17th century, many summation formulas were being discovered, through techniques based on combinations of geometry, algebra, differentiation and integration.

Other examples are: $\arcsin B = B+\frac{B^3}{6R^2}+\frac{B^5}{40R^4}+\frac{B^7}{112R^6}+\cdots$, which in modern terminology (circle radius $R = 1$) gives $\operatorname{arcsin} x = x+\frac{x^3}{6}+\frac{x^5}{40}+\frac{x^7}{112}+\cdots$. There were also sums for $\arctan x$ and $\tan x$. It is also possible that James Gregory (1625 – 1683) was aware of a formula that later was named in honor Brook Taylor (1685 – 1731), namely, the "Taylor series expansion"

$$f(x) = f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\frac{f'''(0)}{3!}x^3+\cdots$$

(published in Taylor's book *The Method of Increments*, 1715.)

Newton independently discovered many summation formulas. It is interesting to note that summations for $\arctan x$, $\sin x$ and $\cos x$ also appeared in Indian works as early as the 14th century [Katz, 494].

Early derivations of series formulas made casual use of the symbol "…" (or some equivalent). In order to identify the extent to which the results were legitimate, one would of course need to investigate those values of x for which the equations and the summations make sense, both algebraically and as short hand when there is a target value. In other words, as is implicit in the approach taken by Archimedes, one would need to determine not only those values of x for which both sides of an equation are defined, but also those values of x for which, as we add more terms to a summation, a target value emerges because the remainder term gets small. Note a further subtly: Newton integrated (and differentiated) series term by term, and in doing so assumed not only the validity of the symbol "…" in the equations that he started with, but that the result on the summation side of an integrated (or differentiated) equation was the same as the result on the algebraic side of an integrated (or differentiated) equation. These issues were not well understood, and free use of the symbolism frequently led to erroneous results.

For example, if we use the geometric series $1+x+x^2+\cdots=\frac{1}{(1-x)}$, and do not restrict x, we can get such things as $1+2+4+8\cdots=\frac{1}{(1-2)}=-1$. Of course this doesn't make any sense. To account for the difficulty, note that in the present case, the full expression that includes the remainder term is $1+2+4+8\cdots+2^n = \frac{2^{n+1}-1}{2-1}$. Clearly, the additional term 2^{n+1} does not decrease in magnitude as the exponent n increases. Consequently, in this geometric series, we cannot make valid use of the symbol "…".

This brings us finally to a key question in the development of calculus.

Question 2.5. How do we define *convergence*?

As Cauchy did [Bressoud, 19], let's take the lead from Archimedes, who always worked with finite sums. The written work of Archimedes though can be difficult to read, because every quantity is represented as some geometric distance, denoted by a pair of points on the plane. This makes for rather complex looking pages of mathematics. (See, for example, [Heath].) We have the modern advantage of having efficient symbolism.

Consider then the geometric series $1+\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)^2\cdots+\left(\frac{1}{2}\right)^n=\frac{\left(\frac{1}{2}\right)^{n+1}-1}{\left(\left(\frac{1}{2}\right)-1\right)}$.

As may now be familiar to the reader, this reduces to

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

Observe that for each n, the sum is less than 2. A further insight is that the short fall to 2 decreases as we add more terms (as n increases). This is the preliminary and basic insight. This insight is not a proof, it is not a definition, and does not regard some "infinity" of terms. Instead, we catch on to a pattern in the sums, and that pattern is that as we add more terms, the sums get closer and closer to 2.

Can we be more precise? The short fall (or remainder) is the quantity $\frac{1}{2^n}$. This quantity obviously gets small, that is, gets close to zero. But, how close? Can we be more precise in our meaning of "close to zero"?

If, for example, we have a sequence of terms $.01+1, .01+\frac{1}{2}, .01+\frac{1}{4}, .01+\frac{1}{8}, \ldots$, then these terms also get increasingly close to zero. But, there is certainly a difference between the way in which the sequence of terms $.1+1, .1+\frac{1}{2}, .1+\frac{1}{4}, .1+\frac{1}{8}, \ldots$ gets closer to zero, as opposed to the way in which the sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ gets closer to zero.

The sequence $.01+1, .01+\frac{1}{2}, .01+\frac{1}{4}, .01+\frac{1}{8}, \ldots$ never gets below $0.01 > 0$! It can be said in fact that the sequence is "bounded below", by $0.01 > 0$. On the other hand, the sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^n}$ would seem to be able to eventually by-pass $0.01 > 0$. Does it? Is there some member of the sequence $1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^n}$ that is smaller than 0.01?

Let's see. Can we find an n such that $\frac{1}{2^n} < 0.01$? To find a solution, we need an exponent n. Solving for n, we get $-n\log 2 < \log 0.01$, and so $n > \frac{\log 0.01}{-\log 2}$. But, this fraction is approximately equal to 3.33. So, as long as $n \geq 4$, we will have $\frac{1}{2^n} < 0.01$.

We can now return to our geometric series $1+\left(\frac{1}{2}\right)+\left(\frac{1}{2}\right)^2\cdots+\left(\frac{1}{2}\right)^n=2-\frac{1}{2^n}$.

As long as $n \geq 4$, the short fall to 2 will necessarily be less than 0.01. Now, thinking of the approach we just took, was there anything special about $0.01 > 0$? Would the same type of argument be possible to show that $\frac{1}{2^n}$ eventually by-passes, say, 0.000001? Or other small numbers?

Exercise 2.27. Use the same approach as above to find out how large n needs to be so that the terms $\frac{1}{2^n}$ eventually by-pass, say, 0.000001.

A key then is to be able to accurately quantify the meaning of "gets close". If something is said to get close to zero, then we can ask, "How close?" Cauchy's words are "less than any assignable" number [Katz, 712]. In the above calculations, we used the assigned number 0.01; and there was also the exercise to try the same thing with the assigned number 0.000001.

Strategic symbols can often better express an idea than wordy phrases. Instead of saying "any assignable number", let's give an "assigned number" a name. Following tradition, let's call this assigned number $\varepsilon > 0$. (In some applications, one thinks of "ε" as "error" from the target value.) To say that the terms $\frac{1}{2^n}$ get closer and closer to 0 (or that the terms "converge" to 0) we need to be able to show that given any assignable number $\varepsilon > 0$, the terms $\frac{1}{2^n}$ eventually by-pass $\varepsilon > 0$ and get even smaller.

For the sequence $\frac{1}{2^n}$, can this always be done? Suppose that $\varepsilon > 0$ is some small positive number, smaller than 1, so $1 > \varepsilon > 0$ Can we show that the terms $\frac{1}{2^n}$ eventually get smaller than $\varepsilon > 0$. In other words, in the present case, can we find n that solves the inequality $\frac{1}{2^n} < \varepsilon$?

Solving for the exponent, $-n\log 2 < \log\varepsilon$, and so $n > \frac{\log\varepsilon}{-\log 2}$. Remember that we are supposing that $\varepsilon > 0$ is a fixed small positive number $1 > \varepsilon > 0$. So, while the right hand side $\frac{\log\varepsilon}{-\log 2}$ may not be an integer, it is a ratio of two negative numbers, and so is some positive number. Now, our sequence $\frac{1}{2^n}$ is easily shown to be a decreasing sequence, that is, $\frac{1}{2^n} > \frac{1}{2^{n+1}} > \frac{1}{2^{n+2}} > \cdots$ Therefore, as soon as we choose some exponent $n > \frac{\log\varepsilon}{-\log 2}$, the rest of the terms of the sequence will also be smaller than $\frac{1}{2^n} < \varepsilon$.

Abstracting the essentials of this argument leads to the following definition:

Definition of a Remainder Converging to Zero: A sequence of positive remainders $r_1, r_2, r_3, \ldots$ converges to 0, if given any assignable number $\varepsilon > 0$, the terms eventually get closer to 0 than $\varepsilon > 0$. In other words, given any $\varepsilon > 0$, it is always possible to find n_0 so that for $n \geq n_0$, $0 \leq r_n < \varepsilon$.

Exercise 2.28. Extend this to define convergence of a sequence $a_1, a_2, a_3, \ldots$ of (not necessarily positive) numbers to a target value (also called *limit*) a. *Clue:* Consider the sequence $|r_n| = |a - a_n| \geq 0$, determined by the absolute value of each remainder.

Exercise 2.29. Use the definition to show that the sums $1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots+\frac{1}{4^n}=\left(\frac{4}{3}\right)-\left(\frac{1}{4^n}\right)\left(\frac{1}{3}\right)$ converge to $\frac{4}{3}$.

Exercise 2.30. (Integral Test) Consider the sums $S_n = 1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots+\frac{1}{n^2}$. Do these sums converge to a target value?

Recall the sums of the form $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}$ and $\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}$ can be obtained as upper and lower bounds for the area under a hyperbola $y=\frac{1}{x}$. We were able to show that these sums do not converge, but rather produce increasingly large positive numbers.

We gathered insight into this problem by relating the sums to areas under a known function $y=\frac{1}{x}$. We then used information about the function $y=\frac{1}{x}$ to deduce features of the sums.

Can something similar work for the sums $S_n=1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots+\frac{1}{n^2}$?

Clue: Use the function $y=\frac{1}{x^2}$. Look at rectangular areas under the graph of $y=\frac{1}{x^2}$, determined by the intervals [1, 2], [2, 3], [3, 4], ..., [1, x]. Recall that if $A(1, x)$ is the area under the graph of the function $y=\frac{1}{x^2}$ over the interval [1, x], then by the Fundamental Theorem of Calculus, the derivative of the area function is $\frac{d}{dx}A(1,x)=\frac{1}{x^2}=x^{-2}$. What then must the area function itself be, as a function of x? If we let n get very large, what happens to the areas $A(1, n)$? Does this sequence converge? Relate this result to the sequence of sums $S_n=1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\cdots+\frac{1}{n^2}$. Note that the presence of the first summand "1" of course does not affect possible convergence.

Exercise 2.31. Think about how the approach of the last example can be generalized to other functions. In particular, suppose that $y = f(x) \geq 0$, and $f(i) = a_i$, for $i = 1, 2, 3, \ldots$, and $S_n = a_1 + a_2 + a_3 + \cdots + a_n$ with $a_i \geq 0$. Based on the last example, might it be possible to establish a relationship between the sums S_n and integrals of the function $y = f(x) \geq 0$?

Some conditions on the function would though be required. For instance, suppose that $y = g(x) \geq 0$ is defined by $g(x) = \begin{cases} 0, n-\frac{1}{4}<x<n+\frac{1}{4} \\ 1, n+\frac{1}{4}\leq x\leq n+\frac{3}{4} \end{cases}$. In this case, $\sum_{i=1}^{i=n} f(i)=0$, but the integral areas (total areas) under the function clearly do not converge.

The Integral Test therefore requires that the function $y = f(x) \geq 0$ be monotone decreasing on the interval $1 \leq x < \infty$. In that case, it can be shown (Exercise for the reader) that the sequence $\int_1^n f(x)\,dx$ and the series $S_n = a_1 + a_2 + a_3 + \cdots + a_n$ either both converge or both diverge.

Example 2.18. (Comparison Test) Consider the sums of the form $S_n=1+\frac{1}{1\cdot 2}+\frac{1}{2\cdot 4}+\frac{1}{3\cdot 8}+\cdots+\frac{1}{n\cdot 2^n}$. Do these sums have a target value?

As a first observation, notice that from one sum to the next we add another positive term. Hence, the sequence of sums S_n is increasing, that is, for each n, $S_n < S_{n+1}$.

Now, do you also see some quantities involved that are familiar? There is the geometric series $G_n=1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^n}=2-\frac{1}{2^n}$. For each n we have $S_n \leq G_n$, since $1\leq 1; \frac{1}{2}\leq\frac{1}{1\cdot 2}; \frac{1}{2\cdot 4}\leq\frac{1}{4}; \frac{1}{3\cdot 8}\leq\frac{1}{8}; \ldots; \frac{1}{n\cdot 2^n}\leq\frac{1}{2^n}$.

Moreover, as we have already determined, the sums G_n increase and converge toward their target value of 2.

We can line up what have so far this way: $0 \leq S_n \leq G_n \leq 2$. See also Fig. 2.20.

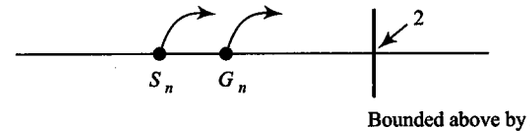


Figure 2.20

The sums S_n are therefore both increasing and bounded above. While the sums increase, they are also trapped. They must therefore, start "crowding", or accumulating. So it makes sense to conjecture that there must be some ceiling value, or target value. In other words, it makes sense to conjecture that the sequence must therefore converge to some target value $T \leq 2$.

This conclusion does make "sense", but we need to be careful here, so that we don't make the same kind of oversight that early geometers made, and Cauchy himself made. One of the oversights in pre-twentieth century Euclidean geometry was to confuse insight into image (grasp of possibility) with mathematical necessity.

There was, for example, the assumption that a line joining a point inside a circle to a point outside a circle necessarily intersects the circle. Certainly, that can be grasped as a possibility, by insight into a diagram. But, as was discovered in the 19th century, there are geometries that do not have this kind of intersection property. The need for some kind of "betweeness axiom" was then confirmed.

We need to be careful in a similar way with sequences of real numbers.

In our case, we have an increasing bounded sequence S_n, of what in fact are fractions. Asserting the existence of a target value is asserting the existence of a real number T that may or may not a fraction, and that is between all of the fractions S_n and 2, that is, $S_n \le T \le 2$. In other words, we would need that the real numbers are sufficiently "complete" in order for such a "between value" to necessarily exist. Note that, in this context, one can also enquire into the existence of a target value for a sequence that is not necessarily increasing and bounded above, but is at least bounded both above and below. For example, consider the sequence $S_n = 2 + \frac{(-1)^n}{n}$.

To go into these matters further is for a course in advanced calculus or analysis, where one would investigate the axiomatic development of the real numbers. The present modest purpose is twofold: (i) To see that having one increasing sequence bounded above by another convergent sequence can lead to the preliminary idea that the lower sequence must also be convergent; and (ii) To be aware that the validity of this result depends on properties of the real numbers that would need further study.

Exercise 2.32. Each of the sums $S_n = 1 + \frac{1}{1\cdot 2} + \frac{1}{2\cdot 4} + \frac{1}{3\cdot 8} + \cdots + \frac{1}{n\cdot 2^n} \le 2$

That is, 2 is a "ceiling" or an "upper bound" for all sums of the form S_n.

Is there a "lower" ceiling, perhaps some "smaller upper bound"? Notice that

$$\begin{aligned} S_n &= 1 + \frac{1}{1\cdot 2} + \frac{1}{2\cdot 4} + \frac{1}{3\cdot 8} + \cdots + \frac{1}{n\cdot 2^n} \\ &\le 1 + \frac{1}{1\cdot 2} + \frac{1}{2\cdot 4} + \left[\frac{1}{3\cdot 8} + \frac{1}{3\cdot 16} \cdots + \frac{1}{3\cdot 2^n}\right] \\ &= \frac{13}{8} + \frac{1}{3\cdot 8}\left[1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n}\right] \\ &\le \frac{13}{8} + \frac{1}{3\cdot 8}[2] \end{aligned}$$

$$\begin{aligned} &= \frac{41}{24} \\ &< 2 \end{aligned}$$

So, there is a "ceiling", an upper bound for the sums S_n that is strictly less than 2. What then could the target value be? Could it be 2? Could it be $\frac{41}{24}$? Might it be less than $\frac{41}{24}$? What about a "lowest possible ceiling", or as it is now called in the textbooks and the literature, a "least upper bound"?

Exercise 2.33. Is a least upper bound unique? Use the approach of Archimedes to see that the *least upper bound* must be the target value. Again, notice that we need a completeness axiom.

Example 2.19. (Cauchy Sequences) Consider two sequences of sums, one harmonic and the other obtained from squares: As we have already discussed, the sequence of harmonic sums $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$ diverges; and the sequence of sums of squares $S_n = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots + \frac{1}{n^2}$ converges.

Use the definition of convergence to show that since the sums of reciprocal squares converge to a target value (integral test), the last terms of the sums $\frac{1}{n^2}$ must converge to zero. While this is necessary, it is not in general sufficient. Take for example, the harmonic series $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$. The last terms $\frac{1}{n}$ certainly converge to zero, but the series does not converge. To characterize convergence of a series, we therefore must look to something more than just the behavior of the last terms.

In the effort to find a usable test for convergence of a series, Cauchy was led to formulate a stronger criterion that included not only the behavior of the last terms, but arbitrary strings of last terms (the "tails" of the series). He asserted that a sequence (of finite sums) is convergent (according to his definition of convergence) if for each k and arbitrary n, the differences $S_{n+k} - S_n$ converge to zero. Later, any sequence with this property was called a *Cauchy sequence.*

Exercise 2.34. Use the definition of convergence to show that a Cauchy sequence must be a convergent sequence. The converse is also true, ..., if we assume completeness of the real numbers!

Exercise 2.35. (Absolute Convergence) Consider the sums $S_n = 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots + (-1)^{n+1} \frac{1}{n^2}$. Does this converge to a target value? *Clues:* Is it a Cauchy sequence? Recall that

$$|a_1 \pm a_2| \leq |a_1| + |a_2|$$

$$|a_1 \pm a_2 \pm a_3| \leq |a_1| + |a_2| + |a_3|$$

ETC !

For another clue: Is $S_n = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots + \frac{1}{n^2}$ a Cauchy sequence?

Exercise 2.36. (Absolute Convergence of a Series of Non-Negative Terms Implies Convergence of the Alternating Series) Suppose that $S_n = a_1 + a_2 + a_3 + \cdots + a_n$. $a_1 \leq 0$. Let $A_n = a_1 - a_2 + a_3 - \cdots + (-1)^{n+1} a_n$ be a sequence of sums obtained by alternating the signs of the summands a_i. If S_n is a convergent sequence of sums, does it follow that $A_n = a_1 - a_2 + a_3 - \cdots + (-1)^{n+1} a_n$ also is a convergent sequence of sums? *Clue:* See the previous Exercise. Is $A_n = a_1 - a_2 + a_3 - \cdots + (-1)^{n+1} a_n$ a Cauchy sequence? Is $S_n = a_1 + a_2 + a_3 + \cdots + a_n$ a Cauchy sequence? See also the clues given in the previous exercise.

Exercise 2.37. (A Non-Negative Sequence and its Alternating Series) Earlier in this Chapter 2 we discussed a series expansion for log $(1 + x)$ and we asked whether or not the sums $1 - \frac{(1)^2}{2} + \frac{(1)^3}{3} - \cdots + (-1)^n \frac{(1)^n}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \cdots (-1)^n \frac{1}{n}$ converge. Is this a Cauchy sequence? Can you generalize this result to an *alternating series*? That is, suppose that $S_n = a_1 - a_2 + a_3 - \cdots + (-1)^{n+1} a_n$ where the sequence $ai \geq 0$ is monotone decreasing and converges to zero.

Exercise 2.38. Think about and take note of the differences between the results of the last two Exercises, both of which regard alternating series.

Example 2.20. (Root Test) Consider the sequence of sums defined by $S_n = \sum_{k=1}^{n} \frac{\sin(2k)}{2^k}$. Is this a convergent sequence of sums? Note that $\left| \frac{\sin(2k)}{2^k} \right| \leq \frac{1}{2^k}$. Do you see that there is therefore a geometric series that dominates the sums? Indeed, we get that $S_n = \sum_{k=1}^{n} \left| \frac{\sin(2k)}{2^k} \right| \leq \sum_{k=1}^{n} \left(\frac{1}{2} \right)^k < 1$. The absolute values therefore converge. It follows that the original series also converges. The general statement

of this approach is called the Root Test. In other words, if $|a_k|^{\frac{1}{k}} \leq r < 1$, then $S_n = \sum_{k=1}^{n} |a_k| \leq \sum_{k=1}^{n} r^k$ is a convergent series. Convergence of a series depends only on the behavior of the tail of the series. There is, therefore, a limit form of the Root Test. See, for example, [Bressoud] 130 – 135.

Example 2.21. (Ratio Test) For the cosine function, the Taylor series is given by the sums $T_n(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!}$. Are these sums convergent? The summands are given by increasing powers of x. However, if we try to compare this directly to some geometric series, evaluating the n^{th} root of a summand leads to the challenging calculation $\left((-1)^n \frac{x^{2n}}{(2n)!} \right)^{\frac{1}{n}} = -\frac{x^2}{((2n)!)^{\frac{1}{n}}}$. How can we estimate $((2n)!)^{\frac{1}{n}}$?

It is a modern and algebraic approach to express a geometric series as a sum of powers. Instead, one may recall the geometric origin of the *geometric* series. In other words, the series is determined by a common geometric ratio (of lengths). This of course is quantitatively equivalent, but includes an additional focus, and gives a different expression. As it happens, it also gives us another way to analyze series. So, toward being able to compare the Taylor series for the cosine function to some geometric series, let's consider the ratios of successive summands. We get

$$\frac{a_{n+1}}{a_n} = \frac{(-1)^{n+1} \frac{x^{2(n+1)}}{(2(n+1))!}}{(-1)^n \frac{x^{2n}}{(2n)!}} = \frac{-x^2}{(2n+1)(2n)}.$$

Clearly, the absolute value of this ratio converges to zero (rapidly, for any x), and so we can easily compare the series to a convergent geometric series. For example, we can compare to the geometric series with common ratio $r = \frac{1}{2}$, by finding the values of n for which $\frac{a_{n+1}}{a_n} = \frac{x^2}{(2n+1)(2n)} < \frac{1}{2}$. One could solve this inequality directly, by using a quadratic equation for n. That would be quite exact. However, the comparison $\frac{a_{n+1}}{a_n} = \frac{x^2}{(2n+1)(2n)} < \frac{1}{2}$ is already only an estimate,

and there is no need for an exact solution at this stage. A common strategy is to solve such an inequality on the conservative side. For example, observe that $\frac{x^2}{(2n+1)(2n)} < \frac{x^2}{(2n)(2n)} < \frac{x^2}{n}$. So, if we can solve $\frac{x^2}{n} < \frac{1}{2}$, then that would be enough to guarantee that $\frac{a_{n+1}}{a_n} = \frac{x^2}{(2n+1)(2n)} < \frac{1}{2}$. A solution then is $n > 2x^2$.

As with the Root Test, there is also a limit form of the Ratio Test. See, for example, [Bressoud] 130–135.

Example 2.22. (Cauchy's Definition of Integral)

Let's briefly review some of what we have so far. Going back to the work of Archimedes and other early mathematicians, the areas of parabolic and other non-rectangular areas were calculated using approximations by constructions and sums of more familiar rectangles and triangles. This basic approach has been used up to the early days of calculus and on to modern times. Intrinsic to this approach is a basic insight grounding the notion of target value or limiting value. When applied to ratios this key insight leads to an initial understanding of "exact ratio", later called "derivative". A special case of that understanding is obtained when we seek the rate of change of a moving area. This led to what later came to be called the Fundamental Theorem of Calculus.

However, as we have mentioned in the notes above, in its initial 17th century formulations, the Fundamental Theorem of Calculus was not rigorously established. In fact, there were main parts of the result that were not yet defined. There was not yet a definition of convergence; there was no completeness axiom stated (although some such axiom is needed to ensure the existence of limits); and there was also no definition of area. Filling these gaps was essential, for otherwise the Fundamental Theorem of Calculus was a statement about a not-necessarily existent undefined derivative of a not-necessarily existent undefined integral.

Cauchy first solved the problem of defining convergence. It was then possible to turn attention back to the integral. Cauchy used his new understanding of convergence to lift the basic insight that grounds the approximation of areas by rectangles into the context of explanatory definition. Indeed, "Cauchy made the approximation into a definition" [Katz, 718]. Note, however, that Cauchy did not recognize the need for a completeness axiom. An axiomatic formulation of the real numbers was to be investigated later in the 19th century. See, for example, the work of Richard Dedekind (1831 to 1916).

Suppose that $f(x)$ is continuous on an interval $[x_0, X]$. Partition the interval into $n-1$ subintervals determined by the intermediate values $x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = X$. When it exists, the Cauchy integral can then be defined as the target value or

limit of sums of the form $f(x_0)(x_1 - x_0) + f(x_1)(x_2 - x_2) + \cdots + f(x_{n-1})(x_n - x_{n-1})$ [Katz, 718 – 719].

The completeness axiom aside, using this definition Cauchy could then easily prove the first rigorous version of the ***Fundamental Theorem of Calculus:***

Suppose that $f(x)$ is continuous on an interval $[x_0, X]$. Let $F(x) = \int_{x_0}^{X} f(x)\,dx$.

Then the derivative is given by $F'(x) = f(x)$.

Remark 2.1. In fact, later, G.P. Lejeune-Dirichlet (1805 – 1859) showed that the Cauchy definition of integral "only guaranteed the existence of the integral for (certain) functions with finitely many discontinuities" [Katz, 725]. Problems posed by Dirichlet and Riemann in the 19th century were with regard to emergent subtleties that ultimately led to the rise of modern analysis. Several of the questions were in response to the work of Fourier on heat flow, which is a topic in our next chapter.

Example 2.23. Earlier in this chapter, we asked about the formula $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$. A traditional derivation of this starts with the geometric series $\frac{1}{1+x} = \frac{1}{1-(-x)} = 1 - x + x^2 - x^3 + \cdots$ for $|x| < 1$. We can now easily justify the convergence of this series by using convergence tests discussed above. To formally obtain the logarithm formula, just integrate the formula term by term. This produces the expression $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$.

As already mentioned, there are at least two questions that need to be addressed. Does the new integrated series converge; and if so, does it converge to $\log(1+x)$? For $0 \le x \le 1$, the right hand side is an alternating series, where the summands clearly converge to zero. So, $0 \le x < 1$ the right hand side of the series does indeed converge.

How, though, might we connect this limit with the logarithm function?

Recall that the symbolism for a convergent series is an abbreviation. When a series converges, the short hand expression excludes writing the remainder term. In order to analyze convergence, evidently we need to know what is going on in that remainder.

Now, the formula $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$ came from integrating the geometric series. So, let's go back to the source. The original geometric series, with remainder

included, is $\frac{1}{1+x}=\frac{1}{1-(-x)}=1-x+x^2-x^3+\cdots(-1)^n x^n+\left[\frac{(-x)^{n+1}}{1-(-x)}\right]$. This is a finite sum, so there is no problem with integrating term by term: We get $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}\cdots(-1)^n\frac{x^{n+1}}{n+1}+\int_0^x\frac{(-1)^{n+1}t^{n+1}}{1+t}dt$. The question then is whether or not the remainder term $R_n(x)=\int_0^x\frac{(-1)^{n+1}t^{n+1}}{1+t}dt$ converges to zero, as n gets large.

If the denominator in the integrand were not present, then this integral could be calculated explicitly. However, we don't need to know an exact value for a remainder. We just need to get an estimate on it, to know that it (or equivalently, its absolute value $|R_n(x)|$) converges to zero as n gets large.

Recall that the larger a denominator, the smaller the ratio. For example, $\frac{\left|(-1)^5x^5\right|}{1+\left(\frac{1}{2}\right)}<\left|(-1)^5x^5\right|=x^5$. Recall also that for the series in question, $|x|<1$.

So, on the interval of integration, as t increases from 0 to $x\geq 0$, the denominator $1+t$ increases from 1 to $1+x$. The end point $t=x$ is fixed, and the variable of integration is t. So, if we look at the integrand, it satisfies $\frac{t^{n+1}}{1+t}\leq\left|\frac{(-1)^{n+1}t^{n+1}}{1+x}\right|\leq t^{n+1}$.

What though is the relationship between $|R_n(x)|=\left|\int_0^x\frac{(-1)^{n+1}t^{n+1}}{1+t}dt\right|$ and the integral $\int_0^x t^{n+1}dt$?

More generally, how does an integral $\int_0^x f(t)\,dt$ relate to an integral $\int_0^x|f(t)|\,dt$.

If we use Cauchy's formula for calculating the integral as a limit of sums, then

wherever $f(t)<0$, we get a summand for the integral of the form $f(t)\,\Delta x<0$. On the other hand, the summands for the absolute value function $|f(t)|\,\Delta x\geq 0$ are always non-negative. Clearly, therefore, $\int_0^x f(t)dt\leq\int_0^x|f(t)|dt$. And similar reasoning gives that for any $g(x)>|f(x)|$, $\int_0^x f(t)\,dt\leq\int_0^x|f(t)|\,dt<\int_0^x g(t)\,dt$. Of course, to prove these results would require adverting to Cauchy's definitions.

For the example in question, we therefore get that $|R_n(x)|=\left|\int_0^x\frac{(-t)^{n+1}}{1+t}dt\right|$ $\leq\int_0^x\left|\frac{(-t)^{n+1}}{1+t}\right|dt<\int_0^x t^{n+1}dt=\frac{x^{n+2}}{(n+2)}<x^{n+2}$. Since x is fixed and less than unity, it immediately follows that the remainder term converges to zero as n gets large. We conclude that $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}\cdots(-1)^n\frac{x^{n+1}}{n+1}+\cdots$, for $|x|<1$.

Example 2.24. Another traditional example is again obtained from a geometric series: $\frac{1}{1-x}=1+x+x^2+\cdots$, for $|x|<1$. If we formally differentiate this term by term, we get the expression $\frac{1}{(1-x)^2}=1+2x+3x^2+\cdots$, for $|x|<1$. Again, is this correct? Does the new differentiated series converge, and if so, to the quantity indicated? As before, in order to abide by the definition of convergence and the meaning of the symbol "…", we need to examine the remainder term.

$\frac{1}{1-x}=1+x+x^2+\cdots+x^n+\frac{x^{n+1}}{1-x}$, for $|x|<1$. This is a finite sum, so we can differe-ntiate all terms. This gives $\frac{1}{(1-x)^2}=1+2x+x^2+\cdots+nx^{n-1}+\frac{nx^{n-1}+(1-n)x^n}{(1-x)^2}$, for $|x|<1$. The remainder in this case is $R_n(x)=\frac{nx^{n-1}-(n-1)x^n}{(1-x)^2}$. Since the

coefficients in the numerator are functions of n, this remainder is a little more subtle to analyze.

Clues: Keeping in mind $|x|<1$, consider the ratio $\left|\frac{R_{n+1}(x)}{R_n(x)}\right|$. Does there exist $0<a<1$ such that the ratios eventually satisfy $\left|\frac{R_{n+1}(x)}{R_n(x)}\right|<a<1$? More precisely, does there exist n_0 such that for $n \geq n_0$ we have $\left|\frac{R_{n+1}(x)}{R_n(x)}\right|<a<1$?

Observe that if this holds, then $|R_{n_0+3}|<a|R_{n_0+2}|<a^2|R_{n_0+1}|<a^3|R_{n_0}|$; and more generally, $|R_{n_0+k}|<a|R_{n_0+k-1}|<a^2|R_{n_0+k-2}|<\cdots<a^k|R_{n_0}|$. What now follows since $0<a<1$?

Remark 2.2. Here are key questions that we have been looking at in the last two examples:

Suppose that $f(x) = S_n(x) + R_n(x)$.

Question 2.6. If the remainder term $R_n(x)$ converges to zero, does it follow that the integral of $R_n(x)$ converges to zero? The term $R_n(x)$ depends on location x within an interval, while the integral is a sum across an interval.

Question 2.7. If the remainder term $R_n(x)$ converges to zero, does that imply that the derivative of $R_n(x)$ converges to zero? If a function is small in value, does that imply that the slope of the function is small?

Clearly, there are subtleties here that need to be sorted out. Other complexities can also arise, as we will mention in the next chapter when we discuss Fourier series. However, it is already evident that theorems are needed that delineate when a series can be integrated or differentiated term by term.

Example 2.25. Let's go back to the series $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}\cdots(-1)^n\frac{x^{n+1}}{n+1}+\int_0^x\frac{(-t)^{n+1}}{1+t}dt$, for $|x|<1$.

What about the end points of the interval $|x|<1$?

For $x=-1$, the left hand side $\log(1+(-1)) = \log 0$ is not defined. So that case is basically a non-starter. All the same, it is interesting to see what the effect is of

that substitution on the right hand side summation. For $x=-1$, the right hand side

$$-1-\frac{1}{2}+\frac{-1}{3}\cdots+\frac{-1}{n}+\int_0^x\frac{(-t)^{n+1}}{1+t}dt$$

which becomes

$$=-\left(1+\frac{1}{2}+\frac{1}{3}\cdots+\frac{1}{n}\right)+\int_0^x\frac{(-t)^{n+1}}{1+t}dt$$

We already know that the harmonic series does not converge - in fact, the sums grow without bound. Note also that the denominator of the integrand is not defined for $x=-1$, so Cauchy's definition does not directly apply.

An extension of Cauchy's definition of integral designed to deal with this type of integrand is to define $\int_0^{-1}\frac{(-t)^{n+1}}{1+t}dt$ as a limit of defined terms. In the present example, however, even if we define $\int_0^{-1}\frac{(-t)^{n+1}}{1+t}dt \overset{\text{Definition}}{=} \lim_{x\to -1^+}\int_0^x\frac{(-t)^{n+1}}{1+t}dt$, it can be shown that the one-sided limit of integrals does not exist. (Note that the notation "$\lim_{x\to -1^+}$" means the limit as x approaches -1, from above.)

For the other end point of the interval $|x|<1$, formal substitution of $x=1$ gives the formula $\log(2)=1-\frac{1^2}{2}+\frac{1^3}{3}-\cdots+(-1)^n\frac{(1)^{n+1}}{n+1}+\int_0^1\frac{(-t)^{n+1}}{1+t}dt$. The left side log(2) is defined; and by the alternating series test, finite sums $1-\frac{1^2}{2}+\frac{1^3}{3}-\cdots+(-1)^n\frac{(1)^{n+1}}{n+1}=1-\frac{1}{2}+\frac{1}{3}-\cdots+(-1)^n\frac{1}{n+1}$ converge to some target value.

Again though, what though about the remainder term $R_n(x)=\int_0^1\frac{(-t)^{n+1}}{1+t}dt$? We have already shown that the remainder term $\left|\int_0^x\frac{(-t)^{n+1}}{1+t}dt\right|<\int_0^1 t^{n+1}dt=\frac{t^{n+2}}{n+2}\Big|_0^1$

$= \frac{1}{(n+2)}$. But, we only have the functional identification $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} \cdots (-1)^n \frac{x^{n+1}}{n+1} + \int_0^x \frac{(-t)^{n+1}}{1+t}\,dt$ for $|x|<1$. This is of the form $\log(1+x) = S_n(x) + R_n(x)$, where for each $|x|<1$, $R_n(x)$ converges to zero. Is it possible to extend this to $x = 1$ by taking a limit in x? Notice that the upper bound we have for the remainder, namely $\frac{1}{n+2}$, is independent of x. This is usually called being "uniformly bonded". Evidently, further interesting questions arise. The next Example and Exercise help illustrate what can happen when the convergence is not independent of the location x

Example 2.26. ("Non-uniform" convergence and discontinuous limits) In the discussion of the series $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} \cdots (-1)^n \frac{x^{n+1}}{n+1} + \int_0^x \frac{(-t)^{n+1}}{1+t}\,dt$, convergence at a particular x was determined by obtaining an upper bound for the remainder. Specifically, $|R_n(x)| = \left|\int_0^x \frac{(-t)^{n+1}}{1+t}\,dt\right| \le \int_0^x \left|\frac{(-t)^{n+1}}{1+t}\right| dt < \int_0^x t^{n+1}\,dt$ $= \frac{x^{n+2}}{(n+2)} < \frac{x^{n+2}}{n+2} < x^{n+2} < x^n$. Clearly, for each $0<x<1$, $|R_n(x)| < x^n$ converges to zero. Since the control terms here are the functions $y = x^n$, it is reasonable to make an independent enquiry into their convergence properties.

Exercise 2.39. On the closed unit interval, namely, $0 \le x \le 1$, graph several of the functions $y = x^n$, enough to get the pattern. Note that for each x satisfying $0 \le x \le 1$, and for each $n \ge 1$, we have $0 \le x^{n+1} \le x^n \le 1$. Observe also that all of the function $y = x^n$ are defined everywhere on the interval and that $y(1) = 1^n = 1$ for all n. Consider $x = \frac{1}{2}$ and suppose that $\varepsilon = .01$. Find n_0 so that for $n \ge n_0$, $\left(\frac{1}{2}\right)^n < .01$. All of the functions reach the point $(x = 1, y = 1)$, and as graphical representation suggests, the n_0 that works for $x = \frac{1}{2}$ might not work for values of

x closer to $x = 1$. Indeed, using the n_0 just obtained, find $1 > x > \frac{1}{2}$ satisfying $(x)^{n_0} > .01$. Repeat these calculations for the general case. That is, suppose $0 < x_1 < 1$, $0 < \varepsilon < 1$ and that for all $n \ge n_0$ we have $(x_1)^n < \varepsilon$. Find $1 > x > x_1$ such that $x^{n_0} > \varepsilon$.

This shows that while at each $0 < x < 1$ the functions $y = x^n$ converge to zero, the rates of convergence depend on the location within the open unit interval. Notice too that if we make use of the entire closed interval $0 \le x \le 1$ over which the functions $y = x^n$ are defined, there is a limit function on the interval, but the limit function is not continuous, for the limit is defined by $f(x) = \begin{cases} 0 & 0 \le x < 1 \\ 1 & x = 1 \end{cases}$.

This last Exercise 2.39 touches on the need for further study into types of convergence. When the rate of convergence can be controlled by the same n_0 for all x in an interval, the convergence is called "uniform". Limit functions obtained in this way inherit many properties of functions from the approximating sequence. A detailed study of these questions though would go beyond the introductory purpose of these notes.

Example 2.27. (Taylor Series) For most of the examples above, ad hoc calculations were used to obtain as well as analyze various series and their remainder terms. A more systematic approach is given by a "Taylor series", which provides explicit formulas for generating series. As mentioned above, James Gregory (1625 – 1683), Newton (1642 – 1727), Euler (1707 – 1783), Lagrange (1736 – 1813), Fourier (1768 – 1830), and others made use of the interpolation formula $f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots$. Because of Taylor's (1685 – 1731) publications on the subject, the formula was later named the "Taylor series" of the function $y = f(x)$.

As we have been doing, we can enquire into the convergence properties of the series. The Taylor series is constructed in a systematic way from a given function. So we can ask, for a particular function $y = f(x)$, does the Taylor series converge, and if so, does it converge to the given function $y = f(x)$? It turns out that this is a non-trivial question. For example, Cauchy discovered that the Taylor series for the function $f(x) = e^{-x^2} + e^{-x^{-2}}$ does not converge to the function [Katz, p. 707]. To study convergence of Taylor series, it helps to look to the origin of the formula, which is the problem of "polynomial interpolation". Polynomial interpolation is the use of polynomials to approximate a given function $y = f(x)$ (usually near a given reference, or "center", $x = a$).

What is the best constant function $T_0(x) = A_0$ to choose so that, at least near $x = a$, the interpolating function $T_0(x) = A_0$ agrees with the given function at the reference point $x = a$? Of course, we are forced to choose $T_0(x) = f(a)$. Next, what is the best linear function $T_1(x) = A_0 + A_1(x - a)$ to choose so that, at $x = a$, both the function value and its rate of change agree with the original function $y = f(x)$? Straightforward calculations show that $T_1(x) = f(a) + f'(a)\,(x - a)$, which of course is the familiar "tangent line". Next require that, at $x = a$, the approximating second degree polynomial $T_2(x) = A_0 + A_1(x-a) + A_2(x-a)^2$ ("hugging parabola") have the same initial value, the same derivative (velocity), and the same second derivative (acceleration). Doing the calculations gives that $T_2(x) = f(a) + f'(a)\,(x-a) + \dfrac{f''(a)}{2!}(x-a)^2$.

If we keep going, and at each stage require that at the reference point $x = a$ the function value and the $n-1$ derivatives of the interpolating polynomial of degree $n-1$ agree with the original function, we get the Taylor series. That is, we get the approximating (or interpolating) Taylor polynomial

$$T_{n-1}(x) = A_0 + A_1(x-a) + A_2(x-a)^2 + A_3(x-a)^3 + \cdots + A_{n-1}(x-a)^{n-1}$$ of degree $(n-1)$ given by

$$T_{n-1}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}$$

As we discussed above, the question of whether or not this series converges to the function value $y = f(x)$ can be formulated in terms of the remainder

$$R_{n-1}(x) = f(x) - \left[f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} \right]$$

or

$$R_{n-1}(x) = f(x) - f(a) - f'(a)(x-a) - \frac{f''(a)}{2!}(x-a)^2 - \frac{f'''(a)}{3!}(x-a)^3 - \cdots - \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}$$

Let's look at a special case of linear approximation, where the Taylor approximation is the equation for the tangent line $T_1(x) = f(a) + f'(a)(x-a)$. The remainder term $R_1(x) = f(x) - f(a) - f'(a)(x-a)$ measures the difference between the height of the tangent line and the height of the given function. See Fig. 2.21.

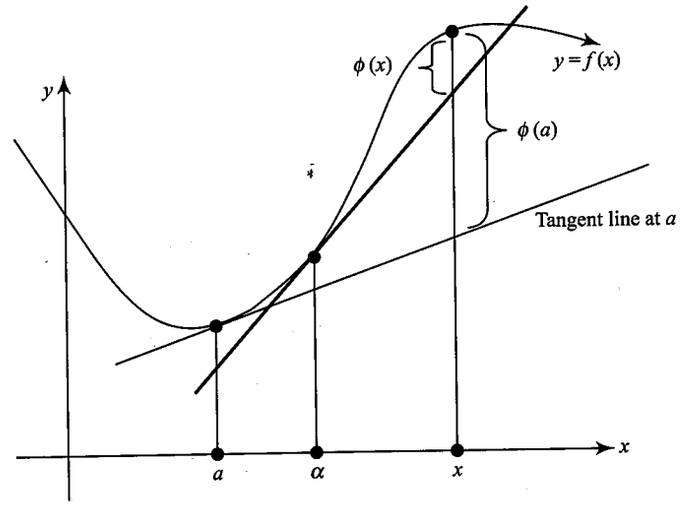


Figure 2.21

Think of the x as the point of interest. That is, suppose that x is fixed, and we need to know the value of the function $y = f(x)$ at x. Formally, the Taylor series can be constructed around any center $x = a$. One way to look at this problem, therefore, is to enquire into what happens to the remainder if we choose different centers a. Does the magnitude of the remainder depend significantly on the choice of center for the approximation? Following Cauchy's lead [Bressoud, 107], we write the remainder term in a way that emphasizes the dependence on the choice of center "a". We write $R_1(\alpha, x) = \phi(\alpha) = f(x) - f(\alpha) - f'(\alpha)(x-\alpha)$, where α represents whatever center is chosen. Note that by suppressing the x and simply writing $\phi(\alpha)$, we have simplified the symbolism somewhat. This is justified since for the purposes of this discussion the x is fixed. (See Figure 2.21 above.)

There are two remainders associated with the problem in a natural way, namely the remainders $\phi(a)$ and $f(x)$ corresponding to the centers $\alpha = a$ and $\alpha = x$. From the diagram, if the center α is close to x, we might expect a smaller remainder term. For $\alpha = x$ we get $\phi(x) = f(x) - f(x) - f'(x)(x-x) = 0$; and for $\alpha = a$ we get $\phi(a) = f(x) - f(a) - f'(a)(x-a)$. How, though, does the remainder depend on the choice of center? Does a strategic choice of center cause the magnitude of the remainder to increase or decrease? The mean value theorem gives us that

$\phi(a)-\phi(x)=(a-x)\phi'(c)$ for some c strictly between a and x. After a few calculations, expressing this back in terms of $y=f(x)$ we get that $R_1(x)=f''(c)(x-a)$.

For the general case, write $\phi(\alpha)=f(x)-f(\alpha)-f'(\alpha)(x-\alpha)-\frac{f''(\alpha)}{2!}(x-\alpha)^2-\frac{f'''(\alpha)}{3!}(x-\alpha)^3-\cdots-\frac{f^{(n-1)}(\alpha)}{(n-1)!}(x-\alpha)^{n-1}$ In exactly the same way as for the linear approximation, the mean value theorem gives that $\phi(a)-\phi(x)=(a-x)\phi'(c)$ for some c strictly between a and x. Again, we have that $f(x)=0$. Using the product rule, straightforward calculations give that the remainder is $R_{n-1}(x)=\frac{f^{(n)}(c)}{(n-1)!}(x-c)^{n-1}(x-a)$. This is called the Cauchy Remainder Theorem.

Exercise 2.40. Use Cauchy's remainder to analyze the convergence of the Taylor series for the function $y=\log(1+x), |x|<1$. Choosing the center $a=0$, the Taylor series is the same as the series developed in the examples above, that is,

$$T_{n+1}=x-\frac{x^2}{2}+\frac{x^3}{3}-\cdots+(-1)^n\frac{x^{n+1}}{n+1}.$$

For $x=\frac{9}{10}$, we obtain $0<c<\frac{9}{10}$ and the Cauchy remainder

$$|R_{n-1}(x)| = \left|\frac{f^{(n)}(c)}{(n-1)!}\left(\frac{9}{10}-c\right)^{n-1}\left(\frac{9}{10}-0\right)\right|$$

$$= \frac{n(1+c)^{-n}}{(n-1)!}\left(\frac{9}{10}-c\right)^{n-1}\left(\frac{9}{10}\right).$$

For $n\geq 4$ we have $\frac{n(1+c)^{-n}}{(n-1)!}=\frac{n}{(n-1)!}\frac{1}{(1+c)^n}<1$ and so

$$\frac{n(1+c)^{-n}}{(n-1)!}\left(\frac{9}{10}-c\right)^{n-1}\left(\frac{9}{10}\right) < \frac{n(1+c)^{-n}}{(n-1)!}\left(\frac{9}{10}-c\right)^{n-1}\left(\frac{9}{10}\right)$$

$$< 1\cdot\left(\frac{9}{10}\right)^{n-1}\left(\frac{9}{10}\right)=\left(\frac{9}{10}\right)^n.$$

Exercise 2.41. Calculate the Taylor series and analyze the convergence for the functions $f(x) = \cos x$, $f(x) = \sin x$, and $f(x) = e^x$.

Notes 2.3. We now have more than one way of calculating remainders for interpolation of the logarithm function $f(x) = \log(1 + x)$. One way is to use the Cauchy remainder. Another approach can be taken by using our results based on geometric series. That is, we may write $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}\cdots(-1)^n\frac{x^{n+1}}{n+1}+\int_0^x\frac{(-t)^{n+1}}{1+t}dt$, for $|x|<1$. How though are the remainders related? Is one formula for the remainder better than the other? Cauchy's remainder formula provides an explicit estimate on the remainder, in terms of the given function. In that way, we obtain a completely self-contained collection of terms determined by the function $y=f(x)$, its derivatives in the finite Taylor series, and the remainder term. There are other formulas for remainder terms. One of these was developed by Lagrange. There is also an integral remainder formula. See, for example, [Spivak, p. 390]. See [Bressoud, p. 108] for a comparison of the Cauchy remainder and the Lagrange remainder, for the function $\log(1+x)$. At some stage in the calculation, the derivation of each well known remainder formula typically appeals to the mean value theorem or the Fundamental Theorem of Calculus, or both.

3

Discovering Real Analysis

Topics: The notion of relative change; relative change as a premise of Newton's laws of motion; d'Alembert's wave equation; d'Alembert's result on a class of solutions of the wave equation; Heat flow, Newton's Law of Cooling and Fourier's Law of Heat Conduction; Fourier's heat equation; Finding solutions of the heat equation by separation of variables; Fourier series; finding solutions of the wave equation; the beginnings of modern analysis.

3.1 CHANGES IN PERSPECTIVE

The purpose of this section is to help enrich appreciation of a change in perspective that occurred in mathematics and physics, due in large part to Newton's discovery of a new approach to physical geometry. It is to be noted that some familiarity with that change in perspective can be most helpful in understanding later developments and topics of this chapter, namely, the wave equation, the heat equation, and Fourier series.

In early mathematics, geometry of the world was studied in terms of ideal objects such as rectangles, circles, conic sections, and so on. This was the main perspective up to the 16^{th} century. However, two scientists of the 16^{th} – 17^{th} centuries helped bring change to the way things were thought about, and this helped prepare the field for Newton's discoveries in calculus and mechanics.

J. Kepler (1571- 1630) discovered the celebrated laws of planetary motion [Burton, 362]:

1. The planets move in elliptical orbits with the sun at one focus.
2. Each planet moves around its orbit, not uniformly, but in a way that a straight line drawn from the sun to the planet sweeps out equal areas in equal time intervals.
3. The squares of the times required for any two planets to make complete orbits about the sun are proportional to the cubes of their mean distances from the sun.

A main feature of these laws is the emphasis on the static structure of a planetary orbit - in continuity with the efforts of the early geometers 2000 years prior. There are, though, aspects of Kepler's laws that were new, when compared with classical geometry. While the emphasis of Kepler's laws is on a static geometry, the laws involve changing areas and time intervals. The perspective though was still primarily classical. In particular the "laws of planetary motion" do not explain the motion as such, and nor do they account for other possible motions such as, say, how an apple falls from a tree.

At roughly the same time as Kepler, but in a more a southern part of Europe, G. Galileo (1564-1642) was also studying the geometry of motion. He was, though, more earthly in his interests. Instead of looking to the heavens, Galileo looked to the puzzle of free-fall, where a free-falling object most noticeably increases its speed while falling toward the ground.

Galileo developed an approach that in fact helped lead to the emergence of modern science. Before Galileo, there were various speculative ideas about the "nature" of motion: An object could be carried by its "principle of motion"; "heavier objects fall more quickly than lighter objects"; and so on. Galileo's objective partly was to obtain a geometry, in the spirit of Euclidean geometry. In that sense, his focus was not on dynamics as such, but on the possibility of obtaining geometric structure. Nevertheless, seeking a geometric system for motion was new; and the follow up of his approach transformed science. Galileo's novel approach had at least three new components: Measure time and distance as simultaneously occurring aspects of a free-fall; try to identify a relationship between the two sets of measurements; and check the hypothesis with experiment.

Certainly, there were those besides Galileo who had made use of experimental data. For instance, Kepler's results were based on the extensive tables of observational results obtained by his senior collaborator Tycho Brahe. Brahe spent many years carefully measuring and recording the motions of the objects of the night sky. Kepler's breakthrough, however, only regarded planetary motions. What was notably unique in Galileo's approach was the deliberate effort not only to find a relationship between different sets of measurements that belonged to the one trajectory, but to place the results within the context of an experimentally verified geometric system. His approach was then the precursor to the later search for geometric explanations of a new type, namely, correlations of measured distance and time, relatively independent of an observer.

There were no high precision clocks in those days. One may wonder what Galileo might have used to measure "time". There is some evidence that Galileo may have used a musician to establish a steady beat [Drake, p. 98]. In order to make measurements accessible to the techniques available, he "slowed down free-fall motion". For, instead of trying to measure distances and times for an actual free-fall, he rolled balls down planes of wood that were tilted at small angles. See Fig. 3.1.

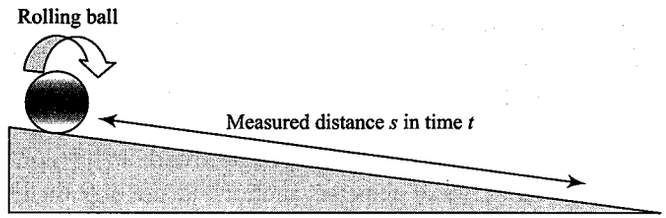


Figure 3.1

The experiments provided tables of measurements. With notable perspicacity, Galileo discovered that the measured distances traveled were proportional to the squares of the measured times. In modern notation, in units of t seconds and s feet, his law turns out to be $s = 16t^2$ feet in t seconds (that is, the proportionality constant is 16). Note that this result expresses a correlation between sets of measurements, and so does not regard the appearance or weight of an object. In particular, it implies that even though a marble slab may have a tremendous weight compared to a small stone, they will both fall at the same rates!

Newton's breakthrough enveloped, subsumed and went beyond the results of both Galileo and Kepler. Newton had what can be called an "inverse insight". While those before him had speculated on the "nature of motion", Newton's insight was that it is not motion as such that is to be understood directly, but change in motion. Instead of trying to explain straight line constant velocity, explain changes from straight line constant velocity. In other words, formulate a theory in terms of rates of change of rates of change.

Recall from Chapter 2 that the rate of change of distance s with respect to time t is the derivative $v = \frac{ds}{dt}$. This is called the *velocity*. In the same way, to calculate the rate at which the velocity changes (the rate of change of the rate of change) we calculate $a = \frac{dv}{dt}$. This is called the *acceleration*. In terms of the distance, $\frac{dv}{dt} = \frac{d\left(\frac{ds}{dt}\right)}{dt}$. Traditionally, the acceleration therefore is denoted short-hand by $\frac{d^2s}{dt^2}$.

Newton had results of collision experiments, from which it was known that in elastic collisions between two objects, the sum of the *momenta* (mass) × (velocity) of the two objects is conserved. That is, if v_1, v_2 are the velocities before a collision, and V_1, V_2 are the velocities after the collision, then $m_1v_1 + m_2v_2 = m_1V_1 + m_2V_2$.

This can be written as $m_1(v_1 - V_1) = -m_2(v_2 - V_2)$. This equation is for a unit of time Δt. Including unit time explicitly, we get $m_1\left(\frac{v_1 - V_1}{\Delta t}\right) = -m_2\left(\frac{v_2 - V_2}{\Delta t}\right)$,

where the ratios are the average accelerations for that time interval. Letting the time interval get small, we obtain $m_1a_1 = -m_2a_2$, where a_1, a_2 are the limiting values of the average accelerations. (In other words, a_1, a_2 are the exact accelerations.)

The product "mass times acceleration" is called *force*. As can be inferred from conservation of momentum, one of Newton's general laws of mechanical motion is that forces occur in equal but opposite pairs. This is sometimes expressed as: "To every action there is an equal but opposite reaction". In this context, note that "action" and "reaction" are not descriptive words for "push" or "pull", but names for forces, where force is mathematically defined as mass times acceleration.

If we include his other law on how to combine "applied" forces, we are lead to the well known result that the net force is equal to the sum of all other forces in the system. In symbols, $(ma)_{Net} = \sum m_i a_i$. For the special case of gravity, he conjectured that the force of gravity is proportional to the product of two interacting masses times the reciprocal distance squared. Newton thereby obtained his universal law of gravitation, $F_{\text{gravity}} = \frac{GMm}{r^2} = Ma_{\text{Mass } M} = -ma_{\text{Mass } m}$, where M, m are the masses of the objects, G is a universal constant, and the force due to gravity is directed along a straight line between the two masses.

Where the results of Kepler and Galileo applied only to certain types of motion, and regarded only certain trajectories, Newton's laws determined a general and universal system of change in terms of accelerations, which applied to all motions. Moreover, in the special case of gravitational force, Newton's laws could be used to re-derive both Kepler's law and Galileo's law.

We can now relate this back to the question of geometry. Early geometers tried to find the static geometry of ideal objects. Even the great Archimedes (who grasped the meaning of limit with a refinement that was not matched for another 2000 years) did not develop a theory of change, but rather used limits of sums to identify target values for static areas. Newton, however, made change a fundamental premise of his theory and consequently developed a general system of dynamics. In particular, Newton discovered general laws of motion for how, relative to each other, physical lengths accelerate in time.

3.2 J. (LE ROND) D'ALEMBERT'S WAVE EQUATION FOR A VIBRATING STRING

Imagine a 10 foot horizontal string fixed at both ends, held taught between two poles. See Fig. 3.2.

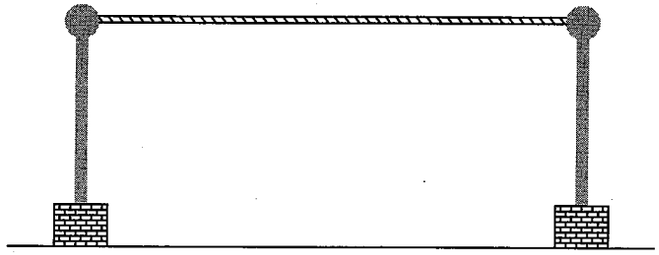


Figure 3.2

Imagine that at one point of the string, someone lifts the string by a couple of inches or so, and then releases it. The string begins to vibrate. See Figure 3.3.

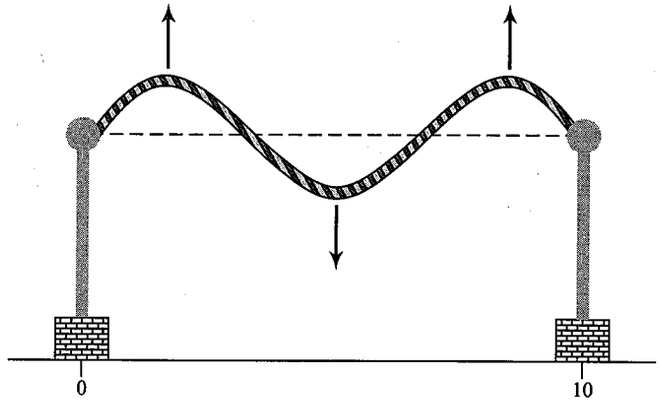


Figure 3.3

The end points are fixed. One finds that segments of the string move up and down. The problem posed is to find the height y as a function of time, where the height y is the distance away from the initial level. The string will be at different heights at different locations. So, the height y will depend on both horizontal distance x, $0 \le x \le 10$ and time t. We therefore write $y = y(x, t)$.

How can we find $y = y(x, t)$?

The height is a quantity that changes in time. The string has mass, and its vertical speeds repeatedly change direction and so both increase and decrease. In other words, there are accelerations.

Newton gave us a law to study such motions, namely, $ma_{Net} = \sum m_i a_i = \sum F_i$. How can we apply that to a length of string that moves differently at different locations?

D'Alembert's approach was to isolate one small segment of string at a time.

See Fig. 3.4.

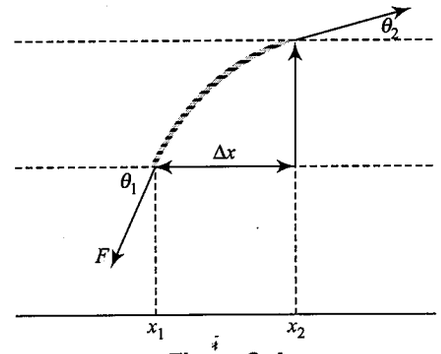


Figure 3.4

Suppose that the horizontal length of the segment is $\Delta x = x_2 - x_1$, and that x is the midpoint with $x_1 < x < x_2$. For simplicity, we suppose that the elastic tension (force) in the string is constant F, and that the string is approximately uniform in its linear mass density, with constant ratio $\frac{\text{mass}}{\text{foot}} = \mu$. From Fig. 3.4, the mass of the segment is therefore $\mu\sqrt{\Delta x^2 + \Delta y^2}$. For simplicity, however, we suppose that the motions have a relatively small displacement, with relatively small Δy. The mass of the segment is then approximately $\mu\sqrt{(\Delta x)^2 + 0} = \mu\Delta x$.

The end point angles θ_1, θ_2 of the segment of string may be different. There are two forces acting on the segments, one from each side due to the tension F. Our question though regards the vertical motion of the string. Newton's laws tell us to add the vertical forces to get $ma_{Net}^{vertical} = \sum F_i^{vertical}$. In our case, we therefore get the approximation $(\mu\Delta x)$ (acceleration of $y(x, t)$ in y direction) $= F \sin \theta_2 - F \sin \theta_1$.

What can we do with the term "(acceleration of $y(x, t)$ in y direction)"?

We are already working under the hypothesis that we are looking at only one segment of string, namely, a segment centered over x. The term "(acceleration of $y(x, t)$ in y direction)" is therefore something that we already know how to calculate. For, if x is fixed, then we need only calculate the second derivative with respect to time t. Note, however, that since the string can move differently at different places, we need a symbolism that expresses that we are calculating the second derivative with respect to time, at a particular location x. The traditional symbolism used to express that we are holding one of the variables fixed is the symbol "∂" ("del" instead of "d").

Our approximation equation for how the height changes can then be written

$$(\mu\Delta x)\frac{\partial^2 y(x,t)}{\partial t^2} = F \sin\theta_2 - F \sin\theta_1 .$$

For simplicity, let's suppose for now that linear mass density and tension are both unity. This last equation then becomes $(\Delta x)\frac{\partial^2 y(x,t)}{\partial t^2} = \sin\theta_2 - \sin\theta_1$.

In keeping with the approach to geometry that was initiated by Galileo and made systematic by Newton, we can try express this equation of change completely in terms of distances and times. We therefore need to re-express the two sine function quantities in terms of distance and time.

We are working under the assumption that the displacement is relatively small. This implies that the angles θ_1, θ_2 are also relatively small. Observe then that $\tan\theta_1 = \frac{\sin\theta_1}{\cos\theta_1} \approx \sin\theta_1$. Note also that the tangent function is another expression for geometric slope. In other words, $\tan\theta_1 \approx \frac{\partial y(x_1,t)}{\partial x}$.

The same reasoning gives us that $\tan\theta_2 \approx \frac{\partial y(x_2,t)}{\partial x}$.

Substituting these ratios of lengths into our equation of change gives us the approximation $(\Delta x)\frac{\partial^2 y(x,t)}{\partial t^2} \approx \left(\frac{\partial y(x_2,t)}{\partial x} - \frac{\partial y(x_1,t)}{\partial x}\right)$.

It is understood that all of the equations are developed as approximations. Just as in calculus, in order to improve the accuracy of the approximation, we can successively reduce the horizontal length Δx. This means trying to identify a target value or limit for the right hand side of the equation $\frac{\partial^2 y(x,t)}{\partial t^2} \approx \frac{\left(\frac{\partial y(x_2,t)}{\partial x} - \frac{\partial y(x_1,t)}{\partial x}\right)}{\Delta x}$.

By definition of derivative, if the limit of the right hand side exists, the ratio $\frac{\left(\frac{\partial y(x_2,t)}{\partial x} - \frac{\partial y(x_1,t)}{\partial x}\right)}{\Delta x}$ converges to $\frac{\partial^2 y(x,t)}{\partial x^2}$.

We therefore obtain the equation $\frac{\partial^2 y(x,t)}{\partial t^2} = \frac{\partial^2 y(x,t)}{\partial x^2}$.

This is an equation for how the height of a vibrating string changes in time. It was derived by applying Newton's laws to a string, under the hypotheses of small displacement, unit linear mass density and constant unit tension.

Partly because the motion of a vibrating string can look like wave motion along a string, this equation has become known as "the wave equation". There are though mathematical reasons for the name as well. There are the results of d'Alembert where he identifies a general class of solutions to this equation. See Section 3.3, below. In order to find particular solutions, there is a technique called "separation of variables", which reveals further connections to "mathematical waves". See Section 3.4.

Exercise 3.1. Using the same or similar rationale, derive the more general equation $\frac{\partial^2 y(x,t)}{\partial t^2} = \left(\frac{F}{\mu}\right)\frac{\partial^2 y(x,t)}{\partial x^2}$, where tension F and linear mass density need not be unity.

Exercise 3.2. For the surface of a drum we can pose the same question. What is an equation of motion for a vibrating drum skin?

Clues: Use an x and y grid for the surface, let $u(x, y, t)$ represent the height as a function of x, y and t. Suppose that there is uniform surface area mass density σ and uniform surface tension S. For each small rectangular segment of dimensions $dx \times dy$, use Newton's laws, first along the x direction; and then along the y direction. Again using Newton's laws, we can add these results to get the net vertical force which is defined in terms of $\frac{\partial^2 u}{\partial t^2}$. Taking limits and simplifying, obtain the 2-D wave equation for a vibrating membrane, namely, $\frac{\partial^2 u}{\partial t^2} = \left(\frac{S}{\sigma}\right)\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right]$.

3.3 D'ALEMBERT'S APPROACH TOWARD CHARACTERIZING SOLUTIONS OF THE 1-D WAVE EQUATION

We now have the 1-D "wave equation" for a vibrating string. (From Exercise 3.2 above, we also have the 2-D wave equation for a vibrating membrane, but will not look at that in detail in these notes. We leave the 2-D case as an Exercise for the reader.) For the vibrating string, what was the original question? It was to find a function $y(x, t)$. We don't yet have a solution to the problem as posed. We do, though, have the "rule of change", or "equation of change", that tells us how $y(x, t)$ changes (accelerates) relative to both changes in time t and changes in horizontal distance x. So, following the approach of d'Alembert, let's see if at least some information can be obtained about possible solutions to the wave equation.

Suppose that a string is held taught, secured at two ends, and is then plucked. The resulting motion as such is vertical. However, over time the displacement pattern steadily shifts back and forth along the string. See Fig. 3.5.

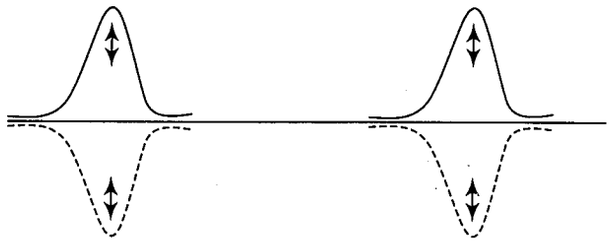


Figure 3.5

The form of the wave equation $\frac{\partial^2 y(x,t)}{\partial t^2} = \left(\frac{F}{\mu}\right)\frac{\partial^2 y(x,t)}{\partial x^2}$ is independent of location x. Indeed, except for the fact that the motions occur at different times, laboratory experiments reveal that the overall pattern of motion at one location x_1 is identical with the pattern of motion at any other location x. Hence, suppose that vertical displacement at x_1 is given by some function $f(t)$ that is independent of location x. Suppose also that the displacement pattern moves along the string with a constant speed c. So, after a time t, the vertical motion $g(x, t)$ at $x = ct$ is identical with the motion determined by $f(t)$. In other words, $g(x, t)$ is simply the translate of $f(t)$, which means that $g(x, t) = f(x - ct)$.

Could a function $g(x, t) = f(x - ct)$ constructed in this way be a solution of the wave equation? In other words, if we hypothesize a motion of a string that is obtained simply by translating some arbitrary function $f(t)$ by a constant speed c, might the overall effect $g(x, t) = f(x - ct)$ be a solution of the wave equation?

Let's see what happens if we calculate second order partial derivatives:

$$\frac{\partial^2 g(x,t)}{\partial t^2} = \frac{\partial^2 f(x-ct)}{\partial t^2} = c^2\frac{d^2 f}{dz^2}; \text{ and}$$

$$\frac{\partial^2 g(x,t)}{\partial x^2} = \frac{\partial^2 f(x-ct)}{\partial x^2} = \frac{d^2 f}{dz^2}, \text{ where } z = x - ct.$$

We therefore get that $\frac{\partial^2 g(x,t)}{\partial t^2} = c^2\frac{\partial^2 g(x,t)}{\partial x^2}$.

In other words, we get a solution of the wave equation with $\left(\frac{F}{\mu}\right) = c^2$. In its time this was a surprising result. See Fig. 3.6.

For, the result tells us that if we use **any!** function $f(t)$ as an initial function, and suppose that its displacement pattern moves ("propagates") in the positive x direction with constant velocity c, then the resulting translate function $g(x, t) = f(x - ct)$ defined across the entire length of the string is a solution of the wave equation.

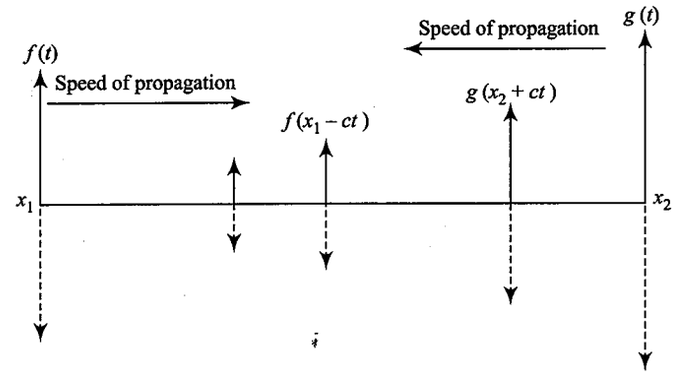


Figure 3.6

Exercise 3.4. Repeat the analysis for a displacement pattern moving to the left.

Clue: $z = x + ct$. See Fig. 3.6.

It now follows that we get a class of solutions of the wave equation: Let $f_1(t)$, $f_2(t)$ be **any two** functions (with second derivatives defined). Construct a combination of two propagations, a propagation of the dynamics of $f_1(t)$ to the right with speed c, and a propagation of the dynamics of $f_2(t)$ to the left also with speed c. Then the new function $y(x, t) = f_1(x - ct) + f_2(x + ct)$ is a solution of the wave equation $\frac{\partial^2 y(x,t)}{\partial t^2} = c^2\frac{\partial^2 y(x,t)}{\partial x^2}$.

It is interesting to see that, at least for an ideal string, this would mean that the speed of propagation is given by $c = \sqrt{\frac{F}{\mu}}$. This seems plausible. Suppose that a string has given mass density. (Without loss of generality, assume that $\mu = 1$). If you do a few experiments, you will find that the more taught a string, the faster the propagation. This is consistent with $c = \sqrt{\frac{F}{\mu}}$, which says that, with $\mu = 1$, the speed of the wave is proportional to $c = \sqrt{F}$. This can of course also be checked accurately in elementary laboratory experiments.

Exercise 3.5. Suppose that $y_1(x, t)$, $y_2(x, t)$ are two solutions of the wave equation. Is $y_1(x, t), y_2(x, t)$ a solution? Can you relate your answer to the origin of the wave equation in Newton's laws? *Clue:* Recall Newton's Law for combining forces.

Remark 3.1. D'Alembert (1717 - 1783) initiated a methodical approach to modeling physical empirical processes by using partial differential equations. In

particular, he discovered the partial differential equation for wave motion, and obtained preliminary results on a general class of solutions. He was not able, though, to systematically produce particular solutions to the wave equation [Katz, 578 ff]. In the present case, we will follow history, and take up the problem of how to find particular solutions of the wave equation, but only after we have first investigated the heat equation, which is another famous partial differential equation.

3.4 HEAT FLOW AND THE HEAT EQUATION

3.4.1 Newton's Law of Cooling

Newton's Law of Cooling is a basic premise that was used in the development of the partial differential equation for heat flow. So, we start with a discussion of a hot drink that is sitting in a cool room. Imagine a cup of some hot drink, perhaps your favorite Darjeeling tea. At first it seems to cool down fairly quickly, but then stays warm almost indefinitely. Is there perhaps a well-defined relationship between changes in temperature and changes in time?

Note that since many of us have observed the same effect in many different settings, we may reasonably conjecture that it is probably not room temperature as such, but the way that the temperature of the tea (or object) changes, relative to the room temperature.

What could our approach be? Newton was developing theories in terms of particle motion. In particular, he had equations for kinetic energy. So, taking a particle approach, we give a sketch for how the Law of Cooling could be derived.

Imagine that the drink consists of particles in motion. In this type of particle model, heat or temperature is taken to be proportional to the average kinetic energy of the particles. Calculations therefore directly regard, not change in temperature, but change in average kinetic energy of the particles. Cooling occurs when the moving particles of the fluid loose their kinetic energy through collisions with nearby air particles. By the same token, this causes these air particles to increase in their average kinetic energy, and so the air near the hot drink is warmed in this process. But, in the particle model, this type of heat transfer continues to occur in the air particles as well, and so the effect spreads throughout the room. Since the room is extremely large compared to the cup, the *average* heat transfer to the air particles of the room is negligible compared to the average heat loss from the much smaller number of particles in the cup of hot drink. Consequently, in the particle model the temperature of the large room is assumed to be approximately constant. Note also that on this approach we can expect that once the average kinetic energy of the hot drink equals the average kinetic energy of the air particles, the average change in kinetic energy will cease. In other words, the temperature of the hot drink will decrease, but not below room temperature.

Let's now try to make this a little more precise. From the last paragraph, we may take the average kinetic energy for the tea to represent the average kinetic energy **above** the average kinetic energy of the room particles. And we are supposing that the temperature/average kinetic energy of the tea decreases through collisions of drink particles with air particles. In our model, this will occur only at the surface of the cup of tea. In other words, we assume that the cup holding the tea is well insulated, except of course at the surface that is open to the air. Suppose for simplicity that there are n particles of tea in the whole cup; that α of these particles are at the surface; and that each tea particle that collides with an air particle looses all of its excess kinetic energy. (Since we are looking to averages relative to ambient room temperature these assumptions are not as strong as they may seem.) At the same time, it is supposed that the tea is reaching its own new average kinetic energy. And since we are looking only to averages, the loss in excess kinetic energy in the tea is therefore approximated by the ratio of particle numbers. That is, $KE_{final} = \left(\frac{n-\alpha}{n}\right) KE_{initial}$. We therefore get that in unit time,

$$KE_{final} - KE_{initial} = \left(\frac{n-\alpha}{n}\right) KE_{initial} - KE_{initial} = \left(\frac{-\alpha}{n}\right) KE_{initial}.$$

However, while we assume that tea particles collide with air particles, we also assume that they are neither destroyed nor leave the "premises" of the tea cup. In other words, for the dimensions involved, we suppose that there is no significant change in the volume of tea. But, it then follows that α and n remain constant for the cup of tea, which means that the ratio $\frac{\alpha}{n}$ is a constant as well. In other words, after a unit of time (at the next stage of cooling), exactly the same argument applies, with the same constants in place - although the tea would be starting from a lower average kinetic energy. Therefore, as the process continues in units of time, the average kinetic energy continues to drop at the same proportionality rate $\frac{-\alpha}{n} KE$. Evidently, the model predicts an exponential decay process. If the time scale is relatively small compared to how long it takes for the temperatures to change, then, based on our arguments so far, we obtain the following differential equation as an approximation to the average change in kinetic energy:

$$\frac{dKE}{dt} = \left(\frac{-\alpha}{n}\right) KE.$$

Now, *Newton's Law of Cooling* is this result, but stated for temperature. Remember that in the present approach, average kinetic energy is assumed to be the kinetic energy above the kinetic energy of the air particles; and is also assumed

to be proportional to temperature ($KE = rT$ for some constant r). So, our result can be written as $\frac{d(rT - rT_{room})}{dt} = \left(\frac{-\alpha}{n}\right)(rT - rT_{room})$. But, the derivative of the constant rT_{room} is zero, and the common term r may be divided from both sides of the equation. Simplifying, we get Newton's formula as it appears in many Calculus books: $\frac{dT}{dt} = -k(T - T_{room})$ for some constant k.

Notice that the larger the temperature difference $T - T_{room}$, the larger the rate of change of T. In some books it is said that the rate of change of temperature is proportional to the temperature drop, or *temperature gradient.* See also Fourier's Law of Heat Conduction, derived in Section 3.4.2 below. Note that under the hypotheses for which we just derived Newton's Law of Cooling, heat is defined as a form of "energy transfer"; and where the energy transferred was assumed to be *kinetic energy*. In nature, there are many forms of energy and many ways that energy can be transferred.

Exercise 3.6. Let's now see how we might use Newton's Law of Cooling in an initial-value problem. For example, suppose that a cup of Darjeeling tea starts out at 98°C, and that the conference room you are in is at a temperature of 72°C. Suppose that after three minutes, the tea temperature has dropped to 90°C. The conference session is to go on for some time. How long will it take for the tea to cool down to 80°C?

3.4.2 From Newton's Law of Cooling to Fourier's Heat Equation

When making pancakes, have you ever used a one-piece cast iron skillet? See Fig. 3.7

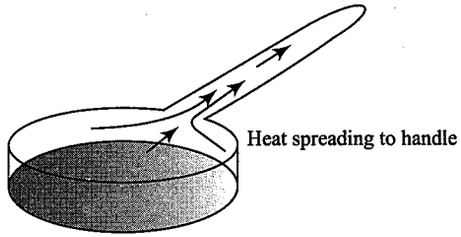


Figure 3.7

Once you've finished cooking the pancakes, you put the skillet aside. Be careful though, for if you reach for the skillet a little while later, heat from the base of the skillet will have spread into the handle. What is happening?

We can follow the approach of Jean-Baptiste Fourier (1768 – 1830). Instead of trying to figure this out for a complicated object like an iron skillet, let's first try to understand this process in something simpler. Suppose that we have a very thin

iron rod, of relatively small radius. To be able study only the effect of the heat spread within the rod, and at the same time to eliminate possible effects of other heat sources, let's suppose that once we heat the rod near the middle say, that we quickly wrap it up in a material so that it becomes perfectly insulated – no heat gets in, and no heat gets out.

Restricting now to this ideal situation, we can ask: How does heat spread along the insulated rod? See Fig. 3.8.

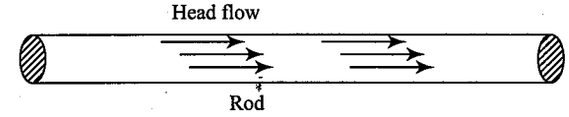


Figure 3.8

Evidently, the spread of heat along the rod involves differences in heat that will depend on both time and location. How can we measure "heat"? As discussed above in the development of Newton's Law of Cooling, heat is a form of energy, and is taken to be proportional to temperature (based on some convenient thermometer scale).

Note also that heat (and therefore temperature) vary along the rod. The *linear heat density* $u(x, t)$ (units of heat per unit length) is therefore a function of both location x and time t.

By definition of linear heat density, for a small length of rod Δx, the amount of heat energy in that segment of rod is approximately $u(x, t)\,\Delta x$. Following d'Alembert (and others), let's try to find the correlated rates of change for heat, location and time by looking at a small segment of the rod, of length $\Delta x = x_2 - x_1$. See Fig. 3.9.

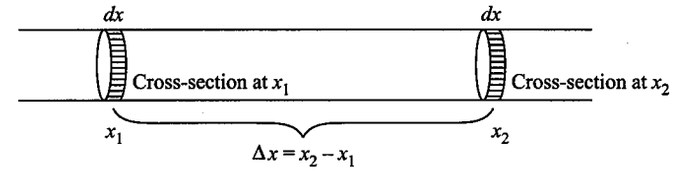


Figure 3.9

Whatever may occur along the length of rod $\Delta x = x_2 - x_1$, as long as we assume that there is conservation of heat energy within the insulated rod, then the only way that the heat energy along the length $\Delta x = x_2 - x_1$ can change is by propagating past the cross-sections at $x = x_1$ and $x = x_2$. We therefore get the master equation

$$\frac{\partial[\text{Volume of Heat in } x_1 \le x \le x_2]}{\partial t} = [\text{Time Rate}]\Big|_{\substack{\text{past cross-}\\ \text{section at } x_2}} - [\text{Time Rate}]\Big|_{\substack{\text{past cross-}\\ \text{section at } x_1}}.$$

Next, let's take one part of this equation at a time, and so try to determine the rate at which heat flows past each cross-sectional area.

Newton's Law of Cooling tells us that the rate of heat transfer through a medium of constant temperature is proportional to the temperature drop. So, for our purposes, let's suppose that the metal rod is small enough in radius that the temperature is constant at every point of a cross-sectional disk of unit area.

As already mentioned above, since we have a linear density function for heat energy given by $u(x, t)$, for dx very small, the product $u(x_1, t)$ is equal to the total heat energy along the small length dx of rod. Note that this small increment dx is based at $x = x_1$, and is therefore different from the length $\Delta x = x_2 - x_1$.

By Newton's Law of Cooling, the time rate of change of heat energy past the cross-section of surface area at $x = x_1$ is approximated by $\frac{\partial}{\partial t}\left[u(x_1,t)\,dx\right] = -k\left[u(x+dx,t)-u(x,t)\right]$, for some constant k. (Typically, the constant k is a property of the type of metal.).

Therefore, $\left.\frac{\partial u(x_1,t)}{\partial t}\right|_{\text{due to cross-sectional flow at } x_1} = -k\left[\frac{u(x_1+dx,t)-u(x_1,t)}{dx}\right]$. Letting dx approach zero, we get that the limiting value for the time rate of change of heat **across the unit surface area of a cross-section** is given by $\left.\frac{\partial u(x_1,t)}{\partial t}\right|_{\text{due to cross-sectional flow}} = -k\frac{\partial u(x_1,t)}{\partial x}$.

This more refined version of the Newton's Law of Cooling is called *Fourier's Law of Heat Conduction*, and tells us that that the time rate of change of heat energy across a sectional area is proportional to the spatial rate of change of the heat energy. Note that the partial derivative $\frac{\text{“}\partial u(x,t)\text{”}}{\partial x}$ is frequently called the *temperature gradient* (meaning gradient with respect to the space coordinate x).

Of course, an exactly similar result is obtained for the other end point, at $x = x_2$.

That is, $\left.\frac{\partial u(x_2,t)}{\partial t}\right|_{\text{due to cross-sectional flow at } x_2} = -k\frac{\partial u(x_2,t)}{\partial x}$.

We can now substitute these values into our master equation to obtain

$$\frac{\partial[\text{Heat in } x_1 \le x \le x_2]}{\partial t} = [\text{Time Rate at } x = x_2] - [\text{Time Rate at } x = x_1]$$

$$\frac{\partial[\text{Heat in } x_1 \le x \le x_2]}{\partial t} = \left[-k\frac{\partial u(x_2,t)}{\partial x}\right] - \left[-k\frac{\partial u(x_1,t)}{\partial x}\right]$$

This leads to the approximation

$$\frac{\partial\left[u(x_1,t)\Delta x\right]}{\partial t} \approx \left[-k\frac{\partial u(x_2,t)}{\partial x}\right] - \left[-k\frac{\partial u(x_1,t)}{\partial x}\right] = -k\left[\frac{\partial u(x_2,t)}{\partial x} - \frac{\partial u(x_1,t)}{\partial x}\right].$$

Dividing by Δx implies

$$\frac{\partial\left[u(x_1,t)\right]}{\partial t} \approx -k\left[\frac{\frac{\partial u(x_2,t)}{\partial x} - \frac{\partial u(x_1,t)}{\partial x}}{\Delta x}\right].$$

Now, the length $\Delta x = x_2 - x_1$ was arbitrary. If we let x_2 approach x_1, and assume that the physical approximations improve accordingly, then the limit gives

$$\frac{\partial\left[u(x_1,t)\right]}{\partial t} = -k\left[\frac{\partial^2 u(x_1,t)}{\partial x^2}\right].$$

Removing the brackets, this can be written as $\frac{\partial u(x_1,t)}{\partial t} = -k\frac{\partial^2 u(x_1,t)}{\partial x^2}$, which is Fourier's famous heat equation.

3.4.3 Summary of Derivation of Fourier's Heat Equation from Newton's Law of Cooling

Assume conservation of energy within an insulated rod. That is, the rod is assumed to be insulated so that energy flow only occurs along the interior of the rod. This allows us to start with a (1-dimensional) *master equation*

$$\frac{\partial[\text{Volume of Heat in } x_1 \le x \le x_2]}{\partial t} = [\text{Time Rate}]\Big|_{\substack{\text{across}\\ \text{surface at } x_2}} - [\text{Time Rate}]\Big|_{\substack{\text{across}\\ \text{surface at } x_1}}$$

By definition of linear heat density, the total heat along the length of rod $\Delta x = x_2 - x_1$ is approximated by $u(x_1, t)\,\Delta x$.

Newton's Law of Cooling/Fourier's Law of Heat Conduction gives the heat flow rates across the sectional surface areas. This gives us the approximation

$$\frac{\partial\left[u(x_1,t)\Delta x\right]}{\partial t} = -k\left[\frac{\partial u(x_2,t)}{\partial x} - \frac{\partial u(x_1,t)}{\partial x}\right].$$

To finish the calculation, divide by Δx to obtain

$$\frac{\partial\left[u(x_1,t)\right]}{\partial t} \approx -k\left[\frac{\dfrac{\partial u(x_2,t)}{\partial x} - \dfrac{\partial u(x_1,t)}{\partial x}}{\Delta x}\right].$$

$$\frac{\partial\left[u(x_1,t)\right]}{\partial t} \approx -k\left[\frac{\dfrac{\partial u(x_2,t)}{\partial x} - \dfrac{\partial u(x_1,t)}{\partial x}}{\Delta x}\right].$$

Now, taking the limit as Δx goes to zero gives the heat equation $\dfrac{\partial u(x_1,t)}{\partial t} = -k\dfrac{\partial^2 u(x_1,t)}{\partial x^2}$.

Exercise 3.7. Develop the heat equation for heat flow through the interior of a plate that is perfectly insulated above and below. *Clues*: We are supposing that heat flux is only horizontal. Look at a small rectangle of the plate, of dimensions Δx by Δy. What is the 2-dimensional *master equation*? What are the rates at which heat flows across the vertical planes perpendicular to the x and y axes respectively. See Fig. 3.10.

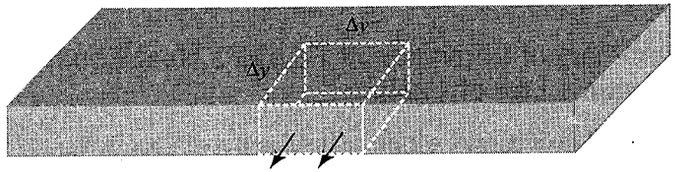


Figure 3.10

Answer: The two dimensional heat equation is $\dfrac{\partial u}{\partial t} = -k\left[\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right]$.

Remark 3.2. In practice, it is not possible to have a rod perfectly insulated along sides of choice. Engineers therefore include additional terms to account for the rates at which heat crosses the remaining bounding surfaces of a volume under consideration. This leads to equations with more terms, but the idea remains the same, and Fourier's Law of Heat Conduction/Newton's Law of Cooling typically is still used to approximate the rates for heat flow across the bounding surface

areas. Note that the right hand side of the master equation gives surface flux rates, while the left hand side is the time rate of change of a volume quantity. Similar results are obtained from the Fundamental Theorem of Calculus, Green's Theorem, Gauss's Theorem, Stokes Theorem (including the Stokes Theorem generalized to n-dimensions), as well as numerous applications of these ideas in physics, engineering and other sciences. In each of these situations, the quantity defined in the higher dimensional interior region is assumed to be conserved. The rate of change of the interior quantity is therefore equal to a flux rate across the lower dimensional bounding surfaces. See Fig. 3.11.

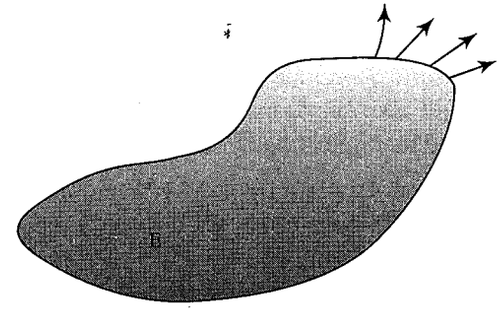


Figure 3.11

Of course, heat flux as defined by Newton and Fourier is only one type of flow. Consequently, different results are obtained in both mathematics and applications when differently defined flux terms are used.

3.5 FINDING SOLUTIONS TO THE HEAT EQUATION

The 1-D heat equation for the rod described above is $\dfrac{\partial u(x_1,t)}{\partial t} = -k\dfrac{\partial^2 u(x_1,t)}{\partial x^2}$.

Finding solutions means finding functions $u(x, t)$ that satisfy the differential equation, and that also satisfy any boundary conditions imposed by the physical situation. We had supposed that the rod, of length L say, had been exposed to a heat source at some initial time $t = 0$. Recall, however, that we are also assuming that rod is perfectly insulated. So, we assume that at the end points of the rod, the temperature drops off to "room temperature", which for present convenience we call zero. Boundary conditions are then $u(0, t) = 0 = u(L, t)$. What about the constant k? Recall that the origin of k is in the derivation of Newton's Law of Cooling/Fourier's Law of Heat Conduction. Since heat is dissipative, we can assume that $-k < 0$, or $k > 0$. Is there a way of keeping track of $k > 0$, without having to keep repeating "$k > 0$"? One way is to replace $k > 0$ by $k = \eta^2$ for some real number η.

We are looking for functions of two variables x and t. What do functions of two variables look like? Where might we look to hopefully find candidates for solutions?

Certainly, any combination such as $x \cdot t$, $7(x^2 \cdot t^5)$, $x \cdot t + 7(x^2 \cdot t^5)$, and so on, will give two-variable functions. Evidently, we can obtain a special class of functions of two variables (x, t) by starting with products of the form $p(x, y) = f(x) \cdot g(t)$. It may not be surprising that the technique of trying to find special solutions of the heat equation of the form $p(x, y) = f(x) \cdot g(t)$ is called "separation of variables".

Recall d'Alembert's result giving a general class of solutions of the wave equation. Part of the derivation included the observation that the sum of two solutions to the wave equation of the same class gives another solution of the wave equation. In a similar way, suppose that we have two solutions to the heat equation, $p_1(x, t)$ and $p_2(x, t)$. Evidently, since derivatives add and scalars factor through derivatives, we also get that $a_1p_1(x, t) + a_2p_2(x, t)$ is a solution for any choice of scalars a_1, a_2. So, if "separation of variables" can provide basic solutions of the form $p(x, y) = f(x) \cdot g(t)$, the strategy also will allow us to use those basic solutions to construct new solutions by addition, and multiplication by scalars.

Let's get started. Substituting $p(x, y) = f(x) \cdot g(t)$ into the heat equation yields the equation $f(x) \cdot g'(t) = -kf''(x) \cdot g(t) = \eta^2 f''(x) \cdot g(t)$.

As long as $f(x) \neq 0$, $g(t) \neq 0$, this gives $\dfrac{f''(x)}{f(x)} = \dfrac{g'(t)}{g(t)} = -\eta^2$.

The left side of this is a function of t and the right side is a function of x. What does this mean about the common ratio? Our assumption in the model was that $k = \eta^2$ is constant. However, to be sure that what we have so far is reasonable, let's look at this last equation a little more carefully. Is it possible that a ratio $\dfrac{f''(x)}{f(x)}$ which is only in terms of x can equal the ratio $\dfrac{g'(t)}{g(t)}$ which is only in terms of t? Changing x on the left side has no effect on the right side. But the right side is identically equal to the left side. So, the left side doesn't change either. It follows that the common ratio does not depend on x. By symmetry of argument, the ratio also does not depend on t. In other words, the only way for a function of "x only", to be identical to a function of "t only", is for both of the functions to be the same constant. But, this is what we have in the equation $\dfrac{f''(x)}{f(x)} = \dfrac{g'(t)}{g(t)} = -\eta^2$.

Hence, so far so good.

Each of these ratios gives a differential equation of a single variable.

$$f''(x) + \eta^2 f(x) = 0$$

$$g'(t) + \eta^2 g(t) = 0.$$

What kind of function has the property that its derivative or derivatives are proportional to the original function value? One type that you may have seen before is an exponential function. We therefore might try for functions of the form $f(x) = e^{\alpha x}$, $g(t) = e^{\beta t}$. See also one of the standard undergraduate texts on differential equations.

From the first equation for $f(x) = e^{\alpha x}$, we get $\alpha^2 + \eta^2 = 0$. The solution of this algebraic equation is complex, $\alpha = \pm \eta\sqrt{-1} = \pm \eta i$. Using $\alpha = \eta i$, the exponential solution to the first equation is $f(x) = e^{\eta i} = \cos \eta x + i \sin \eta x$. (We will say more about Euler's formula $e^{\eta i} = \cos \eta x + i \sin \eta x$ in Chapter 4, Epilogue.)

Using exactly the same approach for $g(t) = e^{\beta t}$ in $g'(t) + \eta^2 g(t) = 0$, we get $\beta + \eta^2 = 0$, so $g(t) = e^{-\eta^2 t}$.

If we substitute $p(x, y) = f(x) \cdot g(t) = (\cos \eta x + i \sin \eta x) e^{-\eta^2 t} = e^{-\eta^2 t} \cos \eta x + ie^{-\eta^2 t} \sin \eta x$ into the heat equation $\dfrac{\partial u(x_1, t)}{\partial t} + \eta^2 \dfrac{\partial^2 u(x_1, t)}{\partial x^2} = 0$, we get that both the real part $e^{-\eta^2 t} \cos \eta x$ and the imaginary part $e^{-\eta^2 t} \sin \eta x$ must also be solutions in their own right. [This is because in addition to having the property that solutions can be added and multiplied by scalars to produce more solutions (see above), the heat equation itself has no complex coefficients.]

Exercise 3.8. Suppose that $u(x, t)$, $v(x, t)$ are real valued functions, $i^2 = -1$, and the complex function $z(x, t) = u(x, t) + iv(x, t)$ is a solution of the heat equation. Show that both $u(x, t)$ and $v(x, t)$ are real valued solutions of the heat equation.

Continuing with the main discussion, since $e^{-\eta^2 t} \cos \eta x + ie^{-\eta^2 t} \sin \eta x$ solves the heat equation, we get that both $e^{-\eta^2 t} \cos \eta x$ and $e^{-\eta^2 t} \sin \eta x$ must also solve the heat equation. We therefore get candidates for basic solutions of the form $p(x, t) = f(x) \cdot g(t)$. That is, $p(x, t) = Ae^{-\eta^2 t} \cos \eta x$, for some constant A, or $p(x, t) = Be^{-\eta^2 t} \sin \eta x$ for some constant B.

We have used separation of variables to determine special basic forms for candidate solutions to the heat equation. We have not yet made use of the boundary conditions. The temperature at the end points is given by $u(0, t) = 0 = u(L, t)$. For $p(x, t) = Ae^{-\eta^2 t} \cos \eta x$, this implies that $0 = p(0, t) = Ae^{-\eta^2 t} \cos 0 = Ae^{-\eta^2 t}$. Therefore, $A = 0$, and so a basic solution of the heat equation for the insulated rod whose left end point is zero degrees can have no cosine functions in its solution. At the other end point of the insulated rod, we get that $Be^{-\eta^2 t} \sin \eta L = 0$. Since B is a freely

chosen constant, it follows that we will need $\sin \eta L = 0$. The length of the rod L is one of the boundary conditions. We obtain, therefore, that $\eta L = n\pi$, where n can be any integer. In other words, we get relationships between the two parameters η and L given by $\eta = \frac{n\pi}{L}$, n an integer.

Recall that the strategy is to build solutions out of sums of scalar multiples of basic solutions. But, we have determined basic solutions $p_n(x, t) = B_n e^{-\left(\frac{n\pi}{L}\right)^2 t} \sin\left(\frac{n\pi x}{L}\right)$, $\eta = \frac{n\pi}{L}$, n an integer. Along with Fourier, we now obtain candidates for solutions of the heat equation of the form $u(x,t) = \sum_{n\in\mathbb{Z}} B_n e^{-\left(\frac{n\pi}{L}\right)^2 t} \sin\left(\frac{n\pi x}{L}\right)$.

To simplify the symbolism somewhat, we can also write $u(x,t) = \sum_{n\in\mathbb{Z}} B_n e^{-\eta_n^2 t} \sin(\eta_n x)$, where $\eta_n = \frac{n\pi}{L}$.

What does this mean about the heat flow in the rod? At time $t = 0$, we get that the initial heat distribution along the rod is given by $u(x,0) = \sum_{n\in\mathbb{Z}} B_n \sin(\eta_n x)$.

As time increases, the exponential factors $e^{-\eta_n^2 t}$ in the series cause the temperature at a location x to decay rapidly. This is consistent with macroscopic experiment. See Fig. 3.12.

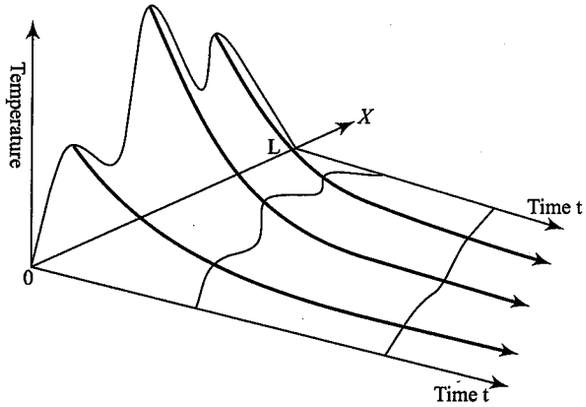


Figure 3.12

3.6 FURTHER QUESTIONS ABOUT FOURIER'S HEAT EQUATION AND FOURIER'S SERIES

Is there anything special about the temperature distribution $h(x) = u(x, 0)$? In fact, we made no initial assumptions about this function. Suppose that instead of being given temperatures at the end points with boundary conditions $u(0, t) = u(L, t) = 0$, we are given initial conditions in terms of the spatial gradient of the temperature. For example, we might suppose that the temperature reaches a local minimum at the end points, and so $\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(L,t) = 0$. In that case, exactly the same approach that we just used will give a somewhat different solution. Instead of a sum with sine functions dependent on x, we would get a sum with cosine functions dependent on x.

Exercise 3.9. Do this calculation.

In other words, different boundary conditions can result in series with sine functions, cosine functions, and even combinations of both. And, since we made no special initial assumptions about the temperature distribution function $h(x) = u(x, 0)$, these observations lead to the following mathematical question:

Question 3.1. Given a function $h(x)$ defined on an interval $0 \le x \le L$, can we find coefficients A_n, B_n such that $h(x) = \sum_{n\in\mathbb{Z}} A_n \cos\left(\frac{n\pi}{L}x\right) + \sum_{n\in\mathbb{Z}} B_n \sin\left(\frac{n\pi}{L}x\right)$?

We can simplify this expression somewhat, by using the symmetries of the sine and cosine functions. Recall that $\cos(-\theta) = \cos(\theta)$ and $\sin(-\theta) = -\sin(\theta)$. Therefore, without loss of generality, we can assume that the sum is over only the non-negative integers. We can also assume that the length of the rod is $L = 2\pi$. Otherwise we can just rescale our units of length. The question then regards the existence of coefficients A_n, B_n such that $h(x) = \sum_{n\in} A_n \cos(nx) + \sum_{n\in} B_n \sin(nx)$.

More explicitly, this is

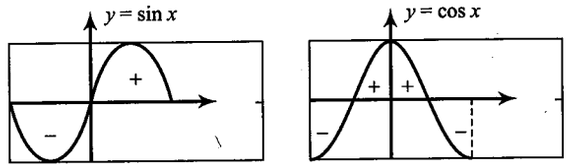
$$h(x) = \left[A_0 + A_1\cos(x) + A_2\cos(2x) + \cdots\right] + \left[B_1\sin(x) + B_2\sin(2x) + \cdots\right]$$

There is no doubt about Fourier's point of view. Suppose that we are given any initial heat distribution function $h(x)$. According to Fourier's conjecture, there exist constants $A_0, A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ such that $h(x) = \left[A_0 + A_1\cos(x) + A_2\cos(2x) + \cdots\right]$ $+ \left[B_1\sin(x) + B_2\sin(2x) + \cdots\right]$.

Fourier's prescription for how to identify the coefficients A_n, B_n was in fact a main catalyst for the development of increasingly adequate theories of integration, through

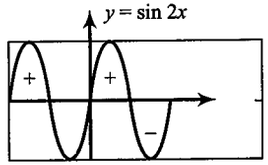
the 19th century, and then into the 20th century with the work of Lebesgue. To get some idea of how this all started, note that part of the problem is to find coefficients A_n, B_n so that we get the interpolation $h(x)=\sum_{n\in\mathbb{N}} A_n \cos(nx)+\sum_{n\in\mathbb{N}} B_n \sin(nx)$. How might we approach this?

In the game of chess, one can only use moves that are part of the game. In a mathematical context, the same principle can apply. The allowed moves are operations, and in the present context, the allowed operations can be those of algebra, differentiation, integration or limits of results of these. But we are looking for a way to produce numbers $A_0, A_1, A_2, \ldots$ and $B_1, B_2, B_3, \ldots$ from the function $h(x)$. Which of the operations just mentioned takes a function and produces a number? That helps focus the problem somewhat. For the question now becomes, "Can we find strategic integrals to produce these coefficients?" Some of the interpolating trigonometric functions are given in Fig. 3.13.

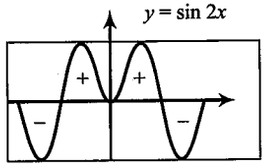


3.13 (a)

3.13 (b)



3.13 (c)



3.13 (d)

Figure 3.13

From parts (a) and (b) of the diagram, one may see that $\int_{-\pi}^{\pi} \cos x dx=0, \int_{-\pi}^{\pi} \sin x dx=0.$

Of course, calculation confirms this.

What about $\int_{-\pi}^{\pi} \sin 2x \sin x dx$? See Figure 3.13 (c) and (d). Again from the diagram, one may see that the integral is zero. The product $y = \sin 2x \sin x$ is (i) even; and (ii) symmetrically out of phase in such a way that total areas above the x – axis are compensated by the total areas below the x – axis. This implies that $\int_{-\pi}^{\pi} \sin 2x \sin x dx=0$. Continuing in the same way, we conjecture that

$$\int_{-\pi}^{\pi} \sin mx \sin nx dx=\begin{cases}0, & m\neq n\\ \pi, & m=n\neq 0\end{cases}.$$

If we multiply a cosine by a sine, then the two functions are again symmetrically out of phase, and we get $\int_{-\pi}^{\pi} \cos mx \sin nx dx=0$. Similarly, we also get that

$$\int_{-\pi}^{\pi} \cos mx \cos nx dx=\begin{cases}0, & m\neq n\\ \pi, & m=n\neq 0\end{cases}.$$

Putting these observations together we obtain the following table that appears in standard discussions of Fourier analysis:

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \cos mx \sin nx dx = 0$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \cos mx \cos nx dx = \begin{cases}0, & m\neq n\\ 1, & m=n\neq 0\end{cases}$$

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \sin mx \sin nx dx = \begin{cases}0, & m\neq n\\ 1, & m=n\neq 0\end{cases}$$

To prove these formulas, one may use the classical trigonometric identities that replace a product of a sine and/or cosine by a sum of a sine and/or cosine. For example, $\cos nx \cos mx=\frac{1}{2}\left(\cos(n+j)x+\cos(n-j)x\right)$.

Exercise 3.10. Verify the formulas in the table just given.

This collection of identities from the table gives us a way to evaluate the coefficients A_n, B_n that we seek. In other words, we obtain candidates for

coefficients for trigonometric interpolation of the function $h(x)$ as a limit of finite sums of trigonometric functions:

$$h(x) = \sum_{n\in} A_n \cos(nx) + \sum_{n\in} B_n \sin(nx).$$

As an example, suppose that know in advance that a function $h(x) = A_1 \cos(x) + B_3 \sin(3x)$. Then, using the above identities it follows that

$$A_1 = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[A_1\cos(x) + B_3\sin(3x)\right]\cos x dx$$

$$B_3 = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[A_1\cos(x) + B_3\sin(3x)\right]\sin 3x dx$$

$$0 = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[A_1\cos(x) + B_3\sin(3x)\right]\cos nx dx$$

for $n \neq 1, 3$

$$0 = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[A_1\cos(x) + B_3\sin(3x)\right]\sin nx dx$$

More generally, if a function $h(x) = \sum_{n\in} A_n \cos(nx) + \sum_{n\in} B_n \sin(nx)$, then we get

$$A_n = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\sum_{n\in} A_n\cos(nx) + \sum_{n\in} B_n\sin(nx)\right]\cos nx dx$$

$$B_n = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\sum_{n\in} A_n\cos(nx) + \sum_{n\in} B_n\sin(nx)\right]\sin nx dx$$

This leads to Fourier's formulas for calculating the coefficients of a trigonometric interpolation:

$$A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(x)\cos nx dx$$

$$B_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(x)\sin nx dx$$

As we have just done, these formulas can be obtained through the techniques and symbolism of basic calculus. But, just as in the early days of infinite series, we need to ask what these symbolic calculations could mean? What is the meaning of these integrals? So far, the function $h(x)$ has been more or less arbitrary. How can

the integrals involving $h(x)$ be evaluated? Fourier was before Cauchy, and so there was not yet any independent definition of integration. Motivated by the early versions of the Fundamental "Theorem" of Calculus, Fourier's contemporaries defined integration as an inverse to differentiation. In other words, the integral was defined as the "anti-derivative". But, using that definition, what is the integral of $h(x) = e^{-x^2}$, or of $h(x) = e^{-x^2} \cdot (\sin x)$? In many cases, finding explicit formulas for some anti-derivatives did not seem possible. This was problematic to the early architects of analysis. Among other things, the very meaning of "function" came under scrutiny. In fact, hindsight reveals that the operational definition of integral is highly significant, for it shows an early (though not as yet matured) grasp of the possibility of defining objects in terms of operations. That way of thinking flowered in the much later developments of abstract algebra. For the immediate purpose of solving the heat equation, though, the need for an independent definition of integral became increasingly evident.

Additional features of the problem are revealed in the interplay between Fourier's heat flow experiments, mathematical modeling through the heat equation, and boundary conditions. Fourier's mathematical conjectures were challenged by Lagrange and others, but the formulas $A_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(x)\cos nx dx$, $B_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(x)\sin nx dx$ allowed Fourier to easily and consistently verify his intriguing formulas to be consistent with laboratory results.

As already mentioned, a real metal rod cannot be perfectly insulated, and the temperature drop takes place within an insulating boundary layer. See Fig. 3.14.

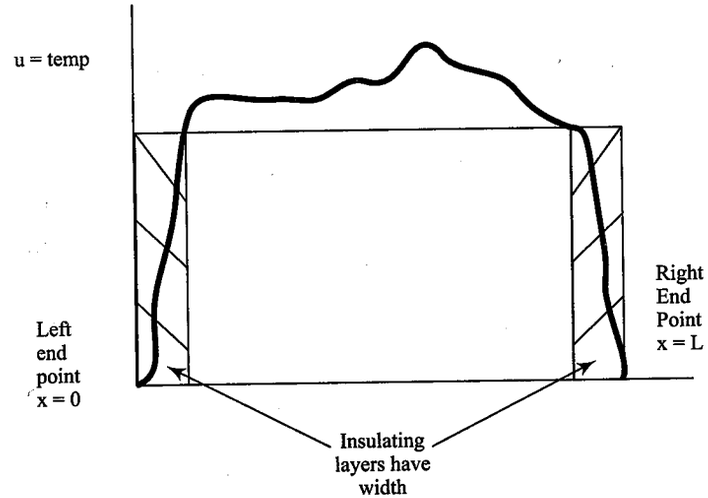


Figure 3.14

The thinner and more efficient the insulating layer, the more extreme the temperature drop. A common mathematical model is for the "ideal" case where the insulating layer has zero width, and the drop in temperature is discontinuous. See Fig. 3.15.

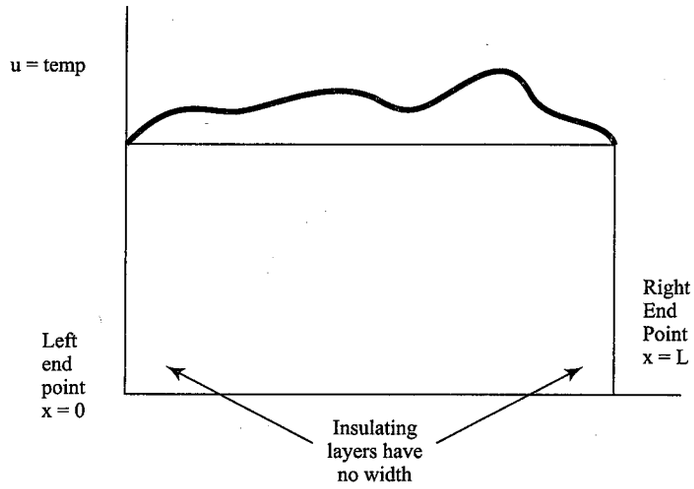


Figure 3.15

Perhaps the simplest test case for modeling this idealization would be a temperature function that is constant within a rod, and zero outside the rod. Keeping to a standard reference interval of length 2π, consider, therefore, the function $f(x)$

$$= \begin{cases} 0 & -\pi \le x < 0 \\ 1 & 0 \le x < \pi \end{cases}.$$

Exercise 3.11. For the function $f(x) = \begin{cases} 0 & -\pi \le x < 0 \\ 1 & 0 \le x < \pi \end{cases}$, calculate and graph the Fourier interpolating functions indicated below. You may find it helpful to consult one of the many websites that have these calculations (and more) well illustrated. See, for example, http://mathworld.wolfram.com/FourierSeries.html. See also Figure 3.16.

(a) $F_1(x) = A_0 + A_1\cos(x) + B_1\sin(x)$

(b) $F_2(x) = A_0 + A_1\cos(x) + B_1\sin(x) + A_2\cos(2x) + B_2\sin(2x)$

(c) $F_3(x) = A_0 + A_1\cos(x) + B_1\sin(x) + \cdots + A_3\cos(3x) + B_3\sin(3x)$

(d) $F_4(x) = A_0 + A_1\cos(x) + B_1\sin(x) + \cdots + A_4\cos(4x) + B_4\sin(4x)$

(e) $F_5(x) = A_0 + A_1\cos(x) + B_1\sin(x) + \cdots + A_5\cos(5x) + B_5\sin(5x)$

(f) $F_6(x) = A_0 + A_1\cos(x) + B_1\sin(x) + \cdots + A_6\cos(6x) + B_6\sin(6x)$

(g) $F_7(x) = A_0 + A_1\cos(x) + B_1\sin(x) + \cdots + A_7\cos(7x) + B_7\sin(7x)$

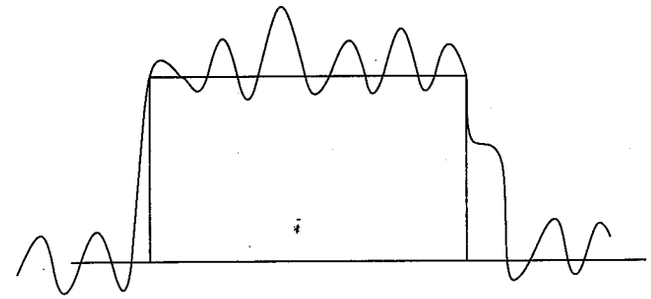


Figure 3.16

There are many other standard examples for heat distribution functions, where an idealized perfectly insulated rod is modeled by a function that has discontinuous jumps at the end points.

Numerous issues arise connected with Fourier's conjecture: How does one prove the convergence of a trigonometric series? There are examples where at $x = \pi$ say, both the left and the right hand limits exist, but are not equal, and neither of these equal the Fourier series when evaluated at $x = \pi$. What kind of function could that be, and how does one integrate such a function. How does one differentiate a Fourier trigonometric series? Even though the summands are all sine and cosine functions, and therefore term by term differentiable, the limit of a Fourier series can be a discontinuous function that is not differentiable anywhere.

Among other things, Fourier's formulas implied the need for a way to integrate functions that have (sometimes infinitely) many discontinuities. To help explore these issues, as well as the very notion of "function", G. L. Dirichlet (1805 – 1859) introduced certain explicitly given test case functions that were differently defined on the rational numbers than on the irrational numbers. For example, there is the now-familiar $f(x) = \begin{cases} 1 & x \in \\ 0 & x \notin \end{cases}$. This example revealed that even though the Riemann integral was an improvement over the Cauchy definition, the Riemann integral also does not exist for all functions that arise in mathematical and physical applications.

You may recall that the Riemann integral is defined as the limit of finite sums of the form $\sum_{i=1}^{n} f(t_i)(x_{i+1} - x_i)$, where $x_0 < x_i < \ldots < x_n$ is a partition of the interval, and for each i, the number t_i satisfies $x_i \le t_i \le x_{i+1}$. This limit is defined relative to

decreasing mesh size (where "mesh size" is defined as the maximum width of the lengths $(x_{i+1} - x_i)$, $(i = 1, 2, \ldots, n)$. Restrict to the unit interval, $0 \le x \le 1$, and let $r_1, r_2, r_3, \ldots$ be a listing of the rational numbers in the interval. If we use the rational numbers as the reference points t_i in the definition of the Riemann integral, then the partial sums of $f(x) = \begin{cases} 1 & x \in \mathbb{Q} \\ 0 & x \notin \mathbb{Q} \end{cases}$ converge to unity. However, in keeping with the definition of the Riemann integral, we may use another partition of arbitrarily small mesh size, where the reference points in the Riemann sum are irrational numbers. In that case, the partial sums are all zero. It follows that the Riemann integral does not exist for this function. At the same time, this is an indication that in order to work with the larger collection of subtleties emergent in Fourier series, a new definition of integral was needed.

Note that a key element in the definition of both the Riemann and the Cauchy integrals is "length" of an interval $(x_{i+1} - x_i)$. Evidently, if there are two intervals, then the total "length" of their union is less than or equal to the sum of their individual lengths. Consider, however, the set of rational numbers $r_1, r_2, r_3, \ldots$ in the unit interval $0 \le x \le 1$. Let $\varepsilon > 0$ be any small positive number, and construct the collection of open intervals $\left(r_i - \frac{\varepsilon}{2^{i+1}}, r_i - \frac{\varepsilon}{2^{i+1}}\right)$. Each of these intervals has length $\frac{\varepsilon}{2^i}$, and so using the geometric series $\sum \frac{1}{2^i}$, the partial sums of their lengths are less than or equal to $\frac{\varepsilon}{2} + \frac{\varepsilon}{2^2} + \cdots + \frac{\varepsilon}{2^n} = \varepsilon\left(\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n}\right) < \varepsilon(1) = \varepsilon$. It would seem that in some sense the rational numbers can be contained in a set that has "total length" less than ε. Since ε was arbitrary, it would follow that the total length of the set of rational numbers would have to zero.

This line of questioning led to Lebesgue's integration theory in the early 1900's, with his definition of integral as a way of "measuring" lengths, areas and volumes that need only be defined "almost everywhere".

In [http://www-history.mcs.st-andrews.ac.uk] we find the following:

Lebesgue formulated the theory of measure in 1901 and in his famous paper *Sur une généralisation de l'intégrale définie,* which appeared in the *Comptes Rendus* on 29 April 1901, he gave the definition of the Lebesgue integral that generalizes the notion of the Riemann integral by extending the concept of the area below a curve to include many discontinuous functions. This generalization of the Riemann integral revolutionized the integral calculus. Up to the end of the 19th century, mathematical analysis was limited to continuous functions, based largely on the Riemann method of integration.

In [Hawkins] we find:

What made the new definition important was that Lebesgue was able to recognize in it an analytic tool capable of dealing with - and to a large extent overcoming - the numerous theoretical difficulties that had arisen in connection with Riemann's theory of integration. In fact, the problems posed by these difficulties motivated all of Lebesgue's major results.

Lebesgue resolved many of these difficulties, and provided finally the analytic results needed for a correct formulation of the Fourier series and their analysis.

Remark 3.1. Using separation of variables led to two differential equations, each of a single variable. Our approach for the single variable equations was to look for a solutions of the form $f(x) = e^{\alpha x}$, $g(x) = e^{\beta t}$. This in fact is a special case of the approach developed by L. Euler (1707 – 1783) in 1739 [Katz, 556-557].

Suppose that we have a differential equation of the form $a_n \frac{d^n y}{dx^n} + a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1 \frac{dy}{dx} + a_0 y = 0$. Looking for a solution of the form $f(x) = e^{\alpha x}$ produces the algebraic equation for the exponent α, namely, $a_n \alpha^n + a_{n-1} \alpha^{n-1} + \cdots + a_1 \alpha + a_0 = 0$. This polynomial equation is now called the "characteristic equation".

Exercise 3.13. Use the method of separation of variables to find solution forms for the wave equations $\frac{\partial^2 y(x,t)}{\partial t^2} = \left(\frac{F}{\mu}\right) \frac{\partial^2 y(x,t)}{\partial x^2}$, $\frac{\partial^2 u}{\partial t^2} = \left(\frac{S}{\sigma}\right)\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right]$ as well as the two variable heat equation $\frac{\partial u}{\partial t} = -k\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right]$.

Notes 3.1: Fourier did not succeed in proving his conjecture. Still, his posthumous notes point in the right direction, and led mathematicians of the day to more adequately investigate length, volume, integration and differentiation. Fourier's results on heat flow, and on using trigonometric series to obtain solutions to his heat equation, revealed the need for a better understanding of the meanings of convergence, differentiation and of integration. His work ultimately led to a new stage in the development in mathematics, namely, the emergence of 19th and 20th century real and harmonic analysis. "Few other works have had such a profound influence on subsequent developments in mathematics as (Fourier's)... Théorie Analytique" [Katz, 629]. See also [Bressoud, 1-6].

4

Epilogue

Complex Numbers, Complex Analysis and Beyond

G. Cardano (1501 – 1576) was an Italian algebraist of the early 16th century. Among other things, he wrote on solving equations, and in particular is known for giving a general solution to the cubic equation in terms of radicals. His work included calculations involving the square roots of negative numbers [Katz, Ch. 9]. Cardano's work was not always expressed in the clearest of terms; and notation was still being developed. Bombelli (1526 – 1572) improved on the presentation of some of Cardano's work, and also made his own contributions to early algebra. Bombelli's work also included calculations involving square roots of negative numbers. For example, he wrote the equivalent of $\sqrt{-3}\sqrt{-3} = -3$.

The square root of a negative number was sometimes called "imaginary". If one imagines numbers as lengths, it becomes challenging to represent the square root of a negative number. And starting from the already known integers, rational numbers, or even irrational numbers, algebraic operations cannot produce a square root of a negative number. Indeed, if a is any resulting rational or irrational number, then $a^2 > 0$. However, the square root of a negative number can be defined in a similar way to how other more familiar numbers are defined.

To see this, first recall that the "0" was introduced historically to balance commerce books. Mathematically this amounted to solving arithmetic problems such as $5 + ? = 5$, and more generally $a + ? = a$. Negative numbers appeared as solutions to problems such as $5 + ? = 0$, and more generally $(a > 0) + b = 0$. Fractions (or "rational" numbers) are solutions to integer ratio equations like $(?)\,5 = 3$, and more generally $xb = a$. It has been known since ancient times that the positive solution $\sqrt{2}$ of $x^2 = 2$ is not rational. More generally, if $b > 0$ is any positive integer that is not a perfect square, then the solutions of $x^2 = b$ are not rational.

There has been a pattern then of gradually adding new numbers, when the new numbers are obtained as solutions of equations. Therefore, allowing square roots of negative numbers was, in that sense, a creative but also natural step to take. Indeed, in the work of both Cardano (1501 – 1576) and Bombelli (1526 – 1572), $\sqrt{-3}$ is essentially defined as a solution to the equation $x^2 = -3$ or $x^2 + 3 = 0$.

De Moivre (1667 – 1754) discovered the formula $(\cos x + i \sin x)^n = \cos nx + i \sin nx$, where n is a non-negative integer and $i = \sqrt{-1}$. This can be proved by induction, a proof technique which was known at least as far back as the 14th century in Europe, and before that in Islamic mathematics [Katz, Ch. 9].

It is fairly straightforward to extend De Moivre's formula to fractional exponents. In other words, for integers p, q we have $(\cos x + i \sin x)^{\frac{p}{q}} = \cos\left(\frac{p}{q}\right)x + i\sin\left(\frac{p}{q}\right)x$. For example, since for any x we have that $(\cos x + i \sin x)^3 = \cos 3x + i \sin 3x$, it follows that $\left(\cos\frac{x}{3} + i\sin\frac{x}{3}\right)^3 = \cos\left[3\left(\frac{x}{3}\right)\right] + i\sin\left[3\left(\frac{x}{3}\right)\right] = \cos x + i\sin x$. We therefore get that $\cos\frac{x}{3} + i\sin\frac{x}{3}$ is a cube root of $\cos x + i \sin x$. For negative exponents, start with the case $(\cos x + i\sin x)^{-1} = \frac{1}{\cos x + i\sin x}$. By complex conjugation, this is $\cos x - i \sin x = \cos(-x) + i \sin(-x)$. For $(\cos x + i \sin x)^{-2}$ we then have $\left[(\cos x + i\sin x)^{-1}\right]^2 = \left[\cos(-x) + i\sin(-x)\right]^2 = \cos(-2x) + i\sin(-2x)$ And so on. There are details of course. It is a good exercise to prove the general case.

At this point in the story, there is the emergence of two distinct (though related) branches of mathematical development. One branch leads to complex analysis; and the other branch leads to algebraic equations, the work of Galois, group theory and abstract algebra.

To get to complex analysis, note that in the early 18th century, L. Euler (1707-1783) showed that $e^{ix} + e^{-ix}$ solved the differential equation as $2 \cos x$; and obtained a corresponding result for $e^{ix} + e^{-ix}$ and $2 \sin x$. Taking solutions of these differential equations to be unique, he made the identifications $e^{ix} + e^{-ix} = 2 \cos x$ and $e^{ix} - e^{-ix} = 2 \sin x$. Straightforward algebra then implies that $e^{ix} = \cos x + i \sin x$ [Katz, 557]. It follows that for any real number α, $(e^{ix})^{\alpha} = e^{i\alpha x} = \cos \alpha x + i \sin \alpha x$, which therefore extends De Moivre's formula to arbitrary real exponents.

However, there is the question of validity. One possible contemporary approach would be to use the fact that any real number can be approximated by a sequence of rational numbers, and apply continuity arguments to De Moivre's formula. Appealing to continuity, however, would require limit theorems for functions that in Euler's time did not yet exist. In fact, Euler's derivation is primarily symbolic technique and algebraic. He formally calculated derivatives of $e^{ix} + e^{-ix}$ and $e^{ix} - e^{-ix}$. The results of his calculations were certainly useful and showed that these exponential terms with complex exponents could be handled consistently with other known real quantities. What was lacking though was an independent definition of the exponential terms e^{ix}.

Later, from Cauchy's (1789-1857) work on convergence, it became possible to define the radius of convergence of a power series of real numbers. But, since a complex number $z = a + ib$ has length $|z| = \sqrt{a^2 + b^2}$, it is also became possible to define the radius of convergence of a power series in the complex variable $z = a + ib$. But that means that we can then define e^{ix} to be the unique complex number determined by substitution $z = ix$ into the absolutely convergent series $\sum_{n=0}^{\infty} \frac{z^n}{n!}$. It can be shown that when there is absolute convergence, the series is well defined, in the sense that rearrangement of terms does not change the limit value of finite sums. Therefore, we may collect the real and imaginary parts and obtain $e^{ix} \overset{\text{Definition}}{=} \sum_{n=0}^{\infty} \frac{(ix)^n}{n!} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots\right)$. One may now observe that the real and imaginary parts e^{ix} are the absolutely convergent Taylor series for the real cosine and sine functions respectively. (See Example 2.27 above.) We therefore get obtain that $e^{ix} = \cos x + i \sin x$, and by the same token extend both De Moivre's formula and Euler's result.

We can add and multiply complex numbers; the operations for complex numbers extend and are compatible with the underlying real numbers; and complex numbers have length. Because of these properties, Cauchy was able to develop a differential and integral calculus for functions of a complex variable $z = x + iy$. To see this, let $f(z) = u(x, y) + iv(x, y)$. Just like the derivative of a real valued function, the derivative of a complex function $\frac{df(z)}{dz}$ was defined as a limit of ratios $\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$.

Note that complex numbers are like vectors in a plane. For, in terms of real numbers, a complex number $z = a + ib$ has two components. This has reaching implications for the differential calculus of complex functions. If the derivative of a complex function exists, then since the limit is defined in terms of the "two

dimensional" distance, the limit $\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$ is independent of the direction of approach $z \to z_0$. In particular, the limit of the ratio along the $z = x + i0 = x$ direction must, by definition, be the same as the limit of the ratio along the $z = 0 + iy = iy$ direction; and both of these must be the same as the limit along any other direction. This independence of direction imposes a symmetry on the real and complex parts of a differentiable complex function $f(z) = u(x, y) + iv(x, y)$. Indeed, following this argument, straightforward calculations show that if the complex function $f(z) = u(x, y) + iv(x, y)$ is differentiable, then the two functions $u(x, y)$ and $v(x, y)$ must satisfy what are now called the Cauchy-Riemann equations $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$. The Cauchy-Riemann equations have played an important role in the development of complex analysis and the theory of analytic functions.

Another branch of mathematics also emerged from the discovery of complex numbers. Continuous with the work of Cardano and other algebraists, there was the on-going development of algebraic equations and algebraic numbers. The young genius Galois (1811 – 1832) discovered the "group of the equation" [Katz, 667]. The implications of Galois' discovery are being explored to this day. The algebraic work of Galois, the complex analysis coming from Cauchy's results, and principles of real analysis have reunited in various branches of modern mathematics of the 20th and 21st centuries. Going further into these matters, however, would take us well beyond the scope of this elementary text.

Bibliography

1. Burton, D.M., The History of Mathematics – An Introduction, WCB McGraw-Hill, 6th Edition. WCB/McGrall Hill Publishers, Boston, Mass., 2007.
2. Bressoud, D.M., A Radical Approach to Real Analysis, 2nd Ed., Mathematical Association of America, Washington, D.C., 2007.
3. Caxton, W., The History and Fables of Aesop, Westminster. Published in 1484. Modern reprint edited by Robert T. Lenaghan, Harvard University Press, Cambridge, 1967.
4. Drake, S., The Role of Music in Galileo's Experiments, Scientific American, June 1975, p. 98.
5. T. Hawkins, Biography, *in* Dictionary of Scientific Biography, American Council of Learned Societies, Scribner Pub., New York, 1970-1990.
6. Katz, V.J., A History of Mathematics – An Introduction, 2nd Ed., Addison Wesley, Reading, Mass., 1998.
7. Heath, T.L., (Ed.), The Works of Archimedes, Dover Pubs. Inc., Mineola, NY, 2002.
8. Marsden, Jerrold E. and Michael J. Hoffman, *Elementary Classical Analysis,* 2nd Ed., Freeman and Co., New York, 1993. (First published in 1974).
9. Simmons, G.F. and S. G. Krantz, Differential Equations – Theory, Technique, and Practice, (Walter Rudin Student Series in Advanced Mathematics), McGraw-Hill Higher Education, Boston, 2007.
10. Spivak, M., Calculus, 2nd Ed., Publish or Perish,, Inc., Berkeley, CA, 1980.
11. http://www-history.mcs.st-andrews.ac.uk/Biographies/Lebesgue.html

Index

T

U

V

W

Z