

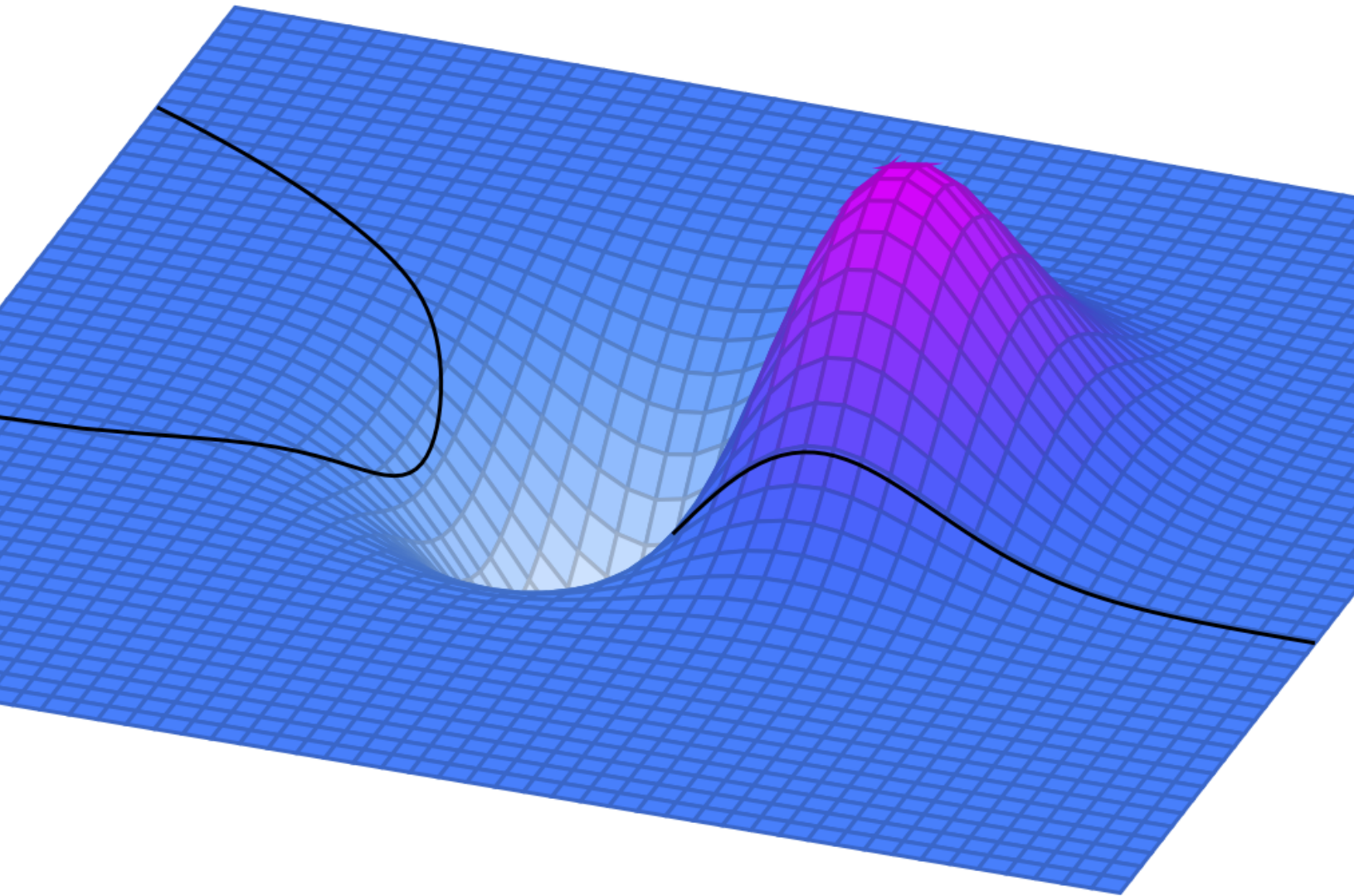
# OPTIMAL, INTEGRAL, LIKELY

OPTIMIZATION, INTEGRAL CALCULUS, AND PROBABILITY  
FOR STUDENTS OF COMMERCE AND THE SOCIAL SCIENCES

---

Prepared by Bruno Belevan, Parham Hamidi, Nisha Malhotra, and Elyse Yeager  
Adapted from *CLP Calculus* by Joel Feldman, Andrew Rechnitzer, and Elyse Yeager

---



---

## ► Licenses and Attributions

Copyright © 2020, 2021 Bruno Belevan, Parham Hamidi, Nisha Malhotra, and Elyse Yeager

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. You can view a copy of the license at <http://creativecommons.org/licenses/by-nc-sa/4.0/>.



Source files can be found at <https://gitlab.math.ubc.ca/ecyeager/OIL>

This textbook contains new material as well as material adapted from open sources.

- Chapters 1 and 2 (and their associated appendix sections) were adapted with minor changes from Chapters 1 and 2 of [CLP 3 – Multivariable Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).
- Chapters 3 and 5 (and their associated appendix sections) and Appendix B were adapted with minor changes from Chapters 1 and 3, Section 2.4, and Appendix A of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).
- Chapter 4 contains content adapted with significant changes from Sections 1.1, 3.1, Ch 4 introduction, 4.1, and 4.2 of [Introductory Statistics](#) by Ilowsky and Dean under a [Creative Commons Attribution License v4.0](#).

## ► Acknowledgements

UBC Point Grey campus sits on the traditional, ancestral and unceded territory of the  $x^w m\theta k^w \acute{a}y\acute{a}m$  (Musqueam). Musqueam and UBC have an ongoing relationship sharing insight, knowledge, and labour. Those interested in learning more about this relationship might start [here](#).

Matt Coles of the University of British Columbia has been an important member of the project to develop quality open resources for Math 105. Thanks to Andrew Rechnitzer at UBC Mathematics for help converting LaTeX to PreTeXt.

The development of this text was supported by an [OER Implementation Grant](#), provided through the UBC Open Educational Resources Fund.

## ► Contact

To report a mistake, or to let us know you're using this book in a course you're teaching, please email [elyse@math.ubc.ca](mailto:elyse@math.ubc.ca)

# CONTENTS

<b>1</b>	<b>Geometry in Three Dimensions</b>	<b>1</b>
1.1	Points	1
1.2	Functions of Two Variables	3
1.3	Sketching Surfaces in 3d	9
1.3.1	Quadric Surfaces	24
<b>2</b>	<b>Partial Derivatives</b>	<b>28</b>
2.1	Partial Derivatives	28
2.2	Higher Order Derivatives	36
2.3	Local Maximum and Minimum Values	39
2.3.1	Critical Points	40
2.3.2	Classifying Critical Points	50
2.4	Absolute Minima and Maxima	61
2.5	Lagrange Multipliers	70
2.5.1	Bounded vs Unbounded Constraints	81
2.6	Utility and Demand Functions	83
2.6.1	Constrained Optimization of the Utility Function	84
2.6.2	Demand Curves	87
<b>3</b>	<b>Integration</b>	<b>95</b>
3.1	Definition of the Integral	95
3.1.1	Summation Notation	103
3.1.2	The Definition of the Definite Integral	107
3.1.3	Using Known Areas to Evaluate Integrals	115
3.1.4	Surplus	118
3.2	Basic Properties of the Definite Integral	122
3.2.1	More Properties of Integration: Even and Odd Functions	129
3.2.2	More Properties of Integration: Inequalities for Integrals	132
3.3	The Fundamental Theorem of Calculus	134
3.3.1	Indefinite Integration	140
3.3.2	Marginal Cost and Marginal Revenue	150

3.4	Substitution . . . . .	153
3.4.1	Substitution and Definite Integrals . . . . .	159
3.4.2	More Substitution Examples . . . . .	163
3.5	Integration by Parts . . . . .	167
3.5.1	Another Technique using Integration by Parts: $dv = dx$ . . . . .	174
3.6	Numerical Integration . . . . .	176
3.6.1	Simpson's Rule . . . . .	179
3.6.2	Error Behaviour . . . . .	183
3.7	Improper Integrals . . . . .	187
3.7.1	Definitions . . . . .	187
3.7.2	Examples . . . . .	193
3.7.3	Convergence Tests for Improper Integrals . . . . .	200
3.8	Overview of Integration Techniques . . . . .	207
3.9	Differential Equations . . . . .	210
3.9.1	(Optional) Logistic Growth . . . . .	220
3.9.2	(Optional) Interest on Investments and Loans . . . . .	226
<b>4</b>	<b>Probability</b> . . . . .	<b>233</b>
4.1	Introduction . . . . .	233
4.1.1	Foundational Vocabulary and Notation . . . . .	233
4.1.2	Discrete vs Continuous . . . . .	237
4.1.3	Combining Events . . . . .	239
4.1.4	Equally Likely Outcomes . . . . .	241
4.2	Probability Mass Function (PMF) . . . . .	244
4.2.1	Limitations of Probability Mass Function (PMF) . . . . .	249
4.3	Cumulative Distribution Function (CDF) . . . . .	251
4.4	Probability Density . . . . .	258
4.4.1	Density Diagrams . . . . .	258
4.4.2	Probability Density Function (PDF) . . . . .	260
4.5	Expected Value . . . . .	267
4.5.1	Motivation: Long-Term Average . . . . .	267
4.5.2	Definition and Examples . . . . .	268
4.5.3	Checking your Expectation Calculation . . . . .	272
4.6	Variance and Standard Deviation . . . . .	278
4.6.1	Motivation: Average difference from the average . . . . .	278
4.6.2	Definitions and Computations . . . . .	280
4.6.3	Checking your Standard Deviation Calculation . . . . .	286
<b>5</b>	<b>Sequences and Series</b> . . . . .	<b>289</b>
5.1	Sequences . . . . .	290
5.1.1	Musical Scales . . . . .	295
5.2	Series . . . . .	300
5.2.1	Geometric Series . . . . .	303
5.2.2	Telescoping Series . . . . .	308
5.2.3	Arithmetic of Series . . . . .	310
5.2.4	(Optional) Intergenerational Cost-Benefit Analysis . . . . .	312
5.3	The Integral and Divergence Tests . . . . .	314

5.4	Comparison Tests	321
5.5	The Ratio Test	327
5.5.1	Convergence Test List	329
5.6	Absolute and Conditional Convergence	330
<b>6</b>	<b>Power Series</b>	<b>333</b>
6.1	Radius of Convergence	334
6.2	Working With Power Series	342
6.3	Extending Taylor Polynomials	349
6.4	Computing with Taylor Series	357
6.5	Evaluating Limits using Taylor Expansions	363
<b>A</b>	<b>Proofs and Supplements</b>	<b>366</b>
A.1	Folding the First Octant of $\mathbb{R}^3$	366
A.2	Vectors	367
A.2.1	Addition of Vectors and Multiplication of a Vector by a Scalar	370
A.2.2	The Dot Product	373
A.3	Conic Sections and Quadric Surfaces	376
A.4	Mixed Partial Derivatives	378
A.4.1	Clairaut: The Proof of Theorem 2.2.5	378
A.4.2	An Example of $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \neq \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$	381
A.5	The (multivariable) chain rule	382
A.5.1	Review of the Proof of $\frac{d}{dt}f(x(t)) = \frac{df}{dx}(x(t)) \frac{dx}{dt}(t)$	384
A.5.2	Proof of Theorem A.5.1	384
A.6	Lagrange Multipliers: Proof of Theorem 2.5.2	388
A.7	A More Rigorous Area Computation	389
A.8	Careful Definition of the Integral	391
A.9	Integrating $\sec x$ and $\csc x$	396
A.10	Further Reading on Numerical Integration	398
A.10.1	The Midpoint Rule	398
A.10.2	The Trapezoidal Rule	401
A.10.3	Error Behaviour	404
A.10.4	An Error Bound for the Midpoint Rule	405
A.11	Comparison Tests Proof	408
A.12	Alternating Series	409
A.12.1	The Alternating Series Test	409
A.12.2	Alternating Series Test Proof	414
A.13	Delicacy of Conditional Convergence	414
<b>B</b>	<b>High school material</b>	<b>418</b>
B.1	Similar Triangles	418
B.2	Pythagoras	419
B.3	Trigonometry — Definitions	419
B.4	Radians, Arcs and Sectors	419
B.5	Trigonometry — Graphs	420
B.6	Trigonometry — Special Triangles	420
B.7	Trigonometry — Simple Identities	420

---

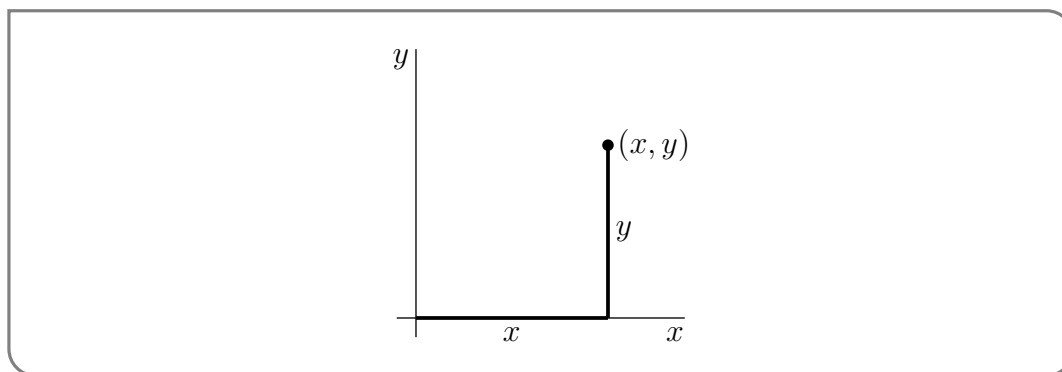
B.8 Trigonometry — Add and Subtract Angles . . . . .	421
B.9 Inverse Trigonometric Functions . . . . .	421
B.10 Areas . . . . .	422
B.11 Volumes . . . . .	423
B.12 Powers . . . . .	423
B.13 Logarithms . . . . .	424
B.14 Highschool Material You Should be Able to Derive . . . . .	425
B.15 Cartesian Coordinates . . . . .	426
B.16 Roots of Polynomials . . . . .	427

# GEOMETRY IN THREE DIMENSIONS

Before we get started doing calculus in two and three dimensions we need to brush up on some basic geometry that we will use a lot. We are already familiar with the Cartesian plane<sup>1</sup>, but we'll start from the beginning.

## 1.1▲ Points

Each point in two dimensions may be labeled by two coordinates<sup>2</sup>  $(x, y)$  which specify the position of the point in some units with respect to some axes as in the figure below.



The set of all points in two dimensions is denoted<sup>3</sup>  $\mathbb{R}^2$ . Observe that

- the distance from the point  $(x, y)$  to the  $x$ -axis is  $|y|$

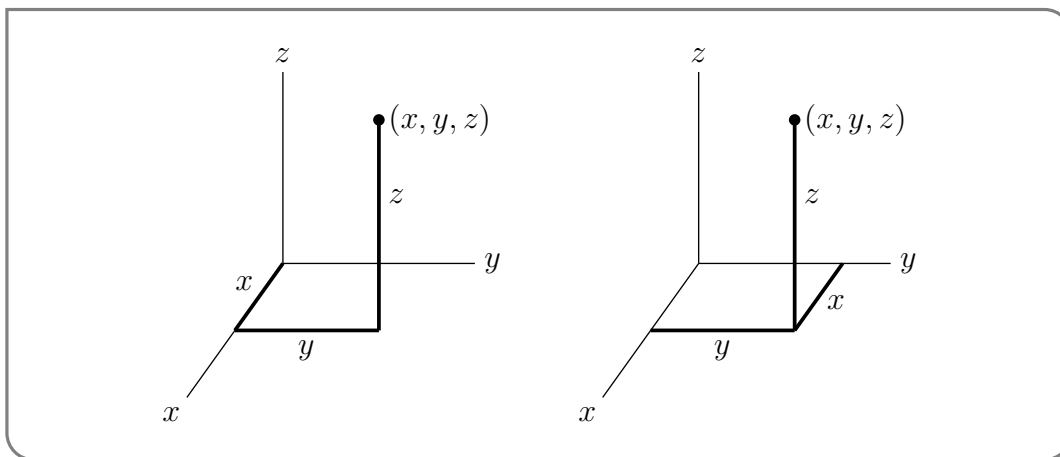
1 René Descartes (1596–1650) was a French scientist and philosopher, who lived in the Dutch Republic for roughly twenty years after serving in the (mercenary) Dutch States Army. He is viewed as the father of analytic geometry, which uses numbers to study geometry.

2 This is why the  $xy$ -plane is called “two dimensional” — the name of each point consists of two real numbers.

3 Not surprisingly, the 2 in  $\mathbb{R}^2$  signifies that each point is labelled by two numbers and the  $\mathbb{R}$  in  $\mathbb{R}^2$  signifies that the numbers in question are real numbers. There are more advanced applications (for example in signal analysis and in quantum mechanics) where complex numbers are used. The space of all pairs  $(z_1, z_2)$ , with  $z_1$  and  $z_2$  complex numbers is denoted  $\mathbb{C}^2$ .

- the distance from the point  $(x, y)$  to the  $y$ -axis is  $|x|$
- the distance from the point  $(x, y)$  to the origin  $(0, 0)$  is  $\sqrt{x^2 + y^2}$

Similarly, each point in three dimensions may be labeled by three coordinates  $(x, y, z)$ , as in the two figures below.

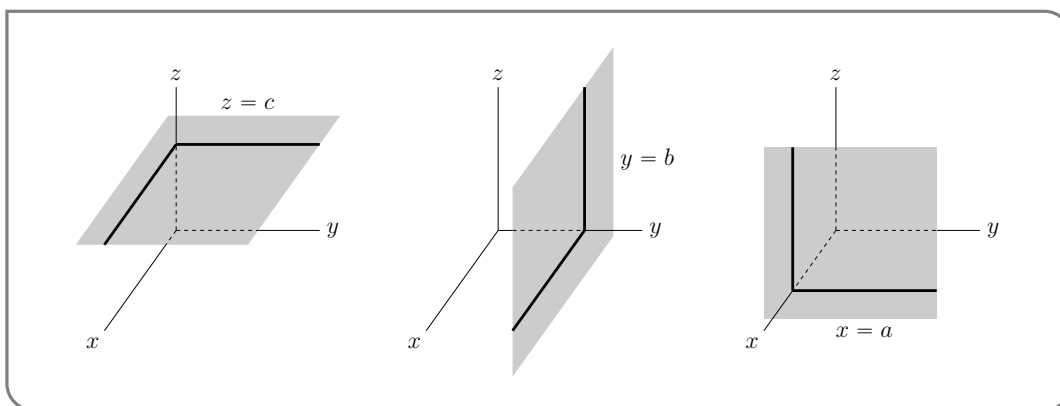


The set of all points in three dimensions is denoted  $\mathbb{R}^3$ . The plane that contains, for example, the  $x$ - and  $y$ -axes is called the  $xy$ -plane.

- The  $xy$ -plane is the set of all points  $(x, y, z)$  that satisfy  $z = 0$ .
- The  $xz$ -plane is the set of all points  $(x, y, z)$  that satisfy  $y = 0$ .
- The  $yz$ -plane is the set of all points  $(x, y, z)$  that satisfy  $x = 0$ .

More generally,

- The set of all points  $(x, y, z)$  that obey  $z = c$  is a plane that is parallel to the  $xy$ -plane and is a distance  $|c|$  from it. If  $c > 0$ , the plane  $z = c$  is above the  $xy$ -plane. If  $c < 0$ , the plane  $z = c$  is below the  $xy$ -plane. We say that the plane  $z = c$  is a signed distance  $c$  from the  $xy$ -plane.
- The set of all points  $(x, y, z)$  that obey  $y = b$  is a plane that is parallel to the  $xz$ -plane and is a signed distance  $b$  from it.
- The set of all points  $(x, y, z)$  that obey  $x = a$  is a plane that is parallel to the  $yz$ -plane and is a signed distance  $a$  from it.



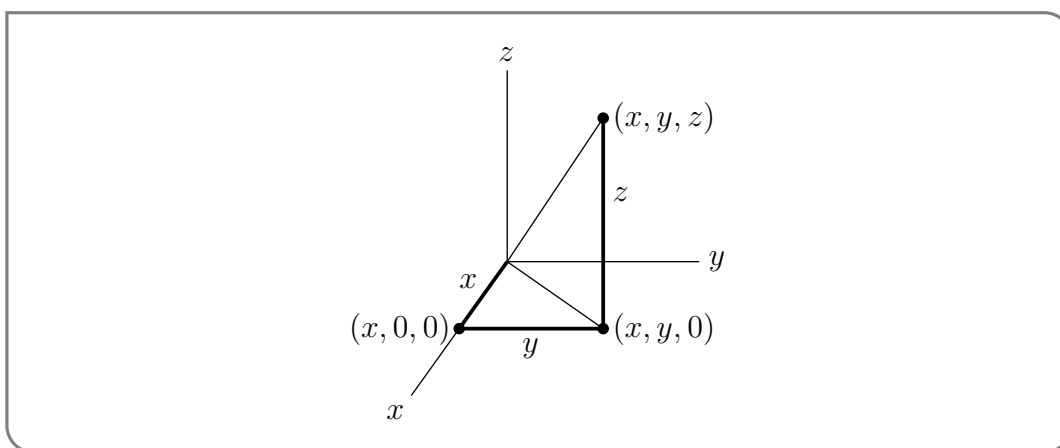
Observe that our 2d distances extend quite easily to 3d.



- the distance from the point  $(x, y, z)$  to the  $xy$ -plane is  $|z|$
- the distance from the point  $(x, y, z)$  to the  $xz$ -plane is  $|y|$
- the distance from the point  $(x, y, z)$  to the  $yz$ -plane is  $|x|$
- the distance from the point  $(x, y, z)$  to the origin  $(0, 0, 0)$  is  $\sqrt{x^2 + y^2 + z^2}$

To see that the distance from the point  $(x, y, z)$  to the origin  $(0, 0, 0)$  is indeed  $\sqrt{x^2 + y^2 + z^2}$ ,

- apply Pythagoras to the right-angled triangle with vertices  $(0, 0, 0)$ ,  $(x, 0, 0)$  and  $(x, y, 0)$  to see that the distance from  $(0, 0, 0)$  to  $(x, y, 0)$  is  $\sqrt{x^2 + y^2}$  and then
- apply Pythagoras to the right-angled triangle with vertices  $(0, 0, 0)$ ,  $(x, y, 0)$  and  $(x, y, z)$  to see that the distance from  $(0, 0, 0)$  to  $(x, y, z)$  is  $\sqrt{(\sqrt{x^2 + y^2})^2 + z^2} = \sqrt{x^2 + y^2 + z^2}$ .



More generally, the distance from the point  $(x, y, z)$  to the point  $(x', y', z')$  is

$$\sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$$

Notice that this gives us the equation for a sphere quite directly. All the points on a sphere are equidistant from the centre of the sphere. So, for example, the equation of the sphere centered on  $(1, 2, 3)$  with radius 4, that is, the set of all points  $(x, y, z)$  whose distance from  $(1, 2, 3)$  is 4, is

$$(x - 1)^2 + (y - 2)^2 + (z - 3)^2 = 16$$

If you're having a hard time picturing the three-dimensional axes, Appendix section [A.1](#) will lead you through folding a model out of a piece of paper.

## 1.2▲ Functions of Two Variables

First, a quick review of dependent and independent variables. *Independent variables* are the variables we think of as changing somehow on their own; the *dependent variables* are the variables whose change we think of as being caused by the independent variables. For example, if you want to describe the relationship between the age of a cup of cottage cheese, and the number of bacteria in that cup, we generally choose age (time) to be the

independent variable and population of bacteria to be the dependent variable: we think of age changing on its own, then that age causing the bacterial population to change.

We could of course go the other way, and write time as a function of bacteria. This could be useful if we were trying to figure out how old the cheese was by counting its bacteria. So the difference between an independent variable and a dependent variable has to do with how we want to interpret a function.

In a single-variable function, by convention we write

$$y = f(x)$$

where  $y$  is the dependent variable and  $x$  is the independent variable. Similarly, in a two-variable function, we generally write

$$z = f(x, y)$$

We think of the variables  $x$  and  $y$  as independent, and the variable  $z$  as dependent.

If we're not too concerned with independent vs dependent variables; or if the relationship between the dependent and independent variables is difficult (or impossible) to write explicitly in this form; then we can also define multivariable functions implicitly. For example, in the equation

$$z^3x + z^2y + xyz - 1 = 0$$

we can think of  $z$  as an implicitly defined function of  $x$  and  $y$ . You've already seen two families of implicitly defined functions: planes and spheres.

**Example 1.2.1**

Which points  $(1, y, 1)$  in  $\mathbb{R}^3$  satisfy the equation

$$z^3x + z^2y + xyz - 1 = 0 ?$$

*Solution.* If  $x = z = 1$ , then the equation becomes

$$1 + y + y - 1 = 0$$

which has solution  $y = 0$ . So the only such point is  $(1, 0, 1)$ .

**Example 1.2.1**

It's common to see a multivariable equation like

$$f(x, y) = \sin(x + y)$$

or

$$g(x, y) = e^{x^2+y^2}$$

and think that the sine and exponential functions are different from the sine and exponential functions we've seen in two dimensions. They aren't! When  $x$  and  $y$  are real numbers, then  $(x + y)$  and  $(x^2 + y^2)$  are real numbers as well. We're taking the sine of a real number in the first equation, and  $e$  to a real power in the second equation, just as we always have.

Functions of two (or more) variables are not so different from functions of one variable in other ways as well.

**Definition 1.2.2** (Domain and Range).

Let  $f(x, y)$  be a function that takes pairs of real numbers as inputs, and gives a real number as its output.

The set of points  $(x, y)$  that can be input to  $f$  is the **domain** of that function. The set of outputs of  $f$  over its entire domain is the **range** of that function.

**Example 1.2.3** (Domain and Range)

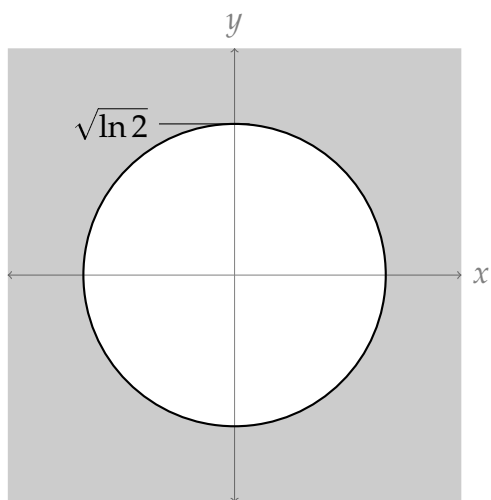
Find the domain and range of the function

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

*Solution.* There are three operations in our function: exponentiation, subtraction, and taking of a square root. We can subtract anything from anything; and we can raise  $e$  to any power. So the only thing that could “break” our function is if we tried to take the square root of a negative number. This tells us that, in order for  $f(x, y)$  to be defined, we need

$$\begin{aligned} (e^{x^2+y^2} - 2) &\geq 0 \\ \implies e^{x^2+y^2} &\geq 2 \\ \implies x^2 + y^2 &\geq \ln 2 \end{aligned}$$

One way of describing the domain of this function is to call it “all points  $(x, y)$  with  $x^2 + y^2 \geq \ln 2$ .” A more standard way is to describe the *shape* this set makes in  $\mathbb{R}^2$ : all points on or outside the circle centred at the origin with radius  $\sqrt{\ln 2} \approx 0.83$ .



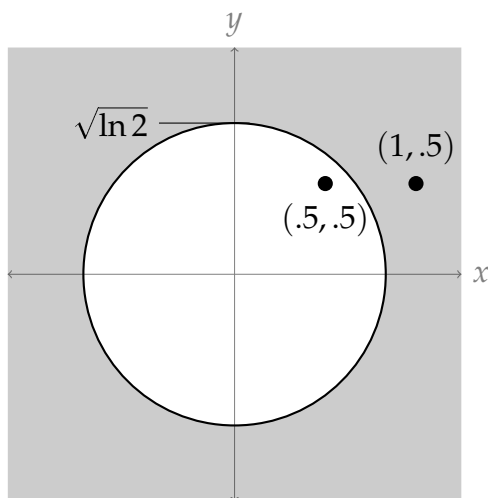
To help you visualize what we mean, take a point in the shaded area above. For example,  $(1, .5)$ . If we plug that into our function, it causes no problems:

$$f(1, .5) = \sqrt{e^{1^2+.5^2} - 1} = \sqrt{e^{1.25} - 2} \approx \sqrt{1.49} \approx 1.22$$

On the other hand, take a point in the white area. For example,  $(.5, .5)$ . If we try to plug this into our function, we end up with

$$f(.5, .75) = \sqrt{e^{.5^2 + .75^2} - 2} = \sqrt{e^{0.5} - 2} \approx \sqrt{1.65 - 2} \approx \sqrt{-0.35}$$

which is not a real number.



Now, let's think about range. By choosing larger and larger values of  $x$  and  $y$ , we can make  $x^2 + y^2$  into larger and larger numbers. So within our restricted domain, the range of  $x^2 + y^2$  is  $[\ln 2, \infty)$ ; so the range of  $e^{x^2 + y^2}$  is  $[e^{\ln 2}, \infty) = [2, \infty)$ ; so the range of  $e^{x^2 + y^2} - 2$  is  $[0, \infty)$ ; so the range of  $f(x, y)$  is  $[0, \infty)$ .

Again, note that the *domain* of  $f$  consists of ordered pairs of real numbers, while its *range* consists of real numbers.

Example 1.2.3

Example 1.2.4

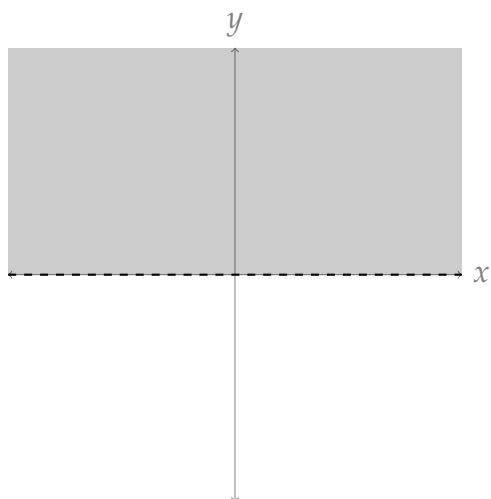
Find the domain and range of the function

$$f(x, y) = \sin\left(\frac{x}{\sqrt{y}}\right)$$

*Solution.* Let's start with domain. We can take the sine of any number we like, so that part of the function doesn't limit the domain. The things limiting the domain are that we cannot take the square root of a negative number, and we can't divide by zero.

- Because we can't take the square root of a negative number, we must have  $y \geq 0$ .
- Because we can't divide by 0, we must have  $\sqrt{y} \neq 0$ , i.e.  $y \neq 0$ .

Combining these restrictions, we can only have values of  $y$  in the interval  $(0, \infty)$ ;  $x$  can be any real number. So, our domain is the upper half of the  $xy$  plane, excluding the  $x$ -axis:



In general, the range of  $\sin x$  is  $[-1, 1]$ . So, we certainly can't get a *larger* range than this. We should check that our range is no smaller. When  $y = 1$ , our function becomes  $f(x, 1) = \sin(x/1) = \sin x$ . Since  $x$  can be any real number, indeed the range of our function is  $[-1, 1]$ .

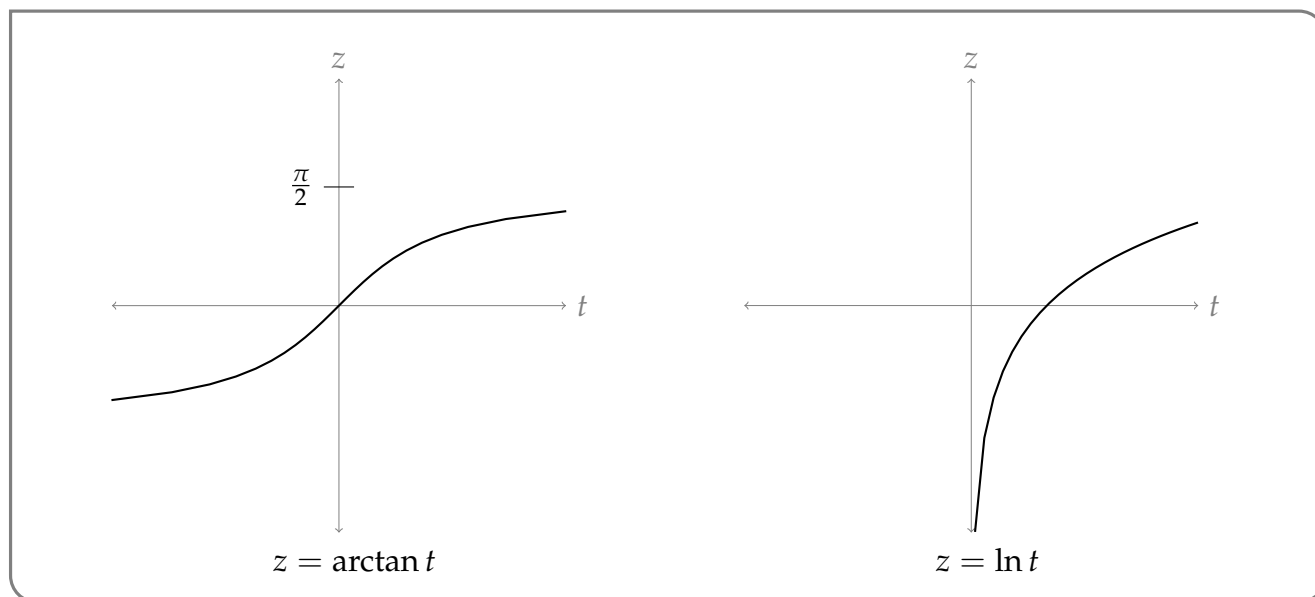
Example 1.2.4

Example 1.2.5

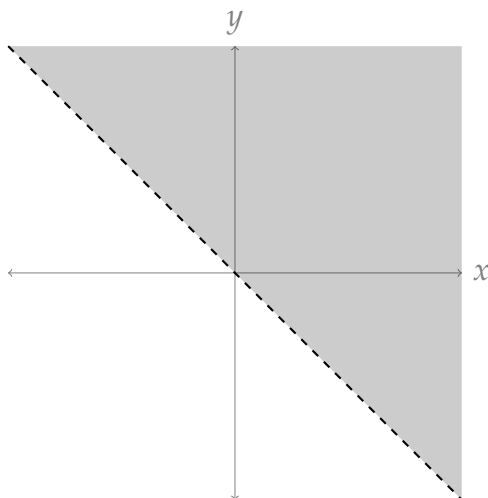
Find the domain and range of the function

$$f(x, y) = \ln(\arctan(x + y))$$

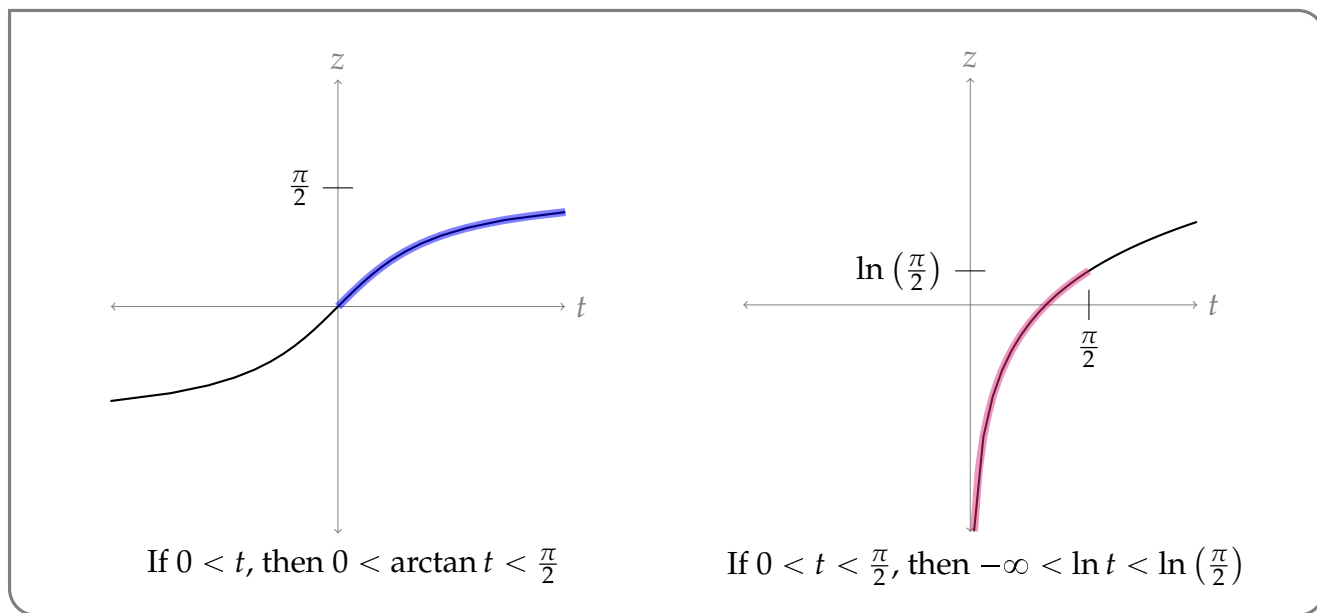
*Solution.* First, let's think about the arctangent and logarithm function in the context of single-variable functions. The domain of arctangent is all real numbers, and its range is  $(-\frac{\pi}{2}, \frac{\pi}{2})$ . The domain of the natural logarithm is all *positive* numbers, and its range is all real numbers.



Since only positive numbers may be input into the natural logarithm, we require  $\arctan(x + y) > 0$ . That requires  $(x + y) > 0$ . So, our domain is the collection of all points  $(x, y)$  such that  $x + y > 0$ ; put another way, all points above the line  $y = -x$ .



If our domain is points  $(x, y)$  such that  $x + y > 0$ , then the range of the function  $(x + y)$  is  $(0, \infty)$ ; so the numbers being plugged into the arctangent function are  $(0, \infty)$ . So, the numbers coming *out* of the arctangent function are  $(0, \frac{\pi}{2})$ . Then the numbers from  $(0, \frac{\pi}{2})$  are being input into the natural logarithm function, leading to a range of the entire function of  $(-\infty, \ln(\frac{\pi}{2}))$ .



Example 1.2.5

We may sometimes restrict the domain of a function more than is mathematically necessary in order for it to make sense in a model. For example, we may have a function that only makes sense in our model when it gives positive values. In this case, we might

restrict the domain to a *model domain*, the set of inputs for which the function is not only defined, but sensible in the context of our model.

Example 1.2.6

A large pharmaceutical company determines its research budget for a new vaccine according to the formula

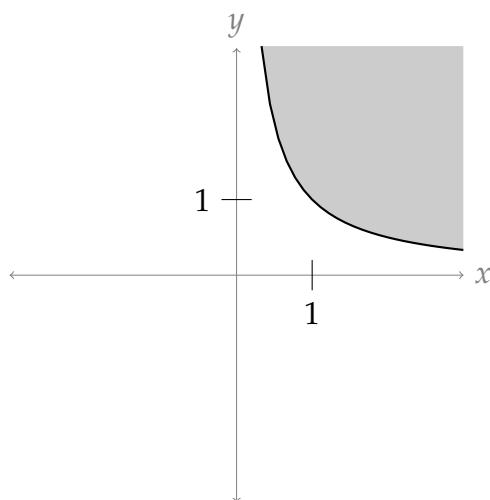
$$R(x, y) = \ln(xy)$$

where  $x$  is the size of the customer base they expect to have and  $y$  is the revenue they expect per dose.

Then for each variable  $x$ ,  $y$ , and  $R$ , negative values don't make sense in the model. So although we *could* compute  $R(-1, -1) = 1$ , and we *could* compute  $R(0.5, 0.5) \approx -1.39$ , they wouldn't be sensible in the context of our model.

- Since  $x$  and  $y$  need to be nonnegative, we will only consider points  $(x, y)$  in the first quadrant of the Cartesian plane:  $x \geq 0$  and  $y \geq 0$ .
- Since  $R$  needs to be nonnegative, we will further restrict  $xy \geq 1$ . That is,  $y \geq \frac{1}{x}$ .

The two restrictions above give us the model domain shaded below.



Depending on the specifics of how the function is being used, the model domain may be restricted even further. For example, perhaps the firm has a maximum budget for any given project; perhaps the amount they can charge is limited by law; etc.

Example 1.2.6

## 1.3<sup>▲</sup> Sketching Surfaces in 3d

In practice students taking multivariable calculus regularly have great difficulty visualising surfaces in three dimensions, despite the fact that we all live in three dimensions. We'll now develop some technique to help us sketch surfaces in three dimensions<sup>4</sup>.

4 Of course you could instead use some fancy graphing software, but part of the point is to build intuition. Not to mention that you can't use fancy graphing software on your exam.

We all have a fair bit of experience drawing curves in two dimensions. Typically the intersection of a surface (in three dimensions) with a plane is a curve lying in the (two dimensional) plane. Such an intersection is usually called a cross-section. In the special case that the plane is one of the coordinate planes, or parallel to one of the coordinate planes, the intersection is sometimes called a trace.

**Definition 1.3.1.**

The trace of a surface is the intersection of that surface with a plane that is parallel to one of the coordinate planes.

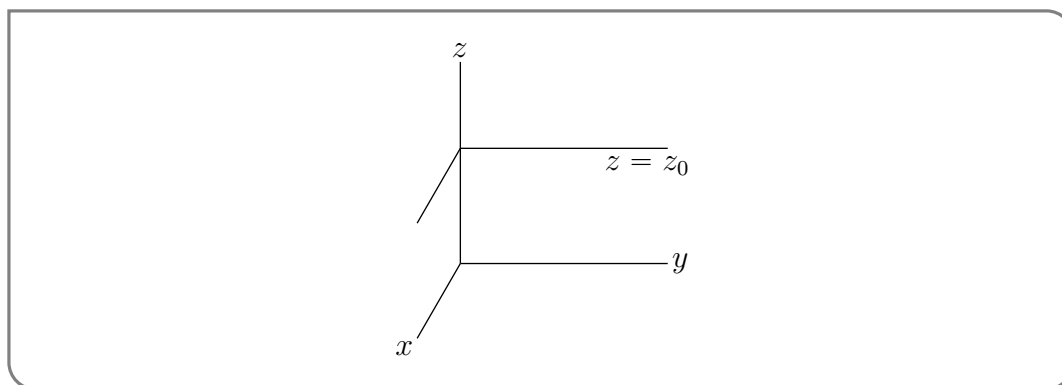
So, one trace (the intersection with the  $xy$  plane) is found by setting  $z$  equal to a constant; another trace (the intersection with the  $yz$  plane) is found by setting  $x$  equal to a constant; and the final trace (the intersection with the  $xz$  plane) is found by setting  $y$  equal to a constant.

One can often get a pretty good idea of what a surface looks like by sketching a bunch of cross-sections. Here are some examples.

**Example 1.3.2** ( $4x^2 + y^2 - z^2 = 1$ )

Sketch the surface that satisfies  $4x^2 + y^2 - z^2 = 1$ .

*Solution.* We'll start by fixing any number  $z_0$  and sketching the part of the surface that lies in the horizontal plane  $z = z_0$ .



The intersection of our surface with that horizontal plane is a horizontal cross-section. Any point  $(x, y, z)$  lying on that horizontal cross-section satisfies both

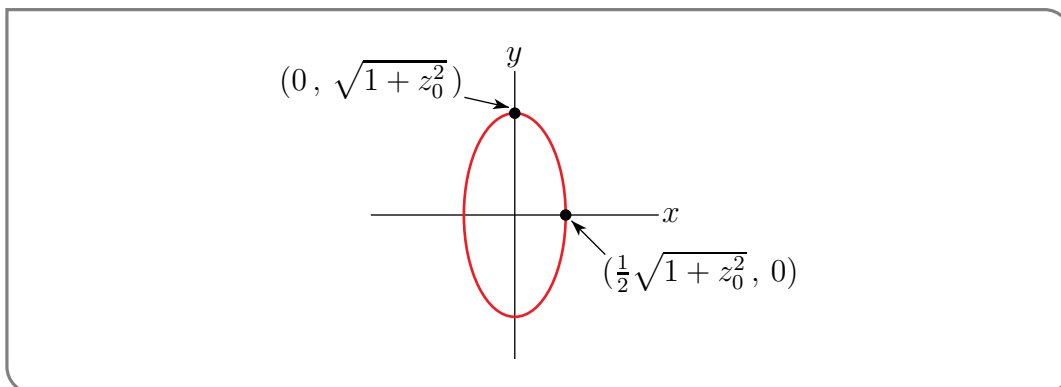
$$\begin{aligned} z = z_0 \text{ and } 4x^2 + y^2 - z^2 = 1 \\ \iff z = z_0 \text{ and } 4x^2 + y^2 = 1 + z_0^2 \end{aligned}$$

Think of  $z_0$  as a constant. Then  $4x^2 + y^2 = 1 + z_0^2$  is a curve in the  $xy$ -plane. As  $1 + z_0^2$  is a constant, the curve is an ellipse. To determine its semi-axes<sup>5</sup>, we observe that when  $y = 0$ ,

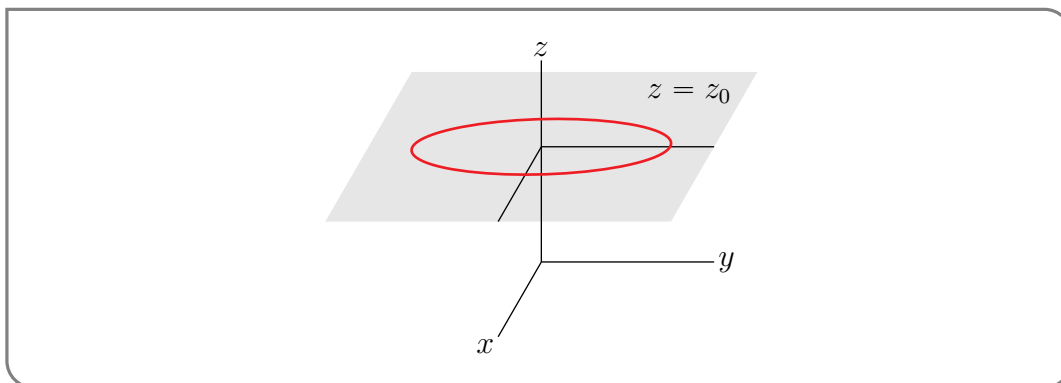
5 The semi-axes of an ellipse are the line segments from the centre of the ellipse to the farthest point on the curve and to the nearest point on the curve. For a circle the lengths of both of these line segments are just the radius.



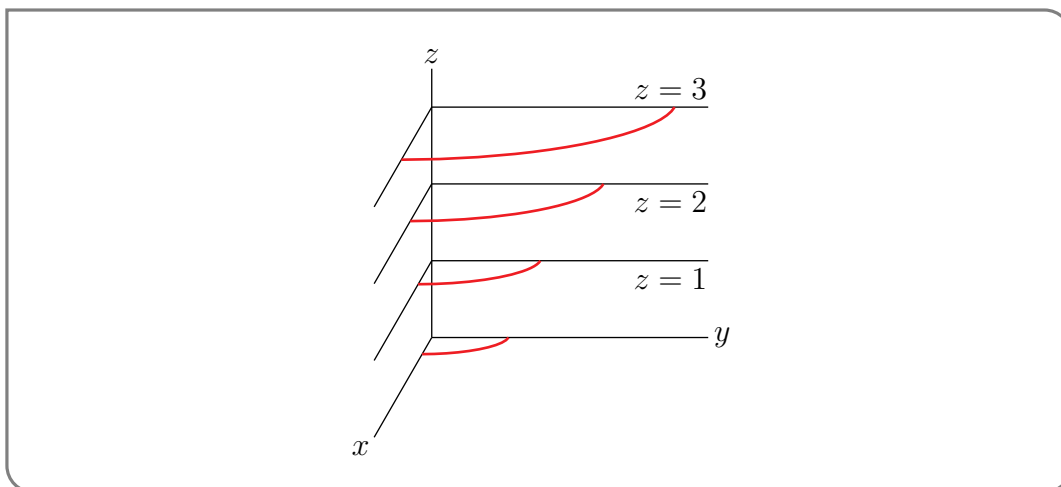
we have  $x = \pm \frac{1}{2}\sqrt{1 + z_0^2}$  and when  $x = 0$ , we have  $y = \pm\sqrt{1 + z_0^2}$ . So the curve is just an ellipse with  $x$  semi-axis  $\frac{1}{2}\sqrt{1 + z_0^2}$  and  $y$  semi-axis  $\sqrt{1 + z_0^2}$ . It's easy to sketch.



Remember that this ellipse is the part of our surface that lies in the plane  $z = z_0$ . Imagine that the sketch of the ellipse is on a single sheet of paper. Lift the sheet of paper up, move it around so that the  $x$ - and  $y$ -axes point in the directions of the three dimensional  $x$ - and  $y$ -axes and place the sheet of paper into the three dimensional sketch at height  $z_0$ . This gives a single horizontal ellipse in 3d, as in the figure below.



We can build up the full surface by stacking many of these horizontal ellipses — one for each possible height  $z_0$ . So we now draw a few of them as in the figure below. To reduce the amount of clutter in the sketch, we have only drawn the first octant (i.e. the part of three dimensions that has  $x \geq 0, y \geq 0$  and  $z \geq 0$ ).



Here is why it is OK, in this case, to just sketch the first octant. Replacing  $x$  by  $-x$  in the equation  $4x^2 + y^2 - z^2 = 1$  does not change the equation. That means that a point  $(x, y, z)$  is on the surface if and only if the point  $(-x, y, z)$  is on the surface. So the surface is invariant under reflection in the  $yz$ -plane. Similarly, the equation  $4x^2 + y^2 - z^2 = 1$  does not change when  $y$  is replaced by  $-y$  or  $z$  is replaced by  $-z$ . Our surface is also invariant reflection in the  $xz$ - and  $yz$ -planes. Once we have the part in the first octant, the remaining octants can be gotten simply by reflecting about the coordinate planes.

We can get a more visually meaningful sketch by adding in some vertical cross-sections. The  $x = 0$  and  $y = 0$  cross-sections (also called traces — they are the parts of our surface that are in the  $yz$ - and  $xz$ -planes, respectively) are

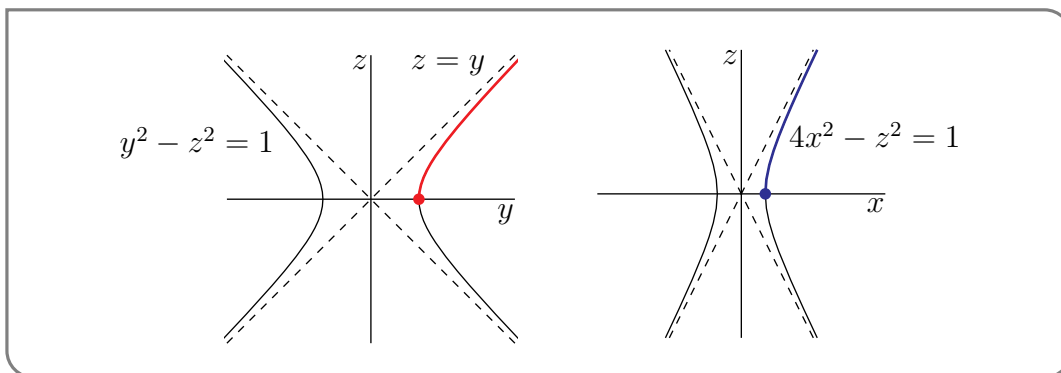
$$x = 0, y^2 - z^2 = 1 \quad \text{and} \quad y = 0, 4x^2 - z^2 = 1$$

These equations describe hyperbolae<sup>6</sup>. If you don't remember how to sketch them, don't worry. We'll do it now. We'll first sketch them in 2d. Since

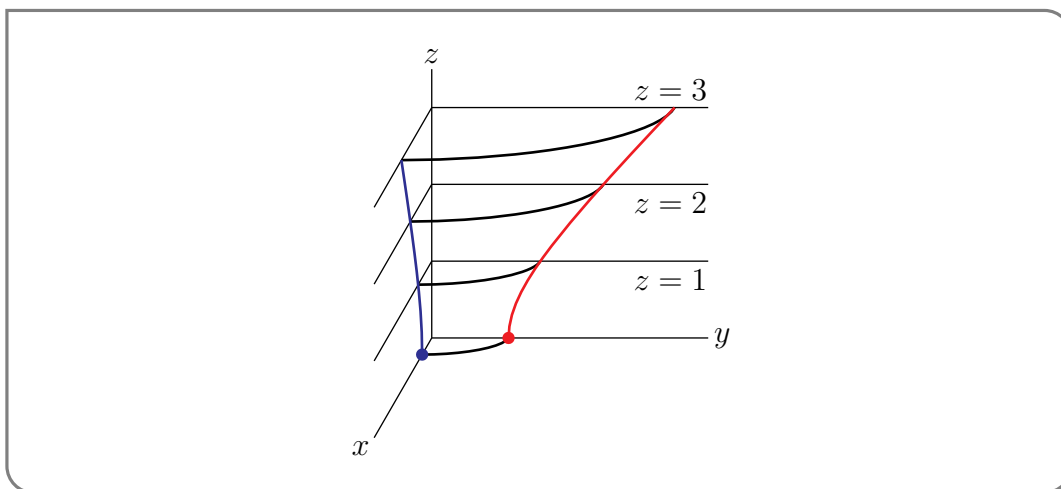
$$y^2 = 1 + z^2 \implies |y| \geq 1 \quad \text{and} \quad y = \pm 1 \text{ when } z = 0 \quad \text{and} \quad \text{for large } z, y \approx \pm z$$

$$4x^2 = 1 + z^2 \implies |x| \geq \frac{1}{2} \quad \text{and} \quad x = \pm \frac{1}{2} \text{ when } z = 0 \quad \text{and} \quad \text{for large } z, x \approx \pm \frac{1}{2}z$$

the sketches are

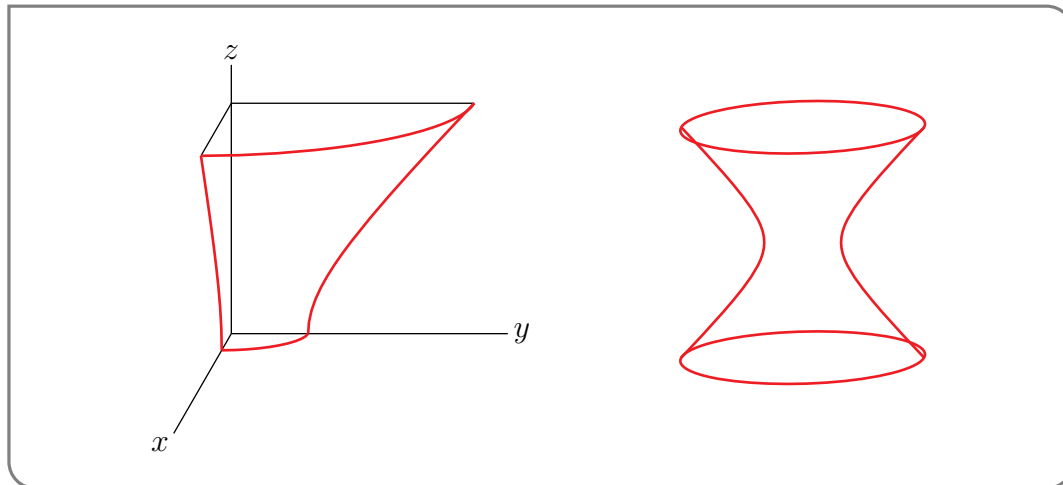


Now we'll incorporate them into the 3d sketch. Once again imagine that each is a single sheet of paper. Pick each up and move it into the 3d sketch, carefully matching up the axes. The red (blue) parts of the hyperbolas above become the red (blue) parts of the 3d sketch below (assuming of course that you are looking at this on a colour screen).



6 It's not just a figure of speech!

Now that we have a pretty good idea of what the surface looks like we can clean up and simplify the sketch. Here are a couple of possibilities.



This type of surface is called a hyperboloid of one sheet.

There are also hyperboloids of two sheets. For example, replacing the +1 on the right hand side of  $x^2 + y^2 - z^2 = 1$  gives  $x^2 + y^2 - z^2 = -1$ , which is a hyperboloid of two sheets. We'll sketch it quickly in the next example.

Example 1.3.2

Example 1.3.3 ( $4x^2 + y^2 - z^2 = -1$ )

Sketch the surface that satisfies  $4x^2 + y^2 - z^2 = -1$ .

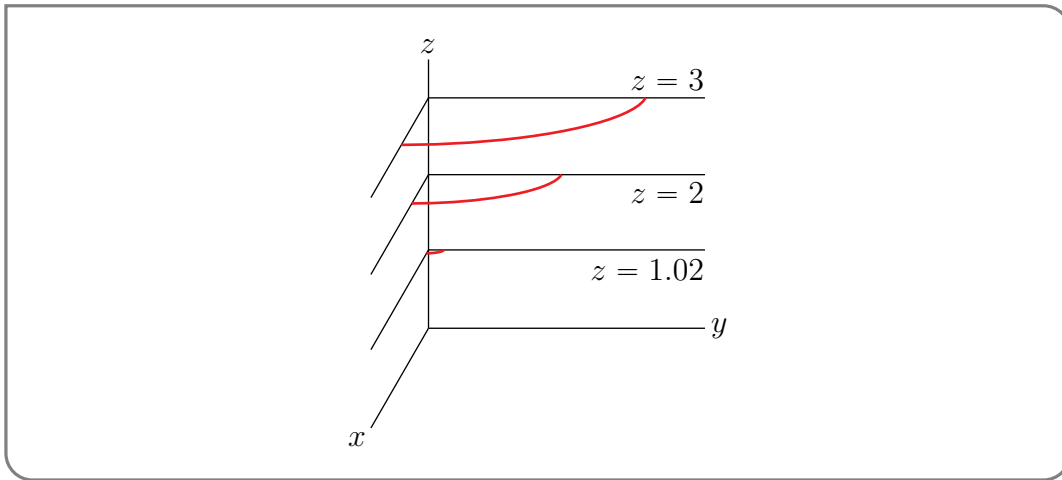
*Solution.* As in the last example, we'll start by fixing any number  $z_0$  and sketching the part of the surface that lies in the horizontal plane  $z = z_0$ . The intersection of our surface with that horizontal plane is

$$z = z_0 \text{ and } 4x^2 + y^2 = z_0^2 - 1$$

Think of  $z_0$  as a constant.

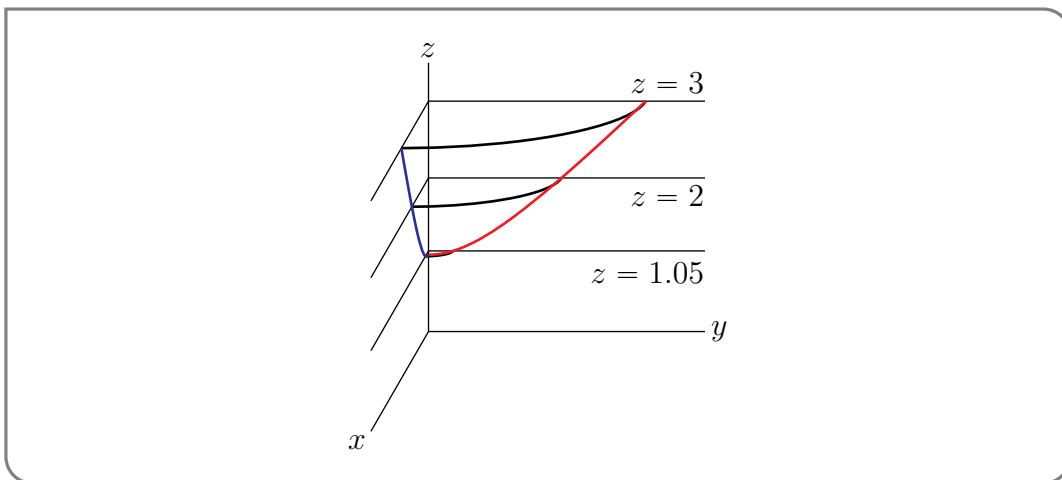
- If  $|z_0| < 1$ , then  $z_0^2 - 1 < 0$  and there are no solutions to  $x^2 + y^2 = z_0^2 - 1$ .
- If  $|z_0| = 1$  there is exactly one solution, namely  $x = y = 0$ .
- If  $|z_0| > 1$  then  $4x^2 + y^2 = z_0^2 - 1$  is an ellipse with  $x$  semi-axis  $\frac{1}{2}\sqrt{z_0^2 - 1}$  and  $y$  semi-axis  $\sqrt{z_0^2 - 1}$ . These semi-axes are small when  $|z_0|$  is close to 1 and grow as  $|z_0|$  increases.

The first octant parts of a few of these horizontal cross-sections are drawn in the figure below.

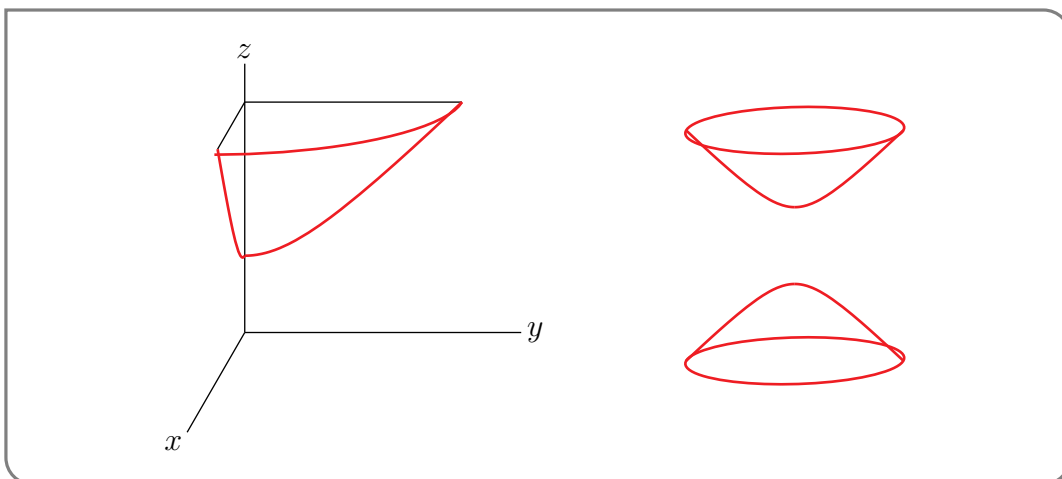


Next we add in the  $x = 0$  and  $y = 0$  cross-sections (i.e. the parts of our surface that are in the  $yz$ - and  $xz$ -planes, respectively)

$$x = 0, z^2 = 1 + y^2 \quad \text{and} \quad y = 0, z^2 = 1 + 4x^2$$



Now that we have a pretty good idea of what the surface looks like we clean up and simplify the sketch.



This type of surface is called a hyperboloid of two sheets.

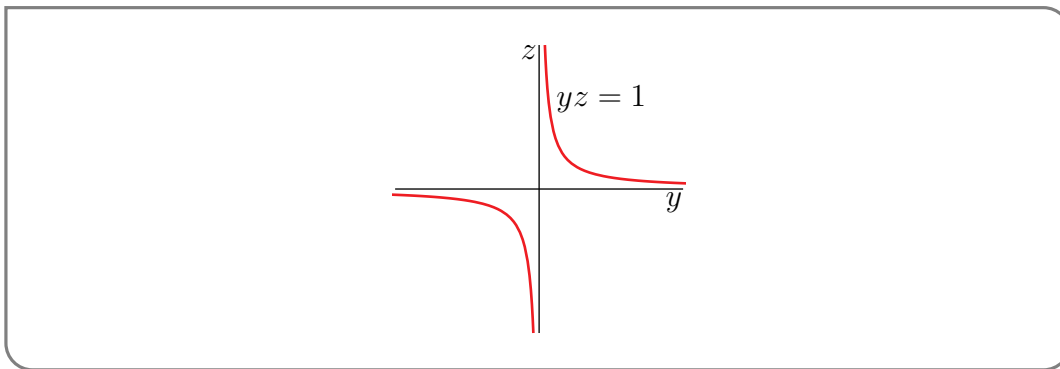
Example 1.3.3

Example 1.3.4 ( $yz = 1$ )

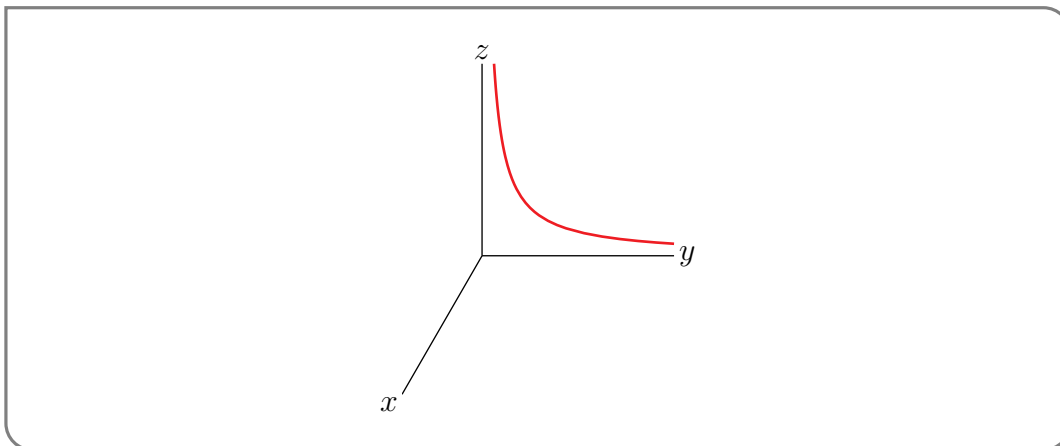
Sketch the surface  $yz = 1$ .

*Solution.* This surface has a special property that makes it relatively easy to sketch. There are no  $x$ 's in the equation  $yz = 1$ . That means that if some  $y_0$  and  $z_0$  obey  $y_0z_0 = 1$ , then the point  $(x, y_0, z_0)$  lies on the surface  $yz = 1$  for all values of  $x$ . As  $x$  runs from  $-\infty$  to  $\infty$ , the point  $(x, y_0, z_0)$  sweeps out a straight line parallel to the  $x$ -axis. So the surface  $yz = 1$  is a union of lines parallel to the  $x$ -axis. It is invariant under translations parallel to the  $x$ -axis. To sketch  $yz = 1$ , we just need to sketch its intersection with the  $yz$ -plane and then translate the resulting curve parallel to the  $x$ -axis to sweep out the surface.

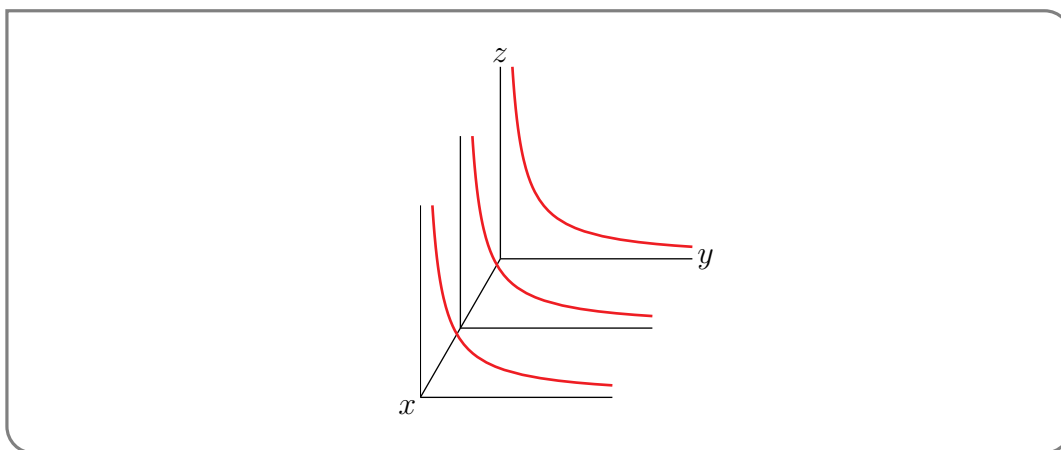
We'll start with a sketch of the hyperbola  $yz = 1$  in two dimensions.



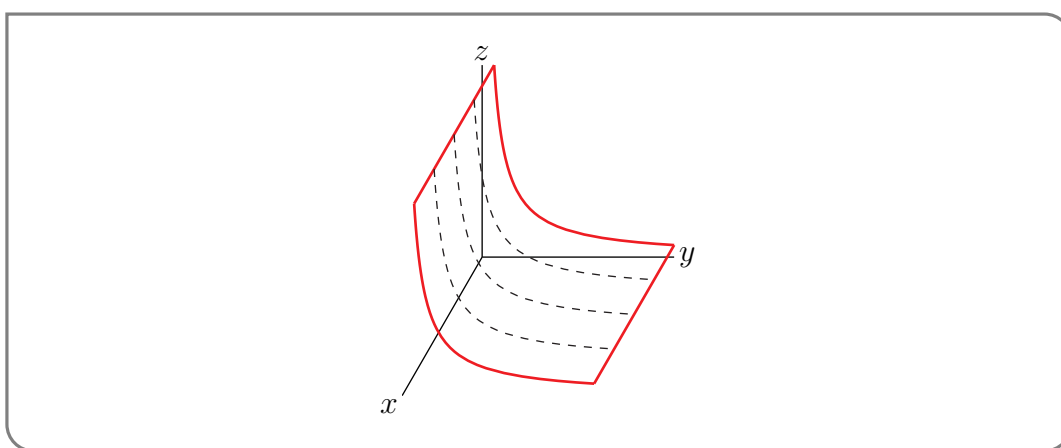
Next we'll move this 2d sketch into the  $yz$ -plane, i.e. the plane  $x = 0$ , in 3d, except that we'll only draw in the part in the first octant.



The we'll draw in  $x = x_0$  cross-sections for a couple of more values of  $x_0$



and clean up the sketch a bit



Example 1.3.4

Example 1.3.5 ( $xyz = 1$ )

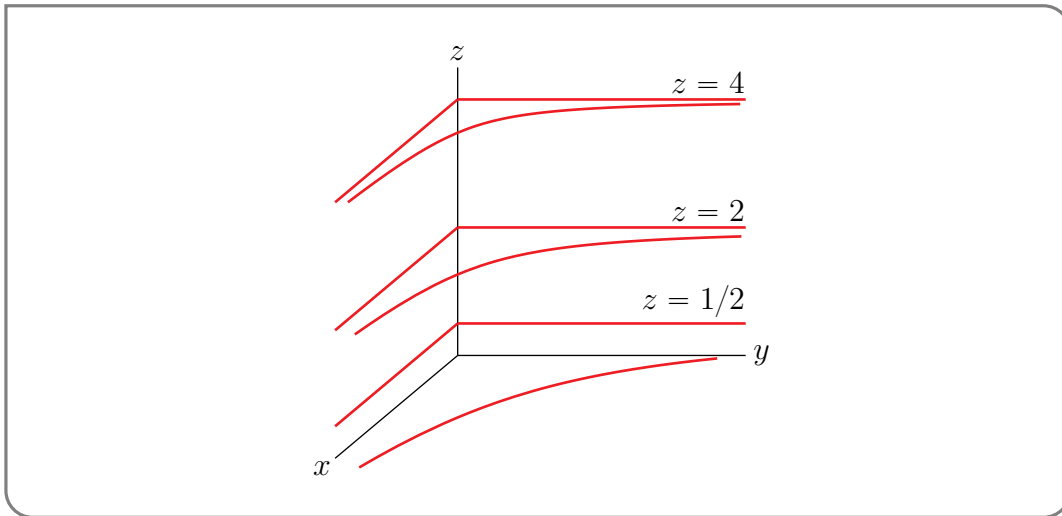
Sketch the surface  $xyz = 4$ .

*Solution.* We'll sketch this surface using much the same procedure as we used in Examples 1.3.2 and 1.3.3. We'll only sketch the part of the surface in the first octant. The remaining parts (in the octants with  $x, y < 0, z \geq 0$ , with  $x, z < 0, y \geq 0$  and with  $y, z < 0, x \geq 0$ ) are just reflections of the first octant part.

As usual, we start by fixing any number  $z_0$  and sketching the part of the surface that lies in the horizontal plane  $z = z_0$ . The intersection of our surface with that horizontal plane is the hyperbola

$$z = z_0 \text{ and } xy = \frac{1}{z_0}$$

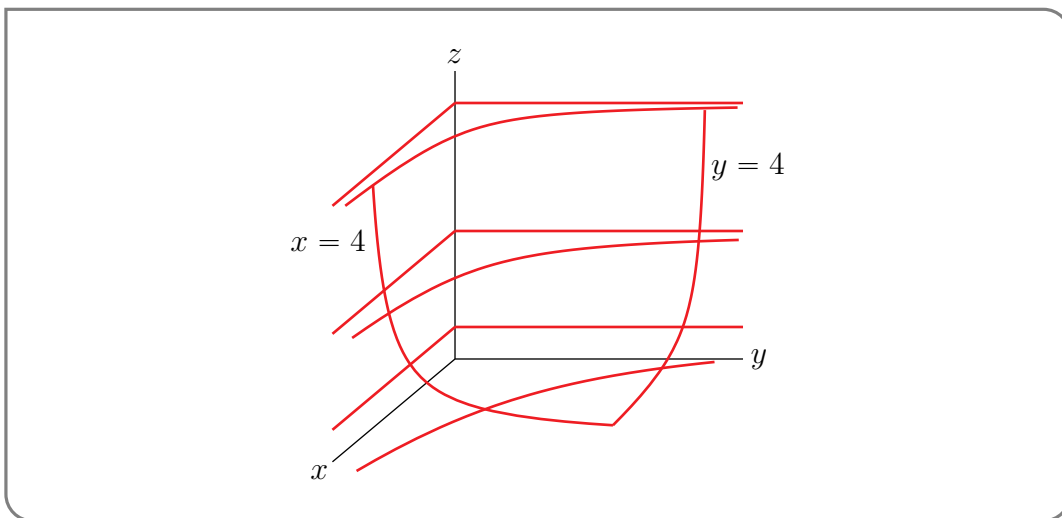
Note that  $x \rightarrow \infty$  as  $y \rightarrow 0$  and that  $y \rightarrow \infty$  as  $x \rightarrow 0$ . So the hyperbola has both the  $x$ -axis and the  $y$ -axis as asymptotes, when drawn in the  $xy$ -plane. The first octant parts of a few of these horizontal cross-sections (namely,  $z_0 = 4, z_0 = 2$  and  $z_0 = \frac{1}{2}$ ) are drawn in the figure below.



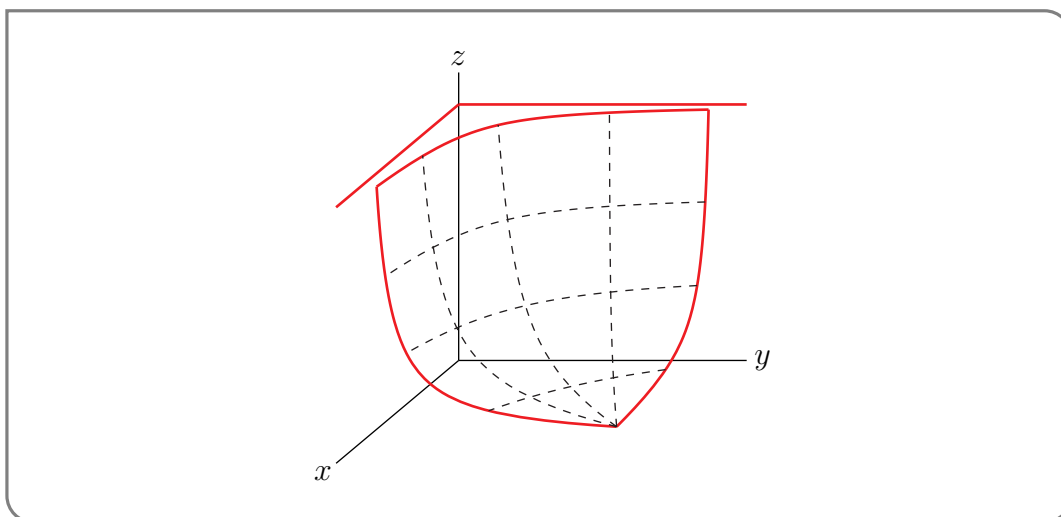
Next we add some vertical cross-sections. We can't use  $x = 0$  or  $y = 0$  because any point on  $xyz = 1$  must have all of  $x, y, z$  nonzero. So we use

$$x = 4, yz = 1 \quad \text{and} \quad y = 4, xz = 1$$

instead. They are again hyperbolae.



Finally, we clean up and simplify the sketch.



Example 1.3.5

Often the reason you are interested in a surface in 3d is that it is the graph  $z = f(x, y)$  of a function of two variables  $f(x, y)$ . Another good way to visualize the behaviour of a function  $f(x, y)$  is to sketch what are called its level curves.

**Definition 1.3.6.**

A level curve of  $f(x, y)$  is a curve whose equation is  $f(x, y) = C$ , for some constant  $C$ .

A level curve is the set of points in the  $xy$ -plane where  $f$  takes the value  $C$ . Because it is a curve in 2d, it is usually easier to sketch than the graph of  $f$ . Here are a couple of examples.

**Example 1.3.7** ( $f(x, y) = x^2 + 4y^2 - 2x + 2$ )

Sketch the level curves of  $f(x, y) = x^2 + 4y^2 - 2x + 2$ .

*Solution.* Fix any real number  $C$ . Then, for the specified function  $f$ , the level curve  $f(x, y) = C$  is the set of points  $(x, y)$  that obey

$$\begin{aligned} x^2 + 4y^2 - 2x + 2 = C &\iff x^2 - 2x + 1 + 4y^2 + 1 = C \\ &\iff (x - 1)^2 + 4y^2 = C - 1 \end{aligned}$$

Now  $(x - 1)^2 + 4y^2$  is the sum of two squares, and so is always at least zero. So if  $C - 1 < 0$ , i.e. if  $C < 1$ , there is no curve  $f(x, y) = C$ . If  $C - 1 = 0$ , i.e. if  $C = 1$ , then  $f(x, y) = C - 1 = 0$  if and only if both  $(x - 1)^2 = 0$  and  $4y^2 = 0$  and so the level curve consists of the single point  $(1, 0)$ . If  $C > 1$ , then  $f(x, y) = C$  become  $(x - 1)^2 + 4y^2 = C - 1 > 0$  which describes an ellipse centred on  $(1, 0)$ . It intersects the  $x$ -axis when  $y = 0$  and

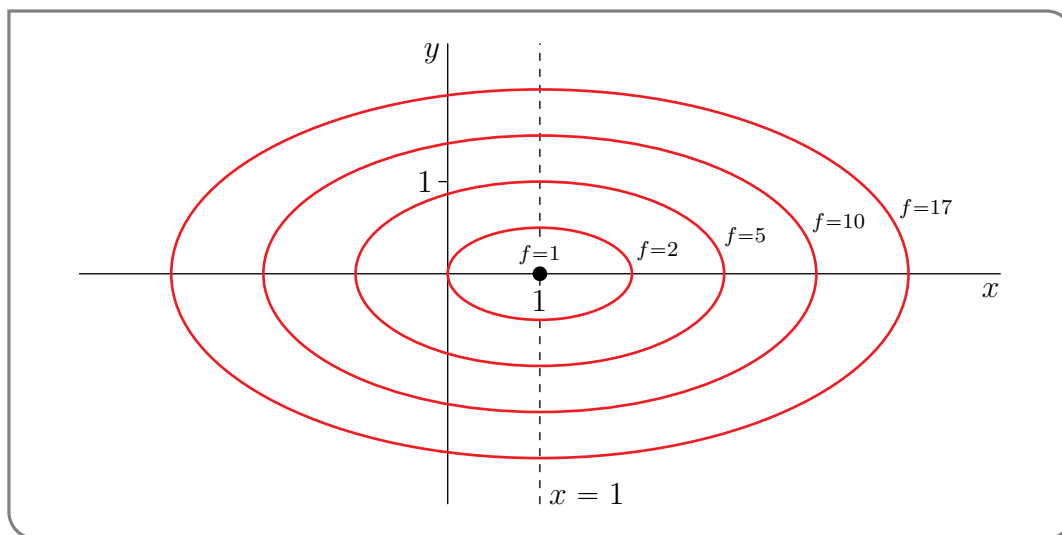
$$(x - 1)^2 = C - 1 \iff x - 1 = \pm\sqrt{C - 1} \iff x = 1 \pm \sqrt{C - 1}$$



and it intersects the line  $x = 1$  (i.e. the vertical line through the centre) when

$$4y^2 = C - 1 \iff 2y = \pm\sqrt{C - 1} \iff y = \pm\frac{1}{2}\sqrt{C - 1}$$

So, when  $C > 1$ ,  $f(x, y) = C$  is the ellipse centred on  $(1, 0)$  with  $x$  semi-axis  $\sqrt{C - 1}$  and  $y$  semi-axis  $\frac{1}{2}\sqrt{C - 1}$ . Here is a sketch of some representative level curves of  $f(x, y) = x^2 + 4y^2 - 2x + 2$ .

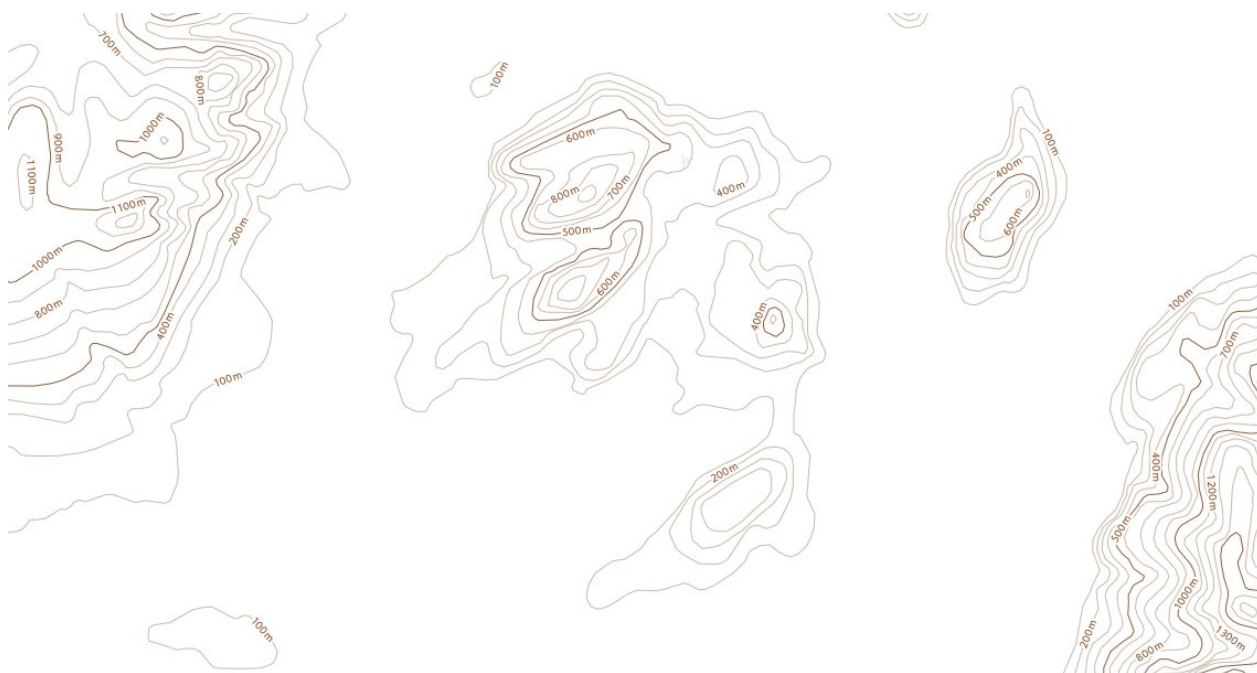


It is often easier to develop an understanding of the behaviour of a function  $f(x, y)$  by looking at a sketch of its level curves, than it is by looking at a sketch of its graph. On the other hand, you can also use a sketch of the level curves of  $f(x, y)$  as the first step in building a sketch of the graph  $z = f(x, y)$ . The next step would be to redraw, for each  $C$ , the level curve  $f(x, y) = C$ , in the plane  $z = C$ , as we did in Example 1.3.2.

Example 1.3.7

If you've ever used a topographic map, you've seen examples of level curves. Modelling the  $z$ -axis as a measure of elevation, with  $z = 0$  as sea level, the contours shown on topographic maps show the level curves associated with different elevations. The example<sup>7</sup> below shows the area around Gambier, Anvil, and Keats Islands, north of UBC. The lines show level curves for  $z = 0$  metres,  $z = 100$  metres,  $z = 200$  metres, etc.

<sup>7</sup> generated by Natural Resources Canada's [Atlas of Canada - Toporama](#), included under an [open government license](#)



Example 1.3.8 ( $e^{x+y+z} = 1$ )

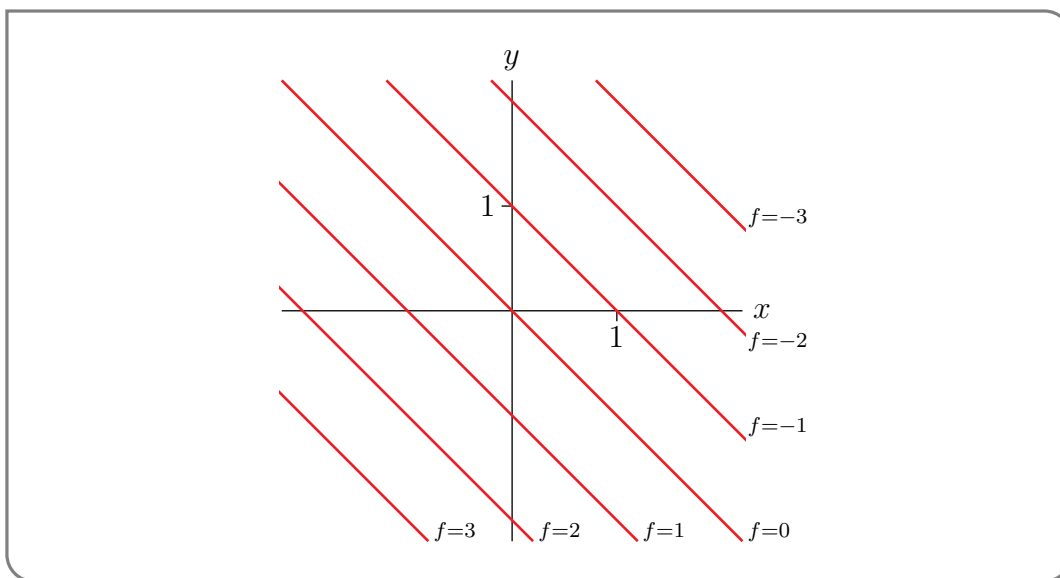
The function  $f(x, y)$  is given implicitly by the equation  $e^{x+y+z} = 1$ . Sketch the level curves of  $f$ .

*Solution.* This one is not as nasty as it appears. That “ $f(x, y)$  is given implicitly by the equation  $e^{x+y+z} = 1$ ” means that, for each  $x, y$ , the solution  $z$  of  $e^{x+y+z} = 1$  is  $f(x, y)$ . So, for the specified function  $f$  and any fixed real number  $C$ , the level curve  $f(x, y) = C$  is the set of points  $(x, y)$  that obey

$$e^{x+y+C} = 1 \iff x + y + C = 0 \quad (\text{by taking the ln of both sides})$$

$$\iff x + y = -C$$

This is of course a straight line. It intersects the  $x$ -axis when  $y = 0$  and  $x = -C$  and it intersects the  $y$ -axis when  $x = 0$  and  $y = -C$ . Here is a sketch of some level curves.

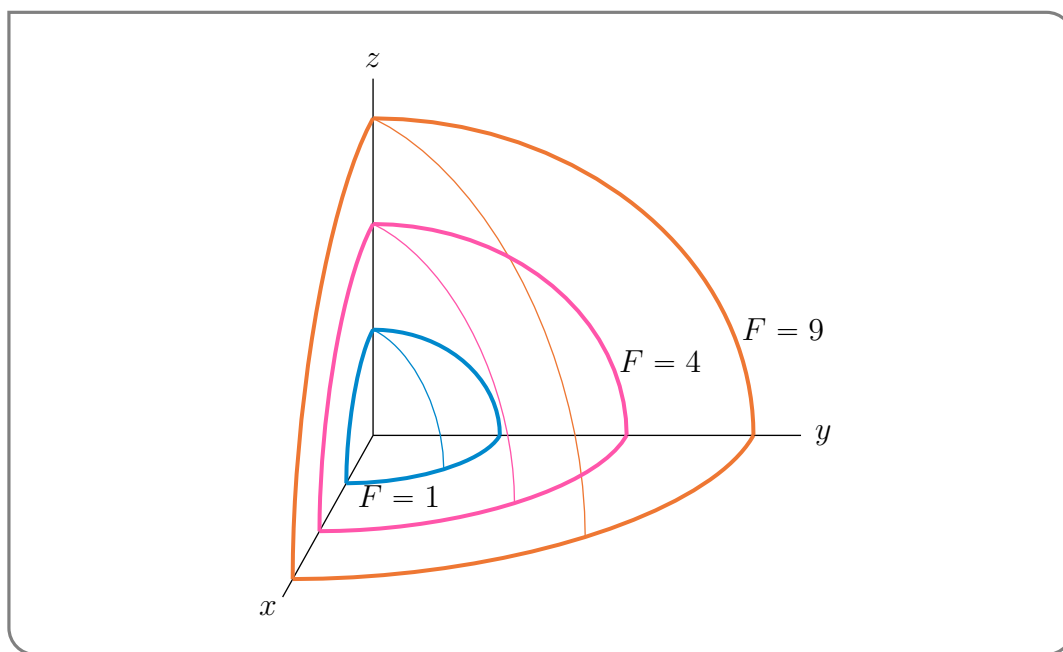


Example 1.3.8

We have just seen that sketching the level curves of a function  $f(x, y)$  can help us understand the behaviour of  $f$ . We can generalise this to functions  $F(x, y, z)$  of three variables. A level surface of  $F(x, y, z)$  is a surface whose equation is of the form  $F(x, y, z) = C$  for some constant  $C$ . It is the set of points  $(x, y, z)$  at which  $F$  takes the value  $C$ .

Example 1.3.9 ( $F(x, y, z) = x^2 + y^2 + z^2$ )

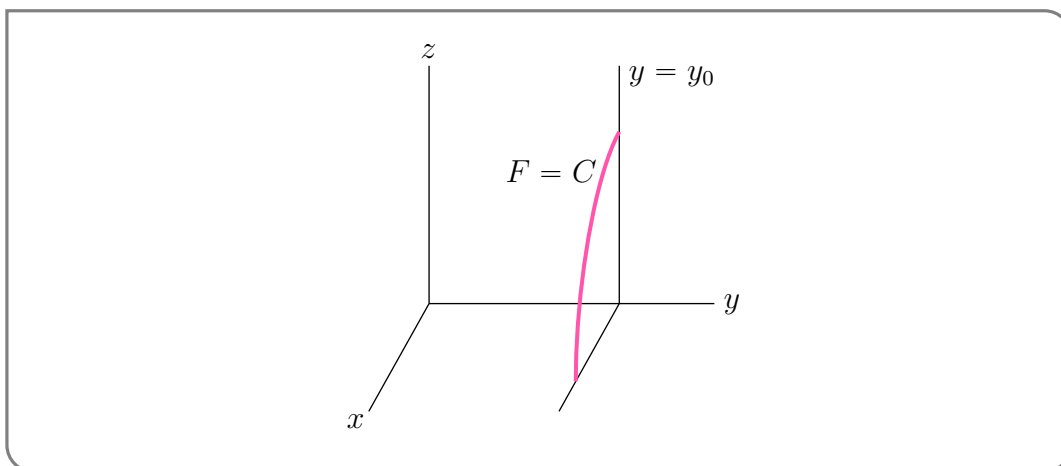
Let  $F(x, y, z) = x^2 + y^2 + z^2$ . If  $C > 0$ , then the level surface  $F(x, y, z) = C$  is the sphere of radius  $\sqrt{C}$  centred on the origin. Here is a sketch of the parts of the level surfaces  $F = 1$  (radius 1),  $F = 4$  (radius 2) and  $F = 9$  (radius 3) that are in the first octant.



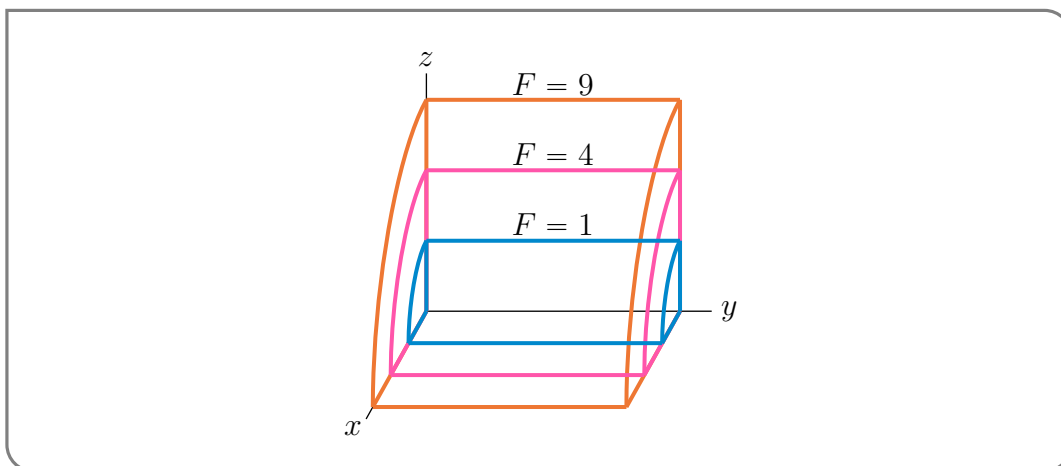
Example 1.3.9

Example 1.3.10 ( $F(x, y, z) = x^2 + z^2$ )

Let  $F(x, y, z) = x^2 + z^2$  and  $C > 0$ . Consider the level surface  $x^2 + z^2 = C$ . The variable  $y$  does not appear in this equation. So for any fixed  $y_0$ , the intersection of our surface  $x^2 + z^2 = C$  with the plane  $y = y_0$  is the circle of radius  $\sqrt{C}$  centred on  $x = z = 0$ . Here is a sketch of the first quadrant part of one such circle.



The full surface is the horizontal stack of all of those circles with  $y_0$  running over  $\mathbb{R}$ . It is the cylinder of radius  $\sqrt{C}$  centred on the  $y$ -axis. Here is a sketch of the parts of the level surfaces  $F = 1$  (radius 1),  $F = 4$  (radius 2) and  $F = 9$  (radius 3) that are in the first octant.



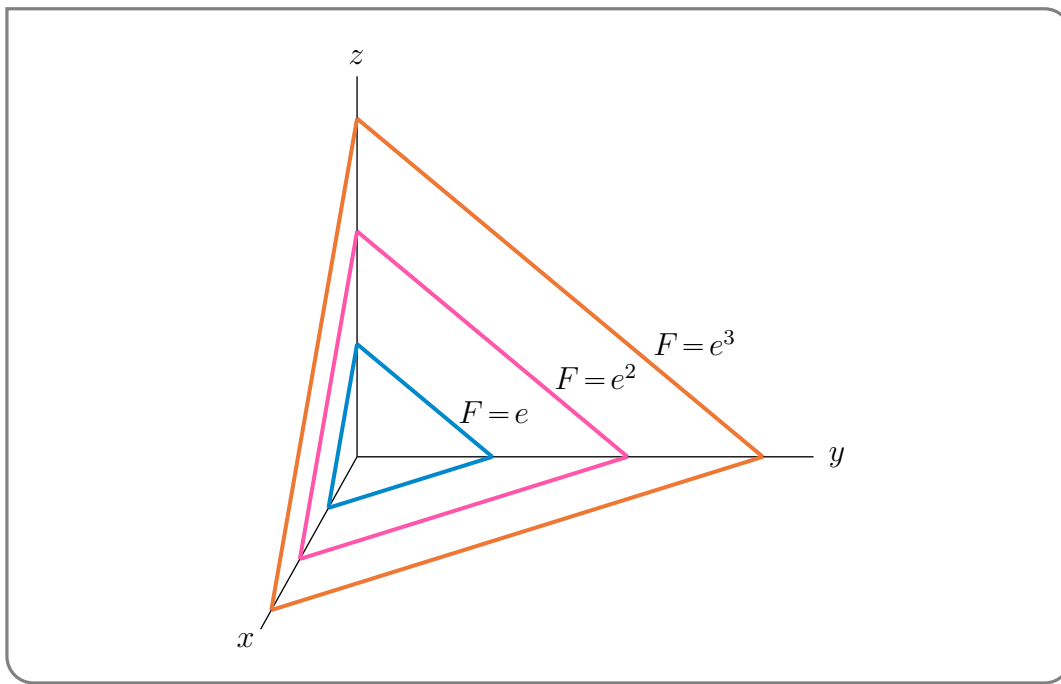
Example 1.3.10

Example 1.3.11 ( $F(x, y, z) = e^{x+y+z}$ )

Let  $F(x, y, z) = e^{x+y+z}$  and  $C > 0$ . Consider the level surface  $e^{x+y+z} = C$ , or equivalently,  $x + y + z = \ln C$ . It is the plane that contains the intercepts  $(\ln C, 0, 0)$ ,  $(0, \ln C, 0)$  and  $(0, 0, \ln C)$ . Here is a sketch of the parts of the level surfaces

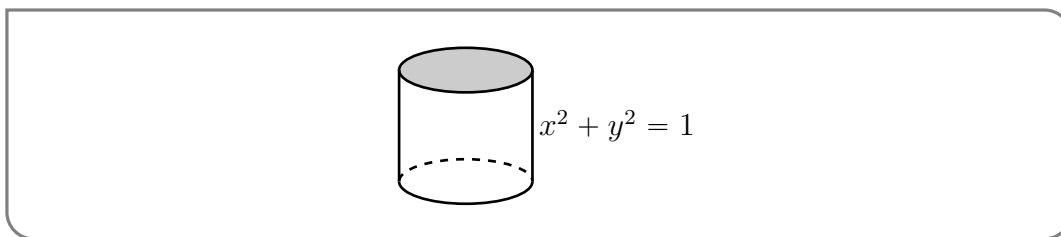
- $F = e$  (intercepts  $(1, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$ ),
- $F = e^2$  (intercepts  $(2, 0, 0)$ ,  $(0, 2, 0)$ ,  $(0, 0, 2)$ ) and
- $F = e^3$  (intercepts  $(3, 0, 0)$ ,  $(0, 3, 0)$ ,  $(0, 0, 3)$ )

that are in the first octant.



Example 1.3.11

There some classes of relatively simple, but commonly occurring, surfaces that are given their own names. One such class is cylindrical surfaces. You are probably used to thinking of a cylinder as being something that looks like  $x^2 + y^2 = 1$ .



In Mathematics the word “cylinder” is given a more general meaning.

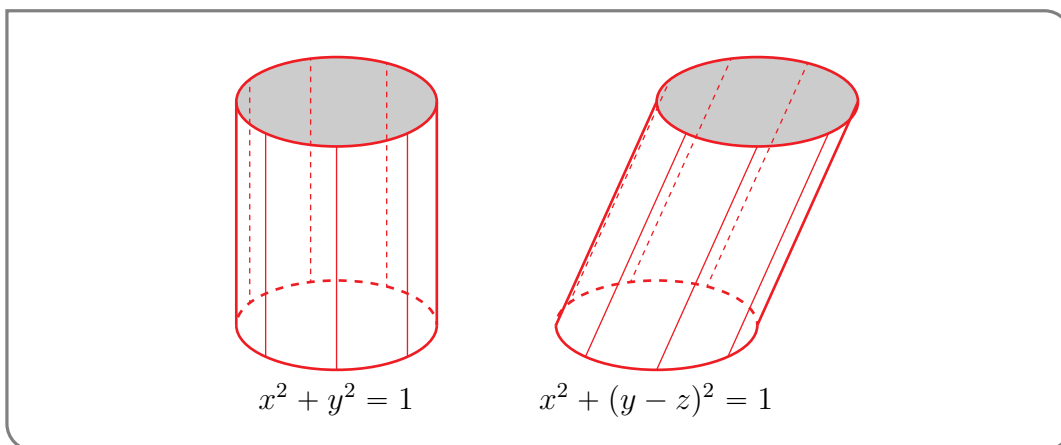
**Definition 1.3.12 (Cylinder).**

A *cylinder* is a surface that consists of all points that are on all lines that are

- parallel to a given line and
- pass through a given fixed plane curve (in a plane not parallel to the given line).

Example 1.3.13

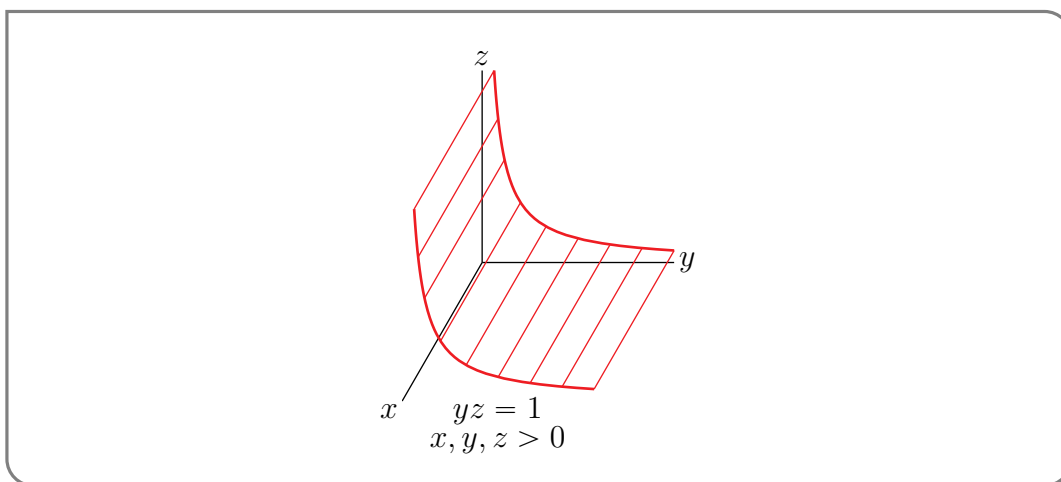
Here are sketches of three cylinders. The familiar cylinder on the left below



is called a right circular cylinder, because the given fixed plane curve ( $x^2 + y^2 = 1, z = 0$ ) is a circle and the given line (the  $z$ -axis) is perpendicular (i.e. at right angles) to the fixed plane curve.

The cylinder on the left above can be thought of as a vertical stack of circles. The cylinder on the right above can also be thought of as a stack of circles, but the centre of the circle at height  $z$  has been shifted rightward to  $(0, z, z)$ . For that cylinder, the given fixed plane curve is once again the circle  $x^2 + y^2 = 1, z = 0$ , but the given line is  $y = z, x = 0$ .

We have already seen the third cylinder



in Example 1.3.4. It is called a hyperbolic cylinder. In this example, the given fixed plane curve is the hyperbola  $yz = 1, x = 0$  and the given line is the  $x$ -axis.

Example 1.3.13

### 1.3.1 ► Quadric Surfaces

Another named class of relatively simple, but commonly occurring, surfaces is the quadric surfaces.

**Definition 1.3.14 (Quadrics).**

A *quadric* surface is surface that consists of all points that obey  $Q(x, y, z) = 0$ , with  $Q$  being a polynomial of degree two<sup>8</sup>.

For  $Q(x, y, z)$  to be a polynomial of degree two, it must be of the form

$$Q(x, y, z) = Ax^2 + By^2 + Cz^2 + Dxy + Eyz + Fxz + Gx + Hy + Iz + J$$

for some constants  $A, B, \dots, J$ . Each constant  $z$  cross section of a quadric surface has an equation of the form

$$Ax^2 + Dxy + By^2 + gx + hy + j = 0, \quad z = z_0$$

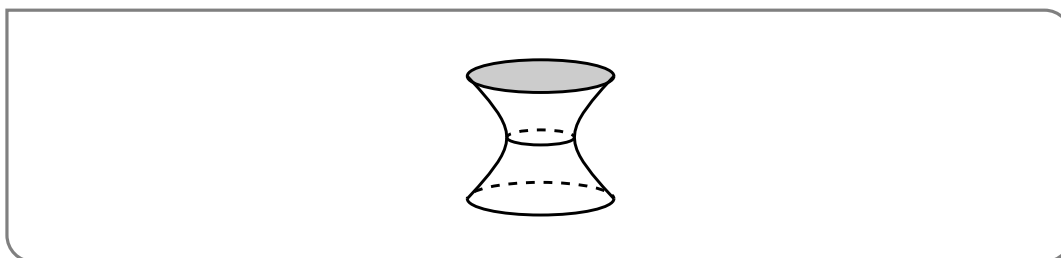
If  $A = B = D = 0$  but  $g$  and  $h$  are not both zero, this is a straight line. If  $A, B,$  and  $D$  are not all zero, then by rotating and translating our coordinate system the equation of the cross section can be brought into one of the forms<sup>9</sup>

- $\alpha x^2 + \beta y^2 = \gamma$  with  $\alpha, \beta > 0$ , which, if  $\gamma > 0$ , is an ellipse (or a circle),
- $\alpha x^2 - \beta y^2 = \gamma$  with  $\alpha, \beta > 0$ , which, if  $\gamma \neq 0$ , is a hyperbola, and if  $\gamma = 0$  is two lines,
- $x^2 = \delta y$ , which, if  $\delta \neq 0$  is a parabola, and if  $\delta = 0$  is a straight line.

There are similar statements for the constant  $y$  cross sections and the constant  $z$  cross sections. Hence quadratic surfaces are built by stacking these three types of curves.

We have already seen a number of quadric surfaces in the last couple of sections.

- We saw the quadric surface  $4x^2 + y^2 - z^2 = 1$  in Example 1.3.2.



Its constant  $z$  cross sections are ellipses and its  $x = 0$  and  $y = 0$  cross sections are hyperbolae. It is called a hyperboloid of one sheet.

- We saw the quadric surface  $x^2 + y^2 = 1$  in Example 1.3.13.

<sup>8</sup> Technically, we should also require that the polynomial can't be factored into the product of two polynomials of degree one.

<sup>9</sup> This statement can be justified using a linear algebra eigenvalue/eigenvector analysis. It is beyond what we can cover here, but is not too difficult for a standard linear algebra course.



Its constant  $z$  cross sections are circles and its  $x = 0$  and  $y = 0$  cross sections are straight lines. It is called a right circular cylinder.

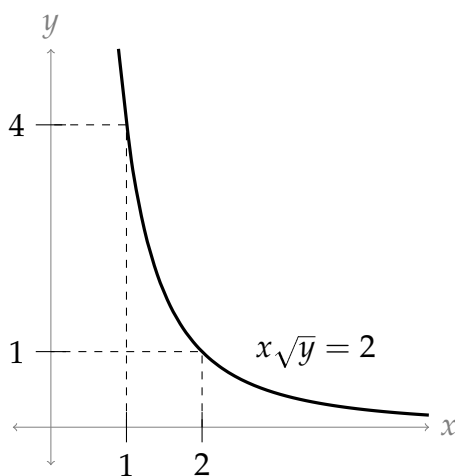
- the quadric surface  $x^2 + (y - z)^2 = 1$  in Example 1.3.13, and
- We saw the quadric surface  $yz = 1$  in Example 1.3.4.

Appendix A.3 contains other quadric surfaces.

Example 1.3.15 (Indifference curves)

Suppose a function  $U(x, y)$  gives the happiness<sup>10</sup> (or *utility*) a consumer gains when they purchase  $x$  units of Good X and  $y$  units of Good Y. The level curves of the surface  $z = U(x, y)$  are called *indifference curves*, because every point along that curve results in the same benefit to the consumer.

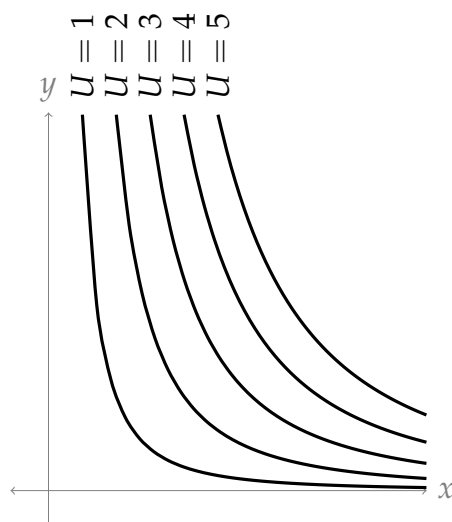
Suppose  $U(x, y) = x\sqrt{y}$ . The purchasing 2 units of Good X and one unit of Good Y produces the same benefit as purchasing 1 unit of Good X and 4 units of Good Y, because both these combinations are on the level curve  $U(x, y) = 2$ .



Let's make a small contour map of our surface  $U(x, y) = x\sqrt{y}$ , plotting several indifference curves. (Note  $x\sqrt{y} = c$  is equivalent to  $y = \frac{c^2}{x^2}$  in our model domain.)

10 An amusing thought experiment is to propose units for measuring happiness. "The one-point increase in GDP was associated with an average increase of 3.7 wrinkly puppy faces of happiness nation-wide."





Not surprisingly, if we move roughly in the direction of the  $(1, 1)$  (that is, increasing both  $x$  and  $y$ ), our happiness  $U(x, y)$  goes up.

Note that none of the indifference curves touch either of the  $x$  or  $y$  axes. It is clear enough from the formula that  $U(0, y) = U(x, 0) = 0$ . This is a common feature of utility functions: that to maximize utility, a consumer will have at least a little of both products, rather than consuming only one type.

Example 1.3.15

Chapter 1 (excluding Section 1.2) was adapted from Chapter 1 of [CLP-3 Multivariable Calculus](#) by Feldman, Reznitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

# PARTIAL DERIVATIVES

In this chapter we are going to generalize the definition of “derivative” to functions of more than one variable, and then we are going to use those derivatives. We can speed things up considerably by recycling what we have already learned in the single-variable case.

## 2.1▲ Partial Derivatives

First, recall how we defined the derivative,  $f'(a)$ , of a function of one variable,  $f(x)$ . We imagined that we were walking along the  $x$ -axis, in the positive direction, measuring, for example, the temperature along the way. We denoted by  $f(x)$  the temperature at  $x$ . The instantaneous rate of change of temperature that we observed as we passed through  $x = a$  was

$$\frac{df}{dx}(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

Next suppose that we are walking in the  $xy$ -plane and that the temperature at  $(x, y)$  is  $f(x, y)$ . We can pass through the point  $(x, y) = (a, b)$  moving in many different directions, and we cannot expect the measured rate of change of temperature if we walk parallel to the  $x$ -axis, in the direction of increasing  $x$ , to be the same as the measured rate of change of temperature if we walk parallel to the  $y$ -axis in the direction of increasing  $y$ . We'll start by considering just those two directions. other directions (like walking parallel to the line  $y = x$ ) later.

Suppose that we are passing through the point  $(x, y) = (a, b)$  and that we are walking parallel to the  $x$ -axis (in the positive direction). Then our  $y$ -coordinate will be constant, always taking the value  $y = b$ . So we can think of the measured temperature as the function of one variable  $B(x) = f(x, b)$  and we will observe the rate of change of temperature

$$\frac{dB}{dx}(a) = \lim_{h \rightarrow 0} \frac{B(a+h) - B(a)}{h} = \lim_{h \rightarrow 0} \frac{f(a+h, b) - f(a, b)}{h}$$

This is called the “partial derivative  $f$  with respect to  $x$  at  $(a, b)$ ” and is denoted  $(\frac{\partial f}{\partial x})_y(a, b)$ .

Here

- the symbol  $\partial$ , which is read “partial”, indicates that we are dealing with a function of more than one variable and
- the subscript  $y$  on  $(\ )_y$  indicates that  $y$  is being held fixed, i.e. being treated as a constant, and
- the  $x$  in  $\frac{\partial f}{\partial x}$  indicates that we are differentiating with respect to  $x$ .
- $\frac{\partial f}{\partial x}$  is read “partial dee  $f$  dee  $x$ ”.

Do not write  $\frac{d}{dx}$  when  $\frac{\partial}{\partial x}$  is appropriate. (There exist situations when  $\frac{d}{dx}f$  and  $\frac{\partial}{\partial x}f$  are both defined and have different meanings.)

If, instead, we are passing through the point  $(x, y) = (a, b)$  and are walking parallel to the  $y$ -axis (in the positive direction), then our  $x$ -coordinate will be constant, always taking the value  $x = a$ . So we can think of the measured temperature as the function of one variable  $A(y) = f(a, y)$  and we will observe the rate of change of temperature

$$\frac{dA}{dy}(b) = \lim_{h \rightarrow 0} \frac{A(b+h) - A(b)}{h} = \lim_{h \rightarrow 0} \frac{f(a, b+h) - f(a, b)}{h}$$

This is called the “partial derivative  $f$  with respect to  $y$  at  $(a, b)$ ” and is denoted  $(\frac{\partial f}{\partial y})_x(a, b)$ .

Just as was the case for the ordinary derivative  $\frac{df}{dx}(x)$ , it is common to treat the partial derivatives of  $f(x, y)$  as functions of  $(x, y)$  simply by evaluating the partial derivatives at  $(x, y)$  rather than at  $(a, b)$ .

### Definition 2.1.1 (Partial Derivatives).

The  $x$ - and  $y$ -partial derivatives of the function  $f(x, y)$  are

$$\left(\frac{\partial f}{\partial x}\right)_y(x, y) = \lim_{h \rightarrow 0} \frac{f(x+h, y) - f(x, y)}{h}$$

$$\left(\frac{\partial f}{\partial y}\right)_x(x, y) = \lim_{h \rightarrow 0} \frac{f(x, y+h) - f(x, y)}{h}$$

respectively. The partial derivatives of functions of more than two variables are defined analogously.

Partial derivatives are used a lot. And there many notations for them.

**Notation 2.1.2.**

The partial derivative  $\left(\frac{\partial f}{\partial x}\right)_y$  of a function  $f(x, y)$  is also denoted

$$\frac{\partial f}{\partial x} \quad f_x \quad D_x f \quad D_1 f$$

The subscript 1 on  $D_1 f$  indicates that  $f$  is being differentiated with respect to its first variable. The partial derivative  $\left(\frac{\partial f}{\partial x}\right)_y(a, b)$  is also denoted

$$\frac{\partial f}{\partial x} \Big|_{(a,b)}$$

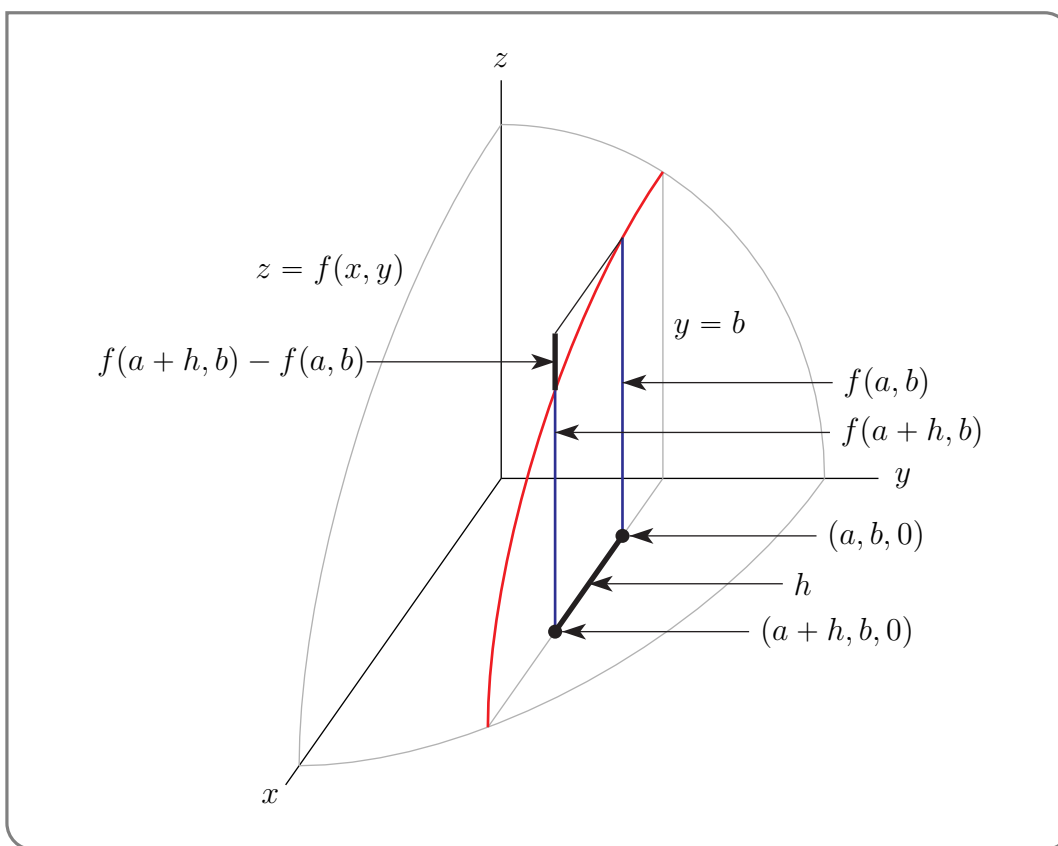
with the subscript  $(a, b)$  indicating that  $\frac{\partial f}{\partial x}$  is being evaluated at  $(x, y) = (a, b)$ . The abbreviated notation  $\frac{\partial f}{\partial x}$  for  $\left(\frac{\partial f}{\partial x}\right)_y$  is extremely commonly used. But it is dangerous to do so, when it is not clear from the context, that it is the variable  $y$  that is being held fixed.

**Remark 2.1.3** (The Geometric Interpretation of Partial Derivatives). We'll now develop a geometric interpretation of the partial derivative

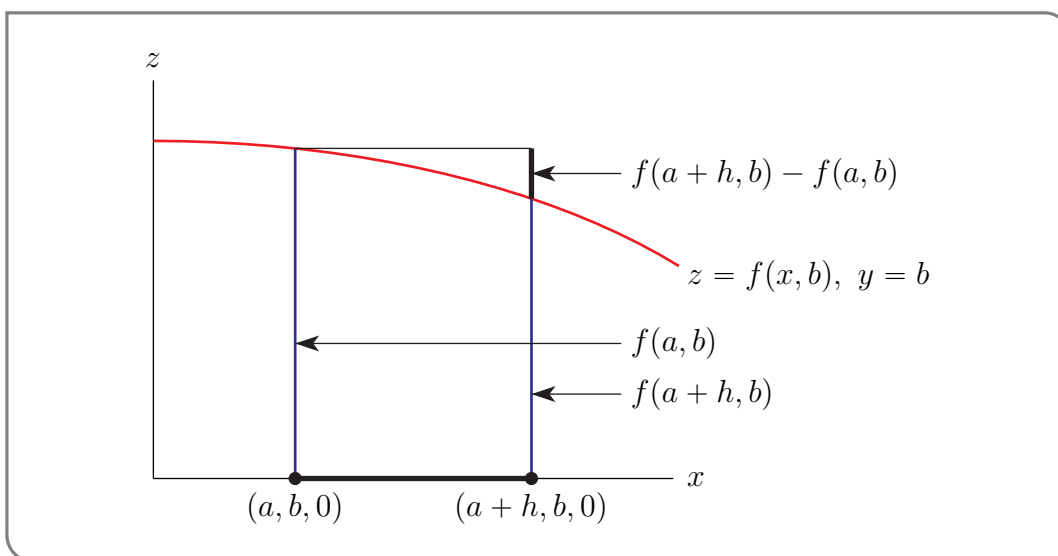
$$\left(\frac{\partial f}{\partial x}\right)_y(a, b) = \lim_{h \rightarrow 0} \frac{f(a+h, b) - f(a, b)}{h}$$

in terms of the shape of the graph  $z = f(x, y)$  of the function  $f(x, y)$ . That graph appears in the figure below. It looks like the part of a deformed sphere that is in the first octant.

The definition of  $\left(\frac{\partial f}{\partial x}\right)_y(a, b)$  concerns only points on the graph that have  $y = b$ . In other words, the curve of intersection of the surface  $z = f(x, y)$  with the plane  $y = b$ . That is the red curve in the figure. The two blue vertical line segments in the figure have heights  $f(a, b)$  and  $f(a+h, b)$ , which are the two numbers in the numerator of  $\frac{f(a+h, b) - f(a, b)}{h}$ .



A side view of the curve (looking from the left side of the  $y$ -axis) is sketched in the figure below. Again, the two blue vertical line segments in the figure have heights  $f(a, b)$



and  $f(a + h, b)$ , which are the two numbers in the numerator of  $\frac{f(a+h,b)-f(a,b)}{h}$ . So the numerator  $f(a + h, b) - f(a, b)$  and denominator  $h$  are the rise and run, respectively, of the curve  $z = f(x, b)$  from  $x = a$  to  $x = a + h$ . Thus  $\left(\frac{\partial f}{\partial x}\right)_y(a, b)$  is exactly the slope of (the tangent to) the curve of intersection of the surface  $z = f(x, y)$  and the plane  $y = b$  at the point  $(a, b, f(a, b))$ . In the same way  $\left(\frac{\partial f}{\partial y}\right)_x(a, b)$  is exactly the slope of (the tangent to) the curve of intersection of the surface  $z = f(x, y)$  and the plane  $x = a$  at the point  $(a, b, f(a, b))$ .

### ►►► Evaluation of Partial Derivatives

From the above discussion, we see that we can readily compute partial derivatives  $\frac{\partial}{\partial x}$  by using what we already know about ordinary derivatives  $\frac{d}{dx}$ . More precisely,

- to evaluate  $\frac{\partial f}{\partial x}(x, y)$ , treat the  $y$  in  $f(x, y)$  as a constant and differentiate the resulting function of  $x$  with respect to  $x$ .
- To evaluate  $\frac{\partial f}{\partial y}(x, y)$ , treat the  $x$  in  $f(x, y)$  as a constant and differentiate the resulting function of  $y$  with respect to  $y$ .
- To evaluate  $\frac{\partial f}{\partial x}(a, b)$ , treat the  $y$  in  $f(x, y)$  as a constant and differentiate the resulting function of  $x$  with respect to  $x$ . Then evaluate the result at  $x = a, y = b$ .
- To evaluate  $\frac{\partial f}{\partial y}(a, b)$ , treat the  $x$  in  $f(x, y)$  as a constant and differentiate the resulting function of  $y$  with respect to  $y$ . Then evaluate the result at  $x = a, y = b$ .

Now for some examples.

#### Example 2.1.4

Let

$$f(x, y) = x^3 + y^2 + 4xy^2$$

Then, since  $\frac{\partial}{\partial x}$  treats  $y$  as a constant,

$$\begin{aligned} \frac{\partial f}{\partial x} &= \left( \frac{\partial f}{\partial x} \right)_y = \frac{\partial}{\partial x}(x^3) + \frac{\partial}{\partial x}(y^2) + \frac{\partial}{\partial x}(4xy^2) \\ &= 3x^2 + 0 + 4y^2 \frac{\partial}{\partial x}(x) \\ &= 3x^2 + 4y^2 \end{aligned}$$

and, since  $\frac{\partial}{\partial y}$  treats  $x$  as a constant,

$$\begin{aligned} \frac{\partial f}{\partial y} &= \left( \frac{\partial f}{\partial y} \right)_x = \frac{\partial}{\partial y}(x^3) + \frac{\partial}{\partial y}(y^2) + \frac{\partial}{\partial y}(4xy^2) \\ &= 0 + 2y + 4x \frac{\partial}{\partial y}(y^2) \\ &= 2y + 8xy \end{aligned}$$

In particular, at  $(x, y) = (1, 0)$  these partial derivatives take the values

$$\begin{aligned} \frac{\partial f}{\partial x}(1, 0) &= 3(1)^2 + 4(0)^2 = 3 \\ \frac{\partial f}{\partial y}(1, 0) &= 2(0) + 8(1)(0) = 0 \end{aligned}$$

#### Example 2.1.4

## Example 2.1.5

Let

$$f(x, y) = y \cos x + xe^{xy}$$

Then, since  $\frac{\partial}{\partial x}$  treats  $y$  as a constant,  $\frac{\partial}{\partial x}e^{yx} = ye^{yx}$  and

$$\frac{\partial f}{\partial x}(x, y) = -y \sin x + e^{xy} + xye^{xy}$$

$$\frac{\partial f}{\partial y}(x, y) = \cos x + x^2e^{xy}$$

## Example 2.1.5

Let's move up to a function of four variables. Things generalize in a quite straight forward way.

## Example 2.1.6

Let

$$f(x, y, z, t) = x \sin(y + 2z) + t^2e^{3y} \ln z$$

Then

$$\frac{\partial f}{\partial x}(x, y, z, t) = \sin(y + 2z)$$

$$\frac{\partial f}{\partial y}(x, y, z, t) = x \cos(y + 2z) + 3t^2e^{3y} \ln z$$

$$\frac{\partial f}{\partial z}(x, y, z, t) = 2x \cos(y + 2z) + t^2e^{3y} / z$$

$$\frac{\partial f}{\partial t}(x, y, z, t) = 2te^{3y} \ln z$$

## Example 2.1.6

Now here is a more complicated example — our function takes a special value at  $(0, 0)$ . To compute derivatives there we have to revert to the definition.

## Example 2.1.7

Set

$$f(x, y) = \begin{cases} \frac{\cos x - \cos y}{x - y} & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

If  $b \neq a$ , then for all  $(x, y)$  sufficiently close to  $(a, b)$ ,  $f(x, y) = \frac{\cos x - \cos y}{x - y}$  and we can compute the partial derivatives of  $f$  at  $(a, b)$  using the familiar rules of differentiation. However that is not the case for  $(a, b) = (0, 0)$ . To evaluate  $f_x(0, 0)$ , we need to set  $y = 0$  and find the derivative of

$$f(x, 0) = \begin{cases} \frac{\cos x - 1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

with respect to  $x$  at  $x = 0$ . To do so, we basically have to apply the definition

$$\begin{aligned}
 f_x(0,0) &= \lim_{h \rightarrow 0} \frac{f(h,0) - f(0,0)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\frac{\cos h - 1}{h} - 0}{h} && \text{(Recall that } h \neq 0 \text{ in the limit.)} \\
 &= \lim_{h \rightarrow 0} \frac{\cos h - 1}{h^2} \\
 &= \lim_{h \rightarrow 0} \frac{-\sin h}{2h} && \text{(By l'Hôpital's rule.)} \\
 &= \lim_{h \rightarrow 0} \frac{-\cos h}{2} && \text{(By l'Hôpital again.)} \\
 &= -\frac{1}{2}
 \end{aligned}$$

Example 2.1.7

Example 2.1.8

Again set

$$f(x, y) = \begin{cases} \frac{\cos x - \cos y}{x - y} & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

We'll now compute  $f_y(x, y)$  for all  $(x, y)$ .

*The case  $y \neq x$ :* When  $y \neq x$ ,

$$\begin{aligned}
 f_y(x, y) &= \frac{\partial}{\partial y} \frac{\cos x - \cos y}{x - y} \\
 &= \frac{(x - y) \frac{\partial}{\partial y} (\cos x - \cos y) - (\cos x - \cos y) \frac{\partial}{\partial y} (x - y)}{(x - y)^2} && \text{by the quotient rule} \\
 &= \frac{(x - y) \sin y + \cos x - \cos y}{(x - y)^2}
 \end{aligned}$$

*The case  $y = x$ :* When  $y = x$ ,

$$\begin{aligned}
 f_y(x, y) &= \lim_{h \rightarrow 0} \frac{f(x, y + h) - f(x, y)}{h} = \lim_{h \rightarrow 0} \frac{f(x, x + h) - f(x, x)}{h} \\
 &= \lim_{h \rightarrow 0} \frac{\frac{\cos x - \cos(x+h)}{x - (x+h)} - 0}{h} && \text{(Recall that } h \neq 0 \text{ in the limit.)} \\
 &= \lim_{h \rightarrow 0} \frac{\cos(x+h) - \cos x}{h^2}
 \end{aligned}$$

Now we apply L'Hôpital's rule, remembering that, in this limit,  $x$  is a constant and  $h$  is



the variable — so we differentiate with respect to  $h$ .

$$\begin{aligned} f_y(x, y) &= \lim_{h \rightarrow 0} \frac{-\sin(x+h)}{2h} \\ &= \lim_{h \rightarrow 0} \frac{-\cos(x+h)}{2} \\ &= -\frac{\cos x}{2} \end{aligned}$$

The conclusion:

$$f(x, y) = \begin{cases} \frac{(x-y)\sin y + \cos x - \cos y}{(x-y)^2} & \text{if } x \neq y \\ -\frac{\cos x}{2} & \text{if } x = y \end{cases}$$

Example 2.1.8

Our next example uses implicit differentiation.

Example 2.1.9

The equation

$$z^5 + y^2 e^z + e^{2x} = 0$$

implicitly determines  $z$  as a function of  $x$  and  $y$ . For example, when  $x = y = 0$ , the equation reduces to

$$z^5 = -1$$

which forces<sup>1</sup>  $z(0, 0) = -1$ . Let's find the partial derivative  $\frac{\partial z}{\partial x}(0, 0)$ .

We are not going to be able to explicitly solve the equation for  $z(x, y)$ . All we know is that

$$z(x, y)^5 + y^2 e^{z(x, y)} + e^{2x} = 0$$

for all  $x$  and  $y$ . We can turn this into an equation for  $\frac{\partial z}{\partial x}(0, 0)$  by differentiating<sup>2</sup> the whole equation with respect to  $x$ , giving

$$5z(x, y)^4 \frac{\partial z}{\partial x}(x, y) + y^2 e^{z(x, y)} \frac{\partial z}{\partial x}(x, y) + 2e^{2x} = 0$$

and then setting  $x = y = 0$ , giving

$$5z(0, 0)^4 \frac{\partial z}{\partial x}(0, 0) + 2 = 0$$

As we already know that  $z(0, 0) = -1$ ,

$$\frac{\partial z}{\partial x}(0, 0) = -\frac{2}{5z(0, 0)^4} = -\frac{2}{5}$$

1 The only real number  $z$  which obeys  $z^5 = -1$  is  $z = -1$ . However there are four other complex numbers which also obey  $z^5 = -1$ .

2 You should have already seen this technique, called implicit differentiation, in your first Calculus course.

## Example 2.1.9

Next we have a partial derivative disguised as a limit.

## Example 2.1.10

In this example we are going to evaluate the limit

$$\lim_{z \rightarrow 0} \frac{(x + y + z)^3 - (x + y)^3}{(x + y)z}$$

The critical observation is that, in taking the limit  $z \rightarrow 0$ ,  $x$  and  $y$  are fixed. They do not change as  $z$  is getting smaller and smaller. Furthermore this limit is exactly of the form of the limits in the Definition 2.1.1 of partial derivative, disguised by some obfuscating changes of notation.

Set

$$f(x, y, z) = \frac{(x + y + z)^3}{(x + y)}$$

Then

$$\begin{aligned} \lim_{z \rightarrow 0} \frac{(x + y + z)^3 - (x + y)^3}{(x + y)z} &= \lim_{z \rightarrow 0} \frac{f(x, y, z) - f(x, y, 0)}{z} = \lim_{h \rightarrow 0} \frac{f(x, y, 0 + h) - f(x, y, 0)}{h} \\ &= \frac{\partial f}{\partial z}(x, y, 0) \\ &= \left[ \frac{\partial}{\partial z} \frac{(x + y + z)^3}{x + y} \right]_{z=0} \end{aligned}$$

Recalling that  $\frac{\partial}{\partial z}$  treats  $x$  and  $y$  as constants, we are evaluating the derivative of a function of the form  $\frac{(\text{const}+z)^3}{\text{const}}$ . So

$$\begin{aligned} \lim_{z \rightarrow 0} \frac{(x + y + z)^3 - (x + y)^3}{(x + y)z} &= 3 \frac{(x + y + z)^2}{x + y} \Big|_{z=0} \\ &= 3(x + y) \end{aligned}$$

## Example 2.1.10

## 2.2▲ Higher Order Derivatives

You have already observed, in your first Calculus course, that if  $f(x)$  is a function of  $x$ , then its derivative,  $\frac{df}{dx}(x)$ , is also a function of  $x$ , and can be differentiated to give the second order derivative  $\frac{d^2f}{dx^2}(x)$ , which can in turn be differentiated yet again to give the third order derivative,  $f^{(3)}(x)$ , and so on.

We can do the same for functions of more than one variable. If  $f(x, y)$  is a function of  $x$  and  $y$ , then both of its partial derivatives,  $\frac{\partial f}{\partial x}(x, y)$  and  $\frac{\partial f}{\partial y}(x, y)$  are also functions of  $x$  and

$y$ . They can both be differentiated with respect to  $x$  and they can both be differentiated with respect to  $y$ . So there are four possible second order derivatives. Here they are, together with various alternate notations.

$$\frac{\partial}{\partial x} \left( \frac{\partial f}{\partial x} \right) (x, y) = \frac{\partial^2 f}{\partial x^2} (x, y) = f_{xx}(x, y)$$

$$\frac{\partial}{\partial y} \left( \frac{\partial f}{\partial x} \right) (x, y) = \frac{\partial^2 f}{\partial y \partial x} (x, y) = f_{xy}(x, y)$$

$$\frac{\partial}{\partial x} \left( \frac{\partial f}{\partial y} \right) (x, y) = \frac{\partial^2 f}{\partial x \partial y} (x, y) = f_{yx}(x, y)$$

$$\frac{\partial}{\partial y} \left( \frac{\partial f}{\partial y} \right) (x, y) = \frac{\partial^2 f}{\partial y^2} (x, y) = f_{yy}(x, y)$$

### Warning 2.2.1.

In  $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial^2 f}{\partial y \partial x}$ , the derivative closest to  $f$ , in this case  $\frac{\partial}{\partial x}$ , is applied first. So we work through the variables in the bottom right-to-left.

In  $f_{xy}$ , the derivative with respect to the variable closest to  $f$ , in this case  $x$ , is applied first. So we work through the subscript variables left-to-right.

The difference in “direction” highlighted in the warning seems confusing at first, but it stems from the way the first partial derivative is written. In the fractional notation, if  $f$  is being differentiated with respect to  $x$ , we write  $\frac{\partial f}{\partial x}$  or  $\frac{\partial}{\partial x} f$ . So the operator  $\frac{\partial}{\partial x}$  is added to the *left* of the function. Now suppose we want to differentiate  $\frac{\partial f}{\partial x}$  with respect to  $y$ . By analogy, we would write  $\frac{\partial}{\partial y} \left[ \frac{\partial f}{\partial x} \right]$ , or  $\frac{\partial^2 f}{\partial y \partial x}$ . This leads to the order of variables being right-to-left.

With the subscript notation, if  $f$  is being differentiated with respect to  $x$ , we write  $f_x$ , with the variable on the *right* of the function. So now if we take the second derivative with respect to  $y$ , it makes sense by analogy to add that new variable to the right:  $(f_x)_y$ , or  $f_{xy}$ , in left-to-right order.

### Example 2.2.2

Let  $f(x, y) = e^{my} \cos(nx)$ . Then

$$f_x = -ne^{my} \sin(nx)$$

$$f_y = me^{my} \cos(nx)$$

$$f_{xx} = -n^2 e^{my} \cos(nx)$$

$$f_{yx} = -mne^{my} \sin(nx)$$

$$f_{xy} = -mne^{my} \sin(nx)$$

$$f_{yy} = m^2 e^{my} \cos(nx)$$

### Example 2.2.2

## Example 2.2.3

Let  $f(x, y) = e^{\alpha x + \beta y}$ . Then

$$\begin{aligned} f_x &= \alpha e^{\alpha x + \beta y} & f_y &= \beta e^{\alpha x + \beta y} \\ f_{xx} &= \alpha^2 e^{\alpha x + \beta y} & f_{yx} &= \beta \alpha e^{\alpha x + \beta y} \\ f_{xy} &= \alpha \beta e^{\alpha x + \beta y} & f_{yy} &= \beta^2 e^{\alpha x + \beta y} \end{aligned}$$

More generally, for any integers  $m, n \geq 0$ ,

$$\frac{\partial^{m+n} f}{\partial x^m \partial y^n} = \alpha^m \beta^n e^{\alpha x + \beta y}$$

## Example 2.2.3

## Example 2.2.4

If  $f(x_1, x_2, x_3, x_4) = x_1^4 x_2^3 x_3^2 x_4$ , then

$$\begin{aligned} \frac{\partial^4 f}{\partial x_1 \partial x_2 \partial x_3 \partial x_4} &= \frac{\partial^3}{\partial x_1 \partial x_2 \partial x_3} (x_1^4 x_2^3 x_3^2) \\ &= \frac{\partial^2}{\partial x_1 \partial x_2} (2 x_1^4 x_2^3 x_3) \\ &= \frac{\partial}{\partial x_1} (6 x_1^4 x_2^2 x_3) \\ &= 24 x_1^3 x_2^2 x_3 \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^4 f}{\partial x_4 \partial x_3 \partial x_2 \partial x_1} &= \frac{\partial^3}{\partial x_4 \partial x_3 \partial x_2} (4 x_1^3 x_2^3 x_3^2 x_4) \\ &= \frac{\partial^2}{\partial x_4 \partial x_3} (12 x_1^3 x_2^2 x_3^2 x_4) \\ &= \frac{\partial}{\partial x_4} (24 x_1^3 x_2^2 x_3 x_4) \\ &= 24 x_1^3 x_2^2 x_3 \end{aligned}$$

## Example 2.2.4

Notice that in Example 2.2.2,

$$f_{xy} = f_{yx} = -mne^{my} \sin(nx)$$

and in Example 2.2.3

$$f_{xy} = f_{yx} = \alpha \beta e^{\alpha x + \beta y}$$

and in Example 2.2.4

$$\frac{\partial^4 f}{\partial x_1 \partial x_2 \partial x_3 \partial x_4} = \frac{\partial^4 f}{\partial x_4 \partial x_3 \partial x_2 \partial x_1} = 24 x_1^3 x_2^2 x_3$$

In all of these examples, it didn't matter what order we took the derivatives in. The following theorem<sup>3</sup> shows that this was no accident.

**Theorem 2.2.5** (Clairaut's Theorem<sup>4</sup> or Schwarz's Theorem<sup>5</sup>).

If the partial derivatives  $\frac{\partial^2 f}{\partial x \partial y}$  and  $\frac{\partial^2 f}{\partial y \partial x}$  exist and are continuous at  $(x_0, y_0)$ , then

$$\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$$

The Proof of Theorem 2.2.5 can be found in Appendix A.4.1. An example of a function  $f(x, y)$  where  $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \neq \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$  can be found in Appendix A.4.2.

We won't use this theorem a whole lot in Math 105. It can occasionally be useful to note that as long as a function is continuous and differentiable, you can differentiate it in any "order."

Example 2.2.6

Let  $f(x, y) = x^5 e^x + y$ . Find  $f_{xxyy}$ .

*Solution.* Since  $f(x, y)$  is continuous and differentiable everywhere, then the order of differentiation doesn't matter. Rather than starting with respect to  $x$  (which is harder), we start with respect to  $y$  (which is easier).

$$\begin{aligned} f_y &= 1 \\ f_{yx} &= 0 \implies f_{xy} = 0 \\ f_{xyx} &= 0 \implies f_{xxy} = 0 \\ f_{xxyx} &= 0 \implies f_{xxyy} = 0 \end{aligned}$$

Example 2.2.6

## 2.3▲ Local Maximum and Minimum Values

One of the core topics in single variable calculus courses is finding the maxima and minima of functions of one variable. We'll now extend that discussion to functions of more than one variable<sup>6</sup>. To keep things simple, we'll focus on functions with two variables.

3 The history of this important theorem is pretty convoluted. See "A note on the history of mixed partial derivatives" by Thomas James Higgins which was published in *Scripta Mathematica* 7 (1940), 59-62.

4 Alexis Clairaut (1713–1765) was a French mathematician, astronomer, and geophysicist.

5 Hermann Schwarz (1843–1921) was a German mathematician.

6 Life is not (always) one-dimensional and sometimes we have to embrace it.

It's worth noting, though, that many of the techniques we use will generalize to functions with even more. To start, we have the following natural extensions to some familiar definitions.

### Definition 2.3.1.

Let the function  $f(x, y)$  be defined for all  $(x, y)$  in some subset  $R$  of  $\mathbb{R}^2$ . Let  $(a, b)$  be a point in  $R$ .

- $(a, b)$  is a *local maximum* of  $f(x, y)$  if  $f(x, y) \leq f(a, b)$  for all  $(x, y)$  close to  $(a, b)$ . More precisely,  $(a, b)$  is a local maximum of  $f(x, y)$  if there is an  $r > 0$  such that  $f(x, y) \leq f(a, b)$  for all points  $(x, y)$  within a distance  $r$  of  $(a, b)$ .
- $(a, b)$  is a *local minimum* of  $f(x, y)$  if  $f(x, y) \geq f(a, b)$  for all  $(x, y)$  close to  $(a, b)$ .
- Local maximum and minimum values are also called extremal values.
- $(a, b)$  is an *absolute maximum* or *global maximum* of  $f(x, y)$  if  $f(x, y) \leq f(a, b)$  for all  $(x, y)$  in  $R$ .
- $(a, b)$  is an *absolute minimum* or *global minimum* of  $f(x, y)$  if  $f(x, y) \geq f(a, b)$  for all  $(x, y)$  in  $R$ .

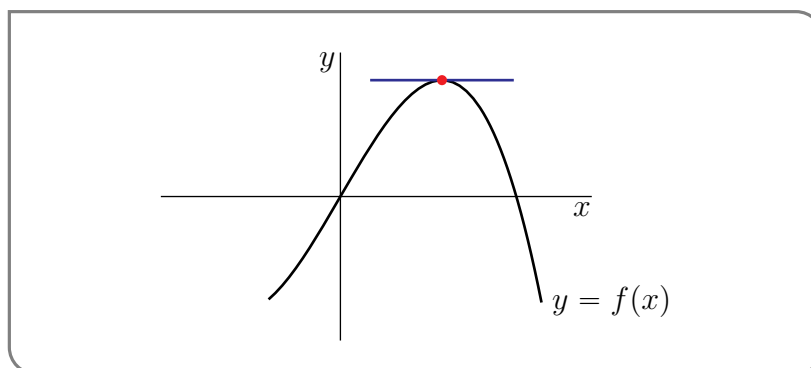
### 2.3.1 ► Critical Points

One of the first things you did when you were developing the techniques used to find the maximum and minimum values of  $f(x)$  was to ask yourself<sup>7</sup>

Suppose that the largest value of  $f(x)$  is  $f(a)$ . What does that tell us about  $a$ ?

After a little thought you answered

If the largest value of  $f(x)$  is  $f(a)$  and  $f$  is differentiable at  $a$ , then  $f'(a) = 0$ .



<sup>7</sup> Or perhaps your instructor asked you.

Let's recall why that's true. Suppose that the largest value of  $f(x)$  is  $f(a)$ . Then for all  $h > 0$ ,

$$f(a+h) \leq f(a) \implies f(a+h) - f(a) \leq 0 \implies \frac{f(a+h) - f(a)}{h} \leq 0 \quad \text{if } h > 0$$

Taking the limit  $h \rightarrow 0$  tells us that  $f'(a) \leq 0$ . Similarly, for all  $h < 0$ ,

$$f(a+h) \leq f(a) \implies f(a+h) - f(a) \leq 0 \implies \frac{f(a+h) - f(a)}{h} \geq 0 \quad \text{if } h < 0$$

Taking the limit  $h \rightarrow 0$  now tells us that  $f'(a) \geq 0$ . So we have both  $f'(a) \geq 0$  and  $f'(a) \leq 0$  which forces  $f'(a) = 0$ .

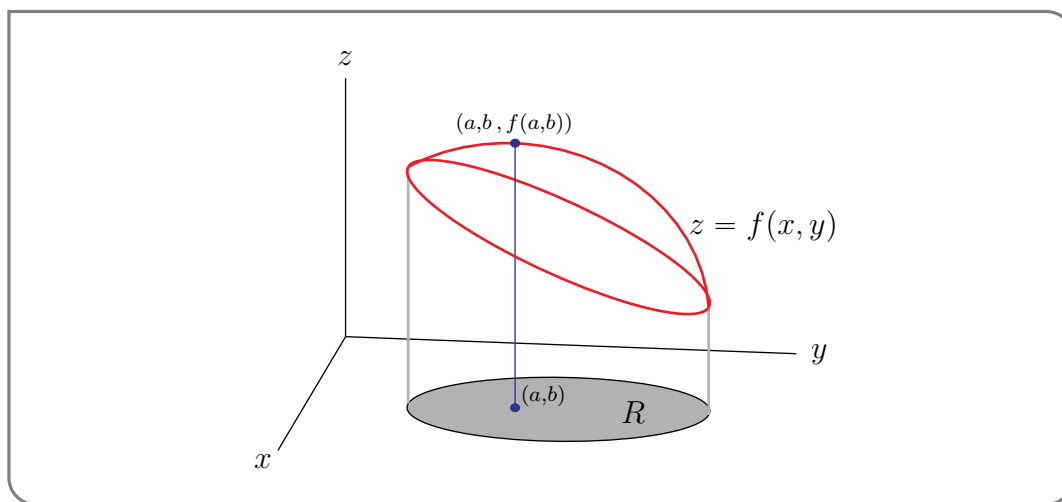
You also observed at the time that for this argument to work, you only need  $f(x) \leq f(a)$  for all  $x$ 's close to  $a$ , not necessarily for all  $x$ 's in the whole world. (In the above inequalities, we only used  $f(a+h)$  with  $h$  small.) Since we care only about  $f(x)$  for  $x$  near  $a$ , we can refine the above statement.

If  $f(a)$  is a local maximum for  $f(x)$  and  $f$  is differentiable at  $a$ , then  $f'(a) = 0$ .

Precisely the same reasoning applies to minima.

If  $f(a)$  is a local minimum for  $f(x)$  and  $f$  is differentiable at  $a$ , then  $f'(a) = 0$ .

Let's use the ideas of the above discourse to extend the study of local maxima and local minima to functions of more than one variable. Suppose that the function  $f(x, y)$  is defined for all  $(x, y)$  in some subset  $R$  of  $\mathbb{R}^2$ , that  $(a, b)$  is point of  $R$  that is not on the boundary of  $R$ , and that  $f$  has a local maximum at  $(a, b)$ . See the figure below.



Then the function  $f(x, y)$  must decrease in value as  $(x, y)$  moves away from  $(a, b)$  in *any* direction. If we change the  $x$ -coordinate a little,  $f(x, y)$  must not increase. So for all  $h > 0$ :

$$f(a+h, b) \leq f(a, b) \implies f(a+h, b) - f(a, b) \leq 0 \implies \frac{f(a+h, b) - f(a, b)}{h} \leq 0 \quad \text{if } h > 0$$

Taking the limit  $h \rightarrow 0$  tells us that  $f_x(a, b) \leq 0$ .

Similarly, for all  $h < 0$ ,

$$f(a+h, b) \leq f(a, b) \implies f(a+h, b) - f(a, b) \leq 0 \implies \frac{f(a+h, b) - f(a, b)}{h} \geq 0 \quad \text{if } h < 0$$

Taking the limit  $h \rightarrow 0$  now tells us that  $f_x(a, b) \geq 0$ . So we have both  $f_x(a, b) \geq 0$  and  $f_x(a, b) \leq 0$  which forces  $f_x(a, b) = 0$ . The same reasoning tells us  $f_y(a, b) = 0$  as well, and that these partial derivatives are zero for minima as well as maxima.

This is an important and useful result, so let's theoremise it.

**Theorem 2.3.2.**

Let the function  $f(x, y)$  be defined for all  $(x, y)$  in some subset  $R$  of  $\mathbb{R}^2$ . Assume that

- $(a, b)$  is a point of  $R$  that is not on the boundary of  $R$  and
- $(a, b)$  is a local maximum or local minimum of  $f$  and that
- the partial derivatives of  $f$  exist at  $(a, b)$ .

Then

$$f_x(a, b) = 0$$

and  $f_y(a, b) = 0$

**Definition 2.3.3.**

Let  $f(x, y)$  be a function and let  $(a, b)$  be a point in its domain. Then we call  $(a, b)$  a *critical point* (or a *stationary point*) of the function if

- $f_x(a, b)$  does not exist, **or**
- $f_y(a, b)$  does not exist, **or**
- $f_x(a, b) = f_y(a, b) = 0$ .

**Warning 2.3.4.**

Note that some people (and texts) do not include the cases where one or both partial derivatives do not exist in the definition of a critical point. These points would (usually) be referred as a singular point of the function. We do not use this terminology.



**Warning 2.3.5.**

Theorem 2.3.2 tells us that every local maximum or minimum (in the interior of the domain of a differentiable function) is a critical point. Beware that it does *not*<sup>8</sup> tell us that every critical point is either a local maximum or a local minimum.

In fact, as we shall see in Example 2.3.12, critical points that are neither local maxima nor a local minima. None-the-less, Theorem 2.3.2 is very useful because often functions have only a small number of critical points. To find local maxima and minima of such functions, we only need to consider its critical points. We'll return later to the question of how to tell if a critical point is a local maximum, local minimum or neither. For now, we'll just practice finding critical points.

Example 2.3.6 ( $f(x, y) = x^2 - 2xy + 2y^2 + 2x - 6y + 12$ )

Find all critical points of  $f(x, y) = x^2 - 2xy + 2y^2 + 2x - 6y + 12$ .

*Solution.* To find the critical points, we need to find the first order partial derivatives. So, as a preliminary calculation, we find the two first order partial derivatives of  $f(x, y)$ .

$$\begin{aligned} f_x(x, y) &= 2x - 2y + 2 \\ f_y(x, y) &= -2x + 4y - 6 \end{aligned}$$

These functions are defined everywhere. So the critical points are the solutions of the pair of equations

$$2x - 2y + 2 = 0 \quad -2x + 4y - 6 = 0$$

or equivalently (dividing by two and moving the constants to the right hand side)

$$x - y = -1 \tag{E1}$$

$$-x + 2y = 3 \tag{E2}$$

This is a system of two equations in two unknowns ( $x$  and  $y$ ). One strategy for solving system like this is to

- First use one of the equations to solve for one of the unknowns in terms of the other unknown. For example, (E1) tells us that  $y = x + 1$ . This expresses  $y$  in terms of  $x$ . We say that we have solved for  $y$  in terms of  $x$ .
- Then substitute the result,  $y = x + 1$  in our case, into the other equation, (E2). In our case, this gives

$$-x + 2(x + 1) = 3 \iff x + 2 = 3 \iff x = 1$$

- We have now found that  $x = 1, y = x + 1 = 2$  is the only solution. So the only critical point is  $(1, 2)$ . Of course it only takes a moment to verify that  $f_x(1, 2) = f_y(1, 2) = 0$ . It is a good idea to do this as a simple check of our work.

8 A very common error of logic that people make is "Affirming the consequent". "If P then Q" is true, does not imply that "If Q then P" is true. The statement "If he is Shakespeare then he is dead" is true. But concluding from "That sheep is dead" that "He must be Shakespeare" is just silly.

An alternative strategy for solving a system of two equations in two unknowns, like (E1) and (E2), is to

- add equations (E1) and (E2) together. This gives

$$(E1) + (E2) : (1 - 1)x + (-1 + 2)y = -1 + 3 \iff y = 2$$

The point here is that adding equations (E1) and (E2) together eliminates the unknown  $x$ , leaving us with one equation in the unknown  $y$ , which is easily solved. For other systems of equations you might have to multiply the equations by some numbers before adding them together.

- We now know that  $y = 2$ . Substituting it into (E1) gives us

$$x - 2 = -1 \implies x = 1$$

- Once again (thankfully) we have found that the only critical point is  $(1, 2)$ .

Example 2.3.6

This was pretty easy because we only had to solve linear equations, which in turn was a consequence of the fact that  $f(x, y)$  was a polynomial of degree two. Here is an example with some slightly more challenging algebra.

Example 2.3.7 ( $f(x, y) = 2x^3 - 6xy + y^2 + 4y$ )

Find all critical points of  $f(x, y) = 2x^3 - 6xy + y^2 + 4y$ .

*Solution.* As in the last example, we need to find where the partial derivatives do not exist or are zero.

$$f_x = 6x^2 - 6y \quad f_y = -6x + 2y + 4$$

These functions are defined everywhere. So the critical points are the solutions of

$$6x^2 - 6y = 0 \quad -6x + 2y + 4 = 0$$

We can rewrite the first equation as  $y = x^2$ , which expresses  $y$  as a function of  $x$ . We can then substitute  $y = x^2$  into the second equation, giving

$$\begin{aligned} -6x + 2y + 4 = 0 &\iff -6x + 2x^2 + 4 = 0 \iff x^2 - 3x + 2 = 0 \iff (x - 1)(x - 2) = 0 \\ &\iff x = 1 \text{ or } 2 \end{aligned}$$

When  $x = 1$ ,  $y = 1^2 = 1$  and when  $x = 2$ ,  $y = 2^2 = 4$ . So, there are two critical points:  $(1, 1)$ ,  $(2, 4)$ .

Alternatively, we could have also used the second equation to write  $y = 3x - 2$ , and then substituted that into the first equation to get

$$6x^2 - 6(3x - 2) = 0 \iff x^2 - 3x + 2 = 0$$

just as above.

## Example 2.3.7

And here is an example for which the algebra requires a bit more thought.

Example 2.3.8 ( $f(x, y) = xy(5x + y - 15)$ )

Find all critical points of  $f(x, y) = xy(5x + y - 15)$ .

*Solution.* The first order partial derivatives of  $f(x, y) = xy(5x + y - 15)$  are

$$f_x(x, y) = y(5x + y - 15) + xy(5) = y(5x + y - 15) + y(5x) = y(10x + y - 15)$$

$$f_y(x, y) = x(5x + y - 15) + xy(1) = x(5x + y - 15) + x(y) = x(5x + 2y - 15)$$

Therefore the partial derivatives of the function exist everywhere in the domain of the function. The critical points are the solutions of  $f_x(x, y) = f_y(x, y) = 0$ . That is, we need to find all  $x, y$  that satisfy the pair of equations

$$y(10x + y - 15) = 0 \tag{E1}$$

$$x(5x + 2y - 15) = 0 \tag{E2}$$

The first equation,  $y(10x + y - 15) = 0$ , is satisfied if at least one of the two factors  $y$ ,  $(10x + y - 15)$  is zero. So the first equation is satisfied if at least one of the two equations

$$y = 0 \tag{E1a}$$

$$10x + y = 15 \tag{E1b}$$

is satisfied. The second equation,  $x(5x + 2y - 15) = 0$ , is satisfied if at least one of the two factors  $x$ ,  $(5x + 2y - 15)$  is zero. So the second equation is satisfied if at least one of the two equations

$$x = 0 \tag{E2a}$$

$$5x + 2y = 15 \tag{E2b}$$

is satisfied.

So both critical point equations (E1) and (E2) are satisfied if and only if at least one of (E1a), (E1b) is satisfied and in addition at least one of (E2a), (E2b) is satisfied. So both critical point equations (E1) and (E2) are satisfied if and only if at least one of the following four possibilities hold.

- (E1a) and (E2a) are satisfied if and only if  $x = y = 0$
- (E1a) and (E2b) are satisfied if and only if  $y = 0, 5x + 2y = 15 \iff y = 0, 5x = 15$
- (E1b) and (E2a) are satisfied if and only if  $10x + y = 15, x = 0 \iff y = 15, x = 0$
- (E1b) and (E2b) are satisfied if and only if  $10x + y = 15, 5x + 2y = 15$ . We can use, for example, the second of these equations to solve for  $x$  in terms of  $y$ :  $x = \frac{1}{5}(15 - 2y)$ . When we substitute this into the first equation we get  $2(15 - 2y) + y = 15$ , which we can solve for  $y$ . This gives  $-3y = 15 - 30$  or  $y = 5$  and then  $x = \frac{1}{5}(15 - 2 \times 5) = 1$ .

In conclusion, the critical points are  $(0, 0)$ ,  $(3, 0)$ ,  $(0, 15)$  and  $(1, 5)$ .

A more compact way to write what we have just done is

$$\begin{aligned}
 & f_x(x, y) = 0 \quad \text{and} \quad f_y(x, y) = 0 \\
 \iff & y(10x + y - 15) = 0 \quad \text{and} \quad x(5x + 2y - 15) = 0 \\
 \iff & \{y = 0 \text{ or } 10x + y = 15\} \quad \text{and} \quad \{x = 0 \text{ or } 5x + 2y = 15\} \\
 \iff & \{y = 0, x = 0\} \text{ or } \{y = 0, 5x + 2y = 15\} \text{ or } \{10x + y = 15, x = 0\} \text{ or} \\
 & \{10x + y = 15, 5x + 2y = 15\} \\
 \iff & \{x = y = 0\} \text{ or } \{y = 0, x = 3\} \text{ or } \{x = 0, y = 15\} \text{ or } \{x = 1, y = 5\}
 \end{aligned}$$

Example 2.3.8

Let's try a more practical example — something from the real world. Well, a mathematician's "real world". The interested reader should search-engine their way to a discussion of "idealisation", "game theory" "Cournot models" and "Bertrand models". But don't spend too long there. A discussion of breweries is about to take place.

Example 2.3.9

In a certain community, there are two breweries in competition<sup>9</sup>, so that sales of each negatively affect the profits of the other. If brewery A produces  $x$  litres of beer per month and brewery B produces  $y$  litres per month, then the profits of the two breweries are given by

$$P = 2x - \frac{2x^2 + y^2}{10^6} \quad Q = 2y - \frac{4y^2 + x^2}{2 \times 10^6}$$

respectively. Find the sum of the two profits if each brewery independently sets its own production level to maximize its own profit and assumes that its competitor does likewise. Then, assuming cartel behaviour, find the sum of the two profits if the two breweries cooperate so as to maximize that sum<sup>10</sup>.

*Solution.* If A adjusts  $x$  to maximize  $P$  (for  $y$  held fixed) and B adjusts  $y$  to maximize  $Q$  (for  $x$  held fixed) then we want to find the  $(x, y)$  using

$$\begin{aligned}
 P_x &= 2 - \frac{4x}{10^6} \\
 Q_y &= 2 - \frac{8y}{2 \times 10^6}
 \end{aligned}$$

Note that  $P_x$  and  $Q_y$  exists everywhere. Then  $x$  and  $y$  are determined by the equations

$$P_x = 0 \tag{E1}$$

$$Q_y = 0 \tag{E2}$$

Equation (E1) yields  $x = \frac{1}{2}10^6$  and equation (E2) yields  $y = \frac{1}{2}10^6$ . Knowing  $x$  and  $y$  we can determine  $P$ ,  $Q$  and the total profit

$$\begin{aligned}
 P + Q &= 2(x + y) - \frac{1}{10^6} \left( \frac{5}{2}x^2 + 3y^2 \right) \\
 &= 10^6 \left( 1 + 1 - \frac{5}{8} - \frac{3}{4} \right) = \frac{5}{8}10^6
 \end{aligned}$$

<sup>9</sup> We have both types of music here — country and western.

<sup>10</sup> This sort of thing is generally illegal.

On the other hand if  $(A, B)$  adjust  $(x, y)$  to maximize  $P + Q = 2(x + y) - \frac{1}{10^6}(\frac{5}{2}x^2 + 3y^2)$ , then  $x$  and  $y$  are determined by

$$(P + Q)_x = 2 - \frac{5x}{10^6} = 0 \quad (\text{E1})$$

$$(P + Q)_y = 2 - \frac{6y}{10^6} = 0 \quad (\text{E2})$$

Equation (E1) yields  $x = \frac{2}{5}10^6$  and equation (E2) yields  $y = \frac{1}{3}10^6$ . Again knowing  $x$  and  $y$  we can determine the total profit

$$\begin{aligned} P + Q &= 2(x + y) - \frac{1}{10^6}(\frac{5}{2}x^2 + 3y^2) \\ &= 10^6(\frac{4}{5} + \frac{2}{3} - \frac{2}{5} - \frac{1}{3}) = \frac{11}{15}10^6 \end{aligned}$$

So cooperating really does help their profits. Unfortunately, like a very small tea-pot, consumers will be a little poorer<sup>11</sup>.

Example 2.3.9

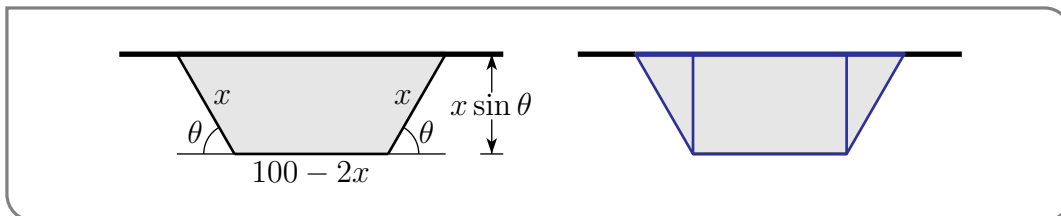
Moving swiftly away from the last pun, let's do something a little more geometric.

Example 2.3.10

Equal angle bends are made at equal distances from the two ends of a 100 metre long fence so the resulting three segment fence can be placed along an existing wall to make an enclosure of trapezoidal shape. What is the largest possible area for such an enclosure?



*Solution.* This is a very geometric problem (fenced off from pun opportunities), and as such we should start by drawing a sketch and introducing some variable names.



The area enclosed by the fence is the area inside the blue rectangle (in the figure on the right above) plus the area inside the two blue triangles.

$$\begin{aligned} A(x, \theta) &= (100 - 2x)x \sin \theta + 2 \cdot \frac{1}{2} \cdot x \sin \theta \cdot x \cos \theta \\ &= (100x - 2x^2) \sin \theta + x^2 \sin \theta \cos \theta \end{aligned}$$

<sup>11</sup> The authors extend their deepest apologies.

To maximize the area, we need to solve

$$0 = \frac{\partial A}{\partial x} = (100 - 4x) \sin \theta + 2x \sin \theta \cos \theta$$

$$0 = \frac{\partial A}{\partial \theta} = (100x - 2x^2) \cos \theta + x^2 \{ \cos^2 \theta - \sin^2 \theta \}$$

Note that  $\frac{\partial A}{\partial x}$  and  $\frac{\partial A}{\partial \theta}$  are defined everywhere in their domain (so here the critical points are the points where both partial derivatives are zero). Both terms in the first equation contain the factor  $\sin \theta$  and all terms in the second equation contain the factor  $x$ . If either  $\sin \theta$  or  $x$  are zero the area  $A(x, \theta)$  will also be zero, and so will certainly not be maximal. So we may divide the first equation by  $\sin \theta$  and the second equation by  $x$ , giving

$$(100 - 4x) + 2x \cos \theta = 0 \quad (\text{E1})$$

$$(100 - 2x) \cos \theta + x \{ \cos^2 \theta - \sin^2 \theta \} = 0 \quad (\text{E2})$$

These equations might look a little scary. But there is no need to panic. They are not as bad as they look because  $\theta$  enters only through  $\cos \theta$  and  $\sin^2 \theta$ , which we can easily write in terms of  $\cos \theta$ . Furthermore we can eliminate  $\cos \theta$  by observing that the first equation forces  $\cos \theta = -\frac{100-4x}{2x}$  and hence  $\sin^2 \theta = 1 - \cos^2 \theta = 1 - \frac{(100-4x)^2}{4x^2}$ . Substituting these into the second equation gives

$$-(100 - 2x) \frac{100 - 4x}{2x} + x \left[ \frac{(100 - 4x)^2}{2x^2} - 1 \right] = 0$$

$$\implies -(100 - 2x)(100 - 4x) + (100 - 4x)^2 - 2x^2 = 0$$

$$\implies 6x^2 - 200x = 0$$

$$\implies x = \frac{100}{3} \quad \cos \theta = -\frac{100/3}{200/3} = -\frac{1}{2} \quad \theta = 60^\circ$$

and the maximum area enclosed is

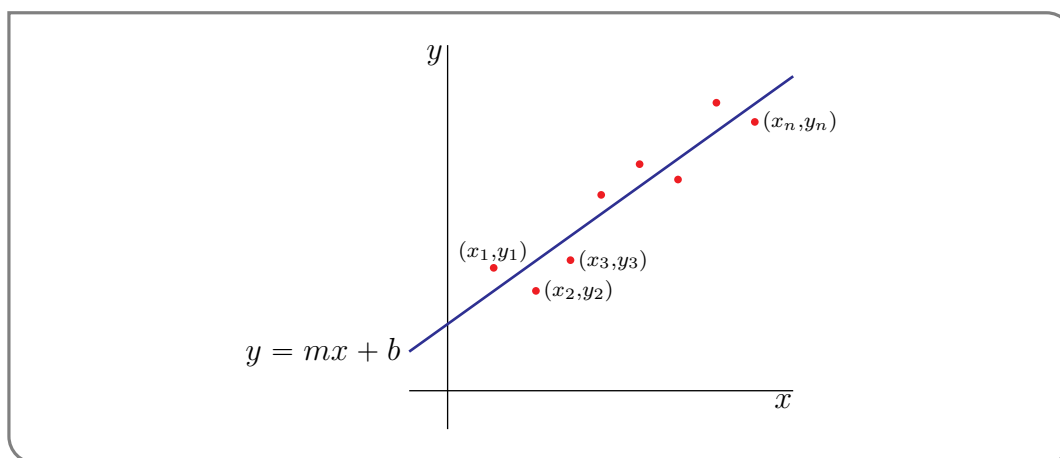
$$A = \left( 100 \frac{100}{3} - 2 \frac{100^2}{3^2} \right) \frac{\sqrt{3}}{2} + \frac{1}{2} \frac{100^2}{3^2} \frac{\sqrt{3}}{2} = \frac{2500}{\sqrt{3}}$$

Example 2.3.10

Now here is a very useful (even practical!) statistical example — finding the line that best fits a given collection of points.

Example 2.3.11 (Linear regression)

An experiment yields  $n$  data points  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . We wish to find the straight line  $y = mx + b$  which “best” fits the data. The definition of “best” is “minimizes the



root mean square error", i.e. minimizes

$$E(m, b) = \sum_{i=1}^n (mx_i + b - y_i)^2$$

Note that

- term number  $i$  in  $E(m, b)$  is the square of the difference between  $y_i$ , which is the  $i^{\text{th}}$  measured value of  $y$ , and  $\left[ mx + b \right]_{x=x_i}$ , which is the approximation to  $y_i$  given by the line  $y = mx + b$ .
- All terms in the sum are positive, regardless of whether the points  $(x_i, y_i)$  are above or below the line.

Our problem is to find the  $m$  and  $b$  that minimizes  $E(m, b)$ . This technique for drawing a line through a bunch of data points is called "linear regression". It is used *a lot*<sup>12 13</sup>. Even in the real world — and not just the real world that you find in mathematics problems. The actual real world that involves jobs.

*Solution.* We wish to choose  $m$  and  $b$  so as to minimize  $E(m, b)$ . So we need to determine where the partial derivatives of  $E$  do not exist, or exist and are equal to zero.

$$\begin{aligned} \frac{\partial E}{\partial m} &= \sum_{i=1}^n 2(mx_i + b - y_i)x_i = m \left[ \sum_{i=1}^n 2x_i^2 \right] + b \left[ \sum_{i=1}^n 2x_i \right] - \left[ \sum_{i=1}^n 2x_i y_i \right] \\ \frac{\partial E}{\partial b} &= \sum_{i=1}^n 2(mx_i + b - y_i) = m \left[ \sum_{i=1}^n 2x_i \right] + b \left[ \sum_{i=1}^n 2 \right] - \left[ \sum_{i=1}^n 2y_i \right] \end{aligned}$$

There are a lot of symbols here. But remember that all of the  $x_i$ 's and  $y_i$ 's are given constants. They come from, for example, experimental data. The only unknowns are  $m$  and  $b$ . To emphasize this, and to save some writing, define the constants

$$S_x = \sum_{i=1}^n x_i \quad S_y = \sum_{i=1}^n y_i \quad S_{x^2} = \sum_{i=1}^n x_i^2 \quad S_{xy} = \sum_{i=1}^n x_i y_i$$

<sup>12</sup> Proof by search engine.

<sup>13</sup> And has been used for a long time. It was introduced by the French mathematician Adreïn-Marie Legendre, 1752–1833, in 1805, and by the German mathematician and physicist Carl Friedrich Gauss, 1777–1855, in 1809.

The partial derivatives of  $E$  exists everywhere so we only need to find where they are equal to zero. The equations which determine the critical points are (after dividing by two)

$$0 = S_{x^2} m + S_x b - S_{xy} \implies S_{x^2} m + S_x b = S_{xy} \quad (\text{E1})$$

$$0 = S_x m + n b - S_y \implies S_x m + n b = S_y \quad (\text{E2})$$

These are two linear equations on the unknowns  $m$  and  $b$ . They may be solved in any of the usual ways. One is to use (E2) to solve for  $b$  in terms of  $m$

$$b = \frac{1}{n}(S_y - S_x m) \quad (\text{E3})$$

and then substitute this into (E1) to get the equation

$$S_{x^2} m + \frac{1}{n} S_x (S_y - S_x m) = S_{xy} \implies (n S_{x^2} - S_x^2) m = n S_{xy} - S_x S_y$$

for  $m$ . We can then solve this equation for  $m$  and substitute back into (E3) to get  $b$ . This gives

$$m = \frac{n S_{xy} - S_x S_y}{n S_{x^2} - S_x^2} \quad b = -\frac{S_x S_{xy} - S_y S_{x^2}}{n S_{x^2} - S_x^2}$$

Another way to solve the system of equations is

$$\begin{aligned} n(\text{E1}) - S_x(\text{E2}) : & \quad [n S_{x^2} - S_x^2] m = n S_{xy} - S_x S_y \\ -S_x(\text{E1}) + S_{x^2}(\text{E2}) : & \quad [n S_{x^2} - S_x^2] b = -S_x S_{xy} + S_y S_{x^2} \end{aligned}$$

which gives the same solution.

So given a bunch of data points, it only takes a quick bit of arithmetic — no calculus required — to apply the above formulae and so to find the best fitting line. Of course while you don't need any calculus to apply the formulae, you do need calculus to understand where they came from. The same technique can be extended to other types of curve fitting problems. For example, polynomial regression.

Example 2.3.11

### 2.3.2 ▶ Classifying Critical Points

Now let's start thinking about how to tell if a critical point is a local minimum, local maximum, or neither. We'll start with an intuitive approach, then introduce the (multivariable) Second Derivative Test.

You have already encountered single variable functions that have a critical point which is neither a local max nor a local min. This can also happen for functions of two variables. We'll start with the simplest possible such example.

Example 2.3.12 ( $f(x, y) = x^2 - y^2$ )

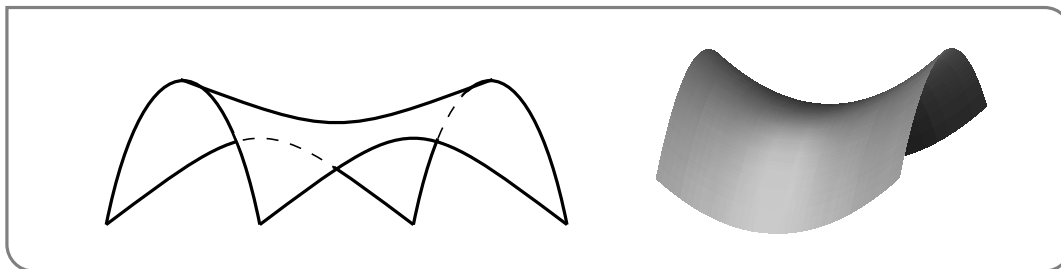
The first partial derivatives of  $f(x, y) = x^2 - y^2$  are  $f_x(x, y) = 2x$  and  $f_y(x, y) = -2y$ . So



the only critical point of this function is  $(0,0)$ . Is this a local minimum or maximum? Well let's start with  $(x,y)$  at  $(0,0)$  and then move  $(x,y)$  away from  $(0,0)$  and see if  $f(x,y)$  gets bigger or smaller. At the origin  $f(0,0) = 0$ . Of course we can move  $(x,y)$  away from  $(0,0)$  in many different directions.

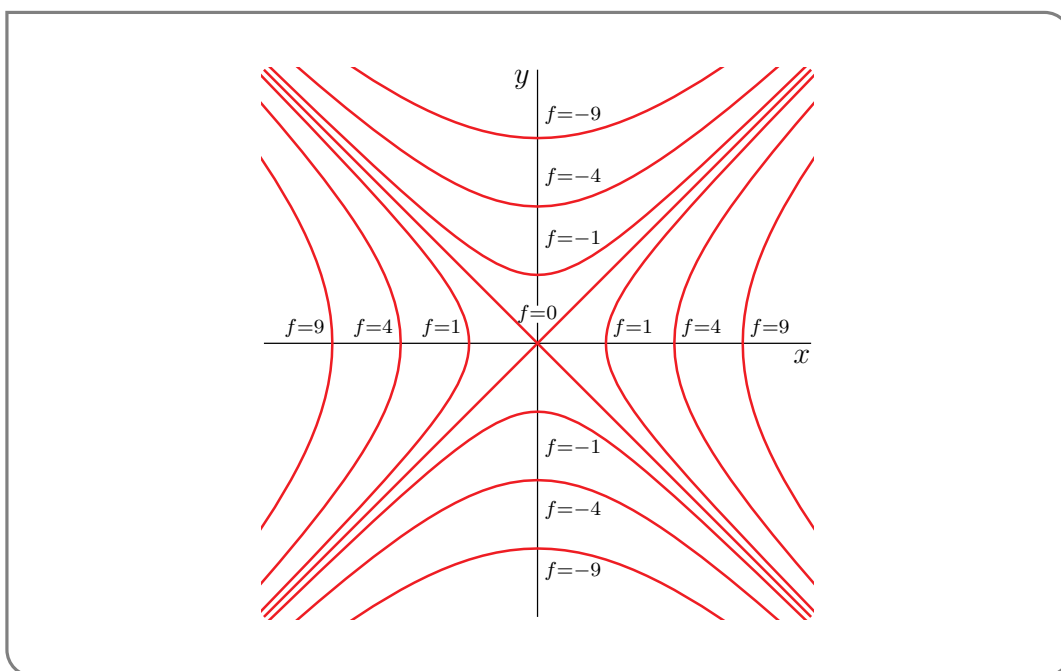
- First consider moving  $(x,y)$  along the  $x$ -axis. Then  $(x,y) = (x,0)$  and  $f(x,y) = f(x,0) = x^2$ . So when we start with  $x = 0$  and then increase  $x$ , the value of the function  $f$  increases — which means that  $(0,0)$  cannot be a local maximum for  $f$ .
- Next let's move  $(x,y)$  away from  $(0,0)$  along the  $y$ -axis. Then  $(x,y) = (0,y)$  and  $f(x,y) = f(0,y) = -y^2$ . So when we start with  $y = 0$  and then increase  $y$ , the value of the function  $f$  decreases — which means that  $(0,0)$  cannot be a local minimum for  $f$ .

So moving away from  $(0,0)$  in one direction causes the value of  $f$  to increase, while moving away from  $(0,0)$  in a second direction causes the value of  $f$  to decrease. Consequently  $(0,0)$  is neither a local minimum or maximum for  $f$ . It is called a saddle point, because the graph of  $f$  looks like a saddle. (The full definition of “saddle point” is given immediately after this example.) Here are some figures showing the graph of  $f$ .



The figure below show some level curves of  $f$ . Observe from the level curves that

- $f$  increases as you leave  $(0,0)$  walking along the  $x$  axis
- $f$  decreases as you leave  $(0,0)$  walking along the  $y$  axis



## Example 2.3.12

Approximately speaking, if a critical point  $(a, b)$  is neither a local minimum nor a local maximum, then it is a saddle point. For  $(a, b)$  to not be a local minimum,  $f$  has to take values smaller than  $f(a, b)$  at some points nearby  $(a, b)$ . For  $(a, b)$  to not be a local maximum,  $f$  has to take values bigger than  $f(a, b)$  at some points nearby  $(a, b)$ . Writing this more mathematically we get the following definition.

**Definition 2.3.13.**

The critical point  $(a, b)$  is called a saddle point for the function  $f(x, y)$  if, for each  $r > 0$ ,

- there is at least one point  $(x, y)$ , within a distance  $r$  of  $(a, b)$ , for which  $f(x, y) > f(a, b)$  and
- there is at least one point  $(x, y)$ , within a distance  $r$  of  $(a, b)$ , for which  $f(x, y) < f(a, b)$ .

Understanding what the graph of a function looks like is a powerful tool for classifying critical points, but it can be very time-consuming. The Second Derivative Test (below) is a more algebraic approach to classification. This test is often faster than graphing, but the drawback is that it is sometimes inconclusive.

**Theorem 2.3.14** (Second Derivative Test).

Let  $r > 0$  and assume that all second order derivatives of the function  $f(x, y)$  are continuous at all points  $(x, y)$  that are within a distance  $r$  of  $(a, b)$ . Assume that  $f_x(a, b) = f_y(a, b) = 0$ . Define

$$D(x, y) = f_{xx}(x, y) f_{yy}(x, y) - f_{xy}(x, y)^2$$

It is called the discriminant of  $f$ . Then

- if  $D(a, b) > 0$  and  $f_{xx}(a, b) > 0$ , then  $f(x, y)$  has a local minimum at  $(a, b)$ ,
- if  $D(a, b) > 0$  and  $f_{xx}(a, b) < 0$ , then  $f(x, y)$  has a local maximum at  $(a, b)$ ,
- if  $D(a, b) < 0$ , then  $f(x, y)$  has a saddle point at  $(a, b)$ , but
- if  $D(a, b) = 0$ , then we cannot draw any conclusions without more work.

The proof of Theorem 2.3.14 is beyond the scope of Math 105, but there is some intuition supporting it that is more accessible. Extremely informally, we can think of saddle points as places with inconsistent concavity: in some directions the surface looks concave

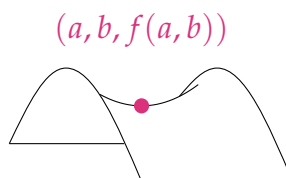
up, in other directions it looks concave down. On the other hand, at a local extremum, the concavity is the same in all directions.

Let's do thought experiments on a few simple cases to expand those ideas.

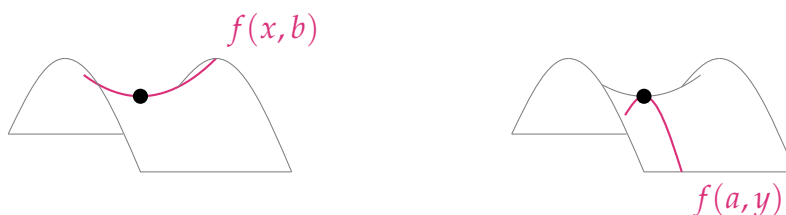
Example 2.3.15 (Second Derivative Test Intuition)

Let  $(a, b)$  be a critical point of the function  $f(x, y)$  with  $f_x(a, b) = f_y(a, b) = 0$ , and assume all second-order derivatives for  $f(x, y)$  are continuous.

1. Suppose at  $(a, b)$ , the surface looks like a minimum if  $y$  is held constant, but it looks like a maximum if  $x$  is held constant. (In particular, this means  $(a, b)$  is the location of a saddle point.)



Holding  $y = b$  constant, we can think of  $z = f(x, b)$  as a one-variable function, in which case  $f_{xx}(a, b) \geq 0$  by the single-variable second derivative test. Holding  $x = a$  constant, we can think of  $z = f(a, y)$  as a one-variable function (whose variable is  $y$ ). In that case,  $f_{yy}(a, b) \leq 0$  by the single-variable second derivative test.



Since  $f_{xx}(a, b)$  and  $f_{yy}(a, b)$  have different signs (or at least one of them is zero):

$$\begin{aligned} f_{xx}(a, b)f_{yy}(a, b) &\leq 0 \\ f_{xx}(a, b)f_{yy}(a, b) - f_{xy}^2(a, b) &\leq -f_{xy}^2(a, b) \leq 0 \\ D(a, b) &\leq 0 \end{aligned}$$

So in this simple saddle-point example, we expect  $D(a, b) \leq 0$ . This accords with the third bullet point in Theorem 2.3.14.

2. Suppose  $D(a, b) > 0$ .

$$\begin{aligned} 0 &< f_{xx}(a, b)f_{yy}(a, b) - f_{xy}^2(a, b) \\ f_{xy}^2(a, b) &< f_{xx}(a, b)f_{yy}(a, b) \end{aligned}$$

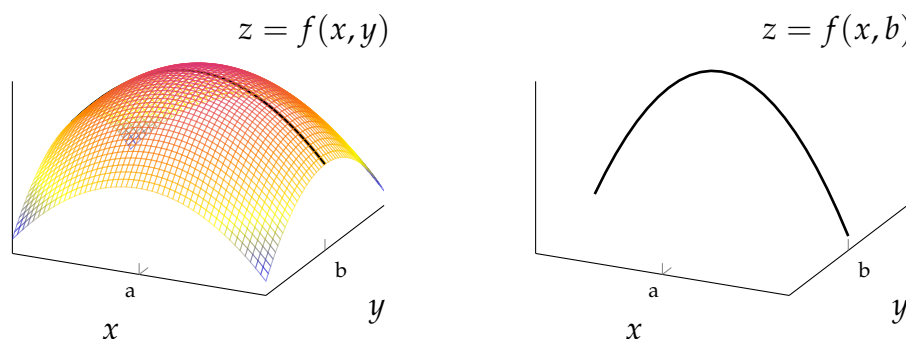
Since  $f_{xy}$  is raised to an even power, it's nonnegative.

$$\begin{aligned} 0 &\leq f_{xy}^2(a, b) < f_{xx}(a, b)f_{yy}(a, b) \\ 0 &< f_{xx}(a, b)f_{yy}(a, b) \end{aligned}$$

This tells us that  $f_{xx}(a, b)$  and  $f_{yy}(a, b)$  have the same sign – either they’re both positive or they’re both negative. So, the function’s concavity is the same whether we hold the  $x$ -value or the  $y$ -value constant. The function might have the same concavity in all directions – unlike the saddle point example we saw above. So, it seems plausible that critical points with positive discriminants are local extrema, rather than saddle points.

3. Suppose the surface has a local maximum at  $(a, b)$ .

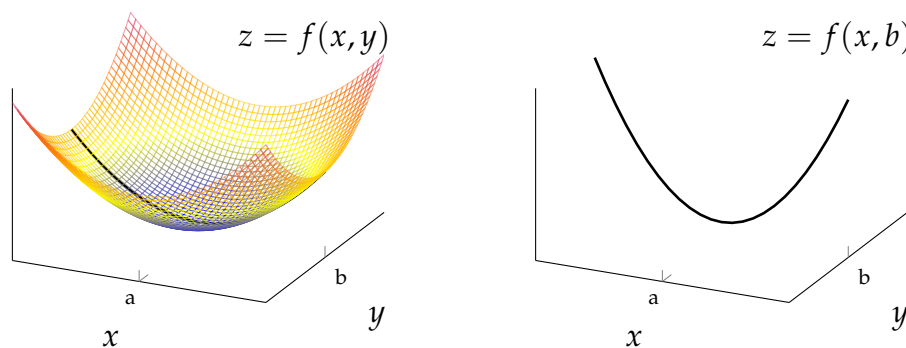
Holding  $y = b$  constant, we can think of  $z = f(x, b)$  as a one-variable function, in which case  $f_{xx}(a, b) \leq 0$  by the single-variable second derivative test.



This doesn’t go so far as to show us that  $D(a, b) \geq 0$ , but it does accord with the test of  $f_{xx}(a, b)$  in the second bullet point of Theorem 2.3.14.

4. Similarly, suppose the surface has a local minimum at  $(a, b)$ .

Holding  $y = b$  constant, we can think of  $z = f(x, b)$  as a one-variable function, in which case  $f_{xx}(a, b) \geq 0$  by the single-variable second derivative test.



Again, although this doesn’t go so far as to show us that  $D(a, b) \geq 0$ , it does accord with the test of  $f_{xx}(a, b)$  in the first bullet point of Theorem 2.3.14.

Example 2.3.15

You might wonder why, in the local maximum/local minimum cases of Theorem 2.3.14,  $f_{xx}(a, b)$  appears rather than  $f_{yy}(a, b)$ . The answer is only that  $x$  is before  $y$  in the

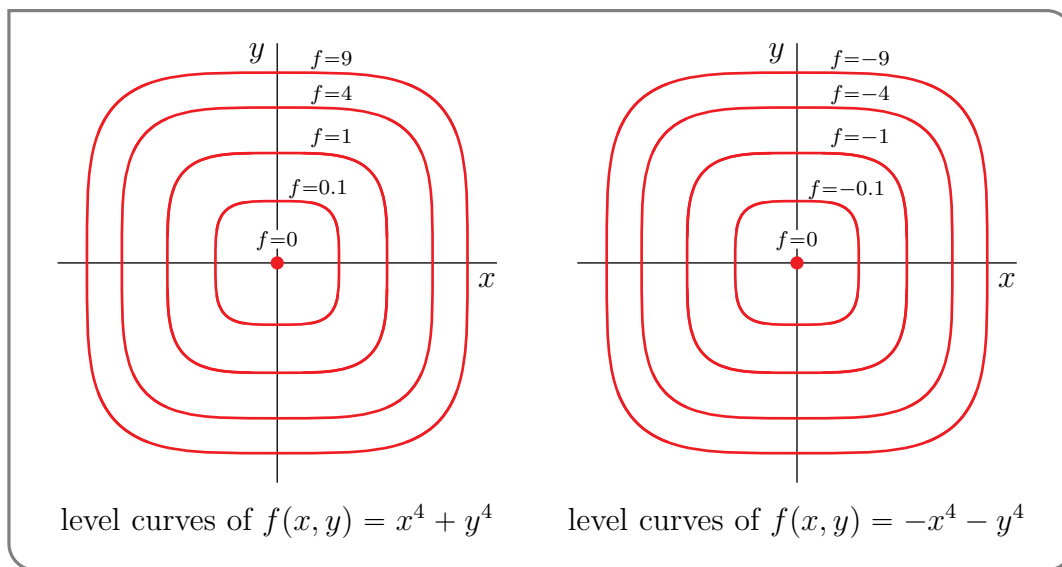
alphabet<sup>14</sup>. You can use  $f_{yy}(a, b)$  just as well as  $f_{xx}(a, b)$ . The reason is that if  $D(a, b) > 0$  (as in the first two bullets of the theorem), then because  $D(a, b) = f_{xx}(a, b)f_{yy}(a, b) - f_{xy}(a, b)^2 > 0$ , we necessarily have  $f_{xx}(a, b)f_{yy}(a, b) > 0$  so that  $f_{xx}(a, b)$  and  $f_{yy}(a, b)$  must have the same sign — either both are positive or both are negative.

You might also wonder why we cannot draw any conclusions when  $D(a, b) = 0$  and what happens then. The second derivative test for functions of two variables was derived in precisely the same way as the second derivative test for functions of one variable is derived — you approximate the function by a polynomial that is of degree two in  $(x - a)$ ,  $(y - b)$  and then you analyze the behaviour of the quadratic polynomial near  $(a, b)$ . For this to work, the contributions to  $f(x, y)$  from terms that are of degree two in  $(x - a)$ ,  $(y - b)$  had better be bigger than the contributions to  $f(x, y)$  from terms that are of degree three and higher in  $(x - a)$ ,  $(y - b)$  when  $(x - a)$ ,  $(y - b)$  are really small. If this is not the case, for example when the terms in  $f(x, y)$  that are of degree two in  $(x - a)$ ,  $(y - b)$  all have coefficients that are exactly zero, the analysis will certainly break down. That's exactly what happens when  $D(a, b) = 0$ . Here are some examples. The functions

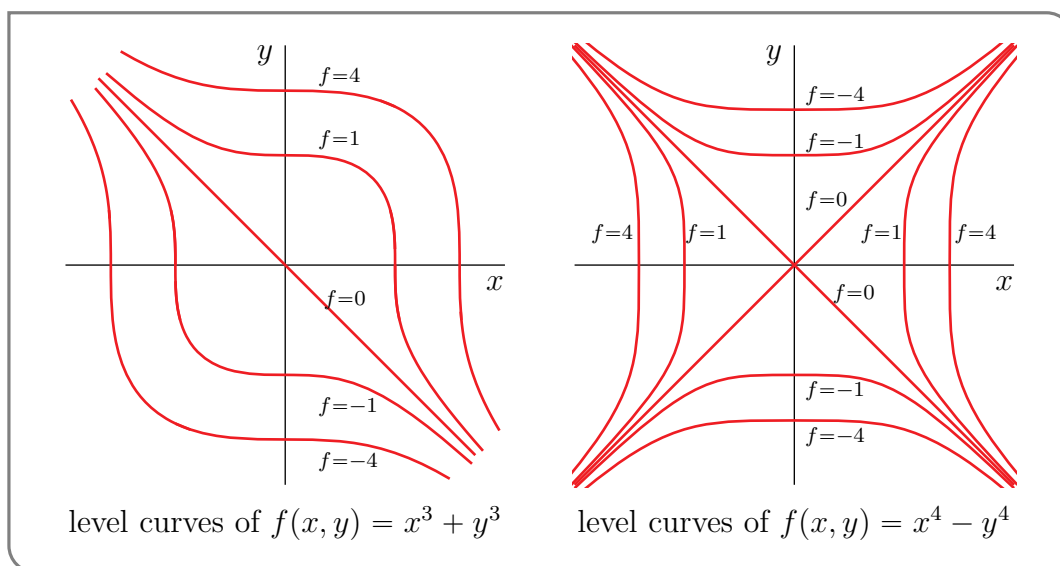
$$f_1(x, y) = x^4 + y^4 \quad f_2(x, y) = -x^4 - y^4 \quad f_3(x, y) = x^3 + y^3 \quad f_4(x, y) = x^4 - y^4$$

all have  $(0, 0)$  as the only critical point and all have  $D(0, 0) = 0$ . The first,  $f_1$  has its minimum there. The second,  $f_2$ , has its maximum there. The third and fourth have a saddle point there.

Here are sketches of some level curves for each of these four functions (with all renamed to simply  $f$ ).



14 The shackles of convention are not limited to mathematics. Election ballots often have the candidates listed in alphabetic order.



Example 2.3.16 ( $f(x, y) = 2x^3 - 6xy + y^2 + 4y$ )

Find and classify all critical points of  $f(x, y) = 2x^3 - 6xy + y^2 + 4y$ .

*Solution.* Thinking a little way ahead, to find the critical points we will need the first order partial derivatives. To apply the second derivative test of Theorem 2.3.14 we will need all second order partial derivatives. So we need all partial derivatives of order up to two. Here they are.

$$\begin{aligned}
 f &= 2x^3 - 6xy + y^2 + 4y \\
 f_x &= 6x^2 - 6y & f_{xx} &= 12x & f_{xy} &= -6 \\
 f_y &= -6x + 2y + 4 & f_{yy} &= 2 & f_{yx} &= -6
 \end{aligned}$$

(Of course,  $f_{xy}$  and  $f_{yx}$  have to be the same. It is still useful to compute both, as a way to catch some mechanical errors.)

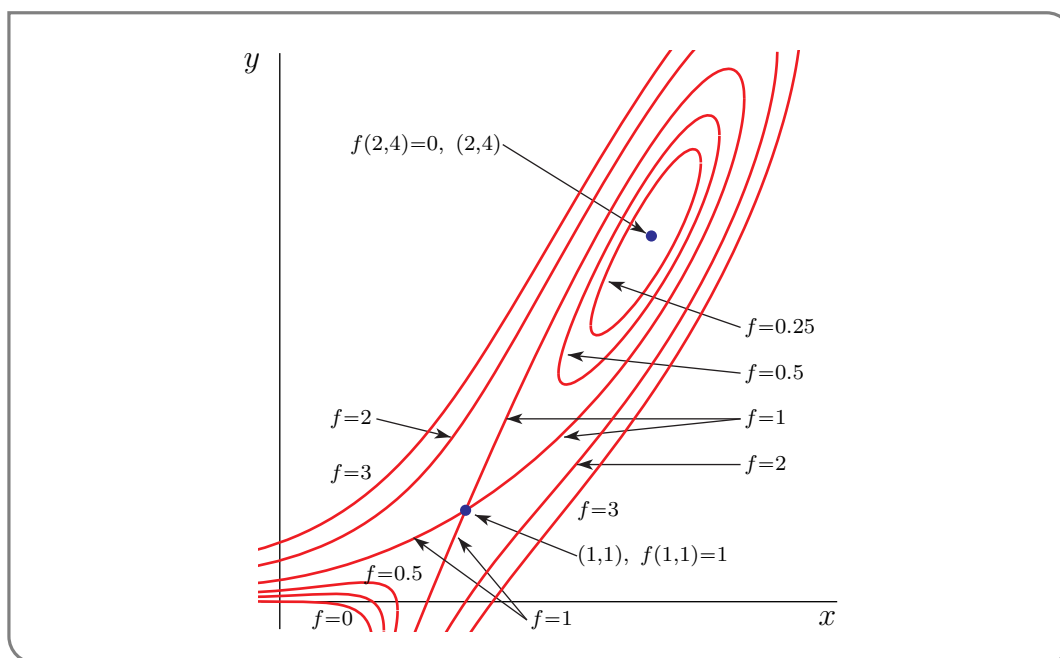
We have already found, in Example 2.3.7, that the critical points are  $(1, 1)$ ,  $(2, 4)$ . The classification is

critical point	$f_{xx}f_{yy} - f_{xy}^2$	$f_{xx}$	type
$(1, 1)$	$12 \times 2 - (-6)^2 < 0$		saddle point
$(2, 4)$	$24 \times 2 - (-6)^2 > 0$	24	local min

We were able to leave the  $f_{xx}$  entry in the top row blank, because

- we knew that  $f_{xx}(1, 1)f_{yy}(1, 1) - f_{xy}^2(1, 1) < 0$ , and
- we knew, from Theorem 2.3.14, that  $f_{xx}(1, 1)f_{yy}(1, 1) - f_{xy}^2(1, 1) < 0$ , by itself, was enough to ensure that  $(1, 1)$  was a saddle point.

Here is a sketch of some level curves of our  $f(x, y)$ . They are not needed to answer this



question, but can give you some idea as to what the graph of  $f$  looks like.

Example 2.3.16

Example 2.3.17 ( $f(x, y) = xy(5x + y - 15)$ )

Find and classify all critical points of  $f(x, y) = xy(5x + y - 15)$ .

*Solution.* We have already computed the first order partial derivatives

$$f_x(x, y) = y(10x + y - 15) \quad f_y(x, y) = x(5x + 2y - 15)$$

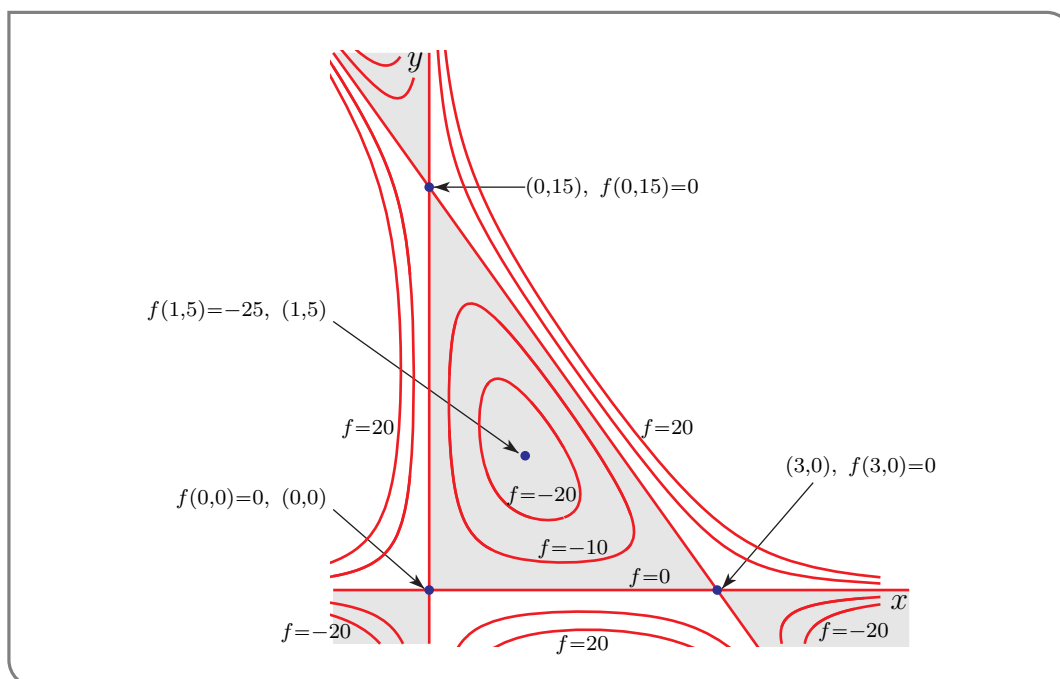
of  $f(x, y)$  in Example 2.3.8. Again, to classify the critical points we need the second order partial derivatives. They are

$$\begin{aligned} f_{xx}(x, y) &= 10y \\ f_{yy}(x, y) &= 2x \\ f_{xy}(x, y) &= (1)(10x + y - 15) + y(1) = 10x + 2y - 15 \\ f_{yx}(x, y) &= (1)(5x + 2y - 15) + x(5) = 10x + 2y - 15 \end{aligned}$$

(Once again, we have computed both  $f_{xy}$  and  $f_{yx}$  to guard against mechanical errors.) We have already found, in Example 2.3.8, that the critical points are  $(0, 0)$ ,  $(0, 15)$ ,  $(3, 0)$  and  $(1, 5)$ . The classification is

critical point	$f_{xx}f_{yy} - f_{xy}^2$	$f_{xx}$	type
$(0, 0)$	$0 \times 0 - (-15)^2 < 0$		saddle point
$(0, 15)$	$150 \times 0 - 15^2 < 0$		saddle point
$(3, 0)$	$0 \times 6 - 15^2 < 0$		saddle point
$(1, 5)$	$50 \times 2 - 5^2 > 0$	50	local min

Here is a sketch of some level curves of our  $f(x, y)$ .  $f$  is negative in the shaded regions and  $f$  is positive in the unshaded regions. Again this is not needed to answer this



question, but can give you some idea as to what the graph of  $f$  looks like.

Example 2.3.17

Example 2.3.18

Find and classify all of the critical points of  $f(x, y) = x^3 + xy^2 - 3x^2 - 4y^2 + 4$ .

*Solution.* We know the drill now. We start by computing all of the partial derivatives of  $f$  up to order 2.

$$\begin{aligned} f &= x^3 + xy^2 - 3x^2 - 4y^2 + 4 \\ f_x &= 3x^2 + y^2 - 6x & f_{xx} &= 6x - 6 & f_{xy} &= 2y \\ f_y &= 2xy - 8y & f_{yy} &= 2x - 8 & f_{yx} &= 2y \end{aligned}$$

$f_x$  and  $f_y$  are defined everywhere. So the critical points are then the solutions of  $f_x = 0$ ,  $f_y = 0$ . That is

$$f_x = 3x^2 + y^2 - 6x = 0 \quad (\text{E1})$$

$$f_y = 2y(x - 4) = 0 \quad (\text{E2})$$

The second equation,  $2y(x - 4) = 0$ , is satisfied if and only if at least one of the two equations  $y = 0$  and  $x = 4$  is satisfied.

- When  $y = 0$ , equation (E1) forces  $x$  to obey

$$0 = 3x^2 + 0^2 - 6x = 3x(x - 2)$$

so that  $x = 0$  or  $x = 2$ .



- When  $x = 4$ , equation (E1) forces  $y$  to obey

$$0 = 3 \times 4^2 + y^2 - 6 \times 4 = 24 + y^2$$

which is impossible.

So, there are two critical points:  $(0,0)$ ,  $(2,0)$ . Here is a table that classifies the critical points.

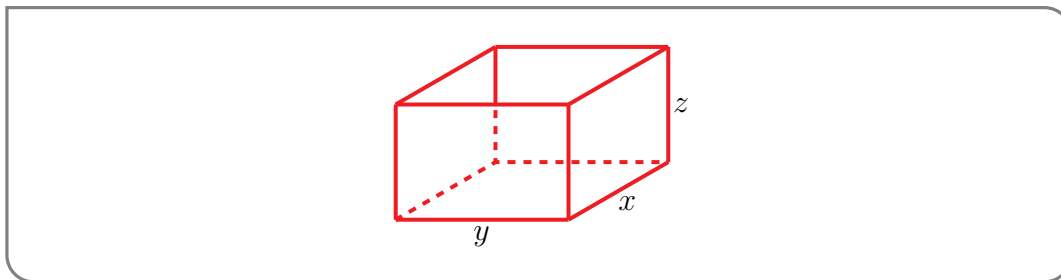
critical point	$f_{xx}f_{yy} - f_{xy}^2$	$f_{xx}$	type
$(0,0)$	$(-6) \times (-8) - 0^2 > 0$	$-6 < 0$	local max
$(2,0)$	$6 \times (-4) - 0^2 < 0$		saddle point

Example 2.3.18

Example 2.3.19

A manufacturer wishes to make an open rectangular box of given volume  $V$  using the least possible material. Find the design specifications.

*Solution.* Denote by  $x$ ,  $y$  and  $z$ , the length, width and height, respectively, of the box.



The box has two sides of area  $xz$ , two sides of area  $yz$  and a bottom of area  $xy$ . So the total surface area of material used is

$$S = 2xz + 2yz + xy$$

However the three dimensions  $x$ ,  $y$  and  $z$  are not independent. The requirement that the box have volume  $V$  imposes the constraint

$$xyz = V$$

We can use this constraint to eliminate one variable. Since  $z$  is at the end of the alphabet (poor  $z$ ), we eliminate  $z$  by substituting  $z = \frac{V}{xy}$ . Note that if  $x$  (or  $y$ ) is equal to zero then the volume of the box would equal zero. What is the point of a box with zero volume?! So if we assume the box has non-zero volume then  $x \neq 0$  and  $y \neq 0$ . So we have find the values of  $x$  and  $y$  that minimize the function

$$S(x, y) = \frac{2V}{y} + \frac{2V}{x} + xy$$

Let's start by finding the critical points of  $S$ . Since

$$S_x(x, y) = -\frac{2V}{x^2} + y$$

$$S_y(x, y) = -\frac{2V}{y^2} + x$$

Note that the partial derivatives are not defined for  $(x, y) = (0, 0)$  but we have already eliminated the case where  $x$  or  $y$  is equal to zero. So  $(x, y)$  is a critical point if and only if

$$x^2 y = 2V \tag{E1}$$

$$x y^2 = 2V \tag{E2}$$

Solving (E1) for  $y$  gives  $y = \frac{2V}{x^2}$ . Substituting this into (E2) gives

$$x \frac{4V^2}{x^4} = 2V \implies x^3 = 2V \implies x = \sqrt[3]{2V} \quad \text{and} \quad y = \frac{2V}{(2V)^{2/3}} = \sqrt[3]{2V}$$

As there is only one critical point, we would expect it to give the minimum<sup>15</sup>. But let's use the second derivative test to verify that at least the critical point is a local minimum. The various second partial derivatives are

$$\begin{aligned} S_{xx}(x, y) &= \frac{4V}{x^3} & S_{xx}(\sqrt[3]{2V}, \sqrt[3]{2V}) &= 2 \\ S_{xy}(x, y) &= 1 & S_{xy}(\sqrt[3]{2V}, \sqrt[3]{2V}) &= 1 \\ S_{yy}(x, y) &= \frac{4V}{y^3} & S_{yy}(\sqrt[3]{2V}, \sqrt[3]{2V}) &= 2 \end{aligned}$$

So

$$S_{xx}(\sqrt[3]{2V}, \sqrt[3]{2V}) S_{yy}(\sqrt[3]{2V}, \sqrt[3]{2V}) - S_{xy}(\sqrt[3]{2V}, \sqrt[3]{2V})^2 = 3 > 0 \quad S_{xx}(\sqrt[3]{2V}, \sqrt[3]{2V}) = 2 > 0$$

and, by Theorem 2.3.14.b,  $(\sqrt[3]{2V}, \sqrt[3]{2V})$  is a local minimum and the desired dimensions are

$$x = y = \sqrt[3]{2V} \quad z = \sqrt[3]{\frac{V}{4}}$$

Note that our solution has  $x = y$ . That's a good thing — the function  $S(x, y)$  is symmetric in  $x$  and  $y$ . Because the box has no top, the symmetry does not extend to  $z$ .

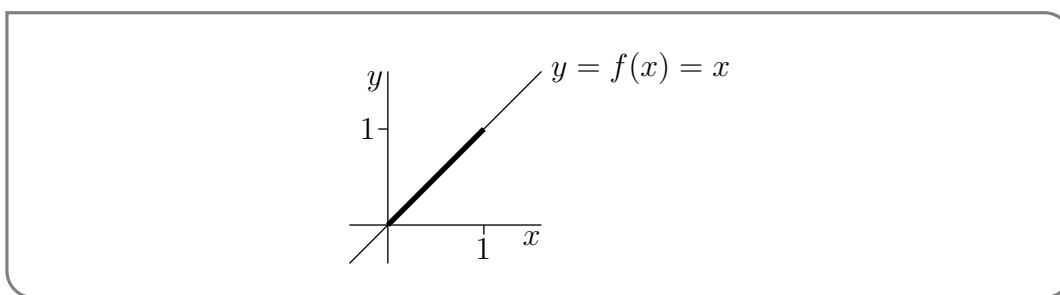
Example 2.3.19

<sup>15</sup> Indeed one can use the facts that  $0 < x < \infty$ , that  $0 < y < \infty$ , and that  $S \rightarrow \infty$  as  $x \rightarrow 0$  and as  $y \rightarrow 0$  and as  $x \rightarrow \infty$  and as  $y \rightarrow \infty$  to prove that the single critical point gives the global minimum.

## 2.4▲ Absolute Minima and Maxima

Of course a local maximum or minimum of a function need not be the absolute maximum or minimum. We'll now consider how to find the absolute maximum and minimum. Let's start by reviewing how one finds the absolute maximum and minimum of a function of one variable on an interval.

For concreteness, let's suppose that we want to find the extremal<sup>16</sup> values of a function  $f(x)$  on the interval  $0 \leq x \leq 1$ . If an extremal value is attained at some  $x = a$  which is in the interior of the interval, i.e. if  $0 < a < 1$ , then  $a$  is also a local maximum or minimum and so has to be a critical point of  $f$ . But if an extremal value is attained at a boundary point  $a$  of the interval, i.e. if  $a = 0$  or  $a = 1$ , then  $a$  need not be a critical point of  $f$ . This happens, for example, when  $f(x) = x$ . The largest value of  $f(x)$  on the interval  $0 \leq x \leq 1$  is 1 and is attained at  $x = 1$ , but  $f'(x) = 1$  is never zero, so that  $f$  has no critical points.



So to find the maximum and minimum of the function  $f(x)$  on the interval  $[0, 1]$ , you:

1. build up a list of all candidate points  $0 \leq a \leq 1$  at which the maximum or minimum could be attained, by finding all  $a$ 's for which either
  - (a)  $0 < a < 1$  and  $f'(a)$  does not exist or
  - (b)  $0 < a < 1$  and  $f'(a) = 0$  or
  - (c)  $a$  is a boundary point, i.e.  $a = 0$  or  $a = 1$ ;
2. and then you evaluate  $f(a)$  at each  $a$  on the list of candidates. The biggest of these candidate values of  $f(a)$  is the absolute maximum and the smallest of these candidate values is the absolute minimum.

The procedure for finding the maximum and minimum of a function of two variables  $f(x, y)$  in a set like, for example, the unit disk  $x^2 + y^2 \leq 1$ , is similar. You again:

1. build up a list of all candidate points  $(a, b)$  in the set at which the maximum or minimum could be attained, by finding all  $(a, b)$ 's for which either<sup>17</sup>
  - (a)  $(a, b)$  is in the interior of the set and  $f_x(a, b)$  or  $f_y(a, b)$  does not exist or
  - (b)  $(a, b)$  is in the interior of the set (for our example,  $a^2 + b^2 < 1$ ) and  $f_x(a, b) = f_y(a, b) = 0$  or

<sup>16</sup> Recall that "extremal value" means "either maximum value or minimum value".

<sup>17</sup> This is probably a good time to review the statement of Theorem 2.3.2.

- (c)  $(a, b)$  is a boundary<sup>18</sup> point, (for our example,  $a^2 + b^2 = 1$ ), and could give the maximum or minimum on the boundary — more about this shortly —
2. and then you evaluate  $f(a, b)$  at each  $(a, b)$  on the list of candidates. The biggest of these candidate values of  $f(a, b)$  is the absolute maximum and the smallest of these candidate values is the absolute minimum.

The boundary of a set in  $\mathbb{R}^2$  (like  $x^2 + y^2 \leq 1$ ) is a curve (like  $x^2 + y^2 = 1$ ). This curve is a one dimensional set, meaning that it is like a deformed  $x$ -axis. We can find the maximum and minimum of  $f(x, y)$  on this curve by converting  $f(x, y)$  into a function of one variable (on the curve) and using the standard function of one variable techniques. This is best explained by some examples.

Example 2.4.1

Find the maximum and minimum values of  $f(x, y) = x^3 + xy^2 - 3x^2 - 4y^2 + 4$  on the disk  $x^2 + y^2 \leq 1$ .

*Solution.* Again, we first find all critical points, and then we analyze the boundary.

*Interior:* If  $f$  takes its maximum or minimum value at a point in the interior,  $x^2 + y^2 < 1$ , then that point must be a critical point of  $f$ . To find the critical points<sup>19</sup> we compute the first order derivatives.

$$f_x = 3x^2 + y^2 - 6x \quad f_y = 2xy - 8y$$

These are polynomials (in two variables) and they are defined everywhere. So the critical points are the solutions of

$$f_x = 3x^2 + y^2 - 6x = 0 \tag{E1}$$

$$f_y = 2y(x - 4) = 0 \tag{E2}$$

The second equation,  $2y(x - 4) = 0$ , is satisfied if and only if at least one of the two equations  $y = 0$  and  $x = 4$  is satisfied.

- When  $y = 0$ , equation (E1) forces  $x$  to obey

$$0 = 3x^2 + 0^2 - 6x = 3x(x - 2)$$

so that  $x = 0$  or  $x = 2$ .

- When  $x = 4$ , equation (E1) forces  $y$  to obey

$$0 = 3 \times 4^2 + y^2 - 6 \times 4 = 24 + y^2$$

which is impossible.

18 It should intuitively obvious from a sketch that the boundary of the disk  $x^2 + y^2 \leq 1$  is the circle  $x^2 + y^2 = 1$ . But if you really need a formal definition, here it is. A point  $(a, b)$  is on the boundary of a set  $S$  if there is a sequence of points in  $S$  that converges to  $(a, b)$  and there is also a sequence of points in the complement of  $S$  that converges to  $(a, b)$ .

19 We actually found the critical points in Example 2.3.18. But, for the convenience of the reader, we'll repeat that here.

So, there are only two critical points:  $(0,0)$ ,  $(2,0)$ .

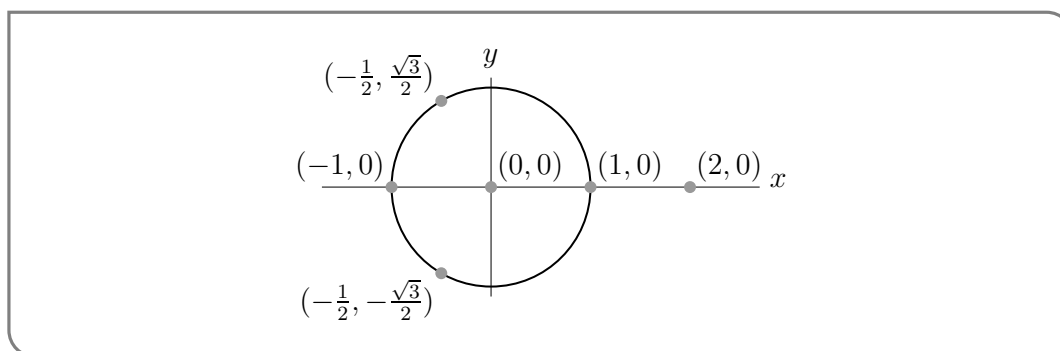
*Boundary:* Our boundary is  $x^2 + y^2 = 1$ . We know that  $(x,y)$  satisfies  $x^2 + y^2 = 1$ , and hence  $y^2 = 1 - x^2$ . Examining the formula for  $f(x,y)$ , we see that it contains only even<sup>20</sup> powers of  $y$ , so we can eliminate  $y$  by substituting  $y^2 = 1 - x^2$  into the formula.

$$f = x^3 + x(1 - x^2) - 3x^2 - 4(1 - x^2) + 4 = x + x^2$$

The max and min of  $x + x^2$  for  $-1 \leq x \leq 1$  must occur either

- when  $x = -1$  ( $\Rightarrow y = f = 0$ ) or
- when  $x = +1$  ( $\Rightarrow y = 0, f = 2$ ) or
- when  $0 = \frac{d}{dx}(x + x^2) = 1 + 2x$  (so  $x = -\frac{1}{2}, y = \pm\sqrt{\frac{3}{4}}, f = -\frac{1}{4}$ ).

Here is a sketch showing all of the points that we have identified.



Note that the point  $(2,0)$  is outside the allowed region<sup>21</sup>. So all together, we have the following candidates for max and min, with the max and min indicated.

point	$(0,0)$	$(-1,0)$	$(1,0)$	$(-\frac{1}{2}, \pm\frac{\sqrt{3}}{2})$
value of $f$	4	2	0	$-\frac{1}{4}$
	max			min

Example 2.4.1

Example 2.4.2

Find the maximum and minimum values of  $f(x,y) = xy - x^3y^2$  when  $(x,y)$  runs over the square  $0 \leq x \leq 1, 0 \leq y \leq 1$ .

<sup>20</sup> If it contained odd powers too, we could consider the cases  $y \geq 0$  and  $y \leq 0$  separately and substitute  $y = \sqrt{1 - x^2}$  in the former case and  $y = -\sqrt{1 - x^2}$  in the latter case.

<sup>21</sup> We found  $(2,0)$  as a solution to the critical point equations (E1), (E2). That's because, in the course of solving those equations, we ignored the constraint that  $x^2 + y^2 \leq 1$ .

*Solution.* As usual, let's examine the critical points and boundary in turn.

*Interior:* If  $f$  takes its maximum or minimum value at a point in the interior,  $0 < x < 1$ ,  $0 < y < 1$ , then that point must be a critical point of  $f$ . To find the critical points we compute the first order derivatives.

$$f_x(x, y) = y - 3x^2y^2 \quad f_y(x, y) = x - 2x^3y$$

Again, these functions are polynomials in two variables and they are smooth everywhere in their domain, so the first order partial derivatives exist everywhere in the interior. This means that the critical points are the solutions of

$$\begin{aligned} f_x = 0 &\iff y(1 - 3x^2y) = 0 &\iff y = 0 \text{ or } 3x^2y = 1 \\ f_y = 0 &\iff x(1 - 2x^2y) = 0 &\iff x = 0 \text{ or } 2x^2y = 1 \end{aligned}$$

- If  $y = 0$ , we cannot have  $2x^2y = 1$ , so we must have  $x = 0$ .
- If  $3x^2y = 1$ , we cannot have  $x = 0$ , so we must have  $2x^2y = 1$ . Dividing gives  $1 = \frac{3x^2y}{2x^2y} = \frac{3}{2}$  which is impossible.

So the only critical point in the square is  $(0, 0)$ . There  $f = 0$ . *Boundary:* The region is a square, so its boundary consists of its four sides.

- First, we look at the part of the boundary with  $x = 0$ . On that entire side  $f = 0$ .
- Next, we look at the part of the boundary with  $y = 0$ . On that entire side  $f = 0$ .
- Next, we look at the part of the boundary with  $y = 1$ . There  $f = f(x, 1) = x - x^3$ . To find the maximum and minimum of  $f(x, y)$  on the part of the boundary with  $y = 1$ , we must find the maximum and minimum of  $x - x^3$  when  $0 \leq x \leq 1$ .

Recall that, in general, the maximum and minimum of a function  $h(x)$  on the interval  $a \leq x \leq b$ , must occur either at  $x = a$  or at  $x = b$  or at an  $x$  for which either  $h'(x) = 0$  or  $h'(x)$  does not exist. In this case,  $\frac{d}{dx}(x - x^3) = 1 - 3x^2$ , so the max and min of  $x - x^3$  for  $0 \leq x \leq 1$  must occur

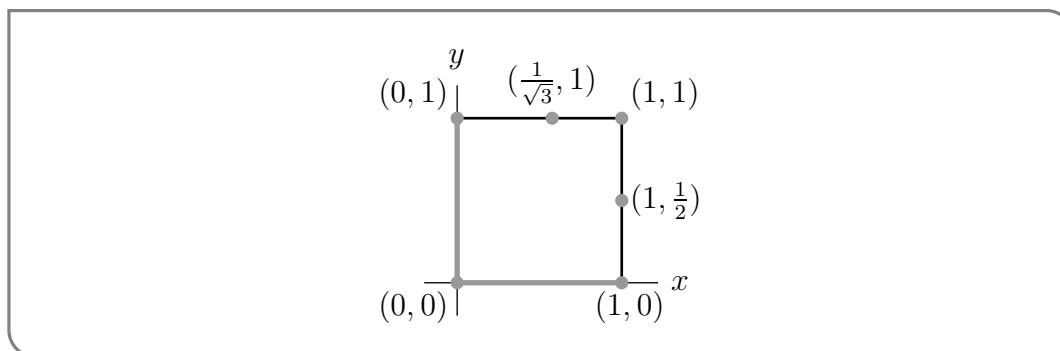
- either at  $x = 0$ , where  $f = 0$ ,
- or at  $x = \frac{1}{\sqrt{3}}$ , where  $f = \frac{2}{3\sqrt{3}}$ ,
- or at  $x = 1$ , where  $f = 0$ .

- Finally, we look at the part of the boundary with  $x = 1$ . There  $f = f(1, y) = y - y^2$ . As  $\frac{d}{dy}(y - y^2) = 1 - 2y$ , the only critical point of  $y - y^2$  is at  $y = \frac{1}{2}$ . So the max and min of  $y - y^2$  for  $0 \leq y \leq 1$  must occur

- either at  $y = 0$ , where  $f = 0$ ,
- or at  $y = \frac{1}{2}$ , where  $f = \frac{1}{4}$ ,
- or at  $y = 1$ , where  $f = 0$ .

All together, we have the following candidates for max and min, with the max and min indicated.

point	$(0, 0)$	$(0, 0 \leq y \leq 1)$	$(0 \leq x \leq 1, 0)$	$(1, 0)$	$(1, \frac{1}{2})$	$(1, 1)$	$(0, 1)$	$(\frac{1}{\sqrt{3}}, 1)$
value of $f$	0	0	0	0	$\frac{1}{4}$	0	0	$\frac{2}{3\sqrt{3}} \approx 0.385$
	min	min	min	min		min	min	max



Example 2.4.2

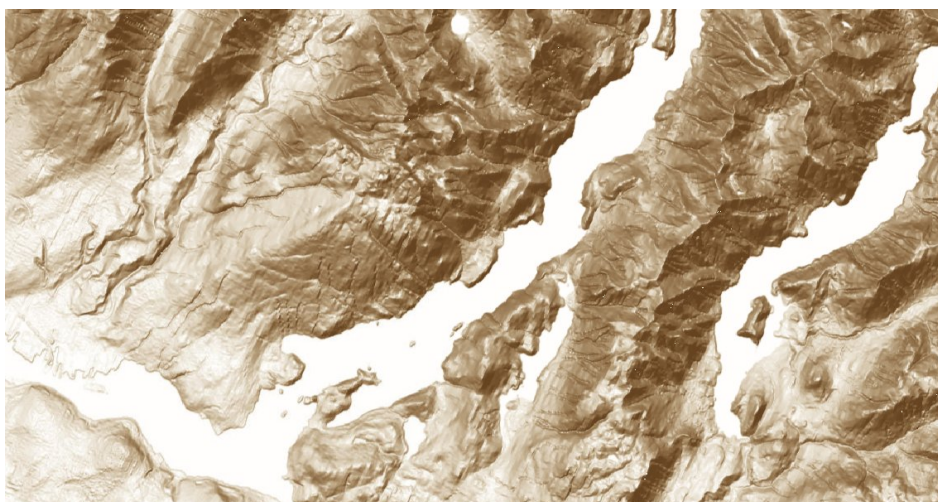
**Warning 2.4.3** (Checking Entire Boundaries).

A common misconception when students are first learning about “checking boundaries” is that the absolute extrema will occur on the “corners” of the boundaries. In the example we just finished, Example 2.4.2, the four corners of our square boundary were indeed points we needed to check. But if we had *only* checked the corners, we wouldn’t have found the absolute maximum.

In your homework, if you notice that the extrema often occur at “corners” of boundaries, or at point with  $x$  or  $y$  equal to 0, you should not take this to be a general rule.

To really see why corners don’t need to be important, consider the image<sup>22</sup> below of an area northeast of UBC. The central body of water in the image is Indian Arm. Indian Arm extends into the ocean, so its elevation is pretty close to sea level. If we’re thinking of the  $z$  axis as height above sea level, the surface of Indian Arm is probably the *global minimum height* in the rectangular region shown. So, the global minimum along the boundary is *not* at a corner. It’s somewhere in the middle of the left vertical boundary segment.

22 image generated by Natural Resources Canada’s [Atlas of Canada - Toporama](#) and shared under the [open government license](#)



Similarly, looking at the mountains in the image, there's no reason to imagine the absolute highest point along the boundary must specifically happen at a *corner*.

#### Example 2.4.4

Find the high and low points of the surface  $z = \sqrt{x^2 + y^2}$  with  $(x, y)$  varying over the square  $|x| \leq 1, |y| \leq 1$ .

*Solution.* The function  $f(x, y) = \sqrt{x^2 + y^2}$  has a particularly simple geometric interpretation — it is the distance from the point  $(x, y)$  to the origin. So

- the minimum of  $f(x, y)$  is achieved at the point in the square that is nearest the origin — namely the origin itself. So  $(0, 0, 0)$  is the lowest point on the surface and is at height 0.
- The maximum of  $f(x, y)$  is achieved at the points in the square that are farthest from the origin — namely the four corners of the square  $(\pm 1, \pm 1)$ . At those four points  $z = \sqrt{2}$ . So the highest points on the surface are  $(\pm 1, \pm 1, \sqrt{2})$ .

Even though we have already answered this question, it will be instructive to see what we would have found if we had followed our usual protocol. The partial derivatives of  $f(x, y) = \sqrt{x^2 + y^2}$  are defined for  $(x, y) \neq (0, 0)$  and are

$$f_x(x, y) = \frac{x}{\sqrt{x^2 + y^2}} \quad f_y(x, y) = \frac{y}{\sqrt{x^2 + y^2}}$$

- As we mentioned above, at the point  $(x, y) = (0, 0)$  the partial derivatives are not defined. But  $(0, 0)$  is inside the interior of the domain of our function. Therefore,  $(0, 0)$  is a critical point.
- There are no other critical points because
  - $f_x = 0$  only for  $x = 0$ , and
  - $f_y = 0$  only for  $y = 0$ .
  - So  $(0, 0)$  is the only critical point because  $f_x$  and  $f_y$  are not defined there.



- The boundary of the square consists of its four sides. One side is

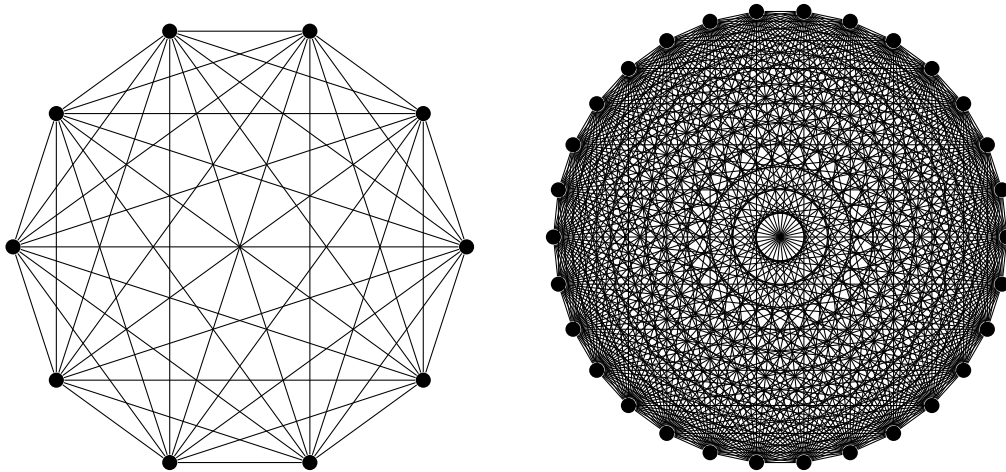
$$\{ (x, y) \mid x = 1, -1 \leq y \leq 1 \}$$

On this side  $f = \sqrt{1 + y^2}$ . As  $\sqrt{1 + y^2}$  increases with  $|y|$ , the smallest value of  $f$  on that side is 1 (when  $y = 0$ ) and the largest value of  $f$  is  $\sqrt{2}$  (when  $y = \pm 1$ ). The same thing happens on the other three sides. The maximum value of  $f$  is achieved at the four corners. Note that  $f_x$  and  $f_y$  are both nonzero at all four corners.

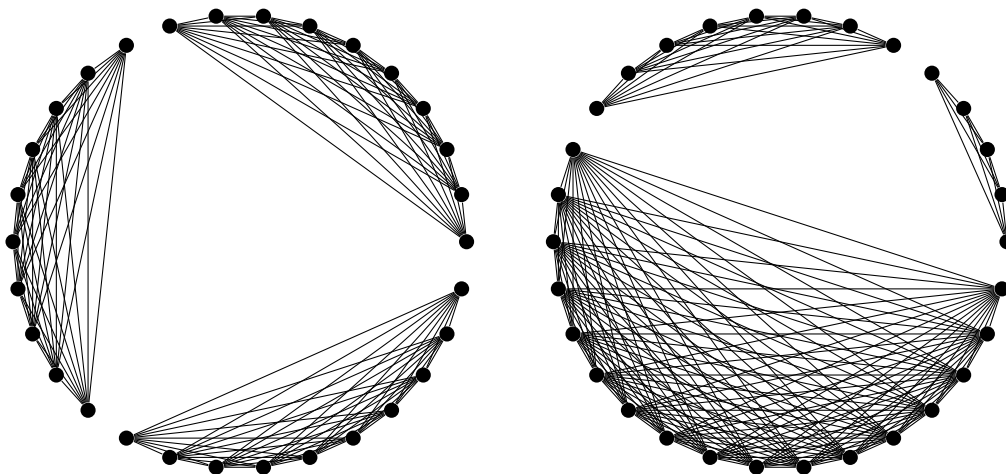
Example 2.4.4

Example 2.4.5 (Disconnecting a Complete Graph)

In graph theory, a *complete graph* is a collection of  $n$  vertices (visualized as dots), every pair of which is connected by an edge (visualized as lines). The complete graphs on 10 vertices and on 30 vertices are shown below.



Suppose you start with the complete graph on 30 vertices. You delete edges (but not vertices) one-by-one until the graph is broken into three parts. Every part has at least one vertex (otherwise it wouldn't be a part, it would be a nothing) and there are no edges between vertices of different parts. Some possibilities are shown below to demonstrate.



What is the minimum number of edges you could have deleted, in order to break the graph into three pieces?

*Solution.* Let's name the pieces  $X$ ,  $Y$ , and  $W$ , and say the numbers of vertices they contain are  $x$ ,  $y$ , and  $w$ , respectively. Then  $x \geq 1$ ,  $y \geq 1$ ,  $w \geq 1$ , and  $x + y + w = 30$ .

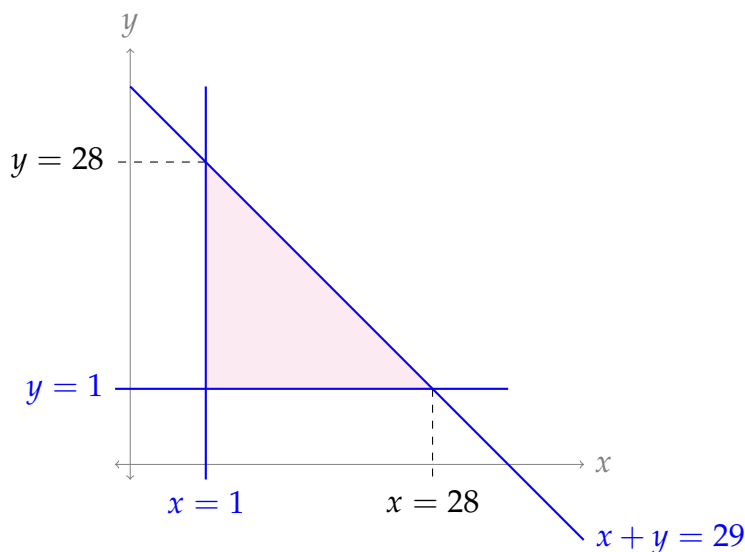
For every vertex in one piece of the broken graph, you must have deleted the edges connecting it to every vertex in every other piece. So, to delete all the edges from  $X$  to  $Y$ , you deleted at least  $xy$  edges; to delete all the edges from  $X$  to  $W$ , you deleted at least  $xw$  edges; and to delete all the edges from  $Y$  to  $W$ , you deleted at least  $yw$  edges. So all together, you deleted at least this many edges:

$$xy + xw + yw$$

Since  $x + y + w = 30$ , we can eliminate one of these from our expression, and say the minimum number of edges deleted was:

$$\begin{aligned} f(x, y) &= xy + x(30 - x - y) + y(30 - x - y) \\ &= 30x + 30y - x^2 - xy - y^2 \end{aligned}$$

The domain of this function is all integer pairs in the region bounded by  $x \geq 1$ ,  $y \geq 1$ , and  $x + y \leq 29$ .



To find the minimum value of  $f(x, y)$  in this region, we should check for critical points, and check all three boundary lines.

- First, let's check for critical points.

$$f(x, y) = 30x + 30y - x^2 - xy - y^2$$

$$f_x = 30 - 2x - y$$

$$f_y = 30 - 2y - x$$

Solving  $f_x = 0$  for  $y$ , we find  $y = 30 - 2x$ . Plugging into the equation  $f_y = 0$ , we get:

$$0 = f_y = 30 - 2(30 - 2x) - x$$

$$= 3x - 30$$

$$x = 10$$

$$y = 30 - 2x = 10$$

So, our only critical point is  $(10, 10)$ , and this is inside our region.

$$f(10, 10) = 300 + 300 - 100 - 100 - 100 = 300$$

- Second, let's check the boundary line  $y = 1$ ,  $1 \leq x \leq 28$ . On this portion of the boundary:

$$\begin{aligned} f(x, y) &= 30x + 30y - x^2 - xy - y^2 \\ &= 30x + 30 - x - x - 1 \\ &= 28x + 29 \end{aligned}$$

This is an increasing function, so its minimum will be at the smallest value of  $x$  in our interval:  $x = 1$ .

$$f(1, 1) = 57$$

- Third, we check the boundary line  $x = 1$ ,  $1 \leq y \leq 28$ . On this portion of the boundary:

$$\begin{aligned} f(x, y) &= 30x + 30y - x^2 - xy - y^2 \\ &= 30 + 30y - 1 - y - y \\ &= 28y + 29 \end{aligned}$$

This is an increasing function, so its minimum will be at the smallest value of  $y$  in our interval:  $y = 1$ .

$$f(1, 1) = 57$$

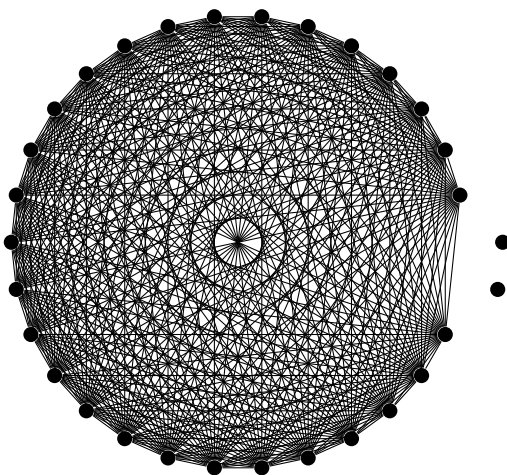
- Fourth, we check the final boundary line,  $y = 29 - x$ ,  $1 \leq x \leq 28$ . On this portion of the boundary:

$$\begin{aligned} f(x, y) &= 30x + 30y - x^2 - xy - y^2 \\ &= 30x + 30(29 - x) - x^2 - x(29 - x) - (29 - x)^2 \\ &= -x^2 + 29x + 29 \end{aligned}$$

The one-variable function  $g(x) = -x^2 + 29x + 29$  is a parabola pointing down, so its minimum will occur at an endpoint of our interval:  $x = 1$  or  $x = 28$ .

$$f(1, 28) = 57 \quad f(28, 1) = 57$$

Comparing the values from the four bullet points, we find the minimum number of edges we could have deleted in order to break the complete graph into 3 pieces is 57. We achieve that minimum by having two pieces of one vertex each, and the remaining piece with all other vertices.



Remark 1: making use of sketching and symmetry can reduce the amount of work involved in solving this problem. If we recognize that  $f(x, y)$  is a paraboloid opening down, then we know its critical point will actually be an absolute max – not the minimum we’re looking for.

We can see the  $x$  and  $y$  are symmetric in  $f(x, y)$  and in our region, so we also could have checked only the boundary  $x = 1$ , and not the boundary  $y = 1$ , understanding that their minimum values would be the same.

Remark 2: Our model domain for this problem actually restricts  $x$  and  $y$  to whole-number values, as opposed to real numbers. We showed that 57 was the minimum value of  $f(x, y)$  over all *real* numbers in the sketched region. Since whole numbers are themselves reals, and the minimum occurred at integer value of  $x$  and  $y$  (i.e. the minimum is in our model domain), we can be sure that 57 is the minimum over all whole numbers in our domain. If the minimum had occurred at, say  $x = \frac{1}{2}$  and  $y = \frac{1}{2}$ , then it wouldn’t have been in our model domain – and this would be a problem for a different course!

Example 2.4.5

## 2.5▲ Lagrange Multipliers

In the last section we had to solve a number of problems of the form “What is the maximum value of the function  $f$  on the curve  $C$ ?” In those examples, the curve  $C$  was simple enough that we could reduce the problem to finding the maximum of a function of one variable. For more complicated problems this reduction might not be possible. In this section, we introduce another method for solving such problems. First some nomenclature.

**Definition 2.5.1.**

A problem of the form

“Find the maximum and minimum values of the function  $f(x, y)$  for  $(x, y)$  on the curve  $g(x, y) = 0$ .”

is one type of *constrained optimization* problem. The function being maximized or minimized,  $f(x, y)$ , is called the *objective function*. The function,  $g(x, y)$ , whose zero set is the curve of interest, is called the *constraint function*.

Such problems are quite common. As we said above, we have already encountered them in the last section on absolute maxima and minima, when we were looking for the extreme values of a function on the boundary of a region. In economics “utility functions” are used to model the relative “usefulness” or “desirability” or “preference” of various economic choices. For example, a utility function  $U(w, \kappa)$  might specify the relative level of satisfaction a consumer would get from purchasing a quantity  $w$  of wine and  $\kappa$  of coffee. If the consumer wants to spend \$100 and wine costs \$20 per unit and coffee costs \$5 per unit, then the consumer would like to maximize  $U(w, \kappa)$  subject to the constraint that  $20w + 5\kappa = 100$ .

To this point we have always solved such constrained optimization problems by solving  $g(x, y) = 0$  for  $y$  as a function of  $x$  (or for  $x$  as a function of  $y$ ). However, quite often the function  $g(x, y)$  is so complicated that one cannot explicitly solve  $g(x, y) = 0$  for  $y$  as a function of  $x$  or for  $x$  as a function of  $y$  and one also cannot explicitly parametrize  $g(x, y) = 0$ . Or sometimes you can, for example, solve  $g(x, y) = 0$  for  $y$  as a function of  $x$ , but the resulting solution is so complicated that it is really hard, or even virtually impossible, to work with. Direct attacks become even harder in higher dimensions when, for example, we wish to optimize a function  $f(x, y, z)$  subject to a constraint  $g(x, y, z) = 0$ .

There is another procedure called the method of “Lagrange<sup>23</sup> multipliers” that comes to our rescue in these scenarios. Here is the two-dimensional version of the method. There are obvious analogues in other dimensions.

23 Joseph-Louis Lagrange was actually born Giuseppe Lodovico Lagrangia in Turin, Italy in 1736. He moved to Berlin in 1766 and then to Paris in 1786. He eventually acquired French citizenship and then the French claimed he was a French mathematician, while the Italians continued to claim that he was an Italian mathematician.

**Theorem 2.5.2 (Lagrange Multipliers).**

Let  $f(x, y)$  and  $g(x, y)$  have continuous first partial derivatives in a region of  $\mathbb{R}^2$  that contains the surface  $S$  given by the equation  $g(x, y) = 0$ . Further assume that  $g(x, y)$  has no critical points on  $S$ .

If  $f$ , restricted to the surface  $S$ , has a local extreme value at the point  $(a, b)$  on  $S$ , then there is a real number  $\lambda$  such that

$$\begin{aligned}f_x(a, b) &= \lambda g_x(a, b) \\f_y(a, b) &= \lambda g_y(a, b)\end{aligned}$$

The number  $\lambda$  is called a *Lagrange multiplier*.

A proof of this theorem can be found in Appendix A.6.

So to find the maximum and minimum values of  $f(x, y)$  on a surface  $g(x, y) = 0$ , assuming that both the objective function  $f(x, y)$  and constraint function  $g(x, y)$  have continuous first partial derivatives, and that  $g(x, y)$  has no critical points, you

1. build up a list of candidate points  $(x, y, z)$  by finding all solutions to the equations

$$\begin{aligned}f_x(x, y) &= \lambda g_x(x, y) \\f_y(x, y) &= \lambda g_y(x, y) \\g(x, y) &= 0\end{aligned}$$

Note that there are three equations and three unknowns, namely  $x$ ,  $y$ , and  $\lambda$ .

2. Then you evaluate  $f(x, y)$  at each  $(x, y)$  on the list of candidates. The biggest of these candidate values is the absolute maximum, if an absolute maximum exists. The smallest of these candidate values is the absolute minimum, if an absolute minimum exists..

Theorem 2.5.2 can be extended to functions of more variables in a natural way. Using higher-dimensional Lagrange isn't in our learning goals, but for interest, we want you to see how easily the method generalizes. The calculus is the same – it's only the algebra that gets longer.

**Theorem 2.5.3** ((Optional) Lagrange Multipliers for Functions of Three Variables).

Let  $f(x, y, z)$  and  $g(x, y, z)$  have continuous first partial derivatives in a region of  $\mathbb{R}^3$  that contains the surface  $S$  given by the equation  $g(x, y, z) = 0$ . Further assume that  $g(x, y, z)$  has no critical points on  $S$ .

If  $f$ , restricted to the surface  $S$ , has a local extreme value at the point  $(a, b, c)$  on  $S$ , then there is a real number  $\lambda$  such that

$$\begin{aligned}f_x(a, b, c) &= \lambda g_x(a, b, c) \\f_y(a, b, c) &= \lambda g_y(a, b, c) \\f_z(a, b, c) &= \lambda g_z(a, b, c)\end{aligned}$$

The number  $\lambda$  is called a *Lagrange multiplier*.

Now for a bunch of examples.

**Example 2.5.4**

Find the maximum and minimum of the function  $x^2 - 10x - y^2$  on the ellipse whose equation is  $x^2 + 4y^2 = 16$ .

*Solution.* For this first example, we'll do out the algebra in truly gory detail. Once you get the hang of it, it'll go much faster.

Our objective function (the one we want to maximize and/or minimize) is  $f(x, y) = x^2 - 10x - y^2$  and the constraint function is  $g(x, y) = x^2 + 4y^2 - 16$ . To apply the method of Lagrange multipliers we start by computing the first-order derivatives of these functions.

$$f_x = 2x - 10 \quad f_y = -2y \quad g_x = 2x \quad g_y = 8y$$

So, according to the method of Lagrange multipliers, we need to find all solutions to the following system of equations.

$$f_x = \lambda g_x \qquad 2x - 10 = \lambda(2x) \qquad \text{(E1)}$$

$$f_y = \lambda g_y \qquad -2y = \lambda(8y) \qquad \text{(E2)}$$

$$g(x, y) = 0 \qquad x^2 + 4y^2 - 16 = 0 \qquad \text{(E3)}$$

**(E1)** In equation (E1), if  $2x$  is nonzero, then we can divide both sides of the equation by it,

to find  $\lambda = \frac{2x-10}{2x}$ , i.e.  $\lambda = \frac{x-5}{x}$ . If  $2x = 0$ , then the equation becomes  $-10 = 0\lambda$ , which is not true for any  $\lambda$ .

**(E2)** In equation (E2), if  $8y$  is nonzero, then we can divide both sides of the equation by it,

to find  $\lambda = \frac{-2y}{8y}$ , i.e.  $\lambda = -\frac{1}{4}$ . If  $8y = 0$ , then we also get a solution  $y = 0$  for any  $\lambda$ .

**(E1)+(E2)** We need all three equations to be true at the same time (that is, for the same values of  $x$ ,  $y$ , and  $\lambda$ ). We've found two ways for both (E1) and (E2) to be true.

- First way:  $\lambda = \frac{x-5}{x}$  and  $\lambda = -\frac{1}{4}$
- Second way:  $\lambda = \frac{x-5}{x}$  and  $y = 0$

(E3) Now we'll see which points make (E1) and (E2) true while *also* making (E3) true.

- First way:  $\lambda = \frac{x-5}{x}$  and  $\lambda = -\frac{1}{4}$

$$\begin{aligned} \lambda &= \frac{x-5}{x} \text{ and } \lambda = -\frac{1}{4} \\ \implies & \frac{x-5}{x} = -\frac{1}{4} \\ \implies & -4x + 20 = x \\ \implies & x = 4 \end{aligned}$$

In order to satisfy (E3):

$$\begin{aligned} 0 &= 4^2 + 4y^2 - 16 \\ 0 &= y \end{aligned}$$

So, the point  $(x, y) = (4, 0)$  satisfies all three equations.

- Second way:  $\lambda = \frac{x-5}{x}$  and  $y = 0$ . If  $y = 0$ , then from E3, we see

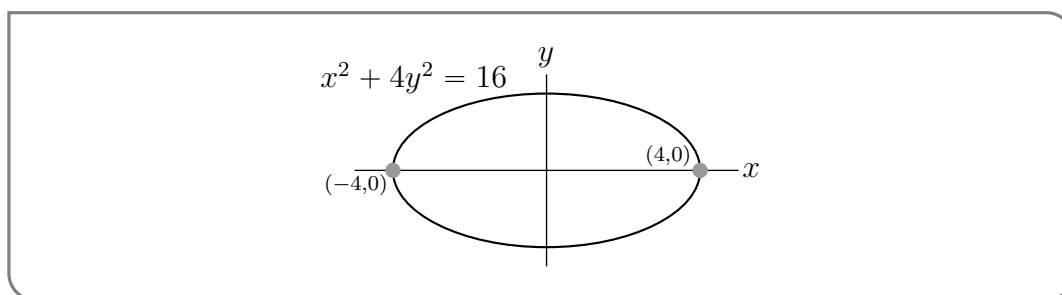
$$\begin{aligned} 0 &= x^2 + 4 \cdot 0^2 - 16 \\ 16 &= x^2 \\ x &= \pm 4 \end{aligned}$$

So the points to consider are  $(x, y) = (\pm 4, 0)$ .

Now we've found the only possible solutions to all three equations:  $(\pm 4, 0)$ . ( $\lambda$  has to exist, but we don't actually care what it is.) So the method of Lagrange multipliers, Theorem 2.5.2, gives that the only possible locations of the maximum and minimum of the function  $f$  are  $(4, 0)$  and  $(-4, 0)$ . To complete the problem, we only have to compute  $f$  at those points.

point	$(4, 0)$	$(-4, 0)$
value of $f$	-24	56
	min	max

Hence the maximum value of  $x^2 - 10x - y^2$  on the ellipse is 56 and the minimum value is -24.





## Example 2.5.4

In the previous example, we had to make a lot of decisions about how to solve for the solutions to the system of three equations. Actually, we can start our Lagrange system-solving the same way every time. The first observation we make is that the partial derivatives of  $g$  can be 0, or nonzero. If they're zero, this may or may not lead to a solution; if they're nonzero, this tells us something about  $\lambda$ .

In the textbook and problem book, we will consistently use the same method to solve the system of equations. It's certainly not the only way, and you are free to use other methods. Once you get used to the computations, you'll probably start finding ways to make them faster based on the specifics of individual problems.

## Example 2.5.5 (Solving Lagrange in General)

Suppose you want to find all points  $(x, y)$  for which a solution exists to the system below.

$$f_x = \lambda g_x \quad (\text{E1})$$

$$f_y = \lambda g_y \quad (\text{E2})$$

$$g(x, y) = 0 \quad (\text{E3})$$

where  $\lambda$  is some real constant. Our method below will hinge on the observation from the last example that we get different solutions for zero vs. nonzero partial derivatives of the constraint.

- If  $g_x \neq 0$  and  $g_y \neq 0$ , then from (E1) we see  $\lambda = \frac{f_x}{g_x}$ , and from (E2) we see  $\lambda = \frac{f_y}{g_y}$ . So, choosing a pair  $(x, y)$  such that

$$\frac{f_x}{g_x} = \frac{f_y}{g_y}$$

means that for some  $\lambda$ , that pair makes (E1) and (E2) true. Simplify the equation above to find the necessary relationship between  $x$  and  $y$ , then find which pairs with that relationship make (E3) true.

- If  $g_x = 0$ , then from (E1) we see also  $f_x = 0$ . Then (E1) is true for any  $\lambda$  that we like. We can check that there exists some  $\lambda$  that makes (E2) true as well. Then, we find the points  $(x, y)$  that make (E3) true as well as  $g_x = f_x = 0$ .
- If  $g_y = 0$ , then from (E2) we see also  $f_y = 0$ . Then (E2) is true for any  $\lambda$  that we like. We can check that there exists some  $\lambda$  that makes (E1) true as well. Then, we find the points  $(x, y)$  that make (E3) true as well as  $g_y = f_y = 0$ .

Sometimes, one or more of these cases won't lead to any solutions. In Example 2.5.4, we were immediately able to discard the possibility  $g_x = 0$ , because it didn't lead to a solution. Once you're practiced with these types of problems, you'll often see quite quickly which cases you get to discard.

## Example 2.5.5

We'll apply our three-case breakdown in subsequent examples.

Example 2.5.6

Find the minimum and maximum values of the objective function

$$f(x, y) = \ln(x^2 - 2x + 5) + \ln(y^2 - 4y + 13)$$

subject to the constraint

$$x^2 - 2x + y^2 - 4y = 20$$

*Solution.* Our constraint function is

$$g(x, y) = x^2 - 2x + y^2 - 4y - 20 = 0$$

We start by setting up the first two equations from the method of Lagrange multipliers.

$$f_x = \lambda g_x \quad \frac{2x - 2}{x^2 - 2x + 5} = \lambda(2x - 2) \quad (\text{E1})$$

$$f_y = \lambda g_y \quad \frac{2y - 4}{y^2 - 4y + 13} = \lambda(2y - 4) \quad (\text{E2})$$

$$g(x, y) = 0 \quad x^2 - 2x + y^2 - 4y = 20 \quad (\text{E3})$$

Now we consider our three cases.

- $g_x \neq 0$  and  $g_y \neq 0$ . From (E1), this means  $\lambda = \frac{1}{x^2 - 2x + 5}$ . From (E2),  $\lambda = \frac{1}{y^2 - 4y + 13}$ .

$$\begin{aligned} \frac{1}{x^2 - 2x + 5} &= \frac{1}{y^2 - 4y + 13} \\ x^2 - 2x + 5 &= y^2 - 4y + 13 \\ x^2 - 2x &= y^2 - 4y + 8 \end{aligned}$$

This gives us the relationship between  $x$  and  $y$  that must hold for (E1) and (E2) to be true under the assumption  $g_x \neq 0$  and  $g_y \neq 0$ . Now, in order for (E3) to be true as well:

$$\begin{aligned} 0 &= (x^2 - 2x) + y^2 - 4y - 20 \\ &= (y^2 - 4y + 8) + y^2 - 4y - 20 \\ &= 2y^2 - 8y - 12 \\ 0 &= y^2 - 4y - 6 \\ y &= \frac{4 \pm \sqrt{16 - 4(1)(-6)}}{2} = \frac{4 \pm \sqrt{40}}{2} = 2 \pm \sqrt{10} \\ \text{So, } 0 &= (x^2 - 2x) + y^2 - 4y - 20 \\ &= x^2 - 2x + (2 \pm \sqrt{10})^2 - 4(2 \pm \sqrt{10}) - 20 \\ &= x^2 - 2x + (4 \pm 4\sqrt{10} + 10) - 8 \mp 4\sqrt{10} - 20 \end{aligned}$$

Note  $\pm 4\sqrt{2} \mp 4\sqrt{2} = 0$

$$\begin{aligned} &= x^2 - 2x + 4 + 10 - 8 - 20 \\ &= x^2 - 2x - 14 \\ x &= \frac{2 \pm \sqrt{4 - 4(-14)}}{2} = \frac{2 \pm 2\sqrt{15}}{2} = 1 \pm \sqrt{15} \end{aligned}$$

This gives us four points to consider:

$(1 + \sqrt{15}, 2 + \sqrt{10})$ ,  $(1 - \sqrt{15}, 2 + \sqrt{10})$ ,  $(1 + \sqrt{15}, 2 - \sqrt{10})$ , and  $(1 - \sqrt{15}, 2 - \sqrt{10})$ .

- If  $g_x = 0$ , then  $x = 1$ , and (E1) is true for any  $\lambda$ . Then we can choose whatever  $\lambda$  is necessary to make (E2) true. By (E3):

$$\begin{aligned} 0 &= x^2 - 2x + y^2 - 4y - 20 \\ &= 1 - 2 + y^2 - 4y - 20 \\ &= y^2 - 4y - 21 \\ &= (y - 7)(y + 3) \\ y &= 7, \quad y = -3 \end{aligned}$$

This gives us two points to consider:  $(1, 7)$  and  $(1, -3)$ .

- If  $g_y = 0$ , then  $y = 2$ , and (E2) is true for any  $\lambda$ . Then we can choose whatever  $\lambda$  is necessary to make (E1) true. By (E3):

$$\begin{aligned} 0 &= x^2 - 2x + y^2 - 4y - 20 \\ &= x^2 - 2x + 4 - 8 - 20 \\ &= x^2 - 2x - 24 \\ &= (x - 6)(x + 4) \\ x &= 6, \quad x = -4 \end{aligned}$$

This gives us two points to consider:  $(-4, 2)$  and  $(6, 2)$ .

So, all together we have eight points that satisfy our three Lagrange equations. It's left only to decide which of those points lead to maxima and to minima.

point	$(1 + \sqrt{15}, 2 + \sqrt{10})$	$(1 - \sqrt{15}, 2 + \sqrt{10})$	$(1 + \sqrt{15}, 2 - \sqrt{10})$	$(1 - \sqrt{15}, 2 - \sqrt{10})$
value of $f$	$\ln 361$	$\ln 361$	$\ln 361$	$\ln 361$
	max	max	max	max

point	$(-4, 2)$	$(6, 2)$	$(1, 7)$	$(1, -3)$
value of $f$	$\ln 261$	$\ln 261$	$\ln 136$	$\ln 136$
			min	min

Our maximum value is  $\ln 361$ , and our minimum value is  $\ln 136$ .

Example 2.5.6

Example 2.5.7

Find the ends of the major and minor axes of the ellipse  $3x^2 - 2xy + 3y^2 = 4$ . They are the points on the ellipse that are farthest from and nearest to the origin.

*Solution.* Let  $(x, y)$  be a point on  $3x^2 - 2xy + 3y^2 = 4$ . This point is at the end of a major axis when it maximizes its distance from the centre of the ellipse,  $(0, 0)$ . It is at the end of a minor axis when it minimizes its distance from  $(0, 0)$ . So we wish to maximize and minimize the distance  $\sqrt{x^2 + y^2}$  subject to the constraint

$$g(x, y) = 3x^2 - 2xy + 3y^2 - 4 = 0$$

Now maximizing/minimizing  $\sqrt{x^2 + y^2}$  is equivalent<sup>24</sup> to maximizing/minimizing its square  $(\sqrt{x^2 + y^2})^2 = x^2 + y^2$ . So we are free to choose the objective function

$$f(x, y) = x^2 + y^2$$

which we will do, because it makes the derivatives cleaner. Again, we use Lagrange multipliers to solve this problem, so we start by finding the partial derivatives.

$$f_x(x, y) = 2x \quad f_y(x, y) = 2y \quad g_x(x, y) = 6x - 2y \quad g_y(x, y) = -2x + 6y$$

We need to find all solutions to

$$2x = \lambda(6x - 2y) \tag{E1}$$

$$2y = \lambda(-2x + 6y) \tag{E2}$$

$$3x^2 - 2xy + 3y^2 - 4 = 0 \tag{E3}$$

- If  $g_x \neq 0$  and  $g_y \neq 0$ , then  $\lambda = \frac{2x}{6x-2y} = \frac{x}{3x-y}$  by (E1), and  $\lambda = \frac{2y}{-2x+6y} = \frac{y}{-x+3y}$  by (E2).

$$\begin{aligned} \frac{x}{3x-y} &= \frac{y}{-x+3y} \\ -x^2 + 3xy &= 3xy - y^2 \\ x^2 &= y^2 \\ x &= \pm y \end{aligned}$$

So if  $x = \pm y$ , then the appropriate  $\lambda$  will make both (E1) and (E2) true. Now let's see

<sup>24</sup> The function  $S(z) = z^2$  is a strictly increasing function for  $z \geq 0$ . So, for  $a, b \geq 0$ , the statement " $a < b$ " is equivalent to the statement " $S(a) < S(b)$ ".

what makes (E3) true.

$$4 = 3x^2 - 2xy + 3y^2$$

$$4 = 3(\pm y)^2 - 2(\pm y)y + 3y^2$$

$$= 3y^2 \mp 2y^2 + 3y^2$$

$$= (6 \mp 2)y^2$$

$$4 = (6 + 2)x^2$$

$$\implies x = \pm \frac{1}{\sqrt{2}} \text{ when } x = -y$$

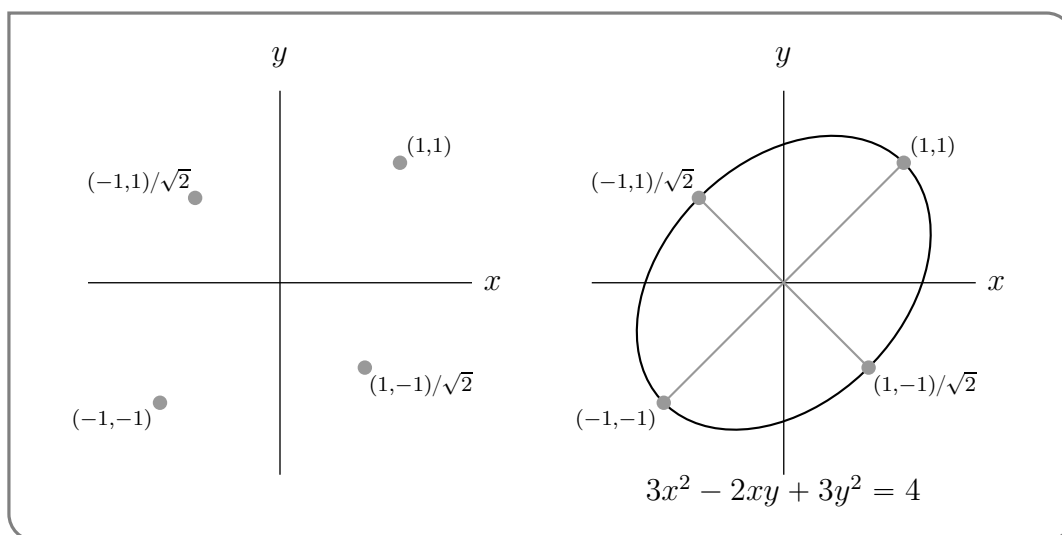
$$4 = (6 - 2)x^2$$

$$\implies x = \pm 1 \text{ when } x = y$$

This gives us four points to check: the two points  $\pm\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$  and the two points  $\pm(1, 1)$

- If  $g_x = 0$ , then  $6x - 2y = 0$ , i.e.  $y = 3x$ . By (E1),  $x = 0$ , so  $y = 0$ . Then (E3) doesn't hold, so this leads to no solutions.
- If  $g_y = 0$ , then  $-2x + 6y = 0$ , i.e.  $x = 3y$ . By (E2),  $y = 0$ , so  $x = 0$ . Then (E3) doesn't hold, so this leads to no solutions.

The distance from  $(0, 0)$  to  $\pm(1, 1)$ , namely  $\sqrt{2}$ , is larger than the distance from  $(0, 0)$  to  $\pm\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ , namely 1. So the ends of the minor axes are  $\pm\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$  and the ends of the major axes are  $\pm(1, 1)$ . Those ends are sketched in the figure on the left below. Once we have the ends, it is an easy matter<sup>25</sup> to sketch the ellipse as in the figure on the right below.



Example 2.5.7

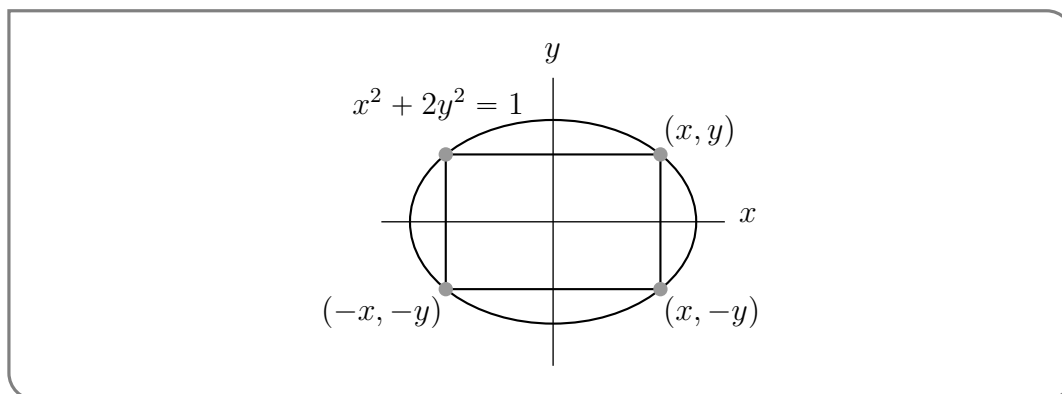
In the previous examples, the objective function and the constraint were specified explicitly. That will not always be the case. In the next example, we have to do a little geometry to extract them.

<sup>25</sup> if you tilt your head so that the line through  $(1, 1)$  and  $(-1, -1)$  appears horizontal

## Example 2.5.8

Find the rectangle of largest area (with sides parallel to the coordinates axes) that can be inscribed in the ellipse  $x^2 + 2y^2 = 1$ .

*Solution.* Since this question is so geometric, it is best to start by drawing a picture.



Call the coordinates of the upper right corner of the rectangle  $(x, y)$ , as in the figure above. Note that  $x \geq 0$  and  $y \geq 0$ ; and if  $x = 0$  or  $y = 0$ , then the area of the rectangle is 0, which is certainly not a maximum. So the global maximum must occur at some point where  $x$  and  $y$  are both positive. This will also be a local maximum, so we should be able to find it using the method of Lagrange multipliers.

The four corners of the rectangle are  $(\pm x, \pm y)$  so the rectangle has width  $2x$  and height  $2y$  and the objective function is  $f(x, y) = 4xy$ . The constraint function for this problem is  $g(x, y) = x^2 + 2y^2 - 1$ . Again, to use Lagrange multipliers we need the first order partial derivatives.

$$f_x = 4y \quad f_y = 4x \quad g_x = 2x \quad g_y = 4y$$

So, according to the method of Lagrange multipliers, we need to find all solutions to

$$4y = \lambda(2x) \tag{E1}$$

$$4x = \lambda(4y) \tag{E2}$$

$$x^2 + 2y^2 - 1 = 0 \tag{E3}$$

- If  $g_x \neq 0$  and  $g_y \neq 0$ , then  $\lambda = \frac{4y}{2x} = \frac{2y}{x}$  from (E1) and  $\lambda = \frac{4x}{4y} = \frac{x}{y}$  from (E2). So,

$$\begin{aligned} \frac{2y}{x} &= \frac{x}{y} \\ 2y^2 &= x^2 \\ x &= (\pm\sqrt{2})y \end{aligned}$$

From (E3),

$$\begin{aligned} ((\pm\sqrt{2})y)^2 + 2y^2 - 1 &= 0 \\ 2y^2 + 2y^2 &= 1 \\ 4y^2 &= 1 \\ y &= \pm\frac{1}{2} \\ x = (\pm\sqrt{2})y &= \pm\frac{1}{\sqrt{2}} \end{aligned}$$

So there are four points to consider:  $(\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{2})$ .

- If  $g_x = 0$ , i.e.  $2x = 0$ , then  $x = 0$ ; by (E1) also  $y = 0$ ; but then (E3) fails. So this doesn't give us any more points to consider.
- If  $g_y = 0$ , i.e.  $4y = 0$ , then  $y = 0$ ; by (E2) also  $x = 0$ ; but then (E3) fails. So this doesn't give us any more points to consider either.

We now have four possible values of  $(x, y)$ , namely  $(1/\sqrt{2}, 1/2)$ ,  $(-1/\sqrt{2}, -1/2)$ ,  $(1/\sqrt{2}, -1/2)$  and  $(-1/\sqrt{2}, 1/2)$ . They are the four corners of a single rectangle. We said that we wanted  $(x, y)$  to be the upper right corner, i.e. the corner in the first quadrant. It is  $(1/\sqrt{2}, 1/2)$ .

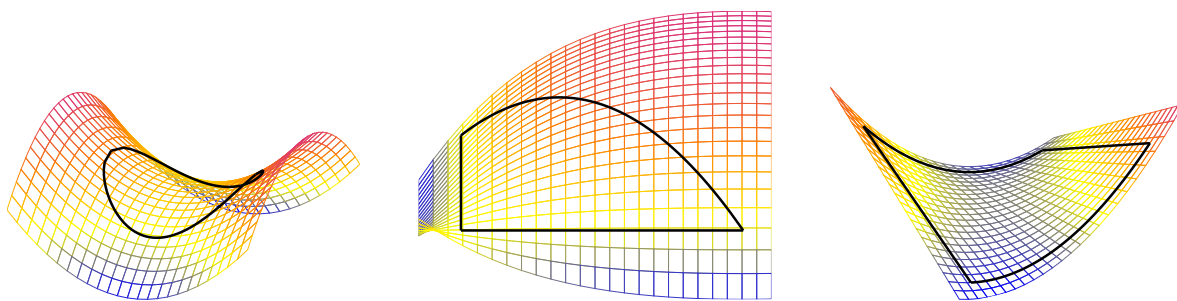
How do we interpret the other three points we found? The global min of the function  $4xy$  subject to the constraint  $x^2 + 2y^2 = 1$  will occur at one of these points, but those points aren't in our model domain. When  $x$  and  $y$  have different signs,  $4xy$  no longer gives the area of a rectangle, since it's negative. Over our model domain, we kind of have "endpoints:"  $x = 0$  and  $y = 0$ . Our maximum occurred somewhere between our endpoints; our model minimum occurs at the endpoints.

Example 2.5.8

## 2.5.1 ▶ Bounded vs Unbounded Constraints

In the last example, we had to think a little extra about whether the solution to the Lagrange equations gave a maximum or minimum. Take a closer look at Theorem 2.5.2: all *local* extrema will occur at a solution point. So when do the solution points definitely also include all *absolute* extrema?

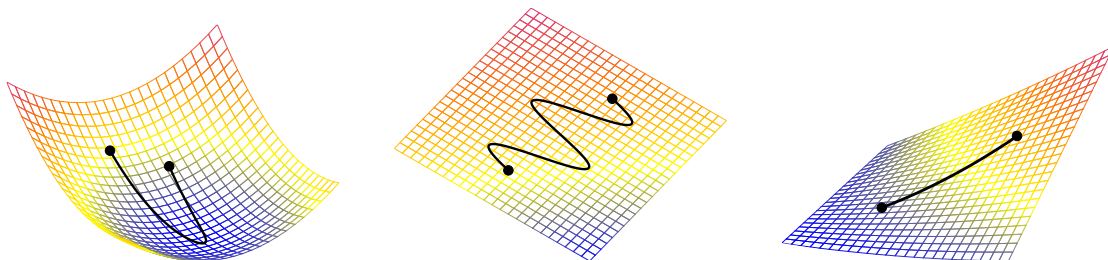
1. If our constraint function is a closed curve (circle, ellipse, square, etc.) and our objective function is continuous over it, then there will certainly be an absolute max and absolute min over the constraint; and these will certainly also be local extrema. So when our constraint is a closed curve, and our objective function is continuous over it, we are guaranteed that the absolute max and min exist, and are at points that satisfy the Lagrange equations.



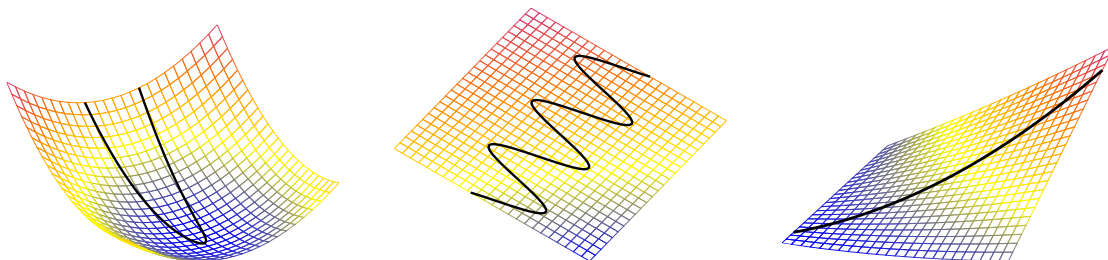
In Section 2.4 we considered domains that were bounded by a closed curve, so we only considered boundaries of this type.

2. If our constraint function is *not* a closed curve (e.g. a line, a line segment, a function like  $xy = 1$ , etc.) then the system is more complicated. Assume that the objective function is continuous over the constraint curve. Since our constraint curve is one-dimensional (like a line, but a line that has some orientation in space), we're in a similar position as we were in single-variable calculus: extrema can occur at endpoints, or at "critical points." In our case, "critical points" translate to solutions to the Lagrange equations; "endpoints" mean pretty much the same thing they always have.

- (a) If the constraint function is bounded, we must consider its endpoints as well as solutions to the Lagrange system. There will be an absolute maximum and minimum, and these will definitely occur at solutions to the Lagrange system *or* at the endpoints of the constraint.



- (b) If the constraint function is unbounded, there may or may not exist absolute extrema. This is where you'll most heavily rely on your understanding of function shape and behaviour. Limits can be useful here.



Example 2.5.9

Find the values of  $w \geq 0$  and  $\kappa \geq 0$  that maximize the utility function

$$U(w, \kappa) = 6w^{2/3}\kappa^{1/3} \quad \text{subject to the constraint} \quad 4w + 2\kappa = 12$$



*Solution.* The constraint  $4w + 2\kappa = 12$  is simple enough that we can easily use it to express  $\kappa$  in terms of  $w$ , then substitute  $\kappa = 6 - 2w$  into  $U(w, \kappa)$ , and then maximize  $U(w, 6 - 2w) = 6w^{2/3}(6 - 2w)^{1/3}$  using the techniques of last semester.

However, for practice purposes, we'll use Lagrange multipliers with the objective function  $U(w, \kappa) = 6w^{2/3}\kappa^{1/3}$  and the constraint function  $g(w, \kappa) = 4w + 2\kappa - 12$ . The first order derivatives of these functions are

$$U_w = 4w^{-1/3}\kappa^{1/3} \quad U_\kappa = 2w^{2/3}\kappa^{-2/3} \quad g_w = 4 \quad g_\kappa = 2$$

The boundary values ("endpoints")  $w = 0$  and  $\kappa = 0$  give utility 0, which is obviously not going to be the maximum utility. So it suffices to consider only local maxima. According to the method of Lagrange multipliers, we need to find all solutions to

$$4w^{-1/3}\kappa^{1/3} = 4\lambda \quad (\text{E1})$$

$$2w^{2/3}\kappa^{-2/3} = 2\lambda \quad (\text{E2})$$

$$4w + 2\kappa - 12 = 0 \quad (\text{E3})$$

Then we see  $g_x \neq 0$  and  $g_w \neq 0$ , so we only have one of our usual three cases.

- equation (E1) gives  $\lambda = w^{-1/3}\kappa^{1/3}$ .
- Substituting this into (E2) gives  $w^{2/3}\kappa^{-2/3} = \lambda = w^{-1/3}\kappa^{1/3}$  and hence  $w = \kappa$ .
- Then substituting  $w = \kappa$  into (E3) gives  $6\kappa = 12$ .

So  $w = \kappa = 2$  and the maximum utility is  $U(2, 2) = 12$ .

Note in this example we had a bounded (but not closed) curve. It has endpoints  $(0, 6)$  and  $(3, 0)$ . Since the maximum didn't occur at the endpoints, then the global maximum was also a local maximum, and so it showed up as a solution to the system of Lagrange equations.

Example 2.5.9

Chapter 2 was adapted from Chapter 2 of [CLP 3 – Multivariable Calculus](#) by Feldman, Reznitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## 2.6▲ Utility and Demand Functions

Economists use the concept of *utility* to define the welfare of an entity or an individual. The utility function measures the level of satisfaction or happiness that a consumer gains from various actions, like consumption, leisure, etc. One such form of utility function, discussed below, represents a constrained optimization problem: where consumers, given their preferences for two goods, maximize welfare or level of happiness from consuming particular combinations of goods or services, given finite resources or income.

The amount consumed should be a non-negative number, so we'll restrict our domains accordingly.

Let  $x$  be a variable representing a quantity of good  $X$ , and let  $u(x)$  be the utility function for that good. If “more is better,” then we expect  $\frac{du}{dx} > 0$  for all nonnegative  $x$  (i.e. for all  $x$  in the model domain). We can think of  $\frac{du}{dx}$  as marginal utility: the gain in happiness of getting just a little more of something.

Suppose good  $X$  is subject to “diminishing returns.” That is, as we get more of the good (i.e. as  $x$  increases) each additional unit brings us less happiness than the last. Then our marginal utility  $\frac{du}{dx}$  is decreasing, meaning  $\frac{d^2u}{dx^2} < 0$ . Since most goods are subject to diminishing returns, we often choose utility functions that are concave down.

Utility functions can encompass more than one good. A multivariable utility function  $u(x, y)$  might give the happiness associated with consuming quantity  $x$  of good  $X$  alongside quantity  $y$  of good  $Y$ . Just like with single-variable utility functions, if “more is better” than  $\frac{\partial u}{\partial x} > 0$  and  $\frac{\partial u}{\partial y} > 0$  everywhere. If there are diminishing returns, then also  $\frac{\partial^2 u}{\partial x^2} < 0$  and  $\frac{\partial^2 u}{\partial y^2} < 0$  everywhere.

### 2.6.1 ► Constrained Optimization of the Utility Function

#### Example 2.6.1

Suppose a consumer’s preferences lead to the utility function

$$u(x, y) = x^2 + 2y$$

for consuming a combination of  $x$  units of good  $X$  and  $y$  units of good  $Y$ .

For these goods, “more is better” (because  $u_x > 0$  and  $u_y > 0$  for all non-negative  $x$  and  $y$ ) without diminishing returns.<sup>26</sup>

Good  $X$  costs 2 dollars per unit, good  $Y$  costs 3 dollars per unit, and the consumer has 10 dollars to spend on these two goods.

Find the combined consumption of goods  $X$  and  $Y$  that maximizes the utility function, subject to the constraint that the consumer spends at most 10 dollars. What is that maximum utility?

*Solution.* This is a constrained optimization problem. Our objective function (what we want to maximize) is

$$u(x, y) = x^2 + 2y$$

Our constraint function comes from our budget and the prices of the two goods:

$$g(x, y) = 2x + 3y - 10 = 0$$

(Since “more is better,” there’s no incentive to spend *less* than our budget of ten dollars.)

We can solve this by substitution. From our constraint, we see  $y = \frac{10-2x}{3}$ . That turns our utility function into the following:

$$u(x, y) = x^2 + 2y = x^2 + \frac{2}{3}(10 - 2x) = x^2 - \frac{4}{3}x + \frac{20}{3}$$

26 Indeed,  $\frac{\partial^2 u}{\partial x^2} > 0$ , meaning each subsequent unit of the good associated with  $x$  brings more happiness than the last – a hallmark that this equation is fabricated for practice purposes, and probably isn’t a realistic utility function of actual goods. The idea of increasing (rather than diminishing) returns can make for an interesting thought experiment, though.

This is a parabola pointing up, so its maximum will be at an endpoint of our interval. Since  $x$  and  $y$  are quantities, we require  $x \geq 0$  and  $y \geq 0$ .

$$0 \leq y = \frac{10 - 2x}{3} \implies x \leq 5$$

Our model domain is  $0 \leq x \leq 5$ . The endpoint  $x = 5$  corresponds to all \$10 going to the first good (and  $y = 0$ ). The endpoint  $x = 0$  corresponds to all \$10 going to the second good (with  $y = \frac{10}{3}$ ).

$$\begin{aligned} u(5, 0) &= 5^2 + 2(0) = 25 \\ u\left(0, \frac{10}{3}\right) &= 0^2 + 2\left(\frac{10}{3}\right) = \frac{20}{3} \end{aligned}$$

Our utility is maximized when we spend all \$10 on the first good, purchasing  $x = 5$  and  $y = 0$ . That maximum utility is 25.

Example 2.6.1

Example 2.6.2

Alejandro has recently found a true passion for baking. He likes making two types of bread: ciabatta ( $c$ ) and pita ( $p$ ). Ciabatta costs 20 dollars per unit to make and pita 10 dollars per unit. Alejandro wants to spend 60 dollars on bread, and his utility function<sup>27</sup> is as follows:

$$u(c, p) = \ln(c) + 2\ln(p)$$

Find the optimal consumption for Alejandro and the corresponding maximum utility.

*Solution.* The utility function will be the objective function and the constraint will be the budget constraint. The budget constraint is  $20c + 10p = 60$ . We can find the maximum utility using substitution or the method of Lagrange multipliers.

**Solution 1: substitution**

Since  $20c + 10p = 60$ , we see  $p = 6 - 2c$ . Then our utility function is:

$$u(c, p) = \ln(c) + 2\ln(p) = \ln(c) + 2\ln(6 - 2c)$$

Using log rules,

$$\begin{aligned} u(c, p) &= \ln(c) + \ln\left((6 - 2c)^2\right) \\ &= \ln\left(c(6 - 2c)^2\right) \end{aligned}$$

<sup>27</sup> We're not averse to having negative utility values. Again, utility doesn't have absolute units, but rather is useful as a relative scale. Higher utility is better, whether the numbers are positive or not. For this particular utility function,  $c = 0$  or  $p = 0$  will *minimize* utility. This is actually a common property of utility functions. It avoids having an optimal solution where one good is not consumed at all.

Much like the square root function, natural logarithm is an increasing function. So, the maximum of  $\ln(c(6-2c)^2)$  will occur at the same place as the maximum of  $c(6-2c)^2$ , provided that maximum is positive (and thus in the domain of the logarithmic function).

$$f(c) = c(6-2c)^2$$

Using the product rule,

$$\begin{aligned} f'(c) &= c \cdot 2(6-2c)(-2) + (6-2c)^2 \\ &= (6-2c)[-4c + (6-2c)] \\ &= 12(3-c)(1-c) \end{aligned}$$

The critical points of  $f(c)$  are  $c = 1$  and  $c = 3$ .

$$f(1) = 16$$

$$f(3) = 0$$

We also need to check the endpoints of our interval. Since  $p \geq 0$ , then:

$$0 \leq p = 6 - 2c \implies c \leq 3$$

The endpoints of our interval are  $c = 0$  (all pita) and  $c = 3$  (all ciabatta). We've already found  $f(3) = 0$ .

$$f(0) = 0$$

The function  $c(6-2c)^2$  has a maximum of 16 when  $c = 1$ , so the function  $\ln(c(6-2c)^2)$  has a maximum of  $\ln 16$  when  $c = 1$ . Since  $c = 1$  means  $p = 4$ , utility is maximized when Alejandro spends \$20 on ciabatta, and \$40 on pita.

### Solution 2: Lagrange

$$\begin{cases} u_c &= \lambda g_c & \implies & \begin{cases} \frac{1}{c} &= \lambda \cdot 20 \\ \frac{2}{p} &= \lambda \cdot 10 \\ 20c + 10p - 60 &= 0 \end{cases} \end{cases}$$

From the first two equations, we see

$$\lambda = \frac{1}{20c} = \frac{2}{10p}$$

$$p = 4c$$

From the constraint equation,

$$0 = 20c + 10p - 60 = 10c + 10(4c) - 60$$

$$c = 1 \quad p = 4$$

So, the point  $c = 1$  is a point to check. We should also check the endpoints of our interval,  $c = 0$  and  $c = 3$ . Note both these cause the utility to go to negative infinity – so they are minima. That tells us  $c = 1$ ,  $p = 4$  gives us our constrained maxima. The utility of spending \$20 on ciabatta and \$40 on pita is  $\ln 1 + 2 \ln 4 = \ln(16)$ .

Example 2.6.2

## 2.6.2 ▶ Demand Curves

A demand curve gives the relationship between the *quantity* of a good a consumer would buy and the *price* of that good. We assume the consumer would buy the quantity that maximizes their utility function, given their budget constraints. In Examples 2.6.1 and 2.6.2 we found “optimal consumption” when the price and budget were fixed numbers. So secretly, we were finding a point on a demand curve.

Instead of keeping price and budget fixed, we can assign them variables. We can still find the amount a consumer would buy to maximize their utility, but now that amount will be a function of price and budget, rather than a fixed number. The consumer’s optimal consumption (as a function of price and budget) gives the general demand function. This is sometimes formally referred to as *Marshallian demand*. That is, Marshallian demand describes the relationship between the price of a good, the budget for that good, and the quantity of that good demanded.

### Definition 2.6.3 (Marshallian demand).

Let  $x$  be a quantity of good  $X$ , let  $y$  be a quantity of good  $Y$ , and let  $u(x, y)$  be the utility function of these two goods.

Let  $p_x$  be the unit price for good  $X$ , and  $p_y$  be the unit price for good  $Y$ . Let  $I$  be the amount of a consumer’s income they budget for buying  $X$  and  $Y$ . Then the function

$$x^m(p_x, p_y, I)$$

giving the optimal consumption of  $x$  to maximize  $u(x, y)$  subject to the budget constraint  $p_x x + p_y y = I$  is called the *Marshallian demand function*.

Note: the superscript  $m$  in the function name  $x^m$  isn’t a power. Rather than denoting a variable,  $m$  simply stands for “Marshallian.”

### Example 2.6.4

Let’s go back to Alejandro and his passion for baking. This weekend he would like to make ciabatta ( $c$ ) and focaccia ( $f$ ). Ciabatta costs  $p_c$  dollars to make and focaccia  $p_f$  dollars. For this weekend, Alejandro wants to spend  $I$  dollars on bread, and his utility function is as follows:

$$u(c, f) = \ln(c) + 2 \ln(f)$$

Find the optimal consumption for Alejandro of each bread type.

*Solution.* The utility function is the objective function, because that’s the function we want to maximize. The constraint is  $p_c c + p_f f = I \Rightarrow b(c, f) = c p_c + f p_f - I$ .

As in Example 2.6.2, the endpoints of our interval ( $c = 0, f = 0$ ) minimize utility, so the maximum will be at some interior point. We can find it using the method of Lagrange multipliers.

$$\begin{cases} u_c &= \lambda \cdot b_c \\ u_f &= \lambda \cdot b_f \\ b(c, f) &= 0 \end{cases} \implies \begin{cases} \frac{1}{c} &= \lambda \cdot p_c \\ \frac{2}{f} &= \lambda \cdot p_f \\ cp_c + fp_f - I &= 0 \end{cases}$$

From the first two equations, we see

$$\lambda = \frac{1}{c \cdot p_c} = \frac{2}{f \cdot p_f}$$

$$f = 2c \left( \frac{p_c}{p_f} \right)$$

From the budget constraint,

$$\begin{aligned} 0 &= cp_c + fp_f - I \\ &= cp_c + 2c \left( \frac{p_c}{p_f} \right) p_f - I \\ &= cp_c + 2cp_c - I \\ I &= 3cp_c \\ c &= \frac{I}{3p_c} \end{aligned}$$

Using the relationship between  $f$  and  $c$ ,

$$f = 2c \left( \frac{p_c}{p_f} \right) = \frac{2I}{3p_f}$$

The point  $c = \frac{I}{3p_c}$ ,  $f = \frac{2I}{3p_f}$  is the only point to consider for a max. Since  $c \geq 0$  and  $f \geq 0$ , the point is within our model domain. So, it gives the optimal consumption of ciabatta and focaccia.

Let's think of the optimal consumption of each bread type (as functions of prices and income allocated to bread), and name these functions  $c^m$  and  $f^m$  ( $m$  for "Marshallian"). Then

$$c^m(I, p_c, p_f) = \frac{I}{3p_c} \quad f^m(I, p_c, p_f) = \frac{2I}{3p_f}$$

give the Marshallian demand curves for ciabatta and focaccia, respectively.

Example 2.6.4

We use the Marshallian demand to define certain *types* of goods. A *normal good* is defined as a product for which quantity demanded increases as income increases. An *inferior good* is defined as a product for which quantity demanded decreases as income increases.

**Definition 2.6.5** (Normal and Inferior Goods).

Let  $x^m(p_x, p_y, I)$  be the Marshallian demand function of a good when the price of that good is  $p_x$ , the price of another good is  $p_y$ , and the amount of a consumer's income budgeted for these goods is  $I$ . If

$$\frac{\partial x^m(p_x, p_y, I)}{\partial I} > 0$$

everywhere then the good is a *normal good*. If

$$\frac{\partial x^m(p_x, p_y, I)}{\partial I} < 0$$

everywhere then the good is an *inferior good*.

In the case of Alejandro's example 2.6.4, you can verify that both goods are normal goods.

**Theorem 2.6.6** (Normal and Inferior Goods).

Let  $X$  and  $Y$  be two goods with positive unit prices  $p_x$  and  $p_y$ , respectively, subject to the budget constraint  $p_x x + p_y y = I$ . If  $X$  is an inferior good, then  $Y$  is a normal good.

*Proof.* Let  $x^m$  and  $y^m$  be the Marshallian demand functions of  $X$  and  $Y$ , respectively.

For any quantities  $x$  and  $y$ , we must satisfy the budget constraint  $p_x x + p_y y = I$ . So:

$$y = \frac{I - p_x x}{p_y}$$

When  $x = x^m$ , then  $y = y^m$ , so:

$$\begin{aligned} y^m &= \frac{I - p_x x^m}{p_y} \\ \Rightarrow \frac{\partial y^m}{\partial I} &= \frac{\partial}{\partial I} \left[ \frac{I - p_x x^m}{p_y} \right] = \frac{\partial}{\partial I} \left[ \frac{1}{p_y} I - \frac{p_x}{p_y} x^m \right] \\ &= \frac{1}{p_y} - \frac{p_x}{p_y} \cdot \frac{\partial x^m}{\partial I} \end{aligned}$$

Since  $X$  is an inferior good, by Definition 2.6.5,  $\frac{\partial x^m}{\partial I} < 0$ . Since also  $p_x$  and  $p_y$  are positive, then the term  $-\frac{p_x}{p_y} \frac{\partial x^m}{\partial I}$  is positive:

$$\Rightarrow \frac{\partial y^m}{\partial I} \geq \frac{1}{p_y} > 0$$

So,  $Y$  is a normal good by Definition 2.6.5. □

Using the first partial derivative, we can also analyse how changes in prices affect the Marshallian Demand.

**Definition 2.6.7 (Price Effect).**

Let  $x^m(p_x, p_y, I)$  be a Marshallian demand function for a good  $X$  whose quantity is given by  $x$  and unit price is given by  $p_x$ , in relation to another good  $Y$  whose quantity is given by  $y$  and whose unit price is given by  $p_y$ .

$$\frac{\partial x^m(p_x, p_y, I)}{\partial p_x}$$

is the rate of change of  $x^m$  (the optimal consumption of  $X$ ) relative to the price of  $X$ . We call this the *price effect* of  $X$  on  $x^m$ . Similarly,

$$\frac{\partial x^m(p_x, p_y, I)}{\partial p_y}$$

the rate of change of  $x^m$  (the optimal consumption of  $X$ ) relative to the price of  $Y$ . This is the *price effect* of  $Y$  on  $x^m$ .

**Example 2.6.8 (Price effects)**

In this example, we revisit Example 2.6.4.

- Are the ciabatta and focaccia from Example 2.6.4 normal goods, or inferior goods?
- If the price of making focaccia increases, how will this effect the amount of *ciabatta* Alejandro makes? (Assume everything else stays the same – the utility function stays the same, the price of ciabatta stays the same, the portion of income  $I$  allotted to bread stays the same, and the assumption remains that Alejandro will maximize utility subject to his budget constraint.)

*Solution.*

To decide whether ciabatta and focaccia are normal or inferior goods, we should take their partial derivatives with respect to  $I$ .

$$c^m(I, p_c, p_f) = \frac{I}{3p_c} \qquad f^m(I, p_c, p_f) = \frac{2I}{3p_f}$$

$$\frac{\partial c^m}{\partial I} = \frac{1}{3p_c} > 0 \qquad \frac{\partial f^m}{\partial I} = \frac{2}{3p_f} > 0$$

Since both derivatives are positive everywhere, ciabatta and focaccia are both normal goods.

Surprisingly, the price of focaccia doesn't affect the consumption of ciabatta at all! The Marshallian demand of ciabatta is

$$c^m(I, p_c, p_f) = \frac{I}{3p_c}$$



Since  $p_f$  doesn't even show up, the derivative is easy to take:

$$\frac{\partial c^m}{\partial p_f} = 0$$

That means the price effect of focaccia on Alejandro's ciabatta baking is zero. (If the price of focaccia goes up, the impact on his baking habits are that he will make less focaccia.)

Example 2.6.8

Example 2.6.9

Kenechukwu is doing groceries for the week, and as usual he has  $I$  dollars to spend on fruits and berries. If he consumes  $a$  kg of apples and  $s$  kg of strawberries, then his utility function is:

$$u(a, s) = a^{1/2}s^{1/4}$$

Apples cost  $p_a$  dollars per kg, and strawberries cost  $p_s$  dollars per kg.

Find Kenechukwu's Marshallian demand function for apples. What is the price effect of  $p_s$  on apples? Are apples normal or inferior goods?

*Solution.* The utility function will be the objective function, because utility is what we want to maximize. As in Example 2.6.2, the endpoints  $a = 0$  and  $s = 0$  are minima of the utility function. (We see this because setting either  $a = 0$  or  $s = 0$  leads to  $u = 0$ ; and since  $u$  involves even roots, it never returns a negative value.) So, the maximum will happen at some internal point, which we can find using Lagrange multipliers.

The budget constraint is  $p_a a + p_s s = I \Rightarrow b(a, s) = p_a a + p_s s - I$ .

$$\begin{cases} u_a &= \lambda \cdot b_a \\ u_b &= \lambda \cdot b_b \\ b(a, s) &= 0 \end{cases} \implies \begin{cases} \frac{1}{2}a^{-1/2}s^{1/4} &= \lambda \cdot p_a \\ \frac{1}{4}a^{1/2}s^{-3/4} &= \lambda \cdot p_s \\ p_a a + p_s s - I &= 0 \end{cases}$$

From the first two equations, we see

$$\begin{aligned} \lambda &= \frac{1}{2p_a}a^{-1/2}s^{1/4} = \frac{1}{4p_s}a^{1/2}s^{-3/4} \\ 2p_s a^{-1/2}s^{1/4} &= p_a a^{1/2}s^{-3/4} \\ 2p_s s^{1/4+3/4} &= p_a a^{1/2+1/2} \\ 2p_s s &= p_a a \\ \frac{2p_s}{p_a} s &= a \end{aligned}$$

Now, to satisfy the budget constraint,

$$\begin{aligned} p_a \left( \frac{2p_s}{p_a} s \right) + p_s s &= I \\ 3p_s s &= I \\ s &= \frac{I}{3p_s}, \quad a = \frac{2p_s}{p_a} s = \frac{2p_s}{p_a} \cdot \frac{I}{3p_s} = \frac{2I}{3p_a} \end{aligned}$$

So, our Marshallian demand functions are

$$a^m(p_a, p_s, I) = \frac{2I}{3p_a} \quad s^m(p_a, p_s, I) = \frac{I}{3p_s}$$

The price effect of  $p_s$  on apples is nothing, since

$$\frac{\partial a^m}{\partial p_s} = 0$$

The goods are both normal, because

$$\frac{\partial a^m}{\partial I} = \frac{2}{3p_a} > 0 \quad \frac{\partial s^m}{\partial I} = \frac{1}{3p_s} > 0$$

The optimal consumption is 1 kg of strawberries and 4 kg of apples. This leads to the maximum utility,  $u(4, 1) = 2$ .

Example 2.6.9

So far, our paradigm has been to optimize happiness, given a fixed budget. We could instead fix the desired amount of utility, and try to minimize the cost required to achieve it. In this paradigm, our utility function is our constraint, while our cost function is the objective function we want to minimize. This gives rise to the Hicksian demand.

**Definition 2.6.10** (Hicksian demand).

Let goods  $X$  and  $Y$  have utility function  $u(x, y)$ , where  $x$  is the quantity of  $X$  and  $y$  is the quantity of  $Y$ . Let  $p_x$  be the unit price of good  $X$ , and  $p_y$  be the unit price of good  $Y$ . Let  $U$  be the minimum level of utility required by the consumer – that is, the consumer requires  $u(x, y) \geq U$ .

The *Hicksian demand function* of good  $X$ , denoted

$$x^h(p_x, p_y, U),$$

gives the value of  $x$  that minimizes the cost function  $f(x, y) = p_x x + p_y y$  subject to the constraint  $u(x, y) \geq U$ . That is, the Hicksian demand function gives the quantity of good  $X$  that minimizes the amount of money spent on the two goods while still achieving some fixed level of utility.

Note: the superscript  $h$  in the function name  $x^h$  isn't a power. Rather than denoting a variable,  $h$  simply stands for "Hicksian."

The definition requires that the utility be *at least* some fixed constant. In practice, we can usually assume that the utility is *equal to* that fixed constant. That's because if we have a higher utility than necessary, we can usually save some money by bringing our utility down to its minimum allowable level. This could fail only if, at some point, our utility function had a negative partial derivative. A negative partial derivative indicates that we might increase utility as we decrease consumption.

Example 2.6.11

Lets go back to Alejandro and his passion for baking. This weekend he would like to make ciabatta (c) and baguettes (B). Ciabatta costs  $p_c$  dollars to make and baguettes  $p_b$  dollars. His utility function is as follows:

$$u(c, b) = \sqrt{cb}$$

Fixing Alejandro's utility as the constant  $u(c, b) = U$ , find his Hicksian demand for both types of bread.

*Solution.* In Hicksian demand, we minimize cost, so cost is our objective function. That is,  $f(c, b) = p_c c + p_b b$ . Our constraint is  $U = \sqrt{cb}$ . We can find the constrained minimum of  $f(c, b)$  using substitution.

$$U = \sqrt{cb}$$

$$U^2 = cb$$

$$c = \frac{U^2}{b}$$

Plugging this into our objective function,

$$f(c, b) = p_c c + p_b b = p_c \left( \frac{U^2}{b} \right) + p_b b = (U^2 p_c) b^{-1} + p_b b$$

This is a function of one variable. Let's find the critical points.

$$0 = - (U^2 p_c) b^{-2} + p_b$$

$$(U^2 p_c) b^{-2} = p_b$$

$$\frac{U^2 p_c}{p_b} = b^2$$

$$b = U \sqrt{\frac{p_c}{p_b}}$$

$$\text{At that point, } c = \frac{U^2}{b} = U \sqrt{\frac{p_b}{p_c}}$$

To verify that this critical point gives a global minimum, consider the second derivative of our one-variable function.

$$\frac{d}{db} \left[ - (U^2 p_c) b^{-2} + p_b \right] = 2 (U^2 p_c) b^{-3}$$

Our model domain only allows for non-negative values of  $b$ , so the second derivative is non-negative everywhere. That means its global minimum is at its sole critical point. In particular, the quantities  $c = U \sqrt{\frac{p_b}{p_c}}$  and  $b = U \sqrt{\frac{p_c}{p_b}}$  minimize the cost function  $f(c, b) = p_c c + p_b b$  subject to the constraint  $u(c, b) = U$ . So, our Hicksian demand functions are:

$$c^h(p_c, p_b, U) = U \sqrt{\frac{p_b}{p_c}} \quad \text{and} \quad b^h(p_c, p_b, U) = U \sqrt{\frac{p_c}{p_b}}$$

## Example 2.6.11

In Example 2.6.11, note  $\frac{\partial c^h}{\partial p_b} \neq 0$ . This is in contrast to examples 2.6.8 and 2.6.9, where the price effects of one good's price on the other good's consumption were both 0. Hicksian demand is sometimes used to study the *substitution effect*, where a change in price in one good causes a change in consumption of another good. This discussion is, however, beyond the scope of the current text.

## Example 2.6.12 (Contrasting Marshallian and Hicksian Demand)

In Marshallian demand, the consumer has a fixed budget, and tries to be as happy as possible. In Hicksian demand, the consumer has a fixed utility need, and tries to be as thrifty as possible. These different models fit different types of transactions.

You have \$1000 to invest in stocks. You choose the combination of stocks that you think will have the best mix of risk and reward, spending your entire budget. You're operating under a generally Marshallian mindset, because you want the most utility for a fixed budget.

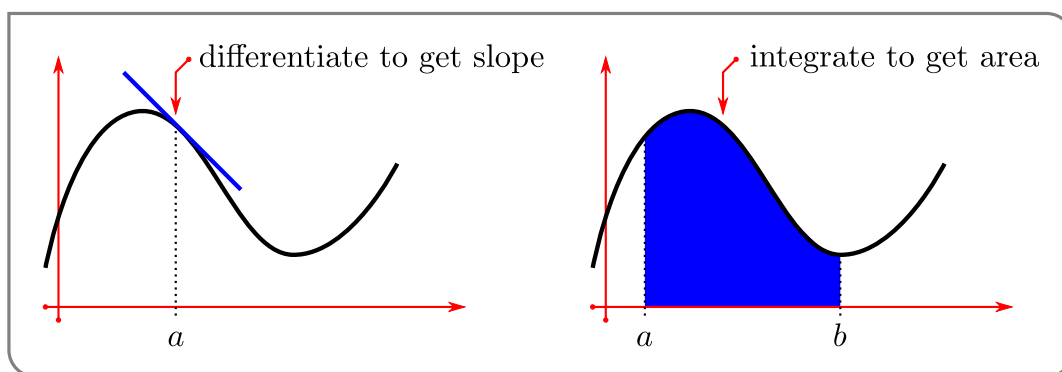
After you purchase your stocks, you're so excited that you accidentally spill ramen on your laptop, and need a new one right away. Your laptop needs to be of a sufficient quality for your needs – some combination of available soon, reliable, fast enough, and so on. You go to your local gadget store and identify all the models that will meet your needs. You buy the cheapest of those options. For the laptop, your demand is more Hicksian – you want to find the cheapest option that still meets your needs.

## Example 2.6.12

# INTEGRATION

Calculus is built on two operations — differentiation and integration.

- Differentiation — as we saw last term, differentiation allows us to compute and study the instantaneous rate of change of quantities. At its most basic it allows us to compute tangent lines and velocities, but it also led us to quite sophisticated applications including approximation of functions through Taylor polynomials and optimisation of quantities by studying critical points.
- Integration — at its most basic, allows us to analyse the area under a curve. Of course, its application and importance extend far beyond areas and it plays a central role in solving differential equations.



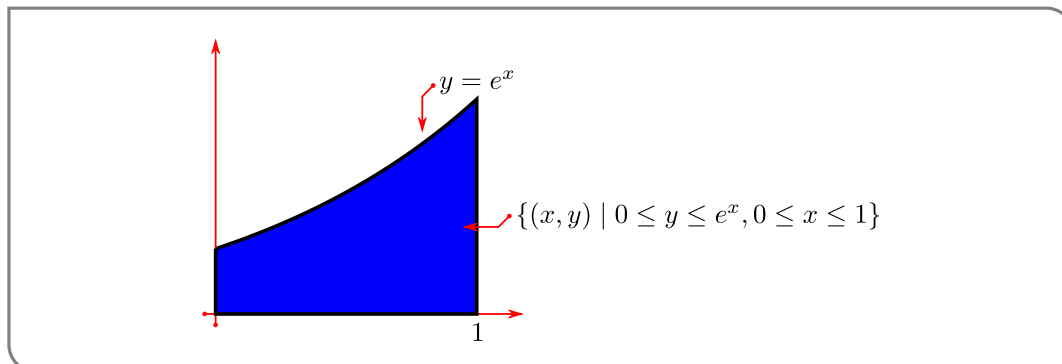
It is not immediately obvious that these two topics are related to each other. However, as we shall see, they are indeed intimately linked.

## 3.1▲ Definition of the Integral

Arguably the easiest way to introduce integration is by considering the area between the graph of a given function and the  $x$ -axis, between two specific vertical lines — such as is shown in the figure above. We'll follow this route by starting with a motivating example.

### ► A Motivating Example

Let us find the area under the curve  $y = e^x$  (and above the  $x$ -axis) for  $0 \leq x \leq 1$ . That is, the area of  $\{(x, y) \mid 0 \leq y \leq e^x, 0 \leq x \leq 1\}$ .



This area is equal to the “definite integral”

$$\text{Area} = \int_0^1 e^x dx$$

Do not worry about this notation or terminology just yet. We discuss it at length below. In different applications this quantity will have different interpretations — not just area. For example, if  $x$  is time and  $e^x$  is your velocity at time  $x$ , then we’ll see later (in Example 3.1.12) that the specified area is the net distance travelled between time 0 and time 1. After we finish with the example, we’ll mimic it to give a general definition of the integral  $\int_a^b f(x)dx$ .

#### Example 3.1.1

We wish to compute the area of  $\{(x, y) \mid 0 \leq y \leq e^x, 0 \leq x \leq 1\}$ . We know, from our experience with  $e^x$  in differential calculus, that the curve  $y = e^x$  is not easily written in terms of other simpler functions, so it is very unlikely that we would be able to write the area as a combination of simpler geometric objects such as triangles, rectangles or circles.

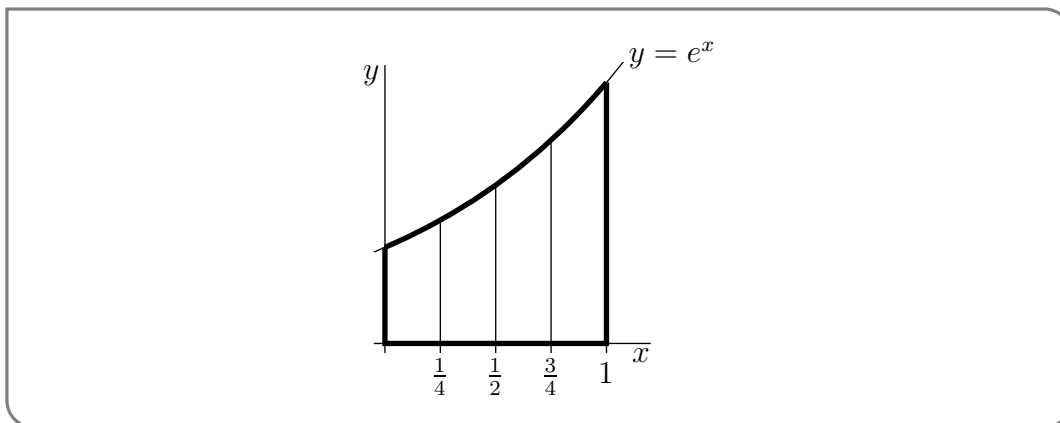
So rather than trying to write down the area exactly, our strategy is to approximate the area and then make our approximation more and more precise<sup>1</sup>. We choose<sup>2</sup> to approximate the area as a union of a large number of tall thin (vertical) rectangles. As we take more and more rectangles we get better and better approximations. Taking the limit as the number of rectangles goes to infinity gives the exact area<sup>3</sup>.

As a warm up exercise, we’ll now just use four rectangles. In Example 3.1.2, below, we’ll consider an arbitrary number of rectangles and then take the limit as the number of rectangles goes to infinity. So

- 1 This should remind the reader of the approach taken to compute the slope of a tangent line way way back at the start of differential calculus.
- 2 Approximating the area in this way leads to a definition of integration that is called Riemann integration. This is the most commonly used approach to integration. However we could also approximate the area by using long thin horizontal strips. This leads to a definition of integration that is called Lebesgue integration. We will not be covering Lebesgue integration in these notes.
- 3 If we want to be more careful here, we should construct two approximations, one that is always a little smaller than the desired area and one that is a little larger. We can then take a limit using the Squeeze Theorem and arrive at the exact area. More on this later.

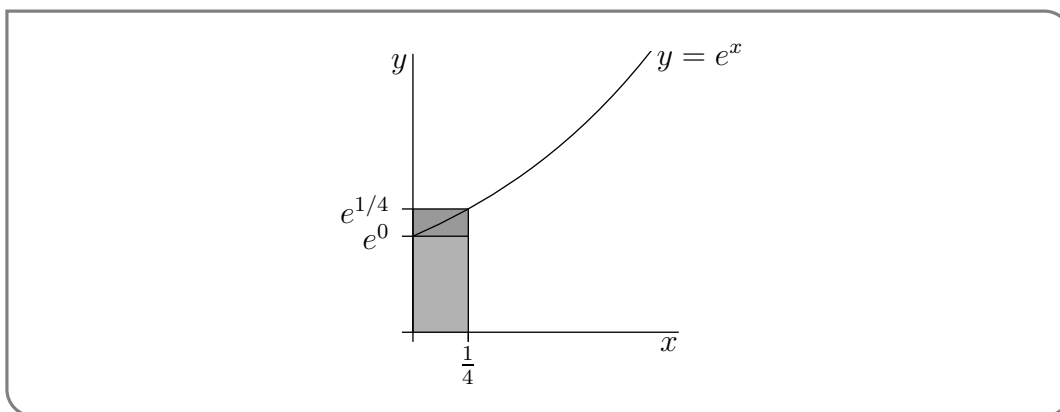
- subdivide the interval  $0 \leq x \leq 1$  into 4 equal subintervals each of width  $1/4$ , and
- subdivide the area of interest into four corresponding vertical strips, as in the figure below.

The area we want is exactly the sum of the areas of all four strips.



Each of these strips is almost, but not quite, a rectangle. While the bottom and sides are fine (the sides are at right-angles to the base), the top of the strip is not horizontal. This is where we must start to approximate. We can replace each strip by a rectangle by just levelling off the top. But now we have to make a choice — at what height do we level off the top?

Consider, for example, the leftmost strip. On this strip,  $x$  runs from 0 to  $1/4$ . As  $x$  runs from 0 to  $1/4$ , the height  $y$  runs from  $e^0$  to  $e^{1/4}$ . It would be reasonable to choose the height of the approximating rectangle to be somewhere between  $e^0$  and  $e^{1/4}$ . Which height



should we choose? Well, actually it doesn't matter. When we eventually take the limit of infinitely many approximating rectangles all of those different choices give exactly the same final answer. We'll say more about this later.

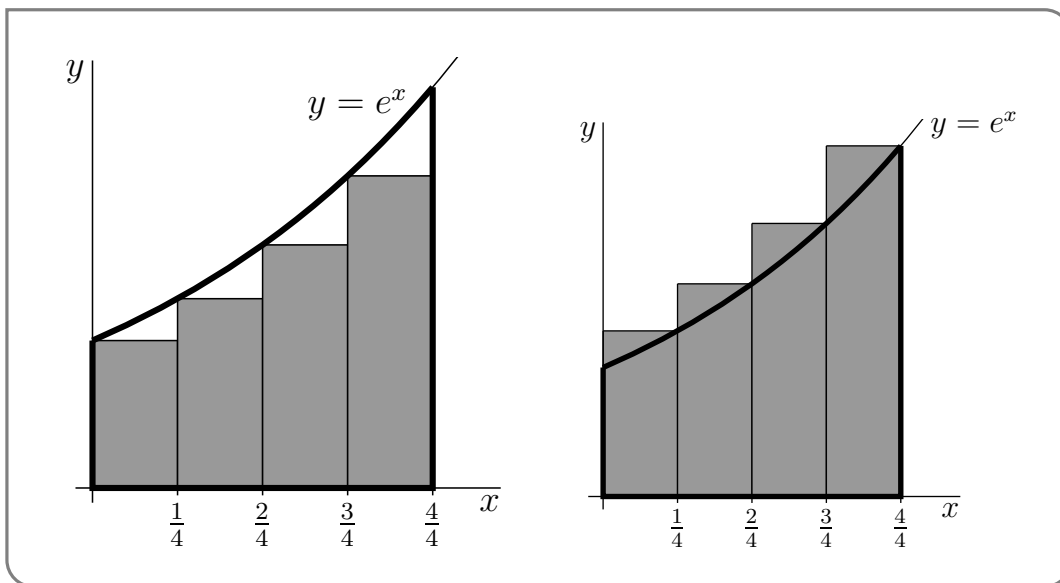
In this example we'll do two sample computations.

- For the first computation we approximate each slice by a rectangle whose height is the height of the *left* hand side of the slice.
  - On the first slice,  $x$  runs from 0 to  $1/4$ , and the height  $y$  runs from  $e^0$ , on the left hand side, to  $e^{1/4}$ , on the right hand side.

- So we approximate the first slice by the rectangle of height  $e^0$  and width  $1/4$ , and hence of area  $\frac{1}{4}e^0 = \frac{1}{4}$ .
- On the second slice,  $x$  runs from  $1/4$  to  $1/2$ , and the height  $y$  runs from  $e^{1/4}$  and  $e^{1/2}$ .
- So we approximate the second slice by the rectangle of height  $e^{1/4}$  and width  $1/4$ , and hence of area  $\frac{1}{4}e^{1/4}$ .
- And so on.
- All together, we approximate the area of interest by the sum of the areas of the four approximating rectangles, which is

$$\left[1 + e^{1/4} + e^{1/2} + e^{3/4}\right] \frac{1}{4} = 1.5124$$

- This particular approximation represents the shaded area in the figure on the left below. Note that, because  $e^x$  increases as  $x$  increases, this approximation is definitely smaller than the true area.



- For the second computation we approximate each slice by a rectangle whose height is the height of the *right* hand side of the slice.
  - On the first slice,  $x$  runs from  $0$  to  $1/4$ , and the height  $y$  runs from  $e^0$ , on the left hand side, to  $e^{1/4}$ , on the right hand side.
  - So we approximate the first slice by the rectangle of height  $e^{1/4}$  and width  $1/4$ , and hence of area  $\frac{1}{4}e^{1/4}$ .
  - On the second slice,  $x$  runs from  $1/4$  to  $1/2$ , and the height  $y$  runs from  $e^{1/4}$  and  $e^{1/2}$ .
  - So we approximate the second slice by the rectangle of height  $e^{1/2}$  and width  $1/4$ , and hence of area  $\frac{1}{4}e^{1/2}$ .
  - And so on.



- All together, we approximate the area of interest by the sum of the areas of the four approximating rectangles, which is

$$\left[ e^{1/4} + e^{1/2} + e^{3/4} + e^1 \right] \frac{1}{4} = 1.9420$$

- This particular approximation represents the shaded area in the figure on the right above. Note that, because  $e^x$  increases as  $x$  increases, this approximation is definitely larger than the true area.

Example 3.1.1

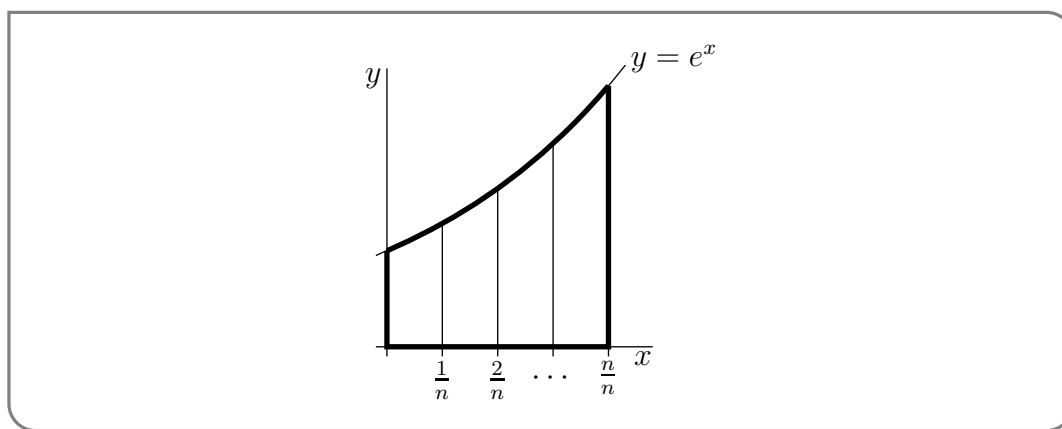
Now for the full computation that gives the exact area.

Example 3.1.2

Recall that we wish to compute the area of  $\{ (x, y) \mid 0 \leq y \leq e^x, 0 \leq x \leq 1 \}$  and that our strategy is to approximate this area by the area of a union of a large number of very thin rectangles, and then take the limit as the number of rectangles goes to infinity. In Example 3.1.1, we used just four rectangles. Now we'll consider a general number of rectangles, that we'll call  $n$ . Then we'll take the limit  $n \rightarrow \infty$ . So

- pick a natural number  $n$  and
- subdivide the interval  $0 \leq x \leq 1$  into  $n$  equal subintervals each of width  $1/n$ , and
- subdivide the area of interest into corresponding thin strips, as in the figure below.

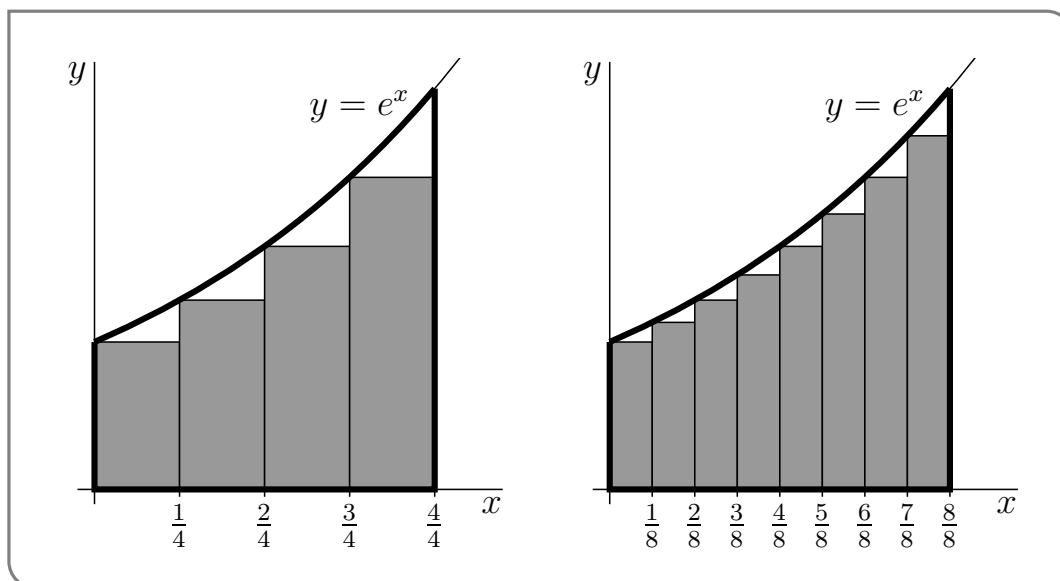
The area we want is exactly the sum of the areas of all of the thin strips.



Each of these strips is almost, but not quite, a rectangle. As in Example 3.1.1, the only problem is that the top is not horizontal. So we approximate each strip by a rectangle, just by levelling off the top. Again, we have to make a choice — at what height do we level off the top?

Consider, for example, the leftmost strip. On this strip,  $x$  runs from 0 to  $1/n$ . As  $x$  runs from 0 to  $1/n$ , the height  $y$  runs from  $e^0$  to  $e^{1/n}$ . It would be reasonable to choose the height of the approximating rectangle to be somewhere between  $e^0$  and  $e^{1/n}$ . Which height should we choose?

Well, as we said in Example 3.1.1, it doesn't matter. We shall shortly take the limit  $n \rightarrow \infty$  and, in that limit, all of those different choices give exactly the same final answer. We won't justify that statement in this example, but Appendix section A.10.4 provides the justification. For this example we just, arbitrarily, choose the height of each rectangle to be the height of the graph  $y = e^x$  at the smallest value of  $x$  in the corresponding strip<sup>4</sup>. The figure on the left below shows the approximating rectangles when  $n = 4$  and the figure on the right shows the approximating rectangles when  $n = 8$ .



Now we compute the approximating area when there are  $n$  strips.

- We approximate the leftmost strip by a rectangle of height  $e^0$ . All of the rectangles have width  $1/n$ . So the leftmost rectangle has area  $\frac{1}{n}e^0$ .
- On strip number 2,  $x$  runs from  $\frac{1}{n}$  to  $\frac{2}{n}$ . So the smallest value of  $x$  on strip number 2 is  $\frac{1}{n}$ , and we approximate strip number 2 by a rectangle of height  $e^{1/n}$  and hence of area  $\frac{1}{n}e^{1/n}$ .
- And so on.
- On the last strip,  $x$  runs from  $\frac{n-1}{n}$  to  $\frac{n}{n} = 1$ . So the smallest value of  $x$  on the last strip is  $\frac{n-1}{n}$ , and we approximate the last strip by a rectangle of height  $e^{(n-1)/n}$  and hence of area  $\frac{1}{n}e^{(n-1)/n}$ .

The total area of all of the approximating rectangles is

$$\begin{aligned} \text{Total approximating area} &= \frac{1}{n}e^0 + \frac{1}{n}e^{1/n} + \frac{1}{n}e^{2/n} + \frac{1}{n}e^{3/n} + \cdots + \frac{1}{n}e^{(n-1)/n} \\ &= \frac{1}{n} \left( 1 + e^{1/n} + e^{2/n} + e^{3/n} + \cdots + e^{(n-1)/n} \right) \end{aligned}$$

Now the sum in the brackets might look a little intimidating because of all the exponentials, but it actually has a pretty simple structure that can be easily seen if we rename  $e^{1/n} = r$ . Then

4 Notice that since  $e^x$  is an increasing function, this choice of heights means that each of our rectangles is smaller than the strip it came from.

- the first term is  $1 = r^0$  and
- the second term is  $e^{1/n} = r^1$  and
- the third term is  $e^{2/n} = r^2$  and
- the fourth term is  $e^{3/n} = r^3$  and
- and so on and
- the last term is  $e^{(n-1)/n} = r^{n-1}$ .

So

$$\text{Total approximating area} = \frac{1}{n} (1 + r + r^2 + \dots + r^{n-1})$$

The sum in brackets is known as a geometric sum and satisfies a nice simple formula:

**Equation 3.1.3**(Geometric sum).

$$1 + r + r^2 + \dots + r^{n-1} = \frac{r^n - 1}{r - 1} \quad \text{provided } r \neq 1$$

The derivation of the above formula is not too difficult. So let's derive it in a little aside.

### ▶▶▶ Geometric Sum

Denote the sum as

$$S = 1 + r + r^2 + \dots + r^{n-1}$$

Notice that if we multiply the whole sum by  $r$  we get back almost the same thing:

$$\begin{aligned} rS &= r(1 + r + r^2 + \dots + r^{n-1}) \\ &= r + r^2 + r^3 + \dots + r^n \end{aligned}$$

This right hand side differs from the original sum  $S$  only in that

- the right hand side is missing the "1+" that  $S$  starts with and
- the right hand side has an extra "+ $r^n$ " at the end that does not appear in  $S$ .

That is

$$rS = S - 1 + r^n$$

Moving this around a little gives

$$\begin{aligned} (r - 1)S &= (r^n - 1) \\ S &= \frac{r^n - 1}{r - 1} \end{aligned}$$

as required. Notice that the last step in the manipulations only works providing  $r \neq 1$  (otherwise we are dividing by zero).

### ▶▶▶ Back to Approximating Areas

Now we can go back to our area approximation armed with the above result about geometric sums.

$$\begin{aligned}
 \text{Total approximating area} &= \frac{1}{n} \left( 1 + r + r^2 + \dots + r^{n-1} \right) \\
 &= \frac{1}{n} \frac{r^n - 1}{r - 1} && \text{remember that } r = e^{1/n} \\
 &= \frac{1}{n} \frac{e^{n/n} - 1}{e^{1/n} - 1} \\
 &= \frac{1}{n} \frac{e - 1}{e^{1/n} - 1}
 \end{aligned}$$

To get the exact area<sup>5</sup> all we need to do is make the approximation better and better by taking the limit  $n \rightarrow \infty$ . The limit will look more familiar if we rename  $1/n$  to  $X$ . As  $n$  tends to infinity,  $X$  tends to 0, so

$$\begin{aligned}
 \text{Area} &= \lim_{n \rightarrow \infty} \frac{1}{n} \frac{e - 1}{e^{1/n} - 1} \\
 &= (e - 1) \lim_{n \rightarrow \infty} \frac{1/n}{e^{1/n} - 1} \\
 &= (e - 1) \lim_{X \rightarrow 0} \frac{X}{e^X - 1} && \text{(with } X = 1/n)
 \end{aligned}$$

Examining this limit we see that both numerator and denominator tend to zero as  $X \rightarrow 0$ , and so we cannot evaluate this limit by computing the limits of the numerator and denominator separately and then dividing the results. Despite this, the limit is not too hard to evaluate; here we give two ways:

- Perhaps the easiest way to compute the limit is by using l'Hôpital's rule<sup>6</sup>. Since both numerator and denominator go to zero, this is a  $0/0$  indeterminate form. Thus

$$\lim_{X \rightarrow 0} \frac{X}{e^X - 1} = \lim_{X \rightarrow 0} \frac{\frac{d}{dX} X}{\frac{d}{dX} (e^X - 1)} = \lim_{X \rightarrow 0} \frac{1}{e^X} = 1$$

- Another way<sup>7</sup> to evaluate the same limit is to observe that it can be massaged into the form of the limit definition of the derivative. First notice that

$$\lim_{X \rightarrow 0} \frac{X}{e^X - 1} = \left[ \lim_{X \rightarrow 0} \frac{e^X - 1}{X} \right]^{-1}$$

5 We haven't proved that this will give us the exact area, but it should be clear that taking this limit will give us a lower bound on the area. To complete things rigorously we also need an upper bound and the Squeeze Theorem. We do this in the next optional subsection.

6 If you do not recall l'Hôpital's rule and indeterminate forms then we recommend you skim over your differential calculus notes on the topic.

7 Say if you don't recall l'Hôpital's rule and have not had time to revise it.

provided this second limit exists and is nonzero. This second limit should look a little familiar:

$$\lim_{X \rightarrow 0} \frac{e^X - 1}{X} = \lim_{X \rightarrow 0} \frac{e^X - e^0}{X - 0}$$

which is just the definition of the derivative of  $e^x$  at  $x = 0$ . Hence we have

$$\begin{aligned} \lim_{X \rightarrow 0} \frac{X}{e^X - 1} &= \left[ \lim_{X \rightarrow 0} \frac{e^X - e^0}{X - 0} \right]^{-1} \\ &= \left[ \frac{d}{dX} e^X \Big|_{X=0} \right]^{-1} \\ &= \left[ e^X \Big|_{X=0} \right]^{-1} \\ &= 1 \end{aligned}$$

So, after this short aside into limits, we may now conclude that

$$\begin{aligned} \text{Area} &= (e - 1) \lim_{X \rightarrow 0} \frac{X}{e^X - 1} \\ &= e - 1 \end{aligned}$$

Example 3.1.2

A more rigorous area computation can be found in Appendix A.7

### 3.1.1 ► Summation Notation

As you can see from the above example (and the more careful rigorous computation), our discussion of integration will involve a fair bit of work with sums of quantities. To this end, we make a quick aside into summation notation. While one can work through the material below without this notation, proper summation notation is well worth learning, so we advise the reader to persevere.

Writing out the summands explicitly can become quite impractical — for example, say we need the sum of the first 11 squares:

$$1 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 + 7^2 + 8^2 + 9^2 + 10^2 + 11^2$$

This becomes tedious. Where the pattern is clear, we will often skip the middle few terms and instead write

$$1 + 2^2 + \cdots + 11^2.$$

A far more precise way to write this is using  $\Sigma$  (capital-sigma) notation. For example, we can write the above sum as

$$\sum_{k=1}^{11} k^2$$

This is read as

The sum from  $k$  equals 1 to 11 of  $k^2$ .

More generally

**Notation 3.1.4.**

Let  $m \leq n$  be integers and let  $f(x)$  be a function defined on the integers. Then we write

$$\sum_{k=m}^n f(k)$$

to mean the sum of  $f(k)$  for  $k$  from  $m$  to  $n$ :

$$f(m) + f(m+1) + f(m+2) + \cdots + f(n-1) + f(n).$$

Similarly we write

$$\sum_{i=m}^n a_i$$

to mean

$$a_m + a_{m+1} + a_{m+2} + \cdots + a_{n-1} + a_n$$

for some set of coefficients  $\{a_m, \dots, a_n\}$ .

Consider the example

$$\sum_{k=3}^7 \frac{1}{k^2} = \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \frac{1}{7^2}$$

It is important to note that the right hand side of this expression evaluates to a number<sup>8</sup>; it does not contain “ $k$ ”. The summation index  $k$  is just a “dummy” variable and it does not have to be called  $k$ . For example

$$\sum_{k=3}^7 \frac{1}{k^2} = \sum_{i=3}^7 \frac{1}{i^2} = \sum_{j=3}^7 \frac{1}{j^2} = \sum_{\ell=3}^7 \frac{1}{\ell^2}$$

Also the summation index has no meaning outside the sum. For example

$$k \sum_{k=3}^7 \frac{1}{k^2}$$

has no mathematical meaning; it is gibberish.

8 Some careful addition shows it is  $\frac{46181}{176400}$ .

A sum can be represented using summation notation in many different ways. If you are unsure as to whether or not two summation notations represent the same sum, just write out the first few terms and the last couple of terms. For example,

$$\sum_{m=3}^{15} \frac{1}{m^2} = \overbrace{\frac{1}{3^2}}^{m=3} + \overbrace{\frac{1}{4^2}}^{m=4} + \overbrace{\frac{1}{5^2}}^{m=5} + \cdots + \overbrace{\frac{1}{14^2}}^{m=14} + \overbrace{\frac{1}{15^2}}^{m=15}$$

$$\sum_{m=4}^{16} \frac{1}{(m-1)^2} = \overbrace{\frac{1}{3^2}}^{m=4} + \overbrace{\frac{1}{4^2}}^{m=5} + \overbrace{\frac{1}{5^2}}^{m=6} + \cdots + \overbrace{\frac{1}{14^2}}^{m=15} + \overbrace{\frac{1}{15^2}}^{m=16}$$

are equal.

Here is a theorem that gives a few rules for manipulating summation notation.

**Theorem 3.1.5** (Arithmetic of Summation Notation).

Let  $n \geq m$  be integers. Then for all real numbers  $c$  and  $a_i, b_i, m \leq i \leq n$ .

- (a)  $\sum_{i=m}^n ca_i = c \left( \sum_{i=m}^n a_i \right)$
- (b)  $\sum_{i=m}^n (a_i + b_i) = \left( \sum_{i=m}^n a_i \right) + \left( \sum_{i=m}^n b_i \right)$
- (c)  $\sum_{i=m}^n (a_i - b_i) = \left( \sum_{i=m}^n a_i \right) - \left( \sum_{i=m}^n b_i \right)$

*Proof.* We can prove this theorem by just writing out both sides of each equation, and observing that they are equal, by the usual laws of arithmetic<sup>9</sup>. For example, for the first equation, the left and right hand sides are

$$\sum_{i=m}^n ca_i = ca_m + ca_{m+1} + \cdots + ca_n \quad \text{and} \quad c \left( \sum_{i=m}^n a_i \right) = c(a_m + a_{m+1} + \cdots + a_n)$$

They are equal by the usual distributive law. The “distributive law” is the fancy name for  $c(a + b) = ca + cb$ . □

Not many sums can be computed exactly<sup>10</sup>. Here are some that can. The first few are used a lot.

- 9 Since all the sums are finite, this isn’t too hard. More care must be taken when the sums involve an infinite number of terms. We will examine this in Chapter 5.
- 10 Of course, any finite sum can be computed exactly — just sum together the terms. What we mean by “computed exactly” in this context, is that we can rewrite the sum as a simple, and easily evaluated, formula involving the terminals of the sum. For example

$$\sum_{k=m}^n r^k = \frac{r^{n+1} - r^m}{r - 1} \quad \text{provided } r \neq 1$$

**Theorem 3.1.6.**

- (a)  $\sum_{i=0}^n ar^i = a \frac{1-r^{n+1}}{1-r}$ , for all real numbers  $a$  and  $r \neq 1$  and all integers  $n \geq 0$ .
- (b)  $\sum_{i=1}^n 1 = n$ , for all integers  $n \geq 1$ .
- (c)  $\sum_{i=1}^n i = \frac{1}{2}n(n+1)$ , for all integers  $n \geq 1$ .
- (d)  $\sum_{i=1}^n i^2 = \frac{1}{6}n(n+1)(2n+1)$ , for all integers  $n \geq 1$ .
- (e)  $\sum_{i=1}^n i^3 = \left[ \frac{1}{2}n(n+1) \right]^2$ , for all integers  $n \geq 1$ .

**▶▶▶ Proof of Theorem 3.1.6**

*Proof.* (a) The first sum is

$$\sum_{i=0}^n ar^i = ar^0 + ar^1 + ar^2 + \cdots + ar^n$$

which is just the left hand side of equation (3.1.3), with  $n$  replaced by  $n+1$  and then multiplied by  $a$ .

(b) The second sum is just  $n$  copies of 1 added together, so of course the sum is  $n$ .

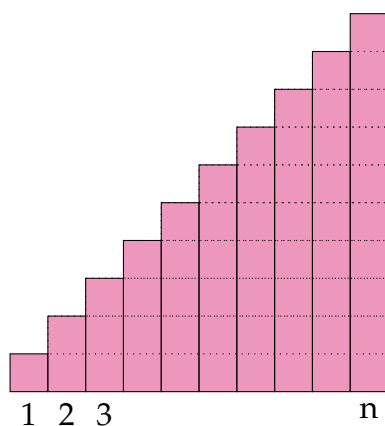
(c) The sum  $\sum_{i=1}^n i = 1 + 2 + 3 + \cdots + n$  can be visualized as the area of the red stairsteps below: the first column has area 1, the second column has area 2, and so on.

No matter what finite integers we choose for  $m$  and  $n$ , we can quickly compute the sum in just a few arithmetic operations. On the other hand, the sums,

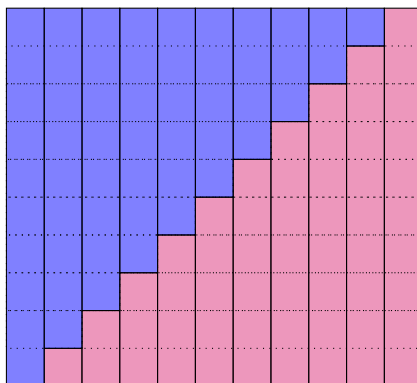
$$\sum_{k=m}^n \frac{1}{k} \qquad \sum_{k=m}^n \frac{1}{k^2}$$

cannot be expressed in such clean formulas (though you can rewrite them quite cleanly using integrals). To explain more clearly we would need to go into a more detailed and careful discussion that is beyond the scope of this course.





If we duplicate those stairsteps and spin them around, we make a rectangle with base  $n + 1$  and height  $n$ .



Since the red stairsteps are exactly half the total area of that rectangle,

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

(d) The last two identities are proved in Question 29 of Section 5.2 of the practice book. □

### 3.1.2 ▶ The Definition of the Definite Integral

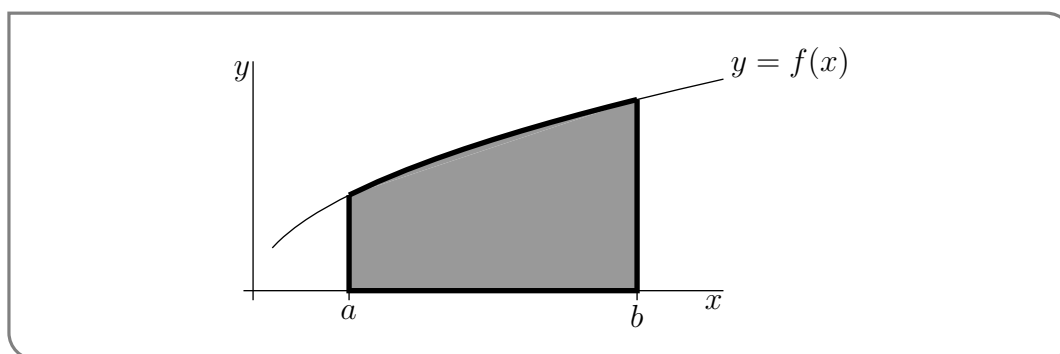
In this section we give a definition of the definite integral  $\int_a^b f(x)dx$  generalising the machinery we used in Example 3.1.1. But first some terminology and a couple of remarks to better motivate the definition.

**Notation 3.1.7.**

The symbol  $\int_a^b f(x)dx$  is read “the definite integral of the function  $f(x)$  from  $a$  to  $b$ ”. The function  $f(x)$  is called the integrand of  $\int_a^b f(x)dx$  and  $a$  and  $b$  are called<sup>11</sup> the limits of integration. The interval  $a \leq x \leq b$  is called the interval of integration and is also called the domain of integration.

Before we explain more precisely what the definite integral actually is, a few remarks (actually — a few interpretations) are in order.

- If  $f(x) \geq 0$  and  $a \leq b$ , one interpretation of the symbol  $\int_a^b f(x)dx$  is “the area of the region  $\{ (x, y) \mid a \leq x \leq b, 0 \leq y \leq f(x) \}$ ”.



In this way we can rewrite the area in Example 3.1.1 as the definite integral  $\int_0^1 e^x dx$ .

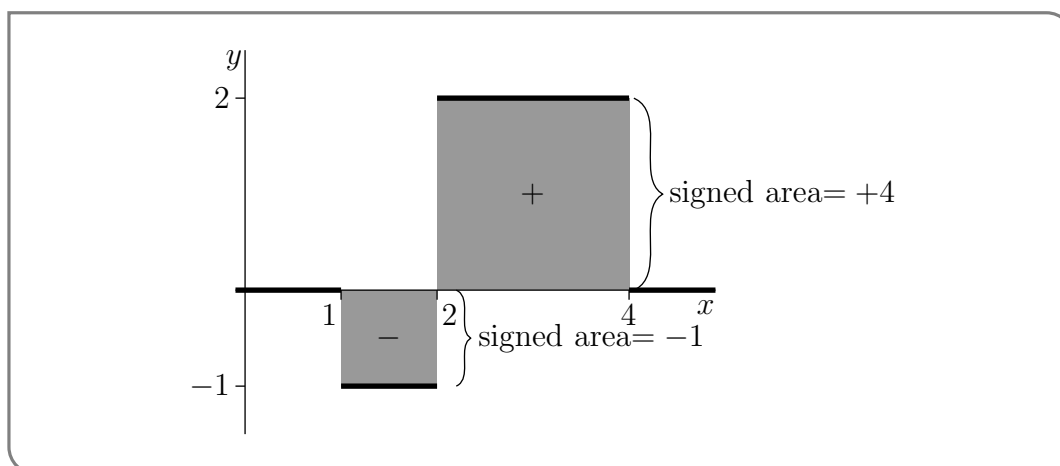
- This interpretation breaks down when either  $a > b$  or  $f(x)$  is not always positive, but it can be repaired by considering “signed areas”.
- If  $a \leq b$ , but  $f(x)$  is not always positive, one interpretation of  $\int_a^b f(x)dx$  is “the signed area between  $y = f(x)$  and the  $x$ -axis for  $a \leq x \leq b$ ”. For “signed area” (which is also called the “net area”), areas above the  $x$ -axis count as positive while areas below the  $x$ -axis count as negative. In the example below, we have the graph of the function

$$f(x) = \begin{cases} -1 & \text{if } 1 \leq x \leq 2 \\ 2 & \text{if } 2 < x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

The  $2 \times 2$  shaded square above the  $x$ -axis has signed area  $+2 \times 2 = +4$ . The  $1 \times 1$  shaded square below the  $x$ -axis has signed area  $-1 \times 1 = -1$ . So, for this  $f(x)$ ,

$$\int_0^5 f(x)dx = +4 - 1 = 3$$

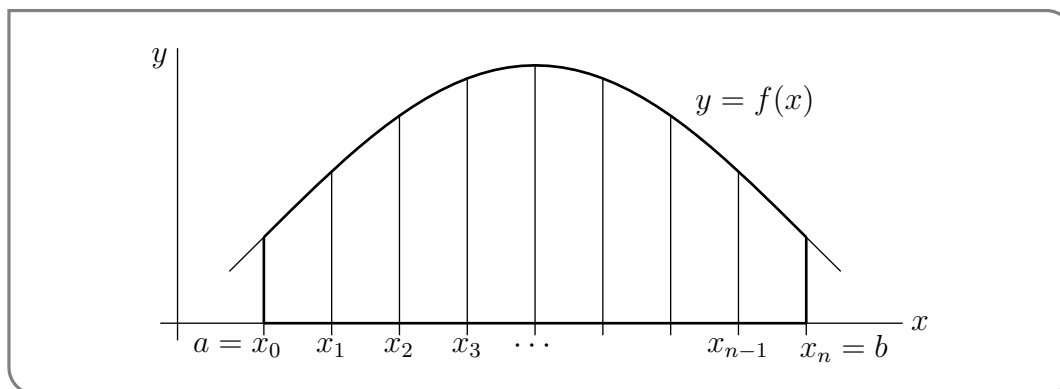
<sup>11</sup>  $a$  and  $b$  are also called the bounds of integration.



- We'll come back to the case  $b < a$  later.

We're now ready to define  $\int_a^b f(x) dx$ . The definition is a little involved, but essentially mimics what we did in Example 3.1.1 (which is why we did the example before the definition). The main differences are that we replace the function  $e^x$  by a generic function  $f(x)$  and we replace the interval from 0 to 1 by the generic interval<sup>12</sup> from  $a$  to  $b$ .

- We start by selecting any natural number  $n$  and subdividing the interval from  $a$  to  $b$  into  $n$  equal subintervals. Each subinterval has width  $\frac{b-a}{n}$ .
- Just as was the case in Example 3.1.1 we will eventually take the limit as  $n \rightarrow \infty$ , which squeezes the width of each subinterval down to zero.
- For each integer  $0 \leq i \leq n$ , define  $x_i = a + i \cdot \frac{b-a}{n}$ . Note that this means that  $x_0 = a$  and  $x_n = b$ . It is worth keeping in mind that these numbers  $x_i$  do depend on  $n$  even though our choice of notation hides this dependence.
- Subinterval number  $i$  is  $x_{i-1} \leq x \leq x_i$ . In particular, on the first subinterval,  $x$  runs from  $x_0 = a$  to  $x_1 = a + \frac{b-a}{n}$ . On the second subinterval,  $x$  runs from  $x_1$  to  $x_2 = a + 2\frac{b-a}{n}$ .



12 We'll eventually allow  $a$  and  $b$  to be any two real numbers, not even requiring  $a < b$ . But it is easier to start off assuming  $a < b$ , and that's what we'll do.

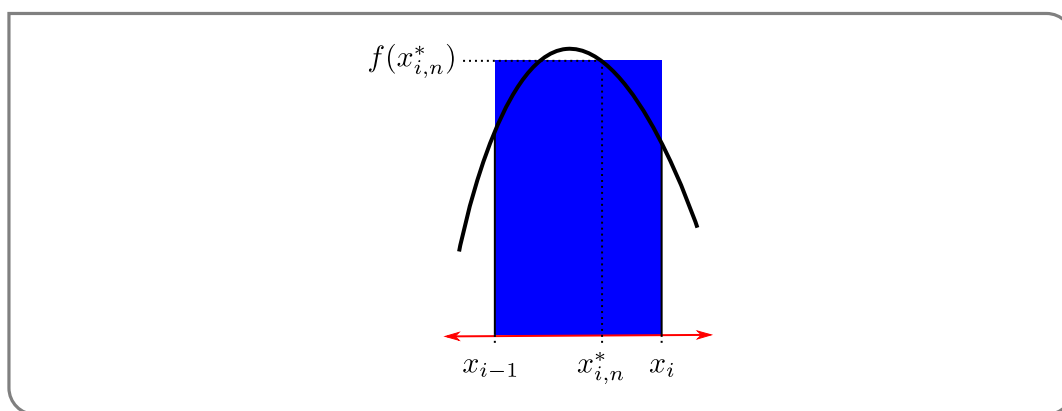
- On each subinterval we now pick  $x_{i,n}^*$  between  $x_{i-1}$  and  $x_i$ . We then approximate  $f(x)$  on the  $i^{\text{th}}$  subinterval by the constant function  $y = f(x_{i,n}^*)$ . We include  $n$  in the subscript to remind ourselves that these numbers depend on  $n$ .

Geometrically, we're approximating the region

$$\{ (x, y) \mid x \text{ is between } x_{i-1} \text{ and } x_i, \text{ and } y \text{ is between } 0 \text{ and } f(x) \}$$

by the rectangle

$$\{ (x, y) \mid x \text{ is between } x_{i-1} \text{ and } x_i, \text{ and } y \text{ is between } 0 \text{ and } f(x_{i,n}^*) \}$$



In Example 3.1.1 we chose  $x_{i,n}^* = x_{i-1}$  and so we approximated the function  $e^x$  on each subinterval by the value it took at the leftmost point in that subinterval.

- So, when there are  $n$  subintervals our approximation to the signed area between the curve  $y = f(x)$  and the  $x$ -axis, with  $x$  running from  $a$  to  $b$ , is

$$\sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n}$$

We interpret this as the signed area since the summands  $f(x_{i,n}^*) \cdot \frac{b-a}{n}$  need not be positive.

- Finally we define the definite integral by taking the limit of this sum as  $n \rightarrow \infty$ .

Oof! This is quite an involved process, but we can now write down the definition we need. (A more mathematically rigorous definition of the definite integral  $\int_a^b f(x)dx$  can be found in Appendix A.8.)

**Definition 3.1.8.**

Let  $a$  and  $b$  be two real numbers and let  $f(x)$  be a function that is defined for all  $x$  between  $a$  and  $b$ . Then we define

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n}$$

when the limit exists and takes the same value for all choices of the  $x_{i,n}^*$ 's. In this case, we say that  $f$  is integrable on the interval from  $a$  to  $b$ .

Of course, it is not immediately obvious when this limit should exist. Thankfully it is easier for a function to be “integrable” than it is for it to be “differentiable”.

**Theorem 3.1.9.**

Let  $f(x)$  be a function on the interval  $[a, b]$ . If

- $f(x)$  is continuous on  $[a, b]$ , or
- $f(x)$  has a finite number of jump discontinuities on  $[a, b]$  (and is otherwise continuous)

then  $f(x)$  is integrable on  $[a, b]$ .

We will not justify this theorem. But a slightly weaker statement is proved in (the optional) Section A.8. Of course this does not tell us how to actually evaluate any definite integrals — but we will get to that in time.

Some comments:

- Note that, in Definition 3.1.8, we allow  $a$  and  $b$  to be any two real numbers. We do not require that  $a < b$ . That is, even when  $a > b$ , the symbol  $\int_a^b f(x) dx$  is still defined by the formula of Definition 3.1.8. We'll get an interpretation for  $\int_a^b f(x) dx$ , when  $a > b$ , later.
- It is important to note that the definite integral  $\int_a^b f(x) dx$  represents a number, not a function of  $x$ . The integration variable  $x$  is another “dummy” variable, just like the summation index  $i$  in  $\sum_{i=m}^n a_i$  (see Section 3.1.1). The integration variable does not have to be called  $x$ . For example

$$\int_a^b f(x) dx = \int_a^b f(t) dt = \int_a^b f(u) du$$

Just as with summation variables, the integration variable  $x$  has no meaning outside of  $f(x) dx$ . For example

$$x \int_0^1 e^x dx \quad \text{and} \quad \int_0^x e^x dx$$

are both gibberish.

The sum inside definition 3.1.8 is named after Bernhard Riemann<sup>13</sup> who made the first rigorous definition of the definite integral and so placed integral calculus on rigorous footings.

**Definition 3.1.10.**

The sum inside definition 3.1.8

$$\sum_{i=1}^n f(x_{i,n}^*) \frac{b-a}{n}$$

is called a Riemann sum. It is also often written as

$$\sum_{i=1}^n f(x_i^*) \Delta x$$

where  $\Delta x = \frac{b-a}{n}$ .

If we choose  $x_{i,n}^* = x_i = a + i\frac{b-a}{n}$  we obtain the approximation

$$\sum_{i=1}^n f\left(a + i\frac{b-a}{n}\right) \frac{b-a}{n}$$

which is called the “right Riemann sum approximation to  $\int_a^b f(x)dx$  with  $n$  subintervals”. The word “right” signifies that, on each subinterval  $[x_{i-1}, x_i]$  we approximate  $f$  by its value at the right-hand end-point,  $x_i = a + i\frac{b-a}{n}$ , of the subinterval.

In order to compute a definite integral using Riemann sums we need to be able to compute the limit of the sum as the number of summands goes to infinity. This approach is not always feasible and we will soon arrive at other means of computing definite integrals based on antiderivatives. However, Riemann sums also provide us with a good means of approximating definite integrals — if we take  $n$  to be a large, but finite, integer, then the corresponding Riemann sum can be a good approximation of the definite integral. Under certain circumstances this can be strengthened to give rigorous bounds on the integral. Let us revisit Example 3.1.1.

**Example 3.1.11**

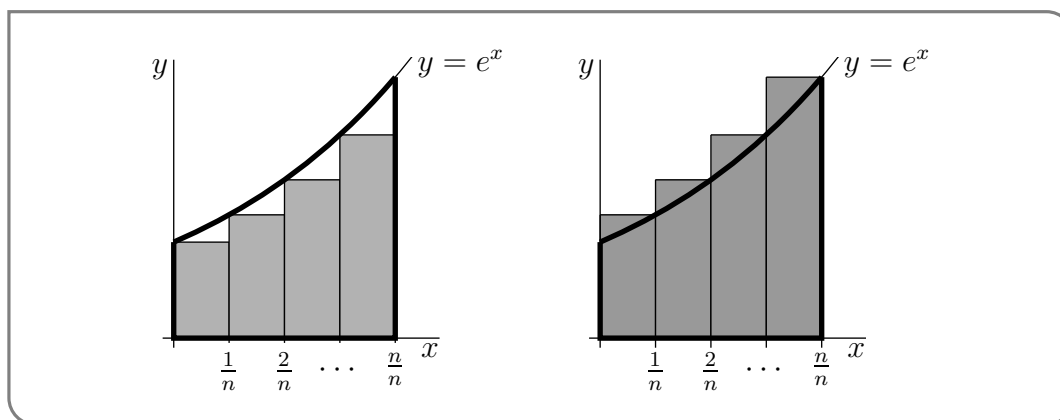
Let’s say we are again interested in the integral  $\int_0^1 e^x dx$ . We can follow the same procedure as we used previously to construct Riemann sum approximations. However since the integrand  $f(x) = e^x$  is an increasing function, we can make our approximations into upper and lower bounds without much extra work.

More precisely, we approximate  $f(x)$  on each subinterval  $x_{i-1} \leq x \leq x_i$

13 Bernhard Riemann was a 19th century German mathematician who made extremely important contributions to many different areas of mathematics — far too many to list here. Arguably two of the most important (after Riemann sums) are now called Riemann surfaces and the Riemann hypothesis (he didn’t name them after himself).

- by its smallest value on the subinterval, namely  $f(x_{i-1})$ , when we compute the Riemann sum approximation using the left endpoint of each interval for  $x_{i,n}^*$ , and
- by its largest value on the subinterval, namely  $f(x_i)$ , when we compute the right Riemann sum approximation.

This is illustrated in the two figures below. The shaded region in the left hand figure is the Riemann sum approximation using the left endpoint of each interval, and the shaded region in the right hand figure is the right Riemann sum approximation.



We can see that exactly because  $f(x)$  is increasing, the first Riemann sum (using the left endpoints of each interval for  $x_{i,n}^*$ ) describes an area smaller than the definite integral, while the right Riemann sum gives an area larger<sup>14</sup> than the integral.

When we approximate the integral  $\int_0^1 e^x dx$  using  $n$  subintervals, then, on interval number  $i$ ,

- $x$  runs from  $\frac{i-1}{n}$  to  $\frac{i}{n}$  and
- $y = e^x$  runs from  $e^{(i-1)/n}$ , when  $x$  is at the left hand end point of the interval, to  $e^{i/n}$ , when  $x$  is at the right hand end point of the interval.

Consequently, the Riemann sum approximation to  $\int_0^1 e^x dx$  using the left endpoint of each interval is  $\sum_{i=1}^n e^{(i-1)/n} \frac{1}{n}$  and the right Riemann sum approximation is  $\sum_{i=1}^n e^{i/n} \cdot \frac{1}{n}$ . So

$$\sum_{i=1}^n e^{(i-1)/n} \frac{1}{n} \leq \int_0^1 e^x dx \leq \sum_{i=1}^n e^{i/n} \cdot \frac{1}{n}$$

Thus  $L_n = \sum_{i=1}^n e^{(i-1)/n} \frac{1}{n}$ , which for any  $n$  can be evaluated by computer, is a lower bound on the exact value of  $\int_0^1 e^x dx$  and  $R_n = \sum_{i=1}^n e^{i/n} \frac{1}{n}$ , which for any  $n$  can also be evaluated by computer, is an upper bound on the exact value of  $\int_0^1 e^x dx$ . For example, when  $n = 1000$ ,

14 When a function is decreasing the situation is reversed — the Riemann sum using left endpoints is always larger than the integral while the right Riemann sum is smaller than the integral. For more general functions that both increase and decrease it is perhaps easiest to study each increasing (or decreasing) interval separately.

$L_n = 1.7174$  and  $R_n = 1.7191$  (both to four decimal places) so that, again to four decimal places,

$$1.7174 \leq \int_0^1 e^x dx \leq 1.7191$$

Recall that the exact value is  $e - 1 = 1.718281828 \dots$

Example 3.1.11

So far, we have only a single interpretation<sup>15</sup> for definite integrals — namely areas under graphs. In the following example, we develop a second interpretation.

Example 3.1.12 (Another Interpretation for Definite Integrals)

Suppose that a particle is moving along the  $x$ -axis and suppose that at time  $t$  its velocity is  $v(t)$  (with  $v(t) > 0$  indicating rightward motion and  $v(t) < 0$  indicating leftward motion). What is the change in its  $x$ -coordinate between time  $a$  and time  $b > a$ ?

We'll work this out using a procedure similar to our definition of the integral. First pick a natural number  $n$  and divide the time interval from  $a$  to  $b$  into  $n$  equal subintervals, each of width  $\frac{b-a}{n}$ . We are working our way towards a Riemann sum (as we have done several times above) and so we will eventually take the limit  $n \rightarrow \infty$ .

- The first time interval runs from  $a$  to  $a + \frac{b-a}{n}$ . If we think of  $n$  as some large number, the width of this interval,  $\frac{b-a}{n}$  is very small and over this time interval, the velocity does not change very much. Hence we can approximate the velocity over the first subinterval as being essentially constant at its value at the end of the time interval —  $v\left(a + \frac{a+b}{n}\right)$ . Over the subinterval the  $x$ -coordinate changes by velocity times time, namely  $v\left(a + \frac{a+b}{n}\right) \cdot \frac{b-a}{n}$ .
- Similarly, the second interval runs from time  $a + \frac{b-a}{n}$  to time  $a + 2\frac{b-a}{n}$ . Again, we can assume that the velocity does not change very much and so we can approximate the velocity as being essentially constant at its value at the end of the subinterval — namely  $v\left(a + 2\frac{b-a}{n}\right)$ . So during the second subinterval the particle's  $x$ -coordinate changes by approximately  $v\left(a + 2\frac{b-a}{n}\right) \frac{b-a}{n}$ .
- In general, time subinterval number  $i$  runs from  $a + (i-1)\frac{b-a}{n}$  to  $a + i\frac{b-a}{n}$  and during this subinterval the particle's  $x$ -coordinate changes, essentially, by

$$v\left(a + i\frac{b-a}{n}\right) \frac{b-a}{n}.$$

<sup>15</sup> If this were the only interpretation then integrals would be a nice mathematical curiosity and unlikely to be the core topic of a large first year mathematics course.



So the net change in  $x$ -coordinate from time  $a$  to time  $b$  is approximately

$$\begin{aligned}
 &v\left(a + \frac{b-a}{n}\right) \frac{b-a}{n} + v\left(a + 2\frac{b-a}{n}\right) \frac{b-a}{n} + \cdots + v\left(a + i\frac{b-a}{n}\right) \frac{b-a}{n} + \cdots \\
 &\qquad\qquad\qquad + v\left(a + n\frac{b-a}{n}\right) \frac{b-a}{n} \\
 &= \sum_{i=1}^n v\left(a + i\frac{b-a}{n}\right) \frac{b-a}{n}
 \end{aligned}$$

This exactly the right Riemann sum approximation to the integral of  $v$  from  $a$  to  $b$  with  $n$  subintervals. The limit as  $n \rightarrow \infty$  is exactly the definite integral  $\int_a^b v(t)dt$ . Following tradition, we have called the (dummy) integration variable  $t$  rather than  $x$  to remind us that it is time that is running from  $a$  to  $b$ .

The conclusion of the above discussion is that if a particle is moving along the  $x$ -axis and its  $x$ -coordinate and velocity at time  $t$  are  $x(t)$  and  $v(t)$ , respectively, then, for all  $b > a$ ,

$$x(b) - x(a) = \int_a^b v(t)dt.$$

Example 3.1.12

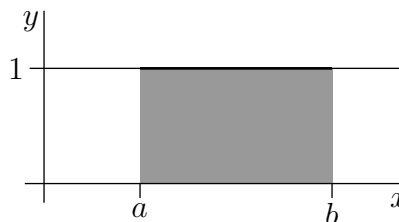
### 3.1.3 ▶ Using Known Areas to Evaluate Integrals

One of the main aims of this course is to build up general machinery for computing definite integrals (as well as interpreting and applying them). We shall start on this soon, but not quite yet. We have already seen one concrete, if laborious, method for computing definite integrals — taking limits of Riemann sums as we did in Example 3.1.1. A second method, which will work for some special integrands, works by interpreting the definite integral as “signed area”. This approach will work nicely when the area under the curve decomposes into simple geometric shapes like triangles, rectangles and circles. Here are some examples of this second method.

Example 3.1.13

The integral  $\int_a^b 1dx$  (which is also written as just  $\int_a^b dx$ ) is the area of the shaded rectangle (of width  $b - a$  and height 1) in the figure on the right below. So

$$\int_a^b dx = (b - a) \times (1) = b - a$$

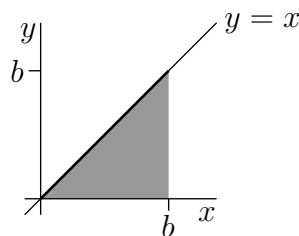


Example 3.1.13

## Example 3.1.14

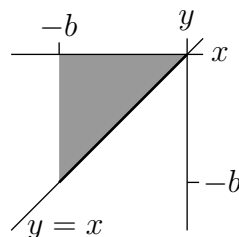
Let  $b > 0$ . The integral  $\int_0^b x dx$  is the area of the shaded triangle (of base  $b$  and of height  $b$ ) in the figure on the right below. So

$$\int_0^b x dx = \frac{1}{2}b \times b = \frac{b^2}{2}$$



The integral  $\int_{-b}^0 x dx$  is the signed area of the shaded triangle (again of base  $b$  and of height  $b$ ) in the figure on the right below. So

$$\int_{-b}^0 x dx = -\frac{b^2}{2}$$



## Example 3.1.14

Notice that it is very easy to extend this example to the integral  $\int_0^b cx dx$  for any real numbers  $b, c > 0$  and find

$$\int_0^b cx dx = \frac{c}{2}b^2.$$

## Example 3.1.15

In this example, we shall evaluate  $\int_{-1}^1 (1 - |x|) dx$ . Recall that

$$|x| = \begin{cases} -x & \text{if } x \leq 0 \\ x & \text{if } x \geq 0 \end{cases}$$

so that

$$1 - |x| = \begin{cases} 1 + x & \text{if } x \leq 0 \\ 1 - x & \text{if } x \geq 0 \end{cases}$$

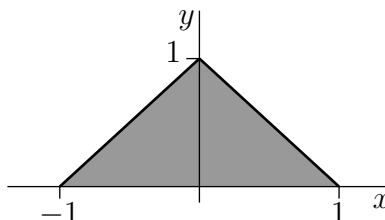
To picture the geometric figure whose area the integral represents observe that

- at the left hand end of the domain of integration  $x = -1$  and the integrand  $1 - |x| = 1 - |-1| = 1 - 1 = 0$  and
- as  $x$  increases from  $-1$  towards  $0$ , the integrand  $1 - |x| = 1 + x$  increases linearly, until

- when  $x$  hits 0 the integrand hits  $1 - |x| = 1 - |0| = 1$  and then
- as  $x$  increases from 0, the integrand  $1 - |x| = 1 - x$  decreases linearly, until
- when  $x$  hits +1, the right hand end of the domain of integration, the integrand hits  $1 - |x| = 1 - |1| = 0$ .

So the integral  $\int_{-1}^1 (1 - |x|) dx$  is the area of the shaded triangle (of base 2 and of height 1) in the figure on the right below and

$$\int_{-1}^1 (1 - |x|) dx = \frac{1}{2} \times 2 \times 1 = 1$$



Example 3.1.15

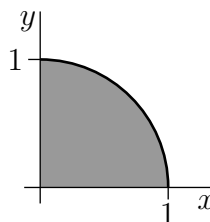
Example 3.1.16

The integral  $\int_0^1 \sqrt{1 - x^2} dx$  has integrand  $f(x) = \sqrt{1 - x^2}$ . So it represents the area under  $y = \sqrt{1 - x^2}$  with  $x$  running from 0 to 1. But we may rewrite

$$y = \sqrt{1 - x^2} \quad \text{as} \quad x^2 + y^2 = 1, y \geq 0$$

But this is the (implicit) equation for a circle — the extra condition that  $y \geq 0$  makes it the equation for the semi-circle centred at the origin with radius 1 lying on and above the  $x$ -axis. Thus the integral represents the area of the quarter circle of radius 1, as shown in the figure on the right below. So

$$\int_0^1 \sqrt{1 - x^2} dx = \frac{1}{4} \pi (1)^2 = \frac{\pi}{4}$$



Example 3.1.16

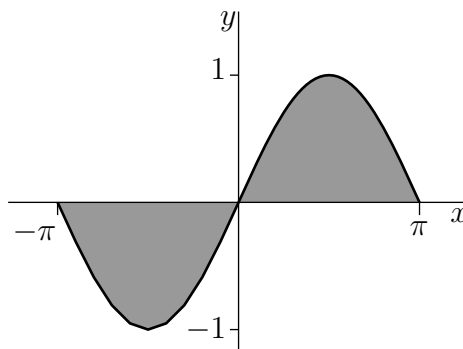
This next one is a little trickier and relies on us knowing the symmetries of the sine function.

Example 3.1.17

The integral  $\int_{-\pi}^{\pi} \sin x dx$  is the signed area of the shaded region in the figure on the right below. It naturally splits into two regions, one on either side of the  $y$ -axis. We don't know the formula for the area of either of these regions (yet), however the two regions are very

nearly the same. In fact, the part of the shaded region below the  $x$ -axis is exactly the reflection, in the  $x$ -axis, of the part of the shaded region above the  $x$ -axis. So the signed area of part of the shaded region below the  $x$ -axis is the negative of the signed area of part of the shaded region above the  $x$ -axis and

$$\int_{-\pi}^{\pi} \sin x dx = 0$$



Example 3.1.17

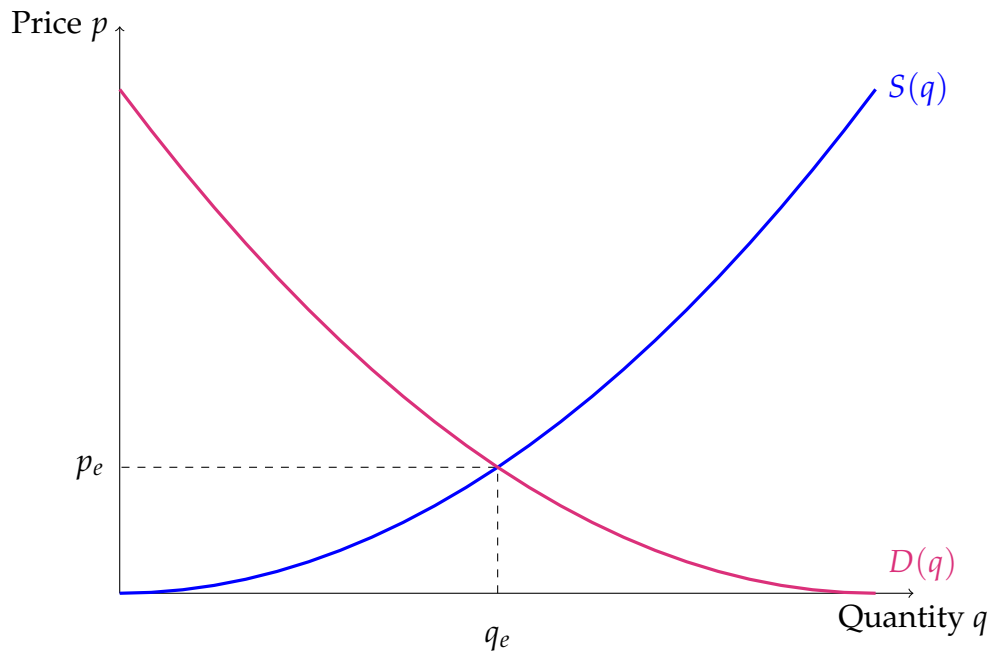
### 3.1.4 ► Surplus

In Section 2.6, we saw demand curves that depended on a consumer's income, their preferences (utility function), and the prices of goods. Now let's use a simplified demand curve:  $D(q)$  is the per-unit price at which a consumer will purchase a quantity  $q$  of a good<sup>16</sup>. Rather than think about individuals' varying utility functions and income, the demand curve imagines a hypothetical "average" consumer.

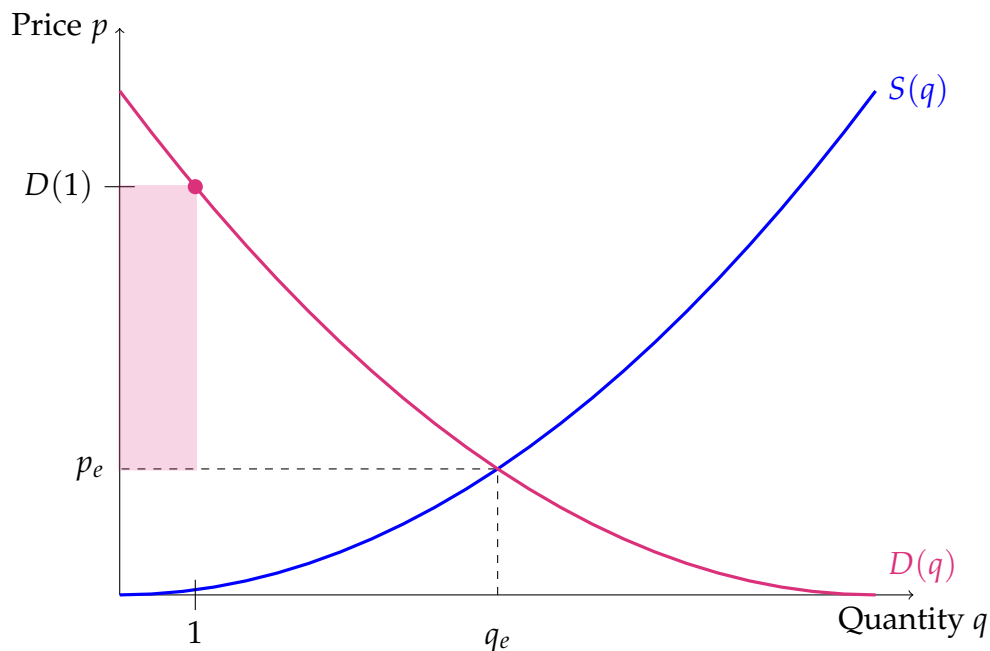
Similarly, we can make a supply curve  $S(q)$  giving the per-unit price at which a supplier is willing to sell  $q$  units.

In simple examples,  $D(q)$  has a negative slope (since, to be motivated to buy a higher quantity, the consumer demands a lower price) and  $S(q)$  has a positive slope (since, to be motivated to sell a higher quantity, the supplier demands a higher price). The quantity and price where the two curves meet are called the *equilibrium quantity* and *equilibrium price*, respectively, and are denoted  $q_e$  resp.  $p_e$ . In theory, suppliers would aim to sell  $q_e$  products at a unit price of  $p_e$ . (If they make more goods, to sell them all they'd have to charge less than they are willing to accept. If they make fewer goods, they will not meet consumer demand.)

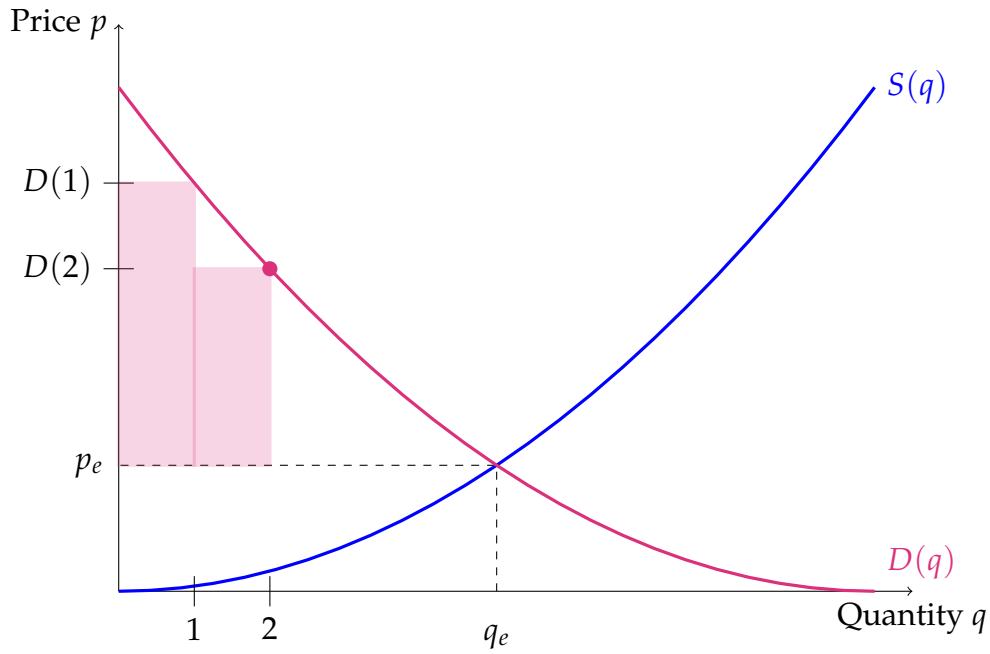
<sup>16</sup> The more natural way of thinking about this is reversed: given the price, how much quantity will the consumer purchase. But formulating the relationship where price is a function of quantity (rather than the other way around) is standard practice in economics texts, so we follow it here.



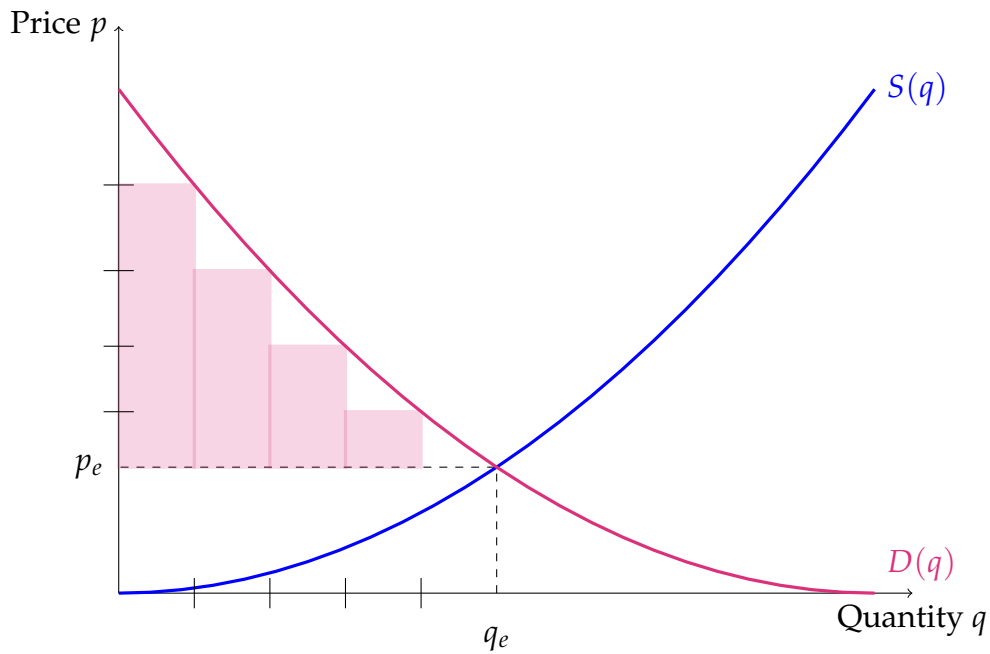
The consumer would have been happy to buy their first good at the price  $D(1)$ . We can say then that the first good has a value of  $D(1)$  for the consumer. If they paid a lower price  $p_e$ , then the number  $D(1) - p_e$  is a surplus to the consumer: they gained  $D(1)$  units of value by paying only  $p_e$  units of value. This surplus can be visualized as the shaded area below.



Similarly, the consumer would have been happy to buy their second good at the unit price  $D(2)$ . If they paid a smaller price  $p_e$ , then their surplus from that second good is  $D(2) - p_e$ : its value to them, minus what they actually paid. Their combined surplus after buying two goods can be visualized as the shaded rectangles below.



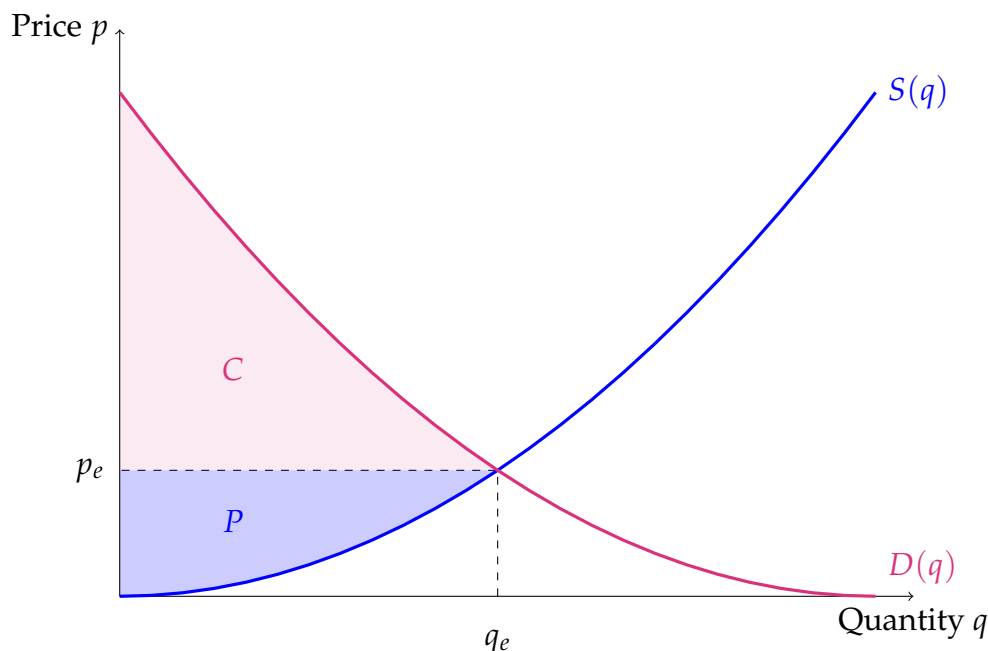
All together, we expect the consumer to buy  $q_e$  units. Their total surplus is represented by the shaded rectangles below.



This motivates the definition of consumer surplus. Producer surplus behaves similarly.

**Definition 3.1.18.**

Consider a supply curve  $S(q)$  and a demand curve  $D(q)$  with intersection point  $(q_e, p_e)$ , graphed on the  $(q, p)$ -plane. The **consumer surplus** is the area from  $q = 0$  to  $q = q_e$  under  $D(q)$  and above the line  $p = p_e$ . The **producer surplus** is the area from  $q = 0$  to  $q = q_e$  over  $S(q)$  and under the line  $p = p_e$ . The **total surplus** is the sum of consumer surplus and producer surplus.



Given a sale of  $q_e$  items at unit price  $p_e$ , we think of the consumer surplus as the net benefit to the consumer, and the producer surplus as the net benefit to the producer. To calculate these, we need a little geometric intuition. The consumer surplus is the area  $\int_0^{q_e} D(q) dq$  minus the area of the rectangle with width  $q_e$  and height  $p_e$ . So, the consumer surplus is

$$C = \int_0^{q_e} D(q) dq - p_e q_e$$

Similarly, the producer surplus is the area of the rectangle with width  $q_e$  and height  $p_e$ , minus the area  $\int_0^{q_e} S(q) dq$ . So, the producer surplus is

$$P = p_e q_e - \int_0^{q_e} S(q) dq$$

Finally, the total surplus is the value gained by everybody, producers and consumers combined:

$$T = C + P = \int_0^{q_e} D(q) dq - \int_0^{q_e} S(q) dq$$

## 3.2▲ Basic Properties of the Definite Integral

When we studied limits and derivatives, we developed methods for taking limits or derivatives of “complicated functions” like  $f(x) = x^2 + \sin(x)$  by understanding how limits and derivatives interact with basic arithmetic operations like addition and subtraction. This allowed us to reduce the problem into one of computing derivatives of simpler functions like  $x^2$  and  $\sin(x)$ . Along the way we established simple rules such as

$$\lim_{x \rightarrow a} (f(x) + g(x)) = \lim_{x \rightarrow a} f(x) + \lim_{x \rightarrow a} g(x) \quad \text{and} \quad \frac{d}{dx} (f(x) + g(x)) = \frac{df}{dx} + \frac{dg}{dx}$$

Some of these rules have very natural analogues for integrals and we discuss them below. Unfortunately the analogous rules for integrals of products of functions or integrals of compositions of functions are more complicated than those for limits or derivatives. We discuss those rules at length in subsequent sections. For now let us consider some of the simpler rules of the arithmetic of integrals.

### Theorem 3.2.1 (Arithmetic of Integration).

Let  $a, b$  and  $A, B, C$  be real numbers. Let the functions  $f(x)$  and  $g(x)$  be integrable on an interval that contains  $a$  and  $b$ . Then

$$\begin{aligned} \text{(a)} \quad & \int_a^b (f(x) + g(x)) \, dx = \int_a^b f(x) \, dx + \int_a^b g(x) \, dx \\ \text{(b)} \quad & \int_a^b (f(x) - g(x)) \, dx = \int_a^b f(x) \, dx - \int_a^b g(x) \, dx \\ \text{(c)} \quad & \int_a^b C f(x) \, dx = C \cdot \int_a^b f(x) \, dx \end{aligned}$$

Combining these three rules we have

$$\text{(d)} \quad \int_a^b (A f(x) + B g(x)) \, dx = A \int_a^b f(x) \, dx + B \int_a^b g(x) \, dx$$

That is, integrals depend linearly on the integrand.

$$\text{(e)} \quad \int_a^b dx = \int_a^b 1 \cdot dx = b - a$$

It is not too hard to prove this result from the definition of the definite integral. Additionally we only really need to prove (d) and (e) since

- (a) follows from (d) by setting  $A = B = 1$ ,
- (b) follows from (d) by setting  $A = 1, B = -1$ , and
- (c) follows from (d) by setting  $A = C, B = 0$ .



*Proof.* As noted above, it suffices for us to prove (d) and (e). Since (e) is easier, we will start with that. It is also a good warm-up for (d).

- The definite integral in (e),  $\int_a^b 1 dx$ , can be interpreted geometrically as the area of the rectangle with height 1 running from  $x = a$  to  $x = b$ ; this area is clearly  $b - a$ . We can also prove this formula from the definition of the integral (Definition 3.1.8):

$$\begin{aligned} \int_a^b dx &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \frac{b-a}{n} && \text{by definition} \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n 1 \frac{b-a}{n} && \text{since } f(x) = 1 \\ &= \lim_{n \rightarrow \infty} (b-a) \sum_{i=1}^n \frac{1}{n} && \text{since } a, b \text{ are constants} \\ &= \lim_{n \rightarrow \infty} (b-a) \\ &= b-a \end{aligned}$$

as required.

- To prove (d) let us start by defining  $h(x) = Af(x) + Bg(x)$  and then we need to express the integral of  $h(x)$  in terms of those of  $f(x)$  and  $g(x)$ . We use Definition 3.1.8 and some algebraic manipulations<sup>17</sup> to arrive at the result.

$$\begin{aligned} \int_a^b h(x) dx &= \sum_{i=1}^n h(x_{i,n}^*) \cdot \frac{b-a}{n} && \text{by Definition 3.1.8} \\ &= \sum_{i=1}^n (Af(x_{i,n}^*) + Bg(x_{i,n}^*)) \cdot \frac{b-a}{n} \\ &= \sum_{i=1}^n \left( Af(x_{i,n}^*) \cdot \frac{b-a}{n} + Bg(x_{i,n}^*) \cdot \frac{b-a}{n} \right) \\ &= \left( \sum_{i=1}^n Af(x_{i,n}^*) \cdot \frac{b-a}{n} \right) + \left( \sum_{i=1}^n Bg(x_{i,n}^*) \cdot \frac{b-a}{n} \right) && \text{by Theorem 3.1.5(b)} \\ &= A \left( \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n} \right) + B \left( \sum_{i=1}^n g(x_{i,n}^*) \cdot \frac{b-a}{n} \right) && \text{by Theorem 3.1.5(a)} \\ &= A \int_a^b f(x) dx + B \int_a^b g(x) dx && \text{by Definition 3.1.8} \end{aligned}$$

as required. □

Using this Theorem we can integrate sums, differences and constant multiples of functions we know how to integrate. For example:

<sup>17</sup> Now is a good time to look back at Theorem 3.1.5.

## Example 3.2.2

In Example 3.1.1 we saw that  $\int_0^1 e^x dx = e - 1$ . So

$$\begin{aligned} \int_0^1 (e^x + 7) dx &= \int_0^1 e^x dx + 7 \int_0^1 1 dx \\ &\quad \text{by Theorem 3.2.1(d) with } A = 1, f(x) = e^x, B = 7, g(x) = 1 \\ &= (e - 1) + 7 \times (1 - 0) \\ &\quad \text{by Example 3.1.1 and Theorem 3.2.1(e)} \\ &= e + 6 \end{aligned}$$

## Example 3.2.2

When we gave the formal definition of  $\int_a^b f(x) dx$  in Definition 3.1.8 we explained that the integral could be interpreted as the signed area between the curve  $y = f(x)$  and the  $x$ -axis on the interval  $[a, b]$ . In order for this interpretation to make sense we required that  $a < b$ , and though we remarked that the integral makes sense when  $a > b$  we did not explain any further. Thankfully there is an easy way to express the integral  $\int_a^b f(x) dx$  in terms of  $\int_b^a f(x) dx$  — making it always possible to write an integral so the lower limit of integration is less than the upper limit of integration. Theorem 3.2.3, below, tells us that, for example,  $\int_7^3 e^x dx = -\int_3^7 e^x dx$ . The same theorem also provides us with two other simple manipulations of the limits of integration.

**Theorem 3.2.3** (Arithmetic for the Domain of Integration).

Let  $a, b, c$  be real numbers. Let the function  $f(x)$  be integrable on an interval that contains  $a, b$  and  $c$ . Then

$$\begin{aligned} \text{(a)} \quad & \int_a^a f(x) dx = 0 \\ \text{(b)} \quad & \int_b^a f(x) dx = -\int_a^b f(x) dx \\ \text{(c)} \quad & \int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx \end{aligned}$$

The proof of this statement is not too difficult.

*Proof.* Let us prove the statements in order.

- Consider the definition of the definite integral

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n}$$

If we now substitute  $b = a$  in this expression we have

$$\begin{aligned} \int_a^a f(x)dx &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \underbrace{\frac{a-a}{n}}_{=0} \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \underbrace{f(x_{i,n}^*)}_{=0} \cdot 0 \\ &= \lim_{n \rightarrow \infty} 0 \\ &= 0 \end{aligned}$$

as required.

- Consider now the definite integral  $\int_A^B f(x)dx$ . Shortly we will substitute  $A = b$  and  $B = a$ , but first let's write down the definition of this integral using Definition 3.1.8.

$$\int_A^B f(x)dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{B-A}{n}$$

Now substitute  $A = b$  and  $B = a$  into this expression:

$$\begin{aligned} \int_b^a f(x)dx &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{a-b}{n} \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot (-1) \cdot \frac{b-a}{n} \\ &= \lim_{n \rightarrow \infty} (-1) \cdot \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n} && \text{by Theorem 3.1.5(a)} \\ &= (-1) \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n} && \text{by arithmetic of limits} \\ &= (-1) \cdot \int_a^b f(x)dx && \text{by Definition 3.1.8} \end{aligned}$$

as required.

(Remark: in the last step, we are glossing over a little fine print about the exact meaning of  $x_{i,n}^*$ .)

- Finally consider (c) — we will not give a formal proof of this, but instead will interpret it geometrically. Indeed one can also interpret (a) geometrically. In both cases these become statements about areas:

$$\int_a^a f(x)dx = 0 \quad \text{and} \quad \int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$$

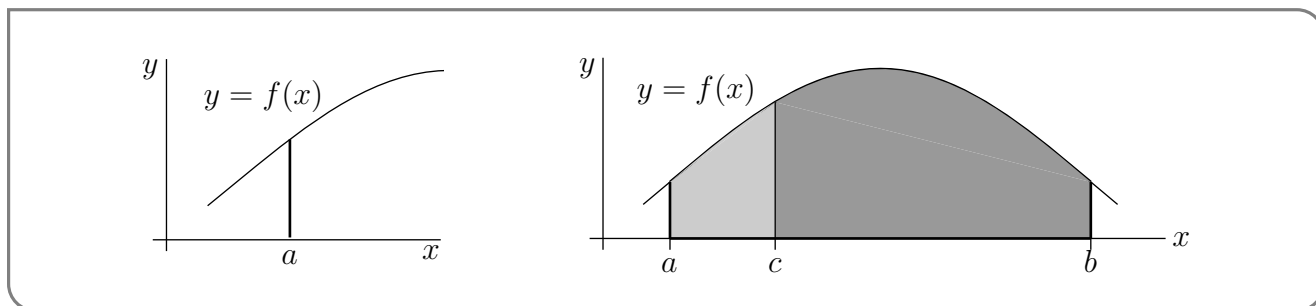
are

$$\text{Area}\{ (x, y) \mid a \leq x \leq a, 0 \leq y \leq f(x) \} = 0$$

and

$$\begin{aligned} \text{Area}\{ (x, y) \mid a \leq x \leq b, 0 \leq y \leq f(x) \} &= \text{Area}\{ (x, y) \mid a \leq x \leq c, 0 \leq y \leq f(x) \} \\ &\quad + \text{Area}\{ (x, y) \mid c \leq x \leq b, 0 \leq y \leq f(x) \} \end{aligned}$$

respectively. Both of these geometric statements are intuitively obvious. See the figures below.



Note that we have assumed that  $a \leq c \leq b$  and that  $f(x) \geq 0$ . One can remove these restrictions and also make the proof more formal, but it becomes quite tedious and less intuitive.

□

Example 3.2.4

Back in Example 3.1.14 we saw that when  $b > 0$   $\int_0^b x dx = \frac{b^2}{2}$ . We'll now verify that  $\int_0^b x dx = \frac{b^2}{2}$  is still true when  $b = 0$  and also when  $b < 0$ .

- First consider  $b = 0$ . Then the statement  $\int_0^b x dx = \frac{b^2}{2}$  becomes

$$\int_0^0 x dx = 0$$

This is an immediate consequence of Theorem 3.2.3(a).

- Now consider  $b < 0$ . Let us write  $B = -b$ , so that  $B > 0$ . In Example 3.1.14 we saw that

$$\int_{-B}^0 x dx = -\frac{B^2}{2}.$$

So we have

$$\begin{aligned} \int_0^b x dx &= \int_0^{-B} x dx = -\int_{-B}^0 x dx && \text{by Theorem 3.2.3(b)} \\ &= -\left(-\frac{B^2}{2}\right) && \text{by Example 3.1.14} \\ &= \frac{B^2}{2} = \frac{b^2}{2} \end{aligned}$$

We have now shown that

$$\int_0^b x dx = \frac{b^2}{2} \quad \text{for all real numbers } b$$

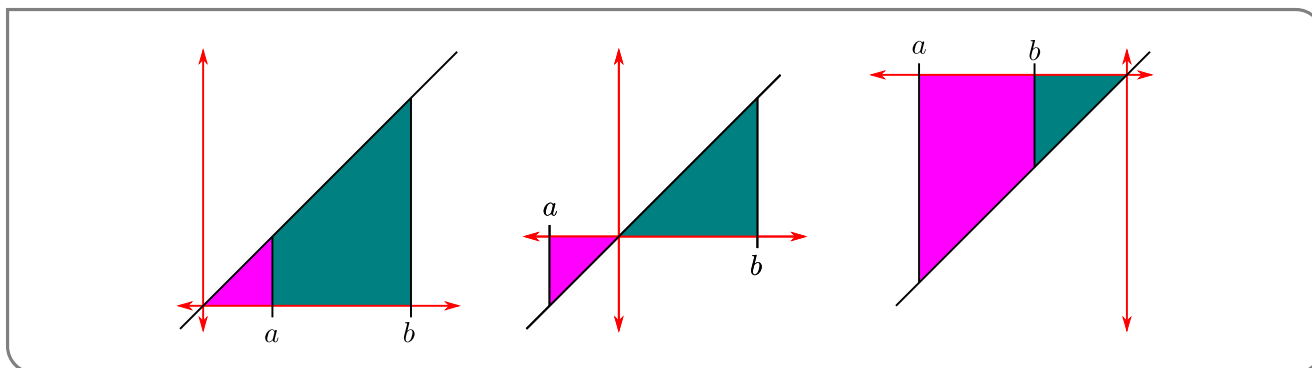
Example 3.2.4

Example 3.2.5

Applying Theorem 3.2.3 yet again, we have, for all real numbers  $a$  and  $b$ ,

$$\begin{aligned} \int_a^b x dx &= \int_a^0 x dx + \int_0^b x dx && \text{by Theorem 3.2.3(c) with } c = 0 \\ &= \int_0^b x dx - \int_0^a x dx && \text{by Theorem 3.2.3(b)} \\ &= \frac{b^2 - a^2}{2} && \text{by Example 3.2.4, twice} \end{aligned}$$

We can also understand this result geometrically.



- (left) When  $0 < a < b$ , the integral represents the area in green which is the difference of two right-angle triangles — the larger with area  $b^2/2$  and the smaller with area  $a^2/2$ .
- (centre) When  $a < 0 < b$ , the integral represents the signed area of the two displayed triangles. The one above the axis has area  $b^2/2$  while the one below has area  $-a^2/2$  (since it is below the axis).
- (right) When  $a < b < 0$ , the integral represents the signed area in purple of the difference between the two triangles — the larger with area  $-a^2/2$  and the smaller with area  $-b^2/2$ .

Example 3.2.5

Theorem 3.2.3(c) shows us how we can split an integral over a larger interval into one over two (or more) smaller intervals. This is particularly useful for dealing with piecewise functions, like  $|x|$ .

**Example 3.2.6**

Using Theorem 3.2.3, we can readily evaluate integrals involving  $|x|$ . First, recall that

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

Now consider (for example)  $\int_{-2}^3 |x| dx$ . Since the integrand changes at  $x = 0$ , it makes sense to split the interval of integration at that point:

$$\begin{aligned} \int_{-2}^3 |x| dx &= \int_{-2}^0 |x| dx + \int_0^3 |x| dx && \text{by Theorem 3.2.3} \\ &= \int_{-2}^0 (-x) dx + \int_0^3 x dx && \text{by definition of } |x| \\ &= -\int_{-2}^0 x dx + \int_0^3 x dx && \text{by Theorem 3.2.1(c)} \\ &= -(-2^2/2) + (3^2/2) = (4 + 9)/2 \\ &= 13/2 \end{aligned}$$

We can go further still — given a function  $f(x)$  we can rewrite the integral of  $f(|x|)$  in terms of the integral of  $f(x)$  and  $f(-x)$ .

$$\begin{aligned} \int_{-1}^1 f(|x|) dx &= \int_{-1}^0 f(|x|) dx + \int_0^1 f(|x|) dx \\ &= \int_{-1}^0 f(-x) dx + \int_0^1 f(x) dx \end{aligned}$$

**Example 3.2.6**

Here is a more concrete example.

**Example 3.2.7**

Let us compute  $\int_{-1}^1 (1 - |x|) dx$  again. In Example 3.1.15 we evaluated this integral by interpreting it as the area of a triangle. This time we are going to use *only* the properties given in Theorems 3.2.1 and 3.2.3 and the facts that

$$\int_a^b dx = b - a \quad \text{and} \quad \int_a^b x dx = \frac{b^2 - a^2}{2}$$

That  $\int_a^b dx = b - a$  is part (e) of Theorem 3.2.1. We saw that  $\int_a^b x dx = \frac{b^2 - a^2}{2}$  in Example 3.2.5.

First we are going to get rid of the absolute value signs by splitting the interval over which we integrate. Recalling that  $|x| = x$  whenever  $x \geq 0$  and  $|x| = -x$  whenever  $x \leq 0$ , we split the interval by Theorem 3.2.3(c),

$$\begin{aligned} \int_{-1}^1 (1 - |x|) dx &= \int_{-1}^0 (1 - |x|) dx + \int_0^1 (1 - |x|) dx \\ &= \int_{-1}^0 (1 - (-x)) dx + \int_0^1 (1 - x) dx \\ &= \int_{-1}^0 (1 + x) dx + \int_0^1 (1 - x) dx \end{aligned}$$

Now we apply parts (a) and (b) of Theorem 3.2.1, and then

$$\begin{aligned} \int_{-1}^1 [1 - |x|] dx &= \int_{-1}^0 1 dx + \int_{-1}^0 x dx + \int_0^1 1 dx - \int_0^1 x dx \\ &= [0 - (-1)] + \frac{0^2 - (-1)^2}{2} + [1 - 0] - \frac{1^2 - 0^2}{2} \\ &= 1 \end{aligned}$$

Example 3.2.7

### 3.2.1 ▶ More Properties of Integration: Even and Odd Functions

Recall<sup>18</sup> the following definition

**Definition 3.2.8.**

Let  $f(x)$  be a function. Then,

- we say that  $f(x)$  is even when  $f(x) = f(-x)$  for all  $x$ , and
- we say that  $f(x)$  is odd when  $f(x) = -f(-x)$  for all  $x$ .

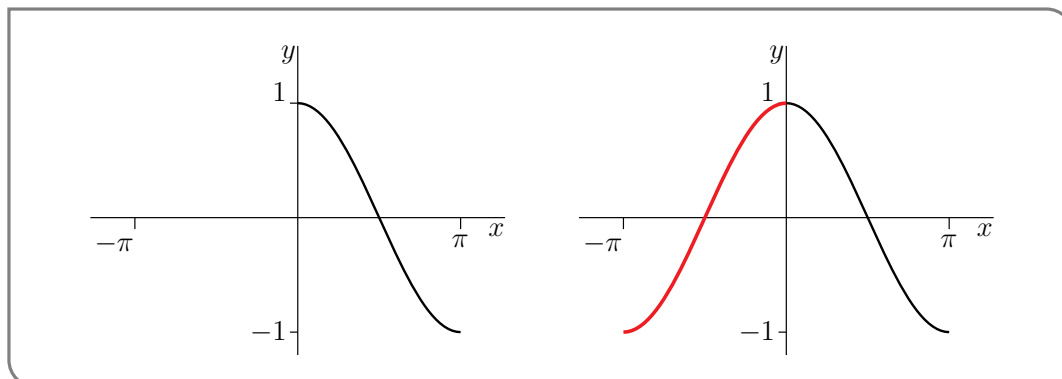
Of course most functions are neither even nor odd, but many of the standard functions you know are.

Example 3.2.9 (Even functions)

- Three examples of even functions are  $f(x) = |x|$ ,  $f(x) = \cos x$  and  $f(x) = x^2$ . In fact, if  $f(x)$  is any even power of  $x$ , then  $f(x)$  is an even function.

<sup>18</sup> We haven't done this in this course, but you should have seen it in your differential calculus course or perhaps even earlier.

- The part of the graph  $y = f(x)$  with  $x \leq 0$ , may be constructed by drawing the part of the graph with  $x \geq 0$  (as in the figure on the left below) and then reflecting it in the  $y$ -axis (as in the figure on the right below).



- In particular, if  $f(x)$  is an even function and  $a > 0$ , then the two sets

$$\{ (x, y) \mid 0 \leq x \leq a \text{ and } y \text{ is between } 0 \text{ and } f(x) \}$$

$$\{ (x, y) \mid -a \leq x \leq 0 \text{ and } y \text{ is between } 0 \text{ and } f(x) \}$$

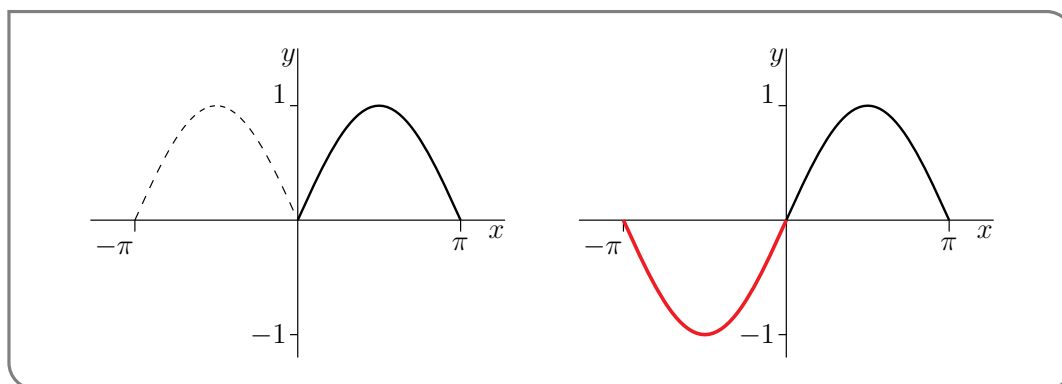
are reflections of each other in the  $y$ -axis and so have the same signed area. That is

$$\int_0^a f(x)dx = \int_{-a}^0 f(x)dx$$

Example 3.2.9

Example 3.2.10 (Odd functions)

- Three examples of odd functions are  $f(x) = \sin x$ ,  $f(x) = \tan x$  and  $f(x) = x^3$ . In fact, if  $f(x)$  is any odd power of  $x$ , then  $f(x)$  is an odd function.
- The part of the graph  $y = f(x)$  with  $x \leq 0$ , may be constructed by drawing the part of the graph with  $x \geq 0$  (like the solid line in the figure on the left below) and then reflecting it in the  $y$ -axis (like the dashed line in the figure on the left below) and then reflecting the result in the  $x$ -axis (i.e. flipping it upside down, like in the figure on the right, below).





- In particular, if  $f(x)$  is an odd function and  $a > 0$ , then the signed areas of the two sets

$$\begin{aligned} & \{ (x, y) \mid 0 \leq x \leq a \text{ and } y \text{ is between } 0 \text{ and } f(x) \} \\ & \{ (x, y) \mid -a \leq x \leq 0 \text{ and } y \text{ is between } 0 \text{ and } f(x) \} \end{aligned}$$

are negatives of each other — to get from the first set to the second set, you flip it upside down, in addition to reflecting it in the  $x$ -axis. That is

$$\int_0^a f(x)dx = - \int_{-a}^0 f(x)dx$$

Example 3.2.10

We can exploit the symmetries noted in the examples above, namely

$$\begin{aligned} \int_0^a f(x)dx &= \int_{-a}^0 f(x)dx && \text{for } f \text{ even} \\ \int_0^a f(x)dx &= - \int_{-a}^0 f(x)dx && \text{for } f \text{ odd} \end{aligned}$$

together with Theorem 3.2.3

$$\int_{-a}^a f(x)dx = \int_{-a}^0 f(x)dx + \int_0^a f(x)dx$$

in order to simplify the integration of even and odd functions over intervals of the form  $[-a, a]$ .

**Theorem 3.2.11 (Even and Odd).**

Let  $a > 0$ .

(a) If  $f(x)$  is an even function, then

$$\int_{-a}^a f(x)dx = 2 \int_0^a f(x)dx$$

(b) If  $f(x)$  is an odd function, then

$$\int_{-a}^a f(x)dx = 0$$

*Proof.* For any function

$$\int_{-a}^a f(x)dx = \int_0^a f(x)dx + \int_{-a}^0 f(x)dx$$

When  $f$  is even, the two terms on the right hand side are equal. When  $f$  is odd, the two terms on the right hand side are negatives of each other. □

### 3.2.2 ▶ More Properties of Integration: Inequalities for Integrals

We are still unable to integrate many functions, however with a little work we can infer bounds on integrals from bounds on their integrands.

#### Theorem 3.2.12 (Inequalities for Integrals).

Let  $a \leq b$  be real numbers and let the functions  $f(x)$  and  $g(x)$  be integrable on the interval  $a \leq x \leq b$ .

(a) If  $f(x) \geq 0$  for all  $a \leq x \leq b$ , then

$$\int_a^b f(x) dx \geq 0$$

(b) If there are constants  $m$  and  $M$  such that  $m \leq f(x) \leq M$  for all  $a \leq x \leq b$ , then

$$m(b-a) \leq \int_a^b f(x) dx \leq M(b-a)$$

(c) If  $f(x) \leq g(x)$  for all  $a \leq x \leq b$ , then

$$\int_a^b f(x) dx \leq \int_a^b g(x) dx$$

(d) We have

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f(x)| dx$$

*Proof.* (a) By interpreting the integral as the signed area, this statement simply says that if the curve  $y = f(x)$  lies above the  $x$ -axis and  $a \leq b$ , then the signed area of  $\{(x, y) \mid a \leq x \leq b, 0 \leq y \leq f(x)\}$  is at least zero. This is quite clear. Alternatively, we could argue more algebraically from Definition 3.1.8. We observe that when we define  $\int_a^b f(x) dx$  via Riemann sums, every summand,  $f(x_{i,n}^*) \frac{b-a}{n} \geq 0$ . Thus the whole sum is nonnegative and consequently, so is the limit, and thus so is the integral.

(b) We can argue this from (a) with a little massaging. Let  $g(x) = M - f(x)$ , then since  $f(x) \leq M$ , we have  $g(x) = M - f(x) \geq 0$  so that

$$\int_a^b (M - f(x)) dx = \int_a^b g(x) dx \geq 0.$$

but we also have

$$\begin{aligned}\int_a^b (M - f(x))dx &= \int_a^b Mdx - \int_a^b f(x)dx \\ &= M(b - a) - \int_a^b f(x)dx\end{aligned}$$

Thus

$$\begin{aligned}M(b - a) - \int_a^b f(x)dx &\geq 0 && \text{rearrange} \\ M(b - a) &\geq \int_a^b f(x)dx\end{aligned}$$

as required. The argument showing  $\int_a^b f(x)dx \geq m(b - a)$  is similar.

- (c) Now let  $h(x) = g(x) - f(x)$ . Since  $f(x) \leq g(x)$ , we have  $h(x) = g(x) - f(x) \geq 0$  so that

$$\int_a^b (g(x) - f(x))dx = \int_a^b h(x)dx \geq 0$$

But we also have that

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

Thus

$$\begin{aligned}\int_a^b g(x)dx - \int_a^b f(x)dx &\geq 0 && \text{rearrange} \\ \int_a^b g(x)dx &\geq \int_a^b f(x)dx\end{aligned}$$

as required.

- (d) For any  $x$ ,  $|f(x)|$  is either  $f(x)$  or  $-f(x)$  (depending on whether  $f(x)$  is positive or negative), so we certainly have

$$f(x) \leq |f(x)| \quad \text{and} \quad -f(x) \leq |f(x)|$$

Applying part (c) to each of those inequalities gives

$$\int_a^b f(x)dx \leq \int_a^b |f(x)|dx \quad \text{and} \quad -\int_a^b f(x)dx \leq \int_a^b |f(x)|dx$$

Now  $|\int_a^b f(x)dx|$  is either equal to  $\int_a^b f(x)dx$  or  $-\int_a^b f(x)dx$  (depending on whether the integral is positive or negative). In either case we can apply the above two inequalities to get the same result, namely

$$\left| \int_a^b f(x)dx \right| \leq \int_a^b |f(x)|dx.$$

□

Example 3.2.13  $\left(\int_0^{\pi/3} \sqrt{\cos x} dx\right)$

Consider the integral

$$\int_0^{\pi/3} \sqrt{\cos x} dx$$

This is not so easy to compute exactly<sup>19</sup>, but we can bound it quite quickly.

For  $x$  between 0 and  $\frac{\pi}{3}$ , the function  $\cos x$  takes values<sup>20</sup> between 1 and  $\frac{1}{2}$ . Thus the function  $\sqrt{\cos x}$  takes values between 1 and  $\frac{1}{\sqrt{2}}$ . That is

$$\frac{1}{\sqrt{2}} \leq \sqrt{\cos x} \leq 1 \quad \text{for } 0 \leq x \leq \frac{\pi}{3}.$$

Consequently, by Theorem 3.2.12(b) with  $a = 0$ ,  $b = \frac{\pi}{3}$ ,  $m = \frac{1}{\sqrt{2}}$  and  $M = 1$ ,

$$\frac{\pi}{3\sqrt{2}} \leq \int_0^{\pi/3} \sqrt{\cos x} dx \leq \frac{\pi}{3}$$

Plugging these expressions into a calculator gives us

$$0.7404804898 \leq \int_0^{\pi/3} \sqrt{\cos x} dx \leq 1.047197551$$

Example 3.2.13

### 3.3▲ The Fundamental Theorem of Calculus

We have spent quite a few pages (and lectures) talking about definite integrals, what they are (Definition 3.1.8), when they exist (Theorem 3.1.9), how to compute some special cases (Section 3.1.3), some ways to manipulate them (Theorem 3.2.1 and 3.2.3) and how to bound them (Theorem 3.2.12). Conspicuously missing from all of this has been a discussion of how to compute them in general. It is high time we rectified that.

The single most important tool used to evaluate integrals is called “the Fundamental Theorem of Calculus”. Its grand name is justified — it links the two branches of calculus by connecting derivatives to integrals. In so doing it also tells us how to compute integrals. Very roughly speaking the derivative of an integral is the original function. This fact allows us to compute integrals using antiderivatives<sup>21</sup>. Of course “very rough” is not enough — let’s be precise.

19 It is not too hard to use Riemann sums and a computer to evaluate it numerically: 0.948025319 . . .

20 You know the graphs of sine and cosine, so you should be able to work this out without too much difficulty.

21 You learned these near the end of your differential calculus course. Now is a good time to revise — but we’ll go over them here since they are so important in what follows.

**Theorem 3.3.1** (Fundamental Theorem of Calculus).

Let  $a < b$  and let  $f(x)$  be a function which is defined and continuous on  $[a, b]$ .

*Part 1:* Let  $F(x) = \int_a^x f(t)dt$  for any  $x$  in the interval  $[a, b]$ . Then the function  $F(x)$  is differentiable and further

$$F'(x) = f(x)$$

*Part 2:* Let  $G(x)$  be any function which is defined and continuous on  $[a, b]$ . Further let  $G(x)$  be differentiable with  $G'(x) = f(x)$  for all  $a < x < b$ . Then

$$\int_a^b f(x)dx = G(b) - G(a) \quad \text{or equivalently} \quad \int_a^b G'(x)dx = G(b) - G(a)$$

Before we prove this theorem and look at a bunch of examples of its application, it is important that we recall one definition from differential calculus — antiderivatives. If  $F'(x) = f(x)$  on some interval, then  $F(x)$  is called an antiderivative of  $f(x)$  on that interval. So Part 2 of the Fundamental Theorem of Calculus tells us how to evaluate the definite integral of  $f(x)$  in terms of any of its antiderivatives — if  $G(x)$  is any antiderivative of  $f(x)$  then

$$\int_a^b f(x)dx = G(b) - G(a)$$

The form  $\int_a^b G'(x)dx = G(b) - G(a)$  of the Fundamental Theorem relates the rate of change of  $G(x)$  over the interval  $a \leq x \leq b$  to the net change of  $G$  between  $x = a$  and  $x = b$ . For that reason, it is sometimes called the “net change theorem”.

We’ll start with a simple example. Then we’ll see why the Fundamental Theorem is true and then we’ll do many more, and more involved, examples.

**Example 3.3.2** (A first example)

Consider the integral  $\int_a^b xdx$  which we have explored previously in Example 3.2.5.

- The integrand is  $f(x) = x$ .
- We can readily verify that  $G(x) = \frac{x^2}{2}$  satisfies  $G'(x) = f(x)$  and so is an antiderivative of the integrand.
- Part 2 of Theorem 3.3.1 then tells us that

$$\int_a^b f(x)dx = G(b) - G(a)$$

$$\int_a^b xdx = \frac{b^2}{2} - \frac{a^2}{2}$$

which is precisely the result we obtained (with more work) in Example 3.2.5.

## Example 3.3.2

We do not give completely rigorous proofs of the two parts of the theorem — that is not really needed for this course. We just give the main ideas of the proofs so that you can understand why the theorem is true.

*Part 1.* We wish to show that if

$$F(x) = \int_a^x f(t) dt \quad \text{then} \quad F'(x) = f(x)$$

- Assume that  $F$  is the above integral and then consider  $F'(x)$ . By definition

$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}$$

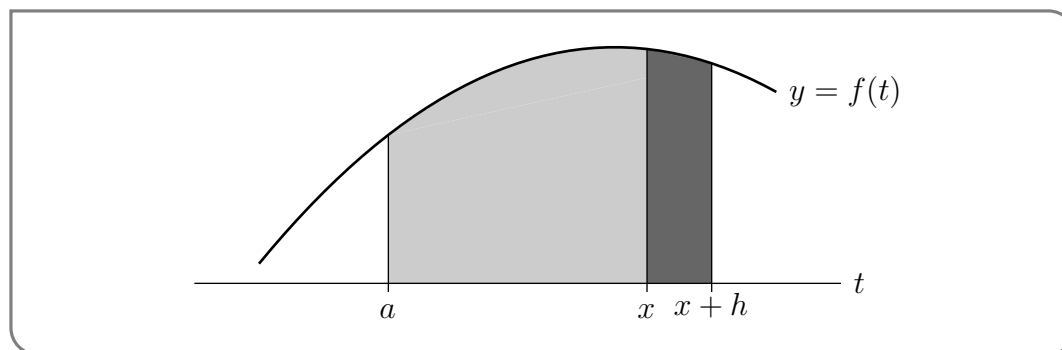
- To understand this limit, we interpret the terms  $F(x)$ ,  $F(x+h)$  as signed areas. To simplify this further, let's only consider the case that  $f$  is always nonnegative and that  $h > 0$ . These restrictions are not hard to remove, but the proof ideas are a bit cleaner if we keep them in place. Then we have

$$\begin{aligned} F(x+h) &= \text{the area of the region } \{ (t, y) \mid a \leq t \leq x+h, 0 \leq y \leq f(t) \} \\ F(x) &= \text{the area of the region } \{ (t, y) \mid a \leq t \leq x, 0 \leq y \leq f(t) \} \end{aligned}$$

- Then the numerator

$$F(x+h) - F(x) = \text{the area of the region } \{ (t, y) \mid x \leq t \leq x+h, 0 \leq y \leq f(t) \}$$

This is just the more darkly shaded region in the figure



- We will be taking the limit  $h \rightarrow 0$ . So suppose that  $h$  is very small. Then, as  $t$  runs from  $x$  to  $x+h$ ,  $f(t)$  runs only over a very narrow range of values<sup>22</sup>, all close to  $f(x)$ .
- So the darkly shaded region is almost a rectangle of width  $h$  and height  $f(x)$  and so has an area which is very close to  $f(x)h$ . Thus  $\frac{F(x+h) - F(x)}{h}$  is very close to  $f(x)$ .

<sup>22</sup> Notice that if  $f$  were discontinuous, then this might be false.

- In the limit  $h \rightarrow 0$ ,  $\frac{F(x+h)-F(x)}{h}$  becomes exactly  $f(x)$ , which is precisely what we want.

□

We can make the above more rigorous using the Mean Value Theorem<sup>23</sup>.

*Part 2.* We want to show that  $\int_a^b f(t)dt = G(b) - G(a)$ . To do this we exploit the fact that the derivative of a constant is zero.

- Let

$$H(x) = \int_a^x f(t)dt - G(x) + G(a)$$

Then the result we wish to prove is that  $H(b) = 0$ . We will do this by showing that  $H(x) = 0$  for all  $x$  between  $a$  and  $b$ .

- We first show that  $H(x)$  is constant by computing its derivative:

$$H'(x) = \frac{d}{dx} \int_a^x f(t)dt - \frac{d}{dx} (G(x)) + \frac{d}{dx} (G(a))$$

Since  $G(a)$  is a constant, its derivative is 0 and by assumption the derivative of  $G(x)$  is just  $f(x)$ , so

$$= \frac{d}{dx} \int_a^x f(t)dt - f(x)$$

Now Part 1 of the theorem tells us that this derivative is just  $f(x)$ , so

$$= f(x) - f(x) = 0$$

Hence  $H$  is constant.

- To determine which constant we just compute  $H(a)$ :

$$\begin{aligned} H(a) &= \int_a^a f(t)dt - G(a) + G(a) \\ &= \int_a^a f(t)dt && \text{by Theorem 3.2.3(a)} \\ &= 0 \end{aligned}$$

as required.

---

23 The MVT tells us that there is a number  $c$  between  $x$  and  $x + h$  so that

$$F'(c) = \frac{F(x+h) - F(x)}{(x+h) - x} = \frac{F(x+h) - F(x)}{h}$$

But since  $F'(x) = f(x)$ , this tells us that

$$\frac{F(x+h) - F(x)}{h} = f(c)$$

where  $c$  is trapped between  $x + h$  and  $x$ . Now when we take the limit as  $h \rightarrow 0$  we have that this number  $c$  is squeezed to  $x$  and the result follows.

□

The simple example we did above (Example 3.3.2), demonstrates the application of part 2 of the Fundamental Theorem of Calculus. Before we do more examples (and there will be many more over the coming sections) we should do some examples illustrating the use of part 1 of the fundamental theorem of calculus. Then we'll move on to part 2.

Example 3.3.3  $\left(\frac{d}{dx} \int_0^x t dt\right)$

Consider the integral  $\int_0^x t dt$ . We know how to evaluate this — it is just Example 3.3.2 with  $a = 0, b = x$ . So we have two ways to compute the derivative. We can evaluate the integral and then take the derivative, or we can apply Part 1 of the Fundamental Theorem. We'll do both, and check that the two answers are the same.

First, Example 3.3.2 gives

$$F(x) = \int_0^x t dt = \frac{x^2}{2}$$

So of course  $F'(x) = x$ . Second, Part 1 of the Fundamental Theorem of calculus tells us that the derivative of  $F(x)$  is just the integrand. That is, Part 1 of the Fundamental Theorem of Calculus also gives  $F'(x) = x$ .

Example 3.3.3

In the previous example we were able to evaluate the integral explicitly, so we did not need the Fundamental Theorem to determine its derivative. Here is an example that really does require the use of the Fundamental Theorem.

Example 3.3.4  $\left(\frac{d}{dx} \int_0^x e^{-t^2} dt\right)$

We would like to find  $\frac{d}{dx} \int_0^x e^{-t^2} dt$ . In the previous example, we were able to compute the corresponding derivative in two ways — we could explicitly compute the integral and then differentiate the result, or we could apply part 1 of the Fundamental Theorem of calculus. In this example we do not know the integral explicitly. Indeed it is not possible to express<sup>24</sup> the integral  $\int_0^x e^{-t^2} dt$  as a finite combination of standard functions such as polynomials, exponentials, trigonometric functions and so on.

Despite this, we can find its derivative by just applying the first part of the Fundamen-

24 The integral  $\int_0^x e^{-t^2} dt$  is closely related to the “error function” which is an extremely important function in mathematics. While we cannot express this integral (or the error function) as a *finite* combination of polynomials, exponentials etc, we can express it as an infinite series

$$\int_0^x e^{-t^2} dt = x - \frac{x^3}{3 \cdot 1} + \frac{x^5}{5 \cdot 2} - \frac{x^7}{7 \cdot 3!} + \frac{x^9}{9 \cdot 4!} + \dots + (-1)^k \frac{x^{2k+1}}{(2k+1) \cdot k!} + \dots$$

But more on this in Chapter 5.



tal Theorem of Calculus with  $f(t) = e^{-t^2}$  and  $a = 0$ . That gives

$$\begin{aligned} \frac{d}{dx} \int_0^x e^{-t^2} dt &= \frac{d}{dx} \int_0^x f(t) dt \\ &= f(x) = e^{-x^2} \end{aligned}$$

Example 3.3.4

Let us ratchet up the complexity of the previous example — we can make the limits of the integral more complicated functions. So consider the previous example with the upper limit  $x$  replaced by  $x^2$ :

Example 3.3.5  $\left(\frac{d}{dx} \int_0^{x^2} e^{-t^2} dt\right)$

Consider the integral  $\int_0^{x^2} e^{-t^2} dt$ . We would like to compute its derivative with respect to  $x$  using part 1 of the fundamental theorem of calculus.

The Fundamental Theorem tells us how to compute the derivative of functions of the form  $\int_a^x f(t) dt$  but the integral at hand is *not* of the specified form because the upper limit we have is  $x^2$ , rather than  $x$ , — so more care is required. Thankfully we can deal with this obstacle with only a little extra work. The trick is to define an auxiliary function by simply changing the upper limit to  $x$ . That is, define

$$E(x) = \int_0^x e^{-t^2} dt$$

Then the integral we want to work with is

$$E(x^2) = \int_0^{x^2} e^{-t^2} dt$$

The derivative  $E'(x)$  can be found via part 1 of the Fundamental Theorem of calculus (as we did in Example 3.3.4) and is  $E'(x) = e^{-x^2}$ . We can then use this fact with the chain rule to compute the derivative we need:

$$\begin{aligned} \frac{d}{dx} \int_0^{x^2} e^{-t^2} dt &= \frac{d}{dx} E(x^2) && \text{use the chain rule} \\ &= 2xE'(x^2) \\ &= 2xe^{-x^4} \end{aligned}$$

Example 3.3.5

What if both limits of integration are functions of  $x$ ? We can still make this work, but we have to split the integral using Theorem 3.2.3.

Example 3.3.6  $\left(\frac{d}{dx} \int_x^{x^2} e^{-t^2} dt\right)$

Consider the integral

$$\int_x^{x^2} e^{-t^2} dt$$

As was the case in the previous example, we have to do a little pre-processing before we can apply the Fundamental Theorem.

This time (by design), not only is the upper limit of integration  $x^2$  rather than  $x$ , but the lower limit of integration also depends on  $x$  — this is different from the integral  $\int_a^x f(t) dt$  in the Fundamental Theorem where the *lower* limit of integration is a constant.

Fortunately we can use the basic properties of integrals (Theorem 3.2.3(b) and (c)) to split  $\int_x^{x^2} e^{-t^2} dt$  into pieces whose derivatives we already know.

$$\begin{aligned} \int_x^{x^2} e^{-t^2} dt &= \int_x^0 e^{-t^2} dt + \int_0^{x^2} e^{-t^2} dt && \text{by Theorem 3.2.3(c)} \\ &= -\int_0^x e^{-t^2} dt + \int_0^{x^2} e^{-t^2} dt && \text{by Theorem 3.2.3(b)} \end{aligned}$$

With this pre-processing, both integrals are of the right form. Using what we have learned in the previous two examples,

$$\begin{aligned} \frac{d}{dx} \int_x^{x^2} e^{-t^2} dt &= \frac{d}{dx} \left( -\int_0^x e^{-t^2} dt + \int_0^{x^2} e^{-t^2} dt \right) \\ &= -\frac{d}{dx} \int_0^x e^{-t^2} dt + \frac{d}{dx} \int_0^{x^2} e^{-t^2} dt \\ &= -e^{-x^2} + 2xe^{-x^4} \end{aligned}$$

Example 3.3.6

### 3.3.1 ► Indefinite Integration

Before we start to work with part 2 of the Fundamental Theorem, we need a little terminology and notation. First some terminology — you may have seen this definition in your differential calculus course.

#### Definition 3.3.7 (Antiderivatives).

Let  $f(x)$  and  $F(x)$  be functions. If  $F'(x) = f(x)$  on an interval, then we say that  $F(x)$  is an antiderivative of  $f(x)$  on that interval.

As we saw above, an antiderivative of  $f(x) = x$  is  $F(x) = x^2/2$  — we can easily verify this by differentiation. Notice that  $x^2/2 + 3$  is also an antiderivative of  $x$ , as is  $x^2/2 + C$  for any constant  $C$ . This observation gives us the following simple lemma.

**Lemma 3.3.8.**

Let  $f(x)$  be a function and let  $F(x)$  be an antiderivative of  $f(x)$ . Then  $F(x) + C$  is also an antiderivative for any constant  $C$ . Further, every antiderivative of  $f(x)$  must be of this form.

*Proof.* There are two parts to the lemma and we prove each in turn.

- Let  $F(x)$  be an antiderivative of  $f(x)$  and let  $C$  be some constant. Then

$$\begin{aligned}\frac{d}{dx}(F(x) + C) &= \frac{d}{dx}(F(x)) + \frac{d}{dx}(C) \\ &= f(x) + 0\end{aligned}$$

since the derivative of a constant is zero, and by definition the derivative of  $F(x)$  is just  $f(x)$ . Thus  $F(x) + C$  is also an antiderivative of  $f(x)$ .

- Now let  $F(x)$  and  $G(x)$  both be antiderivatives of  $f(x)$  — we will show that  $G(x) = F(x) + C$  for some constant  $C$ . To do this let  $H(x) = G(x) - F(x)$ . Then

$$\frac{d}{dx}H(x) = \frac{d}{dx}(G(x) - F(x)) = \frac{d}{dx}G(x) - \frac{d}{dx}F(x) = f(x) - f(x) = 0$$

Since the derivative of  $H(x)$  is zero,  $H(x)$  must be a constant function<sup>25</sup>. Thus  $H(x) = G(x) - F(x) = C$  for some constant  $C$  and the result follows.

□

Based on the above lemma we have the following definition.

<sup>25</sup> This follows from the Mean Value Theorem. Say  $H(x)$  were not constant, then there would be two numbers  $a < b$  so that  $H(a) \neq H(b)$ . Then the MVT tells us that there is a number  $c$  between  $a$  and  $b$  so that

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

Since both numerator and denominator are non-zero, we know the derivative at  $c$  is nonzero. But this would contradict the assumption that derivative of  $H$  is zero. Hence we cannot have  $a < b$  with  $H(a) \neq H(b)$  and so  $H(x)$  must be constant.

**Definition 3.3.9.**

The “indefinite integral of  $f(x)$ ” is denoted by  $\int f(x)dx$  and should be regarded as the general antiderivative of  $f(x)$ . In particular, if  $F(x)$  is an antiderivative of  $f(x)$  then

$$\int f(x)dx = F(x) + C$$

where the  $C$  is an arbitrary constant. In this context, the constant  $C$  is also often called a “constant of integration”.

Now we just need a tiny bit more notation.

**Notation 3.3.10.**

The symbol

$$\int f(x)dx \Big|_a^b$$

denotes the change in an antiderivative of  $f(x)$  from  $x = a$  to  $x = b$ . More precisely, let  $F(x)$  be any antiderivative of  $f(x)$ . Then

$$\int f(x)dx \Big|_a^b = F(x) \Big|_a^b = F(b) - F(a)$$

Notice that this notation allows us to write part 2 of the Fundamental Theorem as

$$\begin{aligned} \int_a^b f(x)dx &= \int f(x)dx \Big|_a^b \\ &= F(x) \Big|_a^b = F(b) - F(a) \end{aligned}$$

Some texts also use an equivalent notation using square brackets:

$$\int_a^b f(x)dx = [F(x)]_a^b = F(b) - F(a).$$

You should be familiar with both notations.

We’ll soon develop some strategies for computing more complicated integrals. But for now, we’ll try a few integrals that are simple enough that we can just guess the answer. Of course, any antiderivative that we can guess we can also check — simply differentiate the guess and verify you get back to the original function:

$$\frac{d}{dx} \int f(x)dx = f(x).$$

We do these examples in some detail to help us become comfortable finding indefinite integrals.

**Example 3.3.11**

Compute the definite integral  $\int_1^2 x dx$ .

*Solution.* We have already seen, in Example 3.2.5, that  $\int_1^2 x dx = \frac{2^2-1^2}{2} = \frac{3}{2}$ . We shall now rederive that result using the Fundamental Theorem of Calculus.

- The main difficulty in this approach is finding the indefinite integral (an antiderivative) of  $x$ . That is, we need to find a function  $F(x)$  whose derivative is  $x$ . So think back to all the derivatives you computed last term<sup>26</sup> and try to remember a function whose derivative was something like  $x$ .
- This shouldn't be too hard — we recall that the derivatives of polynomials are polynomials. More precisely, we know that

$$\frac{d}{dx}x^n = nx^{n-1}$$

So if we want to end up with just  $x = x^1$ , we need to take  $n = 2$ . However this gives us

$$\frac{d}{dx}x^2 = 2x$$

- This is pretty close to what we want except for the factor of 2. Since this is a constant we can just divide both sides by 2 to obtain:

$$\begin{aligned} \frac{1}{2} \cdot \frac{d}{dx}x^2 &= \frac{1}{2} \cdot 2x && \text{which becomes} \\ \frac{d}{dx} \frac{x^2}{2} &= x \end{aligned}$$

which is exactly what we need. It tells us that  $x^2/2$  is an antiderivative of  $x$ .

- Once one has an antiderivative, it is easy to compute the indefinite integral

$$\int x dx = \frac{1}{2}x^2 + C$$

as well as the definite integral:

$$\begin{aligned} \int_1^2 x dx &= \left. \frac{1}{2}x^2 \right|_1^2 && \text{since } x^2/2 \text{ is the antiderivative of } x \\ &= \frac{1}{2}2^2 - \frac{1}{2}1^2 = \frac{3}{2} \end{aligned}$$

<sup>26</sup> Of course, this assumes that you did your differential calculus course last term. If you did that course at a different time then please think back to that point in time. If it is long enough ago that you don't quite remember when it was, then you should probably do some revision of derivatives of simple functions before proceeding further.

## Example 3.3.11

While the previous example could be computed using signed areas, the following example would be very difficult to compute without using the Fundamental Theorem of Calculus.

## Example 3.3.12

Compute  $\int_0^{\pi/2} \sin x dx$ .

*Solution.*

- Once again, the crux of the solution is guessing the antiderivative of  $\sin x$  — that is finding a function whose derivative is  $\sin x$ .
- The standard derivative that comes closest to  $\sin x$  is

$$\frac{d}{dx} \cos x = -\sin x$$

which is the derivative we want, multiplied by a factor of  $-1$ .

- Just as we did in the previous example, we multiply this equation by a constant to remove this unwanted factor:

$$\begin{aligned} (-1) \cdot \frac{d}{dx} \cos x &= (-1) \cdot (-\sin x) && \text{giving us} \\ \frac{d}{dx} (-\cos x) &= \sin x \end{aligned}$$

This tells us that  $-\cos x$  is an antiderivative of  $\sin x$ .

- Now it is straightforward to compute the integral:

$$\begin{aligned} \int_0^{\pi/2} \sin x dx &= -\cos x \Big|_0^{\pi/2} && \text{since } -\cos x \text{ is the antiderivative of } \sin x \\ &= -\cos \frac{\pi}{2} + \cos 0 \\ &= 0 + 1 = 1 \end{aligned}$$

## Example 3.3.12

## Example 3.3.13

Find  $\int_1^2 \frac{1}{x} dx$ .

*Solution.*

- Once again, the crux of the solution is guessing a function whose derivative is  $\frac{1}{x}$ . Our standard way to differentiate powers of  $x$ , namely

$$\frac{d}{dx} x^n = nx^{n-1},$$

doesn't work in this case — since it would require us to pick  $n = 0$  and this would give

$$\frac{d}{dx}x^0 = \frac{d}{dx}1 = 0.$$

- Fortunately, we also know<sup>27</sup> that

$$\frac{d}{dx} \ln x = \frac{1}{x}$$

which is exactly the derivative we want.

- We're now ready to compute the prescribed integral.

$$\begin{aligned} \int_1^2 \frac{1}{x} dx &= \ln x \Big|_1^2 && \text{since } \ln x \text{ is an antiderivative of } 1/x \\ &= \ln 2 - \ln 1 && \text{since } \ln 1 = 0 \\ &= \ln 2 \end{aligned}$$

Example 3.3.13

Example 3.3.14

Find  $\int_{-2}^{-1} \frac{1}{x} dx$ .

*Solution.*

- As we saw in the last example,

$$\frac{d}{dx} \ln x = \frac{1}{x}$$

and if we naively use this here, then we will obtain

$$\int_{-2}^{-1} \frac{1}{x} dx = \ln(-1) - \ln(-2)$$

which makes no sense since the logarithm is only defined for positive numbers<sup>28</sup>.

- We can work around this problem using a slight variation of the logarithm —  $\ln |x|$ .

<sup>27</sup> To align with what you probably saw in high school, we'll use  $\ln x$  to denote the natural logarithm. This is unambiguous —  $\ln x$  is always the same as  $\log_e x$ . On the other hand, the precise meaning of  $\log x$  is not universal. The implied base may be 10 (common in chemistry and physics),  $e$  (common in math and computer languages like Java, C, Python, and MATLAB), or 2 (common in computer science).

<sup>28</sup> This is not entirely true — one can extend the definition of the logarithm to negative numbers, but to do so one needs to understand complex numbers which is a topic beyond the scope of this course.

– When  $x > 0$ , we know that  $|x| = x$  and so we have

$$\begin{aligned} \ln|x| &= \ln x && \text{differentiating gives us} \\ \frac{d}{dx} \ln|x| &= \frac{d}{dx} \ln x = \frac{1}{x}. \end{aligned}$$

– When  $x < 0$  we have that  $|x| = -x$  and so

$$\begin{aligned} \ln|x| &= \ln(-x) && \text{differentiating with the chain rule gives} \\ \frac{d}{dx} \ln|x| &= \frac{d}{dx} \ln(-x) \\ &= \frac{1}{(-x)} \cdot (-1) = \frac{1}{x} \end{aligned}$$

– Indeed, more generally we should write the indefinite integral of  $1/x$  as

$$\int \frac{1}{x} dx = \ln|x| + C$$

which is valid for all positive and negative  $x$ . It is, however, undefined at  $x = 0$ .

• We’re now ready to compute the prescribed integral.

$$\begin{aligned} \int_{-2}^{-1} \frac{1}{x} dx &= \ln|x| \Big|_{-2}^{-1} && \text{since } \ln|x| \text{ is an antiderivative of } 1/x \\ &= \ln|-1| - \ln|-2| = \ln 1 - \ln 2 \\ &= -\ln 2 = \ln^{1/2}. \end{aligned}$$

Example 3.3.14

This next example raises a nasty issue that requires a little care. We know that the function  $1/x$  is not defined at  $x = 0$  — so can we integrate over an interval that contains  $x = 0$  and still obtain an answer that makes sense? More generally can we integrate a function over an interval on which that function has discontinuities?

Example 3.3.15

Find  $\int_{-1}^1 \frac{1}{x^2} dx$ .

*Solution.* Beware that this is a particularly nasty example, which illustrates a booby trap hidden in the Fundamental Theorem of Calculus. The booby trap explodes when the theorem is applied sloppily.

• The sloppy solution starts, as our previous examples have, by finding an antiderivative of the integrand. In this case we know that

$$\frac{d}{dx} \frac{1}{x} = -\frac{1}{x^2}$$

which means that  $-x^{-1}$  is an antiderivative of  $x^{-2}$ .



- This suggests (if we proceed naively) that

$$\begin{aligned} \int_{-1}^1 x^{-2} dx &= -\frac{1}{x} \Big|_{-1}^1 && \text{since } -1/x \text{ is an antiderivative of } 1/x^2 \\ &= -\frac{1}{1} - \left(-\frac{1}{-1}\right) \\ &= -2 \end{aligned}$$

Unfortunately,

- At this point we should really start to be concerned. This answer cannot be correct. Our integrand, being a square, is positive everywhere. So our integral represents the area of a region above the  $x$ -axis and must be positive.
- So what has gone wrong? The flaw in the computation is that the Fundamental Theorem of calculus, which says that

$$\text{if } F'(x) = f(x) \text{ then } \int_a^b f(x) dx = F(b) - F(a),$$

is *only* applicable when  $F'(x)$  exists and equals  $f(x)$  for *all*  $x$  between  $a$  and  $b$ .

- In this case  $F'(x) = \frac{1}{x^2}$  does not exist for  $x = 0$ . So we cannot apply the Fundamental Theorem of Calculus as we tried to above.

An integral, like  $\int_{-1}^1 \frac{1}{x^2} dx$ , whose integrand is undefined somewhere in the domain of integration is called improper. We'll give a more thorough treatment of improper integrals later in the text. For now, we'll just say that the correct way to define (and evaluate) improper integrals is as a limit of well-defined approximating integrals. We shall later see that, not only is  $\int_{-1}^1 \frac{1}{x^2} dx$  not negative, it is infinite.

Example 3.3.15

The above examples have illustrated how we can use the fundamental theorem of calculus to convert knowledge of derivatives into knowledge of integrals. We are now in a position to easily build a table of integrals. Here is a short table of the most important derivatives that we know.

$F(x)$	1	$x^n$	$\sin x$	$\cos x$	$\tan x$	$e^x$	$\ln  x $	$\arcsin x$	$\arctan x$
$f(x) = F'(x)$	0	$nx^{n-1}$	$\cos x$	$-\sin x$	$\sec^2 x$	$e^x$	$\frac{1}{x}$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{1+x^2}$

Of course we know other derivatives, such as those of  $\sec x$  and  $\cot x$ , however the ones listed above are arguably the most important ones. From this table (with a very little massaging) we can write down a short table of indefinite integrals.

**Theorem 3.3.16** (Important indefinite integrals).

$f(x)$	$F(x) = \int f(x)dx$
1	$x + C$
$x^n$	$\frac{1}{n+1}x^{n+1} + C$ provided that $n \neq -1$
$\frac{1}{x}$	$\ln x  + C$
$e^x$	$e^x + C$
$\sin x$	$-\cos x + C$
$\cos x$	$\sin x + C$
$\sec^2 x$	$\tan x + C$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin x + C$
$\frac{1}{1+x^2}$	$\arctan x + C$

**Example 3.3.17**

Find the following integrals

(i)  $\int_2^7 e^x dx$

(ii)  $\int_{-2}^2 \frac{1}{1+x^2} dx$

(iii)  $\int_0^3 (2x^3 + 7x - 2) dx$

*Solution.* We can proceed with each of these as before — find the antiderivative and then apply the Fundamental Theorem. The third integral is a little more complicated, but we can split it up into monomials using Theorem 3.2.1 and do each separately.

(i) An antiderivative of  $e^x$  is just  $e^x$ , so

$$\begin{aligned}\int_2^7 e^x dx &= e^x \Big|_2^7 \\ &= e^7 - e^2 = e^2(e^5 - 1).\end{aligned}$$

(ii) An antiderivative of  $\frac{1}{1+x^2}$  is  $\arctan(x)$ , so

$$\begin{aligned}\int_{-2}^2 \frac{1}{1+x^2} dx &= \arctan(x) \Big|_{-2}^2 \\ &= \arctan(2) - \arctan(-2)\end{aligned}$$

We can simplify this a little further by noting that  $\arctan(x)$  is an odd function, so  $\arctan(-2) = -\arctan(2)$  and thus our integral is

$$= 2 \arctan(2)$$

(iii) We can proceed by splitting the integral using Theorem 3.2.1(d)

$$\begin{aligned}\int_0^3 (2x^3 + 7x - 2) dx &= \int_0^3 2x^3 dx + \int_0^3 7x dx - \int_0^3 2 dx \\ &= 2 \int_0^3 x^3 dx + 7 \int_0^3 x dx - 2 \int_0^3 dx\end{aligned}$$

and because we know that  $x^4/4, x^2/2, x$  are antiderivatives of  $x^3, x, 1$  respectively, this becomes

$$\begin{aligned}&= \left[ \frac{x^4}{2} \right]_0^3 + \left[ \frac{7x^2}{2} \right]_0^3 - [2x]_0^3 \\ &= \frac{81}{2} + \frac{7 \cdot 9}{2} - 6 \\ &= \frac{81 + 63 - 12}{2} = \frac{132}{2} = 66.\end{aligned}$$

We can also just find the antiderivative of the whole polynomial by finding the antiderivatives of each term of the polynomial and then recombining them. This is equivalent to what we have done above, but perhaps a little neater:

$$\begin{aligned}\int_0^3 (2x^3 + 7x - 2) dx &= \left[ \frac{x^4}{2} + \frac{7x^2}{2} - 2x \right]_0^3 \\ &= \frac{81}{2} + \frac{7 \cdot 9}{2} - 6 = 66.\end{aligned}$$

Example 3.3.17

### 3.3.2 ▶ Marginal Cost and Marginal Revenue

#### Definition 3.3.18.

The **total cost** function,  $TC(q)$ , is the cost of producing  $q$  of units of a good.

- We call  $TC(0)$  (the cost incurred for producing  $q = 0$  units) the **fixed cost**,  $FC$ .
- The quantity  $TC(q) - TC(0)$  is the **variable cost**, which we call  $VC(q)$ .

Total cost is, therefore, the sum of fixed and variable costs:

$$TC(q) = FC + VC(q)$$

*Fixed cost* encompasses all expenses that do not change with quantity (such as rent on a factory space, which is the same whether you make 1 or 1000 units). Fixed cost is a constant, and generally nonzero. We can think of these expenses as the cost of setting up a business, incurred before the first unit is ever produced, hence the definition of fixed costs as  $TC(0)$ .

*Variable cost* consists of expenses that depend on the quantity produced. A typical example of such an expense is raw materials: producing more units means using more raw materials. Note that the variable cost varies with the quantity  $q$  produced, while the fixed cost is independent of  $q$ .

Consider the cost of making “one more unit” of output, after having already made  $q$  units:  $TC(q + 1) - TC(q)$ . Using the definition of the derivative, we can approximate this quantity by  $\frac{dTC}{dq}$ :

$$\frac{dTC}{dq} = \lim_{h \rightarrow 0} \frac{TC(q + h) - TC(q)}{h} \approx \frac{TC(q + 1) - TC(q)}{1}$$

This motivates the definition of a *marginal cost*.

#### Definition 3.3.19.

Let  $TC(q)$  be the total cost of producing  $q$  units of output of a particular good. The **marginal cost** of producing the good is defined as

$$MC(q) = \frac{d}{dq} [TC(q)].$$

The marginal cost is generally interpreted as the change in cost due to producing *one*

*additional unit.* This interpretation follows from the definition of the derivative:

$$\begin{aligned} MC(q) &= \frac{d}{dq} [TC(q)] \\ &= \lim_{h \rightarrow 0} \frac{TC(q+h) - TC(q)}{h} \\ &\approx TC(q+1) - TC(q) \end{aligned}$$

where we set  $h = 1$  for our approximation.

Suppose we know the marginal cost function,  $MC(q)$ , and we want to find the total cost function,  $TC(q)$ . By the Fundamental Theorem of Calculus,

$$TC(q) = \int MC(q) dq + C$$

for some constant  $C$ . In order to find  $C$ , we use the initial value

$$TC(0) = FC.$$

Example 3.3.20 (From marginal cost to total cost)

Suppose a product has fixed cost of \$100, and its marginal cost function is  $MC(q) = e^{-q} + 3$ . What is its total cost function?

*Solution.* Using the Fundamental Theorem of Calculus Part 1, given the definition  $MC(q) = \frac{d}{dq} [TC]$ , we see:

$$TC(q) = \int MC dq + C = \int (e^{-q} + 3) dq + C$$

Antidifferentiating by inspection,

$$= -e^{-q} + 3q + C$$

Using  $FC=T(0)$ :

$$\begin{aligned} 100 &= T(0) = -e^{-0} + 3 \cdot 0 + C = -1 + C \\ 101 &= C \end{aligned}$$

All together,

$$TC(q) = -e^{-q} + 3q + 101$$

Example 3.3.20

In addition to considering total and marginal costs, we can consider total and marginal revenue.

**Definition 3.3.21.**

Suppose the total revenue collected from  $q$  units of output is given by the function  $TR(q)$ , with  $TR(0) = 0$  (since selling no products leads to no revenue). We define the **marginal revenue** to be

$$MR(q) = \frac{d}{dq} [TR(q)].$$

We define the **unit price** to be

$$P(q) = \frac{TR(q)}{q}$$

for  $q > 0$ .

We think of marginal revenue as the extra revenue gained by producing *one extra unit* of output. As with the interpretation of marginal cost, this follows from the definition of the derivative, approximated with  $h = 1$ .

For unit price, we assume that each unit is sold for the same amount, but that amount is determined by the number of units of output  $q$ . So if 100 units are sold, each unit is priced at  $P(100)$ ; but a larger production of 1000 units would lead to each unit being priced at  $P(1000)$ .

**Example 3.3.22**

Suppose the marginal revenue function for a product is  $MR(q) = 10 - \frac{3}{1+q^2}$ . Let  $P$  be the price at which each unit is sold, and suppose 10 units are sold. Find  $P$ .

*Solution.* First, we use the Fundamental Theorem of Calculus Part 1 to find the total revenue function.

$$TR = \int MR dq + C = \int \left( 10 - \frac{3}{1+q^2} \right) dq + C$$

Referring to Theorem 3.3.16,

$$= 10q - 3 \arctan q + C$$

Now we use the initial value  $TR(0) = 0$ .

$$\begin{aligned} 0 &= TR(0) = 10(0) - 3 \arctan 0 + C \\ 0 &= C \end{aligned}$$

All together,

$$TR(q) = 10q - 3 \arctan q$$

If 10 units are sold, the unit price is

$$P(10) = \frac{TR(10)}{10} = \frac{10(10) - 3 \arctan(10)}{10} = 10 - 0.3 \arctan(10) \approx 10.44$$

### 3.4▲ Substitution

In the previous section we explored the Fundamental Theorem of Calculus and the link it provides between definite integrals and antiderivatives. Indeed, integrals with simple integrands are usually evaluated via this link. In this section we start to explore methods for integrating more complicated integrals. We have already seen — via Theorem 3.2.1 — that integrals interact very nicely with addition, subtraction and multiplication by constants:

$$\int_a^b (Af(x) + Bg(x)) dx = A \int_a^b f(x) dx + B \int_a^b g(x) dx$$

for  $A, B$  constants. By combining this with the list of indefinite integrals in Theorem 3.3.16, we can compute integrals of linear combinations of simple functions. For example

$$\begin{aligned} \int_1^4 (e^x - 2 \sin x + 3x^2) dx &= \int_1^4 e^x dx - 2 \int_1^4 \sin x dx + 3 \int_1^4 x^2 dx \\ &= \left( e^x + (-2) \cdot (-\cos x) + 3 \frac{x^3}{3} \right) \Big|_1^4 \quad \text{and so on} \end{aligned}$$

Of course there are a great many functions that can be approached in this way, however there are some very simple examples that cannot.

$$\int \sin(\pi x) dx \qquad \int x e^x dx \qquad \int \frac{x}{x^2 - 5x + 6} dx$$

In each case the integrands are not linear combinations of simpler functions; in order to compute them we need to understand how integrals (and antiderivatives) interact with compositions, products and quotients. We reached a very similar point in our differential calculus course where we understood the linearity of the derivative,

$$\frac{d}{dx} (Af(x) + Bg(x)) = A \frac{df}{dx} + B \frac{dg}{dx},$$

but had not yet seen the chain, product and quotient rules<sup>29</sup>. While we will develop tools to find the second and third integrals in later sections, we should really start with how to integrate compositions of functions.

It is important to state up front, that in general one cannot write down the integral of the composition of two functions — even if those functions are simple. This is not because the integral does not exist. Rather it is because the integral cannot be written down as a finite combination of the standard functions we know. A very good example of this,

<sup>29</sup> If your memory of these rules is a little hazy then you really should go back and revise them before proceeding. You will definitely need a good grasp of the chain rule for what follows in this section.

which we encountered in Example 3.3.4, is the composition of  $e^x$  and  $-x^2$ . Even though we know

$$\int e^x dx = e^x + C \quad \text{and} \quad \int -x^2 dx = -\frac{1}{3}x^3 + C$$

there is no simple function that is equal to the indefinite integral

$$\int e^{-x^2} dx.$$

even though the indefinite integral exists. In this way integration is very different from differentiation.

With that caveat out of the way, we can introduce the substitution rule. The substitution rule is obtained by antidifferentiating the chain rule. In some sense it is the chain rule in reverse. For completeness, let us restate the chain rule:

**Theorem 3.4.1** (The chain rule).

Let  $F(u)$  and  $u(x)$  be differentiable functions and form their composition  $F(u(x))$ . Then

$$\frac{d}{dx}F(u(x)) = F'(u(x)) \cdot u'(x)$$

Equivalently, if  $y(x) = F(u(x))$ , then

$$\frac{dy}{dx} = \frac{dF}{du} \cdot \frac{du}{dx}.$$

Consider a function  $f(u)$ , which has antiderivative  $F(u)$ . Then we know that

$$\int f(u) du = \int F'(u) du = F(u) + C$$

Now take the above equation and substitute into it  $u = u(x)$  — i.e. replace the variable  $u$  with any (differentiable) function of  $x$  to get

$$\int f(u) du \Big|_{u=u(x)} = F(u(x)) + C$$

But now the right-hand side is a function of  $x$ , so we can differentiate it with respect to  $x$  to get

$$\frac{d}{dx}F(u(x)) = F'(u(x)) \cdot u'(x)$$

This tells us that  $F(u(x))$  is an antiderivative of the function  $F'(u(x)) \cdot u'(x) = f(u(x))u'(x)$ . Thus we know

$$\int f(u(x)) \cdot u'(x) dx = F(u(x)) + C = \int f(u) du \Big|_{u=u(x)}$$

This is the substitution rule for indefinite integrals.



**Theorem 3.4.2** (The substitution rule — indefinite integral version).

For any differentiable function  $u(x)$ :

$$\int f(u(x))u'(x)dx = \int f(u)du \Big|_{u=u(x)}$$

In order to apply the substitution rule successfully we will have to write the integrand in the form  $f(u(x)) \cdot u'(x)$ . To do this we need to make a good choice of the function  $u(x)$ ; after that it is not hard to then find  $f(u)$  and  $u'(x)$ . Unfortunately there is no one strategy for choosing  $u(x)$ . This can make applying the substitution rule more art than science<sup>30</sup>. Here we suggest two possible strategies for picking  $u(x)$ :

- (1) Factor the integrand and choose one of the factors to be  $u'(x)$ . For this to work, you must be able to easily find the antiderivative of the chosen factor. The antiderivative will be  $u(x)$ .
- (2) Look for a factor in the integrand that is a function with an argument that is more complicated than just “ $x$ ”. That factor will play the role of  $f(u(x))$ . Choose  $u(x)$  to be the complicated argument.

Here are two examples which illustrate each of those strategies in turn.

**Example 3.4.3**

Consider the integral

$$\int 9 \sin^8(x) \cos(x) dx$$

We want to massage this into the form of the integrand in the substitution rule — namely  $f(u(x)) \cdot u'(x)$ . Our integrand can be written as the product of the two factors

$$\underbrace{9 \sin^8(x)}_{\text{first factor}} \cdot \underbrace{\cos(x)}_{\text{second factor}}$$

and we start by determining (or guessing) which factor plays the role of  $u'(x)$ . We can choose  $u'(x) = 9 \sin^8(x)$  or  $u'(x) = \cos(x)$ .

- If we choose  $u'(x) = 9 \sin^8(x)$ , then antidifferentiating this to find  $u(x)$  is really not very easy. So it is perhaps better to investigate the other choice before proceeding further with this one.
- If we choose  $u'(x) = \cos(x)$ , then we know (Theorem 3.3.16) that  $u(x) = \sin(x)$ . This also works nicely because it makes the other factor simplify quite a bit  $9 \sin^8(x) = 9u^8$ . This looks like the right way to go.

<sup>30</sup> Thankfully this does become easier with experience and we recommend that the reader read some examples and then practice a LOT.

So we go with the second choice. Set  $u'(x) = \cos(x)$ ,  $u(x) = \sin(x)$ , then

$$\begin{aligned}\int 9 \sin^8(x) \cos(x) dx &= \int 9u(x)^8 \cdot u'(x) dx \\ &= \int 9u^8 du \Big|_{u=\sin(x)} && \text{by the substitution rule}\end{aligned}$$

We are now left with the problem of antidifferentiating a monomial; this we can do with Theorem 3.3.16.

$$\begin{aligned}&= (u^9 + C) \Big|_{u=\sin(x)} \\ &= \sin^9(x) + C\end{aligned}$$

Note that  $9 \sin^8(x) \cos(x)$  is a function of  $x$ . So our answer, which is the indefinite integral of  $9 \sin^8(x) \cos(x)$ , must also be a function of  $x$ . This is why we have substituted  $u = \sin(x)$  in the last step of our solution — it makes our solution a function of  $x$ .

Example 3.4.3

Example 3.4.4

Evaluate the integral

$$\int 3x^2 \cos(x^3) dx$$

*Solution.* Again we are going to use the substitution rule and helpfully our integrand is a product of two factors

$$\underbrace{3x^2}_{\text{first factor}} \cdot \underbrace{\cos(x^3)}_{\text{second factor}}$$

The second factor,  $\cos(x^3)$  is a function, namely  $\cos$ , with a complicated argument, namely  $x^3$ . So we try  $u(x) = x^3$ . Then  $u'(x) = 3x^2$ , which is the other factor in the integrand. So the integral becomes

$$\begin{aligned}\int 3x^2 \cos(x^3) dx &= \int u'(x) \cos(u(x)) dx && \text{just swap order of factors} \\ &= \int \cos(u(x)) u'(x) dx && \text{by the substitution rule} \\ &= \int \cos(u) du \Big|_{u=x^3} \\ &= (\sin(u) + C) \Big|_{u=x^3} && \text{using Theorem 3.3.16} \\ &= \sin(x^3) + C\end{aligned}$$

Example 3.4.4

Now let's look at a definite integral.

Example 3.4.5  $\left(\int_0^1 e^x \sin(e^x) dx\right)$

Compute

$$\int_0^1 e^x \sin(e^x) dx.$$

*Solution.* Again we use the substitution rule.

- The integrand is again the product of two factors and we can choose  $u'(x) = e^x$  or  $u'(x) = \sin(e^x)$ .
- If we choose  $u'(x) = e^x$  then  $u(x) = e^x$  and the other factor becomes  $\sin(u)$  — this looks promising. Notice that if we applied the other strategy of looking for a complicated argument then we would arrive at the same choice.
- So we try  $u'(x) = e^x$  and  $u(x) = e^x$ . This gives (if we ignore the limits of integration for a moment)

$$\begin{aligned} \int e^x \sin(e^x) dx &= \int \sin(u(x)) u'(x) dx && \text{apply the substitution rule} \\ &= \int \sin(u) du \Big|_{u=e^x} \\ &= (-\cos(u) + C) \Big|_{u=e^x} \\ &= -\cos(e^x) + C \end{aligned}$$

- But what happened to the limits of integration? We can incorporate them now. We have just shown that the indefinite integral is  $-\cos(e^x)$ , so by the fundamental theorem of calculus

$$\begin{aligned} \int_0^1 e^x \sin(e^x) dx &= [-\cos(e^x)]_0^1 \\ &= -\cos(e^1) - (-\cos(e^0)) \\ &= -\cos(e) + \cos(1) \end{aligned}$$

Example 3.4.5

The example below introduces a special case where the “inside” function is linear.

Example 3.4.6

Compute the indefinite integrals

$$\int \sqrt{2x+1} dx \quad \text{and} \quad \int e^{3x-2} dx$$

*Solution.*

- Starting with the first integral, we see that it is not too hard to spot the complicated argument. If we set  $u(x) = 2x + 1$  then the integrand is just  $\sqrt{u}$ .
- Hence we substitute  $2x + 1 \rightarrow u$  and  $dx \rightarrow \frac{1}{u'(x)} du = \frac{1}{2} du$ :

$$\begin{aligned} \int \sqrt{2x+1} dx &= \int \sqrt{u} \frac{1}{2} du \\ &= \int u^{1/2} \frac{1}{2} du \\ &= \left( \frac{2}{3} u^{3/2} \cdot \frac{1}{2} + C \right) \Big|_{u=2x+1} \\ &= \frac{1}{3} (2x+1)^{3/2} + C \end{aligned}$$

- We can evaluate the second integral in much the same way. Set  $u(x) = 3x - 2$  and replace  $dx$  by  $\frac{1}{u'(x)} du = \frac{1}{3} du$ :

$$\begin{aligned} \int e^{3x-2} dx &= \int e^u \frac{1}{3} du \\ &= \left( \frac{1}{3} e^u + C \right) \Big|_{u=3x-2} \\ &= \frac{1}{3} e^{3x-2} + C \end{aligned}$$

Example 3.4.6

This last example illustrates that substitution can be used to easily deal with arguments of the form  $ax + b$ , i.e. that are linear functions of  $x$ , and suggests the following theorem.

**Theorem 3.4.7.**

Let  $F(u)$  be an antiderivative of  $f(u)$  and let  $a, b$  be constants. Then

$$\int f(ax+b) dx = \frac{1}{a} F(ax+b) + C$$

*Proof.* We can show this using the substitution rule. Let  $u(x) = ax + b$  so  $u'(x) = a$ , then

$$\begin{aligned} \int f(ax + b) dx &= \int f(u) \cdot \frac{1}{u'(x)} du \\ &= \int \frac{1}{a} f(u) du \\ &= \frac{1}{a} \int f(u) du && \text{since } a \text{ is a constant} \\ &= \frac{1}{a} F(u) \Big|_{u=ax+b} + C && \text{since } F(u) \text{ is an antiderivative of } f(u) \\ &= \frac{1}{a} F(ax + b) + C. \end{aligned}$$

□

### 3.4.1 ▶ Substitution and Definite Integrals

Theorem 3.4.2, the substitution rule for indefinite integrals, tells us that if  $F(u)$  is any antiderivative for  $f(u)$ , then  $F(u(x))$  is an antiderivative for  $f(u(x))u'(x)$ . So the Fundamental Theorem of Calculus gives us

$$\begin{aligned} \int_a^b f(u(x))u'(x) dx &= F(u(x)) \Big|_{x=a}^{x=b} \\ &= F(u(b)) - F(u(a)) \\ &= \int_{u(a)}^{u(b)} f(u) du && \text{since } F(u) \text{ is an antiderivative for } f(u) \end{aligned}$$

and we have just found

**Theorem 3.4.8** (The substitution rule — definite integral version).

For any differentiable function  $u(x)$ :

$$\int_a^b f(u(x))u'(x) dx = \int_{u(a)}^{u(b)} f(u) du$$

Notice that to get from the integral on the left hand side to the integral on the right hand side you

- substitute<sup>31</sup>  $u(x) \rightarrow u$  and  $u'(x)dx \rightarrow du$ ,
- set the lower limit for the  $u$  integral to the value of  $u$  (namely  $u(a)$ ) that corresponds to the lower limit of the  $x$  integral (namely  $x = a$ ), and

31 A good way to remember this last step is that we replace  $\frac{du}{dx}dx$  by just  $du$  — which looks like we cancelled out the  $dx$  terms:  $\frac{du}{dx}dx = du$ . While using “cancel the  $dx$ ” is a good mnemonic (memory aid), you should not think of the derivative  $\frac{du}{dx}$  as a fraction — you are not dividing  $du$  by  $dx$ .

- set the upper limit for the  $u$  integral to the value of  $u$  (namely  $u(b)$ ) that corresponds to the upper limit of the  $x$  integral (namely  $x = b$ ).

Also note that we now have two ways to evaluate definite integrals of the form  $\int_a^b f(u(x))u'(x) dx$ .

- We can find the indefinite integral  $\int f(u(x))u'(x) dx$ , using Theorem 3.4.2, and then evaluate the result between  $x = a$  and  $x = b$ . This is what was done in Example 3.4.5.
- Or we can apply Theorem 3.4.2. This entails finding the indefinite integral  $\int f(u) du$  and evaluating the result between  $u = u(a)$  and  $u = u(b)$ . This is what we will do in the following example.

Example 3.4.9  $\left(\int_0^1 x^2 \sin(x^3 + 1) dx\right)$

Compute

$$\int_0^1 x^2 \sin(x^3 + 1) dx$$

*Solution.*

- In this example the integrand is already neatly factored into two pieces. While we could deploy either of our two strategies, it is perhaps easier in this case to choose  $u(x)$  by looking for a complicated argument.
- The second factor of the integrand is  $\sin(x^3 + 1)$ , which is the function  $\sin$  evaluated at  $x^3 + 1$ . So set  $u(x) = x^3 + 1$ , giving  $u'(x) = 3x^2$  and  $f(u) = \sin(u)$
- The first factor of the integrand is  $x^2$  which is not quite  $u'(x)$ , however we can easily massage the integrand into the required form by multiplying and dividing by 3:

$$x^2 \sin(x^3 + 1) = \frac{1}{3} \cdot 3x^2 \cdot \sin(x^3 + 1).$$

- We want this in the form of the substitution rule, so we do a little massaging:

$$\begin{aligned} \int_0^1 x^2 \sin(x^3 + 1) dx &= \int_0^1 \frac{1}{3} \cdot 3x^2 \cdot \sin(x^3 + 1) dx \\ &= \frac{1}{3} \int_0^1 \sin(x^3 + 1) \cdot 3x^2 dx && \text{by Theorem 3.2.1(c)} \end{aligned}$$

- Now we are ready for the substitution rule:

$$\begin{aligned}
 \frac{1}{3} \int_0^1 \sin(x^3 + 1) \cdot 3x^2 dx &= \frac{1}{3} \int_0^1 \underbrace{\sin(x^3 + 1)}_{=f(u(x))} \cdot \underbrace{3x^2}_{=u'(x)} dx \\
 &= \frac{1}{3} \int_0^1 f(u(x))u'(x) dx && \text{with } u(x) = x^3 + 1 \text{ and } f(u) = \sin(u) \\
 &= \frac{1}{3} \int_{u(0)}^{u(1)} f(u) du && \text{by the substitution rule} \\
 &= \frac{1}{3} \int_1^2 \sin(u) du && \text{since } u(0) = 1 \text{ and } u(1) = 2 \\
 &= \frac{1}{3} [-\cos(u)]_1^2 \\
 &= \frac{1}{3} (-\cos(2) - (-\cos(1))) \\
 &= \frac{\cos(1) - \cos(2)}{3}.
 \end{aligned}$$

Example 3.4.9

There is another, and perhaps easier, way to view the manipulations in the previous example. Once you have chosen  $u(x)$  you

- make the substitution  $u(x) \rightarrow u$ ,
- replace  $dx \rightarrow \frac{1}{u'(x)} du$ .

In so doing, we take the integral

$$\begin{aligned}
 \int_a^b f(u(x)) \cdot u'(x) dx &= \int_{u(a)}^{u(b)} f(u) \cdot u'(x) \cdot \frac{1}{u'(x)} du \\
 &= \int_{u(a)}^{u(b)} f(u) du && \text{exactly the substitution rule}
 \end{aligned}$$

but we do not have to manipulate the integrand so as to make  $u'(x)$  explicit. Let us redo the previous example by this approach.

Example 3.4.10 (*Example 3.4.9 revisited*)

Compute the integral

$$\int_0^1 x^2 \sin(x^3 + 1) dx$$

*Solution.*

- We have already observed that one factor of the integrand is  $\sin(x^3 + 1)$ , which is  $\sin$  evaluated at  $x^3 + 1$ . Thus we try setting  $u(x) = x^3 + 1$ .
- This makes  $u'(x) = 3x^2$ , and we replace  $u(x) = x^3 + 1 \rightarrow u$  and  $dx \rightarrow \frac{1}{u'(x)} du = \frac{1}{3x^2} du$ :

$$\begin{aligned} \int_0^1 x^2 \sin(x^3 + 1) dx &= \int_{u(0)}^{u(1)} \underbrace{x^2 \sin(x^3 + 1)}_{=\sin(u)} \frac{1}{3x^2} du \\ &= \int_1^2 \sin(u) \frac{x^2}{3x^2} du \\ &= \int_1^2 \frac{1}{3} \sin(u) du \\ &= \frac{1}{3} \int_1^2 \sin(u) du \end{aligned}$$

which is precisely the integral we found in Example 3.4.9.

Example 3.4.10

We can do the following example using the substitution rule or Theorem 3.4.7:

Example 3.4.11  $\left( \int_0^{\pi/2} \cos(3x) dx \right)$

Compute  $\int_0^{\pi/2} \cos(3x) dx$ .

- In this example we should set  $u = 3x$ , and substitute  $dx \rightarrow \frac{1}{u'(x)} du = \frac{1}{3} du$ . When we do this we also have to convert the limits of the integral:  $u(0) = 0$  and  $u(\pi/2) = 3\pi/2$ . This gives

$$\begin{aligned} \int_0^{\pi/2} \cos(3x) dx &= \int_0^{3\pi/2} \cos(u) \frac{1}{3} du \\ &= \left[ \frac{1}{3} \sin(u) \right]_0^{3\pi/2} \\ &= \frac{\sin(3\pi/2) - \sin(0)}{3} \\ &= \frac{-1 - 0}{3} = -\frac{1}{3}. \end{aligned}$$

- We can also do this example more directly using the above theorem. Since  $\sin(x)$  is an antiderivative of  $\cos(x)$ , Theorem 3.4.7 tells us that  $\frac{\sin(3x)}{3}$  is an antiderivative of



$\cos(3x)$ . Hence

$$\begin{aligned}\int_0^{\pi/2} \cos(3x) dx &= \left[ \frac{\sin(3x)}{3} \right]_0^{\pi/2} \\ &= \frac{\sin(3\pi/2) - \sin(0)}{3} \\ &= -\frac{1}{3}.\end{aligned}$$

Example 3.4.11

### 3.4.2 ▶ More Substitution Examples

The rest of this section is just more examples of the substitution rule. We recommend that you after reading these that you practice many examples by yourself under exam conditions. Practice is integral to the learning process – there is no substitution for it.

Example 3.4.12  $\left( \int_0^1 x^2 \sin(1 - x^3) dx \right)$

This integral looks a lot like that of Example 3.4.9. It makes sense to try  $u(x) = 1 - x^3$  since it is the argument of  $\sin(1 - x^3)$ . We

- substitute  $u = 1 - x^3$  and
- replace  $dx$  with  $\frac{1}{u'(x)} du = \frac{1}{-3x^2} du$ ,
- when  $x = 0$ , we have  $u = 1 - 0^3 = 1$  and
- when  $x = 1$ , we have  $u = 1 - 1^3 = 0$ .

So

$$\begin{aligned}\int_0^1 x^2 \sin(1 - x^3) \cdot dx &= \int_1^0 x^2 \sin(u) \cdot \frac{1}{-3x^2} du \\ &= \int_1^0 -\frac{1}{3} \sin(u) du.\end{aligned}$$

Note that the lower limit of the  $u$ -integral, namely 1, is larger than the upper limit, which is 0. There is absolutely nothing wrong with that. We can simply evaluate the  $u$ -integral in the normal way. Since  $-\cos(u)$  is an antiderivative of  $\sin(u)$ :

$$\begin{aligned}&= \left[ \frac{\cos(u)}{3} \right]_1^0 \\ &= \frac{\cos(0) - \cos(1)}{3} \\ &= \frac{1 - \cos(1)}{3}.\end{aligned}$$

## Example 3.4.12

Example 3.4.13  $\left(\int_0^1 \frac{1}{(2x+1)^3} dx\right)$ 

Compute  $\int_0^1 \frac{1}{(2x+1)^3} dx$ .

We could do this one using Theorem 3.4.7, but it's not too hard to do without. We can think of the integrand as the function “one over a cube” with the argument  $2x + 1$ . So it makes sense to substitute  $u = 2x + 1$ . That is

- set  $u = 2x + 1$  and
- replace  $dx \rightarrow \frac{1}{u'(x)} du = \frac{1}{2} du$ .
- When  $x = 0$ , we have  $u = 2 \times 0 + 1 = 1$  and
- when  $x = 1$ , we have  $u = 2 \times 1 + 1 = 3$ .

So

$$\begin{aligned} \int_0^1 \frac{1}{(2x+1)^3} dx &= \int_1^3 \frac{1}{u^3} \cdot \frac{1}{2} du \\ &= \frac{1}{2} \int_1^3 u^{-3} du \\ &= \frac{1}{2} \left[ \frac{u^{-2}}{-2} \right]_1^3 \\ &= \frac{1}{2} \left( \frac{1}{-2} \cdot \frac{1}{9} - \frac{1}{-2} \cdot \frac{1}{1} \right) \\ &= \frac{1}{2} \left( \frac{1}{2} - \frac{1}{18} \right) = \frac{1}{2} \cdot \frac{8}{18} \\ &= \frac{2}{9} \end{aligned}$$

## Example 3.4.13

Example 3.4.14  $\left(\int_0^1 \frac{x}{1+x^2} dx\right)$ 

Evaluate  $\int_0^1 \frac{x}{1+x^2} dx$ .

*Solution.*

- The integrand can be rewritten as  $x \cdot \frac{1}{1+x^2}$ . This second factor suggests that we should try setting  $u = 1 + x^2$  — and so we interpret the second factor as the function “one over” evaluated at argument  $1 + x^2$ .
- With this choice we

- set  $u = 1 + x^2$ ,
- substitute  $dx \rightarrow \frac{1}{2x} du$ , and
- translate the limits of integration: when  $x = 0$ , we have  $u = 1 + 0^2 = 1$  and when  $x = 1$ , we have  $u = 1 + 1^2 = 2$ .

- The integral then becomes

$$\begin{aligned} \int_0^1 \frac{x}{1+x^2} dx &= \int_1^2 \frac{x}{u} \frac{1}{2x} du \\ &= \int_1^2 \frac{1}{2u} du \\ &= \frac{1}{2} [\ln |u|]_1^2 \\ &= \frac{\ln 2 - \ln 1}{2} = \frac{\ln 2}{2}. \end{aligned}$$

Example 3.4.14

Example 3.4.15 ( $\int x^3 \cos(x^4 + 2) dx$ )

Compute the integral  $\int x^3 \cos(x^4 + 2) dx$ .

*Solution.*

- The integrand is the product of  $\cos$  evaluated at the argument  $x^4 + 2$  times  $x^3$ , which aside from a factor of 4, is the derivative of the argument  $x^4 + 2$ .
- Hence we set  $u = x^4 + 2$  and then substitute  $dx \rightarrow \frac{1}{u'(x)} du = \frac{1}{4x^3} du$ .
- Before proceeding further, we should note that this is an indefinite integral so we don't have to worry about the limits of integration. However we do need to make sure our answer is a function of  $x$  — we cannot leave it as a function of  $u$ .
- With this choice of  $u$ , the integral then becomes

$$\begin{aligned} \int x^3 \cos(x^4 + 2) dx &= \int x^3 \cos(u) \frac{1}{4x^3} du \Big|_{u=x^4+2} \\ &= \int \frac{1}{4} \cos(u) du \Big|_{u=x^4+2} \\ &= \left( \frac{1}{4} \sin(u) + C \right) \Big|_{u=x^4+2} \\ &= \frac{1}{4} \sin(x^4 + 2) + C. \end{aligned}$$

Example 3.4.15

The next two examples are more involved and require more careful thinking.

Example 3.4.16  $\left(\int \sqrt{1+x^2} x^3 dx\right)$

Compute  $\int \sqrt{1+x^2} x^3 dx$ .

- An obvious choice of  $u$  is the argument inside the square root. So substitute  $u = 1 + x^2$  and  $dx \rightarrow \frac{1}{2x} du$ .
- When we do this we obtain

$$\begin{aligned} \int \sqrt{1+x^2} \cdot x^3 dx &= \int \sqrt{u} \cdot x^3 \cdot \frac{1}{2x} du \\ &= \int \frac{1}{2} \sqrt{u} \cdot x^2 du \end{aligned}$$

Unlike all our previous examples, we have not cancelled out all of the  $x$ 's from the integrand. However before we do the integral with respect to  $u$ , the integrand must be expressed solely in terms of  $u$  — no  $x$ 's are allowed. (Look that integrand on the right hand side of Theorem 3.4.2.)

- But all is not lost. We can rewrite the factor  $x^2$  in terms of the variable  $u$ . We know that  $u = 1 + x^2$ , so this means  $x^2 = u - 1$ . Substituting this into our integral gives

$$\begin{aligned} \int \sqrt{1+x^2} \cdot x^3 dx &= \int \frac{1}{2} \sqrt{u} \cdot x^2 du \\ &= \int \frac{1}{2} \sqrt{u} \cdot (u-1) du \\ &= \frac{1}{2} \int (u^{3/2} - u^{1/2}) du \\ &= \frac{1}{2} \left( \frac{2}{5} u^{5/2} - \frac{2}{3} u^{3/2} \right) \Big|_{u=x^2+1} + C \\ &= \left( \frac{1}{5} u^{5/2} - \frac{1}{3} u^{3/2} \right) \Big|_{u=x^2+1} + C \\ &= \frac{1}{5} (x^2+1)^{5/2} - \frac{1}{3} (x^2+1)^{3/2} + C. \end{aligned}$$

Oof!

- Don't forget that you can always check the answer by differentiating:

$$\begin{aligned} \frac{d}{dx} \left( \frac{1}{5} (x^2+1)^{5/2} - \frac{1}{3} (x^2+1)^{3/2} + C \right) &= \frac{d}{dx} \left( \frac{1}{5} (x^2+1)^{5/2} \right) - \frac{d}{dx} \left( \frac{1}{3} (x^2+1)^{3/2} \right) \\ &= \frac{1}{5} \cdot 2x \cdot \frac{5}{2} \cdot (x^2+1)^{3/2} - \frac{1}{3} \cdot 2x \cdot \frac{3}{2} \cdot (x^2+1)^{1/2} \\ &= x(x^2+1)^{3/2} - x(x^2+1)^{1/2} \\ &= x[(x^2+1) - 1] \cdot \sqrt{x^2+1} \\ &= x^3 \sqrt{x^2+1}. \end{aligned}$$

which is the original integrand ✓.

Example 3.4.16

Example 3.4.17 ( $\int \tan x dx$ )Evaluate the indefinite integral  $\int \tan(x) dx$ .*Solution.*

- At first glance there is nothing to manipulate here and so very little to go on. However we can rewrite  $\tan x$  as  $\frac{\sin x}{\cos x}$ , making the integral  $\int \frac{\sin x}{\cos x} dx$ . This gives us more to work with.
- Now think of the integrand as being the product  $\frac{1}{\cos x} \cdot \sin x$ . This suggests that we set  $u = \cos x$  and that we interpret the first factor as the function “one over” evaluated at  $u = \cos x$ .
- Substitute  $u = \cos x$  and  $dx \rightarrow \frac{1}{-\sin x} du$  to give:

$$\begin{aligned} \int \frac{\sin x}{\cos x} dx &= \int \frac{\sin x}{u} \frac{1}{-\sin x} du \Big|_{u=\cos x} \\ &= \int -\frac{1}{u} du \Big|_{u=\cos x} \\ &= -\ln |\cos x| + C && \text{and if we want to go further} \\ &= \ln \left| \frac{1}{\cos x} \right| + C \\ &= \ln |\sec x| + C. \end{aligned}$$

Example 3.4.17

### 3.5▲ Integration by Parts

The Fundamental Theorem of Calculus tells us that it is very easy to integrate a derivative. In particular, we know that

$$\int \frac{d}{dx} (F(x)) dx = F(x) + C$$

We can exploit this in order to develop another rule for integration — in particular a rule to help us integrate products of simpler function such as

$$\int xe^x dx$$

In so doing we will arrive at a method called “integration by parts”.

To do this we start with the product rule and integrate. Recall that the product rule says

$$\frac{d}{dx}u(x)v(x) = u'(x)v(x) + u(x)v'(x)$$

Integrating this gives

$$\begin{aligned}\int [u'(x)v(x) + u(x)v'(x)]dx &= [\text{a function whose derivative is } u'v + uv'] + C \\ &= u(x)v(x) + C\end{aligned}$$

Now this, by itself, is not terribly useful. In order to apply it we need to have a function whose integrand is a sum of products that is in exactly this form  $u'(x)v(x) + u(x)v'(x)$ . This is far too specialised.

However if we tease this apart a little:

$$\int [u'(x)v(x) + u(x)v'(x)]dx = \int u'(x)v(x)dx + \int u(x)v'(x)dx$$

Bring one of the integrals to the left-hand side

$$u(x)v(x) - \int u'(x)v(x)dx = \int u(x)v'(x)dx$$

Swap left and right sides

$$\int u(x)v'(x)dx = u(x)v(x) - \int u'(x)v(x)dx$$

In this form we take the integral of one product and express it in terms of the integral of a different product. If we express it like that, it doesn't seem too useful. However, if the second integral is easier, then this process helps us.

Let us do a simple example before explaining this more generally.

Example 3.5.1 ( $\int xe^x dx$ )

Compute the integral  $\int xe^x dx$ .

*Solution.*

- We start by taking the equation above

$$\int u(x)v'(x)dx = u(x)v(x) - \int u'(x)v(x)dx$$

- Now set  $u(x) = x$  and  $v'(x) = e^x$ . How did we know how to make this choice? We will explain some strategies later. For now, let us just accept this choice and keep going.
- In order to use the formula we need to know  $u'(x)$  and  $v(x)$ . In this case it is quite straightforward:  $u'(x) = 1$  and  $v(x) = e^x$ .

- Plug everything into the formula:

$$\int xe^x dx = xe^x - \int e^x dx$$

So our original more difficult integral has been turned into a question of computing an easy one.

$$= xe^x - e^x + C$$

- We can check our answer by differentiating:

$$\begin{aligned} \frac{d}{dx}(xe^x - e^x + C) &= \underbrace{xe^x + 1 \cdot e^x}_{\text{by product rule}} - e^x + 0 \\ &= xe^x \end{aligned}$$

as required.

Example 3.5.1

The process we have used in the above example is called “integration by parts”. When our integrand is a product we try to write it as  $u(x)v'(x)$  — we need to choose one factor to be  $u(x)$  and the other to be  $v'(x)$ . We then compute  $u'(x)$  and  $v(x)$  and then apply the following theorem:

**Theorem 3.5.2** (Integration by parts).

Let  $u(x)$  and  $v(x)$  be continuously differentiable. Then

$$\int u(x) v'(x) dx = u(x) v(x) - \int v(x) u'(x) dx$$

If we write  $dv$  for  $v'(x)dx$  and  $du$  for  $u'(x)dx$  (as the substitution rule suggests), then the formula becomes

$$\int u dv = uv - \int v du$$

The application of this formula is known as integration by parts.

The corresponding statement for definite integrals is

$$\int_a^b u(x) v'(x) dx = u(b) v(b) - u(a) v(a) - \int_a^b v(x) u'(x) dx$$

Integration by parts is not as easy to apply as the product rule for derivatives. This is because it relies on us

- (1) judiciously choosing  $u(x)$  and  $v'(x)$ , then
- (2) computing  $u'(x)$  and  $v(x)$  — which requires us to antidifferentiate  $v'(x)$ , and finally

(3) that the integral  $\int u'(x)v(x)dx$  is easier than the integral we started with.

Notice that any antiderivative of  $v'(x)$  will do. All antiderivatives of  $v'(x)$  are of the form  $v(x) + A$  with  $A$  a constant. Putting this into the integration by parts formula gives

$$\begin{aligned} \int u(x)v'(x)dx &= u(x)(v(x) + A) - \int u'(x)(v(x) + A) dx \\ &= u(x)v(x) + Au(x) - \int u'(x)v(x)dx - A \underbrace{\int u'(x)dx}_{=Au(x)+C} \\ &= u(x)v(x) - \int u'(x)v(x)dx + C \end{aligned}$$

So that constant  $A$  will always cancel out.

In most applications (but not all) our integrand will be a product of two factors so we have two choices for  $u(x)$  and  $v'(x)$ . Typically one of these choices will be “good” (in that it results in a simpler integral) while the other will be “bad” (we cannot antidifferentiate our choice of  $v'(x)$  or the resulting integral is harder). Let us illustrate what we mean by returning to our previous example.

Example 3.5.3 ( $\int xe^x dx$  — again)

Our integrand is the product of two factors

$$x \qquad \text{and} \qquad e^x$$

This gives us two obvious choices of  $u$  and  $v'$ :

$$\begin{array}{ll} u(x) = x & v'(x) = e^x \\ \text{or} & \\ u(x) = e^x & v'(x) = x \end{array}$$

We should explore both choices:

1. If take  $u(x) = x$  and  $v'(x) = e^x$ . We then quickly compute

$$u'(x) = 1 \qquad \text{and} \qquad v(x) = e^x$$

which means we will need to integrate (in the right-hand side of the integration by parts formula)

$$\int u'(x)v(x)dx = \int 1 \cdot e^x dx$$

which looks straightforward. This is a good indication that this is the right choice of  $u(x)$  and  $v'(x)$ .

2. But before we do that, we should also explore the other choice, namely  $u(x) = e^x$  and  $v'(x) = x$ . This implies that

$$u'(x) = e^x \qquad \text{and} \qquad v(x) = \frac{1}{2}x^2$$



which means we need to integrate

$$\int u'(x)v(x)dx = \int \frac{1}{2}x^2 \cdot e^x dx.$$

This is at least as hard as the integral we started with. Hence we should try the first choice.

With our choice made, we integrate by parts to get

$$\begin{aligned} \int xe^x dx &= xe^x - \int e^x dx \\ &= xe^x - e^x + C. \end{aligned}$$

The above reasoning is a very typical workflow when using integration by parts.

Example 3.5.3

Integration by parts is often used

- to eliminate factors of  $x$  from an integrand like  $xe^x$  by using that  $\frac{d}{dx}x = 1$  and
- to eliminate a  $\ln x$  from an integrand by using that  $\frac{d}{dx} \ln x = \frac{1}{x}$  and
- to eliminate inverse trig functions, like  $\arctan x$ , from an integrand by using that, for example,  $\frac{d}{dx} \arctan x = \frac{1}{1+x^2}$ .

Example 3.5.4 ( $\int x \sin x dx$ )

*Solution.*

- Again we have a product of two factors giving us two possible choices.

(1) If we choose  $u(x) = x$  and  $v'(x) = \sin x$ , then we get

$$u'(x) = 1 \quad \text{and} \quad v(x) = -\cos x$$

which is looking promising.

(2) On the other hand if we choose  $u(x) = \sin x$  and  $v'(x) = x$ , then we have

$$u'(x) = \cos x \quad \text{and} \quad v(x) = \frac{1}{2}x^2$$

which is looking worse — we'd need to integrate  $\int \frac{1}{2}x^2 \cos x dx$ .

- So we stick with the first choice. Plugging  $u(x) = x$ ,  $v(x) = -\cos x$  into integration by parts gives us

$$\begin{aligned} \int x \sin x dx &= -x \cos x - \int 1 \cdot (-\cos x) dx \\ &= -x \cos x + \sin x + C \end{aligned}$$

- Again we can check our answer by differentiating:

$$\begin{aligned}\frac{d}{dx}(-x \cos x + \sin x + C) &= -\cos x + x \sin x + \cos x + 0 \\ &= x \sin x \checkmark\end{aligned}$$

Once we have practised this a bit we do not really need to write as much. Let us solve it again, but showing only what we need to.

*Solution.*

- We use integration by parts to solve the integral.
- Set  $u(x) = x$  and  $v'(x) = \sin x$ . Then  $u'(x) = 1$  and  $v(x) = -\cos x$ , and

$$\begin{aligned}\int x \sin x dx &= -x \cos x + \int \cos x dx \\ &= -x \cos x + \sin x + C.\end{aligned}$$

Example 3.5.4

It is pretty standard practice to reduce the notation even further in these problems. As noted above, many people write the integration by parts formula as

$$\int u dv = uv - \int v du$$

where  $du, dv$  are shorthand for  $u'(x)dx, v'(x)dx$ . Let us write up the previous example using this notation.

Example 3.5.5 ( $\int x \sin x dx$  yet again)

*Solution.* Using integration by parts, we set  $u = x$  and  $dv = \sin x dx$ . This makes  $du = 1 dx$  and  $v = -\cos x$ . Consequently

$$\begin{aligned}\int x \sin x dx &= \int u dv \\ &= uv - \int v du \\ &= -x \cos x + \int \cos x dx \\ &= -x \cos x + \sin x + C\end{aligned}$$

You can see that this is a very neat way to write up these problems and we will continue using this shorthand in the examples that follow below.

Example 3.5.5

We can also use integration by parts to eliminate higher powers of  $x$ . We just need to apply the method more than once.

Example 3.5.6 ( $\int x^2 e^x dx$ )

*Solution.*

- Let  $u = x^2$  and  $dv = e^x dx$ . This then gives  $du = 2x dx$  and  $v = e^x$ , and

$$\int x^2 e^x dx = x^2 e^x - \int 2x e^x dx$$

- So we have reduced the problem of computing the original integral to one of integrating  $2x e^x$ . We know how to do this — just integrate by parts again:

$$\begin{aligned} \int x^2 e^x dx &= x^2 e^x - \int 2x e^x dx && \text{set } u = 2x, dv = e^x dx \\ &= x^2 e^x - \left( 2x e^x - \int 2e^x dx \right) && \text{since } du = 2dx, v = e^x \\ &= x^2 e^x - 2x e^x + 2e^x + C \end{aligned}$$

- We can, if needed, check our answer by differentiating:

$$\begin{aligned} \frac{d}{dx} (x^2 e^x - 2x e^x + 2e^x + C) &= (x^2 e^x + 2x e^x) - (2x e^x + 2e^x) + 2e^x + 0 \\ &= x^2 e^x \checkmark \end{aligned}$$

A similar iterated application of integration by parts will work for integrals

$$\int P(x) (Ae^{ax} + B \sin(bx) + C \cos(cx)) dx$$

where  $P(x)$  is a polynomial and  $A, B, C, a, b, c$  are constants.

Example 3.5.6

Now let us look at integrands containing logarithms. We don't know the antiderivative of  $\ln x$ , but we can eliminate  $\ln x$  from an integrand by using integration by parts with  $u = \ln x$ .

Example 3.5.7 ( $\int x \ln x dx$ )

*Solution.*

- We have two choices for  $u$  and  $dv$ .
  - (1) Set  $u = x$  and  $dv = \ln x dx$ . This gives  $du = dx$  but  $v$  is hard to compute — we haven't done it yet<sup>32</sup>. Before we go further along this path, we should look to see what happens with the other choice.

<sup>32</sup> We will soon.

- (2) Set  $u = \ln x$  and  $dv = x dx$ . This gives  $du = \frac{1}{x} dx$  and  $v = \frac{1}{2}x^2$ , and we have to integrate

$$\int v du = \int \frac{1}{x} \cdot \frac{1}{2}x^2 dx$$

which is easy.

- So we proceed with the second choice.

$$\begin{aligned} \int x \ln x dx &= \frac{1}{2}x^2 \ln x - \int \frac{1}{2}x dx \\ &= \frac{1}{2}x^2 \ln x - \frac{1}{4}x^2 + C \end{aligned}$$

- We can check our answer quickly:

$$\frac{d}{dx} \left( \frac{x^2}{2} \ln x - \frac{x^2}{4} + C \right) = x \ln x + \frac{x^2}{2} \frac{1}{x} - \frac{x}{2} + 0 = x \ln x$$

Example 3.5.7

### 3.5.1 ▶ Another Technique using Integration by Parts: $dv = dx$

It's tempting to think of integration by parts as a product rule for integration. It can be used that way, but it also shows up in more surprising contexts. Integration by parts lets us find the antiderivatives of the natural logarithm and inverse trigonometric functions, despite these not being the product of two functions in any obvious way.

Example 3.5.8 ( $\int \ln x dx$ )

It is not immediately obvious that one should use integration by parts to compute the integral

$$\int \ln x dx$$

since the integrand is not a product. But we should persevere — indeed this is a situation where our shorter notation helps to clarify how to proceed.

*Solution.*

- In the previous example we saw that we could remove the factor  $\ln x$  by setting  $u = \ln x$  and using integration by parts. Let us try repeating this. When we make this choice, we are then forced to take  $dv = dx$  — that is we choose  $v'(x) = 1$ . Once we have made this sneaky move everything follows quite directly.

- We then have  $du = \frac{1}{x}dx$  and  $v = x$ , and the integration by parts formula gives us

$$\begin{aligned}\int \ln x dx &= x \ln x - \int \frac{1}{x} \cdot x dx \\ &= x \ln x - \int 1 dx \\ &= x \ln x - x + C\end{aligned}$$

- As always, it is a good idea to check our result by verifying that the derivative of the answer really is the integrand.

$$\frac{d}{dx}(x \ln x - x + C) = \ln x + x \frac{1}{x} - 1 + 0 = \ln x$$

Example 3.5.8

The same method works almost exactly to compute the antiderivatives of  $\arcsin(x)$  and  $\arctan(x)$ :

Example 3.5.9 ( $\int \arctan(x)dx$  and  $\int \arcsin(x)dx$ )

Compute the antiderivatives of the inverse sine and inverse tangent functions.

*Solution.*

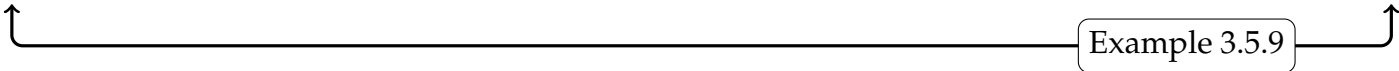
- Again neither of these integrands are products, but that is no impediment. In both cases we set  $dv = dx$  (ie  $v'(x) = 1$ ) and choose  $v(x) = x$ .
- For inverse tan we choose  $u = \arctan(x)$ , so  $du = \frac{1}{1+x^2}dx$ :

$$\begin{aligned}\int \arctan(x)dx &= x \arctan(x) - \int x \cdot \frac{1}{1+x^2}dx && \text{now use substitution rule} \\ &= x \arctan(x) - \int \frac{w'(x)}{2} \cdot \frac{1}{w}dx && \text{with } w(x) = 1+x^2, w'(x) = 2x \\ &= x \arctan(x) - \frac{1}{2} \int \frac{1}{w}dw \\ &= x \arctan(x) - \frac{1}{2} \ln |w| + C \\ &= x \arctan(x) - \frac{1}{2} \ln |1+x^2| + C && \text{but } 1+x^2 > 0, \text{ so} \\ &= x \arctan(x) - \frac{1}{2} \ln(1+x^2) + C\end{aligned}$$

- Similarly for inverse sine we choose  $u = \arcsin(x)$  so  $du = \frac{1}{\sqrt{1-x^2}}dx$ :

$$\begin{aligned} \int \arcsin(x)dx &= x \arcsin(x) - \int \frac{x}{\sqrt{1-x^2}}dx && \text{now use substitution rule} \\ &= x \arcsin(x) - \int \frac{-w'(x)}{2} \cdot w^{-1/2}dx && \text{with } w(x) = 1-x^2, w'(x) = -2x \\ &= x \arcsin(x) + \frac{1}{2} \int w^{-1/2}dw \\ &= x \arcsin(x) + \frac{1}{2} \cdot 2w^{1/2} + C \\ &= x \arcsin(x) + \sqrt{1-x^2} + C \end{aligned}$$

- Both can be checked quite quickly by differentiating — but we leave that as an exercise for the reader.



Example 3.5.9

### 3.6▲ Numerical Integration

By now the reader will have come to appreciate that integration is generally quite a bit more difficult than differentiation. There are a great many simple-looking integrals, such as  $\int e^{-x^2} dx$ , that are either very difficult or even impossible to express in terms of standard functions<sup>33</sup>. Such integrals are not merely mathematical curiosities, but arise very naturally in many contexts. For example, the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

is extremely important in many areas of mathematics, and also in many practical applications of statistics.

In such applications we need to be able to evaluate this integral (and many others) at a given numerical value of  $x$ . In this section we turn to the problem of how to find (approximate) numerical values for integrals, without having to evaluate them algebraically.

We have already seen that a definite integral can be approximated by a Riemann sum. Returning to Definition 3.1.8, we defined

$$\int_a^b f(x)dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n}$$

(subject to some fine print). That means we can approximate the actual value of  $\int_a^b f(x)dx$  by evaluating a Riemann sum  $\sum_{i=1}^n f(x_{i,n}^*) \cdot \frac{b-a}{n}$  for a large-enough value of  $n$ .

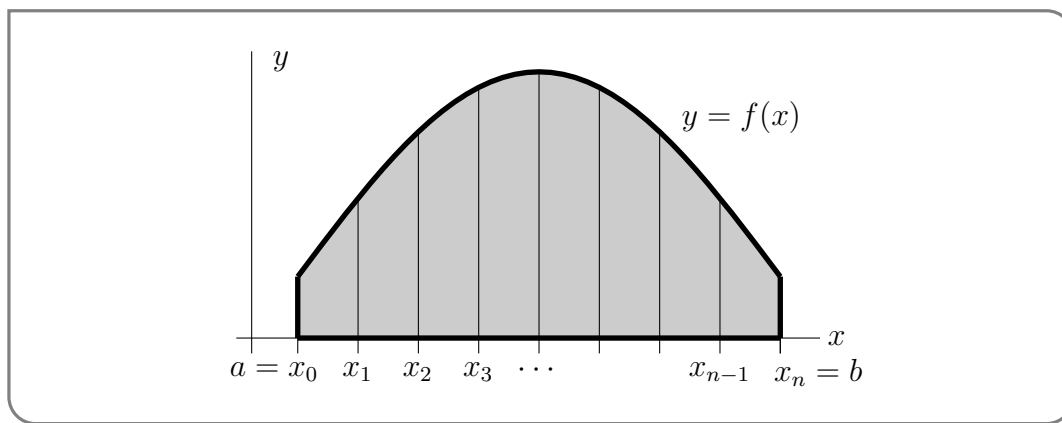
---

33 We apologise for being a little sloppy here — but we just want to say that it can be very hard or even impossible to write some integrals as some finite sized expression involving polynomials, exponentials, logarithms and trigonometric functions. We don't want to get into a discussion of computability, though that is a very interesting topic.

In practice, the phrase “large enough” in the above paragraph is quite important. If a sufficiently accurate estimation requires a huge value of  $n$ , it might take a lot of computing power to evaluate. So, we will introduce a method that uses the same basic ideas as Riemann sums, but that is often more efficient.

Let’s take a moment to remember where Riemann sums come from, and think about how we might have made different choices to end up at different methods of approximating signed areas.

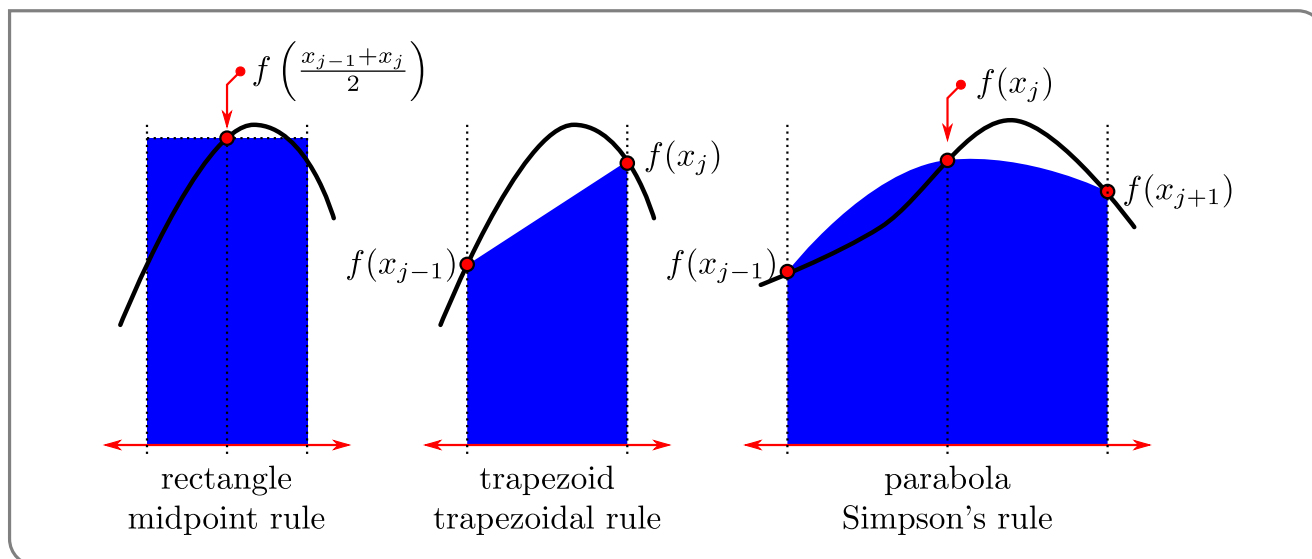
- We first select an integer  $n > 0$ , called the “number of steps”.
- We then divide the interval of integration,  $a \leq x \leq b$ , into  $n$  equal subintervals, each of length  $\Delta x = \frac{b-a}{n}$ . The first subinterval runs from  $x_0 = a$  to  $x_1 = a + \Delta x$ . The second runs from  $x_1$  to  $x_2 = a + 2\Delta x$ , and so on. The last runs from  $x_{n-1} = b - \Delta x$  to  $x_n = b$ .



This splits the original integral into  $n$  pieces:

$$\int_a^b f(x) \, dx = \int_{x_0}^{x_1} f(x) \, dx + \int_{x_1}^{x_2} f(x) \, dx + \cdots + \int_{x_{n-1}}^{x_n} f(x) \, dx$$

Each subintegral  $\int_{x_{j-1}}^{x_j} f(x) \, dx$  is approximated by the area of a simple geometric figure. For a Riemann sum, we used a rectangle, but we could have used different shapes. Three algorithms commonly considered approximate the area by rectangles, trapezoids and parabolas, respectively.



We will focus our attention on Simpson's rule, but brief overviews of the three methods are below. (For more detailed information on the Midpoint and Trapezoid Rules, see Appendix A.10.)

- (1) The midpoint rule approximates each subintegral by the area of a rectangle of height given by the value of the function at the midpoint of the subinterval

$$\int_{x_{j-1}}^{x_j} f(x) dx \approx f\left(\frac{x_{j-1} + x_j}{2}\right) \Delta x$$

This is illustrated in the leftmost figure above.

- (2) The trapezoidal rule approximates each subintegral by the area of a trapezoid with vertices at  $(x_{j-1}, 0)$ ,  $(x_{j-1}, f(x_{j-1}))$ ,  $(x_j, f(x_j))$ ,  $(x_j, 0)$ :

$$\int_{x_{j-1}}^{x_j} f(x) dx \approx \frac{1}{2} (f(x_{j-1}) + f(x_j)) \Delta x$$

The trapezoid is illustrated in the middle figure above.

- (3) Simpson's rule approximates two adjacent subintegrals by the area under a parabola that passes through the points  $(x_{j-1}, f(x_{j-1}))$ ,  $(x_j, f(x_j))$  and  $(x_{j+1}, f(x_{j+1}))$ :

$$\int_{x_{j-1}}^{x_{j+1}} f(x) dx \approx \frac{1}{3} (f(x_{j-1}) + 4f(x_j) + f(x_{j+1})) \Delta x$$

The parabola is illustrated in the right hand figure above. We shall derive the formula for the area shortly.

One thing that makes these rules useful is that we have methods of bounding the errors involved. So, let's define exactly what we mean by "error."



**Definition 3.6.1.**

Suppose that  $\alpha$  is an approximation to  $A$ . This approximation has

- absolute error  $|A - \alpha|$  and
- relative error  $\frac{|A - \alpha|}{A}$  and
- percentage error  $100 \frac{|A - \alpha|}{A}$

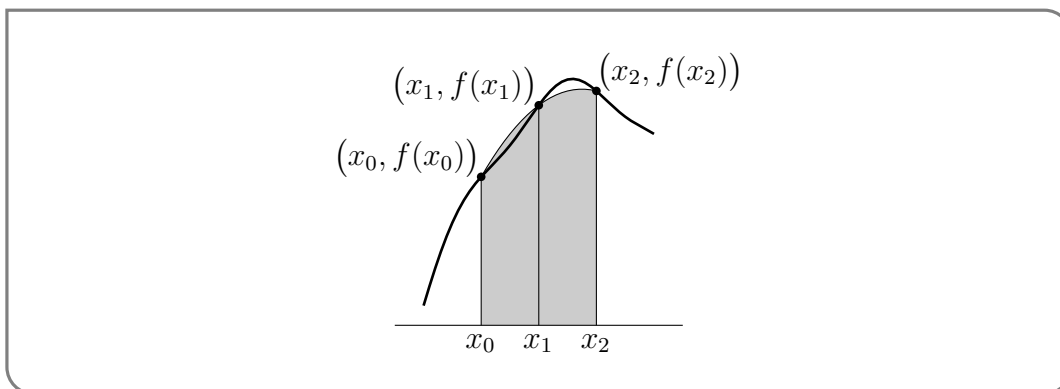
We will discuss errors further in Section 3.6.2.

**3.6.1 ▶ Simpson's Rule**

Simpson's<sup>34</sup> rule approximates the integral over two neighbouring subintervals by the area between a parabola and the  $x$ -axis. In order to describe this parabola we need 3 distinct points (which is why we approximate two subintegrals at a time). That is, we approximate

$$\int_{x_0}^{x_1} f(x) \, dx + \int_{x_1}^{x_2} f(x) \, dx = \int_{x_0}^{x_2} f(x) \, dx$$

by the area bounded by the parabola that passes through the three points  $(x_0, f(x_0))$ ,  $(x_1, f(x_1))$  and  $(x_2, f(x_2))$ , the  $x$ -axis, and the vertical lines  $x = x_0$  and  $x = x_2$ . We repeat



this on the next pair of subintervals and approximate  $\int_{x_2}^{x_4} f(x) \, dx$  by the area between the  $x$ -axis and the part of a parabola with  $x_2 \leq x \leq x_4$ . This parabola passes through the three points  $(x_2, f(x_2))$ ,  $(x_3, f(x_3))$  and  $(x_4, f(x_4))$ . And so on. Because Simpson's rule does the approximation two slices at a time,  $n$  must be even.

To derive Simpson's rule formula, we first find the equation of the parabola that passes through the three points  $(x_0, f(x_0))$ ,  $(x_1, f(x_1))$  and  $(x_2, f(x_2))$ . Then we find the area between the  $x$ -axis and the part of that parabola with  $x_0 \leq x \leq x_2$ . To simplify this computation consider a parabola passing through the points  $(-h, y_{-1})$ ,  $(0, y_0)$  and  $(h, y_1)$ .

34 Simpson's rule is named after the 18th century English mathematician Thomas Simpson, despite its use a century earlier by the German mathematician and astronomer Johannes Kepler. In many German texts the rule is often called Kepler's rule.

Write the equation of the parabola as

$$y = Ax^2 + Bx + C$$

Then the area between it and the  $x$ -axis with  $x$  running from  $-h$  to  $h$  is

$$\begin{aligned} \int_{-h}^h [Ax^2 + Bx + C] dx &= \left[ \frac{A}{3}x^3 + \frac{B}{2}x^2 + Cx \right]_{-h}^h \\ &= \frac{2A}{3}h^3 + 2Ch && \text{it is helpful to write it as} \\ &= \frac{h}{3} (2Ah^2 + 6C) \end{aligned}$$

Now, the three points  $(-h, y_{-1})$ ,  $(0, y_0)$  and  $(h, y_1)$  lie on this parabola if and only if

$$\begin{aligned} Ah^2 - Bh + C &= y_{-1} && \text{at } (-h, y_{-1}) \\ C &= y_0 && \text{at } (0, y_0) \\ Ah^2 + Bh + C &= y_1 && \text{at } (h, y_1) \end{aligned}$$

Adding the first and third equations together gives us

$$2Ah^2 + (B - B)h + 2C = y_{-1} + y_1$$

To this we add four times the middle equation

$$2Ah^2 + 6C = y_{-1} + 4y_0 + y_1.$$

This means that

$$\begin{aligned} \text{area} &= \int_{-h}^h [Ax^2 + Bx + C] dx = \frac{h}{3} (2Ah^2 + 6C) \\ &= \frac{h}{3} (y_{-1} + 4y_0 + y_1) \end{aligned}$$

Note that here

- $h$  is one half of the length of the  $x$ -interval under consideration
- $y_{-1}$  is the height of the parabola at the left hand end of the interval under consideration
- $y_0$  is the height of the parabola at the middle point of the interval under consideration
- $y_1$  is the height of the parabola at the right hand end of the interval under consideration

So Simpson's rule approximates

$$\int_{x_0}^{x_2} f(x) dx \approx \frac{1}{3} \Delta x [f(x_0) + 4f(x_1) + f(x_2)]$$

and

$$\int_{x_2}^{x_4} f(x) \, dx \approx \frac{1}{3} \Delta x [f(x_2) + 4f(x_3) + f(x_4)]$$

and so on. Summing these all together gives:

$$\begin{aligned} \int_a^b f(x) \, dx &= \int_{x_0}^{x_2} f(x) \, dx + \int_{x_2}^{x_4} f(x) \, dx + \int_{x_4}^{x_6} f(x) \, dx + \cdots + \int_{x_{n-2}}^{x_n} f(x) \, dx \\ &\approx \frac{\Delta x}{3} [f(x_0) + 4f(x_1) + f(x_2)] + \frac{\Delta x}{3} [f(x_2) + 4f(x_3) + f(x_4)] \\ &\quad + \frac{\Delta x}{3} [f(x_4) + 4f(x_5) + f(x_6)] + \cdots + \frac{\Delta x}{3} [f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)] \\ &= \left[ f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n) \right] \frac{\Delta x}{3} \end{aligned}$$

In summary:

**Equation 3.6.2**(Simpson's rule).

The Simpson's rule approximation is

$$\int_a^b f(x) \, dx \approx \left[ f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots \right. \\ \left. \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n) \right] \frac{\Delta x}{3}$$

where  $n$  is even and

$$\Delta x = \frac{b-a}{n}, \quad x_0 = a, \quad x_1 = a + \Delta x, \quad x_2 = a + 2\Delta x, \quad \cdots, \quad x_{n-1} = b - \Delta x, \quad x_n = b$$

Notice that Simpson's rule requires only slightly more work than a right Riemann sum. In both cases we must evaluate  $f(x)$  at  $x = x_1, \dots, x_n$ , but we add those terms multiplied by different constants<sup>35</sup>, and for Simpson's rule, we must also find  $f(x_0)$ .

**Example 3.6.3** (Approximating  $\int_0^1 \frac{4}{1+x^2} \, dx$  using Simpson's rule)

We approximate the above integral using Simpson's rule with  $n = 8$  steps.

*Solution.*

- First we set up all the  $x$ -values that we will need. Note that  $a = 0, b = 1, \Delta x = \frac{1}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{1}{8} \quad x_2 = \frac{2}{8} \quad \cdots \quad x_7 = \frac{7}{8} \quad x_8 = \frac{8}{8} = 1$$

35 There is an easy generalisation of Simpson's rule that uses cubics instead of parabolas. It leads to the formula

$$\int_a^b f(x) \, dx = [f(x_0) + 3f(x_1) + 3f(x_2) + 2f(x_3) + 2f(x_4) + 3f(x_5) + 3f(x_6) + 2f(x_7) + \cdots + f(x_n)] \frac{3\Delta x}{8}$$

where  $n$  is a multiple of 3. This result is known as Simpson's second rule and Simpson's  $3/8$  rule. While one can push this approach further (using quartics, quintics etc), it can sometimes lead to larger errors — the interested reader should look up Runge's phenomenon.

- Applying Equation 3.6.2 gives

$$\begin{aligned} \int_0^1 \frac{4}{1+x^2} dx &\approx \left[ \frac{4}{1+0^2} + 4\frac{4}{1+\frac{1}{8^2}} + 2\frac{4}{1+\frac{2^2}{8^2}} + 4\frac{4}{1+\frac{3^2}{8^2}} \right. \\ &\quad \left. + 2\frac{4}{1+\frac{4^2}{8^2}} + 4\frac{4}{1+\frac{5^2}{8^2}} + 2\frac{4}{1+\frac{6^2}{8^2}} + 4\frac{4}{1+\frac{7^2}{8^2}} + \frac{4}{1+\frac{8^2}{8^2}} \right] \frac{1}{8 \times 3} \\ &= \left[ 4 + 4 \times 3.938461538 + 2 \times 3.764705882 + 4 \times 3.506849315 \right. \\ &\quad \left. + 2 \times 3.2 + 4 \times 2.876404494 + 2 \times 2.56 + 4 \times 2.265486726 + 2 \right] \frac{1}{8 \times 3} \\ &= 3.14159250 \end{aligned}$$

(correct to eight decimal places).

- In this case we can compute the integral exactly (which is one of the reasons it was chosen as a first example):

$$\int_0^1 \frac{4}{1+x^2} dx = 4 \arctan x \Big|_0^1 = \pi$$

- Our approximation agrees with  $\pi$  (the exact value of the integral) to six decimal places. So the error in the approximation generated by eight steps of Simpson's rule is  $|3.14159250 - \pi| = 1.5 \times 10^{-7}$ , which is a percent error of only  $100 \frac{|3.14159250 - \pi|}{\pi} \% = 5 \times 10^{-6} \%$ .

Example 3.6.3

When a computer algebra system approximates  $\pi$ , or any other number, it has to approximate it by using other numbers that it either already knows or can compute. Note that the calculation above consisted of multiplying, dividing, adding, and subtracting integers. These are all operations that we can do by hand, and can also easily program a computer to do.

Example 3.6.4 ( $\int_0^\pi \sin x dx$  — Simpson's rule)

Apply Simpson's rule with  $n = 8$  steps to the above integral.

*Solution.*

- We again start by setting up all the  $x$ -values that we will need. So  $a = 0$ ,  $b = \pi$ ,  $\Delta x = \frac{\pi}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{\pi}{8} \quad x_2 = \frac{2\pi}{8} \quad \cdots \quad x_7 = \frac{7\pi}{8} \quad x_8 = \frac{8\pi}{8} = \pi$$

- Applying Equation 3.6.2 gives

$$\begin{aligned}
 \int_0^{\pi} \sin x \, dx &\approx \left[ \sin(x_0) + 4 \sin(x_1) + 2 \sin(x_2) + \cdots + 4 \sin(x_7) + \sin(x_8) \right] \frac{\Delta x}{3} \\
 &= \left[ \sin(0) + 4 \sin\left(\frac{\pi}{8}\right) + 2 \sin\left(\frac{2\pi}{8}\right) + 4 \sin\left(\frac{3\pi}{8}\right) + 2 \sin\left(\frac{4\pi}{8}\right) \right. \\
 &\quad \left. + 4 \sin\left(\frac{5\pi}{8}\right) + 2 \sin\left(\frac{6\pi}{8}\right) + 4 \sin\left(\frac{7\pi}{8}\right) + \sin\left(\frac{8\pi}{8}\right) \right] \frac{\pi}{8 \times 3} \\
 &= \left[ 0 + 4 \times 0.382683 + 2 \times 0.707107 + 4 \times 0.923880 + 2 \times 1.0 \right. \\
 &\quad \left. + 4 \times 0.923880 + 2 \times 0.707107 + 4 \times 0.382683 + 0 \right] \frac{\pi}{8 \times 3} \\
 &= 15.280932 \times 0.130900 \\
 &= 2.00027
 \end{aligned}$$

- Again, we have chosen this example so that we can compare it against the exact value:

$$\int_0^{\pi} \sin x \, dx = [-\cos x]_0^{\pi} = -\cos \pi + \cos 0 = 2.$$

- With only eight steps of Simpson's rule, we achieved  $100 \frac{2.00027-2}{2} = 0.014\%$  error.

Example 3.6.4

Although it is possible to exactly find numbers like  $\sin\left(\frac{\pi}{8}\right)$  using geometry, we used a calculator to find the values in the above example. Once again, it's worth thinking about how a calculator "finds" values like this. One reason why we cover numerical integration is to give you a sense of where calculator answers might be coming from. (Many computer algebra systems are not transparent about how they do their computations, leaving users to guess—and hope that the programmers haven't left any bugs.)

So far, we have only found errors in approximations by comparing them to known values. But in general, if we know an exact value, we have no need of an approximation. One reason why Simpson's rule is useful is that we can say something about our error *without* knowing the exact value we're approximating. The next section explains how.

### 3.6.2 ▶ Error Behaviour

Now we are armed with a (relatively simple) method for numerical integration we should give thought to how practical it might be in the real world<sup>36</sup>. Two obvious considerations when deciding whether or not a given algorithm is of any practical value are

- the amount of computational effort required to execute the algorithm and
- the accuracy that this computational effort yields.

<sup>36</sup> Indeed, even beyond the "real world" of many applications in first year calculus texts, some of the methods we have described are used by actual people (such as ship builders, engineers and surveyors) to estimate areas and volumes of actual objects!

For algorithms like Simpson's rule, the bulk of the computational effort usually goes into evaluating the function  $f(x)$ . The number of evaluations of  $f(x)$  required for  $n$  steps of Simpson's rules is  $n + 1$ . So the amount of effort required is essentially one evaluation of  $f(x)$  per step. So if we double the value of  $n$ , we can expect to roughly double the computation time needed for an evaluation.

Let's revisit Example 3.6.4 to investigate<sup>37</sup> how the error changes as we increase  $n$ . In the table below, we approximated<sup>38</sup>  $\int_0^\pi \sin x dx$  using  $n$  intervals. Recall the exact value of this integral is 2.

n	approximation	error	# evals
10	2.00011	$1.1 \times 10^{-4}$	11
100	2.000000011	$1.1 \times 10^{-8}$	101
1000	2.0000000000011	$1.1 \times 10^{-12}$	1001

39

It seems that when we increase  $n$  by a factor of 10, we decrease the error by a factor of  $10^4$ . That is, it seems like the error

$$e_n = |\text{exact value} - \text{approximate value}|$$

with  $n$  steps is (roughly) of the form

$$e_n = K \frac{1}{n^4}$$

for some constant  $K \approx 1.1$ .

This intuition does indeed accord with the general behaviour of Simpson's rule when  $f(x)$  is reasonably smooth. A proof of the following theorem is beyond the scope of this course, but Appendix A.10.4 can give you in idea of what is involved.

37 For more detail, see Appendix A.10.3.

38 Even in this simple case, we notice the opacity of computer algebra systems. The author evaluated the  $n = 10$  approximation three ways: using Python, Open Office Calc, and WolframAlpha. Python returned 2.000109517315004. Open Office returned 2.0001095173, using fewer decimal places but agreeing in all of them. However, **WolframAlpha** gave 2.000006784441801 (accessed August 20, 2021). It seems that WolframAlpha interprets the number of intervals in a different way. It tracks the number of approximating parabolas, as opposed to the number of  $x$ -values where we evaluate our function. So, to get the same approximation as the other programs, we have to ask WolframAlpha for 5 intervals instead of 10. Then WolframAlpha gives 2.000109517315004, exactly agreeing with Python.

39 If you need a high degree of accuracy, it's good to know how many decimals your computer algebra system keeps track of. When the author asked Python to evaluate  $\sin(\pi)$ , it returned about  $1.2 \times 10^{-16}$  (instead of the actual value, 0). That's a good indicator that values close to or smaller than  $10^{-16}$  will not be accurately calculated (without telling the program to use more decimal places).

**Theorem 3.6.5** (Numerical integration errors).

Assume that  $|f^{(4)}(x)| \leq L$  for all  $a \leq x \leq b$ . Then the total error introduced by Simpson's rule when approximating  $\int_a^b f(x) dx$  is bounded by

$$\frac{L}{180} \frac{(b-a)^5}{n^4}$$

Here are some examples which illustrate the error bound is used. First, let us check that the above result is consistent with our observations about  $\int_0^\pi \sin x dx$ .

**Example 3.6.6** (Simpson's rule error approximating  $\int_0^\pi \sin x dx$ )

- The integral  $\int_0^\pi \sin x dx$  has  $b - a = \pi$ .
- To find  $L$ , we should first find the fourth derivative of  $f(x) = \sin x$ :

$$\begin{aligned} f'(x) &= \cos x \\ f''(x) &= -\sin x \\ f^{(3)}(x) &= -\cos x \\ f^{(4)}(x) &= \sin x \end{aligned}$$

So the fourth derivative of the integrand satisfies

$$\left| \frac{d^4}{dx^4} \sin x \right| = |\sin x| \leq 1$$

for all values of  $x$ . Therefore, we take  $L = 1$ .

- So the error,  $e_n$ , introduced when  $n$  steps are used is bounded by

$$\begin{aligned} |e_n| &\leq \frac{L}{180} \frac{(b-a)^5}{n^4} \\ &= \frac{1}{180} \frac{\pi^5}{n^4} \\ &\approx 1.7 \frac{1}{n^4} \end{aligned}$$

- Our guess for the error was around  $1.1 \frac{1}{n^4}$ , which is indeed less than  $1.7 \frac{1}{n^4}$ . Remember the bounds given in the theorem are worst-case scenarios. They represent the biggest the error could possibly be, but are not exactly the same as the actual error. So our experiments do accord with the theorem.

## Example 3.6.6

In a typical application we would be asked to evaluate a given integral to some specified accuracy. For example, if you are manufacturer and your machinery can only cut materials to an accuracy of  $\frac{1}{10}$ <sup>th</sup> of a millimeter, there is no point in making design specifications more accurate than  $\frac{1}{10}$ <sup>th</sup> of a millimeter.

## Example 3.6.7

Suppose, for example, that we wish to use Simpson's rule to evaluate<sup>40</sup>

$$\int_0^1 e^{-x^2} dx$$

to within an accuracy of  $10^{-6}$ .

*Solution.* In order to use Simpson's rule, we need to decide how many intervals to use. To find the number of intervals, we'll take the worst-case error from Theorem 3.6.5, set it to be less than  $10^{-6}$ , and solve for  $n$ . First, we'll need to find  $a$ ,  $b$ , and  $L$ .

- The integral has  $a = 0$  and  $b = 1$ .
- The first four derivatives of the integrand are:

$$\frac{d}{dx} [e^{-x^2}] = -2xe^{-x^2}$$

$$\frac{d^2}{dx^2} [e^{-x^2}] = \frac{d}{dx} [-2xe^{-x^2}] = -2e^{-x^2} + 4x^2e^{-x^2} = 2(2x^2 - 1)e^{-x^2}$$

$$\frac{d^3}{dx^3} [e^{-x^2}] = \frac{d}{dx} [2(2x^2 - 1)e^{-x^2}] = 2[(2x^2 - 1)(-2x)e^{-x^2} + 4xe^{-x^2}] = 4(-2x^3 + 3x)e^{-x^2}$$

$$\begin{aligned} \frac{d^4}{dx^4} [e^{-x^2}] &= \frac{d}{dx} [4(-2x^3 + 3x)e^{-x^2}] = 4[(-2x^3 + 3x)(-2x)e^{-x^2} + (-6x^2 + 3)e^{-x^2}] \\ &= 4(4x^4 - 12x^2 + 3)e^{-x^2} \end{aligned}$$

- As  $x$  runs from 0 to 1,  $2x^2 - 1$  increases from  $-1$  to 1, so that

$$0 \leq x \leq 1 \implies |2x^2 - 1| \leq 1, e^{-x^2} \leq 1 \implies |2(2x^2 - 1)e^{-x^2}| \leq 2$$

- Now, for any  $x$ ,  $e^{-x^2} \leq 1$ . Also, for  $0 \leq x \leq 1$ ,

$$0 \leq x^2, x^4 \leq 1 \quad \text{so}$$

$$3 \leq 4x^4 + 3 \leq 7 \quad \text{and}$$

$$-12 \leq -12x^2 \leq 0 \quad \text{adding these together gives}$$

$$-9 \leq 4x^4 - 12x^2 + 3 \leq 7$$

40 This is our favourite running example of an integral that cannot be evaluated algebraically — we need to use numerical methods.



Consequently,  $|4x^4 - 12x^2 + 3|$  is bounded by 9 and so

$$\left| \frac{d^4}{dx^4} e^{-x^2} \right| \leq 4 \times 9 = 36$$

So take  $L = 36$ .

- The error introduced by the  $n$  step Simpson's rule is at most

$$\begin{aligned} e_n &\leq \frac{L}{180} \frac{(b-a)^5}{n^4} \\ &\leq \frac{36}{180} \frac{(1-0)^5}{n^4} = \frac{1}{5n^4} \end{aligned}$$

- In order for this error to be no more than  $10^{-6}$  we require  $n$  to satisfy

$$\begin{aligned} e_n &\leq \frac{1}{5n^4} \leq 10^{-6} && \text{and so} \\ 5n^4 &\geq 10^6 \\ n^4 &\geq 200000 && \text{take fourth root} \\ n &\geq 21.15 \end{aligned}$$

So 22 steps of Simpson's rule will do the job.

- $n = 22$  steps actually results in an error of  $3.5 \times 10^{-8}$ . The reason that we get an error so much smaller than we need is that we have overestimated the number of steps required. This, in turn, occurred because we made quite a rough bound of  $\left| \frac{d^4}{dx^4} f(x) \right| \leq 36$ . If we are more careful then we will get a slightly smaller  $n$ . It actually turns out<sup>41</sup> that you only need  $n = 10$  to approximate within  $10^{-6}$ .

Example 3.6.7

## 3.7▲ Improper Integrals

### 3.7.1 ► Definitions

To this point we have only considered nicely behaved integrals  $\int_a^b f(x)dx$ . Though the algebra involved in some of our examples was quite difficult, all the integrals had

- finite limits of integration  $a$  and  $b$ , and
- a bounded integrand  $f(x)$  (and in fact continuous except possibly for finitely many jump discontinuities).

Not all integrals we need to study are quite so nice.

<sup>41</sup> The authors tested this empirically.

**Definition 3.7.1.**

An integral having either an infinite limit of integration or an unbounded integrand is called an improper integral.

Two examples are

$$\int_0^{\infty} \frac{dx}{1+x^2} \quad \text{and} \quad \int_0^1 \frac{dx}{x}$$

The first has an infinite domain of integration and the integrand of the second tends to  $\infty$  as  $x$  approaches the left end of the domain of integration. We'll start with an example that illustrates the traps that you can fall into if you treat such integrals sloppily. Then we'll see how to treat them carefully.

**Example 3.7.2**  $\left(\int_{-1}^1 \frac{1}{x^2} dx\right)$

Consider the integral

$$\int_{-1}^1 \frac{1}{x^2} dx$$

If we “do” this integral completely naively then we get

$$\begin{aligned} \int_{-1}^1 \frac{1}{x^2} dx &= \left. \frac{x^{-1}}{-1} \right|_{-1}^1 \\ &= \frac{1}{-1} - \frac{-1}{-1} \\ &= -2 \end{aligned}$$

which is *wrong*<sup>42</sup>. In fact, the answer is ridiculous. The integrand  $\frac{1}{x^2} > 0$ , so the integral has to be positive.

The flaw in the argument is that the Fundamental Theorem of Calculus, which says that

$$\text{if } F'(x) = f(x) \text{ then } \int_a^b f(x) dx = F(b) - F(a)$$

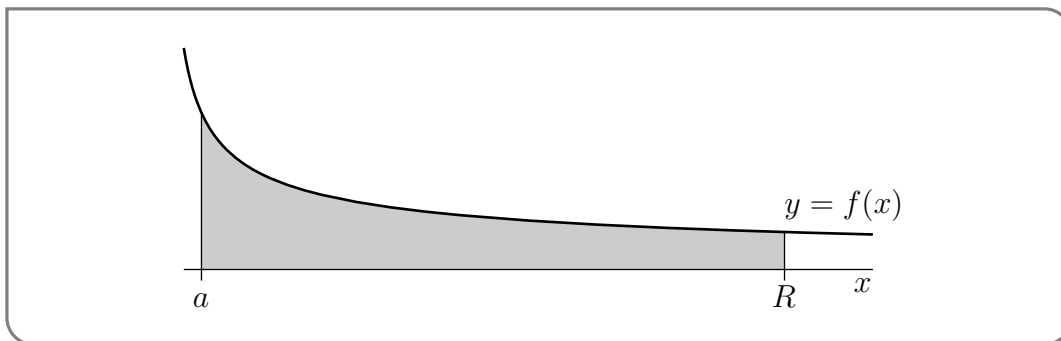
is applicable only when  $F'(x)$  exists and equals  $f(x)$  for *all*  $a \leq x \leq b$ . In this case  $F'(x) = \frac{1}{x^2}$  does not exist for  $x = 0$ . The given integral is improper. We'll see later that the correct answer is  $+\infty$ .

**Example 3.7.2**

Let us put this example to one side for a moment and turn to the integral  $\int_a^{\infty} \frac{dx}{1+x^2}$ . In this

42 Very wrong. But it is not an example of “not even wrong” — which is a phrase attributed to the physicist Wolfgang Pauli who was known for his harsh critiques of sloppy arguments. The phrase is typically used to describe arguments that are so incoherent that not only can one not prove they are true, but they lack enough coherence to be able to show they are false. The interested reader should do a little searchengineering and look at the concept of falsifiability.

case, the integrand is bounded but the domain of integration extends to  $+\infty$ . We can evaluate this integral by sneaking up on it. We compute it on a bounded domain of integration, like  $\int_a^R \frac{dx}{1+x^2}$ , and then take the limit  $R \rightarrow \infty$ . Let us put this into practice:



Example 3.7.3  $\left(\int_a^\infty \frac{dx}{1+x^2}\right)$

*Solution.*

- Since the domain extends to  $+\infty$  we first integrate on a finite domain

$$\begin{aligned}\int_a^R \frac{dx}{1+x^2} &= \arctan x \Big|_a^R \\ &= \arctan R - \arctan a\end{aligned}$$

- We then take the limit as  $R \rightarrow +\infty$ :

$$\begin{aligned}\int_a^\infty \frac{dx}{1+x^2} &= \lim_{R \rightarrow \infty} \int_a^R \frac{dx}{1+x^2} \\ &= \lim_{R \rightarrow \infty} [\arctan R - \arctan a] \\ &= \frac{\pi}{2} - \arctan a.\end{aligned}$$

Example 3.7.3

To be more precise, we actually formally *define* an integral with an infinite domain as the limit of the integral with a finite domain as we take one or more of the limits of integration to infinity.

**Definition 3.7.4** (Improper integral with infinite domain of integration).

(a) If the integral  $\int_a^R f(x) dx$  exists for all  $R > a$ , then

$$\int_a^\infty f(x) dx = \lim_{R \rightarrow \infty} \int_a^R f(x) dx$$

when the limit exists (and is finite).

(b) If the integral  $\int_r^b f(x) dx$  exists for all  $r < b$ , then

$$\int_{-\infty}^b f(x) dx = \lim_{r \rightarrow -\infty} \int_r^b f(x) dx$$

when the limit exists (and is finite).

(c) If the integral  $\int_r^R f(x) dx$  exists for all  $r < R$ , then

$$\int_{-\infty}^\infty f(x) dx = \lim_{r \rightarrow -\infty} \int_r^c f(x) dx + \lim_{R \rightarrow \infty} \int_c^R f(x) dx$$

when both limits exist (and are finite). Any  $c$  can be used.

When the limit(s) exist, the integral is said to be convergent. Otherwise it is said to be divergent.

We must also be able to treat an integral like  $\int_0^1 \frac{dx}{x}$  that has a finite domain of integration but whose integrand is unbounded near one limit of integration<sup>43</sup> Our approach is similar — we sneak up on the problem. We compute the integral on a smaller domain, such as  $\int_t^1 \frac{dx}{x}$ , with  $t > 0$ , and then take the limit  $t \rightarrow 0+$ .

**Example 3.7.5**  $\left(\int_0^1 \frac{1}{x} dx\right)$

*Solution.*

- Since the integrand is unbounded near  $x = 0$ , we integrate on the smaller domain  $t \leq x \leq 1$  with  $t > 0$ :

$$\int_t^1 \frac{1}{x} dx = \ln|x| \Big|_t^1 = -\ln|t|$$

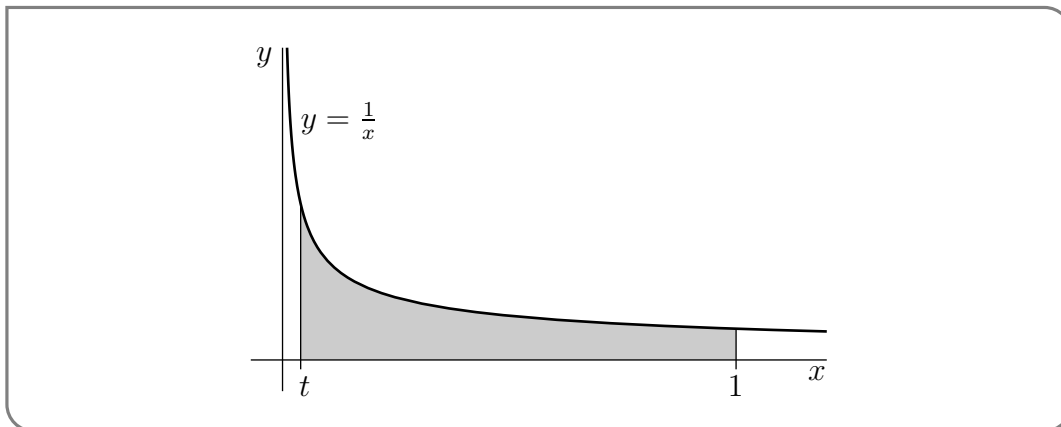
- We then take the limit as  $t \rightarrow 0+$  to obtain

$$\int_0^1 \frac{1}{x} dx = \lim_{t \rightarrow 0+} \int_t^1 \frac{1}{x} dx = \lim_{t \rightarrow 0+} -\ln|t| = +\infty$$

<sup>43</sup> This will, in turn, allow us to deal with integrals whose integrand is unbounded somewhere inside the domain of integration.

Thus this integral diverges to  $+\infty$ .

Example 3.7.5



Indeed, we *define* integrals with unbounded integrands via this process:

**Definition 3.7.6** (Improper integral with unbounded integrand).

(a) If the integral  $\int_t^b f(x) dx$  exists for all  $a < t < b$ , then

$$\int_a^b f(x) dx = \lim_{t \rightarrow a^+} \int_t^b f(x) dx$$

when the limit exists (and is finite).

(b) If the integral  $\int_a^T f(x) dx$  exists for all  $a < T < b$ , then

$$\int_a^b f(x) dx = \lim_{T \rightarrow b^-} \int_a^T f(x) dx$$

when the limit exists (and is finite).

(c) Let  $a < c < b$ . If the integrals  $\int_a^T f(x) dx$  and  $\int_t^b f(x) dx$  exist for all  $a < T < c$  and  $c < t < b$ , then

$$\int_a^b f(x) dx = \lim_{T \rightarrow c^-} \int_a^T f(x) dx + \lim_{t \rightarrow c^+} \int_t^b f(x) dx$$

when both limit exist (and are finite).

When the limit(s) exist, the integral is said to be convergent. Otherwise it is said to be divergent.

Notice that (c) is used when the integrand is unbounded at some point in the middle of the domain of integration, such as was the case in our original example

$$\int_{-1}^1 \frac{1}{x^2} dx$$

A quick computation shows that this integral diverges to  $+\infty$

$$\begin{aligned} \int_{-1}^1 \frac{1}{x^2} dx &= \lim_{a \rightarrow 0^-} \int_{-1}^a \frac{1}{x^2} dx + \lim_{b \rightarrow 0^+} \int_b^1 \frac{1}{x^2} dx \\ &= \lim_{a \rightarrow 0^-} \left[ 1 - \frac{1}{a} \right] + \lim_{b \rightarrow 0^+} \left[ \frac{1}{b} - 1 \right] \\ &= +\infty \end{aligned}$$

More generally, if an integral has more than one “source of impropriety” (for example an infinite domain of integration and an integrand with an unbounded integrand or multiple infinite discontinuities) then you split it up into a sum of integrals with a single “source of impropriety” in each. For the integral, as a whole, to converge every term in that sum has to converge.

For example

Example 3.7.7  $\left( \int_{-\infty}^{\infty} \frac{dx}{(x-2)x^2} \right)$

Consider the integral

$$\int_{-\infty}^{\infty} \frac{dx}{(x-2)x^2}$$

- The domain of integration that extends to both  $+\infty$  and  $-\infty$ .
- The integrand is singular (i.e. becomes infinite) at  $x = 2$  and at  $x = 0$ .
- So we would write the integral as

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{dx}{(x-2)x^2} &= \int_{-\infty}^a \frac{dx}{(x-2)x^2} + \int_a^0 \frac{dx}{(x-2)x^2} + \int_0^b \frac{dx}{(x-2)x^2} \\ &\quad + \int_b^2 \frac{dx}{(x-2)x^2} + \int_2^c \frac{dx}{(x-2)x^2} + \int_c^{\infty} \frac{dx}{(x-2)x^2} \end{aligned}$$

where

- $a$  is any number strictly less than 0,
- $b$  is any number strictly between 0 and 2, and
- $c$  is any number strictly bigger than 2.

So, for example, take  $a = -1, b = 1, c = 3$ .

- When we examine the right-hand side we see that

- the first integral has domain of integration extending to  $-\infty$
  - the second integral has an integrand that becomes unbounded as  $x \rightarrow 0-$ ,
  - the third integral has an integrand that becomes unbounded as  $x \rightarrow 0+$ ,
  - the fourth integral has an integrand that becomes unbounded as  $x \rightarrow 2-$ ,
  - the fifth integral has an integrand that becomes unbounded as  $x \rightarrow 2+$ , and
  - the last integral has domain of integration extending to  $+\infty$ .
- Each of these integrals can then be expressed as a limit of an integral on a small domain.

Example 3.7.7

### 3.7.2 ▶ Examples

With the more formal definitions out of the way, we are now ready for some (important) examples.

Example 3.7.8  $\left(\int_1^\infty \frac{dx}{x^p} \text{ with } p > 0\right)$

*Solution.*

- Fix any  $p > 0$ .
- The domain of the integral  $\int_1^\infty \frac{dx}{x^p}$  extends to  $+\infty$  and the integrand  $\frac{1}{x^p}$  is continuous and bounded on the whole domain.
- So we write this integral as the limit

$$\int_1^\infty \frac{dx}{x^p} = \lim_{R \rightarrow \infty} \int_1^R \frac{dx}{x^p}$$

- The antiderivative of  $1/x^p$  changes when  $p = 1$ , so we will split the problem into three cases,  $p > 1$ ,  $p = 1$  and  $p < 1$ .
- When  $p > 1$ ,

$$\begin{aligned} \int_1^R \frac{dx}{x^p} &= \frac{1}{1-p} x^{1-p} \Big|_1^R \\ &= \frac{R^{1-p} - 1}{1-p} \end{aligned}$$

Taking the limit as  $R \rightarrow \infty$  gives

$$\begin{aligned} \int_1^\infty \frac{dx}{x^p} &= \lim_{R \rightarrow \infty} \int_1^R \frac{dx}{x^p} \\ &= \lim_{R \rightarrow \infty} \frac{R^{1-p} - 1}{1-p} \\ &= \frac{-1}{1-p} = \frac{1}{p-1} \end{aligned}$$

since  $1 - p < 0$ .

- Similarly when  $p < 1$  we have

$$\int_1^\infty \frac{dx}{x^p} = \lim_{R \rightarrow \infty} \int_1^R \frac{dx}{x^p} = \lim_{R \rightarrow \infty} \frac{R^{1-p} - 1}{1-p} = +\infty$$

because  $1 - p > 0$  and the term  $R^{1-p}$  diverges to  $+\infty$ .

- Finally when  $p = 1$

$$\int_1^R \frac{dx}{x} = \ln |R| - \ln 0 = \ln R$$

Then taking the limit as  $R \rightarrow \infty$  gives us

$$\int_1^\infty \frac{dx}{x^p} = \lim_{R \rightarrow \infty} \ln |R| = +\infty.$$

- So summarising, we have

$$\int_1^\infty \frac{dx}{x^p} = \begin{cases} \text{divergent} & \text{if } p \leq 1 \\ \frac{1}{p-1} & \text{if } p > 1 \end{cases}$$

Example 3.7.8

Example 3.7.9 ( $\int_0^1 \frac{dx}{x^p}$  with  $p > 0$ )

*Solution.*

- Again fix any  $p > 0$ .
- The domain of integration of the integral  $\int_0^1 \frac{dx}{x^p}$  is finite, but the integrand  $\frac{1}{x^p}$  becomes unbounded as  $x$  approaches the left end, 0, of the domain of integration.
- So we write this integral as

$$\int_0^1 \frac{dx}{x^p} = \lim_{t \rightarrow 0^+} \int_t^1 \frac{dx}{x^p}$$

- Again, the antiderivative changes at  $p = 1$ , so we split the problem into three cases.
- When  $p > 1$  we have

$$\begin{aligned} \int_t^1 \frac{dx}{x^p} &= \frac{1}{1-p} x^{1-p} \Big|_t^1 \\ &= \frac{1 - t^{1-p}}{1-p} \end{aligned}$$



Since  $1 - p < 0$  when we take the limit as  $t \rightarrow 0$  the term  $t^{1-p}$  diverges to  $+\infty$  and we obtain

$$\int_0^1 \frac{dx}{x^p} = \lim_{t \rightarrow 0^+} \frac{1 - t^{1-p}}{1 - p} = +\infty$$

- When  $p = 1$  we similarly obtain

$$\begin{aligned} \int_0^1 \frac{dx}{x} &= \lim_{t \rightarrow 0^+} \int_t^1 \frac{dx}{x} \\ &= \lim_{t \rightarrow 0^+} (-\ln |t|) \\ &= +\infty \end{aligned}$$

- Finally, when  $p < 1$  we have

$$\begin{aligned} \int_0^1 \frac{dx}{x^p} &= \lim_{t \rightarrow 0^+} \int_t^1 \frac{dx}{x^p} \\ &= \lim_{t \rightarrow 0^+} \frac{1 - t^{1-p}}{1 - p} = \frac{1}{1 - p} \end{aligned}$$

since  $1 - p > 0$ .

- In summary

$$\int_0^1 \frac{dx}{x^p} = \begin{cases} \frac{1}{1-p} & \text{if } p < 1 \\ \text{divergent} & \text{if } p \geq 1 \end{cases}$$

Example 3.7.9

Example 3.7.10  $\left( \int_0^\infty \frac{dx}{x^p} \text{ with } p > 0 \right)$

*Solution.*

- Yet again fix  $p > 0$ .
- This time the domain of integration of the integral  $\int_0^\infty \frac{dx}{x^p}$  extends to  $+\infty$ , and in addition the integrand  $\frac{1}{x^p}$  becomes unbounded as  $x$  approaches the left end, 0, of the domain of integration.
- So we split the domain in two — given our last two examples, the obvious place to cut is at  $x = 1$ :

$$\int_0^\infty \frac{dx}{x^p} = \int_0^1 \frac{dx}{x^p} + \int_1^\infty \frac{dx}{x^p}$$

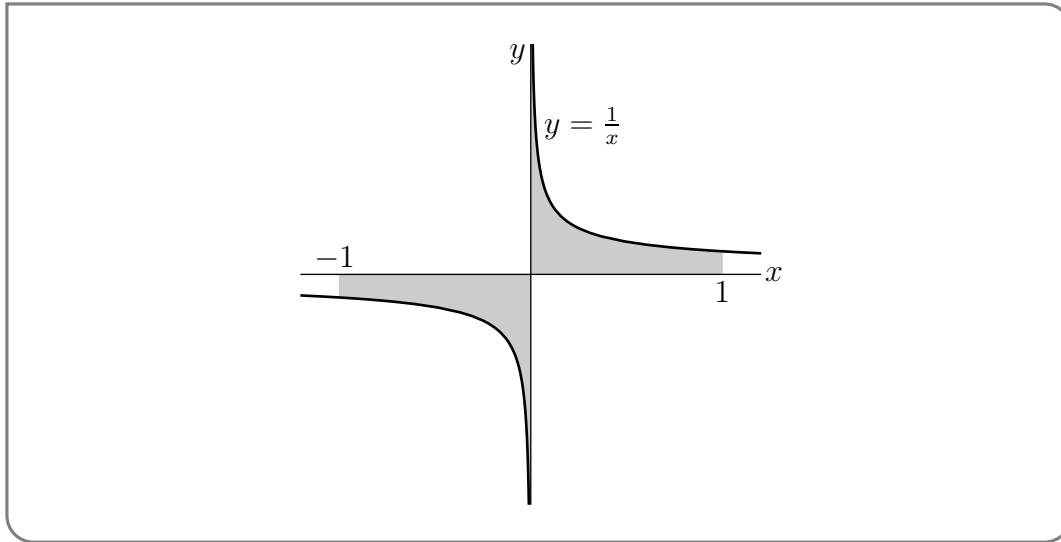
- We saw, in Example 3.7.9, that the first integral diverged whenever  $p \geq 1$ , and we also saw, in Example 3.7.8, that the second integral diverged whenever  $p \leq 1$ .

- So the integral  $\int_0^{\infty} \frac{dx}{x^p}$  diverges for all values of  $p$ .

Example 3.7.10

Example 3.7.11  $\left(\int_{-1}^1 \frac{dx}{x}\right)$

This is a pretty subtle example. Look at the sketch below: This suggests that the signed



area to the left of the  $y$ -axis should exactly cancel the area to the right of the  $y$ -axis making the value of the integral  $\int_{-1}^1 \frac{dx}{x}$  exactly zero.

But both of the integrals

$$\int_0^1 \frac{dx}{x} = \lim_{t \rightarrow 0^+} \int_t^1 \frac{dx}{x} = \lim_{t \rightarrow 0^+} [\ln x]_t^1 = \lim_{t \rightarrow 0^+} \ln \frac{1}{t} = +\infty$$

$$\int_{-1}^0 \frac{dx}{x} = \lim_{T \rightarrow 0^-} \int_{-1}^T \frac{dx}{x} = \lim_{T \rightarrow 0^-} [\ln |x|]_{-1}^T = \lim_{T \rightarrow 0^-} \ln |T| = -\infty$$

diverge so  $\int_{-1}^1 \frac{dx}{x}$  diverges. Don't make the mistake of thinking that  $\infty - \infty = 0$ . It is undefined. And it is undefined for good reason.

For example, we have just seen that the area to the right of the  $y$ -axis is

$$\lim_{t \rightarrow 0^+} \int_t^1 \frac{dx}{x} = +\infty$$

and that the area to the left of the  $y$ -axis is (substitute  $-7t$  for  $T$  above)

$$\lim_{t \rightarrow 0^+} \int_{-1}^{-7t} \frac{dx}{x} = -\infty$$

If  $\infty - \infty = 0$ , the following limit should be 0.

$$\begin{aligned} \lim_{t \rightarrow 0^+} \left[ \int_t^1 \frac{dx}{x} + \int_{-1}^{-7t} \frac{dx}{x} \right] &= \lim_{t \rightarrow 0^+} \left[ \ln \frac{1}{t} + \ln |-7t| \right] \\ &= \lim_{t \rightarrow 0^+} \left[ \ln \frac{1}{t} + \ln(7t) \right] \\ &= \lim_{t \rightarrow 0^+} \left[ -\ln t + \ln 7 + \ln t \right] = \lim_{t \rightarrow 0^+} \ln 7 \\ &= \ln 7 \end{aligned}$$

This appears to give  $\infty - \infty = \ln 7$ . Of course the number 7 was picked arbitrarily. You can make  $\infty - \infty$  be any number at all, by making a suitable replacement for 7.

Example 3.7.11

Example 3.7.12 (Example 3.7.2 revisited)

The careful computation of the integral of Example 3.7.2 is

$$\begin{aligned} \int_{-1}^1 \frac{1}{x^2} dx &= \lim_{T \rightarrow 0^-} \int_{-1}^T \frac{1}{x^2} dx + \lim_{t \rightarrow 0^+} \int_t^1 \frac{1}{x^2} dx \\ &= \lim_{T \rightarrow 0^-} \left[ -\frac{1}{x} \right]_{-1}^T + \lim_{t \rightarrow 0^+} \left[ -\frac{1}{x} \right]_t^1 \\ &= \infty + \infty \end{aligned}$$

Hence the integral diverges to  $+\infty$ .

Example 3.7.12

**Warning 3.7.13 (Sneaky Divergence).**

If you don't realize that an integral diverges, you can generate answers that look plausible but are secretly nonsense. For example, attempting to use the Fundamental Theorem of Calculus on Example 3.7.2 gives  $\int_{-1}^1 \frac{1}{x^2} dx$  as  $-2$ : a poor approximation for positive infinity.

This mistake can be especially dangerous using computer algebra systems, where you spend less time thinking about the integral and so have fewer chances to notice that something is awry. As of this writing,<sup>44</sup> **WolframAlpha** gives no warnings when you ask it to approximate  $\int_{-1}^1 \frac{1}{x^2} dx$  using Simpson's Rule: it tells you the approximation with one interval is  $\frac{2}{3}$ .

Example 3.7.14  $\left( \int_{-\infty}^{\infty} \frac{dx}{1+x^2} \right)$

Since

$$\begin{aligned} \lim_{R \rightarrow \infty} \int_0^R \frac{dx}{1+x^2} &= \lim_{R \rightarrow \infty} [\arctan x]_0^R = \lim_{R \rightarrow \infty} \arctan R = \frac{\pi}{2} \\ \lim_{r \rightarrow -\infty} \int_r^0 \frac{dx}{1+x^2} &= \lim_{r \rightarrow -\infty} [\arctan x]_r^0 = \lim_{r \rightarrow -\infty} -\arctan r = \frac{\pi}{2} \end{aligned}$$

The integral  $\int_{-\infty}^{\infty} \frac{dx}{1+x^2}$  converges and takes the value  $\pi$ .

Example 3.7.14

Example 3.7.15

For what values of  $p$  does  $\int_e^{\infty} \frac{dx}{x(\ln x)^p}$  converge?

*Solution.*

- For  $x \geq e$ , the denominator  $x(\ln x)^p$  is never zero. So the integrand is bounded on the entire domain of integration and this integral is improper only because the domain of integration extends to  $+\infty$  and we proceed as usual.
- We have

$$\begin{aligned} \int_e^{\infty} \frac{dx}{x(\ln x)^p} &= \lim_{R \rightarrow \infty} \int_e^R \frac{dx}{x(\ln x)^p} && \text{use substitution} \\ &= \lim_{R \rightarrow \infty} \int_1^{\ln R} \frac{du}{u^p} && \text{with } u = \ln x, du = \frac{dx}{x} \\ &= \lim_{R \rightarrow \infty} \begin{cases} \frac{1}{1-p} [(\ln R)^{1-p} - 1] & \text{if } p \neq 1 \\ \ln(\ln R) & \text{if } p = 1 \end{cases} \\ &= \begin{cases} \text{divergent} & \text{if } p \leq 1 \\ \frac{1}{p-1} & \text{if } p > 1 \end{cases} \end{aligned}$$

In this last step we have used similar logic that that used in Example 3.7.8, but with  $R$  replaced by  $\ln R$ .

Example 3.7.15

Example 3.7.16 (the gamma function)

The gamma function  $\Gamma(x)$  is defined by the improper integral

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$

We shall now compute  $\Gamma(n)$  for all natural numbers  $n$ .

- To get started, we'll compute

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = \lim_{R \rightarrow \infty} \int_0^R e^{-x} dx = \lim_{R \rightarrow \infty} \left[ -e^{-x} \right]_0^R = 1$$

- Then compute

$$\begin{aligned} \Gamma(2) &= \int_0^{\infty} x e^{-x} dx \\ &= \lim_{R \rightarrow \infty} \int_0^R x e^{-x} dx && \text{use integration by parts with} \\ & && u = x, dv = e^{-x} dx, \\ & && v = -e^{-x}, du = dx \\ &= \lim_{R \rightarrow \infty} \left[ -x e^{-x} \Big|_0^R + \int_0^R e^{-x} dx \right] \\ &= \lim_{R \rightarrow \infty} \left[ -x e^{-x} - e^{-x} \right]_0^R \\ &= 1 \end{aligned}$$

For the last equality, we used that  $\lim_{x \rightarrow \infty} x e^{-x} = 0$ .

- Now we move on to general  $n$ , using the same type of computation as we just used to evaluate  $\Gamma(2)$ . For any natural number  $n$ ,

$$\begin{aligned} \Gamma(n+1) &= \int_0^{\infty} x^n e^{-x} dx \\ &= \lim_{R \rightarrow \infty} \int_0^R x^n e^{-x} dx && \text{again integrate by parts with} \\ & && u = x^n, dv = e^{-x} dx, \\ & && v = -e^{-x}, du = n x^{n-1} dx \\ &= \lim_{R \rightarrow \infty} \left[ -x^n e^{-x} \Big|_0^R + \int_0^R n x^{n-1} e^{-x} dx \right] \\ &= \lim_{R \rightarrow \infty} n \int_0^R x^{n-1} e^{-x} dx \\ &= n \Gamma(n) \end{aligned}$$

To get to the third row, we used that  $\lim_{x \rightarrow \infty} x^n e^{-x} = 0$ .

- Now that we know  $\Gamma(2) = 1$  and  $\Gamma(n+1) = n\Gamma(n)$ , for all  $n \in \mathbb{N}$ , we can compute

all of the  $\Gamma(n)$ 's.

$$\begin{aligned} \Gamma(2) &= 1 \\ \Gamma(3) &= \Gamma(2 + 1) = 2\Gamma(2) = 2 \cdot 1 \\ \Gamma(4) &= \Gamma(3 + 1) = 3\Gamma(3) = 3 \cdot 2 \cdot 1 \\ \Gamma(5) &= \Gamma(4 + 1) = 4\Gamma(4) = 4 \cdot 3 \cdot 2 \cdot 1 \\ &\vdots \\ \Gamma(n) &= (n - 1) \cdot (n - 2) \cdots 4 \cdot 3 \cdot 2 \cdot 1 = (n - 1)! \end{aligned}$$

That is, the factorial is just<sup>45</sup> the Gamma function shifted by one.

Example 3.7.16

### 3.7.3 ▶ Convergence Tests for Improper Integrals

It is very common to encounter integrals that are too complicated to evaluate explicitly. Numerical approximation schemes, evaluated by computer, are often used instead (see Section 3.6). You want to be sure that at least the integral converges before feeding it into a computer<sup>46</sup>. Fortunately it is usually possible to determine whether or not an improper integral converges even when you cannot evaluate it explicitly.

**Remark 3.7.17.** For pedagogical purposes, we are going to concentrate on the problem of determining whether or not an integral  $\int_a^\infty f(x) \, dx$  converges, when  $f(x)$  has no singularities for  $x \geq a$ . Recall that the first step in analyzing any improper integral is to write it as a sum of integrals each of has only a single “source of impropriety” — either a domain of integration that extends to  $+\infty$ , or a domain of integration that extends to  $-\infty$ , or an integrand which is singular at one end of the domain of integration. So we are now going to consider only the first of these three possibilities. But the techniques that we are about to see have obvious analogues for the other two possibilities.

Now let’s start. Imagine that we have an improper integral  $\int_a^\infty f(x) \, dx$ , that  $f(x)$  has no singularities for  $x \geq a$  and that  $f(x)$  is complicated enough that we cannot evaluate the integral explicitly<sup>47</sup>. The idea is find another improper integral  $\int_a^\infty g(x) \, dx$

- with  $g(x)$  simple enough that we can evaluate the integral  $\int_a^\infty g(x) \, dx$  explicitly, or at least determine easily whether or not  $\int_a^\infty g(x) \, dx$  converges, and

45 The Gamma function is far more important than just a generalisation of the factorial. It appears all over mathematics, physics, statistics and beyond. It has all sorts of interesting properties and its definition can be extended from natural numbers  $n$  to all numbers excluding  $0, -1, -2, -3, \dots$ . For example, one can show that

$$\Gamma(1 - z)\Gamma(z) = \frac{\pi}{\sin \pi z}.$$

46 Applying numerical integration methods to a divergent integral may result in perfectly reasonably looking but very wrong answers.

47 You could, for example, think of something like our running example  $\int_a^\infty e^{-t^2} \, dt$ .

- with  $g(x)$  behaving enough like  $f(x)$  for large  $x$  that the integral  $\int_a^\infty f(x) dx$  converges if and only if  $\int_a^\infty g(x) dx$  converges.

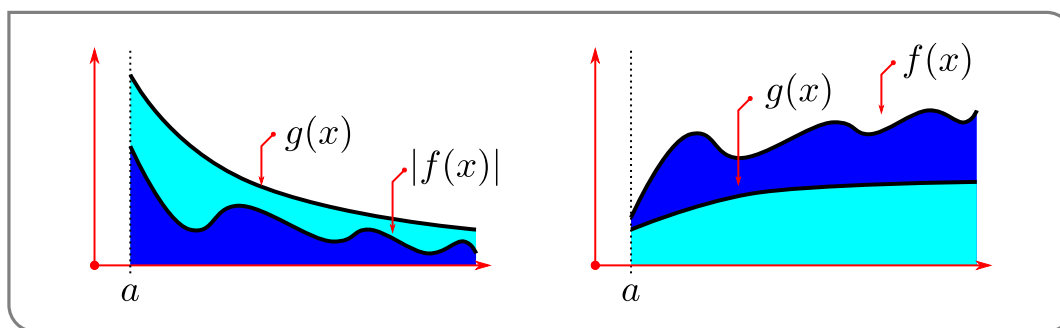
So far, this is a pretty vague strategy. Here is a theorem which starts to make it more precise.

**Theorem 3.7.18 (Comparison).**

Let  $a$  be a real number. Let  $f$  and  $g$  be functions that are defined and continuous for all  $x \geq a$  and assume that  $g(x) \geq 0$  for all  $x \geq a$ .

- (a) If  $|f(x)| \leq g(x)$  for all  $x \geq a$  and if  $\int_a^\infty g(x) dx$  converges then  $\int_a^\infty f(x) dx$  also converges.
- (b) If  $f(x) \geq g(x)$  for all  $x \geq a$  and if  $\int_a^\infty g(x) dx$  diverges then  $\int_a^\infty f(x) dx$  diverges.

We will not prove this theorem, but, hopefully, the following supporting arguments should at least appear reasonable to you. Consider the figure below:



- If  $\int_a^\infty g(x) dx$  converges, then the area of

$$\{ (x, y) \mid x \geq a, 0 \leq y \leq g(x) \} \text{ is finite.}$$

When  $|f(x)| \leq g(x)$ , the region

$$\{ (x, y) \mid x \geq a, 0 \leq y \leq |f(x)| \} \text{ is contained inside } \{ (x, y) \mid x \geq a, 0 \leq y \leq g(x) \}$$

and so must also have finite area. Consequently the areas of both the regions

$$\{ (x, y) \mid x \geq a, 0 \leq y \leq f(x) \} \text{ and } \{ (x, y) \mid x \geq a, f(x) \leq y \leq 0 \}$$

are finite too<sup>48</sup>.

48 We have separated the regions in which  $f(x)$  is positive and negative, because the integral  $\int_a^\infty f(x) dx$  represents the signed area of the union of  $\{ (x, y) \mid x \geq a, 0 \leq y \leq f(x) \}$  and  $\{ (x, y) \mid x \geq a, f(x) \leq y \leq 0 \}$ .

- If  $\int_a^\infty g(x) dx$  diverges, then the area of

$$\{ (x, y) \mid x \geq a, 0 \leq y \leq g(x) \} \text{ is infinite.}$$

When  $f(x) \geq g(x)$ , the region

$$\{ (x, y) \mid x \geq a, 0 \leq y \leq f(x) \} \text{ contains the region } \{ (x, y) \mid x \geq a, 0 \leq y \leq g(x) \}$$

and so also has infinite area.

Example 3.7.19  $\left( \int_1^\infty e^{-x^2} dx \right)$

We cannot evaluate the integral  $\int_1^\infty e^{-x^2} dx$  explicitly<sup>49</sup>, however we would still like to understand if it is finite or not — does it converge or diverge?

*Solution.* We will use Theorem 3.7.18 to answer the question.

- So we want to find another integral that we can compute and that we can compare to  $\int_1^\infty e^{-x^2} dx$ . To do so we pick an integrand that looks like  $e^{-x^2}$ , but whose indefinite integral we know — such as  $e^{-x}$ .
- When  $x \geq 1$ , we have  $x^2 \geq x$  and hence  $e^{-x^2} \leq e^{-x}$ . Thus we can use Theorem 3.7.18 to compare

$$\int_1^\infty e^{-x^2} dx \text{ with } \int_1^\infty e^{-x} dx$$

- The integral

$$\begin{aligned} \int_1^\infty e^{-x} dx &= \lim_{R \rightarrow \infty} \int_1^R e^{-x} dx \\ &= \lim_{R \rightarrow \infty} \left[ -e^{-x} \right]_1^R \\ &= \lim_{R \rightarrow \infty} \left[ e^{-1} - e^{-R} \right] = e^{-1} \end{aligned}$$

converges.

- So, by Theorem 3.7.18, with  $a = 1$ ,  $f(x) = e^{-x^2}$  and  $g(x) = e^{-x}$ , the integral  $\int_1^\infty e^{-x^2} dx$  converges too (it is approximately equal to 0.1394).

Example 3.7.19

Example 3.7.20  $\left( \int_{1/2}^\infty e^{-x^2} dx \right)$

*Solution.*

<sup>49</sup> It has been the subject of many remarks and footnotes.



- The integral  $\int_{1/2}^{\infty} e^{-x^2} dx$  is quite similar to the integral  $\int_1^{\infty} e^{-x^2} dx$  of Example 3.7.19. But we cannot just repeat the argument of Example 3.7.19 because it is not true that  $e^{-x^2} \leq e^{-x}$  when  $0 < x < 1$ .
- In fact, for  $0 < x < 1$ ,  $x^2 < x$  so that  $e^{-x^2} > e^{-x}$ .
- However the difference between the current example and Example 3.7.19 is

$$\int_{1/2}^{\infty} e^{-x^2} dx - \int_1^{\infty} e^{-x^2} dx = \int_{1/2}^1 e^{-x^2} dx$$

which is clearly a well defined finite number (its actually about 0.286). It is important to note that we are being a little sloppy by taking the difference of two integrals like this — we are assuming that both integrals converge. More on this below.

- So we would expect that  $\int_{1/2}^{\infty} e^{-x^2} dx$  should be the sum of the proper integral  $\int_{1/2}^1 e^{-x^2} dx$  and the convergent integral  $\int_1^{\infty} e^{-x^2} dx$  and so should be a convergent integral. This is indeed the case. The Theorem below provides the justification.

Example 3.7.20

**Theorem 3.7.21.**

Let  $a$  and  $c$  be real numbers with  $a < c$  and let the function  $f(x)$  be continuous for all  $x \geq a$ . Then the improper integral  $\int_a^{\infty} f(x) dx$  converges if and only if the improper integral  $\int_c^{\infty} f(x) dx$  converges.

*Proof.* By definition the improper integral  $\int_a^{\infty} f(x) dx$  converges if and only if the limit

$$\begin{aligned} \lim_{R \rightarrow \infty} \int_a^R f(x) dx &= \lim_{R \rightarrow \infty} \left[ \int_a^c f(x) dx + \int_c^R f(x) dx \right] \\ &= \int_a^c f(x) dx + \lim_{R \rightarrow \infty} \int_c^R f(x) dx \end{aligned}$$

exists and is finite. (Remember that, in computing the limit,  $\int_a^c f(x) dx$  is a finite constant independent of  $R$  and so can be pulled out of the limit.) But that is the case if and only if the limit  $\lim_{R \rightarrow \infty} \int_c^R f(x) dx$  exists and is finite, which in turn is the case if and only if the integral  $\int_c^{\infty} f(x) dx$  converges.  $\square$

Example 3.7.22

Does the integral  $\int_1^{\infty} \frac{\sqrt{x}}{x^2+x} dx$  converge or diverge?

*Solution.*

- Our first task is to identify the potential sources of impropriety for this integral.

- The domain of integration extends to  $+\infty$ , but we must also check to see if the integrand contains any singularities. On the domain of integration  $x \geq 1$  so the denominator is never zero and the integrand is continuous. So the only problem is at  $+\infty$ .
- Our second task is to develop some intuition<sup>50</sup>. As the only problem is that the domain of integration extends to infinity, whether or not the integral converges will be determined by the behavior of the integrand for very large  $x$ .
- When  $x$  is very large,  $x^2$  is much much larger than  $x$  (which we can write as  $x^2 \gg x$ ) so that the denominator  $x^2 + x \approx x^2$  and the integrand

$$\frac{\sqrt{x}}{x^2 + x} \approx \frac{\sqrt{x}}{x^2} = \frac{1}{x^{3/2}}$$

- By Example 3.7.8, with  $p = 3/2$ , the integral  $\int_1^\infty \frac{dx}{x^{3/2}}$  converges. So we would expect that  $\int_1^\infty \frac{\sqrt{x}}{x^2+x} dx$  converges too.
- Our final task is to verify that our intuition is correct. To do so, we want to apply part (a) of Theorem 3.7.18 with  $f(x) = \frac{\sqrt{x}}{x^2+x}$  and  $g(x)$  being  $\frac{1}{x^{3/2}}$ , or possibly some constant times  $\frac{1}{x^{3/2}}$ . That is, we need to show that for all  $x \geq 1$  (i.e. on the domain of integration)

$$\frac{\sqrt{x}}{x^2 + x} \leq \frac{A}{x^{3/2}}$$

for some constant  $A$ . Let's try this.

- Since  $x \geq 1$  we know that

$$x^2 + x > x^2$$

Now take the reciprocal of both sides:

$$\frac{1}{x^2 + x} < \frac{1}{x^2}$$

Multiply both sides by  $\sqrt{x}$  (which is always positive, so the sign of the inequality does not change)

$$\frac{\sqrt{x}}{x^2 + x} < \frac{\sqrt{x}}{x^2} = \frac{1}{x^{3/2}}$$

- So Theorem 3.7.18(a) and Example 3.7.8, with  $p = 3/2$  do indeed show that the integral  $\int_1^\infty \frac{\sqrt{x}}{x^2+x} dx$  converges.

---

<sup>50</sup> This takes practice, practice and more practice. At the risk of alliteration — please perform plenty of practice problems.


 Example 3.7.22

Notice that in this last example we managed to show that the integral exists by finding an integrand that behaved the same way for large  $x$ . Our intuition then had to be bolstered with some careful inequalities to apply the comparison Theorem 3.7.18. It would be nice to avoid this last step and be able jump from the intuition to the conclusion without messing around with inequalities. Thankfully there is a variant of Theorem 3.7.18 that is often easier to apply and that also fits well with the sort of intuition that we developed to solve Example 3.7.22.

A key phrase in the previous paragraph is “behaves the same way for large  $x$ ”. A good way to formalise this expression — “ $f(x)$  behaves like  $g(x)$  for large  $x$ ” — is to require that the limit

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} \text{ exists and is a finite nonzero number.}$$

Suppose that this is the case and call the limit  $L \neq 0$ . Then

- the ratio  $\frac{f(x)}{g(x)}$  must approach  $L$  as  $x$  tends to  $+\infty$ .
- So when  $x$  is very large — say  $x > B$ , for some big number  $B$  — we must have that

$$\frac{1}{2}L \leq \frac{f(x)}{g(x)} \leq 2L \quad \text{for all } x > B$$

Equivalently,  $f(x)$  lies between  $\frac{1}{2}Lg(x)$  and  $2Lg(x)$ , for all  $x \geq B$ .

- Consequently, the integral of  $f(x)$  converges if and only if the integral of  $g(x)$  converges, by Theorems 3.7.18 and 3.7.21.

These considerations lead to the following variant of Theorem 3.7.18.

**Theorem 3.7.23** (Limiting comparison).

Let  $-\infty < a < \infty$ . Let  $f$  and  $g$  be functions that are defined and continuous for all  $x \geq a$  and assume that  $g(x) \geq 0$  for all  $x \geq a$ .

(a) If  $\int_a^\infty g(x) \, dx$  converges and the limit

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)}$$

exists, then  $\int_a^\infty f(x) \, dx$  converges.

(b) If  $\int_a^\infty g(x) \, dx$  diverges and the limit

$$\lim_{x \rightarrow \infty} \frac{f(x)}{g(x)}$$

exists and is nonzero, then  $\int_a^\infty f(x) \, dx$  diverges.

Note that in (b) the limit must exist and be nonzero, while in (a) we only require that the limit exists (it can be zero).

Here is an example of how Theorem 3.7.23 is used.

Example 3.7.24  $\left( \int_1^\infty \frac{x + \sin x}{e^{-x} + x^2} \, dx \right)$

Does the integral  $\int_1^\infty \frac{x + \sin x}{e^{-x} + x^2} \, dx$  converge or diverge?

*Solution.*

- Our first task is to identify the potential sources of impropriety for this integral.
- The domain of integration extends to  $+\infty$ . On the domain of integration the denominator is never zero so the integrand is continuous. Thus the only problem is at  $+\infty$ .
- Our second task is to develop some intuition about the behavior of the integrand for very large  $x$ . A good way to start is to think about the size of each term when  $x$  becomes big.
- When  $x$  is very large:
  - $e^{-x} \ll x^2$ , so that the denominator  $e^{-x} + x^2 \approx x^2$ , and
  - $|\sin x| \leq 1 \ll x$ , so that the numerator  $x + \sin x \approx x$ , and
  - the integrand  $\frac{x + \sin x}{e^{-x} + x^2} \approx \frac{x}{x^2} = \frac{1}{x}$ .

Notice that we are using  $A \ll B$  to mean that “ $A$  is much much smaller than  $B$ ”. Similarly  $A \gg B$  means “ $A$  is much much bigger than  $B$ ”. We don’t really need to be too precise about its meaning beyond this in the present context.

- Now, since  $\int_1^\infty \frac{dx}{x}$  diverges, we would expect  $\int_1^\infty \frac{x+\sin x}{e^{-x}+x^2} dx$  to diverge too.
- Our final task is to verify that our intuition is correct. To do so, we set

$$f(x) = \frac{x + \sin x}{e^{-x} + x^2} \qquad g(x) = \frac{1}{x}$$

and compute

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} &= \lim_{x \rightarrow \infty} \frac{x + \sin x}{e^{-x} + x^2} \div \frac{1}{x} \\ &= \lim_{x \rightarrow \infty} \frac{(1 + \sin x/x)x}{(e^{-x}/x^2 + 1)x^2} \times x \\ &= \lim_{x \rightarrow \infty} \frac{1 + \sin x/x}{e^{-x}/x^2 + 1} \\ &= 1 \end{aligned}$$

- Since  $\int_1^\infty g(x) dx = \int_1^\infty \frac{dx}{x}$  diverges, by Example 3.7.8 with  $p = 1$ , Theorem 3.7.23(b) now tells us that  $\int_1^\infty f(x) dx = \int_1^\infty \frac{x+\sin x}{e^{-x}+x^2} dx$  diverges too.

Example 3.7.24

### 3.8▲ Overview of Integration Techniques

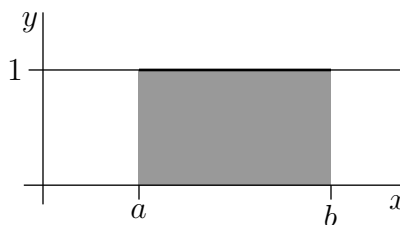
We have now learned many fancy methods of integration. Below we give a short review of the methods we've learned and a general idea of when you might want to choose each one. This section has no new mathematical content.

Up till now, you could often guess the method to use on integrals in the practice book by noticing which section they were in. Section 3.8 in the practice book has lots of integrals to work on, without that contextual hint.

#### ► Known Areas (Section 3.1.3)

A definite integral gives the area underneath a curve. If that area makes a simple shape, you might be able to use a formula from geometry. This is particularly convenient for rectangles.

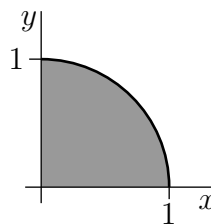
$$\int_a^b 1 dx = (b - a) \times (1) = b - a$$



A special case where this method is useful is with half and quarter circles. If we wanted to use the Fundamental Theorem of Calculus to evaluate the integral below, we'd need a

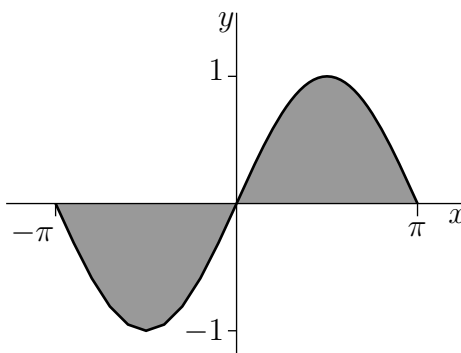
trigonometric substitution. It's much easier to recognize that the area in question is one quarter of the unit circle.

$$\int_0^1 \sqrt{1-x^2} dx = \frac{1}{4}\pi(1)^2 = \frac{\pi}{4}$$



You can also take advantage of a function's symmetry. For example,  $\int_{-\pi}^{\pi} \sin x dx = 0$  because the positive area on the right exactly cancels out the negative (net) area on the left.

$$\int_{-\pi}^{\pi} \sin x dx = 0$$



### ► Substitution (Section 3.4)

Substitution is doing the chain rule in reverse. If you see some “inside” function whose derivative shows up multiplied to the rest of the integrand, you might want to try substitution. An obvious example would be something like this:

$$\int (4x + 3) \cos(2x^2 + 3x) dx$$

The “inside function”  $2x^2 + 3x$  has derivative  $(4x + 3)$ , and we see precisely that derivative multiplied to the rest of the integrand. So this is a great candidate for the substitution  $u = 2x^2 + 3x$ .

Substitution is sometimes a first step to get a function in a better form for a second technique. For example, the function  $\frac{e^x + e^{3x}}{e^x(1-e^x)(2-e^x)}$  is a rational function (and a candidate for the method of partial fractions to antidifferentiate) if you consider  $e^x$  to be your variable. So a first step in antidifferentiation would be to use  $u = e^x$ ,  $du = e^x dx$ :

$$\int \frac{e^x + e^{3x}}{e^x(1-e^x)(2-e^x)} dx = \int \frac{1 + e^{2x}}{e^x(1-e^x)(2-e^x)} \cdot e^x dx = \int \frac{1 + u^2}{u(1-u)(2-u)} du$$

### ►► Integration by Parts (Section 3.5)

If you see the product of two functions, and you would like to swap one with its derivative and the other with its antiderivative, then integration by parts is the method for you. One standard example is the integral

$$\int xe^x dx.$$

We let  $u = x$  and  $dv = e^x dx$ . Then  $du = dx$  and  $v = e^x$ . The function  $v = e^x$  isn't much of an improvement over  $dv = e^x dx$ , but the function  $du = dx$  is an improvement over  $u = x$ . We get to replace our integrand with a simpler product of functions:

$$\int xe^x dx = xe^x - \int e^x dx$$

(Contrast this with the substitution rule: both often operate on integrands that are the product of functions.)

There is one special case you should recognize where integration by parts is useful although the integrand doesn't obviously consist of the product of functions: the antiderivatives of logarithms and inverse trig functions. (See Examples 3.5.8 and 3.5.9 for details.) For example, to antidifferentiate the natural logarithm, we use integration by parts with  $u = \ln x$  and  $dv = dx$ .

$$\int \ln x dx = x \ln x - \int x \cdot \frac{1}{x} dx$$

### ►► Numerical Integration (Section 3.6)

Some integrals, such as

$$\int e^{x^2} dx \quad \text{and} \quad \int \sin(x^2) dx$$

cannot be evaluated using the techniques we've learned so far. Their definite integrals, however, can be approximated using Simpson's Rule. This rule comes with error bounds, so we can make sure our error is within a given tolerance.

One special application of numerical integration is finding a decimal approximation for an irrational number. In Question 23 of Section 3.6 in the practice book, we find a decimal approximation of  $\ln 2$  by applying Simpson's Rule to the integral

$$\int_1^2 \frac{1}{x} dx.$$

### ►► Improper Integrals (Section 3.7)

Improper integrals have infinite discontinuities in their integrands or infinite intervals of integration. The two integrals

$$\int_1^{\infty} e^{-x} dx \quad \text{and} \quad \int_{-1}^2 \frac{1}{x} dx$$

are both improper. The second one is a dangerous type: it's easy to try to apply the Fundamental Theorem of Calculus to evaluate it, without realizing that your computation

is nonsense. Both types of improper integrals are evaluated with limits. If at least one of these limits don't exist (including limits going to infinity), then we say the integral diverges. That means, roughly, that we don't have a sensible way of assigning a number to that definite integral.

### 3.9▲ Differential Equations

A differential equation is an equation for an unknown function that involves the *derivative* of the unknown function. Differential equations play a central role in modelling a huge number of different phenomena. Here is a table giving a bunch of named differential equations and what they are used for. It is far from complete.

Newton's Law of Motion	describes motion of particles
Maxwell's equations	describes electromagnetic radiation
Navier–Stokes equations	describes fluid motion
Heat equation	describes heat flow
Wave equation	describes wave motion
Schrödinger equation	describes atoms, molecules and crystals
Stress-strain equations	describes elastic materials
Black–Scholes models	used for pricing financial options
Predator–prey equations	describes ecosystem populations
Einstein's equations	connects gravity and geometry
Ludwig–Jones–Holling's equation	models spruce budworm/Balsam fir ecosystem
Zeeman's model	models heart beats and nerve impulses
Sherman–Rinzel–Keizer model	for electrical activity in Pancreatic $\beta$ -cells
Hodgkin–Huxley equations	models nerve action potentials

We are just going to scratch the surface of the study of differential equations. Most universities offer half a dozen different undergraduate courses on various aspects of differential equations. We'll focus here on one important type of differential equation: separable differential equations.

We've already seen one type of differential equation: finding an antiderivative.

#### Example 3.9.1

Suppose  $y(x)$  is a function satisfying

$$\frac{dy}{dx} = e^x.$$

What is  $y$ ?



*Solution.* We know the derivative of our function  $y$ , as a function of  $x$ , so we just antidifferentiate.

$$y(x) = e^x + C$$

for some constant  $C$ .

Note the answer to the question is a *function*.

Example 3.9.1

Before we talk about *solving* more complicated differential equations, let's get more practice working with them. The biggest paradigm shift between solving a differential equation, and the type of equation-solving you're used to, is that we're solving for a *function* instead of a variable.

Example 3.9.2

Choose the function(s) listed below that solve this differential equation:

$$\frac{dy}{dx} + x^2 - 1 = y$$

A.  $y = x^2 + 2x + 1$

B.  $y = x^2 + 1$

*Solution.* We want to check whether  $y$  is a solution, so we replace  $y$  and  $\frac{dy}{dx}$  in the differential equation, and check whether the equation is true.

In the case  $y = x^2 + 2x + 1$ , then  $\frac{dy}{dx} = 2x + 2$ . We plug these into our differential equation:

$$\begin{aligned} \frac{dy}{dx} + x^2 - 1 &= y \\ 2x + 2 + x^2 - 1 &= x^2 + 2x + 1 \end{aligned}$$

Simplifying the left-hand side,

$$x^2 + 2x + 1 = x^2 + 2x + 1$$

This is true – the function on the left and the function on the right are the same. So the equation  $y = x^2 + 2x + 1$  is a solution to the differential equation  $\frac{dy}{dx} + x^2 - 1 = y$ .

Now let's think about the other function we were asked to consider,  $y = x^2 + 1$ . For this function,  $\frac{dy}{dx} = 2x$ . We plug these into our differential equation:

$$\begin{aligned} \frac{dy}{dx} + x^2 - 1 &= y \\ 2x + x^2 - 1 &= x^2 + 2x + 1 \end{aligned}$$

Rearranging the left-hand side,

$$x^2 + 2x - 1 = x^2 + 2x + 1$$

This is **not** true – the function on the left and the function on the right are **not** the same. So the equation  $y = x^2 + 2x$  is **not** a solution to the differential equation  $\frac{dy}{dx} + x^2 - 1 = y$ .

You don't have enough tools yet to come up with solutions like  $y = x^2 + 2x + 1$  – those will come shortly. For this example, we only want you to understand what it means for a function to be a solution to a differential equation.

Example 3.9.2

### Definition 3.9.3.

A *separable differential equation* is an equation for a function  $y(x)$  that can be written in the form

$$g(y(x)) \frac{dy}{dx}(x) = f(x)$$

It may take some rearranging to get a differential equation in this form. The “separable” refers to the mechanics of getting all terms containing  $y$  and  $y'$  on side of the equation, and all terms containing  $x$  on the other side.

We'll start by developing a recipe for solving separable differential equations. Then we'll look at many examples. Usually one suppresses the argument of  $y(x)$  and writes the equation as below:

$$g(y) \frac{dy}{dx} = f(x)$$

If the left and right functions side of the equation are the same (and they should be – otherwise that equals sign has no business being there) then their *antiderivatives with respect to  $x$*  should be the same as well, up to the usual additive constant.

$$\int \left( g(y) \frac{dy}{dx} \right) dx = \int f(x) dx$$

The left-hand side of the equation above is in a perfect form for a substitution.

$$\int g(y) dy = \int f(x) dx \quad (*)$$

In this way, we've turned the problem of finding solutions to our separable differential equation into the problem of finding two antiderivatives.

Note the work above didn't really depend on what, exactly,  $f(x)$  and  $g(y)$  were. So, to skip to the end, we use the following mnemonic algorithm. It looks strange, but you can simply think of it as shorthand for the work we just did above.

$$g(y) \cdot \frac{dy}{dx} = f(x)$$

$$g(y) dy = f(x) dx \quad (1)$$

$$\int g(y) dy = \int f(x) dx \quad (2)$$

In Step (1), we separate all  $x$ 's and  $y$ 's, including in  $\frac{dy}{dx}$ , by “multiplying” both sides of the equation by  $dx$ . In Step (2), we add an integral side to both sides of the equation.

This looks illegal, and indeed is illegal —  $\frac{dy}{dx}$  is not a fraction. Again, this procedure is simply a mnemonic device to help you remember the result (\*).

Example 3.9.4

The differential equation

$$\frac{dy}{dx} = xe^{-y}$$

is separable, and we now find all of its solutions by using our mnemonic device. We start by cross-multiplying so as to move all  $y$ 's to the left hand side and all  $x$ 's (including  $dx$ ) to the right hand side.

$$e^y dy = x dx$$

Then we integrate both sides.

$$\int e^y dy = \int x dx \iff e^y = \frac{x^2}{2} + C$$

The  $C$  on the right hand side contains both the arbitrary constant for the indefinite integral  $\int e^y dy$  and the arbitrary constant for the indefinite integral  $\int x dx$ . Finally, we solve for  $y$ , which is really a function of  $x$ .

$$y(x) = \ln \left( \frac{x^2}{2} + C \right)$$

Note that  $C$  is an arbitrary constant. It can take any value. It cannot be determined by the differential equation itself. In applications  $C$  is usually determined by a requirement that  $y$  take some prescribed value (determined by the application) when  $x$  is some prescribed value. (We call these types of problems “initial value” problems. The given constants are “initial conditions.”) For example, suppose that we wish to find a function  $y(x)$  that obeys both

$$\frac{dy}{dx} = xe^{-y} \quad \text{and} \quad y(0) = 1$$

We know that, to have  $\frac{dy}{dx} = xe^{-y}$  satisfied, we must have  $y(x) = \ln \left( \frac{x^2}{2} + C \right)$ , for some constant  $C$ . To also have the initial condition  $y(0) = 1$ , we must have

$$1 = y(0) = \ln \left( \frac{x^2}{2} + C \right) \Big|_{x=0} = \ln C \iff \ln C = 1 \iff C = e$$

So our final solution is  $y(x) = \ln\left(\frac{x^2}{2} + e\right)$ .

Example 3.9.4

Example 3.9.5

Solve  $\frac{dy}{dx} = y^2$

*Solution.* When  $y \neq 0$ , we can use our mnemonic.

$$\begin{aligned}\frac{dy}{dx} &= y^2 \\ \frac{dy}{y^2} &= dx \\ \int \frac{dy}{y^2} &= \int dx \\ \frac{y^{-1}}{-1} &= x + C \\ y &= -\frac{1}{x + C}\end{aligned}$$

When  $y = 0$ , this computation breaks down because  $\frac{dy}{y^2}$  contains a division by 0. We can check if the function  $y(x) = 0$  satisfies the differential equation by just subbing it in:

$$y(x) = 0 \implies y'(x) = 0, \quad y(x)^2 = 0 \implies y'(x) = y(x)^2$$

So  $y(x) = 0$  is a solution and the full solution is:

$$y(x) = 0 \text{ or } y(x) = -\frac{1}{x + C}, \text{ for any constant } C$$

Example 3.9.5

Example 3.9.6 (War Moods)

In the article *War Moods: 1*<sup>51</sup>, researcher Lewis Richardson models the proportion of a population eager for war using a model previously applied to the spread of infectious diseases. (We note here an important quote from the paper: “To describe a phenomenon is not to praise it.” Understanding the social psychology of public support for war may lead to strategies for preventing conflicts.)

A simplified version of Richardson’s model for the lead-up to hostilities is as follows. Let  $y$  be the proportion of a population that supports going to war, with the rest of the population against going to war. Then the rate of change of  $y$  over time is proportional to the product of the proportion of people who are pro-war and the proportion of people

51 War Moods: 1 by Richardson, PSYCHOMETRIKA–Vol. 13, no. 3 September, 1948. You can access the full text with your UBC CWL at [this link](#).

who are anti-war. The reasoning is roughly<sup>52</sup> that  $y$  changes as pro-war people encounter anti-war people.

That corresponds to the differential equation

$$\frac{dy}{dt} = Cy(1 - y)$$

where  $1 - y$  is the proportion of people who are anti-war.

Let's solve this differential equation.

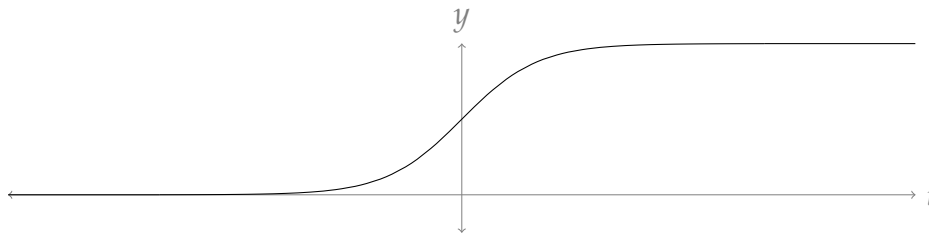
$$\begin{aligned}\frac{dy}{dt} &= Cy(1 - y) \\ \frac{1}{y(1 - y)} dy &= C dt \\ \int \frac{1}{y(1 - y)} dy &= \int C dt\end{aligned}$$

Here, we'll use a slightly unfair trick. Note that  $\frac{1}{y(1-y)} = \frac{1}{y} + \frac{1}{1-y}$ . (Finding such convenient equalities in order to integrate is called the method of partial fractions; you can read about it in many integral calculus textbooks. You will not be expected to come up with this equality on your own for Math 105.)

$$\begin{aligned}\int \left( \frac{1}{y} + \frac{1}{1-y} \right) dy &= \int C dt \\ \ln y - \ln(1 - y) &= Ct + D \\ \ln \left( \frac{y}{1-y} \right) &= Ct + D \\ \frac{y}{1-y} &= e^{Ct+D} \\ y &= (1-y)e^{Ct+D} = e^{Ct+D} - ye^{Ct+D} \\ y(1 + e^{Ct+D}) &= e^{Ct+D} \\ y &= \frac{e^{Ct+D}}{1 + e^{Ct+D}}\end{aligned}$$

where  $D$  is some constant.

The graph of this function has the shape below.



52 The actual paper has more subtlety, including considering populations of rival nations, and the progression of public sentiment as a war drags on.

In this model, there is a quick change from low support for war ( $y \approx 0$ ) to high support for war ( $y \approx 1$ ). The paper notes, regarding the first world war: “There is evidence ... that the majority of Britishers changed their opinions about war with Germany during a week in 1914 between July 24 and August 4.”

Example 3.9.6

Example 3.9.7 (Fish growth)

Professor Daniel Pauly of the UBC Institute for the Oceans and Fisheries considered the following model of fish growth in the paper *A précis of Gill-Oxygen Limitation Theory (GOLT), with some Emphasis on the Eastern Mediterranean*<sup>53</sup>.

Let  $w(t)$  be the weight of an individual fish over time. The rate at which it is able to synthesize proteins (and other necessary substances) is proportional to  $w^d$ , while the rate at which its proteins need to be replaced is proportional to  $w$ . So,

$$\frac{dw}{dt} = Hw^d - kw$$

where  $Hw^d$  is the rate at which new proteins are built, and  $kw$  is the rate at which they need to be replaced. Because the rate of production is limited to the rate of oxygen intake (which itself is proportional to gill size), the exponent  $d$  is less than one.

The paper notes that researchers often neglect oxygen impacts on fish growth—a decision not supported by this model.

Suppose  $d = 0.5$  for a particular small species of fish. What is  $w(t)$ ? How large would the fish grow, if it grew indefinitely?

*Solution.* The differential equation is separable.

$$\begin{aligned} \frac{dw}{dt} &= H\sqrt{w} - kw \\ \frac{1}{H\sqrt{w} - kw} dw &= dt \\ \int \frac{1}{H\sqrt{w} - kw} dw &= \int 1 dt \\ \int \frac{1}{\sqrt{w}(H - k\sqrt{w})} dw &= \int 1 dt \end{aligned}$$

53 PAULY, D. (2019). A précis of Gill-Oxygen Limitation Theory (GOLT), with some Emphasis on the Eastern Mediterranean. *Mediterranean Marine Science*, 20(4), 660-668. doi:<http://dx.doi.org/10.12681/mms.19285>

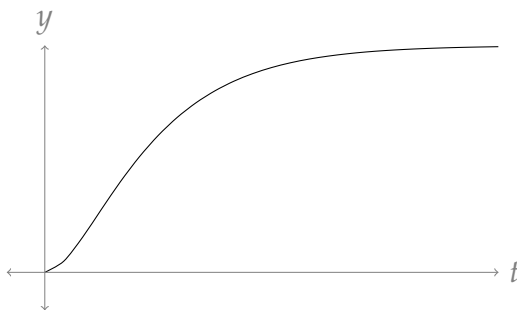
For the left-hand side, we use the substitution  $u = H - k\sqrt{w}$ ,  $-\frac{2}{k}du = \frac{1}{\sqrt{w}}dw$

$$\begin{aligned} -\frac{2}{k} \int \frac{1}{u} du &= \int 1 dt \\ -\frac{2}{k} \ln |u| &= t + C \\ \ln |u| &= -\frac{k}{2}t + C && \text{(remember } C \text{ is an arbitrary constant)} \\ |u| &= e^{-\frac{k}{2}t+C} \\ |H - k\sqrt{w}| &= e^{-\frac{k}{2}t+C} \end{aligned}$$

If we're studying young fish growing to adulthood, then  $\frac{dw}{dt} > 0$ , because the fish are getting bigger. Under this assumption,  $H - k\sqrt{w} > 0$ , so we can drop the absolute value signs.

$$\begin{aligned} H - k\sqrt{w} &= e^{-\frac{k}{2}t+C} \\ k\sqrt{w} &= H - e^{-\frac{k}{2}t+C} \\ \sqrt{w} &= \frac{1}{k} \left( H - e^{-\frac{k}{2}t+C} \right) \\ w &= \frac{1}{k^2} \left( H - e^{-\frac{k}{2}t+C} \right)^2 \end{aligned}$$

The shape of this function is shown below.



The graph shape suggests the existence of a horizontal asymptote. Indeed:

$$\lim_{t \rightarrow \infty} \frac{1}{k^2} \left( H - e^{-\frac{k}{2}t+C} \right)^2 = \left( \frac{H}{k} \right)^2$$

So, aging fish who grow according to our model approach the weight  $\left(\frac{H}{k}\right)^2$ .

Example 3.9.7

**Definition 3.9.8.**

A differential equation of the form

$$\frac{dy}{dx} = a(y - b)$$

where  $a$  and  $b$  are constants is called a **first-order linear** differential equation.

“First-order” means the equation has a first derivative, but no higher-order derivatives (e.g. no second derivatives). The right hand side is a linear expression in the variable  $y$ .

**Example 3.9.9**

Let  $a$  and  $b$  be any two constants. We’ll now solve the family of differential equations

$$\frac{dy}{dx} = a(y - b)$$

using our mnemonic device.

$$\begin{aligned} \frac{dy}{y - b} = a \, dx &\implies \int \frac{dy}{y - b} = \int a \, dx \implies \ln|y - b| = ax + c \implies |y - b| = e^{ax+c} = e^c e^{ax} \\ &\implies y - b = C e^{ax} \end{aligned}$$

where  $C$  is either  $+e^c$  or  $-e^c$ . Note that as  $c$  runs over all real numbers,  $+e^c$  runs over all strictly positive real numbers and  $-e^c$  runs over all strictly negative real numbers. So, so far,  $C$  can be any real number except 0. But we were a bit sloppy here. We implicitly assumed that  $y - b$  was nonzero, so that we could divide it across. None-the-less, the constant function  $y = b$ , which corresponds to  $C = 0$ , is a perfectly good solution — when  $y$  is the constant function  $y = b$ , both  $\frac{dy}{dx}$  and  $a(y - b)$  are zero. So the general solution to  $\frac{dy}{dx} = a(y - b)$  is  $y(x) = C e^{ax} + b$ , where the constant  $C$  can be any real number. Note that when  $y(x) = C e^{ax} + b$  we have  $y(0) = C + b$ . So  $C = y(0) - b$  and the general solution is

$$y(x) = (y(0) - b)e^{ax} + b$$

**Example 3.9.9**

This is worth stating as a theorem.

**Theorem 3.9.10.**

Let  $a$  and  $b$  be constants. The differentiable function  $y(x)$  obeys the differential equation

$$\frac{dy}{dx} = a(y - b)$$

if and only if

$$y(x) = (y(0) - b)e^{ax} + b$$



One solution to the differential equation

$$\frac{dy}{dt} = a(y - b)$$

is the constant equation

$$y = b.$$

We call this the “steady state” solution. Steady, because  $y$  is never changing – it’s always  $b$ .

**Definition 3.9.11.**

A constant function  $y = b$  that satisfies a differential equation of the form

$$\frac{dy}{dt} = g(y)$$

is a *steady state solution*. The constant  $b$  is a *steady state* of the differential equation

$$\text{if } \left. \frac{dy}{dt} \right|_{y=b} = g(b) = 0$$

**Example 3.9.12**

A glucose solution is administered intravenously into the bloodstream at a constant rate  $r$ . As the glucose is added, it is converted into other substances at a rate that is proportional to the concentration at that time. The concentration,  $C(t)$ , of the glucose in the bloodstream at time  $t$  obeys the differential equation

$$\frac{dC}{dt} = r - kC$$

where  $k$  is a positive constant of proportionality.

- Express  $C(t)$  in terms of  $k$  and  $C(0)$ .
- Find  $\lim_{t \rightarrow \infty} C(t)$ .
- Find the steady-state solution to the differential equation.

*Solution.* (a) Since  $r - kC = -k(C - \frac{r}{k})$  the given equation is

$$\frac{dC}{dt} = -k\left(C - \frac{r}{k}\right)$$

which is of the form solved in Theorem 3.9.10 with  $a = -k$  and  $b = \frac{r}{k}$ . So the solution is

$$C(t) = \frac{r}{k} + \left(C(0) - \frac{r}{k}\right)e^{-kt}$$

(b) For any  $k > 0$ ,  $\lim_{t \rightarrow \infty} e^{-kt} = 0$ . Consequently, for any  $C(0)$  and any  $k > 0$ ,  $\lim_{t \rightarrow \infty} C(t) = \frac{r}{k}$ . We could have predicted this limit without solving for  $C(t)$ . If we assume that  $C(t)$  approaches some equilibrium value  $C_e$  as  $t$  approaches infinity, then taking the limits of both sides of  $\frac{dC}{dt} = r - kC$  as  $t \rightarrow \infty$  gives

$$0 = r - kC_e \implies C_e = \frac{r}{k}$$

(c)  $\frac{dC}{dt} = 0$  when  $C = \frac{r}{k}$ , so  $C = \frac{r}{k}$  is the steady-state solution. That is, if we have a blood concentration of glucose of  $\frac{r}{k}$ , the concentration is staying the same even as the glucose is injected and metabolized.

Example 3.9.12

### 3.9.1 ▶ (Optional) Logistic Growth

Suppose that we wish to predict the size  $P(t)$  of a population as a function of the time  $t$ . In the most naive model of population growth, each couple produces  $\beta$  offspring (for some constant  $\beta$ ) and then dies. Thus over the course of one generation  $\beta \frac{P(t)}{2}$  children are produced and  $P(t)$  parents die so that the size of the population grows from  $P(t)$  to

$$P(t + t_g) = \underbrace{P(t) + \beta \frac{P(t)}{2}}_{\text{parents+offspring}} - \underbrace{P(t)}_{\text{parents die}} = \frac{\beta}{2} P(t)$$

where  $t_g$  denotes the lifespan of one generation. The rate of change of the size of the population per unit time is

$$\frac{P(t + t_g) - P(t)}{t_g} = \frac{1}{t_g} \left[ \frac{\beta}{2} P(t) - P(t) \right] = bP(t)$$

where  $b = \frac{\beta - 2}{2t_g}$  is the net birthrate per member of the population per unit time. If we approximate

$$\frac{P(t + t_g) - P(t)}{t_g} \approx \frac{dP}{dt}(t)$$

we get the differential equation

$$\frac{dP}{dt} = bP(t) \tag{3.9.1}$$

By Theorem 3.9.10,

$$P(t) = P(0) \cdot e^{bt} \tag{3.9.2}$$

This is called the Malthusian<sup>54</sup> growth model. It is, of course, very simplistic. One of its main characteristics is that, since  $P(t + T) = P(0) \cdot e^{b(t+T)} = P(t) \cdot e^{bT}$ , every time you *add*

54 This is named after Rev. Thomas Robert Malthus. He described this model in a 1798 paper called "An essay on the principle of population".

$T$  to the time, the population size is *multiplied* by  $e^{bT}$ . In particular, the population size doubles every  $\frac{\ln 2}{b}$  units of time.

Example 3.9.13

In 1927 the population of the world was about 2 billion. In 1974 it was about 4 billion. Estimate when it reached 6 billion. What will the population of the world be in 2100, assuming the Malthusian growth model?

*Solution.*

- Let  $P(t)$  be the world's population, in billions,  $t$  years after 1927. Note that 1974 corresponds to  $t = 1974 - 1927 = 47$ .
- We are assuming that  $P(t)$  obeys equation (3.9.1). So, by (3.9.2)

$$P(t) = P(0) \cdot e^{bt}$$

Notice that there are 2 unknowns here —  $b$  and  $P(0)$  — so we need two pieces of information to find them.

- We are told  $P(0) = 2$ , so

$$P(t) = 2 \cdot e^{bt}$$

- We are also told  $P(47) = 4$ , which gives

$$4 = 2 \cdot e^{47b} \quad \text{clean up}$$

$$e^{47b} = 2 \quad \text{take the log and clean up}$$

$$b = \frac{\ln 2}{47} = 0.0147 \quad \text{to 3 decimal places}$$

- We now know  $P(t)$  completely, so we can easily determine the predicted population<sup>55</sup> in 2100, i.e. at  $t = 2100 - 1927 = 173$ .

$$P(173) = 2e^{173b} = 2e^{173 \times 0.0147} = 12.7 \text{ billion}$$

- Finally, our crude model predicts that the population is 6 billion at the time  $t$  that obeys

$$P(t) = 2e^{bt} = 6 \quad \text{clean up}$$

$$e^{bt} = 3 \quad \text{take the log and clean up}$$

$$t = \frac{\ln 3}{b} = 47 \frac{\ln 3}{\ln 2} = 74.5$$

which corresponds<sup>56</sup> to the middle of 2001.

55 The 2015 Revision of World Population, a publication of the United Nations, predicts that the world's population in 2100 will be about 11 billion. But "about" covers a pretty large range. They give an 80% confidence interval running from 10 billion to 12.5 billion.

56 The world population really reached 6 billion in about 1999.

## Example 3.9.13

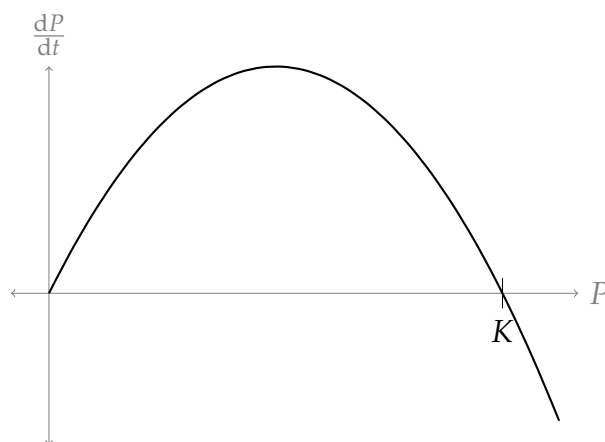
The Malthusian growth model can be a reasonably good model only when the population size is very small compared to its environment<sup>57</sup>. A more sophisticated model of population growth takes into account the “carrying capacity of the environment.”

Logistic growth adds one more wrinkle to the simple population model. It assumes that the population only has access to limited resources. As the size of the population grows the amount of food available to each member decreases. This in turn causes the net birth rate  $b$  to decrease. In the logistic growth model  $b = b_0 \left(1 - \frac{P}{K}\right)$ , where  $K$  is called the carrying capacity of the environment, so that

$$P'(t) = b_0 \left(1 - \frac{P(t)}{K}\right) P(t)$$

Figure 3.9.1.

Below is a graph of  $P'(t) = b_0 \left(1 - \frac{P(t)}{K}\right) P(t)$ . Pay attention to the axis labels: the independent (horizontal) axis is population,  $P$ . It is *not* time. The dependent (vertical) axis is rate of change of population,  $\frac{dP}{dt}$ .



- When  $P = 0$ , there are no individuals in the population, so its growth rate is zero. (Extinct animals do not usually reproduce.)
- When  $0 < P < K$ , the population is less than the carrying capacity of its environment, so the population grows ( $\frac{dP}{dt} > 0$ ).
- When  $P > K$ , the population has outgrown the capacity of the environment to support it. Then  $\frac{dP}{dt} < 0$ , as the population experiences a higher death rate than birth rate.

This is a separable differential equation and we can solve it explicitly. We shall do so shortly (in Example 3.9.14, below). But, before doing that, we'll see what we can learn

57 That is, the population has plenty of food and space to grow.

about the behaviour of solutions to differential equations like this without finding formulae for the solutions. It turns out that we can learn a lot just by watching the sign of  $P'(t)$ . For concreteness, we'll look at solutions of the differential equation

$$\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$$

We'll sketch the graphs of four functions  $P(t)$  that obey this equation.

- For the first function,  $P(0) = 0$ .
- For the second function,  $P(0) = 1000$ .
- For the third function,  $P(0) = 2000$ .
- For the fourth function,  $P(0) = 3000$ .

The sketches will be based on the observation that  $(6000 - 3P)P = 3(2000 - P)P$

- is zero for  $P = 0, 2000$ ,
- is strictly positive for  $0 < P < 2000$  and
- is strictly negative for  $P > 2000$ .

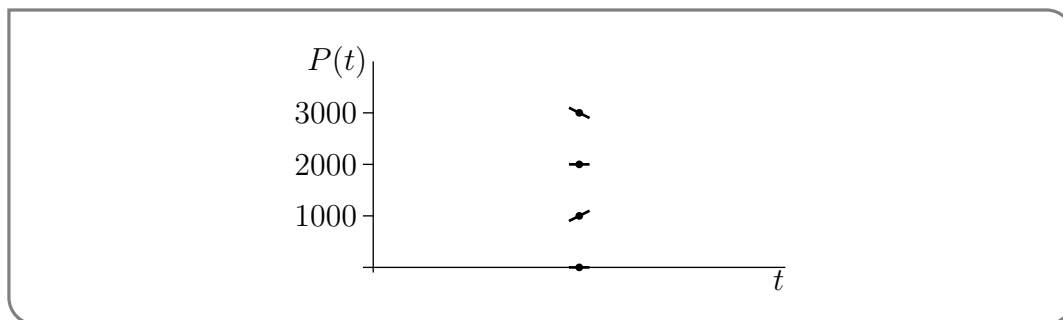
Consequently

$$\frac{dP}{dt}(t) \begin{cases} = 0 & \text{if } P(t) = 0 \\ > 0 & \text{if } 0 < P(t) < 2000 \\ = 0 & \text{if } P(t) = 2000 \\ < 0 & \text{if } P(t) > 2000 \end{cases}$$

Thus if  $P(t)$  is some function that obeys  $\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$ , then as the graph of  $P(t)$  passes through the point  $(t, P(t))$

$$\text{the graph has } \begin{cases} \text{slope zero,} & \text{i.e. is horizontal, if } P(t) = 0 \\ \text{positive slope,} & \text{i.e. is increasing, if } 0 < P(t) < 2000 \\ \text{slope zero,} & \text{i.e. is horizontal, if } P(t) = 2000 \\ \text{negative slope,} & \text{i.e. is decreasing, if } P(t) > 2000 \end{cases}$$

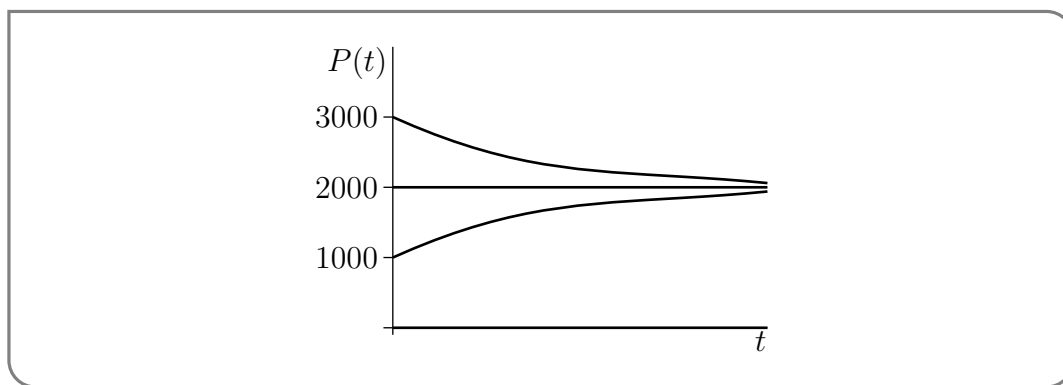
as illustrated in the figure



As a result,

- if  $P(0) = 0$ , the graph starts out horizontally. In other words, as  $t$  starts to increase,  $P(t)$  remains at zero, so the slope of the graph remains at zero. The population size remains zero for all time. As a check, observe that the function  $P(t) = 0$  obeys  $\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$  for all  $t$ .
- Similarly, if  $P(0) = 2000$ , the graph again starts out horizontally. So  $P(t)$  remains at 2000 and the slope remains at zero. The population size remains 2000 for all time. Again, the function  $P(t) = 2000$  obeys  $\frac{dP}{dt}(t) = (6000 - 3P(t))P(t)$  for all  $t$ .
- If  $P(0) = 1000$ , the graph starts out with positive slope. So  $P(t)$  increases with  $t$ . As  $P(t)$  increases towards 2000, the slope  $(6000 - 3P(t))P(t)$ , while remaining positive, gets closer and closer to zero. As the graph approaches height 2000, it becomes more and more horizontal. The graph cannot actually cross from below 2000 to above 2000, because to do so it would have to have strictly positive slope for some value of  $P$  above 2000, which is not allowed.
- If  $P(0) = 3000$ , the graph starts out with negative slope. So  $P(t)$  decreases with  $t$ . As  $P(t)$  decreases towards 2000, the slope  $(6000 - 3P(t))P(t)$ , while remaining negative, gets closer and closer to zero. As the graph approaches height 2000, it becomes more and more horizontal. The graph cannot actually cross from above 2000 to below 2000, because to do so it would have to have negative slope for some value of  $P$  below 2000, which is not allowed.

These curves are sketched in the figure below. We conclude that for any initial population size  $P(0)$ , except  $P(0) = 0$ , the population size approaches 2000 as  $t \rightarrow \infty$ .



Now we'll do an example in which we explicitly solve the logistic growth equation.

#### Example 3.9.14

In 1986, the population of the world was 5 billion and was increasing at a rate of 2% per year. Using the logistic growth model with an assumed maximum population of 100 billion, predict the population of the world in the years 2000, 2100 and 2500.

*Solution.* Let  $y(t)$  be the population of the world, in billions of people, at time  $1986 + t$ . The logistic growth model assumes

$$y' = ay(K - y)$$

where  $K$  is the carrying capacity and  $a = \frac{b_0}{K}$ .

First we'll determine the values of the constants  $a$  and  $K$  from the given data.

- We know that, if at time zero the population is below  $K$ , then as time increases the population increases, approaching the limit  $K$  as  $t$  tends to infinity. So in this problem  $K$  is the maximum population. That is,  $K = 100$ .
- We are also told that, at time zero, the percentage rate of change of population,  $100\frac{y'}{y}$ , is 2, so that, at time zero,  $\frac{y'}{y} = 0.02$ . But, from the differential equation,  $\frac{y'}{y} = a(K - y)$ . Hence at time zero,  $0.02 = a(100 - 5)$ , so that  $a = \frac{2}{9500}$ .

We now know  $a$  and  $K$  and can solve the (separable) differential equation

$$\begin{aligned} \frac{dy}{dt} = ay(K - y) &\implies \frac{dy}{y(K - y)} = a dt \implies \int \frac{1}{K} \left[ \frac{1}{y} - \frac{1}{y - K} \right] dy = \int a dt \\ &\implies \frac{1}{K} [\ln |y| - \ln |y - K|] = at + C \\ &\implies \ln \frac{|y|}{|y - K|} = aKt + CK \implies \left| \frac{y}{y - K} \right| = De^{aKt} \end{aligned}$$

with  $D = e^{CK}$ . We know that  $y$  remains between 0 and  $K$ , so that  $\left| \frac{y}{y - K} \right| = \frac{y}{K - y}$  and our solution obeys

$$\frac{y}{K - y} = De^{aKt}$$

At this stage, we know the values of the constants  $a$  and  $K$ , but not the value of the constant  $D$ . We are given that at  $t = 0$ ,  $y = 5$ . Subbing in this, and the values of  $K$  and  $a$ ,

$$\frac{5}{100 - 5} = De^0 \implies D = \frac{5}{95}$$

So the solution obeys the algebraic equation

$$\frac{y}{100 - y} = \frac{5}{95} e^{2t/95}$$

which we can solve to get  $y$  as a function of  $t$ .

$$\begin{aligned} y = (100 - y) \frac{5}{95} e^{2t/95} &\implies 95y = (500 - 5y) e^{2t/95} \\ &\implies (95 + 5e^{2t/95})y = 500e^{2t/95} \\ &\implies y = \frac{500e^{2t/95}}{95 + 5e^{2t/95}} = \frac{100e^{2t/95}}{19 + e^{2t/95}} = \frac{100}{1 + 19e^{-2t/95}} \end{aligned}$$

Finally,

- In the year 2000,  $t = 14$  and  $y = \frac{100}{1 + 19e^{-28/95}} \approx 6.6$  billion.
- In the year 2100,  $t = 114$  and  $y = \frac{100}{1 + 19e^{-228/95}} \approx 36.7$  billion.
- In the year 2200,  $t = 514$  and  $y = \frac{100}{1 + 19e^{-1028/95}} \approx 100$  billion.

Example 3.9.14

### 3.9.2 ▶ (Optional) Interest on Investments and Loans

Suppose that you deposit  $\$P$  in a bank account at time  $t = 0$ . The account pays  $r\%$  interest per year compounded  $n$  times per year.

- The first interest payment is made at time  $t = \frac{1}{n}$ . Because the balance in the account during the time interval  $0 < t < \frac{1}{n}$  is  $\$P$  and interest is being paid for  $(\frac{1}{n})^{\text{th}}$  of a year, that first interest payment is  $\frac{1}{n} \times \frac{r}{100} \times P$ . After the first interest payment, the balance in the account is  $P + \frac{1}{n} \times \frac{r}{100} \times P = (1 + \frac{r}{100n})P$ .
- The second interest payment is made at time  $t = \frac{2}{n}$ . Because the balance in the account during the time interval  $\frac{1}{n} < t < \frac{2}{n}$  is  $(1 + \frac{r}{100n})P$  and interest is being paid for  $(\frac{1}{n})^{\text{th}}$  of a year, the second interest payment is  $\frac{1}{n} \times \frac{r}{100} \times (1 + \frac{r}{100n})P$ . After the second interest payment, the balance in the account is  $(1 + \frac{r}{100n})P + \frac{1}{n} \times \frac{r}{100} \times (1 + \frac{r}{100n})P = (1 + \frac{r}{100n})^2 P$ .
- And so on.

In general, at time  $t = \frac{m}{n}$  (just after the  $m^{\text{th}}$  interest payment), the balance in the account is

$$B(t) = \left(1 + \frac{r}{100n}\right)^m P = \left(1 + \frac{r}{100n}\right)^{nt} P \tag{3.9.3}$$

Three common values of  $n$  are 1 (interest is paid once a year), 12 (i.e. interest is paid once a month) and 365 (i.e. interest is paid daily). The limit  $n \rightarrow \infty$  is called continuous compounding<sup>58</sup>. Under continuous compounding, the balance at time  $t$  is

$$B(t) = \lim_{n \rightarrow \infty} \left(1 + \frac{r}{100n}\right)^{nt} P$$

You may have already seen the limit

$$\lim_{x \rightarrow 0} (1 + x)^{a/x} = e^a \tag{3.9.4}$$

If so, you can evaluate  $B(t)$  by applying (3.9.4) with  $x = \frac{r}{100n}$  and  $a = \frac{rt}{100}$  (so that  $\frac{a}{x} = nt$ ). As  $n \rightarrow \infty$ ,  $x \rightarrow 0$  so that

$$B(t) = \lim_{n \rightarrow \infty} \left(1 + \frac{r}{100n}\right)^{nt} P = \lim_{x \rightarrow 0} (1 + x)^{a/x} P = e^a P = e^{rt/100} P \tag{3.9.5}$$

If you haven't seen (3.9.4) before, that's OK. In the following example, we rederive (3.9.5) using a differential equation instead of (3.9.4).

**Example 3.9.15**

Suppose, again, that you deposit  $\$P$  in a bank account at time  $t = 0$ , and that the account pays  $r\%$  interest per year compounded  $n$  times per year, and denote by  $B(t)$  the balance at time  $t$ . Suppose that you have just received an interest payment at time  $t$ . Then the next

58 There are banks that advertise continuous compounding. You can find some by googling "interest is compounded continuously and paid"



interest payment will be made at time  $t + \frac{1}{n}$  and will be  $\frac{1}{n} \times \frac{r}{100} \times B(t) = \frac{r}{100n} B(t)$ . So, calling  $\frac{1}{n} = h$ ,

$$B(t+h) = B(t) + \frac{r}{100} B(t)h \quad \text{or} \quad \frac{B(t+h) - B(t)}{h} = \frac{r}{100} B(t)$$

To get continuous compounding we take the limit  $n \rightarrow \infty$  or, equivalently,  $h \rightarrow 0$ . This gives

$$\lim_{h \rightarrow 0} \frac{B(t+h) - B(t)}{h} = \frac{r}{100} B(t) \quad \text{or} \quad \frac{dB}{dt}(t) = \frac{r}{100} B(t)$$

By Theorem 3.9.10, with  $a = \frac{r}{100}$  and  $b = 0$ ,

$$B(t) = e^{rt/100} B(0) = e^{rt/100} P$$

once again.

Example 3.9.15

Example 3.9.16

- (a) A bank advertises that it compounds interest continuously and that it will double your money in ten years. What is the annual interest rate?
- (b) A bank advertises that it compounds monthly and that it will double your money in ten years. What is the annual interest rate?

*Solution.* (a) Let the interest rate be  $r\%$  per year. If you start with  $\$P$ , then after  $t$  years, you have  $Pe^{rt/100}$ , under continuous compounding. This was (3.9.5). After 10 years you have  $Pe^{r/10}$ . This is supposed to be  $2P$ , so

$$Pe^{r/10} = 2P \implies e^{r/10} = 2 \implies \frac{r}{10} = \ln 2 \implies r = 10 \ln 2 = 6.93\%$$

(b) Let the interest rate be  $r\%$  per year. If you start with  $\$P$ , then after  $t$  years, you have  $P(1 + \frac{r}{100 \times 12})^{12t}$ , under monthly compounding. This was (3.9.3). After 10 years you have  $P(1 + \frac{r}{100 \times 12})^{120}$ . This is supposed to be  $2P$ , so

$$\begin{aligned} P(1 + \frac{r}{100 \times 12})^{120} = 2P &\implies (1 + \frac{r}{1200})^{120} = 2 \implies 1 + \frac{r}{1200} = 2^{1/120} \\ \implies \frac{r}{1200} = 2^{1/120} - 1 &\implies r = 1200(2^{1/120} - 1) = 6.95\% \end{aligned}$$

Example 3.9.16

Example 3.9.17

A 25 year old graduate of UBC is given  $\$50,000$  which is invested at  $5\%$  per year compounded continuously. The graduate also intends to deposit money continuously at the rate of  $\$2000$  per year.

- (a) Find a differential equation that  $A(t)$  obeys, assuming that the interest rate remains 5%.
- (b) Determine the amount of money in the account when the graduate is 65.
- (c) At age 65, the graduate will start withdrawing money continuously at the rate of  $W$  dollars per year. If the money must last until the person is 85, what is the largest possible value of  $W$ ?

*Solution.* (a) Let's consider what happens to  $A$  over a very short time interval from time  $t$  to time  $t + \Delta t$ . At time  $t$  the account balance is  $A(t)$ . During the (really short) specified time interval the balance remains very close to  $A(t)$  and so earns interest of  $\frac{5}{100} \times \Delta t \times A(t)$ . During the same time interval, the graduate also deposits an additional  $\$2000\Delta t$ . So

$$A(t + \Delta t) \approx A(t) + 0.05A(t)\Delta t + 2000\Delta t \implies \frac{A(t + \Delta t) - A(t)}{\Delta t} \approx 0.05A(t) + 2000$$

In the limit  $\Delta t \rightarrow 0$ , the approximation becomes exact and we get

$$\frac{dA}{dt} = 0.05A + 2000$$

- (b) The amount of money at time  $t$  obeys

$$\frac{dA}{dt} = 0.05A(t) + 2,000 = 0.05(A(t) + 40,000)$$

So by Theorem 3.9.10 (with  $a = 0.05$  and  $b = -40,000$ ),

$$A(t) = (A(0) + 40,000)e^{0.05t} - 40,000$$

At time 0 (when the graduate is 25),  $A(0) = 50,000$ , so the amount of money at time  $t$  is

$$A(t) = 90,000e^{0.05t} - 40,000$$

In particular, when the graduate is 65 years old,  $t = 40$  and

$$A(40) = 90,000e^{0.05 \times 40} - 40,000 = \$625,015.05$$

- (c) When the graduate stops depositing money and instead starts withdrawing money at a rate  $W$ , the equation for  $A$  becomes

$$\frac{dA}{dt} = 0.05A - W = 0.05(A - 20W)$$

assuming that the interest rate remains 5%. This time, Theorem 3.9.10 (with  $a = 0.05$  and  $b = 20W$ ) gives

$$A(t) = (A(0) - 20W)e^{0.05t} + 20W$$

If we now reset our clock so that  $t = 0$  when the graduate is 65,  $A(0) = 625,015.05$ . So the amount of money at time  $t$  is

$$A(t) = 20W + e^{0.05t}(625,015.05 - 20W)$$

We want the account to be depleted when the graduate is 85. So, we want  $A(20) = 0$ . This is the case if

$$\begin{aligned} 20W + e^{0.05 \times 20}(625,015.05 - 20W) &= 0 \implies 20W + e(625,015.05 - 20W) = 0 \\ &\implies 20(e - 1)W = 625,015.05e \\ &\implies W = \frac{625,015.05e}{20(e - 1)} = \$49,437.96 \end{aligned}$$

Example 3.9.17

Example 3.9.18 (Loan Repayment)

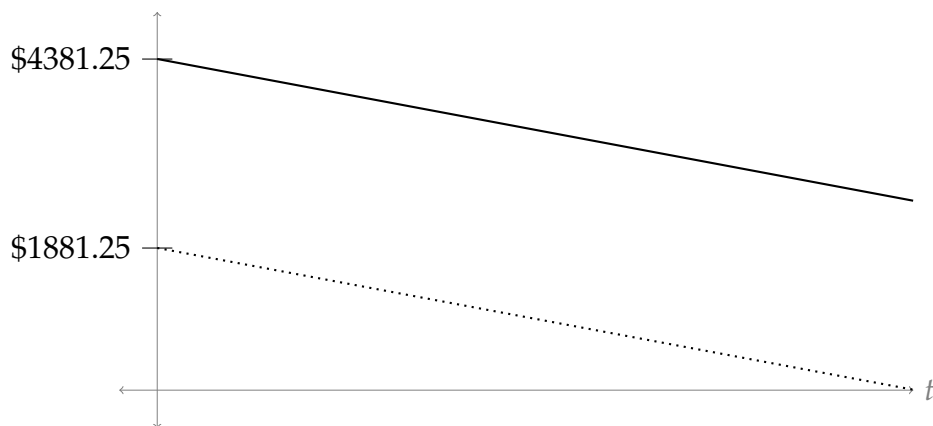
Suppose you borrow \$750,000 from the bank under the following conditions:

1. You make payments to the bank once a month.
2. Every month, you pay 0.25% of the remaining portion of the loan as interest.
3. Your last payment will be after 300 months (25 years)

Let's consider different ways to structure your payments.

**Option 1:** In the most naive option, let's assume you pay off  $\frac{1}{300}$  of the loan each month. Then your payments towards the loan itself are always \$2500. However, the interest you pay changes month by month.

Suppose  $P(t)$  is the amount of the loan still owed to the bank after  $t$  monthly payments. Then  $P(t) = 750,000 - 2500t$ , since after  $t$  months you've repaid \$2500t. The interest you pay in month  $t$  is then  $\frac{25}{100}P(t-1) = \frac{1}{4}(750,000 - 2500(t-1)) = \frac{25}{4}(301 - t)$ .



The dotted line shows monthly interest payments; the solid line shows monthly payments (\$2500+interest).

In this option, your actual monthly payments to the bank vary quite a bit over the 25 years of the loan. If you expect your salary to grow over time, you pay the highest payments early on, when you make the least amount of money. So, this option is not ideal.

**Option 2:** Let's figure out how to pay off the loan in such a way that your monthly payments are the same each month, for all 300 months. Again, let  $P(t)$  be the amount of the loan left to repay the bank after you've made  $t$  monthly payments. Each month, you pay back some portion of the loan, plus an interest payment of  $\frac{.25}{100}P(t)$ .

The amount of the loan you've paid back in month  $t$  is  $P(t-1) - P(t)$ . In particular,

$$P(t-1) - P(t) = -\frac{P(t) - P(t-1)}{1} \approx -\lim_{h \rightarrow 0} \frac{P(t) - P(t-h)}{h} = -P'(t)$$

Thinking of our monthly payments on the loan as *how fast*  $P(t)$  changes, it makes sense to approximate them by the rate of change of  $P$ . The important detail is that  $P(t)$  is decreasing as you pay positive amounts, which is why we use  $-P'(t)$  as the approximation of the amount you paid.

All together, the amount you pay each month is about:

$$\text{loan payment} + \text{interest} = -P'(t) + \frac{.25}{100}P(t)$$

In Option 2, we want this amount to be constant. Let's call that constant monthly payment  $C$ . This gives us a linear differential equation,  $C = -P'(t) + \frac{.25}{100}P(t)$ , or

$$P'(t) = \frac{1}{400}(P(t) - 400C)$$

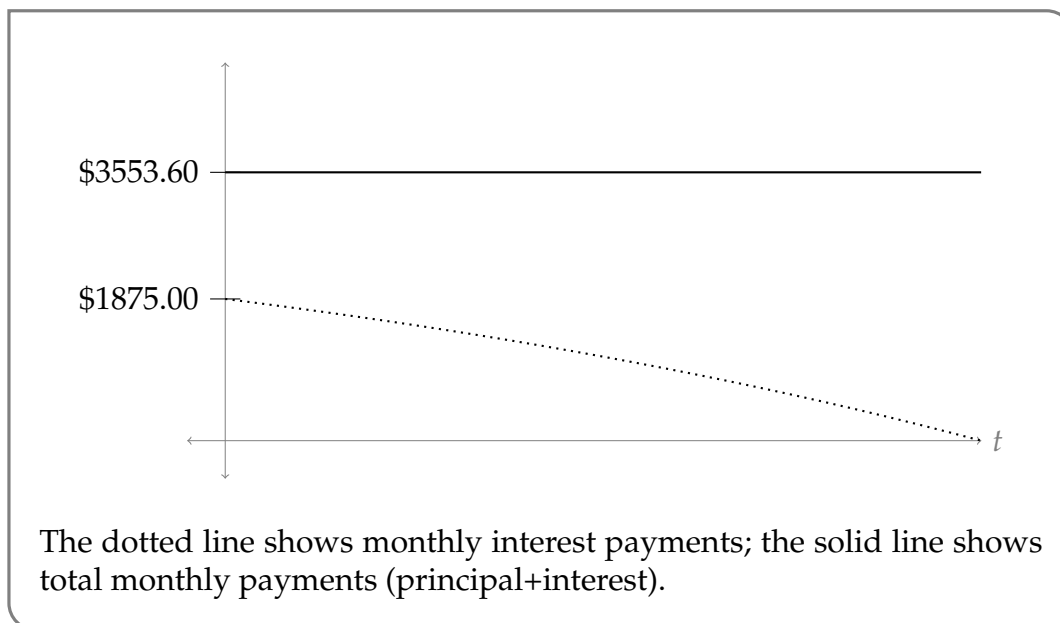
By Theorem 3.9.10,

$$\begin{aligned} P(t) &= (P(0) - 400C)e^{t/400} + 400C \\ &= (750,000 - 400C)e^{t/400} + 400C \end{aligned}$$

To find  $C$ , recall  $P(300) = 0$ .

$$\begin{aligned} 0 &= (750,000 - 400C)e^{300/400} + 400C \\ 400C(e^{3/4} - 1) &= 750,000e^{3/4} \\ C &= \frac{750,000e^{3/4}}{400(e^{3/4} - 1)} = 1875 \frac{e^{3/4}}{(e^{3/4} - 1)} \approx 3553.60 \end{aligned}$$

So, a monthly payment of roughly \$3553.60 would be sufficient to pay off the loan in 25 years. The amount of that monthly payment that goes to the loan itself is about  $P'(t) = (1875 - C)e^{t/400} = \left(\frac{1875}{e^{3/4}-1}\right)e^{t/400}$ , while the rest is interest.



Initial payments consist of roughly equal parts interest and principal. Over time, payments consist of more and more principal, with less and less interest.

We note here that the *Government of Canada mortgage calculator* gives a monthly payment of \$3,549.34 for a mortgage of \$750,000 with annual rate of 3% ( $0.25\% \times 12$ ) and amortization period 25 years. It also mentions that the total interest paid will be \$314,802.37.

Aside from monthly payments, we can also look at the total amount of interest paid in the two scenarios. In Option 1, the amount of interest paid in month  $t$  was  $\frac{25}{4}(301 - t)$ . So, over 300 months, the total interest paid was:

$$\begin{aligned} \sum_{t=1}^{301} \frac{25}{4}(301 - t) &= \frac{25}{4} \left( 301 \cdot 300 - \sum_{t=1}^{301} t \right) \\ &= \frac{25}{4} \left( 301 \cdot 300 - \frac{301 \cdot 302}{2} \right) = \frac{25 \cdot 301 \cdot (300 - 151)}{4} \\ &= 280,306.25 \end{aligned}$$

For Option 2, in month  $t$ , interest paid was approximately  $\frac{1}{400}P(t) = \frac{1875}{e^{3/4}-1} (e^{3/4} - e^{t/400})$ . Total interest is then approximately:

$$\sum_{t=1}^{300} \frac{1875}{e^{3/4}-1} (e^{3/4} - e^{t/400}) = \frac{1875}{e^{3/4}-1} \left( 300 \cdot e^{3/4} - \sum_{t=1}^{300} e^{t/400} \right)$$

For lack of a nice formula, we'll interpret the sum as a Riemann sum. It corresponds to the right-hand Riemann sum for the area under the curve  $f(t) = e^{t/400}$  from  $t = 0$  to  $t = 300$ , using 300 intervals.

$$\begin{aligned}
 &\approx \frac{1875}{e^{3/4} - 1} \left( 300 \cdot e^{3/4} - \int_{t=0}^{300} e^{t/400} dt \right) \\
 &= \frac{1875}{e^{3/4} - 1} \left( 300 \cdot e^{3/4} - \left[ 400e^{t/400} \right]_{t=0}^{300} \right) \\
 &= \frac{1875}{e^{3/4} - 1} \left( 300 \cdot e^{3/4} - 400 \left( e^{3/4} - 1 \right) \right) \\
 &\approx 316081.01
 \end{aligned}$$

Option 2 is more expensive than Option 1.

Example 3.9.18

Chapter 3 of this work was adapted from Chapter 1 and Section 2.4 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International](#) license.

# PROBABILITY

## 4.1▲ Introduction

Before we start, a note. Most terms in this introductory section (probability, event, value) accord pretty well with their usage in everyday life. However, later on in the chapter we will introduce new vocabulary and notation (PDF, CDF,  $\mathbb{E}$ ) whose interpretations are far less obvious. Keeping track of definitions will be key to understanding what's going on. Make flashcards if you have a hard time remembering different terms. If you read a term whose meaning you've forgotten, look it up! If we don't have the same vocabulary, then we aren't speaking the same language – so it will be difficult to explain things.

### 4.1.1 ► Foundational Vocabulary and Notation

#### Definition 4.1.1.

A **probability** is a number between 0 and 1. We interpret it as a likelihood.

If an outcome of an event has probability 1, it will certainly happen. If an outcome has probability 0, it will certainly not happen. If an outcome has probability  $\frac{1}{2}$ , it has an equal chance of happening and not happening. If we have an event with an outcome with probability  $\frac{1}{2}$  and the event happens a large number of times, we expect the outcome to occur in roughly half of those trials.

A random variable is a lot like the ordinary variables you're used to using in functions, in that it is a kind of placeholder that can take on different values. The "random" part of the name explains that the values taken on are the result of an event – some experiment or random process.

**Definition 4.1.2.**

A **random variable** (or just **variable**) is a characteristic or measurement that can be determined for each outcome of some event.

Random variables are usually denoted with capital letters, like  $X$  or  $Y$ . When they correspond to a clear event, we may also give them names like “*Flip*” (for a coin flip) or “*Roll*” (for a dice roll).

Events result in **values**. The event of rolling a dice might result in the value 1; the event of flipping a coin might result in the value heads; and the event of choosing a person might result in the value Parham. We usually use lower-case letters as variables specifying values.

We’ll mostly use events that result in numerical values, although coin flips are a handy experiment as well. Unless otherwise specified, you can assume values will be numbers. (Otherwise our formulas become quite abstract – we won’t ask you to average people or integrate colours.)

**Example 4.1.3**

Let *Roll* be the random variable corresponding to the event of rolling a standard 6-sided dice<sup>1</sup>. *Roll* can result in any of the values 1, 2, 3, 4, 5, or 6.

Suppose we are playing a game and our points are determined by doubling the number rolled. We might write the following:

If  $\text{Roll} = x$ , the number of points earned is  $2x$ .

**Example 4.1.3****Notation 4.1.4.**

We’ll use the shorthand  $\text{Pr}(A)$  to mean “the probability that  $A$  happens.” For example:

$$\text{Pr}(E = x)$$

denotes “the probability that the event  $E$  results in the value  $x$ .”

The equation  $E = x$  can take some getting used to. Remember that  $E$  corresponds to the event (like rolling a dice), while  $x$  corresponds to the outcome of that event (e.g. 5).

**Example 4.1.5**

To express “the probability of a dice roll being 5 is  $1/6$ ,” we write:

$$\text{Pr}(\text{Roll} = 5) = \frac{1}{6}$$

<sup>1</sup> In the interest of clarity, we’ll use “dice” as its own singular (as is common in colloquial English), rather than “die” (which is more standard in academic English).



where  $Roll$  is the event (dice roll), 5 is the value, and  $\frac{1}{6}$  is the probability.

Example 4.1.5

Example 4.1.6

If  $R$  is the roll of a fair dice, then

$$Pr(R = 1 \text{ or } R = 2) = \frac{1}{3}$$

Example 4.1.6

Example 4.1.7

Let  $F$  be the event of flipping a fair coin (that is, a coin that is equally likely to come up heads or tails), and let  $x$  be one of the values "heads" or "tails." Then:

$$Pr(F = x) = \frac{1}{2}$$

Example 4.1.7

**Definition 4.1.8.**

The **sample space** of an event is the set of all possible outcomes. We will use  $\mathcal{S}$  to denote the sample space.

Example 4.1.9

If you roll a standard dice,  $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ .

Example 4.1.9

**Warning 4.1.10.**

This seemingly straightforward definition can cause some confusion, especially when measured data is involved.

For example: suppose our random variable  $X$  is the mileage of a car picked at random out of a parking lot. If there are, say, 100 cars in that lot, then there are (at most) 100 values possible for  $X$  to take on. Unfortunately, we do not know those values.

When we use the context of a supposedly measured variable, we'll pretend that it could be anything theoretically sensible. In the case of the cars, we do know that all of those values will be nonnegative numbers. So, in this case, we could use the sample space  $[0, \infty)$ . Alternately we could say that each of those values is less than some arbitrarily huge number like  $10^{12}$  km<sup>2</sup> and use the sample space  $[0, 10^{12}]$ .

**Example 4.1.11 (A Probabilistic Model in Linguistics)**

These introductory concepts are enough to start understanding probabilistic models in a wide array of fields. Here we'll consider a paragraph from a short paper about people's decisions to change the language they use over time.

The following quote is taken from the article: Abrams, D., Strogatz, S. Modelling the dynamics of language death. *Nature* 424, 900 (2003). DOI: <https://doi.org/10.1038/424900a>

Consider a system of two competing languages,  $X$  and  $Y$ , in which the attractiveness of a language increases with both its number of speakers and its perceived status (a parameter that reflects the social or economic opportunities afforded to its speakers). Suppose an individual converts from  $Y$  to  $X$  with a probability, per unit of time, of  $P_{yx}(x, s)$ , where  $x$  is the fraction of the population speaking  $X$ , and  $0 \leq s \leq 1$  is a measure of  $X$ 's relative status. A minimal model for language change is therefore

$$\frac{dx}{dt} = yP_{yx}(x, s) - xP_{xy}(x, s)$$

where  $y = 1 - x$  is the complementary fraction of the population speaking  $Y$  at time  $t$ .

Let's parse this quote in terms of vocabulary that is familiar to us.

- $x$  is the fraction of a population speaking language  $X$  at a given time. So if everyone is speaking  $X$ , then  $x = 1$ ; if half the population is speaking  $x$ , then  $x = \frac{1}{2}$ .

2 that's farther than driving at 100 kph around the clock for one million years, so it's safe to say no car has more mileage than this

- $y$  is the fraction of the population speaking language  $Y$  at a given time. Under the simplified assumptions of the model in the paper, everyone speaks either  $X$  or  $Y$ , but not both, at a particular time. So,  $y = 1 - x$ .
- $\frac{dx}{dt}$  is the rate of change of speakers of language  $X$  over time. So if  $\frac{dx}{dt}$  is positive, then  $\frac{dx}{dt}$  is increasing and so people are changing from language  $Y$  to language  $X$ ; if  $\frac{dx}{dt}$  is negative, then people are changing from language  $X$  to language  $Y$ .
- The random event in question is *a person changing their language*. The three values in its sample space are: person does not change; person changes from  $X$  to  $Y$ ; and person changes from  $Y$  to  $X$ .
- The paper uses notation that is different from this textbook. They write  $P_{yx}$  for “the probability that a person changes from  $Y$  to  $X$ ,” and they write  $P_{xy}$  for “the probability that a person changes from  $X$  to  $Y$ .”
- The probabilities come with arguments:  $P_{yx}(x, s)$  and  $P_{xy}(x, s)$ . The variables inside the parentheses are function variables. How likely someone is to switch languages is not a fixed constant, but rather a function depending on how many people speak the language, and how much status that language is perceived to have. So,  $P_{yx}$  and  $P_{xy}$  are functions of multiple variables, like the functions we worked with in Chapter 2.

Now that we understand all the notation, we can figure out where the equation in the quote came from.

- $x$  increases as people switch from speaking  $Y$  to speaking  $X$ .  $P_{yx}$  is the proportion of speakers of  $Y$  that we expect to change to  $X$ . The number of speakers of  $Y$  is  $y$ . So, we expect  $yP_{yx}$  people to change from  $Y$  to  $X$ .
- $x$  decreases as people switch from speaking  $X$  to speaking  $Y$ .  $P_{xy}$  is the proportion of speakers of  $X$  that we expect to change to  $Y$ . The number of speakers of  $X$  is  $x$ . So, we expect  $xP_{xy}$  people to change from  $X$  to  $Y$ .
- All together, the change in  $x$  is (number of people coming to  $X$  from  $Y$ ) minus (number of people going to  $Y$  from  $X$ ), or

$$\frac{dx}{dt} = yP_{yx} - xP_{xy}$$

which is exactly the equation from the article.

Example 4.1.11

## 4.1.2 ▶ Discrete vs Continuous

The distinction between discrete and continuous variables plays an important role in the way we calculate properties of variables. The formal definition of a continuous variable will have to wait until Definition 4.3.6. For now, we’ll think of continuous simply as the alternative to discrete.

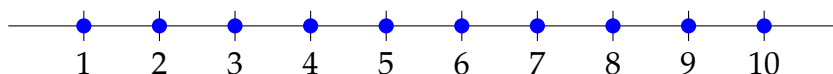
**Definition 4.1.12.**

If the sample space of a random variable can be written as a list (as opposed to existing on a continuum), then the sample space and the random variable are **discrete**.

“Written as a list” is an informal description of the mathematical term “countable,” whose definition<sup>3</sup> is beyond the scope of this class. We’ll explain what we mean with examples.

**Example 4.1.13**

Let  $X$  be the random variable corresponding to choosing a whole number in  $[1, 10]$ .



The values in the sample space can be listed: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. So,  $X$  is discrete.

**Example 4.1.13****Example 4.1.14**

Let  $Y$  be the random variable corresponding to choosing *any* real number number in  $[1, 10]$



$S = [1, 10]$ . There are infinitely many possible values, along a continuum, that could result.  $Y$  is not discrete, it is continuous.

**Example 4.1.14****Example 4.1.15**

For each of the following events, describe the sample space as discrete or continuous, where we are still using “continuous” informally as the opposite of “discrete.”

1. Roll three standard dice, add the values.
2. Number of pets you have.<sup>4</sup>
3. Your exact age at noon today.<sup>5</sup>
4. Volume of a box.

<sup>3</sup> a set is countable if there exists an injective (one-to-one) function from that set to the natural numbers.

<sup>4</sup> We aren’t monsters—there’s no such thing as half a pet.

<sup>5</sup> Imagine you have have perfect precision.

*Solution.*

1. The sample space is discrete,  $\mathcal{S} = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18\}$ .
2. This is also discrete. The number of pets you have is a whole number: 0, 1, 2, 3, etc.
3. You can be any age from 0 to, say, 500 years. If we have exact precision, your age is a real number: you might be 19.015 years old, or 19.016 years old, or somewhere in between. So with exact precision, we're taking numbers along a continuum. This is a continuous sample space.
4. Similar to the above, if we have exact precision, our answer can be any non-negative real number, so this is a continuous sample space.

Example 4.1.15

When our variables are describing physical processes, the line between discrete and continuous can be somewhat blurry. For example, suppose we're measuring the amount of water in a container. Volumes in general exist on continuums (or continua), like the volume of a box in Example 4.1.15, so we could think of this as a continuous sample space. Alternately, we could think of the amount of water as a discrete quantity, because the number of molecules is in integer.

Remembering back to our definition of the definite integral (Definition 3.1.8), we approximated a curvy area with lots of small rectangles. In a similar way, continuous sample spaces can be approximated with discrete sample spaces. The reason we need the distinction is less important for actual measurements, and more important for deciding how to perform calculations. You'll see as we progress through the chapter that some calculations only make sense in one type of variable, and not the other.

### 4.1.3 ▶ Combining Events

#### Definition 4.1.16.

Two outcomes of an event are **disjoint** if no value in the sample space can be described by both outcomes.

For example, consider the event of rolling a dice, corresponding as usual to the discrete random variable  $X$ . The outcomes  $X = 1$  and  $X = 2$  are disjoint, because no dice roll will result in both of them being true. On the other hand, the outcomes  $X > 2$  and " $X$  is even" are not disjoint, because a roll of 4 or 6 makes both of them true.

Example 4.1.17

Let  $X$  be a continuous random variable with sample space  $[0, 10]$ . For each collection of outcomes below, decide whether the outcomes are disjoint or not.

1.  $X < 5$ ;  $X \geq 5$

2.  $X \geq 9$ ;  $X \geq 8$
3.  $1 < X < 2$ ;  $X$  even;  $X$  odd

*Solution.*

1. These are disjoint; no number is both less than five, and also greater than or equal to five.
2. These are not disjoint:  $X = 9$ , for example, makes both true. (So does  $X = 9.5$ ,  $X = 10$ , etc.)
3. These are disjoint. If  $X$  is an integer, then it is even or odd but not both, and it is not in the interval  $(1, 2)$ . If  $X$  is in the interval  $(1, 2)$ , then it is not an integer, so it is not even and not odd.

Example 4.1.17

**Theorem 4.1.18.**

Suppose  $A$  and  $B$  represent disjoint outcomes of the same event. Then

$$Pr(A \text{ happens OR } B \text{ happens}) = Pr(A \text{ happens}) + Pr(B \text{ happens})$$

**Warning 4.1.19.**

In a mathematical context, “or” has a slightly different meaning from its colloquial use. When we say “ $A$  or  $B$ ,” we mean “ $A$ ,  $B$ , or both.”

An old joke is that a mathematician is told they may have a peanut butter cookie or a chocolate cookie, and takes one of each.

*Proof.* This comes from the interpretation of a probability as the proportion of trials where an outcome occurs. If  $A$  and  $B$  never occur at the same trial, then the proportion where one or the other occurs is simply the sum of the proportions where one occurs.  $\square$

Example 4.1.20

If  $X$  is the random variable corresponding to a dice throw, then  $Pr(X \leq 3) = Pr(X = 1 \text{ OR } X = 2 \text{ OR } X = 3)$ . Since the events  $X = 1$ ,  $X = 2$ , and  $X = 3$  are disjoint, this probability is equal to:

$$Pr(X = 1) + Pr(X = 2) + Pr(X = 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Example 4.1.20

Example 4.1.21

Suppose there is a lottery where you pick five numbers, and you win a prize if at least three of your five picks accord with the winning five numbers. Suppose you know that the probability of matching exactly three numbers is  $\frac{1}{100}$ ; the probability of matching exactly four numbers is  $\frac{1}{1000}$ ; and the probability of matching exactly five numbers is  $\frac{1}{10000}$ .

Then the probability of winning something is the probability of matching 3, 4, or 5 numbers:  $\frac{1}{100} + \frac{1}{1000} + \frac{1}{10000}$ .

Example 4.1.21

Example 4.1.22

Suppose for the province of British Columbia, the probability that a randomly chosen adult resident will apply for employment insurance (EI) benefits in 2021 is  $\frac{3}{100}$ , while the probability that a randomly chosen adult resident will be laid off from their job in 2021 is  $\frac{7}{100}$ .

True or false: the probability that a randomly chosen adult resident will apply for EI or be laid off is  $\frac{1}{10}$ .

*Solution.* Not necessarily true (and almost certainly false). These are not disjoint events, so Theorem 4.1.18 does not apply.

Example 4.1.22

#### 4.1.4 ▶ Equally Likely Outcomes

**Equally likely** means that each outcome of a discrete-valued experiment occurs with equal probability. For example, if you roll a fair, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head ( $H$ ) and a Tail ( $T$ ) are equally likely to occur.

Example 4.1.23

Suppose you roll one fair six-sided die, with the numbers  $\{1, 2, 3, 4, 5, 6\}$  on its faces, and you need to roll at least 5 to win a game.

There are two values that win you the game, 5 and 6. Each is expected to occur  $\frac{1}{6}$  of the time. So,

$$Pr(X \geq 5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, roughly (if not exactly)  $\frac{1}{3}$  of the rolls would result in an outcome of "at least five".

## Example 4.1.23

It is important to realize that in many situations, the outcomes are not equally likely. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur.

The continuous analog of “equally likely” is *uniformly distributed*.

**Definition 4.1.24.**

Intuitively, a continuous random variable is **uniformly distributed** on an interval if the variable doesn't favour one region of the interval over any other region.

More formally:

Let  $X$  be a continuous random variable.  $X$  is **uniformly distributed** on the interval  $[a, b]$  if there exists some constant  $c$  such that for any interval  $[a_1, b_1]$  in  $[a, b]$ ,  $\Pr(a_1 \leq X \leq b_1) = c(b_1 - a_1)$ . That is, the probability that  $X$  is in a particular interval within  $[a, b]$  depends only on the length of that interval.

## Example 4.1.25

Suppose  $X$  is a continuous random variable that is **uniformly distributed** on the interval  $[0, 10]$ .

The intervals  $[2, 5]$  and  $[3, 6]$  have the same length, so  $\Pr(2 \leq X \leq 5) = \Pr(3 \leq X \leq 6)$ .

The intervals  $[2, 3]$ ,  $[3, 4]$ , and  $[7, 8]$  have equal length, so  $\Pr(2 \leq X \leq 3) = \Pr(3 \leq X \leq 4) = \Pr(7 \leq X \leq 8)$ . So,  $X$  is twice as likely to be in the interval  $[2, 4]$  as it is to be in the interval  $[7, 8]$ .

## Example 4.1.25

**Corollary 4.1.26.**

Suppose  $X$  is a continuous random variable that is uniformly distributed on its sample space, the interval  $[a, b]$ . Then for any interval  $[a_1, b_1]$  with  $a \leq a_1 \leq b_1 \leq b$ ,

$$\Pr(a_1 \leq X \leq b_1) = \frac{b_1 - a_1}{b - a}$$

That is, the probability that  $X$  is in the interval  $[a_1, b_1]$  is the ratio of the length of that interval to the sample space interval.



*Proof.* Since the sample space of  $X$  is  $[a, b]$ ,

$$\Pr(a \leq X \leq b) = 1$$

Since  $X$  is uniformly distributed on  $[a, b]$ , there exists a constant  $c$  such that  $\Pr(a_1 \leq X \leq b_1) = c(b_1 - a_1)$  for any interval  $[a_1, b_1]$  inside the interval  $[a, b]$ . So,

$$1 = \Pr(a \leq X \leq b) = c(b - a) \implies c = \frac{1}{b - a}$$

Then:

$$\Pr(a_1 \leq X \leq b_1) = c(b_1 - a_1) = \frac{b_1 - a_1}{b - a}$$

□

#### Example 4.1.27

Let  $X$  be a continuous random variable that is uniformly distributed on its sample space, the interval  $[0, 10]$ . What is  $\Pr(7 \leq X \leq 9)$ ?

*Solution.*

The interval  $[7, 9]$  has length 2; the sample space interval  $[0, 10]$  has length 10. So,

$$\Pr(7 \leq X \leq 9) = \frac{2}{10} = \frac{1}{5}$$

#### Example 4.1.27

#### Example 4.1.28

Let  $X$  be a continuous random variable that is uniformly distributed across its sample space  $[-8, 17]$ . Calculate the probabilities below.

1.  $\Pr(1 \leq X \leq 2)$
2.  $\Pr(-5 \leq X)$
3.  $\Pr(-10 \leq X \leq 10)$

*Solution.*

1. By Corollary 4.1.26,  $\Pr(1 \leq X \leq 2) = \frac{2-1}{17-(-8)} = \frac{1}{25}$
2. Since  $X$  only takes on values in its sample space  $[-8, 17]$ :  $\Pr(-5 \leq X) = \Pr(-5 \leq X \leq 17)$ . By Corollary 4.1.26,  $\Pr(-5 \leq X \leq 17) = \frac{17-(-5)}{17-(-8)} = \frac{22}{25}$
3. Since  $X$  only takes on values in its sample space  $[-8, 17]$ :  $\Pr(-10 \leq X \leq 10) = \Pr(-8 \leq X \leq 10)$ . Now the interval  $[-8, 10]$  is inside our sample space, unlike the interval  $[-10, 10]$ , so we can apply Corollary 4.1.26.  
 $\Pr(-8 \leq X \leq 10) = \frac{10-(-8)}{17-(-8)} = \frac{18}{25}$

Example 4.1.28

Example 4.1.29

Suppose the continuous variable  $X$  is the age of a randomly chosen living person, measured in years with exact precision. Then  $X$  is more likely to be near 50 than it is to be near 110. So,  $X$  is *not* uniformly distributed.

Example 4.1.29

## 4.2▲ Probability Mass Function (PMF)

For a discrete random variable, the description of the probabilities of all events in its sample space is its probability mass function.

### Definition 4.2.1.

A **probability mass function (PMF)** for a discrete random variable  $X$  is the function  $f(x)$  from  $\mathbb{R}$  to  $[0, 1]$ , where

$$f(x) = \Pr(X = x)$$

Often  $f(x)$  formally takes the form of a piecewise function, e.g.

$$f(x) = \begin{cases} \frac{1}{6} & x = 1, 2, 3, 4, 5, \text{ or } 6 \\ 0 & \text{else} \end{cases}$$

for a dice roll. In particular,  $f(x) = 0$  for every value  $x$  not in the sample space of the random variable.

**Notation 4.2.2.**

Rather than writing a piecewise function every time, we will represent the probability mass function (PMF) of a random variable  $X$  using a table, set up like this:

$x$	$Pr(X = x)$
1	$Pr(X = 1) = \frac{1}{6}$
2	$Pr(X = 2) = \frac{1}{6}$
3	$Pr(X = 3) = \frac{1}{6}$
4	$Pr(X = 4) = \frac{1}{6}$
5	$Pr(X = 5) = \frac{1}{6}$
6	$Pr(X = 6) = \frac{1}{6}$

where events not in the sample space do not show up in the table.

**Theorem 4.2.3.**

For any probability mass function (PMF)  $f(x)$ :

- $f(x)$  is a number in  $[0, 1]$  for every real number  $x$ , and
- the sum of the probabilities of all values in the sample space is one.

*Proof.* •  $f(x) = Pr(X = x)$ , and probabilities are defined (4.1.1) to be in  $[0, 1]$ .

- $X$  is guaranteed to be a value in the sample space, so using Theorem 4.1.18,  $1 = Pr(X \text{ is in } \mathcal{S})$ , which is the sum of  $Pr(X = x)$  for every  $x$  in  $\mathcal{S}$ .

□

**Warning 4.2.4.**

If  $\Pr(X = x) = 0$ , then often  $f(x)$  is omitted from the probability mass function (PMF). For example, in Notation 4.2.2, we don't bother writing that  $\Pr(X = 17) = 0$ , or  $\Pr(X = 18) = 0$ , or  $\Pr(X = 107.4) = 0$ .

You might also see this omission in a probability mass function (PMF) written as a piecewise function. Instead of writing this:

$$f(x) = \begin{cases} \frac{1}{6} & x = 1, 2, 3, 4, 5, \text{ or } 6 \\ 0 & \text{else} \end{cases}$$

you could write this:

$$f(x) = \frac{1}{6} \text{ for all } x \text{ in } \{1, 2, 3, 4, 5, 6\}.$$

**Notation 4.2.5.**

The notation

$$\sum_x f(x)$$

means we take the sum of  $f(x)$  for every value  $x$  in some set. In this context, that set is understood to be the sample space of a random variable.

We may also omit the bound, writing simply

$$\sum f(x)$$

The sample space may or may not be a range of integers, which is why this notation is slightly different from the sigma notation we use in the other chapters of this book.

**Example 4.2.6**

A child psychologist is interested in the number of times per night a newborn baby's crying wakes its parent. The record this number for 100 different parents.

$x$	number of parents woken $x$ times
0	5
1	5
2	40
3	23
4	13
5	10
6	0
7	3
8	1

Suppose we pick one parent uniformly at random. Let  $X$  be the number of times per night that parent is woken up.  $X$  takes on the values 0, 1, 2, 3, 4, 5, 6, 7, 8.

$x$	$P(X = x)$
0	$P(X = 0) = \frac{5}{100}$
1	$P(X = 1) = \frac{5}{100}$
2	$P(X = 2) = \frac{40}{100}$
3	$P(X = 3) = \frac{23}{100}$
4	$P(X = 4) = \frac{13}{100}$
5	$P(X = 5) = \frac{10}{100}$
6	$P(X = 6) = \frac{0}{100}$
7	$P(X = 7) = \frac{3}{100}$
8	$P(X = 8) = \frac{1}{100}$

This is a probability mass function (PMF) because:

- Each probability is in the interval  $[0, 1]$ .
- The sum of the probabilities is one, that is,

$$\sum_x Pr(X = x) = \frac{5}{100} + \frac{5}{100} + \frac{40}{100} + \frac{23}{100} + \frac{13}{100} + \frac{10}{100} + \frac{0}{100} + \frac{3}{100} + \frac{1}{100} = 1$$

Example 4.2.6

Example 4.2.7

A hospital researcher is interested in the number of times an average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained.

Let  $X$  be the number of times a patient rings the nurse during a 12-hour shift.

$x$	$P(X = x)$
0	$P(x = 0) = \frac{4}{50}$
1	$P(x = 1) = \frac{8}{50}$
2	$P(x = 2) = \frac{16}{50}$
3	$P(x = 3) = \frac{14}{50}$
4	$P(x = 4) = \frac{6}{50}$
5	$P(x = 5) = \frac{2}{50}$

Why is this a probability mass function (PMF)?

*Solution.* Yes, each probability is a number from the interval  $[0, 1]$ , and their sum is 1:

$$\sum_x Pr(X = x) = \frac{4}{50} + \frac{8}{50} + \frac{16}{50} + \frac{14}{50} + \frac{6}{50} + \frac{2}{50} = 1$$

Example 4.2.7

Example 4.2.8

Suppose Nancy has classes three days a week. She attends classes three days a week 80% of the time, two days 15% of the time, one day 4% of the time, and no days 1% of the time. Suppose one week is randomly selected.

- a. What is the random variable in this case? Call it  $X$ .
- b. What values does  $X$  take on?
- c. Construct a probability mass table (called a PM table) like the one in Example 4.2.6. The table should have two columns, labelled  $x$  and  $P(X = x)$ .
- d. What does the  $P(x)$  column sum to?

*Solution.*

- a.  $X$  is the number of days Nancy went to class on the randomly selected week.
- b. From the description,  $X$  has sample space  $\{0, 1, 2, 3\}$ .

$x$	$P(X = x)$
0	$P(x = 0) = 0.01$
c. 1	$P(x = 1) = 0.04$
2	$P(x = 2) = 0.15$
3	$P(x = 3) = 0.8$

d.  $\sum_{x=0}^3 Pr(X = x) = 0.01 + 0.04 + 0.15 + 0.8 = 1$ , which accords with Definition 4.2.1.

Example 4.2.8

Example 4.2.9

Suppose a person is chosen at random from a group. Let  $X$  be the discrete random variable describing the number of siblings that person has, and suppose the following probabilities hold for  $X$ :

$x$	$P(X = x)$
0	$P(x = 0) = 0.25$
1	$P(x = 1) = 0.3$
2	$P(x = 2) = 0.25$
3	$P(x = 3) = 0.1$
4	$P(x = 4) = 0.05$

If we sum up the right column, we get

$$\sum_{x=0}^4 Pr(X = x) = 0.25 + 0.3 + 0.25 + 0.1 + 0.05 = 0.95 < 1$$

That tells us this is not a probability mass function (PMF). Since all probabilities are numbers in the interval  $[0, 1]$ , it must be the case that we haven't summed over all values in the sample space. That is, in our sample of people, there must be some people who haven't been described here, e.g. people with more than four siblings. (Indeed, these folks would make up 5% of the group.)

Example 4.2.9

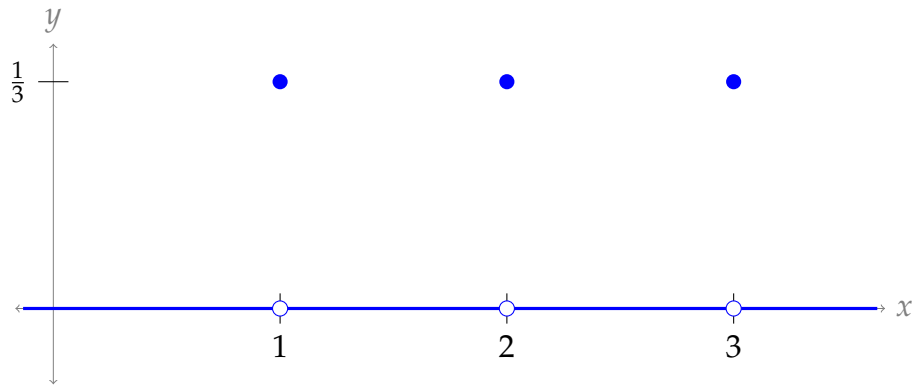
### 4.2.1 ► Limitations of Probability Mass Function (PMF)

Let's imagine we're choosing numbers from 1 to 3 uniformly at random. The number chosen is called  $X$ . In the examples below, we'll investigate the difference between *discrete* choices and a *continuous* choice.

- If we choose an *integer* from 1 to 3 uniformly at random, then our probability mass function (PMF) is:

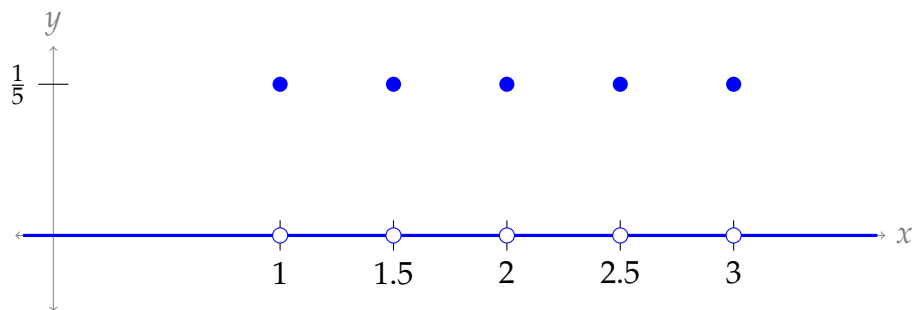
$$f(x) = Pr(X = x) = \begin{cases} \frac{1}{3} & x \text{ is } 1, 2, \text{ or } 3 \\ 0 & \text{otherwise} \end{cases}$$

Graphed, it looks like this:



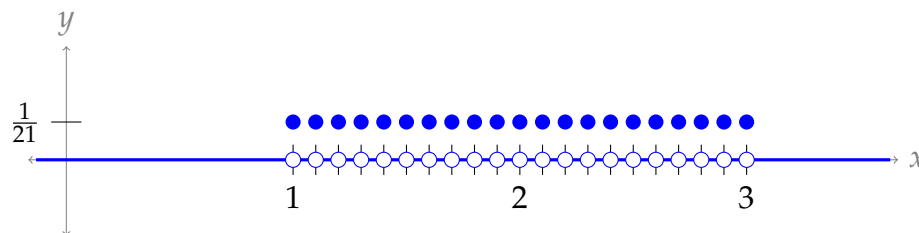
Furthermore,  $Pr(X \leq 2) = \frac{2}{3}$ .

- If we choose a number from 1 to 3 uniformly at random, choosing numbers of the form  $\frac{n}{2}$  where  $n$  is an integer (e.g. we can choose 1 and 1.5 but not 1.15), then there are five numbers to choose from. Our probability mass function (PMF) is:



For example,  $Pr(X = 2) = \frac{1}{5}$  and  $Pr(X = 7) = 0$ . Furthermore,  $Pr(X \leq 2) = \frac{3}{5}$ .

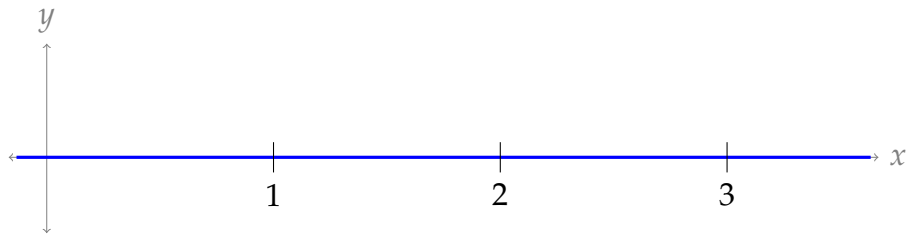
- If we choose a number from 1 to 3 uniformly at random, choosing numbers of the form  $\frac{n}{10}$  where  $n$  is an integer (e.g. we can choose 1 and 1.1 but not 1.15), then there are 21 numbers to choose from. Our probability mass function (PMF) is:



Furthermore,  $Pr(X \leq 2) = \frac{11}{21}$ .

- So far, all the examples have been discrete systems. What if we want  $X$  to be a continuous variable? We want to be able to choose *any real number* from 1 to 3. In this case, there are *infinitely many* numbers to choose from. So, the probability of choosing any of them is... zero!





This is a problem! We know we're choosing numbers between 1 and 3, but we have  $Pr(X = 1) = 0$  and  $Pr(X = 4) = 0$ . So the probability mass function (PMF) is not useful for describing continuous random variables. We need a different tool.

On the other hand, it's easy to imagine that  $Pr(X \leq 2) = \frac{1}{2}$ . So somehow this calculation didn't break when we moved from a discrete system to a continuous system.

### 4.3▲ Cumulative Distribution Function (CDF)

In the final example above,  $Pr(X = x) = 0$  for every number  $x$ . Looking at individual numbers isn't very enlightening. Instead of looking at individual numbers, then, we can look at *ranges* of numbers. These behave more nicely. With that in mind, we make the following definition.

#### Definition 4.3.1.

Given a random variable  $X$ , the **cumulative distribution function (CDF)** of  $X$ , usually denoted by  $F(x)$ , is

$$Pr(X \leq x)$$

This might seem like a weirdly specific definition. Secretly, our main purpose in creating this function is to use it as a tool to define two other things: a continuous random variable, and the probability density function. Our motivation for defining the cumulative distribution function (CDF) may lie with continuous random variables, but the definition applies to discrete random variables as well.

#### Example 4.3.2

Suppose a random variable  $X$  has cumulative distribution function (CDF)  $F(x)$ , given by

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{10^4} & 0 \leq x \leq 100 \\ 1 & x > 100 \end{cases}$$

Evaluate the following probabilities:

1.  $Pr(X \leq 50)$
2.  $Pr(X > 10)$

3.  $Pr(X \leq 0)$
4.  $Pr(X \geq 200)$
5.  $Pr(10 < X \leq 20)$

*Solution.*

1. By Definition 4.3.1:  $Pr(X \leq 50) = F(50)$ ; using the formula given for  $F(x)$ , this is  $\frac{50^2}{10^4} = \frac{1}{4}$

2.  $Pr(X > 10)$  is the probability that  $X$  is *not* less than or equal to 10, so

$$Pr(X > 10) = 1 - Pr(X \leq 10) = 1 - F(10) = 1 - \frac{10^2}{10^4} = 1 - \frac{1}{100} = \frac{99}{100}$$

3.  $Pr(X \leq 0) = F(0) = 0$ . Note this tells us that  $X$  never takes negative values.

4. Note  $Pr(X \leq 100) = F(100) = 1$ . That tells us that  $X$  always takes values less than or equal to 100. Combined with our last note, that means the only values  $X$  ever takes are in the interval  $[0, 100]$ . So,  $Pr(X \geq 200) = 0$ .

5. We can think of the interval  $(10, 20]$  as “numbers that are less than equal to 20 *except* numbers less than or equal to 10.” We rewrite  $Pr(10 < X \leq 20)$  in a manner similar to Problem 2:

$$\begin{aligned} Pr(10 < X \leq 20) &= Pr(X \leq 20 \text{ and } X \not\leq 10) = Pr(X \leq 20) - Pr(X \leq 10) \\ &= F(20) - F(10) = \frac{20^2}{10^4} - \frac{10^2}{10^4} = \frac{3}{100} \end{aligned}$$

Example 4.3.2

The ideas in the calculations of 2 and 5 above give us the following corollary.

**Corollary 4.3.3.**

Let  $X$  be a random variable with cumulative distribution function (CDF)  $F(x)$ . Then

1.  $Pr(X > a) = 1 - F(a)$ , and
2.  $Pr(a < X \leq b) = F(b) - F(a)$

*Proof.* The probability  $Pr(X > a)$  is the probability that  $X$  is *not* less than or equal to  $a$ , so

$$Pr(X > a) = 1 - Pr(X \leq a) = 1 - F(a)$$

The probability  $Pr(a < X \leq b)$  is the probability that  $X$  is less than or equal to  $b$  and  $X$  is *not* less than or equal to  $a$ .

$$Pr(a < X \leq b) = Pr(X \leq b) - Pr(X \leq a) = F(b) - F(a)$$

□

Example 4.3.4

Let  $X$  be a discrete random variable with probability mass function (PMF) below.

$x$	$Pr(X = x)$
10	$\frac{1}{16}$
20	$\frac{3}{16}$
30	$\frac{5}{16}$
40	$\frac{7}{16}$

Note that the only values taken on by  $X$  are the numbers 10, 20, 30, and 40.

Let  $F(X)$  be the cumulative distribution function (CDF) of  $X$ .

- By Definition 4.3.1,  $F(10) = Pr(X \leq 10)$ . Looking at the probability mass function (PMF),  $X \leq 10$  only when  $X = 10$ , which happens  $\frac{1}{16}$  of the time. So, in this case:

$$F(10) = Pr(X \leq 10) = Pr(X = 10) = \frac{1}{16}$$

- By Definition 4.3.1,  $F(20) = Pr(X \leq 20)$ . Looking at the probability mass function (PMF),  $X \leq 20$  only when  $X = 10$  or  $X = 20$ , which happens  $\frac{1}{16} + \frac{3}{16}$  of the time. So, in this case:

$$F(20) = Pr(X \leq 20) = Pr(X = 10 \text{ or } X = 20) = \frac{1}{16} + \frac{3}{16} = \frac{4}{16} = \frac{1}{4}$$

- Similarly,

$$F(30) = Pr(X \leq 30) = Pr(X = 10 \text{ or } X = 20 \text{ or } X = 30) = \frac{1}{16} + \frac{3}{16} + \frac{5}{16} = \frac{9}{16}$$

and

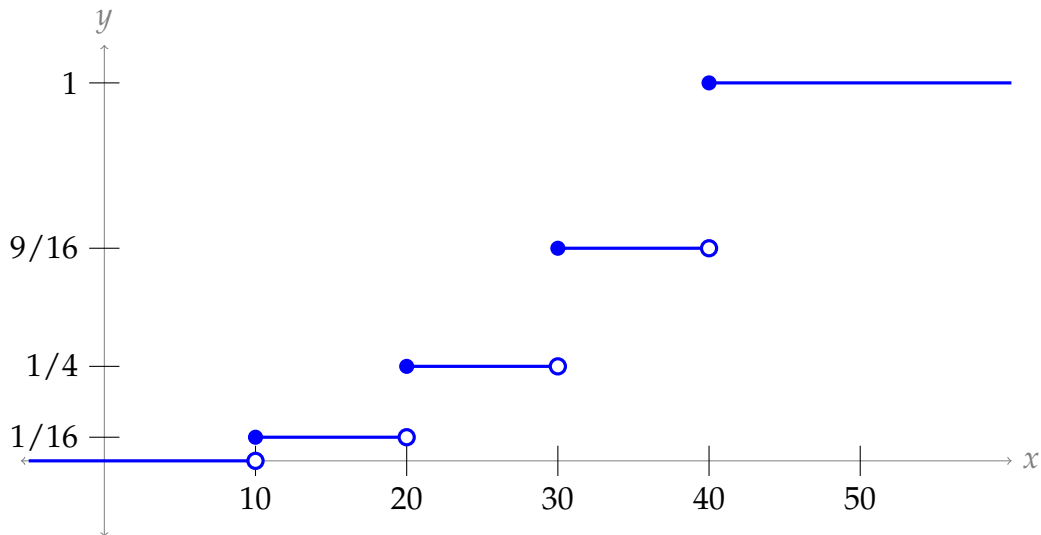
$$\begin{aligned} F(40) &= Pr(X \leq 40) = Pr(X = 10 \text{ or } X = 20 \text{ or } X = 30 \text{ or } X = 40) \\ &= \frac{1}{16} + \frac{3}{16} + \frac{5}{16} + \frac{7}{16} = 1 \end{aligned}$$

The cumulative distribution function (CDF) has all real numbers as its domain, so we aren't quite finished determining the function  $F(x)$ . However, after doing a few examples, the rest of the function is easy to figure out.

- $F(9) = Pr(X \leq 9) = 0$ ; indeed,  $F(x) = 0$  for all  $x < 10$ .
- $F(11) = Pr(X \leq 11) = Pr(X = 10)$ , since 10 is the only number ever taken by  $X$  that is less than or equal to 11. So,  $F(11) = F(10) = \frac{1}{16}$ . Indeed,  $F(x) = F(10)$  for all  $x$  in the interval  $[10, 20)$

- Following this line of reasoning:

$$F(x) = \begin{cases} 0 & x < 10 \\ \frac{1}{16} & 10 \leq x < 20 \\ \frac{1}{4} & 20 \leq x < 30 \\ \frac{9}{16} & 30 \leq x < 40 \\ 1 & 40 \leq x \end{cases}$$



Example 4.3.4

Example 4.3.5

Let  $U$  be a random variable that is chosen uniformly at random from all real numbers in the interval  $[0, 1]$ . Understanding the cumulative distribution function (CDF)  $F(x)$  of  $U$  can help us understand what “uniformly” means in this case.

As we saw in section 4.2.1, it’s not useful to note that  $Pr(U = x)$  is the same for every number in  $[0, 1]$ , because that probability is 0. We can get at the meaning of “uniformly” in a more useful way by examining *ranges* of numbers.

If we were to divide our interval<sup>6</sup> in half, then the uniformity of distribution tells us that half the time,  $U$  is in one half, and half the time,  $U$  is in the other half. In particular,

$$Pr\left(0 \leq U \leq \frac{1}{2}\right) = Pr\left(\frac{1}{2} \leq U \leq 1\right) = \frac{1}{2}$$

So, for the cumulative distribution function (CDF),

$$F\left(\frac{1}{2}\right) = \frac{1}{2}$$

6 Since  $Pr(U = \frac{1}{2}) = 0$ , it won’t matter whether we use the interval  $[0, 1/2]$  or  $[0, 1/2)$ .

If we were to divide our interval into equal tenths, then the uniformity of distribution tells us that  $U$  should fall in each interval one-tenth of the time. For example,

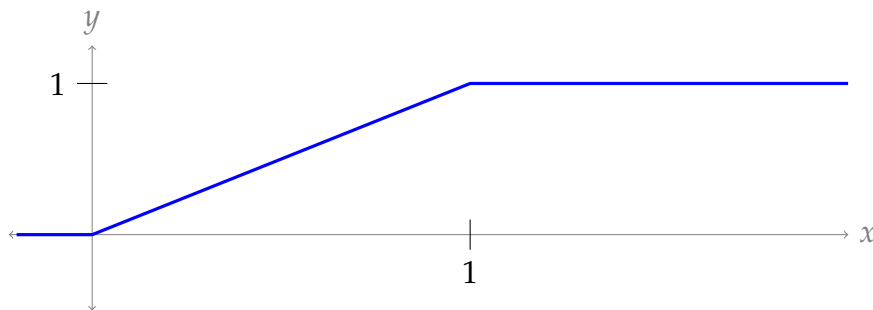
$$\Pr\left(0 \leq U \leq \frac{1}{10}\right) = \Pr\left(\frac{1}{10} \leq U \leq \frac{2}{10}\right) = \Pr\left(\frac{2}{10} \leq U \leq \frac{3}{10}\right) = \frac{1}{10}$$

So,

$$F\left(\frac{1}{10}\right) = \frac{1}{10}$$

In general, if  $x$  is a number in the interval  $[0, 1]$ , then  $x$  describes the proportion of  $[0, 1]$  taken up by the interval  $[0, x]$ , so  $F(x) = x$ .

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & 1 < x \end{cases}$$



Example 4.3.5

The cumulative distribution function (CDF) will give us our actual definition of a continuous random variable. Thinking of “continuous” as the opposite of “discrete” is not sufficiently accurate.

**Definition 4.3.6.**

A random variable  $X$  is **continuous** if its cumulative distribution function (CDF) is continuous.

Example 4.3.7

The random variables from Examples 4.3.2 and 4.3.5 are continuous random variables. The random variable from Examples 4.3.4 is not a continuous random variable.

Example 4.3.7

**Corollary 4.3.8.**

Let  $X$  be a continuous random variable. For any real number  $a$ ,

$$\Pr(X = a) = 0$$

Furthermore,

$$\Pr(X < a) = \Pr(X \leq a) \quad \text{and} \quad \Pr(X > a) = \Pr(X \geq a)$$

*Proof.* Let  $F(x)$  be the cumulative distribution function (CDF) of  $X$ .

$$\lim_{x \rightarrow a^-} F(x) = \lim_{x \rightarrow a^-} \Pr(X \leq x) = \Pr(X < a)$$

By the definition of a continuous function,

$$\lim_{x \rightarrow a^-} F(x) = F(a)$$

So,

$$\begin{aligned} \Pr(X < a) &= \Pr(X \leq a) \\ \Pr(X \leq a) - \Pr(X < a) &= 0 \\ \Pr(X = a) &= 0 \end{aligned}$$

The “furthermore” statements follow.

$$\begin{aligned} \Pr(X \leq a) &= \Pr(X < a) + \Pr(X = a) = \Pr(X < a) \\ \Pr(X \geq a) &= \Pr(X > a) + \Pr(X = a) = \Pr(X > a) \end{aligned}$$

□

**Example 4.3.9**

$V$  is a number chosen at random from all real numbers in the intervals  $[-3, -1]$  or  $[1, 3]$  as follows:

- First, a fair 6-sided dice is rolled. If the outcome of the roll is 1 or 2, then  $V$  is chosen to be in the interval  $[-3, -1]$ . If the outcome of the roll is 3, 4, 5, or 6, then  $V$  is chosen to be in the interval  $[1, 3]$ .
- Within the selected interval,  $V$  is chosen uniformly at random.

Determine the cumulative distribution function (CDF) of  $V$  and decide whether or not  $V$  is continuous.

*Solution.* From the first step, we see that  $V$  is in the interval  $[-3, -1]$  one-third of the time, and in the interval  $[1, 3]$  two-thirds of the time.

$$\Pr(-3 \leq V \leq -1) = \frac{1}{3}, \quad \Pr(1 \leq V \leq 3) = \frac{2}{3}$$

Within these intervals,  $V$  has a uniform distribution. As in Example 4.3.5, we consider intervals. For example,  $V$  is equally likely to be in the interval  $[-3, -2]$  and the interval  $[-2, -1]$ . So,

$$\Pr(-3 \leq V \leq -2) = \Pr(-2 \leq V \leq -1)$$

Also,

$$\Pr(-3 \leq V \leq -2) + \Pr(-2 \leq V \leq -1) = \Pr(-3 \leq V \leq -1) = \frac{1}{3}$$

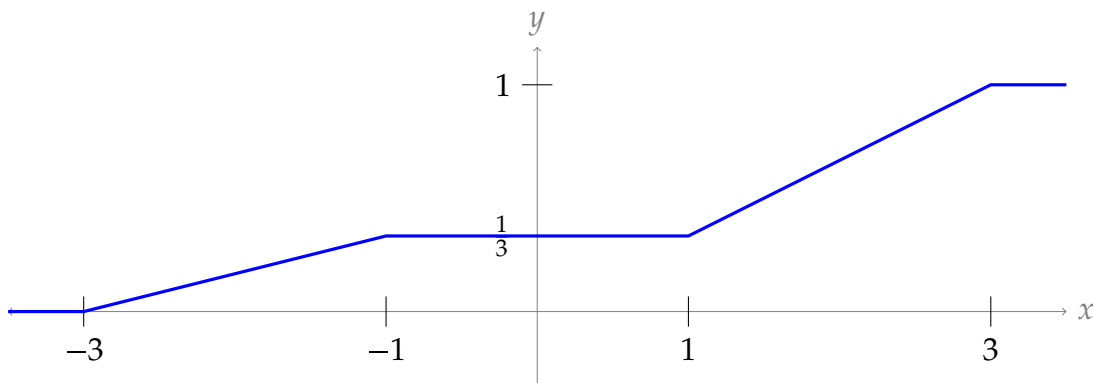
So,

$$\Pr(-3 \leq V \leq -2) = \Pr(-2 \leq V \leq -1) = \frac{1}{6}$$

Following the reasoning in Example 4.3.5, we see on the interval  $[-3, -1]$ , the function  $F(x)$  is a straight line from  $F(-3) = 0$  to  $F(-1) = \frac{1}{3}$ .

When  $-1 < x < 1$ , then  $F(x) = \Pr(X \leq x) = \Pr(X \leq -1) = F(-1)$ , since no values of  $V$  are ever less than 1 without also being less than or equal to  $-1$ . Then, by Corollary 4.3.8, also  $F(1) = \Pr(X \leq 1) = \Pr(X < 1) = \Pr(X \leq -1) = F(-1)$ .

On the interval  $[1, 3]$ ,  $V$  is uniformly distributed. Following the familiar line of reasoning, the function  $F(x)$  is a straight line from  $F(1) = \frac{1}{3}$  to  $F(3) = 1$ . All together:



Using the graph, we can find  $F(x)$  in equation form:

$$F(x) = \begin{cases} 0 & x < -3 \\ \frac{x}{6} + \frac{1}{2} & -3 \leq x < -1 \\ \frac{1}{3} & -1 \leq x < 1 \\ \frac{1}{3}x & 1 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

Since  $F(x)$  is continuous,  $V$  is a continuous random variable.

Example 4.3.9

**Corollary 4.3.10** (Properties of the cumulative distribution functions (CDF)).

If  $F(x)$  is the cumulative distribution function (CDF) of a continuous random variable  $X$ , then:

1.  $0 \leq F(x) \leq 1$  for all real  $x$
2.  $F(x)$  is nondecreasing
3.  $\lim_{x \rightarrow \infty} F(x) = 1$
4.  $\lim_{x \rightarrow -\infty} F(x) = 0$

*Proof.* 1.  $F(x)$  is a probability, and all probabilities are numbers between 0 and 1.

2. Suppose  $a < b$ .

$$F(a) = Pr(X \leq a) \leq Pr(X \leq a) + Pr(a < X \leq b) = Pr(X \leq b) = F(b)$$

That is, if  $a < b$ , then  $F(a) \leq F(b)$ .

3. Rather than a rigorous proof, we offer the following hand-wavy intuition: if infinity were a number, we'd expect  $F(\infty) = Pr(X \leq \infty) = 1$ .

4. Rather than a rigorous proof, we offer the following hand-wavy intuition: if negative infinity were a number, we'd expect  $F(-\infty) = Pr(X \leq -\infty) = 0$ . □

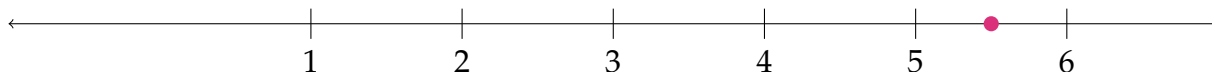
## 4.4▲ Probability Density

### 4.4.1 ▶ Density Diagrams

We're going to introduce a tool for visualizing random processes that will hopefully help topics in continuous random variables be more intuitive. We'll call that tool a *density diagram*.

Let  $X$  be some continuous random variable. If  $X$  is a process (like choosing the height, in feet, of a random student), we can imagine performing that process again and again and again. Suppose we do just that. Every time we get a new value of  $X$ , we put a mark on a number line. For example:

1. The first randomly-chosen student has height 5.5 feet:



2. The second randomly-chosen student has height 6.1 feet:





3. The third randomly-chosen student has height 5.2 feet:



4. The third randomly-chosen student has height 5.4 feet:



5. After 20 choices, our results might look like this:



6. After 100 choices, our marks would start being so close together, they would be indistinguishable, so we might choose to make the marks slightly transparent. Then darker regions represent ranges where more heights have been chosen.



**Example 4.4.1**

The continuous random variable  $V$  from Example 4.3.9 is chosen as follows:

- First, a fair 6-sided dice is rolled. If the outcome of the roll is 1 or 2, then  $V$  is chosen to be in the interval  $[-3, -1]$ . If the outcome of the roll is 3, 4, 5, or 6, then  $V$  is chosen to be in the interval  $[1, 3]$ .
- Within the selected interval,  $V$  is chosen uniformly at random.

If we were to perform this trial 100 times, and record the number each time, our results might look like this:



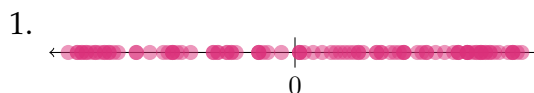
The marks (trial outcomes) are twice as dense on the right interval. Inside the right interval, and inside the left interval, the marks are fairly evenly distributed.

Example 4.4.1

Example 4.4.2

Match the density diagrams to the variable descriptions so that every description corresponds to exactly one density diagram.

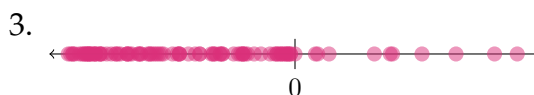
A.  $Pr(X \leq 0) = Pr(X \geq 0)$ .



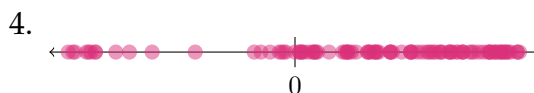
B.  $X$  is uniformly distributed.



C.  $Pr(X \leq 0) < Pr(X \geq 0)$ .



D.  $Pr(X \leq 0) > Pr(X \geq 0)$ .



*Solution.* In both 1 and 2, it seems like (roughly) the same number of trials resulted in positive and negative values of  $X$ . So in both cases, A holds. However, in 2, the distribution is not uniform: trials are more likely to have large absolute values than to be near 0. So, we match B to 1 and A to 2.

In 3, more trials gave  $X \leq 0$  than  $X \geq 0$ , so we match that to D.

In 4, more trials gave  $X \geq 0$  than  $X \leq 0$ , so we match that to C.

Example 4.4.2

### 4.4.2 ▶ Probability Density Function (PDF)

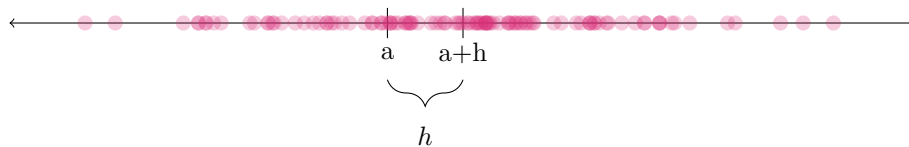
As we saw in Corollary 4.3.8, if  $X$  is a continuous random variable, then  $Pr(X = a) = 0$  for *any* real number  $a$ . However, that doesn't mean that all number ranges are equally likely. In Example 4.3.5, we saw a continuous random variable  $U$  that only existed in the range  $[0, 1]$ ; so getting a value near  $\frac{1}{2}$  is more likely than getting a value near 2.

When looking at density diagrams, areas with more "hits" show up as having a higher *density* of marks. This idea will be central to this section: measuring the *density* of a continuous random variable.

A usual definition of density is something like

$$\frac{\text{how much stuff}}{\text{how much space}}$$

Population density might be measured in people per square kilometre, liquid density might be measured in grams per mL, etc. Probability density follows a similar pattern: we'll measure *how likely a variable is to be in a given interval* and divide it by the size (length) of that interval.



Suppose the density diagram above represents some continuous random variable  $X$ , and we want to measure the probability density near the indicated point  $a$ . We start by defining a small interval around  $a$ . As is tradition, we take the interval between  $a$  and  $a + h$ , where  $h$  is some small<sup>7</sup> real number.

It doesn't make sense to count the marks in this interval, since the actual number will change as we do different trials, so instead we measure the likeliness our random variable is to be in this interval:  $Pr(a \leq X \leq a + h)$ . The length of the interval is  $h$ . So, our probability density around  $a$  is:

$$\frac{Pr(a \leq X \leq a + h)}{h}$$

If  $F(x)$  is our cumulative distribution function (CDF), then we can re-write this using Corollaries 4.3.3 and 4.3.8.

$$= \frac{F(a + h) - F(a)}{h}$$

Since we only consider small values of  $h$ , we recognize the definition of a derivative.

$$\lim_{h \rightarrow 0} \frac{F(a + h) - F(a)}{h} = F'(a)$$

This motivates our definition of the probability density function (PDF) of a continuous random variable. Probability mass functions (PMFs) and probability density functions (PDFs) serve similar purposes: describing which values a variable tends to take on.

#### Definition 4.4.3.

The **probability density function (PDF)** of a continuous random variable, usually written  $f(x)$ , is the derivative of the cumulative distribution function (CDF), where it exists.

The observant reader will note that the conventional use of  $F(x)$  as an antiderivative of  $f(x)$  squares nicely with our use of  $F(x)$  for a cumulative distribution function (CDF) and  $f(x)$  for a probability density function (PDF).

<sup>7</sup> By "small," we mean  $|h| \approx 0$ . In the discussion that follows, we're considering the case  $h > 0$ ; the case  $h < 0$  proceeds in the same way.

**Warning 4.4.4.**

Some textbooks use the term “probability distribution function” instead of “probability mass function,” and then use the abbreviation PDF in both a continuous and a discrete context. This reflects the similar roles probability density functions (PDFs) and probability mass functions (PMFs) play.

**Example 4.4.5**

In Example 4.3.2, we considered a continuous random variable with cumulative distribution function (CDF) given by

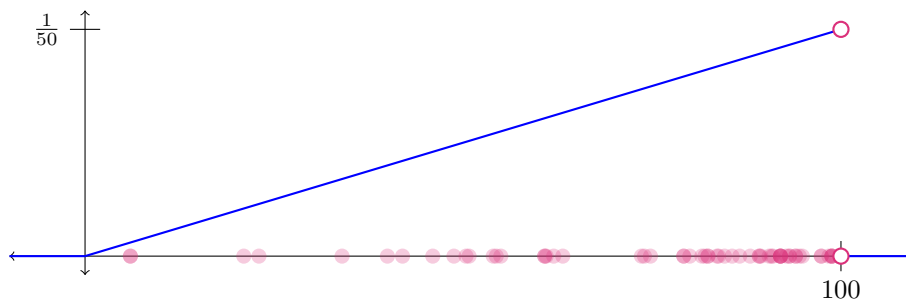
$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{10^4} & 0 \leq x \leq 100 \\ 1 & x > 100 \end{cases}$$

The probability density function (PDF) of this variable is  $F'(x)$ , namely

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{5000} & 0 \leq x < 100 \\ 0 & x > 100 \end{cases}$$

Translating  $f(x)$  into a density diagram, to help build intuition about the behaviour of this variable, we expect to see

- no marks except in the interval  $[0, 100]$ , and
- an increasing density of marks from left to right on the interval  $[0, 100]$ .

**Example 4.4.5**

**Notation 4.4.6.**

As with probability mass functions (PMFs), it is common to suppress the regions where a probability density function (PDF) is zero or doesn't exist. Instead of writing

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{5000} & 0 \leq x < 100 \\ 0 & x > 100 \end{cases}$$

as in Example 4.4.5, we may also write

$$f(x) = \begin{cases} \frac{x}{5000} & 0 \leq x < 100 \end{cases}$$

and it is understood that  $f(x)$  is zero or doesn't exist when  $x$  *not* in the interval  $[0, 100)$ .

Another time-saving measure is to use the words "else" or "otherwise" in a piecewise-defined function. In the context of this function:

$$f(x) = \begin{cases} \frac{x}{5000} & 0 \leq x < 100 \\ 0 & \text{else} \end{cases}$$

"else" means "for all values of  $x$  *other than* the ones that have already been defined," i.e. for all values of  $x$  outside the interval  $[0, 100)$ .

**Example 4.4.7**

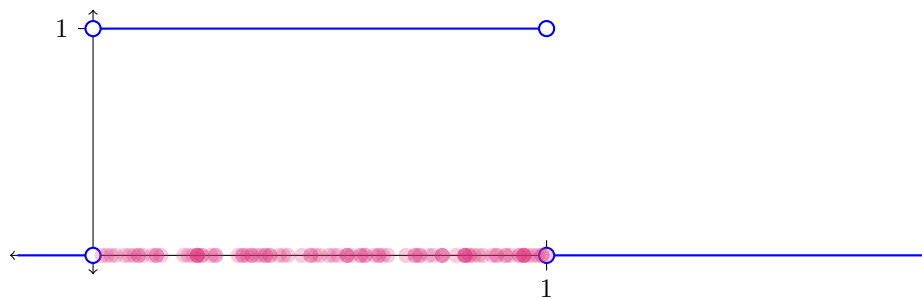
In Example 4.3.5, we considered a continuous random variable with cumulative distribution function (CDF) given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1 & 1 < x \end{cases}$$

The probability density function (PDF) of this variable is  $F'(x)$ , namely

$$f(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 < x < 1 \\ 0 & 1 < x \end{cases}$$

Notice that the density is constant on the interval  $(0, 1)$ . This is a hallmark of uniformly distributed variables: in the interval in question, no one region is denser than any other region.



Note  $f(x)$  is not defined at  $x = 0$  and  $x = 1$  because  $F(x)$  is not differentiable at these points<sup>8</sup>.

Example 4.4.7

Example 4.4.8

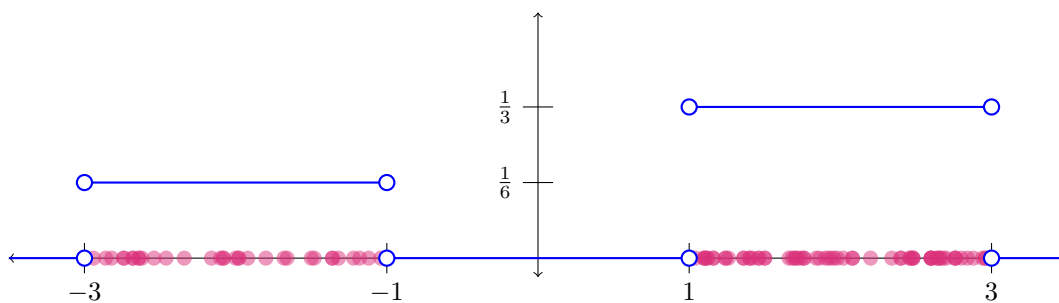
In Example 4.3.9, we considered a continuous random variable with cumulative distribution function (CDF) given by the function

$$F(x) = \begin{cases} 0 & x < -3 \\ \frac{x}{6} + \frac{1}{2} & -3 \leq x < -1 \\ \frac{1}{3} & -1 \leq x < 1 \\ \frac{1}{3}x & 1 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

The probability density function (PDF) of this variable is  $F'(x)$ , namely

$$f(x) = \begin{cases} 0 & x < -3 \\ \frac{1}{6} & -3 < x < -1 \\ 0 & -1 < x < 1 \\ \frac{1}{3} & 1 < x < 3 \\ 0 & x > 3 \end{cases}$$

The places where the probability density function (PDF) is 0 are telling: these are the regions where our variable never reaches (impossible to occur).



8 You can see this by comparing the right and left limits of the limit definition of the derivative of  $F(x)$  at these points.

Our intuition about  $f(x)$  is that higher  $f(x)$  means more “hits” near  $x$ . In the density diagram above,  $f(2) > f(-2)$ , and indeed the marks are denser in the area 2 than near -2.

Example 4.4.8

**Warning 4.4.9.**

Let  $f(x)$  be the probability density function (PDF) of a continuous random variable. If  $f(x) > f(y)$ , it is not correct to say that  $x$  is more likely than  $y$ , because it is still the case that  $\Pr(X = x) = \Pr(X = y) = 0$ .

**Corollary 4.4.10.**

From Definition 4.4.3, given a continuous random variable  $X$  with probability density function (PDF)  $f(x)$ :

1.  $\Pr(a \leq X \leq b) = \int_a^b f(x) dx$
2.  $f(x) \geq 0$  for all real  $x$  in the domain of  $f$ .
3.  $\int_{-\infty}^{\infty} f(x) = 1$

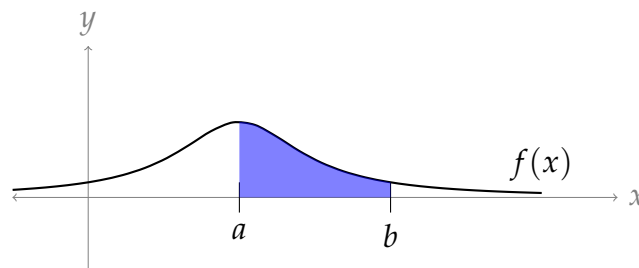
*Proof.* 1. By Corollaries 4.3.3 and 4.3.8,  $\Pr(a \leq X \leq b) = F(b) - F(a)$ . By the Fundamental Theorem of Calculus Part 2,  $\int_a^b f(x) dx = F(b) - F(a)$ , since  $F'(x) = f(x)$ .

For this property, we are glossing over some details in assuming  $f(x)$  exists on  $(a, b)$ . If it does not, then we partition  $(a, b)$  into intervals where it does exist, and apply the Fundamental Theorem of Calculus to those intervals separately.

2. By Part 2 of Corollary 4.3.10,  $F(x)$  is nondecreasing, so its derivative is nonnegative.
3. From the property above,  $\int_{-\infty}^{\infty} f(x) = \Pr(-\infty \leq X \leq \infty) = 1$

□

The first property of Corollary 4.4.10 is a key piece of intuition for working with probability density functions (PDFs): the probability density function (PDF) of a continuous random variable  $X$  is a function  $f(x)$  with the property that the area under the curve of  $f(x)$  from  $a$  to  $b$  is equal to the probability that  $X$  lies between  $a$  and  $b$ .



shaded area:  $\Pr(a \leq X \leq b)$

Example 4.4.11

A continuous random variable  $X$  has probability density function (PDF)

$$f(x) = \frac{a}{x^2 + 1}$$

for some constant  $a$ .

1. Find  $a$ .
2. Find  $Pr(0 \leq X \leq 10)$ .
3. Find the cumulative distribution function (CDF) of  $X$ .

*Solution.*

1. By Corollary 4.4.10,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \frac{a}{x^2 + 1} dx = a \int_{-\infty}^{\infty} \frac{1}{x^2 + 1} dx \\ &= a \left[ \lim_{b \rightarrow -\infty} \int_b^0 \frac{1}{x^2 + 1} dx + \lim_{c \rightarrow \infty} \int_0^c \frac{1}{x^2 + 1} dx \right] \\ &= a \left[ \left( \lim_{b \rightarrow -\infty} (\arctan 0 - \arctan b) \right) + \left( \lim_{c \rightarrow \infty} (\arctan c - \arctan 0) \right) \right] \\ &= a \left[ 0 - \frac{-\pi}{2} + \frac{\pi}{2} + 0 \right] = a \cdot \pi \end{aligned}$$

So,  $a = \frac{1}{\pi}$ .

2. By Corollary 4.4.10,

$$Pr(0 \leq X \leq 10) = \int_0^{10} f(x) dx = \int_0^{10} \frac{1/\pi}{x^2 + 1} dx = \frac{1}{\pi} [\arctan 10 - \arctan 0] = \frac{\arctan(10)}{\pi} \approx 0.47$$

Note: because  $f(x)$  has even symmetry, we know  $Pr(X \leq 0) = Pr(X \geq 0) = \frac{1}{2}$ . Also,  $Pr(0 \leq X \leq 10) \leq Pr(0 \leq X)$ , so it stands to reason that our answer would be less than one-half.

3. Let  $F(x)$  be the cumulative distribution function (CDF) of  $X$ .

$$\begin{aligned} F(x) &= Pr(X \leq x) && \text{(definition of CDF)} \\ &= Pr(-\infty < X \leq x) \\ &= \int_{-\infty}^x f(t) dt = \lim_{b \rightarrow -\infty} \int_b^x \frac{1/\pi}{t^2 + 1} dt \\ &= \frac{1}{\pi} \lim_{b \rightarrow -\infty} [\arctan x - \arctan b] = \frac{1}{\pi} \left[ \arctan x + \frac{\pi}{2} \right] \\ &= \frac{1}{\pi} \arctan x + \frac{1}{2} \end{aligned}$$

Note: it's nice to do a quick sanity check by comparing  $F(x)$  to the properties of a cumulative distribution function (CDF) in Corollary 4.3.10. This is a great way to root out calculation errors, sign errors, and so on.



We can formalize the last part of the previous exercise as a corollary to Corollary 4.4.10.

**Corollary 4.4.12.**

Let  $X$  be a continuous random variable with probability density function (PDF)  $f(x)$ . Then the cumulative distribution function (CDF) of  $X$  is

$$F(x) = \int_{-\infty}^x f(t) dt$$

*Proof.* The CDF is defined as  $F(x) = Pr(X \leq x)$ , i.e.  $Pr(-\infty < X \leq x)$ . By Corollary 4.4.10,

$$Pr(-\infty < X \leq x) = \int_{-\infty}^x f(t) dt$$

□

## 4.5▲ Expected Value

### 4.5.1 ► Motivation: Long-Term Average

Suppose I throw a 4-sided dice a large number of times, and record the number that comes up each time. What will the average (mean) of those numbers be?

To calculate the mean, I'll add up the results of my rolls and divide by the number of rolls I took.

$$\text{mean} = \frac{(\text{result of first roll}) + (\text{result of second roll}) + \cdots + (\text{result of last roll})}{\text{total number of rolls}}$$

The numerator will consist of the numbers 1 through 4, since these are the numbers resulting from a 4-sided dice roll. Let's regroup the numerator so we add up all the 1s first, then all the 2s second, etc.

$$\begin{aligned}
 &= \frac{(1 + 1 + \cdots) + (2 + 2 + \cdots) + (3 + 3 + \cdots) + (4 + 4 + \cdots)}{\text{total number of rolls}} \\
 &= \frac{(1 + 1 + \cdots)}{\text{total rolls}} + \frac{(2 + 2 + \cdots)}{\text{total rolls}} + \frac{(3 + 3 + \cdots)}{\text{total rolls}} + \frac{(4 + 4 + \cdots)}{\text{total rolls}} \\
 &= \frac{1 \cdot (\text{number of times 1 was rolled})}{\text{total rolls}} + \frac{2 \cdot (\text{number of times 2 was rolled})}{\text{total rolls}} \\
 &+ \frac{3 \cdot (\text{number of times 3 was rolled})}{\text{total rolls}} + \frac{4 \cdot (\text{number of times 4 was rolled})}{\text{total rolls}} \\
 &= 1 \cdot (\text{proportion of rolls resulting in 1}) + 2 \cdot (\text{proportion of rolls resulting in 2}) \\
 &+ 3 \cdot (\text{proportion of rolls resulting in 3}) + 4 \cdot (\text{proportion of rolls resulting in 4})
 \end{aligned}$$

If we've rolled the dice a large number of times, we expect the proportion of rolls resulting in 1 to closely approximate  $Pr(X = 1)$ , and so on.

$$\approx 1 \cdot Pr(X = 1) + 2 \cdot Pr(X = 2) + 3 \cdot Pr(X = 3) + 4 \cdot Pr(X = 4) = \sum_{x=1}^4 x \cdot Pr(X = x)$$

This calculation, what we expect to have as our average if we perform the dice roll a large number of times, motivates Definition 4.5.1 below.

## 4.5.2 ▶ Definition and Examples

The **expected value** or **expectation** of a random variables is often referred to as the “**long-term**” average. This means that over the long term of doing an experiment over and over, you would expect this average.

### Definition 4.5.1.

Given a discrete random variable  $X$ , the **expected value of  $X$** , denoted  $\mathbb{E}(X)$ , is given by

$$\sum x \cdot Pr(X = x)$$

where the sum is taken over every possible value of  $X$ .

Given a continuous random variable  $X$  with probability density function (PDF)  $f(x)$ , the expected value of  $X$  is given by

$$\int_{-\infty}^{\infty} x \cdot f(x) dx$$

Note the similarities between the continuous and discrete cases. A sum in the discrete cases turns into an integral in the continuous case;  $Pr(X = x)$  turns into the probability density function (PDF)  $f(x)$ ; and “every possible value of  $X$ ” turns into the range  $(-\infty, \infty)$ .

Example 4.5.2

In Example 4.2.6, we saw the following probability mass function (PMF) for the random variable  $X$ :

$x$	$P(X = x)$
0	$P(x = 0) = \frac{5}{100}$
1	$P(x = 1) = \frac{5}{100}$
2	$P(x = 2) = \frac{40}{100}$
3	$P(x = 3) = \frac{23}{100}$
4	$P(x = 4) = \frac{13}{100}$
5	$P(x = 5) = \frac{10}{100}$
6	$P(x = 6) = \frac{0}{100}$
7	$P(x = 7) = \frac{3}{100}$
8	$P(x = 8) = \frac{1}{100}$

The expected value of this discrete random variable is

$$\begin{aligned}
 \mathbb{E}(X) &= \sum_{x=0}^8 x \cdot Pr(X = x) \\
 &= 0 \cdot Pr(X = 0) + 1 \cdot Pr(X = 1) + 2 \cdot Pr(X = 2) + 3 \cdot Pr(X = 3) + 4 \cdot Pr(X = 4) \\
 &\quad + 5 \cdot Pr(X = 5) + 6 \cdot Pr(X = 6) + 7 \cdot Pr(X = 7) + 8 \cdot Pr(X = 8) \\
 &= 0 \cdot \frac{5}{100} + 1 \cdot \frac{5}{100} + 2 \cdot \frac{40}{100} + 3 \cdot \frac{23}{100} + 4 \cdot \frac{13}{100} + 5 \cdot \frac{10}{100} + 6 \cdot \frac{0}{100} + 7 \cdot \frac{3}{100} + 8 \cdot \frac{1}{100} \\
 &= \frac{285}{100} = 2.85
 \end{aligned}$$

The most literal interpretation of expected value in this context is this:

Suppose we choose a parent from a list at random many times, and each time record the number of awakenings,  $X$ . After a large number of trials, we expect the average of these  $X$  values to approach 2.85.

We can also interpret the calculation like this:

The average number of times a parent was woken up in our trial was 2.85.

Of course, no parent was woken up exactly 2.85 times in the night. Expected values refer to *averages*, and do not necessarily accord well with individual trials.

Example 4.5.2

Probability does not describe the short-term results of an experiment. It gives information about what can be expected *in the long term*. The **Law of Large Numbers** states that, as the number of trials in a probability experiment increases, the difference between the theoretical probability of an event and the relative frequency approaches zero (the theoretical probability and the relative frequency get closer and closer together).

Example 4.5.3

Suppose we flip a fair coin a large number of times. We want to record the average number of times the flip resulted in heads.

Let  $X$  be the random variable corresponding to a coin flip, with  $X = 1$  when the flip is heads and  $X = 0$  when the flip is tails. Using these assignments, if we add up the values of  $X$  from each experiment, that sum tells us how many flips were heads. The expected value of  $X$  is

$$\mathbb{E}(X) = \sum_{x=1}^2 x \cdot Pr(X = x) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

Consider interpreting the expected value as a long-term average, using the law of large numbers. If we were to flip a fair coin a large number of times, we would expect the average value of  $X$  to be  $\frac{1}{2}$ . That is, we would expect roughly  $\frac{1}{2}$  of the tosses to result in heads.

In 2009, intrepid undergraduate students at Berkeley tossed coins 40,000 times<sup>9</sup>. The tosses resulted in 20,217 heads. The fraction of coin tosses resulting in heads, therefore, was

$$\frac{20,217}{40,000} = 0.505425$$

which is indeed fairly close to  $\frac{1}{2}$ .

Example 4.5.3

Example 4.5.4

Let  $X$  be a continuous random variable with probability density function (PDF)

$$f(x) = ax^2(10 - x), \quad 0 \leq x \leq 10$$

where  $a$  is a constant.

Find  $a$  and  $\mathbb{E}(X)$ .

<sup>9</sup> A writeup is here: [https://www.stat.berkeley.edu/~aldous/Real-World/coin\\_tosses.html](https://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html). They were actually trying to determine whether the starting orientation of a coin had an impact on the result of a toss. There's actually a bit of a cart-before-the-horse problem in using this example here: if we tossed a coin a large number of times and it didn't result in very close to half heads and half tails, the conclusion would be that the probability of tossing heads was not actually  $\frac{1}{2}$ .

*Solution.*

From Corollary 4.4.10 part 3:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x) dx = 0 + \int_0^{10} ax^2(10-x) dx = a \int_0^{10} (10x^2 - x^3) dx \\ &= a \left[ \frac{10}{3}x^3 - \frac{1}{4}x^4 \right]_0^{10} = a \left[ \frac{10^4}{3} - \frac{10^4}{4} \right] = a \frac{10^4}{12} \\ a &= \frac{12}{10^4} \end{aligned}$$

From Definition 4.5.1,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Note where  $f(x) = 0$ , we have  $\int_a^b x \cdot f(x) dx = \int_a^b 0 dx = 0$ .

$$\begin{aligned} &= 0 + \int_0^{10} ax^3(10-x) dx = a \int_0^{10} (10x^3 - x^4) dx \\ &= a \left[ \frac{10}{4}x^4 - \frac{1}{5}x^5 \right]_0^{10} = a \left[ \frac{10^5}{4} - \frac{10^5}{5} \right] = a \frac{10^5}{20} \\ &= \frac{12}{10^4} \cdot \frac{10^5}{20} = 6 \end{aligned}$$

Example 4.5.4

Example 4.5.5

Suppose  $Y$  is a continuous random variable with probability density function (PDF)

$$f(x) = e^x, \quad x \leq 0$$

Find  $\mathbb{E}(Y)$ .

*Solution.*

From Definition 4.5.1,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} x \cdot f(x) dx = 0 + \int_{-\infty}^0 x \cdot e^x dx = \lim_{a \rightarrow -\infty} \left[ \int_a^0 x \cdot e^x dx \right]$$

We use integration by parts with  $u = x$ ,  $dv = e^x dx$ ;  $du = dx$ ,  $v = e^x$

$$= \lim_{a \rightarrow -\infty} \left[ [xe^x]_a^0 - \int_a^0 e^x dx \right] = \lim_{a \rightarrow -\infty} \left[ -ae^a - [e^x]_a^0 \right] = \lim_{a \rightarrow -\infty} [-ae^a - 1 + e^a]$$

Note  $\lim_{a \rightarrow -\infty} e^a = 0$ , so  $\lim_{a \rightarrow -\infty} -ae^a$  has the indeterminate form  $0 \cdot \infty$ . We use l'Hôpital's rule.

$$= \lim_{a \rightarrow -\infty} \left[ \underbrace{\frac{-a}{e^{-a}}}_{\substack{\text{num} \rightarrow \infty \\ \text{den} \rightarrow \infty}} \right] - 1 + 0 = \lim_{a \rightarrow -\infty} \left[ \frac{1}{e^{-a}} \right] - 1 = \lim_{a \rightarrow -\infty} [e^a] - 1 = -1$$

So,  $\mathbb{E}(Y) = -1$ .

Example 4.5.5

Example 4.5.6

Let  $Z$  be a continuous random variable with probability density function (PDF)  $f(x) = \frac{1}{x^2}$ ,  $x \geq 1$ . Find  $\mathbb{E}(Z)$ .

*Solution.*

From Definition 4.5.1,

$$\begin{aligned} \mathbb{E}(Z) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = 0 + \int_1^{\infty} x \cdot x^{-2} dx = \int_1^{\infty} x^{-1} dx \\ &= \lim_{b \rightarrow \infty} \left[ \int_1^b x^{-1} dx \right] = \lim_{b \rightarrow \infty} [\ln b] = \infty \end{aligned}$$

It is sometimes the case that the expectation of a continuous random variable is infinite. How should we interpret that?

A random variable  $Z$  with the given probability density function (PDF) has sample space is  $[1, \infty)$ . It takes on finite values, but there is no limit to how large those values can be. (It is true that smaller values are more likely, since  $f(x) = x^{-2}$  is a decreasing function. However,  $Z$  also takes on extremely large values from time to time.)  $\mathbb{E}(Z) = \infty$  tells us that if we run our experiment  $Z$  a lot of times, over time the average will increase without bound.

Example 4.5.6

### 4.5.3 ▶ Checking your Expectation Calculation

The expectation of a random variable has several intuitive properties that can be used to quickly check that your answer is reasonable.

#### Theorem 4.5.7.

Let  $a, b$  be real numbers or  $\pm\infty$  with  $a < b$ . Suppose a (discrete or continuous) random variable  $X$  takes values from the interval  $[a, b]$ . Then  $\mathbb{E}(X)$  will be some number in the interval  $[a, b]$ .

*Proof.* First, suppose  $X$  is continuous, with probability density function (PDF)  $f(x)$ .

$$\mathbb{E}(X) = \int_a^b xf(x)dx \leq \int_a^b b \cdot f(x)dx = b \int_a^b f(x)dx = b$$

$$\mathbb{E}(X) = \int_a^b xf(x)dx \geq \int_a^b a \cdot f(x)dx = a \int_a^b f(x)dx = a$$

Next, suppose  $X$  is discrete

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x) \leq \sum_x b \cdot \Pr(X = x) = b \sum_x \Pr(X = x) = b$$

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x) \geq \sum_x a \cdot \Pr(X = x) = a \sum_x \Pr(X = x) = a$$

□

#### Theorem 4.5.8.

Let  $a$  and  $b$  be real numbers with  $a < b$ . Suppose a continuous (resp. discrete) random variable  $X$  takes values from the interval  $[a, b]$ , and its probability density function (PDF) (resp. probability mass function (PMF)) is *increasing* on the interval  $[a, b]$ . Then  $\mathbb{E}(X) > \frac{a+b}{2}$ .

Similarly, suppose a continuous (resp. discrete) random variable  $X$  takes values from the interval  $[a, b]$ , with  $a < b$ , and its probability density function (PDF) (resp. probability mass function (PMF)) is *decreasing* on the interval  $[a, b]$ . Then  $\mathbb{E}(X) < \frac{a+b}{2}$ .

*Proof.* Intuitively, an increasing  $f(x)$  means we have more high values than low values, so when we average them together, the average will be high. Similarly, decreasing  $f(x)$  means we have more low values than high values, so when we average them together, the average will be low.

More rigorously:

$$\begin{aligned} \int_a^b xf(x)dx - \frac{a+b}{2} &= \int_a^b xf(x)dx - \frac{a+b}{2} \int_a^b f(x)dx \\ &= \int_a^b \left(x - \frac{a+b}{2}\right) f(x)dx \\ &= \int_a^{\frac{a+b}{2}} \left(x - \frac{a+b}{2}\right) f(x)dx + \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) f(x)dx \\ &= \int_{\frac{a+b}{2}}^a \left(\frac{a+b}{2} - x\right) f(x)dx + \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) f(x)dx \end{aligned}$$

Using the substitution  $y = a + b - x$  in the first integral and noting that  $dy = -dx$ ,

$$= - \int_{\frac{a+b}{2}}^b \left(y - \frac{a+b}{2}\right) f(a+b-y)dy + \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) f(x)dx$$

Changing  $y$ 's into  $x$ 's,

$$\begin{aligned} &= - \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) f(a+b-x) dx + \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) f(x) dx \\ &= \int_{\frac{a+b}{2}}^b \left(x - \frac{a+b}{2}\right) [-f(a+b-x) + f(x)] dx \end{aligned}$$

For values of  $x$  in the interval  $\left[\frac{a+b}{2}, b\right]$ , the term  $\left(x - \frac{a+b}{2}\right)$  is positive except at  $x = \frac{a+b}{2}$  where it is zero. So, all together,

$$\int_a^b xf(x) dx - \frac{a+b}{2} = \int_{\frac{a+b}{2}}^b \underbrace{\left(x - \frac{a+b}{2}\right)}_{\text{positive}} [f(a+b-x) - f(x)] dx$$

Furthermore,  $a+b-x \leq a+b - \left(\frac{a+b}{2}\right) = \frac{a+b}{2} \leq x$ .

If  $f(x)$  is increasing, then  $f(a+b-x) < f(x)$ :

$$\begin{aligned} \int_a^b xf(x) dx - \frac{a+b}{2} &= \int_{\frac{a+b}{2}}^b \underbrace{\left(x - \frac{a+b}{2}\right)}_{\text{positive}} \underbrace{[-f(a+b-x) + f(x)]}_{\text{positive}} dx > 0 \\ \text{so } \int_a^b xf(x) dx &> \frac{a+b}{2} \end{aligned}$$

If  $f(x)$  is decreasing, then  $f(a+b-x) > f(x)$  whenever  $x \in \left(\frac{a+b}{2}, b\right]$ :

$$\begin{aligned} \int_a^b xf(x) dx - \frac{a+b}{2} &= \int_{\frac{a+b}{2}}^b \underbrace{\left(x - \frac{a+b}{2}\right)}_{\text{positive}} \underbrace{[-f(a+b-x) + f(x)]}_{\text{negative}} dx > 0 \\ \text{so } \int_a^b xf(x) dx &> \frac{a+b}{2} \end{aligned}$$

The discrete case proceeds in a similar fashion. □

**Example 4.5.9**

Suppose  $X$  is a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} e^{x-a} & \text{if } 1 \leq x \leq 5 \\ 0 & \text{else} \end{cases}$$

for some appropriate constant  $a$ . Using the two theorems in this section, give a range for  $\mathbb{E}(X)$ .



*Solution.*  $X$  only takes values from  $[1, 5]$ , so by Theorem 4.5.7,  $1 \leq \mathbb{E}(X) \leq 5$ .

The probability density function (PDF)  $f(x) = e^{x-a}$  is an increasing function, so by Theorem 4.5.8,  $\mathbb{E}(X) > \frac{5+1}{2} = 3$ .

So,  $\mathbb{E}(X)$  is in the interval  $(3, 5]$ .

Note: There is a unique value of  $a$  for which  $f(x)$  is a probability density function (PDF). It is the value of  $a$  that satisfies the following equality:

$$1 = \int_1^5 e^{x-a} dx$$

i.e.  $a = \ln(e^5 - e)$ .

Example 4.5.9

Example 4.5.10

You calculate expected values for the various random variables described below. Which of the values can you immediately, with very little computation, say are wrong? Which seem reasonable?

1.  $W$  is a random variable that takes values from  $[4, 5]$ , and you calculate  $\mathbb{E}(W) = 4.75$ .
2.  $X$  is a random variable that takes values from  $[-1, 0]$ , and you calculate  $\mathbb{E}(X) = 0.5$ .
3.  $Y$  is a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} \frac{1}{x} & 1 \leq x \leq e \\ 0 & \text{else} \end{cases}$$

and you calculate  $\mathbb{E}(Y) = 1.9$ .

4.  $Z$  is a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} \frac{1}{x^2} & 1 \leq x \\ 0 & x < 1 \end{cases}$$

and you calculate  $\mathbb{E}(Z) = -1$ .

5.  $A$  is a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} \frac{1}{x^3} & \frac{1}{\sqrt{2}} \leq x \\ 0 & x < \frac{1}{\sqrt{2}} \end{cases}$$

and you calculate  $\mathbb{E}(A) = \sqrt{2}$ .

*Solution.* From Theorem 4.5.7,  $\mathbb{E}(X)$  and  $\mathbb{E}(Z)$  are incorrect.

The PDF of  $Y$  is decreasing on  $[1, e]$ , so  $\mathbb{E}(Y) < \frac{e+1}{2} \approx 3.72 = 1.85$  by Theorem 4.5.8. Therefore the result  $\mathbb{E}(Y) = 1.9$  is incorrect.

For  $\mathbb{E}(W)$ , we don't have enough information to apply Theorem 4.5.8. However, it passes the test of Theorem 4.5.7. So  $\mathbb{E}(W)$  is reasonable, though we have no way of knowing whether it is correct.

For  $\mathbb{E}(A)$ , Theorem 4.5.8 doesn't apply, since the values of  $A$  do not lie in a finite interval. However, it passes the test of Theorem 4.5.7. So  $\mathbb{E}(W)$  is reasonable. (Indeed, if you go through the calculation, it is correct.)

Example 4.5.10

Example 4.5.11 (Conspiracy Theories)

The paper *On the Viability of Conspiratorial Beliefs*<sup>10</sup> investigates a probabilistic model<sup>11</sup> for the length of time a conspiracy theory can remain secret. In particular, the author uses the formula

$$L(t) = 1 - e^{-t(1-(1-p)^{N(t)})}$$

where  $L(t)$  is the probability that, after  $t$  years, a leak has occurred that would cause the conspiracy to be exposed;  $N(t)$  is the number of people involved in the conspiracy at time  $t$ ; and  $p$  is the probability that a person involved will cause a leak in any particular year. (It is implied that  $L(t) = 0$  for negative values of  $t$ .)

For this example, we'll only use a very basic version of the full model. Suppose there are 100 (immortal) people involved in a conspiracy, no new people are ever brought into the conspiracy, and each person has a 1% chance of causing a leak in one year.

- Using the model above, what is the expected amount of time it will take for a leak to occur?
- Using the model above, what is the probability that the conspiracy can survive without a leak for at least 5 years?

*Solution.*

- $L(t)$  is the probability that, at time  $t$ , at least one leak has occurred. Let  $T$  be the time that the first leak occurs. Then  $L(t) = \Pr(T \leq t)$ . So, the function  $L(t)$  is the *cumulative distribution function* of  $T$ . In order to find  $\mathbb{E}(T)$ , we'll need the probability density function of  $T$ , which will be  $L'(t)$  (by Definition 4.4.3).

Let's start by filling in our constants:  $N(t) = 100$  and  $p = \frac{1}{100}$ .

$$L(t) = 1 - e^{-t(1-(1-p)^{N(t)})} = 1 - e^{-t(1-0.99^{100})} = 1 - e^{t(0.99^{100}-1)}$$

10 Grimes DR (2016) On the Viability of Conspiratorial Beliefs. PLoS ONE 11(1): e0147905. <https://doi.org/10.1371/journal.pone.0147905>

11 The assumptions made that lead to this model are that every member of the conspiracy is equally likely to cause a leak (whether by negligence or on purpose); that leak events are independent of one another; and that the probability of a conspirator causing a leak in any given year is constant. The full derivation is beyond the scope of the text, but the interested reader may look up "Poisson distribution."

The paper goes on to approximate  $p$  using conspiracy theories that have been exposed. They also use demographic data to approximate  $N(t)$ . They apply the model to famous conspiracy theories (e.g. the moon landing being faked) to discuss whether such a plot could realistically remain secret until present day.

Note that  $(0.99^{100} - 1)$  is a constant. In order to make the work below clearer, we'll replace it with  $c$ .

$$L(t) = 1 - e^{ct} \text{ where } c = (0.99^{100} - 1)$$

We find the PDF of  $T$  by differentiating  $L(t)$ .

$$L'(t) = -ce^{ct}$$

Now we use the definition of expected value, Definition 4.5.1

$$\mathbb{E}(T) = \int_{-\infty}^{\infty} t \cdot L'(t) dt = \int_0^{\infty} t \cdot (-ce^{tc}) dt = \lim_{b \rightarrow \infty} \int_0^b t \cdot (-ce^{ct}) dt$$

We use integration by parts with  $u = t$ ,  $dv = -ce^{ct} dt$ ;  $du = dt$ ,  $v = -e^{ct}$

$$\begin{aligned} &= \lim_{b \rightarrow \infty} \left[ [-te^{ct}]_0^b - \int_0^b -e^{ct} dt \right] \\ &= \lim_{b \rightarrow \infty} \left[ -be^{bc} + \left[ \frac{1}{c} e^{ct} \right]_0^b \right] \\ &= \lim_{b \rightarrow \infty} \left[ -be^{bc} + \frac{1}{c} e^{bc} - \frac{1}{c} \right] \\ &= \lim_{b \rightarrow \infty} \left[ \left( \frac{1}{c} - b \right) e^{bc} \right] - \frac{1}{c} \end{aligned}$$

Since  $c < 0$ ,  $\lim_{b \rightarrow \infty} e^{bc} = 0$  (\*). So,  $\left(\frac{1}{c} - b\right) e^{bc}$  has the indeterminate form  $-\infty \cdot 0$ . We will re-write this in order to use l'Hôpital's rule.

$$\begin{aligned} &= \lim_{b \rightarrow \infty} \left[ \frac{\frac{1}{c} - b}{e^{-bc}} \right] - \frac{1}{c} \\ &= \lim_{b \rightarrow \infty} \left[ \frac{\frac{d}{db} \left[ \frac{1}{c} - b \right]}{\frac{d}{db} [e^{-bc}]} \right] - \frac{1}{c} \\ &= \lim_{b \rightarrow \infty} \left[ \frac{-1}{-ce^{-bc}} \right] - \frac{1}{c} \\ &= \lim_{b \rightarrow \infty} \left[ \frac{1}{c} e^{bc} \right] - \frac{1}{c} \\ &= 0 - \frac{1}{c} \quad \text{using (*)} \\ &= -\frac{1}{.99^{100} - 1} = \frac{1}{1 - .99^{100}} \approx 1.58 \end{aligned}$$

So, the expected value of the time it would take for this conspiracy theory to be leaked is about 19 months.

(b) The probability that no leak has occurred at time  $t = 5$  is  $1 - L(5)$ :

$$1 - L(5) = e^{5(.99^{100} - 1)} \approx 0.04$$

So, there's about a 4% chance that the conspiracy would survive at least 5 years without any leaks.

Example 4.5.11

## 4.6▲ Variance and Standard Deviation

### 4.6.1 ► Motivation: Average difference from the average

In Example 4.5.2, we found that if we chose one of our 100 parents at random, the expected number of nightly awakenings was 2.85. If we choose a parent at random in this way, **how can we determine whether that parent had a “usual” or “unusual” experience?** Let's get our head around this problem with some preliminary observations.

- The expected value is not an integer. So no matter who we choose, we are guaranteed to *not* choose a parent with the expected number of awakenings. So, a “usual” experience is not the same as actually achieving the expected value.
- If we choose a parent with 3 awakenings, that's as close as we can get to the expectation. It seems reasonable that when  $X \approx \mathbb{E}(X)$ , that's a fairly “usual” trial.
- Parents with two awakenings are the most numerous. So although these parents are farther from average, we are more likely to choose one of them than we are to choose any other. So it is not enough to look for value of  $X$  that are *closest* to  $\mathbb{E}(X)$ .
- Suppose we choose a parent with 4 awakenings. Is this so far above average that is is very unusual (and so possibly a cause for concern) or is it still within a reasonably common range? This question will bring us to the heart of the matter: **how far from  $\mathbb{E}(X)$  is still “usual”?**

To quantify the last bullet point, let's compare each parent's experience to the expected value. If your baby woke you up twice during the night, then your experience differs from the average by 0.85. If your baby woke you up three times during the night, then your experience differs from the average by 0.15. Let's give that difference its own variable name,  $Y$ . Larger values of  $Y$  mean a larger difference between the individual experience and the expectation. So parents with a high  $Y$  value are “less usual” than parents with a low  $Y$ -value.

$x$	proportion of parents woken up $x$ times	$Y =  x - 2.85 $
0	$\frac{5}{100}$	2.85
1	$\frac{5}{100}$	1.85
2	$\frac{40}{100}$	0.85
3	$\frac{23}{100}$	0.15
4	$\frac{13}{100}$	1.15
5	$\frac{10}{100}$	2.15
6	$\frac{0}{100}$	3.15
7	$\frac{3}{100}$	4.15
8	$\frac{1}{100}$	5.15

The expectation of  $Y$  is about 2.38:

$$\begin{aligned}
 \mathbb{E}(Y) &= \sum_{x=0}^8 \underbrace{|x - 2.85|}_Y \cdot Pr(Y = y) \\
 &= 2.85 \cdot \frac{5}{100} + 1.85 \cdot \frac{5}{100} + 0.85 \cdot \frac{40}{100} + 0.15 \cdot \frac{23}{100} + 1.15 \cdot \frac{13}{100} \\
 &\quad + 2.15 \cdot \frac{10}{100} + 3.15 \cdot \frac{0}{100} + 4.15 \cdot \frac{3}{100} + 5.15 \cdot \frac{1}{100} \\
 &\approx 1.15
 \end{aligned}$$

That is, when we choose parents at random, on average their number of awakenings differs from the expected number of awakenings by 1.15.

With that in mind, we might say a parent who wakes up between  $1.15 - 2.38 = -1.23$  and  $1.15 + 2.38 = 3.53$  times wakes up a “usual” number of times, which the other parents have experiences that are “unusual”. A parent whose baby wakes then up four times during the night is “unusual,” in that their experience is quite different from the expectation, but a parent whose baby never wakes them up is still in the range of “usual”.

To generalize what we just computed:

- $X$  is a random variable
- $Y = |X - \mathbb{E}(X)|$  tells us how different  $X$  is from its expectation
- So,  $\mathbb{E}(Y) = \mathbb{E}(|X - \mathbb{E}(X)|)$  is the expected difference from the expectation. We used this as a measure of how far off from  $\mathbb{E}(X)$  our variable  $X$  could be and still be considered “usual”.

## 4.6.2 ► Definitions and Computations

**Definition 4.6.1.**

The **variance** of a random variable  $X$ , denoted  $\text{Var}(X)$ , is:

$$\mathbb{E} \left[ \left( X - \mathbb{E}(X) \right)^2 \right].$$

The **standard deviation** of  $X$ , written  $\sigma(X)$ , is the square root of the variance:

$$\sigma(X) = \sqrt{\text{Var}(X)}$$

Recall from Definition 4.5.1 that the definition of  $\mathbb{E}(X)$  depends on whether  $X$  is continuous or discrete.

**Corollary 4.6.2.**

If  $X$  is a discrete random variable, then

$$\text{Var}(X) = \sum_x (x - \mathbb{E}(X))^2 \cdot \text{Pr}(X = x)$$

where the sum is taken over every possible value of  $X$ .

If  $X$  is a continuous random variable, then

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 \cdot f(x) dx$$

where  $f(x)$  is the probability density function (PDF) of  $X$ .

Note the similarities between  $\text{Var}(X)$  and  $\mathbb{E}(Y)$  from the end of the last subsection, 4.6.1. Their interpretations are similar:  $\text{Var}(X)$  measures the expected *squared* difference between  $X$  and  $\mathbb{E}(X)$ <sup>12</sup>.

One reason we replace  $|X - \mathbb{E}(X)|$  with  $(X - \mathbb{E}(X))^2$  is that  $f(X) = |x - \mathbb{E}(X)|$  is not differentiable, while  $f(x) = (x - \mathbb{E}(X))^2$  is differentiable. We want to be able to use calculus tools, so differentiability is desirable.

**Example 4.6.3**

Consider the random variable  $X$  with probability mass function (PMF) given below.

$x$	$\text{Pr}(X = x)$
0	$\frac{1}{2}$
10	$\frac{1}{2}$

<sup>12</sup> To explore why we need absolute values or squares, see Question 14 in Section 4.6 of the practice book.

$X$  takes on values from  $[0, 10]$ , with  $\mathbb{E}(X) = 5$ . Every value of  $X$  differs from  $\mathbb{E}(X)$  by 5. However,

$$\text{Var}(X) = \sum_x (x - 5)^2 \cdot \text{Pr}(X = x) = (0 - 5)^2 \cdot \frac{1}{2} + (10 - 5)^2 \cdot \frac{1}{2} = 25$$

A drawback to replacing  $|X - \mathbb{E}(X)|$  with  $(X - \mathbb{E}(X))^2$  is that the variance may no longer be in a meaningful range. In this case, 25 is not in the range of numbers we're considering, so it's hard to interpret this as a "usual difference" between  $X$  and  $\mathbb{E}(X)$ . That's why we define the standard deviation:

$$\sigma(G) = \sqrt{25} = 5$$

We take the square root of  $\text{Var}(X)$  to somehow atone for our previous transgression of squaring  $|X - \mathbb{E}(X)|$ . Informally, we think of the standard deviation as the "usual" difference between  $X$  and  $\mathbb{E}(X)$ .

Example 4.6.3

Example 4.6.4

One thousand students take a midterm, and we choose one student uniformly at random.  $X$  is the mark the student got on the midterm, out of 100. For this particular group of 1000 students,  $\mathbb{E}(X) = 65$  and  $\sigma(X) = 15$ .

- Suppose we select Student A, who earned 60 points. Although this is below the class average, it is *within one standard deviation of the expectation*. That is,

$$|X - \mathbb{E}(X)| = 5 < 15 = \sigma(X).$$

So this student is not below average in a really significant way.

- If we select Student B who scored 90, not only are they above the class average, they are *well above* the class average. The *difference* between  $X$  and  $\mathbb{E}(X)$  is greater than usual.
- If we select Student C who scored 45, not only are they below the class average, they are *well below* the class average. The *difference* between  $X$  and  $\mathbb{E}(X)$  is worse than usual.
- In general, we think of students with grades from 50 to 80 as having a "usual" score. Those numbers come from  $\mathbb{E}(X) - \sigma(X) = 50$  and  $\mathbb{E}(X) + \sigma(X) = 80$ .

Example 4.6.4

The variance of a random variable is often calculated in the manner below:

**Corollary 4.6.5.**

$$\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

*Proof.* In the continuous case, from Corollary 4.6.2:

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f(x) dx \\ &= \int_{-\infty}^{\infty} (x^2 - 2x \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2) f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2 \cdot \mathbb{E}(X) \int_{-\infty}^{\infty} x f(x) dx + [\mathbb{E}(X)]^2 \int_{-\infty}^{\infty} f(x) dx \end{aligned}$$

By the definition of  $\mathbb{E}(X)$ , Definition 4.5.1:

$$= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2 \int_{-\infty}^{\infty} f(x) dx$$

By property 3 of Corollary 4.4.10,

$$\begin{aligned} &= \mathbb{E}(X^2) - 2 \cdot [\mathbb{E}(X)]^2 + [\mathbb{E}(X)]^2 \cdot 1 \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \end{aligned}$$

The discrete case progresses similarly. From Corollary 4.6.2:

$$\begin{aligned} \text{Var}(X) &= \sum_x (x - \mathbb{E}(x))^2 \cdot \text{Pr}(X = x) \\ &= \sum_x (x^2 - 2 \cdot \mathbb{E}(x) + [\mathbb{E}(X)]^2) \cdot \text{Pr}(X = x) \\ &= \sum_x x^2 \cdot \text{Pr}(X = x) - 2 \cdot \mathbb{E}(X) \cdot \sum_x x \text{Pr}(X = x) + [\mathbb{E}(X)]^2 \sum_x \text{Pr}(X = x) \end{aligned}$$

By the definition of  $\mathbb{E}(X)$ , Definition 4.5.1:

$$= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2 \sum_x \text{Pr}(X = x)$$

By the definition of a PDF, Definition 4.2.1,

$$\begin{aligned} &= \mathbb{E}(X^2) - 2 \cdot \mathbb{E}(X) \cdot \mathbb{E}(X) + [\mathbb{E}(X)]^2 \cdot 1 \\ &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \end{aligned}$$

□

**Example 4.6.6**

Suppose  $X$  is a continuous random variable with probability density function (PDF)

$$f(x) = \begin{cases} \frac{x}{50} & \text{if } 0 \leq x \leq 10 \\ 0 & \text{else} \end{cases}$$

We will calculate  $\text{Var}(X)$  two ways.



- To calculate  $\text{Var}(X)$ , we first need to know  $\mathbb{E}(X)$ .

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_0^{10} x \cdot \frac{x}{50} dx \\ &= \int_0^{10} \frac{x^2}{50} dx = \frac{x^3}{150} \Big|_0^{10} = \frac{10^3}{150} = \frac{20}{3}\end{aligned}$$

- Using Corollary 4.6.2,

$$\begin{aligned}\text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(x))^2 \cdot f(x) dx = \int_0^{10} \left(x - \frac{20}{3}\right)^2 \cdot \frac{x}{50} dx \\ &= \int_0^{10} \left(x^2 - \frac{40}{3}x + \frac{400}{9}\right) \cdot \frac{x}{50} dx = \frac{1}{50} \int_0^{10} \left(x^3 - \frac{40}{3}x^2 + \frac{400}{9}x\right) dx \\ &= \frac{1}{50} \left(\frac{x^4}{4} - \frac{40x^3}{9} + \frac{200x^2}{9}\right) \Big|_0^{10} = \frac{1}{50} \left(\frac{10^4}{4} - \frac{40 \cdot 10^3}{9} + \frac{200 \cdot 10^2}{9}\right) \\ &= \frac{10^4}{50} \left(\frac{1}{4} - \frac{4}{9} + \frac{2}{9}\right) = \frac{50}{9}\end{aligned}$$

- Using Corollary 4.6.5,

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\ &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx - \left(\frac{20}{3}\right)^2 = \int_0^{10} x^2 \cdot \frac{x}{50} dx - \frac{400}{9} \\ &= \frac{x^4}{50 \cdot 4} \Big|_0^{10} - \frac{400}{9} = \frac{10^4}{200} - \frac{400}{9} = \frac{50}{9}\end{aligned}$$

Example 4.6.6

Example 4.6.7

Calculate the variance (two ways) and standard deviation of a dice roll.

*Solution.* (Since we'll be evaluating sums, Theorem 3.1.6 comes in handy.)

Let  $X$  be the random variable that takes on the number rolled. By Definition 4.5.1,

$$\mathbb{E}(X) = \sum_{x=1}^6 x \cdot \text{Pr}(X = x) = \frac{1}{6} \sum_{x=1}^6 x = \frac{1}{6} \left(\frac{6 \cdot 7}{2}\right) = \frac{7}{2}$$

Using Corollary 4.6.2,

$$\begin{aligned}\text{Var}(X) &= \sum_x (x - \mathbb{E}(X))^2 \cdot \Pr(X = x) = \sum_{x=1}^6 \left(x - \frac{7}{2}\right)^2 \cdot \frac{1}{6} \\ &= \sum_{x=1}^6 \frac{1}{6} \left(x^2 - 7x + \frac{49}{4}\right) = \frac{1}{6} \sum_{x=1}^6 x^2 - \frac{7}{6} \sum_{x=1}^6 x + \frac{1}{6} \sum_{x=1}^6 \frac{49}{4} \\ &= \frac{1}{6} \left(\frac{6 \cdot 7 \cdot 13}{6}\right) - \frac{7}{6} \left(\frac{7 \cdot 6}{2}\right) + \frac{49}{4} \\ &= \frac{35}{12}\end{aligned}$$

Using Corollary 4.6.5 to calculate a second way,

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \sum_{x=1}^6 x^2 \cdot \Pr(X = x) - \left[\frac{7}{2}\right]^2 \\ &= \sum_{x=1}^6 \frac{1}{6} x^2 - \frac{49}{4} = \frac{1}{6} \left(\frac{6 \cdot 7 \cdot 13}{6}\right) - \frac{49}{4} = \frac{35}{12}\end{aligned}$$

(Computing the variance two different ways is not usually necessary, but it can be a good way to double-check your work.)

Using Definition 4.6.1,  $\sigma(X) = \sqrt{\text{Var}(X)} = \sqrt{\frac{35}{12}} \approx 1.7$

Example 4.6.7

Example 4.6.8

A continuous random variable  $W$  has cumulative distribution function (CDF)

$$F(x) = \begin{cases} 0 & x < 0 \\ e^x - 1 & 0 \leq x \leq \ln 2 \\ 1 & x > \ln 2 \end{cases}$$

Calculate the variance and standard deviation of  $W$ . For practice, use both methods discussed in this section for computing variance.

*Solution.*

We use the variance to calculate the standard deviation; we use expected value to calculate variance; we use probability density function (PDF) to calculate expected value; and we use cumulative distribution function (CDF) to define probability density function (PDF). Working backwards, this gives us a plan for performing the necessary calculations.

$$F(x) \xrightarrow{\text{Step 1}} f(x) \xrightarrow{\text{Step 2}} \mathbb{E}(W) \xrightarrow{\text{Step 3}} \text{Var}(W) \xrightarrow{\text{Step 4}} \sigma(W)$$

**Step 1** Definition 4.4.3 tells us the probability density function (PDF) is the derivative of the cumulative distribution function (CDF).

$$F(x) = \begin{cases} 0 & x < 0 \\ e^x - 1 & 0 \leq x \leq \ln 2 \\ 1 & x > \ln 2 \end{cases}$$

$$f(x) = \begin{cases} 0 & x < 0 \\ e^x & 0 < x < \ln 2 \\ 0 & x > \ln 2 \end{cases}$$

$$= \begin{cases} e^x & 0 < x < \ln 2 \\ 0 & \text{else} \end{cases}$$

**Step 2** Using Definition 4.5.1

$$\begin{aligned} \mathbb{E}(W) &= \int_{-\infty}^{\infty} x \cdot f(x) dx \\ &= \int_0^{\ln 2} x \cdot e^x dx \end{aligned}$$

We use integration by parts with  $u = x$ ,  $dv = e^x dx$ ;  $du = dx$ ,  $v = e^x$

$$= \left[ x e^x \right]_0^{\ln 2} - \int_0^{\ln 2} e^x dx = 2 \ln 2 - [2 - 1] = 2 \ln 2 - 1 \approx 0.39$$

We can do a quick reliability check using Theorems 4.5.7 and 4.5.8. Our variable  $W$  spends its entire life between 0 and  $\ln 2 \approx 0.69$ , so we expect  $\mathbb{E}(W)$  to be in that same interval, which is true. Since  $f(x)$  is increasing on the relevant interval,  $W$  spends more time near the larger numbers, so we also expect  $\mathbb{E}(W) > \frac{\ln 2}{2} \approx 0.35$ . This accords with our calculation.

**Step 3** We'll be using the constant  $\mathbb{E}(W) = 2 \ln 2 - 1$  a lot in the calculations below, so we'll use logarithm rules to write it more compactly:

$$2 \ln 2 - 1 = \ln(2^2) - \ln e = \ln 4 - \ln e = \ln(4/e)$$

The benefit of this equivalent expression is that when we square it, there are no binomials to expand. (It is of course perfectly possible to do the computations with  $2 \ln 2 - 1$ .)

Using Corollary 4.6.2,

$$\begin{aligned} \text{Var}(W) &= \int_{-\infty}^{\infty} (x - \mathbb{E}(W))^2 \cdot f(x) dx = \int_0^{\ln 2} (x - \ln(4/e))^2 \cdot e^x dx \\ &= \int_0^{\ln 2} \left( x^2 - 2x \ln(4/e) + \ln^2(4/e) \right) e^x dx \\ &= \int_0^{\ln 2} x^2 e^x dx - 2 \ln(4/e) \int_0^{\ln 2} x \cdot e^x dx + \ln^2(4/e) \int_0^{\ln 2} e^x dx \\ &= \int_0^{\ln 2} x^2 e^x dx - 2 \ln(4/e) \mathbb{E}(W) + \ln^2(4/e) (2 - 1) \end{aligned}$$

We'll use integration by parts on the remaining integral:  $u = x^2$ ,  $dv = e^x dx$ ;  $du = 2x dx$ ,  $v = e^x$

$$\begin{aligned} &= \left[ x^2 e^x \right]_0^{\ln 2} - \int_0^{\ln 2} 2x \cdot e^x dx - 2 \ln^2(4/e) + \ln^2(4/e) \\ &= 2 \ln^2 2 - 2 \mathbb{E}(W) - \ln^2(4/e) \\ &= 2 \ln^2 2 - 2 \ln(4/e) - \ln^2(4/e) \end{aligned}$$

To simplify, we'll revert the arguments of our logarithms to  $2 \ln 2 - 1$ , rather than  $\ln(4/e)$ .

$$\begin{aligned} &= 2 \ln^2 2 - 2(2 \ln 2 - 1) - (2 \ln 2 - 1)^2 \\ &= 2 \ln^2 2 - 4 \ln 2 + 2 - (4 \ln^2 2 - 4 \ln 2 + 1) \\ &= \boxed{1 - 2 \ln^2 2} \approx 0.039 \end{aligned}$$

Using Corollary 4.6.5,

$$\text{Var}(W) = \mathbb{E}(W^2) - [\mathbb{E}(W)]^2 = \int_0^{\ln 2} x^2 e^x dx - (2 \ln 2 - 1)^2$$

The integral was already computed in the work above.

$$\begin{aligned} &= (2 \ln^2 2 - 4 \ln 2 + 2) - (2 \ln 2 - 1)^2 \\ &= \boxed{1 - 2 \ln^2 2} \approx 0.039 \end{aligned}$$

Step 4 By Definition 4.6.1,

$$\sigma(W) = \sqrt{\text{Var}(W)} = \sqrt{1 - 2 \ln^2 2} \approx 0.198$$

Example 4.6.8

### 4.6.3 ▶ Checking your Standard Deviation Calculation

#### Corollary 4.6.9.

Let  $a, b$  be real numbers with  $a < b$  and suppose a random variable  $X$  takes values from the interval  $[a, b]$ . Then

$$0 \leq \sigma(X) \leq \frac{b-a}{2}$$

*Proof.* First, consider what happens when we replace  $\mathbb{E}(X)$  with  $\frac{b+a}{2}$  (the midpoint of the sample space) in the definition of variance (Definition 4.6.1).

$$\begin{aligned}
 \mathbb{E} \left( \left( X - \frac{b+a}{2} \right)^2 \right) &= \int_a^b \left( x - \frac{b+a}{2} \right)^2 \cdot f(x) dx \\
 &= \int_a^b \left( x^2 - (b+a)x + \left( \frac{b+a}{2} \right)^2 \right) \cdot f(x) dx \\
 &= \int_a^b x^2 f(x) dx - (b+a) \int_a^b x f(x) dx + \left( \frac{b+a}{2} \right)^2 \int_a^b f(x) dx \\
 &= \mathbb{E}(X^2) - (b+a)\mathbb{E}(X) + \left( \frac{b+a}{2} \right)^2 \\
 &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 + [\mathbb{E}(X)]^2 - (b+a)\mathbb{E}(X) + \left( \frac{b+a}{2} \right)^2 \\
 &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 + \left( \mathbb{E}(X) - \frac{b+a}{2} \right)^2 \\
 &\geq \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \text{Var}(X) \tag{*}
 \end{aligned}$$

Since  $X$  takes values in the interval  $[a, b]$ :

$$\begin{aligned}
 & a \leq X \leq b \\
 \implies & a - \frac{b+a}{2} \leq X - \frac{b+a}{2} \leq b - \frac{b+a}{2} \\
 \implies & -\frac{b-a}{2} \leq X - \frac{b+a}{2} \leq \frac{b-a}{2} \\
 \implies & 0 \leq \left( X - \frac{b+a}{2} \right)^2 \leq \left( \frac{b-a}{2} \right)^2
 \end{aligned}$$

By Theorem 4.5.7,

$$0 \leq \mathbb{E} \left( \left( X - \frac{b+a}{2} \right)^2 \right) \leq \left( \frac{b-a}{2} \right)^2$$

So, with our previous result (\*),

$$\text{Var}(X) \leq \mathbb{E} \left( \left( X - \frac{b+a}{2} \right)^2 \right) \leq \left( \frac{b-a}{2} \right)^2$$

So, 
$$\sigma(X) \leq \frac{b-a}{2}$$

□

**Example 4.6.10**

If the random variable  $X$  takes on values from the interval  $[1, 5]$ , then  $0 \leq \sigma(X) \leq 2$ . Since  $\sigma(X) = \sqrt{\text{Var}(X)}$ , then  $0 \leq \text{Var}(X) \leq 4$ .

## Example 4.6.10

Chapter 4 contains content adapted by Bruno Belevan, Parham Hamidi, and Elyse Yeager from Sections 1.1, 3.1, Ch 4 introduction, 4.1, and 4.2 of *Introductory Statistics* by Ilowsky and Dean under a [Creative Commons Attribution License v4.0](#).

# SEQUENCES AND SERIES

You have probably learned about Taylor polynomials<sup>1</sup> and, in particular, that

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + E_n(x)$$

where  $E_n(x)$  is the error introduced when you approximate  $e^x$  by its Taylor polynomial of degree  $n$ . You may have even seen a formula for  $E_n(x)$ . We are now going to ask what happens as  $n$  goes to infinity? Does the error go zero, giving an exact formula for  $e^x$ ? We shall later see that it does and that

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

At this point we haven't defined, or developed any understanding of, this infinite sum. How do we compute the sum of an infinite number of terms? Indeed, when does a sum of an infinite number of terms even make sense? Clearly we need to build up foundations to deal with these ideas. Along the way we shall also see other functions for which the corresponding error obeys  $\lim_{n \rightarrow \infty} E_n(x) = 0$  for some values of  $x$  and not for other values of  $x$ .

To motivate the next section, consider using the above formula with  $x = 1$  to compute the number  $e$ :

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}$$

As we stated above, we don't yet understand what to make of this infinite number of terms, but we might try to sneak up on it by thinking about what happens as we take

<sup>1</sup> Now would be an excellent time to quickly read over your notes on the topic.

more and more terms.

1 term	$1 = 1$
2 terms	$1 + 1 = 2$
3 terms	$1 + 1 + \frac{1}{2} = 2.5$
4 terms	$1 + 1 + \frac{1}{2} + \frac{1}{6} = 2.666666\dots$
5 terms	$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} = 2.708333\dots$
6 terms	$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} = 2.716666\dots$

By looking at the infinite sum in this way, we naturally obtain a sequence of numbers

$$\{ 1, 2, 2.5, 2.666666, \dots, 2.708333, \dots, 2.716666, \dots, \dots \}.$$

The key to understanding the original infinite sum is to understand the behaviour of this sequence of numbers — in particular, what do the numbers do as we go further and further? Does it settle down<sup>2</sup> to a given limit?

## 5.1▲ Sequences

In the discussion above we used the term “sequence” without giving it a precise mathematical meaning. Let us rectify this now.

**Definition 5.1.1.**

A sequence is a list of infinitely<sup>3</sup> many numbers with a specified order. It is denoted

$$\{ a_1, a_2, a_3, \dots, a_n, \dots \} \quad \text{or} \quad \{ a_n \} \quad \text{or} \quad \{ a_n \}_{n=1}^{\infty}$$

We will often specify a sequence by writing it more explicitly, like

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

where  $f(n)$  is some function from the natural numbers to the real numbers.

2 You will notice a great deal of similarity between the results of the next section and “limits at infinity” which was covered last term.

3 For the more pedantic reader, here we mean a list of countably infinitely many numbers. The interested (pedantic or otherwise) reader should look up countable and uncountable sets.



Example 5.1.2

Here are three sequences.

$$\begin{aligned} \left\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, \dots\right\} & \quad \text{or} \quad \left\{a_n = \frac{1}{n}\right\}_{n=1}^{\infty} \\ \left\{1, 2, 3, \dots, n, \dots\right\} & \quad \text{or} \quad \left\{a_n = n\right\}_{n=1}^{\infty} \\ \left\{1, -1, 1, -1, \dots, (-1)^{n-1}, \dots\right\} & \quad \text{or} \quad \left\{a_n = (-1)^{n-1}\right\}_{n=1}^{\infty} \end{aligned}$$

It is not necessary that there be a simple explicit formula for the  $n^{\text{th}}$  term of a sequence. For example the decimal digits of  $\pi$  is a perfectly good sequence

$$\{3, 1, 4, 1, 5, 9, 2, 6, 5, 3, 5, 8, 9, 7, 9, 3, 2, 3, 8, 4, 6, 2, 6, 4, 3, 3, 8, \dots\}$$

but there is no simple formula<sup>4</sup> for the  $n^{\text{th}}$  digit.

Example 5.1.2

Our primary concern with sequences will be the behaviour of  $a_n$  as  $n$  tends to infinity and, in particular, whether or not  $a_n$  “settles down” to some value as  $n$  tends to infinity.

**Definition 5.1.3.**

A sequence  $\{a_n\}_{n=1}^{\infty}$  is said to converge to the limit  $A$  if  $a_n$  approaches  $A$  as  $n$  tends to infinity. If so, we write

$$\lim_{n \rightarrow \infty} a_n = A \quad \text{or} \quad a_n \rightarrow A \text{ as } n \rightarrow \infty$$

A sequence is said to converge if it converges to some limit. Otherwise it is said to diverge.

The reader should immediately recognise the similarity with limits at infinity

$$\lim_{x \rightarrow \infty} f(x) = L \quad \text{if} \quad f(x) \rightarrow L \text{ as } x \rightarrow \infty$$

Example 5.1.4

Three of the four sequences in Example 5.1.2 diverge:

- The sequence  $\{a_n = n\}_{n=1}^{\infty}$  diverges because  $a_n$  grows without bound, rather than approaching some finite value, as  $n$  tends to infinity.

4 There is, however, a remarkable result due to Bailey, Borwein and Plouffe that can be used to compute the  $n^{\text{th}}$  binary digit of  $\pi$  (i.e. writing  $\pi$  in base 2 rather than base 10) without having to work out the preceding digits.

- The sequence  $\{a_n = (-1)^{n-1}\}_{n=1}^{\infty}$  diverges because  $a_n$  oscillates between  $+1$  and  $-1$  rather than approaching a single value as  $n$  tends to infinity.
- The sequence of the decimal digits of  $\pi$  also diverges, though the proof that this is the case is a bit beyond us right now<sup>5</sup>.

The other sequence in Example 5.1.2 has  $a_n = \frac{1}{n}$ . As  $n$  tends to infinity,  $\frac{1}{n}$  tends to zero. So

$$\lim_{n \rightarrow \infty} \frac{1}{n} = 0$$

Example 5.1.4

Example 5.1.5  $\left(\lim_{n \rightarrow \infty} \frac{n}{2n+1}\right)$

Here is a little less trivial example. To study the behaviour of  $\frac{n}{2n+1}$  as  $n \rightarrow \infty$ , it is a good idea to write it as

$$\frac{n}{2n+1} = \frac{1}{2 + \frac{1}{n}}$$

As  $n \rightarrow \infty$ , the  $\frac{1}{n}$  in the denominator tends to zero, so that the denominator  $2 + \frac{1}{n}$  tends to 2 and  $\frac{1}{2 + \frac{1}{n}}$  tends to  $\frac{1}{2}$ . So

$$\lim_{n \rightarrow \infty} \frac{n}{2n+1} = \lim_{n \rightarrow \infty} \frac{1}{2 + \frac{1}{n}} = \frac{1}{2}$$

Example 5.1.5

Notice that in this last example, we are really using techniques that we used before to study infinite limits like  $\lim_{x \rightarrow \infty} f(x)$ . This experience can be easily transferred to dealing with  $\lim_{n \rightarrow \infty} a_n$  limits by using the following result.

**Theorem 5.1.6.**

If

$$\lim_{x \rightarrow \infty} f(x) = L$$

and if  $a_n = f(n)$  for all positive integers  $n$ , then

$$\lim_{n \rightarrow \infty} a_n = L$$

5 If the digits of  $\pi$  were to converge, then  $\pi$  would have to be a rational number. The irrationality of  $\pi$  (that it cannot be written as a fraction) was first proved by Lambert in 1761. Niven's 1947 proof is more accessible and we invite the interested reader to use their favourite search engine to find step-by-step guides to that proof.

Example 5.1.7  $\left(\lim_{n \rightarrow \infty} e^{-n}\right)$

Set  $f(x) = e^{-x}$ . Then  $e^{-n} = f(n)$  and

since  $\lim_{x \rightarrow \infty} e^{-x} = 0$  we know that  $\lim_{n \rightarrow \infty} e^{-n} = 0$

Example 5.1.7

The bulk of the rules for the arithmetic of limits of functions that you already know also apply to the limits of sequences. That is, the rules you learned to work with limits such as  $\lim_{x \rightarrow \infty} f(x)$  also apply to limits like  $\lim_{n \rightarrow \infty} a_n$ .

**Theorem 5.1.8** (Arithmetic of limits).

Let  $A, B$  and  $C$  be real numbers and let the two sequences  $\{a_n\}_{n=1}^{\infty}$  and  $\{b_n\}_{n=1}^{\infty}$  converge to  $A$  and  $B$  respectively. That is, assume that

$$\lim_{n \rightarrow \infty} a_n = A \qquad \lim_{n \rightarrow \infty} b_n = B$$

Then the following limits hold.

- (a)  $\lim_{n \rightarrow \infty} [a_n + b_n] = A + B$   
(The limit of the sum is the sum of the limits.)
- (b)  $\lim_{n \rightarrow \infty} [a_n - b_n] = A - B$   
(The limit of the difference is the difference of the limits.)
- (c)  $\lim_{n \rightarrow \infty} Ca_n = CA$ .
- (d)  $\lim_{n \rightarrow \infty} a_n b_n = AB$   
(The limit of the product is the product of the limits.)
- (e) If  $B \neq 0$  then  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \frac{A}{B}$   
(The limit of the quotient is the quotient of the limits *provided* the limit of the denominator is not zero.)

We use these rules to evaluate limits of more complicated sequences in terms of the limits of simpler sequences — just as we did for limits of functions.

Example 5.1.9

Combining Examples 5.1.5 and 5.1.7,

$$\begin{aligned} \lim_{n \rightarrow \infty} \left[ \frac{n}{2n+1} + 7e^{-n} \right] &= \lim_{n \rightarrow \infty} \frac{n}{2n+1} + \lim_{n \rightarrow \infty} 7e^{-n} && \text{by Theorem 5.1.8.a} \\ &= \lim_{n \rightarrow \infty} \frac{n}{2n+1} + 7 \lim_{n \rightarrow \infty} e^{-n} && \text{by Theorem 5.1.8.c} \\ &= \frac{1}{2} + 7 \cdot 0 && \text{by Examples 5.1.5 and 5.1.7} \\ &= \frac{1}{2} \end{aligned}$$

Example 5.1.9

There is also a Squeeze Theorem for sequences.

**Theorem 5.1.10 (Squeeze Theorem).**

If  $a_n \leq c_n \leq b_n$  for all natural numbers  $n$ , and if

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} b_n = L$$

then

$$\lim_{n \rightarrow \infty} c_n = L$$

Example 5.1.11

In this example we use the Squeeze Theorem to evaluate

$$\lim_{n \rightarrow \infty} \left[ 1 + \frac{\pi_n}{n} \right]$$

where  $\pi_n$  is the  $n^{\text{th}}$  decimal digit of  $\pi$ . That is,

$$\pi_1 = 3 \quad \pi_2 = 1 \quad \pi_3 = 4 \quad \pi_4 = 1 \quad \pi_5 = 5 \quad \pi_6 = 9 \quad \dots$$

We do not have a simple formula for  $\pi_n$ . But we do know that

$$0 \leq \pi_n \leq 9 \implies 0 \leq \frac{\pi_n}{n} \leq \frac{9}{n} \implies 1 \leq 1 + \frac{\pi_n}{n} \leq 1 + \frac{9}{n}$$

and we also know that

$$\lim_{n \rightarrow \infty} 1 = 1 \quad \lim_{n \rightarrow \infty} \left[ 1 + \frac{9}{n} \right] = 1$$

So the Squeeze Theorem with  $a_n = 1$ ,  $b_n = 1 + \frac{9}{n}$ , and  $c_n = 1 + \frac{\pi_n}{n}$  gives

$$\lim_{n \rightarrow \infty} \left[ 1 + \frac{\pi_n}{n} \right] = 1$$

Example 5.1.11

Finally, recall that we can compute the limit of the composition of two functions using continuity. In the same way, we have the following result:

**Theorem 5.1.12** (Continuous functions of limits).

If  $\lim_{n \rightarrow \infty} a_n = L$  and if the function  $g(x)$  is continuous at  $L$ , then

$$\lim_{n \rightarrow \infty} g(a_n) = g(L)$$

Example 5.1.13  $\left( \lim_{n \rightarrow \infty} \sin \frac{\pi n}{2n+1} \right)$

Write  $\sin \frac{\pi n}{2n+1} = g\left(\frac{n}{2n+1}\right)$  with  $g(x) = \sin(\pi x)$ . We saw, in Example 5.1.5 that

$$\lim_{n \rightarrow \infty} \frac{n}{2n+1} = \frac{1}{2}$$

Since  $g(x) = \sin(\pi x)$  is continuous at  $x = \frac{1}{2}$ , which is the limit of  $\frac{n}{2n+1}$ , we have

$$\lim_{n \rightarrow \infty} \sin \frac{\pi n}{2n+1} = \lim_{n \rightarrow \infty} g\left(\frac{n}{2n+1}\right) = g\left(\frac{1}{2}\right) = \sin \frac{\pi}{2} = 1$$

Example 5.1.13

### 5.1.1 ▶ Geometric and harmonic sequences in musical scales

Lists of numbers don't always get added together<sup>6</sup>, so sequences (that are not worked into series) can be interesting in their own right. We present here an application of sequences to music theory.

First, some musical preliminaries. Sound is caused by waves, and the frequency of a sound wave determines its pitch – how high or low it sounds. Higher frequencies lead to higher pitches, so the sound wave made by the chirp of a sparrow has a higher frequency than the sound wave made by the growl of a dog. We measure frequency in Hz (Hertz), which corresponds to periods per second. We often leave out the units, so you might see a frequency referred to as “100” instead of “100 Hz.” We'll picture the waves like the graph of a sine function, although this is not how they would actually appear.

We'll use the word “note” to mean a specified pitch. For example, the note named A4 usually corresponds to a frequency of 440 Hz. An interval is the “distance” between two notes, quantified as the ratio of their frequencies. The way we perceive the “distance”

<sup>6</sup> or they do but they shouldn't be, e.g. [this amusing sign](#)

between two notes relies on the ratio of their two frequencies, which is why we use a ratio and not a difference when measuring intervals.

**Example 5.1.14**

Consider the three pairs of notes below. Which pairs will sound roughly the same distance from each other, and which will sound different?

1. 110 Hz and 193.25 Hz
2. 440 Hz and 523.25 Hz
3. 587.33 Hz and 698.46 Hz

*Solution.* To quantify how far apart two notes sound, we take the ratio of their frequencies.

- 1 110 Hz and 193.25 Hz have a ratio of  $\frac{193.25}{110} \approx 1.75682$
- 2 440 Hz and 523.25 Hz have a ratio of  $\frac{523.25}{440} \approx 1.18920$
- 3 587.33 Hz and 698.46 Hz have a ratio of  $\frac{698.46}{587.33} \approx 1.18921$

The last two pairs of notes sound about the same distance away from one another, because their ratios are nearly identical. The first pair of notes will sound farther apart from one another than the other pairs.

Incidentally, the interval spanned in 2 and 3 has a name: a minor third. For listeners in the Western tradition, the sound of two notes of such an interval being played together is often evocative of a melancholy or enigmatic mood.

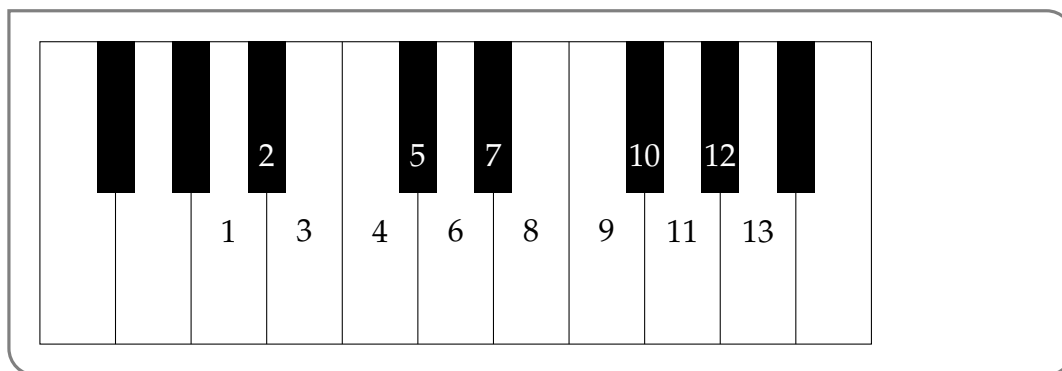
**Example 5.1.14**

A scale is a collection of notes. There are many different scales that are used, and many more that are theoretically possible. Scales in context usually refer to the collection of notes that make up most of a single piece of music. So, one song might mainly consist of notes from a scale named “B Minor,” and another song might mainly consist of notes from a scale named “G major pentatonic.” Generally speaking<sup>7</sup>, standardized scales consist of notes that people have decided they like hearing played together.

**Example 5.1.15**

The interval between some frequency  $a$  and the frequency  $2a$  is called an octave. Some popular musical scales divide the octave into twelve intervals. (In the partial piano schematic below, the key labelled 13 would produce a note with twice the frequency of the key labelled 1.)

<sup>7</sup> Precision in describing the things that people do is much harder to attain than precision in mathematics.



We call a scale “even-tempered” if consecutive notes always sound like they’re the same distance apart from one another. Since the sound of notes in relation to each other is determined by the ratio of their frequencies, this means that the ratio of the frequencies of two consecutive notes is the same, no matter which two consecutive notes we’re considering.

Suppose the key labelled 1 makes the note 440Hz, and the key labelled 13 makes the note 880 Hz (one octave above 440). If the piano is tuned to an even-tempered scale, what are the frequencies associated with the keys labelled 2 through 12?

*Solution.*

Let the notes on the piano form the first part<sup>8</sup> of a sequence, with key 1 making note  $a_1$ , key 2 making note  $a_2$ , and so on. We know three pieces of information:

1.  $a_1 = 440$
2.  $a_{13} = 880$
3.  $\frac{a_2}{a_1} = \frac{a_3}{a_2} = \frac{a_4}{a_3} = \dots = \frac{a_{13}}{a_{12}}$

(3 comes from the description of even-tempering.) Let’s give the number  $\frac{a_2}{a_1}$  the name  $r$  (because it’s a ratio). This gives us a recurrence relation to describe our partial sequence: since  $\frac{a_{n+1}}{a_n} = r$ , then  $a_{n+1} = ra_n$ . We can now write out each element of the partial sequence in terms of  $r$ .

$$\begin{aligned} a_1 &= 440 \\ a_2 &= 440r \\ a_3 &= (440r)r = 440r^2 \\ a_4 &= (440r^2)r = 440r^3 \\ &\vdots \\ a_{12} &= 440r^{11} \\ a_{13} &= 440r^{12} \end{aligned}$$

Since we’re given  $a_{13} = 880$ , we can solve for  $r$ .

$$\begin{aligned} 880 &= 440r^{12} \\ r &= 2^{\frac{1}{12}} \end{aligned}$$

8 we defined sequences to be infinite, but pianos have only finitely many keys

Now, we can write down each note frequency.

1. 440	5. $440 \cdot 2^{4/12} \approx 554.365$	9. $440 \cdot 2^{8/12} \approx 698.456$
2. $440 \cdot 2^{1/12} \approx 466.163$	6. $440 \cdot 2^{5/12} \approx 587.330$	10. $440 \cdot 2^{2/12} \approx 739.989$
3. $440 \cdot 2^{2/12} \approx 493.883$	7. $440 \cdot 2^{6/12} \approx 622.254$	11. $440 \cdot 2^{10/12} \approx 783.991$
4. $440 \cdot 2^{1/12} \approx 523.251$	8. $440 \cdot 2^{7/12} \approx 659.255$	12. $440 \cdot 2^{11/12} \approx 830.609$

Example 5.1.15

When we say “the interval between consecutive notes is the same,” we mean “the *ratio* between consecutive notes is the same.” Having a common ratio between consecutive terms is the defining characteristic of a geometric sequence.

**Definition 5.1.16.**

A geometric sequence is a sequence of the form

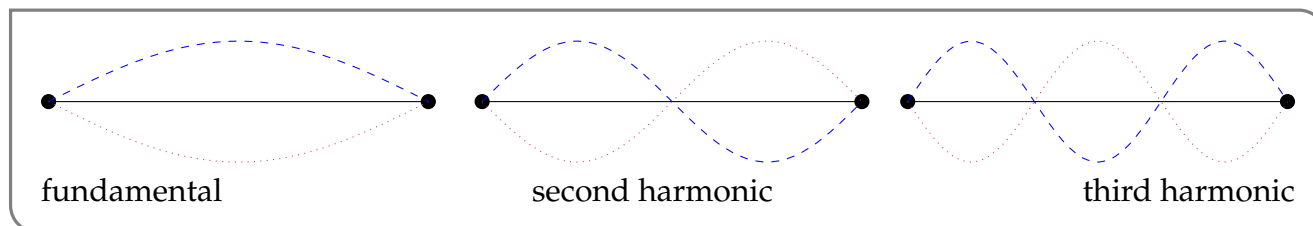
$$\{a, ar, ar^2, \dots, ar^n, \dots\} \quad \text{or} \quad \{a_n = ar^n\}_{n=1}^{\infty}$$

where  $a$  and  $r$  are any two fixed real numbers with  $a \neq 0$ .

Note  $\frac{a_{n+1}}{a_n} = r$  for every whole number  $n$ . We call  $r$  the *common ratio*.

(If we were to “add up” the terms of a geometric sequences, we’d get a geometric series – see Example 5.2.4.)

When a tone is made by a vibrating physical<sup>9</sup> object, although we may primarily pick up on one frequency (the “fundamental”), usually waves of many different frequencies are being generated. If we make a tone by causing a string to vibrate, as on a violin or guitar, the waves that make noise have frequencies that are whole-number multiples of the fundamental frequency. To explain this behaviour, note that the ends of the string are fixed, so they can’t move up and down. So, the only waves that can occur on the string are waves that keep these points fixed. The fundamental is the longest wave. The other waves that are generated are called harmonics. The  $n$ th harmonic has frequency  $n$  times the fundamental.



<sup>9</sup> For the following “physical” discussion, we’re relying on a very simplified model. However, the results are indeed relevant to how actual musical instruments sound.



In the figure above, a string<sup>10</sup> is fixed between two dots. We imagine it vibrating up and down in a wave pattern, moving between the positions shown by the dashed and dotted lines. The wavelength of these waves is inversely proportional to the frequency they generate – so dividing a wavelength by (say) three causes the frequency to triple.

Example 5.1.17

A string, when played, has a fundamental tone of 100 Hz, with a wavelength of 1 m. Let  $\{f_n\}$  be the sequence of frequencies of the harmonics of the string, organized by increasing pitch (with  $f_1 = 100$ ). Let  $\{\ell_n\}$  be the sequence of corresponding wavelengths (so  $\ell_1 = 1$ ). What are  $\{\ell_n\}$  and  $\{f_n\}$ ?

*Solution.* The frequencies of harmonic tones are integer multiples of the fundamental, so

$$f_1 = 100, \quad f_2 = 200, \quad f_3 = 300, \quad \dots, \quad f_n = 100n$$

The wavelengths are inversely proportional to the frequencies. So, if frequency  $f_n$  is  $f_1 \cdot n$ , then wavelength  $\ell_n$  is  $\frac{\ell_1}{n}$ .

$$\ell_1 = 1, \quad \ell_2 = \frac{1}{2}, \quad \ell_3 = \frac{1}{3}, \quad \dots, \quad \ell_n = \frac{1}{n}$$

Example 5.1.17

The sequence  $\{\frac{1}{n}\}_{n=1}^\infty$  is called the harmonic sequence. (We'll consider the harmonic series in Example 5.3.4.) In music textbooks, you might see the sequence of harmonic notes referred to as the "harmonic series." This isn't because the notes are added together, it's simply a different use of the word "series."

Example 5.1.18

Consider an even-tempered musical scale with twelve intervals in each octave, the lowest note of which is 250 Hz.

Suppose we have a string whose fundamental tone is 250 Hz. Which harmonics of the string are also notes of the even-tempered scale?

*Solution.*

The even-tempered musical scale is given by the geometric sequence  $\{e_n = a \cdot r^n\}_{n=0}^\infty$  where  $a = 250$  and  $r = 2^{\frac{1}{12}}$ . The harmonic sequence of the string is  $\{h_n = 250n\}_{n=1}^\infty$ .

All frequencies in the harmonic sequence are integer multiples of 250, and so are whole numbers. The only whole numbers in the geometric sequence  $e_n$  occur when 2 is raised to a whole-number powers, i.e. when  $n$  is a multiple of 12. So our only candidates for frequencies that appear in both sequences have the form  $250 \cdot 2^k$ . It's quick to see that these occur in both:  $250 \cdot 2^k = g_n$  when  $n = 12k$ , and  $250 \cdot 2^k = h_n$  when  $n = 2^k$ .

So, the only intervals from the even-tempered scale that perfectly line up with the natural harmonics of the string are octaves: the fundamental, twice the fundamental, twice that frequency, etc.

10 Similar wave behaviour occurs in tubes of air, like you might find in a brass instrument or woodwind. Brass players can emphasize different harmonic notes by changing they way they blow into their instrument.

Example 5.1.18

Harmonics are produced naturally, so it's nice if they're "in tune" with the scale notes. The dearth of overlap between harmonic and geometric sequences is one reason that even-tempered scales are sometimes unpopular. However, many harmonic notes are *approximated* by the even-tempered scale above. For example,  $2^{\frac{19}{12}} \approx 2.9966 \approx 3$ , so  $g_{19}$  is a fair approximation to  $e_3$ .

Example 5.1.19

Suppose we were to make a scale that consisted only of harmonics. The frequencies would make up the sequence  $\{h_n = an\}_{n=1}^{\infty}$ , where  $a$  is the fundamental.

How would such a scale sound if we played the notes one after the other? Remember, the way two notes sound depends on the ratio of their frequencies. A bigger ratio sounds like a bigger "step" from one note to the next. So, let's define a sequence  $\{r_n\}_{n=2}^{\infty}$  to be the ratio of the  $n$ th harmonic to the note before it. A value of  $r_n$  that is close to 1 means the two notes sound the same. A value of  $r_n$  that is far from 1 means the two notes sound different.

frequency:	$a$	$2a$	$3a$	$4a$	$5a$	$6a$	$7a$	$\dots$	$na$
ratio:		$2$	$\frac{3}{2}$	$\frac{4}{3}$	$\frac{5}{4}$	$\frac{6}{5}$	$\frac{7}{6}$	$\dots$	$\frac{n}{n-1}$

The sequences  $h_n$  and  $r_n$  have different limits, each with a musical interpretation.

- $\lim_{n \rightarrow \infty} h_n = \infty$  tells us that the notes of this sequence have no upper bound. We can find notes as high as we please in this scale.
- $\lim_{n \rightarrow \infty} r_n = 1$  tells us that notes of the scale sound more and more alike as we go higher.

The picture painted by these two limits is that the scale climbs higher and higher, but does so in tiny increments, so that many different high-pitched notes are virtually indistinguishable from one another. (On the other hand, the first step is huge: an entire octave!)

Example 5.1.19

With this introduction to sequences and some tools to determine their limits, we can now return to the problem of understanding infinite sums.

The content of Section 5.1.1 is original, but the authors would like to acknowledge the open textbook used for fact-checking: Catherine Schmidt-Jones, Sound, Physics and Music. OpenStax CNX. Mar. 27, 2013

<http://cnx.org/contents/18e41aa3-0133-4bd1-84ae-2975f4d0ddaf>.

## 5.2▲ Series

A series is a sum

$$a_1 + a_2 + a_3 + \dots + a_n + \dots$$

of infinitely many terms. In summation notation, it is written

$$\sum_{n=1}^{\infty} a_n$$

You already have a lot of experience with series, though you might not realise it. When you write a number using its decimal expansion you are really expressing it as a series. Perhaps the simplest example of this is the decimal expansion of  $\frac{1}{3}$ :

$$\frac{1}{3} = 0.3333 \dots$$

Recall that the expansion written in this way actually means

$$0.333333 \dots = \frac{3}{10} + \frac{3}{100} + \frac{3}{1000} + \frac{3}{10000} + \dots = \sum_{n=1}^{\infty} \frac{3}{10^n}$$

The summation index  $n$  is of course a dummy index. You can use any symbol you like (within reason) for the summation index.

$$\sum_{n=1}^{\infty} \frac{3}{10^n} = \sum_{i=1}^{\infty} \frac{3}{10^i} = \sum_{j=1}^{\infty} \frac{3}{10^j} = \sum_{\ell=1}^{\infty} \frac{3}{10^\ell}$$

A series can be expressed using summation notation in many different ways. For example the following expressions all represent the same series:

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{3}{10^n} &= \overbrace{\frac{3}{10}}^{n=1} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \dots \\ \sum_{j=2}^{\infty} \frac{3}{10^{j-1}} &= \overbrace{\frac{3}{10}}^{j=2} + \overbrace{\frac{3}{100}}^{j=3} + \overbrace{\frac{3}{1000}}^{j=4} + \dots \\ \sum_{\ell=0}^{\infty} \frac{3}{10^{\ell+1}} &= \overbrace{\frac{3}{10}}^{\ell=0} + \overbrace{\frac{3}{100}}^{\ell=1} + \overbrace{\frac{3}{1000}}^{\ell=2} + \dots \\ \frac{3}{10} + \sum_{n=2}^{\infty} \frac{3}{10^n} &= \frac{3}{10} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \dots \end{aligned}$$

We can get from the first line to the second line by substituting  $n = j - 1$  — don't forget to also change the limits of summation (so that  $n = 1$  becomes  $j - 1 = 1$  which is rewritten as  $j = 2$ ). To get from the first line to the third line, substitute  $n = \ell + 1$  everywhere, including in the limits of summation (so that  $n = 1$  becomes  $\ell + 1 = 1$  which is rewritten as  $\ell = 0$ ).

Whenever you are in doubt as to what series a summation notation expression represents, it is a good habit to write out the first few terms, just as we did above.

Of course, at this point, it is not clear whether the sum of infinitely many terms adds up to a finite number or not. In order to make sense of this we will recast the problem in terms of the convergence of sequences (hence the discussion of the previous section). Before we proceed more formally let us illustrate the basic idea with a few simple examples.

Example 5.2.1  $\left( \sum_{n=1}^{\infty} \frac{3}{10^n} \right)$

As we have just seen above the series  $\sum_{n=1}^{\infty} \frac{3}{10^n}$  is

$$\sum_{n=1}^{\infty} \frac{3}{10^n} = \overbrace{\frac{3}{10}}^{n=1} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \dots$$

Notice that the  $n^{\text{th}}$  term in that sum is

$$3 \times 10^{-n} = 0.\overbrace{00 \dots 0}^{n-1 \text{ zeroes}} 3$$

So the sum of the first 5, 10, 15 and 20 terms in that series are

$$\sum_{n=1}^5 \frac{3}{10^n} = 0.33333$$

$$\sum_{n=1}^{10} \frac{3}{10^n} = 0.3333333333$$

$$\sum_{n=1}^{15} \frac{3}{10^n} = 0.333333333333333$$

$$\sum_{n=1}^{20} \frac{3}{10^n} = 0.33333333333333333333$$

It sure looks like that, as we add more and more terms, we get closer and closer to  $0.\dot{3} = \frac{1}{3}$ . So it is very reasonable<sup>11</sup> to define  $\sum_{n=1}^{\infty} \frac{3}{10^n}$  to be  $\frac{1}{3}$ .

Example 5.2.1

Example 5.2.2  $\left( \sum_{n=1}^{\infty} 1 \text{ and } \sum_{n=1}^{\infty} (-1)^n \right)$

Every term in the series  $\sum_{n=1}^{\infty} 1$  is exactly 1. So the sum of the first  $N$  terms is exactly  $N$ . As we add more and more terms this grows unboundedly. So it is very reasonable to say that the series  $\sum_{n=1}^{\infty} 1$  diverges.

The series

$$\sum_{n=1}^{\infty} (-1)^n = \overbrace{(-1)}^{n=1} + \overbrace{1}^{n=2} + \overbrace{(-1)}^{n=3} + \overbrace{1}^{n=4} + \overbrace{(-1)}^{n=5} + \dots$$

So the sum of the first  $N$  terms is 0 if  $N$  is even and  $-1$  if  $N$  is odd. As we add more and more terms from the series, the sum alternates between 0 and  $-1$  for ever and ever. So the

11 Of course we are free to define the series to be whatever we want. The hard part is defining it to be something that makes sense and doesn't lead to contradictions. We'll get to a more systematic definition shortly.

sum of all infinitely many terms does not make any sense and it is again reasonable to say that the series  $\sum_{n=1}^{\infty} (-1)^n$  diverges.

Example 5.2.2

In the above examples we have tried to understand the series by examining the sum of the first few terms and then extrapolating as we add in more and more terms. That is, we tried to sneak up on the infinite sum by looking at the limit of (partial) sums of the first few terms. This approach can be made into a more formal rigorous definition. More precisely, to define what is meant by the infinite sum  $\sum_{n=1}^{\infty} a_n$ , we approximate it by the sum of its first  $N$  terms and then take the limit as  $N$  tends to infinity.

**Definition 5.2.3.**

The  $N^{\text{th}}$  partial sum of the series  $\sum_{n=1}^{\infty} a_n$  is the sum of its first  $N$  terms

$$S_N = \sum_{n=1}^N a_n.$$

The partial sums form a sequence  $\{S_N\}_{N=1}^{\infty}$ . If this sequence of partial sums converges  $S_N \rightarrow S$  as  $N \rightarrow \infty$  then we say that the series  $\sum_{n=1}^{\infty} a_n$  converges to  $S$  and we write

$$\sum_{n=1}^{\infty} a_n = S$$

If the sequence of partial sums diverges, we say that the series diverges.

**5.2.1 ▶ Geometric Series**

Example 5.2.4 (Geometric Series)

Let  $a$  and  $r$  be any two fixed real numbers with  $a \neq 0$ . The series

$$a + ar + ar^2 + \dots + ar^n + \dots = \sum_{n=0}^{\infty} ar^n$$

is called the geometric series with first term  $a$  and ratio  $r$ .

Notice that we have chosen to start the summation index at  $n = 0$ . That's fine. The first<sup>12</sup> term is the  $n = 0$  term, which is  $ar^0 = a$ . The second term is the  $n = 1$  term, which is  $ar^1 = ar$ . And so on. We could have also written the series  $\sum_{n=1}^{\infty} ar^{n-1}$ . That's exactly the same series — the first term is  $ar^{n-1}|_{n=1} = ar^{1-1} = a$ , the second term is

12 It is actually quite common in computer science to think of 0 as the first integer. In that context, the set of natural numbers is defined to contain 0:

$$\mathbb{N} = \{0, 1, 2, \dots\}$$

$ar^{n-1}|_{n=2} = ar^{2-1} = ar$ , and so on<sup>13</sup>. Regardless of how we write the geometric series,  $a$  is the first term and  $r$  is the ratio between successive terms.

Geometric series have the extremely useful property that there is a very simple formula for their partial sums. Denote the partial sum by

$$S_N = \sum_{n=0}^N ar^n = a + ar + ar^2 + \dots + ar^N.$$

The secret to evaluating this sum is to see what happens when we multiply it  $r$ :

$$\begin{aligned} rS_N &= r(a + ar + ar^2 + \dots + ar^N) \\ &= ar + ar^2 + ar^3 + \dots + ar^{N+1} \end{aligned}$$

Notice that this is almost the same<sup>14</sup> as  $S_N$ . The only differences are that the first term,  $a$ , is missing and one additional term,  $ar^{N+1}$ , has been tacked on the end. So

$$\begin{aligned} S_N &= a + ar + ar^2 + \dots + ar^N \\ rS_N &= ar + ar^2 + \dots + ar^N + ar^{N+1} \end{aligned}$$

Hence taking the difference of these expressions cancels almost all the terms:

$$(1 - r)S_N = a - ar^{N+1} = a(1 - r^{N+1})$$

Provided  $r \neq 1$  we can divide both side by  $1 - r$  to isolate  $S_N$ :

$$S_N = a \cdot \frac{1 - r^{N+1}}{1 - r}.$$

On the other hand, if  $r = 1$ , then

$$S_N = \underbrace{a + a + \dots + a}_{N+1 \text{ terms}} = a(N + 1)$$

So in summary:

$$S_N = \begin{cases} a \frac{1-r^{N+1}}{1-r} & \text{if } r \neq 1 \\ a(N + 1) & \text{if } r = 1 \end{cases} \tag{5.2.1}$$

while the notation

$$\mathbb{Z}^+ = \{1, 2, 3, \dots\}$$

is used to denote the (strictly) positive integers. Remember that in this text, as is more standard in mathematics, we define the set of natural numbers to be the set of (strictly) positive integers.

13 This reminds the authors of the paradox of Hilbert’s hotel. The hotel with an infinite number of rooms is completely full, but can always accommodate one more guest. The interested reader should use their favourite search engine to find more information on this.

14 One can find similar properties of other special series, that allow us, with some work, to cancel many terms in the partial sums. We will shortly see a good example of this. The interested reader should look up “creative telescoping” to see how this idea might be used more generally, though it is somewhat beyond this course.

Now that we have this expression we can determine whether or not the series converges. If  $|r| < 1$ , then  $r^{N+1}$  tends to zero as  $N \rightarrow \infty$ , so that  $S_N$  converges to  $\frac{a}{1-r}$  as  $N \rightarrow \infty$  and

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r} \text{ provided } |r| < 1. \tag{5.2.2}$$

On the other hand if  $|r| \geq 1$ ,  $S_N$  diverges. To understand this divergence, consider the following 4 cases:

- If  $r > 1$ , then  $r^N$  grows to  $\infty$  as  $N \rightarrow \infty$ .
- If  $r < -1$ , then the magnitude of  $r^N$  grows to  $\infty$ , and the sign of  $r^N$  oscillates between  $+$  and  $-$ , as  $N \rightarrow \infty$ .
- If  $r = +1$ , then  $N + 1$  grows to  $\infty$  as  $N \rightarrow \infty$ .
- If  $r = -1$ , then  $r^N$  just oscillates between  $+1$  and  $-1$  as  $N \rightarrow \infty$ .

In each case the sequence of partial sums does not converge and so the series does not converge.

Example 5.2.4

Equations 5.2.1 and 5.2.2 are worth stating as a theorem.

**Theorem 5.2.5 (Geometric Series and Partial Sums).**

Let  $a$  and  $r$  be fixed real numbers, and let  $N$  be a positive integer. Then

$$\sum_{n=0}^N ar^n = \begin{cases} a \cdot \frac{1-r^{N+1}}{1-r} & \text{if } r \neq 1 \\ a(N+1) & \text{if } r = 1 \end{cases}$$

and

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r} \text{ provided } |r| < 1.$$

If  $|r| \geq 1$  and  $a \neq 0$ , then the series  $\sum_{n=0}^{\infty} ar^n$  diverges.

Example 5.2.6 (Bitcoin Supply)

Bitcoin is a virtual currency that mimics traditional currencies in a number of ways. One of those ways is controlled supply<sup>15</sup>. That is, new bitcoins enter circulation over time in a controlled manner.

15 Source for the specifics in this example: *Controlled Supply*, Bitcoin Wiki, url [https://en.bitcoin.it/wiki/Controlled\\_supply](https://en.bitcoin.it/wiki/Controlled_supply) accessed 16 Aug 2020

New *blocks*<sup>16</sup> are searched for by computers. When a block is found, it is converted into a set number of new bitcoins (owned by the finder). This is the *reward* for finding a block.

This process is analogous to mining precious metals which then are added to the currency supply, so the process of finding new blocks is often called mining. Importantly, the bitcoins given in the reward are new bitcoins that did not exist before the block was found. So, finding blocks is how bitcoins are *created*.

The reward for finding a block started at 50 bitcoins, but it halves every 210,000 blocks. The miners who found block 0, block 1, and block 209,999 each got a reward of 50 bitcoins. Then, the miners who found block 210,000 through block 419,999 each got a reward of 25 bitcoins, and so on.

For the purposes of this example, we will assume that miners will always be able to find blocks. (That is, blocks never run out.) We will also assume that rewards for finding blocks are the only ways bitcoins are ever created, and that bitcoins are never destroyed.

- (a) Suppose bitcoins are infinitely divisible. (That is, you can have an arbitrarily small portion of a bitcoin, such as one-trillionth of a bitcoin, without a limit on how small that portion can be.) If miners continue finding blocks for an infinite period of time, what will happen to the total supply of bitcoins?
- (b) One *Satoshi* (or one *sat*) is equal to  $1/100,000,000$  bitcoin. Suppose when the reward for a block is scheduled to be less than one satoshi, the block finder actually gets a reward of 0 bitcoins. That is, there are no more bitcoins created when the reward for finding a new block dips below one satoshi. If miners continue finding blocks for an infinite period of time, what will happen to the total supply of bitcoins?

*Solution.*

- (a) Let's model the number of bitcoins by grouping together collections of 210,000 blocks.
- For the first collection of 210,000 blocks, the number of bitcoins created is 50 each, for a total of  $210,000 \cdot 50$  bitcoins created.
  - For the second collection of 210,000 blocks, the number of bitcoins created is  $\frac{50}{2} = 25$  each, for a total of  $210,000 \cdot \frac{50}{2}$  bitcoins created.
  - For the third collection of 210,000 blocks, the number of bitcoins created is  $\frac{50}{4} = \frac{25}{2} = 12.5$  each, for a total of  $210,000 \cdot \frac{50}{4}$  bitcoins created.
  - In general, for the  $n$ th collection of 210,000 blocks, the total number of bitcoins created by those blocks is  $210,000 \cdot \frac{50}{2^{n-1}}$  bitcoins.
  - All together, the number of bitcoins created by an infinite collection of blocks is

$$\sum_{n=1}^{\infty} 210,000 \cdot \frac{50}{2^{n-1}} = 210,000 \cdot 50 \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^{n-1}$$

<sup>16</sup> For the purposes of this question, the technical details are not important. What you need to know about blocks is that you find them and they get turned into currency.



This series almost, but not exactly, looks like the series from Theorem 5.2.5. We'll expand the series<sup>17</sup> in order to see how we might have indexed the terms differently.

$$\begin{aligned} 210,000 \cdot 50 \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^{n-1} &= 210,000 \cdot 50 \left( \left(\frac{1}{2}\right)^0 + \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \dots \right) \\ &= 210,000 \cdot 50 \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n \end{aligned}$$

Now we can apply Theorem 5.2.5 with  $r = \frac{1}{2}$ .

$$= 210,000 \cdot 50 \cdot \frac{1}{1 - \frac{1}{2}} = 210,000 \cdot 50 \cdot 2 = 21,000,000$$

As blocks are mined, the total number of bitcoins will approach 21 million. It will never exceed 21 million.

- (b) For this part we assume that after a certain number of blocks, no more bitcoin are created. So, we will look at a finite sum, rather than an infinite series. Let's start by figuring out when the reward for a block drops below 1 satoshi.

The  $n$ th batch of 210,000 blocks earns  $\frac{50}{2^{n-1}}$  bitcoins, as long as that number is greater than or equal to one satoshi. That is, we create bitcoins as long as:

$$\frac{50}{2^{n-1}} \geq \frac{1}{100,000,000} = \frac{1}{10^8}$$

Solving for  $n$ :

$$\begin{aligned} 5 \cdot 10^9 &\geq 2^{n-1} \\ \log_2(5 \cdot 10^9) &\geq n - 1 \\ 1 + \log_2(5 \cdot 10^9) &\geq n \end{aligned}$$

Note  $n$  only makes sense as an integer. Using a calculator,  $1 + \log_2(5 \cdot 10^9) \approx 33.2$ . So when  $n = 33$ , blocks earn rewards, but when  $n \geq 34$ , they do not.

The means the total supply of bitcoins that could ever be created under this system is:

$$\begin{aligned} \sum_{n=1}^{33} 210,000 \cdot \frac{50}{2^{n-1}} &= 210,000 \cdot 50 \left( \left(\frac{1}{2}\right)^0 + \left(\frac{1}{2}\right)^1 + \left(\frac{1}{2}\right)^2 + \dots + \left(\frac{1}{2}\right)^{32} \right) \\ &= 210,000 \cdot 50 \sum_{n=0}^{32} \frac{1}{2^n} \end{aligned}$$

---

17 indexing from 0 (starting with the 0th collection, then the 1st collection in the bullet list) would have eliminated this upcoming step. We described the creation of the series using the indexing that we thought would be most intuitive to our readers, rather than the indexing that would lead to the least amount of algebra.

Now we can apply Theorem 5.2.5 with  $r = \frac{1}{2}$  and  $N = 32$ .

$$= 210,000 \cdot 50 \cdot \frac{1 - \left(\frac{1}{2}\right)^{33}}{1 - \frac{1}{2}} = 210,000 \cdot 100 \cdot \left(1 - \frac{1}{2^{33}}\right)$$

Using a calculator,

$$\approx 20,999,999.997555278$$

So the total supply of bitcoins approaches 20,999,999 bitcoins and 99,755,528 satoshi, but never exceeds this amount.

Example 5.2.6

### 5.2.2 ▶ Telescoping Series

Typically, it is quite difficult to write down a neat closed form expression for the partial sums of a series. Geometric series are very notable exceptions to this. Another family of series for which we can write down partial sums is called “telescoping series”. These series have the desirable property that many of the terms in the sum cancel each other out rendering the partial sums quite simple.

Example 5.2.7 (Telescoping Series)

In this example, we are going to study the series  $\sum_{n=1}^{\infty} \frac{1}{n(n+1)}$ . This is a rather artificial series<sup>18</sup> that has been rigged to illustrate a phenomenon call “telescoping”. Notice that the  $n^{\text{th}}$  term can be rewritten as

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$$

and so we have

$$a_n = b_n - b_{n+1} \quad \text{where } b_n = \frac{1}{n}.$$

Because of this we get big cancellations when we add terms together. This allows us to get a simple formula for the partial sums of this series.

$$\begin{aligned} S_N &= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{N \cdot (N+1)} \\ &= \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{N} - \frac{1}{N+1}\right) \end{aligned}$$

<sup>18</sup> Well... this sort of series does show up when you start to look at the Maclaurin polynomial of functions like  $(1-x)\ln(1-x)$ . So it is not *totally* artificial. At any rate, it illustrates the basic idea of telescoping very nicely, and the idea of “creative telescoping” turns out to be extremely useful in the study of series — though it is well beyond the scope of this course.

The second term of each bracket exactly cancels the first term of the following bracket. So the sum “telescopes” leaving just

$$S_N = 1 - \frac{1}{N+1}$$

and we can now easily compute

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{N \rightarrow \infty} S_N = \lim_{N \rightarrow \infty} \left(1 - \frac{1}{N+1}\right) = 1$$

Example 5.2.7

More generally, if we can write

$$a_n = b_n - b_{n+1}$$

for some other known sequence  $b_n$ , then the series telescopes and we can compute partial sums using

$$\begin{aligned} \sum_{n=1}^N a_n &= \sum_{n=1}^N (b_n - b_{n+1}) \\ &= \sum_{n=1}^N b_n - \sum_{n=1}^N b_{n+1} \\ &= b_1 - b_{N+1}. \end{aligned}$$

and hence

$$\sum_{n=1}^{\infty} a_n = b_1 - \lim_{N \rightarrow \infty} b_{N+1}$$

provided this limit exists. Often  $\lim_{N \rightarrow \infty} b_{N+1} = 0$  and then  $\sum_{n=1}^{\infty} a_n = b_1$ . But this does not always happen. Here is an example.

Example 5.2.8 (A Divergent Telescoping Series)

In this example, we are going to study the series  $\sum_{n=1}^{\infty} \log\left(1 + \frac{1}{n}\right)$ . (We don't specify the base — any base greater than one will behave the same way.) Let's start by just writing out the first few terms.

$$\begin{aligned} \sum_{n=1}^{\infty} \log\left(1 + \frac{1}{n}\right) &= \overbrace{\log\left(1 + \frac{1}{1}\right)}^{n=1} + \overbrace{\log\left(1 + \frac{1}{2}\right)}^{n=2} + \overbrace{\log\left(1 + \frac{1}{3}\right)}^{n=3} + \overbrace{\log\left(1 + \frac{1}{4}\right)}^{n=4} + \dots \\ &= \log(2) + \log\left(\frac{3}{2}\right) + \log\left(\frac{4}{3}\right) + \log\left(\frac{5}{4}\right) + \dots \end{aligned}$$

This is pretty suggestive since

$$\log(2) + \log\left(\frac{3}{2}\right) + \log\left(\frac{4}{3}\right) + \log\left(\frac{5}{4}\right) = \log\left(2 \times \frac{3}{2} \times \frac{4}{3} \times \frac{5}{4}\right) = \log(5)$$

So let's try using this idea to compute the partial sum  $S_N$ :

$$\begin{aligned} S_N &= \sum_{n=1}^N \log\left(1 + \frac{1}{n}\right) \\ &= \overbrace{\log\left(1 + \frac{1}{1}\right)}^{n=1} + \overbrace{\log\left(1 + \frac{1}{2}\right)}^{n=2} + \overbrace{\log\left(1 + \frac{1}{3}\right)}^{n=3} + \cdots + \overbrace{\log\left(1 + \frac{1}{N-1}\right)}^{n=N-1} + \overbrace{\log\left(1 + \frac{1}{N}\right)}^{n=N} \\ &= \log(2) + \log\left(\frac{3}{2}\right) + \log\left(\frac{4}{3}\right) + \cdots + \log\left(\frac{N}{N-1}\right) + \log\left(\frac{N+1}{N}\right) \\ &= \log\left(2 \times \frac{3}{2} \times \frac{4}{3} \times \cdots \times \frac{N}{N-1} \times \frac{N+1}{N}\right) \\ &= \log(N+1) \end{aligned}$$

Uh oh!

$$\lim_{N \rightarrow \infty} S_N = \lim_{N \rightarrow \infty} \log(N+1) = +\infty$$

This telescoping series diverges! There is an important lesson here. Telescoping series *can* diverge. They do not always converge to  $b_1$ .

Example 5.2.8

### 5.2.3 ▶ Arithmetic of Series

As was the case for limits, differentiation and antidifferentiation, we can compute more complicated series in terms of simpler ones by understanding how series interact with the usual operations of arithmetic. It is, perhaps, not so surprising that there are simple rules for addition and subtraction of series and for multiplication of a series by a constant. Unfortunately there are no simple general rules for computing products or ratios of series.

**Theorem 5.2.9** (Arithmetic of series).

Let  $A, B$  and  $C$  be real numbers and let the two series  $\sum_{n=1}^{\infty} a_n$  and  $\sum_{n=1}^{\infty} b_n$  converge to  $S$  and  $T$  respectively. That is, assume that

$$\sum_{n=1}^{\infty} a_n = S \qquad \sum_{n=1}^{\infty} b_n = T$$

Then the following hold.

(a)  $\sum_{n=1}^{\infty} [a_n + b_n] = S + T$       and       $\sum_{n=1}^{\infty} [a_n - b_n] = S - T$

(b)  $\sum_{n=1}^{\infty} C a_n = CS.$

**Example 5.2.10**

As a simple example of how we use the arithmetic of series Theorem 5.2.9, consider

$$\sum_{n=1}^{\infty} \left[ \frac{1}{7^n} + \frac{2}{n(n+1)} \right]$$

We recognize that we know how to compute parts of this sum. We know that

$$\sum_{n=1}^{\infty} \frac{1}{7^n} = \frac{1/7}{1 - 1/7} = \frac{1}{6}$$

because it is a geometric series (Example 5.2.4) with first term  $a = \frac{1}{7}$  and ratio  $r = \frac{1}{7}$ . And we know that

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = 1$$

by Example 5.2.7. We can now use Theorem 5.2.9 to build the specified “complicated” series out of these two “simple” pieces.

$$\begin{aligned} \sum_{n=1}^{\infty} \left[ \frac{1}{7^n} + \frac{2}{n(n+1)} \right] &= \sum_{n=1}^{\infty} \frac{1}{7^n} + \sum_{n=1}^{\infty} \frac{2}{n(n+1)} && \text{by Theorem 5.2.9.a} \\ &= \sum_{n=1}^{\infty} \frac{1}{7^n} + 2 \sum_{n=1}^{\infty} \frac{1}{n(n+1)} && \text{by Theorem 5.2.9.b} \\ &= \frac{1}{6} + 2 \cdot 1 = \frac{13}{6} \end{aligned}$$

**Example 5.2.10**

### 5.2.4 ▶ (Optional) Intergenerational Cost-Benefit Analysis

This subsection presents ideas from the article<sup>19</sup> *Intergenerational cost–benefit analysis and marine ecosystem restoration* by UBC Institute for the Oceans and Fisheries Professor Ussif Rashid Sumaila.

Generally we value the promise of money in the future less than we value the possession of money in the present. The *discounting rate* describes the loss of value that occurs with time, and is calculated like interest. For example, suppose we have a discounting rate of 10%. That means  $D$  dollars in our possession today has the same value to us as  $D(1 + 0.1) = 1.1 \cdot D$  dollars promised to us in one year. These both have the same value to us as  $(1.1)(1.1 \cdot D) = 1.1^2 \cdot D$  dollars in two years, or  $1.1^t \cdot D$  dollars in  $t$  years:

$$D \text{ present-day dollars} = (1.1^t \cdot D) \text{ future dollars}$$

Dividing both sides of the equation by  $1.1^t$ , we see that the promise of  $D$  dollars in  $t$  years is worth the same to us as the possession of  $\frac{D}{1.1^t}$  dollars in the present:

$$D \text{ future dollars} = \frac{D \text{ present-day dollars}}{1.1^t}$$

In a conventional cost-benefit analysis (CBA), returns that will happen in the future are subject to precisely this form of discounting. To quantify the value of a project, units of Present Value (PV) are used. Given a discounting rate<sup>20</sup> of  $\delta$ , possession of  $D$  dollars today has the same value as a gain of  $(1 + \delta)^t D$  dollars  $t$  years from now. Rearranged, the present value of  $D$  dollars that will be gained  $t$  years in the future is given by

$$\text{PV}(D, t) = \frac{D}{(1 + \delta)^t}$$

Future discounting is human nature, but it doesn't always make for good policy. In particular, "high discount rates favour myopic fisheries policies resulting in global overfishing" (p. 334) since the model makes the health of an ecosystem one hundred years from now worth almost nothing today.

Sumaila proposes an intergenerational model, where discounting still happens within a generation of people, but different generations are considered together. Quoting the article:

"The benefits to the current generation from the use of ecosystem resources today would never have appeared in the conventional CBA[Cost-Benefit Analysis] of the generations that were here a hundred years ago. Similarly, the generation that will be here in a hundred years time, would receive benefits from restored marine ecosystems that would mean much to them but would not appear in the current generation's conventional CBA. Therefore, to capture the benefits to all generations from ecosystem restoration projects, it is necessary to use [an intergenerational] CBA approach" (p. 336).

19 Sumaila UR. Intergenerational cost–benefit analysis and marine ecosystem restoration. *Fish and fisheries* (Oxford, England). 2004;5(4):329-43. You can access the full text online with your UBC CWL (campus-wide login) here: [https://libkey.io/libraries/498/articles/30981866/full-text-file?utm\\_source=api\\_542](https://libkey.io/libraries/498/articles/30981866/full-text-file?utm_source=api_542).

20 To better understand the rate, note that if  $\delta = 0$ , then \$1 today is worth the same to us as \$1 one year from now, 100 years from now, or at any other time in the future.

The approach proposed by Sumaila is as follows. We divide up the future into distinct generations, each of which reigns over a (non-overlapping) interval of time. Each generation has its own Present Value calculation, measured from the start of its reign. So the Present Value of the promise of  $D$  dollars in year  $t$ , to a generation that started its reign in year  $t_0$ , is

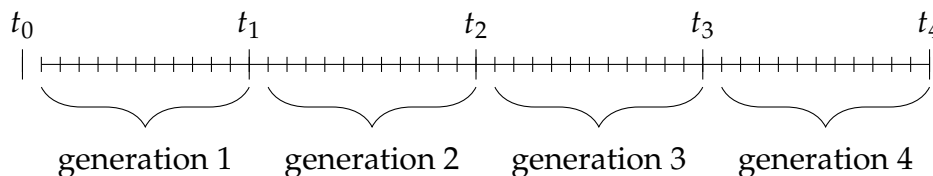
$$PV(D, t) = \frac{D}{(1 + \delta)^{t-t_0}}$$

The difference between this calculation and the conventional PV calculation is that “present” is relative for each generation.

Now that we have these components, we can create an expression for a cost-benefit analysis (CBA) of a long-term project.

Suppose in year  $t$ , the costs incurred by the project are given by  $C_t$ , and the benefits are given by  $V_t$ . The *net* value gained in that year is  $V_t - C_t$ , before future discounting is applied. If the generation started its reign in year  $t = t_0$ , then the present value of  $(V_t - C_t)$  to that generation is  $\frac{V_t - C_t}{(1 + \delta)^{t-t_0}}$ . If the generation reigns from  $t = t_0$  to  $t = t_1$ , then we combine the net present value of each of those years to find the net present value to the generation of the entire project.

To include a *collection* of generations, we add up each generation’s Net Present Value. To express this in sigma notation, let  $NPV_k$  be the Net Present Value for the  $k$ th generation. We’ll index years as follows. The first generation reigns from  $t = t_0 + 1$  to  $t = t_1$ ; the second generation reigns from  $t = t_1 + 1$  to  $t = t_2$ ; and (in general) the  $k$ th generation reigns from  $t = t_{k-1} + 1$  to  $t = t_k$ . (Considering the first year to be  $t = t_0 + 1$  looks weird, but makes the indices more consistent with one another.)



Then generation  $k$  has its personal Net Present Value given by

$$NPV_k = \sum_{t=t_{k-1}+1}^{t_k} \frac{V_t - C_t}{(1 + \delta)^{t-t_n}}$$

All together, the intergenerational Net Present Value of a project, from generation 1 to generation  $L$ , is

$$\begin{aligned} NPV &= \sum_{k=1}^L NPV_k \\ &= \sum_{k=1}^L \left( \sum_{t=t_{k-1}+1}^{t_k} \frac{V_t - C_t}{(1 + \delta)^{t-t_n}} \right) \end{aligned}$$

If the NPV is positive, then the project is a good investment: adjusting for discounting, but considering future generations, the benefits will exceed the costs. If the NPV is negative, then the project is a bad investment.

### 5.3▲ The Integral and Divergence Tests

It is very common to encounter series for which it is difficult, or even virtually impossible, to determine the sum exactly. Often you try to evaluate the sum approximately by truncating it, i.e. having the index run only up to some finite  $N$ , rather than infinity. But there is no point in doing so if the series diverges. So you like to at least know if the series converges or diverges. Furthermore you would also like to know what error is introduced when you approximate  $\sum_{n=1}^{\infty} a_n$  by the “truncated series”  $\sum_{n=1}^N a_n$ . That’s called the truncation error. There are a number of “convergence tests” to help you with this.

Our first test is very easy to apply, but it is also rarely useful. It just allows us to quickly reject some “trivially divergent” series. It is based on the observation that

- by definition, a series  $\sum_{n=1}^{\infty} a_n$  converges to  $S$  when the partial sums  $S_N = \sum_{n=1}^N a_n$  converge to  $S$ .
- Then, as  $N \rightarrow \infty$ , we have  $S_N \rightarrow S$  and, because  $N - 1 \rightarrow \infty$  too, we also have  $S_{N-1} \rightarrow S$ .
- So  $a_N = S_N - S_{N-1} \rightarrow S - S = 0$ .

This tells us that, if we already know that a given series  $\sum a_n$  is convergent, then the  $n^{\text{th}}$  term of the series,  $a_n$ , must converge to 0 as  $n$  tends to infinity. In this form, the test is not so useful. However the contrapositive<sup>21</sup> of the statement is a useful test for *divergence*.

**Theorem 5.3.1 (Divergence Test).**

If the sequence  $\{a_n\}_{n=1}^{\infty}$  fails to converge to zero as  $n \rightarrow \infty$ , then the series  $\sum_{n=1}^{\infty} a_n$  diverges.

**Example 5.3.2**

Let  $a_n = \frac{n}{n+1}$ . Then

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} \frac{n}{n+1} = \lim_{n \rightarrow \infty} \frac{1}{1 + 1/n} = 1 \neq 0$$

So the series  $\sum_{n=1}^{\infty} \frac{n}{n+1}$  diverges.

**Example 5.3.2**

21 Given a statement of the form “If A is true, then B is true” the contrapositive is “If B is not true, then A is not true”. The two statements in quotation marks are logically equivalent — if one is true, then so is the other. In the present context we have

If  $(\sum a_n \text{ converges})$  then  $(a_n \text{ converges to } 0)$ .

The contrapositive of this statement is then

If  $(a_n \text{ does not converge to } 0)$  then  $(\sum a_n \text{ does not converge})$ .



**Warning 5.3.3.**

The divergence test is a “one way test”. It tells us that if  $\lim_{n \rightarrow \infty} a_n$  is nonzero, or fails to exist, then the series  $\sum_{n=1}^{\infty} a_n$  diverges. But it tells us *absolutely nothing* when  $\lim_{n \rightarrow \infty} a_n = 0$ . In particular, it is perfectly possible for a series  $\sum_{n=1}^{\infty} a_n$  to *diverge* even though  $\lim_{n \rightarrow \infty} a_n = 0$ . An example is  $\sum_{n=1}^{\infty} \frac{1}{n}$ . We’ll show in Example 5.3.7, below, that it diverges.

Now while convergence or divergence of series like  $\sum_{n=1}^{\infty} \frac{1}{n}$  can be determined using some clever tricks, it would be much better to have methods that are more systematic and rely less on being sneaky. Over the next subsections we will discuss several methods for testing series for convergence.

Note that while these tests will tell us whether or not a series converges, they do not (except in rare cases) tell us what the series adds up to. For example, the test we will see in the next subsection tells us quite immediately that the series

$$\sum_{n=1}^{\infty} \frac{1}{n^3}$$

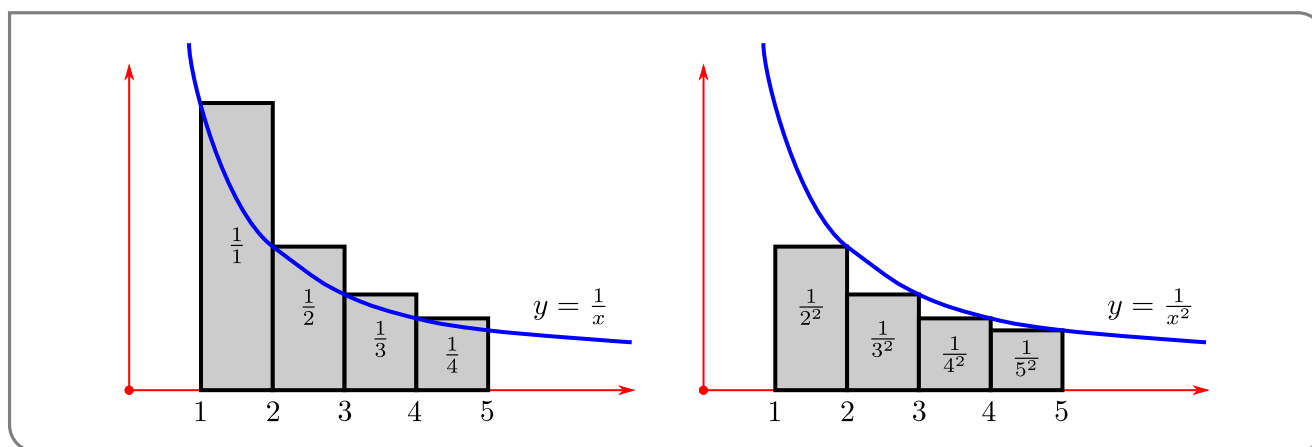
converges. However it does not tell us its value<sup>22</sup>.

In the integral test, we think of a series  $\sum_{n=1}^{\infty} a_n$ , that we cannot evaluate explicitly, as the area of a union of rectangles, with  $a_n$  representing the area of a rectangle of width one and height  $a_n$ . Then we compare that area with the area represented by an integral, that we can evaluate explicitly, much as we did in Theorem 3.7.18, the comparison test for improper integrals. We’ll start with a simple example, to illustrate the idea. Then we’ll move on to a formulation of the test in general.

**Example 5.3.4**

Visualise the terms of the harmonic series  $\sum_{n=1}^{\infty} \frac{1}{n}$  as a bar graph — each term is a rectangle of height  $\frac{1}{n}$  and width 1. The limit of the series is then the limiting area of this union of rectangles. Consider the sketch on the left below.

22 This series converges to Apéry’s constant  $1.2020569031 \dots$ . The constant is named for Roger Apéry (1916–1994) who proved that this number must be irrational. This number appears in many contexts including the following cute fact — the reciprocal of Apéry’s constant gives the probability that three positive integers, chosen at random, do not share a common prime factor.



It shows that the area of the shaded columns,  $\sum_{n=1}^4 \frac{1}{n}$ , is bigger than the area under the curve  $y = \frac{1}{x}$  with  $1 \leq x \leq 5$ . That is

$$\sum_{n=1}^4 \frac{1}{n} \geq \int_1^5 \frac{1}{x} dx$$

If we were to continue drawing the columns all the way out to infinity, then we would have

$$\sum_{n=1}^{\infty} \frac{1}{n} \geq \int_1^{\infty} \frac{1}{x} dx$$

We are able to compute this improper integral exactly:

$$\int_1^{\infty} \frac{1}{x} dx = \lim_{R \rightarrow \infty} \left[ \ln |x| \right]_1^R = +\infty$$

That is the area under the curve diverges to  $+\infty$  and so the area represented by the columns must also diverge to  $+\infty$ .

It should be clear that the above argument can be quite easily generalised. For example the same argument holds *mutatis mutandis*<sup>23</sup> for the series

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

Indeed we see from the sketch on the right above that

$$\sum_{n=2}^N \frac{1}{n^2} \leq \int_1^N \frac{1}{x^2} dx$$

and hence

$$\sum_{n=2}^{\infty} \frac{1}{n^2} \leq \int_1^{\infty} \frac{1}{x^2} dx$$

<sup>23</sup> Latin for “Once the necessary changes are made”. This phrase still gets used a little, but these days mathematicians tend to write something equivalent in English. Indeed, English is pretty much the *lingua franca* for mathematical publishing. *Quidquid erit.*

This last improper integral is easy to evaluate:

$$\begin{aligned} \int_2^\infty \frac{1}{x^2} dx &= \lim_{R \rightarrow \infty} \left[ -\frac{1}{x} \right]_2^R \\ &= \lim_{R \rightarrow \infty} \left( \frac{1}{2} - \frac{1}{R} \right) = \frac{1}{2} \end{aligned}$$

Thus we know that

$$\sum_{n=1}^\infty \frac{1}{n^2} = 1 + \sum_{n=2}^\infty \frac{1}{n^2} \leq \frac{3}{2}.$$

and so the series must converge.

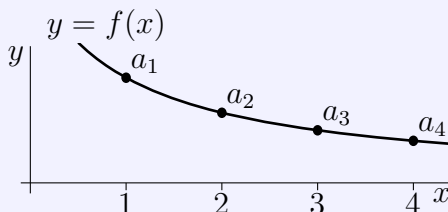
Example 5.3.4

The above arguments are formalised in the following theorem.

**Theorem 5.3.5 (The Integral Test).**

Let  $N_0$  be any natural number. If  $f(x)$  is a function which is defined and continuous for all  $x \geq N_0$  and which obeys

- (i)  $f(x) \geq 0$  for all  $x \geq N_0$ , and
- (ii)  $f(x)$  decreases or stays the same as  $x$  increases, and
- (iii)  $f(n) = a_n$  for all  $n \geq N_0$ .



Then

$$\sum_{n=1}^\infty a_n \text{ converges} \iff \int_{N_0}^\infty f(x) dx \text{ converges}$$

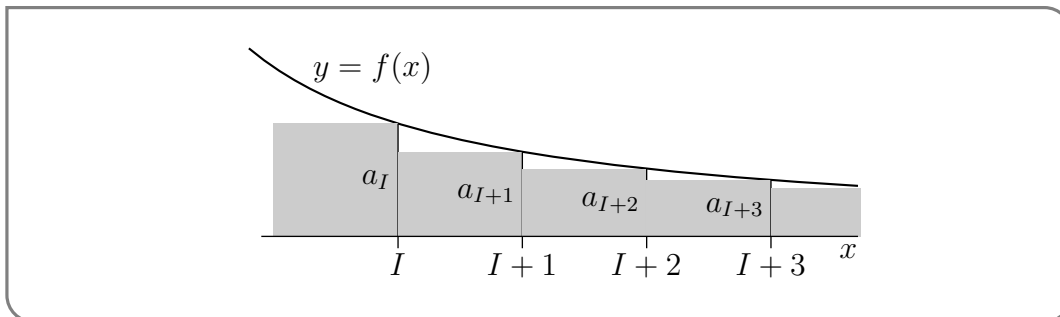
Furthermore, when the series converges, the truncation error

$$\left| \sum_{n=1}^\infty a_n - \sum_{n=1}^N a_n \right| \leq \int_N^\infty f(x) dx \quad \text{for all } N \geq N_0$$

*Proof.* Let  $I$  be any fixed integer with  $I > N_0$ . Then

- $\sum_{n=1}^\infty a_n$  converges if and only if  $\sum_{n=I}^\infty a_n$  converges — removing a fixed finite number of terms from a series cannot impact whether or not it converges.

- Since  $a_n \geq 0$  for all  $n \geq I > N_0$ , the sequence of partial sums  $s_\ell = \sum_{n=I}^{\ell} a_n$  obeys  $s_{\ell+1} = s_\ell + a_{\ell+1} \geq s_\ell$ . That is,  $s_\ell$  increases as  $\ell$  increases.
- So  $\{s_\ell\}$  must either converge to some finite number or increase to infinity. That is, either  $\sum_{n=I}^{\infty} a_n$  converges to a finite number or it is  $+\infty$ .



Look at the figure above. The shaded area in the figure is  $\sum_{n=I}^{\infty} a_n$  because

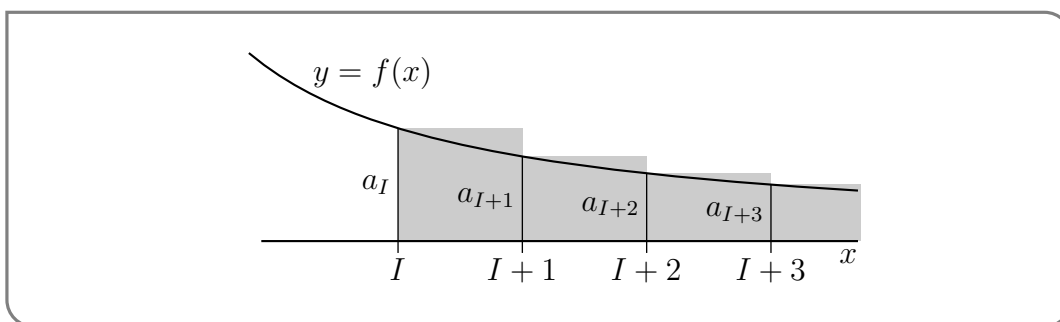
- the first shaded rectangle has height  $a_I$  and width 1, and hence area  $a_I$  and
- the second shaded rectangle has height  $a_{I+1}$  and width 1, and hence area  $a_{I+1}$ , and so on

This shaded area is smaller than the area under the curve  $y = f(x)$  for  $I - 1 \leq x < \infty$ . So

$$\sum_{n=I}^{\infty} a_n \leq \int_{I-1}^{\infty} f(x) dx$$

and, if the integral is finite, the sum  $\sum_{n=I}^{\infty} a_n$  is finite too. Furthermore, the desired bound on the truncation error is just the special case of this inequality with  $I = N + 1$ :

$$\sum_{n=1}^{\infty} a_n - \sum_{n=1}^N a_n = \sum_{n=N+1}^{\infty} a_n \leq \int_N^{\infty} f(x) dx$$



For the “divergence case” look at the figure above. The (new) shaded area in the figure is again  $\sum_{n=I}^{\infty} a_n$  because

- the first shaded rectangle has height  $a_I$  and width 1, and hence area  $a_I$  and
- the second shaded rectangle has height  $a_{I+1}$  and width 1, and hence area  $a_{I+1}$ , and so on

This time the shaded area is larger than the area under the curve  $y = f(x)$  for  $I \leq x < \infty$ . So

$$\sum_{n=I}^{\infty} a_n \geq \int_I^{\infty} f(x) dx$$

and, if the integral is infinite, the sum  $\sum_{n=I}^{\infty} a_n$  is infinite too. □

Now that we have the integral test, it is straightforward to determine for which values of  $p$  the series<sup>24</sup>

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges.

**Remark 5.3.6.** Theorem 5.3.5 requires  $f(x)$  to be non-increasing. If  $f(x)$  is increasing (or constant) while  $x$  increases, and  $f(x) > 0$  for all sufficiently large  $x$ , then  $\sum_{n=1}^{\infty} a_n$  (where  $a_n = f(n)$ ) is divergent by the divergence test. So if you feel the desire to use the integral test for an *increasing* function, remember that an easier option is available.

In some texts, the integral test is defined to allow increasing functions. The test as stated would indeed work with increasing functions, but as noted above, there is *always*<sup>25</sup> an easier way.

Example 5.3.7 (The  $p$  test:  $\sum_{n=1}^{\infty} \frac{1}{n^p}$ )

Let  $p > 0$ . We'll now use the integral test to determine whether or not the series  $\sum_{n=1}^{\infty} \frac{1}{n^p}$  (which is sometimes called the  $p$ -series) converges.

- To do so, we need a function  $f(x)$  that obeys  $f(n) = a_n = \frac{1}{n^p}$  for all  $n$  bigger than some  $N_0$ . Certainly  $f(x) = \frac{1}{x^p}$  obeys  $f(n) = \frac{1}{n^p}$  for all  $n \geq 1$ . So let's pick this  $f$  and try  $N_0 = 1$ . (We can always increase  $N_0$  later if we need to.)
- This function also obeys the other two conditions of Theorem 5.3.5:

24 This series, viewed as a function of  $p$ , is called the Riemann zeta function,  $\zeta(p)$ , or the Euler-Riemann zeta function. It is extremely important because of its connections to prime numbers (among many other things). Indeed Euler proved that

$$\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p} = \prod_{P \text{ prime}} (1 - P^{-p})^{-1}$$

Riemann showed the connections between the zeros of this function (over complex numbers  $p$ ) and the distribution of prime numbers. Arguably the most famous unsolved problem in mathematics, the Riemann hypothesis, concerns the locations of zeros of this function.

25 OK, OK, "always" is a strong term to a mathematician. If you know a function is positive and nondecreasing, you automatically know the associated series is divergent. But perhaps a scenario could be invented where you knew a function was either increasing or decreasing, but you didn't know which one (the general term is "monotone"), and somehow you could also integrate that function. We suppose in that case you might want the extended version of the integral test.

- (i)  $f(x) > 0$  for all  $x \geq N_0 = 1$  and
- (ii)  $f(x)$  decreases as  $x$  increases because  $f'(x) = -p \frac{1}{x^{p+1}} < 0$  for all  $x \geq N_0 = 1$ .

- So the integral test tells us that the series  $\sum_{n=1}^{\infty} \frac{1}{n^p}$  converges if and only if the integral  $\int_1^{\infty} \frac{dx}{x^p}$  converges.
- We have already seen, in Example 3.7.8, that the integral  $\int_1^{\infty} \frac{dx}{x^p}$  converges if and only if  $p > 1$ .

So we conclude that  $\sum_{n=1}^{\infty} \frac{1}{n^p}$  converges if and only if  $p > 1$ . This is sometimes called the  $p$ -test.

- In particular, the series  $\sum_{n=1}^{\infty} \frac{1}{n}$ , which is called the harmonic series, has  $p = 1$  and so diverges. As we add more and more terms of this series together, the terms we add, namely  $\frac{1}{n}$ , get smaller and smaller and tend to zero, but they tend to zero so slowly that the full sum is still infinite.
- On the other hand, the series  $\sum_{n=1}^{\infty} \frac{1}{n^{1.000001}}$  has  $p = 1.000001 > 1$  and so converges. This time as we add more and more terms of this series together, the terms we add, namely  $\frac{1}{n^{1.000001}}$ , tend to zero (just) fast enough that the full sum is finite. Mind you, for this example, the convergence takes place very slowly — you have to take a huge number of terms to get a decent approximation to the full sum. If we approximate  $\sum_{n=1}^{\infty} \frac{1}{n^{1.000001}}$  by the truncated series  $\sum_{n=1}^N \frac{1}{n^{1.000001}}$ , we make an error of at most

$$\int_N^{\infty} \frac{dx}{x^{1.000001}} = \lim_{R \rightarrow \infty} \int_N^R \frac{dx}{x^{1.000001}} = \lim_{R \rightarrow \infty} -\frac{1}{0.000001} \left[ \frac{1}{R^{0.000001}} - \frac{1}{N^{0.000001}} \right] = \frac{10^6}{N^{0.000001}}$$

This does tend to zero as  $N \rightarrow \infty$ , but really slowly.

Example 5.3.7

We now know that the dividing line between convergence and divergence of  $\sum_{n=1}^{\infty} \frac{1}{n^p}$  occurs at  $p = 1$ . We can dig a little deeper and ask ourselves how much more quickly than  $\frac{1}{n}$  the  $n^{\text{th}}$  term needs to shrink in order for the series to converge. We know that for large  $x$ , the function  $\log x$  (of any base) is smaller than  $x^a$  for any positive  $a$  — you can convince yourself of this with a quick application of L'Hôpital's rule. So it is not unreasonable to ask whether the series

$$\sum_{n=2}^{\infty} \frac{1}{n \ln n}$$

converges. Notice that we sum from  $n = 2$  because when  $n = 1$ ,  $n \ln n = 0$ . And we don't need to stop there<sup>26</sup>. We can analyse the convergence of this sum with any power of  $\ln n$ .

Example 5.3.8  $\left( \sum_{n=2}^{\infty} \frac{1}{n(\ln n)^p} \right)$

Let  $p > 0$ . We'll now use the integral test to determine whether or not the series  $\sum_{n=2}^{\infty} \frac{1}{n(\ln n)^p}$  converges.

<sup>26</sup> We could go even further and see what happens if we include powers of  $\ln(\ln(n))$  and other more exotic slow-growing functions.

- As in the last example, we start by choosing a function that obeys  $f(n) = a_n = \frac{1}{n(\ln n)^p}$  for all  $n$  bigger than some  $N_0$ . Certainly  $f(x) = \frac{1}{x(\ln x)^p}$  obeys  $f(n) = \frac{1}{n(\ln n)^p}$  for all  $n \geq 2$ . So let's use that  $f$  and try  $N_0 = 2$ .
- Now let's check the other two conditions of Theorem 5.3.5:
  - (i) Both  $x$  and  $\ln x$  are positive for all  $x > 1$ , so  $f(x) > 0$  for all  $x \geq N_0 = 2$ .
  - (ii) As  $x$  increases both  $x$  and  $\ln x$  increase and so  $x(\ln x)^p$  increases and  $f(x)$  decreases.
- So the integral test tells us that the series  $\sum_{n=2}^{\infty} \frac{1}{n(\ln n)^p}$  converges if and only if the integral  $\int_2^{\infty} \frac{dx}{x(\ln x)^p}$  converges.
- To test the convergence of the integral, we make the substitution  $u = \ln x$ ,  $du = \frac{dx}{x}$ .

$$\int_2^R \frac{dx}{x(\ln x)^p} = \int_{\ln 2}^{\ln R} \frac{du}{u^p}$$

We already know that the integral  $\int_1^{\infty} \frac{du}{u^p}$ , and hence the integral  $\int_2^R \frac{dx}{x(\ln x)^p}$ , converges if and only if  $p > 1$ .

So we conclude that  $\sum_{n=2}^{\infty} \frac{1}{n(\ln n)^p}$  converges if and only if  $p > 1$ .

Example 5.3.8

## 5.4 Comparison Tests

Our next convergence test is the comparison test. It is much like the comparison test for improper integrals (see Theorem 3.7.18) and is true for much the same reasons. The rough idea is quite simple. A sum of larger terms must be bigger than a sum of smaller terms. So if we know the big sum converges, then the small sum must converge too. On the other hand, if we know the small sum diverges, then the big sum must also diverge. Formalising this idea gives the following theorem.

### Theorem 5.4.1 (The Comparison Test).

Let  $N_0$  be a natural number and let  $K > 0$ .

(a) If  $|a_n| \leq Kc_n$  for all  $n \geq N_0$  and  $\sum_{n=0}^{\infty} c_n$  converges, then  $\sum_{n=0}^{\infty} a_n$  converges.

(b) If  $a_n \geq Kd_n \geq 0$  for all  $n \geq N_0$  and  $\sum_{n=0}^{\infty} d_n$  diverges, then  $\sum_{n=0}^{\infty} a_n$  diverges.

“Proof”. We will not prove this theorem here. We’ll just observe that it is very reasonable. That’s why there are quotation marks around “Proof”. For an actual proof see the appendix section A.11.

- (a) If  $\sum_{n=0}^{\infty} c_n$  converges to a finite number and if the terms in  $\sum_{n=0}^{\infty} a_n$  are smaller than the terms in  $\sum_{n=0}^{\infty} c_n$ , then it is no surprise that  $\sum_{n=0}^{\infty} a_n$  converges too.
- (b) If  $\sum_{n=0}^{\infty} d_n$  diverges (i.e. adds up to  $\infty$ ) and if the terms in  $\sum_{n=0}^{\infty} a_n$  are larger than the terms in  $\sum_{n=0}^{\infty} d_n$ , then of course  $\sum_{n=0}^{\infty} a_n$  adds up to  $\infty$ , and so diverges, too.

□

The comparison test for series is also used in much the same way as is the comparison test for improper integrals. Of course, one needs a good series to compare against, and often the series  $\sum n^{-p}$  (from Example 5.3.7), for some  $p > 0$ , turns out to be just what is needed.

Example 5.4.2  $\left( \sum_{n=1}^{\infty} \frac{1}{n^2+2n+3} \right)$

Whether or not any series converges is determined by the behaviour of the summand<sup>27</sup> for very large  $n$ . So the first step in tackling such a problem is to develop some intuition about the behaviour of  $a_n$  when  $n$  is very large.

- *Step 1: Develop intuition.* In this case, when  $n$  is very large<sup>28</sup>  $n^2 \gg 2n \gg 3$  so that  $\frac{1}{n^2+2n+3} \approx \frac{1}{n^2}$ . We already know, from Example 5.3.7, that  $\sum_{n=1}^{\infty} \frac{1}{n^p}$  converges if and only if  $p > 1$ . So  $\sum_{n=1}^{\infty} \frac{1}{n^2}$ , which has  $p = 2$ , converges, and we would expect that  $\sum_{n=1}^{\infty} \frac{1}{n^2+2n+3}$  converges too.
- *Step 2: Verify intuition.* We can use the comparison test to confirm that this is indeed the case. For any  $n \geq 1$ ,  $n^2 + 2n + 3 > n^2$ , so that  $\frac{1}{n^2+2n+3} \leq \frac{1}{n^2}$ . So the comparison test, Theorem 5.4.1, with  $a_n = \frac{1}{n^2+2n+3}$  and  $c_n = \frac{1}{n^2}$ , tells us that  $\sum_{n=1}^{\infty} \frac{1}{n^2+2n+3}$  converges.

27 To understand this consider any series  $\sum_{n=1}^{\infty} a_n$ . We can always cut such a series into two parts — pick some huge number like  $10^6$ . Then

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{10^6} a_n + \sum_{n=10^6+1}^{\infty} a_n$$

The first sum, though it could be humongous, is finite. So the left hand side,  $\sum_{n=1}^{\infty} a_n$ , is a well-defined finite number if and only if  $\sum_{n=10^6+1}^{\infty} a_n$ , is a well-defined finite number. The convergence or divergence of the series is determined by the second sum, which only contains  $a_n$  for “large”  $n$ .

28 The symbol “ $\gg$ ” means “much larger than”. Similarly, the symbol “ $\ll$ ” means “much less than”. Good shorthand symbols can be quite expressive.



Example 5.4.2

Of course the previous example was “rigged” to give an easy application of the comparison test. It is often relatively easy, using arguments like those in Example 5.4.2, to find a “simple” series  $\sum_{n=1}^{\infty} b_n$  with  $b_n$  almost the same as  $a_n$  when  $n$  is large. However it is pretty rare that  $a_n \leq b_n$  for all  $n$ . It is much more common that  $a_n \leq Kb_n$  for some constant  $K$ . This is enough to allow application of the comparison test. Here is an example.

Example 5.4.3  $\left(\sum_{n=1}^{\infty} \frac{n+\cos n}{n^3-1/3}\right)$

As in the previous example, the first step is to develop some intuition about the behaviour of  $a_n$  when  $n$  is very large.

- *Step 1: Develop intuition.* When  $n$  is very large,
  - $n \gg |\cos n|$  so that the numerator  $n + \cos n \approx n$  and
  - $n^3 \gg 1/3$  so that the denominator  $n^3 - 1/3 \approx n^3$ .

So when  $n$  is very large

$$a_n = \frac{n + \cos n}{n^3 - 1/3} \approx \frac{n}{n^3} = \frac{1}{n^2}$$

We already know from Example 5.3.7, with  $p = 2$ , that  $\sum_{n=1}^{\infty} \frac{1}{n^2}$  converges, so we would expect that  $\sum_{n=1}^{\infty} \frac{n+\cos n}{n^3-1/3}$  converges too.

- *Step 2: Verify intuition.* We can use the comparison test to confirm that this is indeed the case. To do so we need to find a constant  $K$  such that  $|a_n| = \frac{|n+\cos n|}{n^3-1/3} = \frac{n+\cos n}{n^3-1/3}$  is smaller than  $\frac{K}{n^2}$  for all  $n$ . A good way<sup>29</sup> to do that is to factor the dominant term (in this case  $n$ ) out of the numerator and also factor the dominant term (in this case  $n^3$ ) out of the denominator.

$$a_n = \frac{n + \cos n}{n^3 - 1/3} = \frac{n}{n^3} \frac{1 + \frac{\cos n}{n}}{1 - \frac{1}{3n^3}} = \frac{1}{n^2} \frac{1 + \frac{\cos n}{n}}{1 - \frac{1}{3n^3}}$$

So now we need to find a constant  $K$  such that  $\frac{1+(\cos n)/n}{1-1/3n^3}$  is smaller than  $K$  for all  $n \geq 1$ .

- First consider the numerator  $1 + (\cos n)\frac{1}{n}$ . For all  $n \geq 1$ 
  - \*  $\frac{1}{n} \leq 1$  and
  - \*  $|\cos n| \leq 1$

So the numerator  $1 + (\cos n)\frac{1}{n}$  is always smaller than  $1 + (1)\frac{1}{1} = 2$ .
- Next consider the denominator  $1 - 1/3n^3$ .
  - \* When  $n \geq 1$ ,  $\frac{1}{3n^3}$  lies between  $\frac{1}{3}$  and 0 so that
  - \*  $1 - \frac{1}{3n^3}$  is between  $\frac{2}{3}$  and 1 and consequently

29 This is very similar to how we computed limits at infinity way way back near the beginning of first-semester calculus.

- \*  $\frac{1}{1-1/3n^3}$  is between  $\frac{3}{2}$  and 1.
- o As the numerator  $1 + (\cos n)\frac{1}{n}$  is always smaller than 2 and  $\frac{1}{1-1/3n^3}$  is always smaller than  $\frac{3}{2}$ , the fraction

$$\frac{1 + \frac{\cos n}{n}}{1 - \frac{1}{3n^3}} \leq 2\left(\frac{3}{2}\right) = 3$$

We now know that

$$|a_n| = \frac{1}{n^2} \frac{1 + 2/n}{1 - 1/3n^3} \leq \frac{3}{n^2}$$

and, since we know  $\sum_{n=1}^{\infty} n^{-2}$  converges, the comparison test tells us that  $\sum_{n=1}^{\infty} \frac{n + \cos n}{n^3 - 1/3}$  converges.

Example 5.4.3

The last example was actually a relatively simple application of the Comparison Theorem — finding a suitable constant  $K$  can be *really* tedious. Fortunately, there is a variant of the comparison test that completely eliminates the need to explicitly find  $K$ .

The idea behind this isn't too complicated. We have already seen that the convergence or divergence of a series depends not on its first few terms, but just on what happens when  $n$  is really large. Consequently, if we can work out how the series terms behave for really big  $n$  then we can work out if the series converges. So instead of comparing the terms of our series for all  $n$ , just compare them when  $n$  is big.

**Theorem 5.4.4 (Limit Comparison Theorem).**

Let  $\sum_{n=1}^{\infty} a_n$  and  $\sum_{n=1}^{\infty} b_n$  be two series with  $b_n > 0$  for all  $n$ . Assume that

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = L$$

exists.

- (a) If  $\sum_{n=1}^{\infty} b_n$  converges, then  $\sum_{n=1}^{\infty} a_n$  converges too.
- (b) If  $L \neq 0$  and  $\sum_{n=1}^{\infty} b_n$  diverges, then  $\sum_{n=1}^{\infty} a_n$  diverges too.

In particular, if  $L \neq 0$ , then  $\sum_{n=1}^{\infty} a_n$  converges if and only if  $\sum_{n=1}^{\infty} b_n$  converges.

*Proof.* (a) Because we are told that  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = L$ , we know that,

- when  $n$  is large,  $\frac{a_n}{b_n}$  is very close to  $L$ , so that  $\left| \frac{a_n}{b_n} \right|$  is very close to  $|L|$ .
- In particular, there is some natural number  $N_0$  so that  $\left| \frac{a_n}{b_n} \right| \leq |L| + 1$ , for all  $n \geq N_0$ , and hence

- $|a_n| \leq Kb_n$  with  $K = |L| + 1$ , for all  $n \geq N_0$ .
- The Comparison Theorem 5.4.1 now implies that  $\sum_{n=1}^{\infty} a_n$  converges.

(b) Let's suppose that  $L > 0$ . (If  $L < 0$ , just replace  $a_n$  with  $-a_n$ .) Because we are told that  $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = L$ , we know that,

- when  $n$  is large,  $\frac{a_n}{b_n}$  is very close to  $L$ .
- In particular, there is some natural number  $N$  so that  $\frac{a_n}{b_n} \geq \frac{L}{2}$ , and hence
- $a_n \geq Kb_n$  with  $K = \frac{L}{2} > 0$ , for all  $n \geq N$ .
- The Comparison Theorem 5.4.1 now implies that  $\sum_{n=1}^{\infty} a_n$  diverges.

□

The next two examples illustrate how much of an improvement the above theorem is over the straight comparison test (though of course, we needed the comparison test to develop the limit comparison test).

Example 5.4.5  $\left( \sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{n^2-2n+3} \right)$

Set  $a_n = \frac{\sqrt{n+1}}{n^2-2n+3}$ . We first try to develop some intuition about the behaviour of  $a_n$  for large  $n$  and then we confirm that our intuition was correct.

- *Step 1: Develop intuition.* When  $n \gg 1$ , the numerator  $\sqrt{n+1} \approx \sqrt{n}$ , and the denominator  $n^2 - 2n + 3 \approx n^2$  so that  $a_n \approx \frac{\sqrt{n}}{n^2} = \frac{1}{n^{3/2}}$  and it looks like our series should converge by Example 5.3.7 with  $p = \frac{3}{2}$ .
- *Step 2: Verify intuition.* To confirm our intuition we set  $b_n = \frac{1}{n^{3/2}}$  and compute the limit

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} \frac{\frac{\sqrt{n+1}}{n^2-2n+3}}{\frac{1}{n^{3/2}}} = \lim_{n \rightarrow \infty} \frac{n^{3/2}\sqrt{n+1}}{n^2-2n+3}$$

Again it is a good idea to factor the dominant term out of the numerator and the dominant term out of the denominator.

$$\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = \lim_{n \rightarrow \infty} \frac{n^2\sqrt{1+1/n}}{n^2(1-2/n+3/n^2)} = \lim_{n \rightarrow \infty} \frac{\sqrt{1+1/n}}{1-2/n+3/n^2} = 1$$

We already know that the series  $\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} \frac{1}{n^{3/2}}$  converges by Example 5.3.7 with  $p = \frac{3}{2}$ . So our series converges by the limit comparison test, Theorem 5.4.4.

Example 5.4.5

Example 5.4.6 ( $\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{n^2-2n+3}$ , again)

We can also try to deal with the series of Example 5.4.5, using the comparison test directly. But that requires us to find  $K$  so that

$$\frac{\sqrt{n+1}}{n^2-2n+3} \leq \frac{K}{n^{3/2}}$$

We might do this by examining the numerator and denominator separately:

- The numerator isn't too bad since for all  $n \geq 1$ :

$$\begin{aligned} n+1 &\leq 2n && \text{and so} \\ \sqrt{n+1} &\leq \sqrt{2n} \end{aligned}$$

- The denominator is quite a bit more tricky, since we need a *lower* bound, rather than an upper bound, and we cannot just write  $|n^2 - 2n + 3| \geq n^2$ , which is false. Instead we have to make a more careful argument. In particular, we'd like to find  $N_0$  and  $K'$  so that  $n^2 - 2n + 3 \geq K'n^2$ , i.e.  $\frac{1}{n^2-2n+3} \leq \frac{1}{K'n^2}$  for all  $n \geq N_0$ . For  $n \geq 4$ , we have  $2n = \frac{1}{2}4n \leq \frac{1}{2}n \cdot n = \frac{1}{2}n^2$ . So for  $n \geq 4$ ,

$$n^2 - 2n + 3 \geq n^2 - \frac{1}{2}n^2 + 3 \geq \frac{1}{2}n^2$$

Putting the numerator and denominator back together we have

$$\frac{\sqrt{n+1}}{n^2-2n+3} \leq \frac{\sqrt{2n}}{n^2/2} = 2\sqrt{2} \frac{1}{n^{3/2}} \quad \text{for all } n \geq 4$$

and the comparison test then tells us that our series converges. It is pretty clear that the approach of Example 5.4.5 was much more straightforward.

Example 5.4.6

Example 5.4.7 (Alternating Harmonic Series)

We've seen by the integral test that the harmonic series,  $\sum_{n=1}^{\infty} \frac{1}{n}$ , diverges. Now we'll consider the alternating harmonic series,

$$\sum_{n=1}^{\infty} \frac{(-1)^n}{n}$$

Since we have negative<sup>30</sup> terms, we can't immediately use a comparison test.

<sup>30</sup> There's a really convenient test for convergence of series that alternate signs every term, the aptly-named Alternating Series Test. You can find more information in Appendix A.12.1. The Alternating Series Test, however, is not on our syllabus.

We'd like to re-write our series. The fine print is that there are only certain circumstances where re-writing a series of this type preserves its convergence. (See Section 5.6 and Appendix A.13.) You can take our word for it that the rearrangement below does not impact the convergence of this particular series.

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{(-1)^n}{n} &= -1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} - \frac{1}{5} + \frac{1}{6} - \dots \\ &= \left(-1 + \frac{1}{2}\right) + \left(-\frac{1}{3} + \frac{1}{4}\right) + \left(-\frac{1}{5} + \frac{1}{6}\right) + \dots\end{aligned}$$

We'll get a common denominator for each bracketed pair.

$$\begin{aligned}&= \left(-\frac{2}{1 \cdot 2} + \frac{1}{1 \cdot 2}\right) + \left(-\frac{4}{3 \cdot 4} + \frac{3}{3 \cdot 4}\right) + \left(-\frac{6}{5 \cdot 6} + \frac{5}{5 \cdot 6}\right) + \dots \\ &= \left(-\frac{1}{1 \cdot 2}\right) + \left(-\frac{1}{3 \cdot 4}\right) + \left(-\frac{1}{5 \cdot 6}\right) - \dots \\ &= \sum_{n=1}^{\infty} \frac{-1}{2n(2n-1)}\end{aligned}$$

We can compare  $\sum \frac{-1}{2n(2n-1)}$  with  $\sum \frac{1}{n^2}$  using the Limit Comparison Test:

$$\begin{aligned}a_n &= \frac{-1}{2n(2n-1)} & b_n &= \frac{1}{n^2} \\ L &= \lim_{n \rightarrow \infty} \frac{a_n}{b_n} \\ &= \lim_{n \rightarrow \infty} \frac{\frac{-1}{2n(2n-1)}}{\frac{1}{n^2}} \\ &= \lim_{n \rightarrow \infty} \frac{-n^2}{2n(2n-1)} \\ &= -\frac{1}{4}\end{aligned}$$

Since  $\sum_{n=1}^{\infty} b_n$  converges, and  $L = -\frac{1}{4}$  exists,  $\sum_{n=1}^{\infty} a_n$  converges as well. That is, the alternating harmonic series converges by the limit comparison test (and by trust in your authors that the rearrangement we started with is, indeed, allowed).

Example 5.4.7

## 5.5▲ The Ratio Test

The idea behind the ratio test comes from a reexamination of the geometric series. Recall that the geometric series

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} ar^n$$

converges when  $|r| < 1$  and diverges otherwise. So the convergence of this series is completely determined by the number  $r$ . This number is just the ratio of successive terms — that is  $r = a_{n+1}/a_n$ .

In general the ratio of successive terms of a series,  $\frac{a_{n+1}}{a_n}$ , is not constant, but depends on  $n$ . However, as we have noted above, the convergence of a series  $\sum a_n$  is determined by the behaviour of its terms when  $n$  is large. In this way, the behaviour of this ratio when  $n$  is small tells us nothing about the convergence of the series, but the limit of the ratio as  $n \rightarrow \infty$  does. This is the basis of the ratio test.

**Theorem 5.5.1 (Ratio Test).**

Let  $N$  be any positive integer and assume that  $a_n \neq 0$  for all  $n \geq N$ .

(a) If  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = L < 1$ , then  $\sum_{n=1}^{\infty} a_n$  converges.

(b) If  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = L > 1$ , or  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = +\infty$ , then  $\sum_{n=1}^{\infty} a_n$  diverges.

**Warning 5.5.2.**

Beware that the ratio test provides absolutely no conclusion about the convergence or divergence of the series  $\sum_{n=1}^{\infty} a_n$  if  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = 1$ .

*Proof.* (a) Pick any number  $R$  obeying  $L < R < 1$ . We are assuming that  $\left| \frac{a_{n+1}}{a_n} \right|$  approaches  $L$  as  $n \rightarrow \infty$ . In particular there must be some natural number  $M$  so that  $\left| \frac{a_{n+1}}{a_n} \right| \leq R$  for all  $n \geq M$ . So  $|a_{n+1}| \leq R|a_n|$  for all  $n \geq M$ . In particular

$$\begin{aligned} |a_{M+1}| &\leq R |a_M| \\ |a_{M+2}| &\leq R |a_{M+1}| \leq R^2 |a_M| \\ |a_{M+3}| &\leq R |a_{M+2}| \leq R^3 |a_M| \\ &\vdots \\ |a_{M+\ell}| &\leq R^\ell |a_M| \end{aligned}$$

for all  $\ell \geq 0$ . The series  $\sum_{\ell=0}^{\infty} R^\ell |a_M|$  is a geometric series with ratio  $R$  smaller than one in magnitude and so converges. Consequently, by the comparison test with  $a_n$  replaced by  $A_\ell = a_{n+\ell}$  and  $c_n$  replaced by  $C_\ell = R^\ell |a_M|$ , the series  $\sum_{\ell=1}^{\infty} a_{M+\ell} = \sum_{n=M+1}^{\infty} a_n$  converges. So the series  $\sum_{n=1}^{\infty} a_n$  converges too.

(b) We are assuming that  $\left| \frac{a_{n+1}}{a_n} \right|$  approaches  $L > 1$  as  $n \rightarrow \infty$ . In particular there must be some natural number  $M > N$  so that  $\left| \frac{a_{n+1}}{a_n} \right| \geq 1$  for all  $n \geq M$ . So  $|a_{n+1}| \geq |a_n|$  for all

$n \geq M$ . That is,  $|a_n|$  increases as  $n$  increases as long as  $n \geq M$ . So  $|a_n| \geq |a_M|$  for all  $n \geq M$  and  $a_n$  cannot converge to zero as  $n \rightarrow \infty$ . So the series diverges by the divergence test.  $\square$

**Example 5.5.3** ( $\sum_{n=0}^{\infty} a n x^{n-1}$ )

Fix any two nonzero real numbers  $a$  and  $x$ . We have already seen in Example 5.2.4 — we have just renamed  $r$  to  $x$  — that the geometric series  $\sum_{n=0}^{\infty} a x^n$  converges when  $|x| < 1$  and diverges when  $|x| \geq 1$ . We are now going to consider a new series, constructed by differentiating<sup>31</sup> each term in the geometric series  $\sum_{n=0}^{\infty} a x^n$ . This new series is

$$\sum_{n=0}^{\infty} a_n \quad \text{with} \quad a_n = a n x^{n-1}$$

Let's apply the ratio test.

$$\left| \frac{a_{n+1}}{a_n} \right| = \left| \frac{a(n+1)x^n}{a n x^{n-1}} \right| = \frac{n+1}{n} |x| = \left(1 + \frac{1}{n}\right) |x| \rightarrow L = |x| \quad \text{as } n \rightarrow \infty$$

The ratio test now tells us that the series  $\sum_{n=0}^{\infty} a n x^{n-1}$  converges if  $|x| < 1$  and diverges if  $|x| > 1$ . It says nothing about the cases  $x = \pm 1$ . But in both of those cases  $a_n = a n (\pm 1)^n$  does not converge to zero as  $n \rightarrow \infty$  and the series diverges by the divergence test.

**Example 5.5.3**

Notice that in the above example, we had to apply another convergence test in addition to the ratio test. This will be commonplace when we reach power series and Taylor series — the ratio test will tell us something like

The series converges for  $|x| < R$  and diverges for  $|x| > R$ .

We generally won't bother with the cases  $x = +R, -R$ .

### 5.5.1 ► Convergence Test List

We now have a handful of convergence tests:

- *Divergence Test*
  - works well when the  $n^{\text{th}}$  term in the series *fails* to converge to zero as  $n$  tends to infinity
- *Integral Test*

<sup>31</sup> We shall see later, in Theorem 6.2.1, that the function  $\sum_{n=0}^{\infty} a n x^{n-1}$  is indeed the derivative of the function  $\sum_{n=0}^{\infty} a x^n$ . Of course, such a statement only makes sense where these series converge — how can you differentiate a divergent series? (This is not an allusion to a popular series of dystopian novels.) Actually, there is quite a bit of interesting and useful mathematics involving divergent series, but it is well beyond the scope of this course.

- works well when, if you substitute  $x$  for  $n$  in the  $n^{\text{th}}$  term you get a function,  $f(x)$ , that you can integrate
- don't forget to check that  $f(x) \geq 0$  and that  $f(x)$  decreases as  $x$  increases
- *Ratio Test*
  - works well when  $\frac{a_{n+1}}{a_n}$  simplifies enough that you can easily compute  $\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = L$
  - this often happens when  $a_n$  contains powers, like  $7^n$ , or factorials, like  $n!$
  - don't forget that  $L = 1$  tells you nothing about the convergence/divergence of the series
- *Comparison Test and Limit Comparison Test*
  - works well when, for very large  $n$ , the  $n^{\text{th}}$  term  $a_n$  is approximately the same as a simpler term  $b_n$  (see Example 5.4.3) and it is easy to determine whether or not  $\sum_{n=1}^{\infty} b_n$  converges
  - don't forget to check that  $b_n \geq 0$
  - A particular comparison series may work with one comparison test but not the other. The Direct Comparison Test is usually only easier when series have fairly simple terms, like  $\sum \frac{1}{2n^2+5}$ . For series with more complicated terms, like  $\sum \frac{2n^2+\sin n}{4n^3+n^2-n}$ , Limit Comparison Test is often the easier choice.

## 5.6▲ Absolute and Conditional Convergence

We have now seen examples of series that converge and of series that diverge. But we haven't really discussed how robust the convergence of series is — that is, can we tweak the coefficients in some way while leaving the convergence unchanged. A good example of this is the series

$$\sum_{n=1}^{\infty} \left(\frac{1}{3}\right)^n$$

This is a simple geometric series and we know it converges. We have also seen, as Example 5.5.3 showed us, that we can multiply or divide the  $n^{\text{th}}$  term by  $n$  and it will still converge. We can even multiply the  $n^{\text{th}}$  term by  $(-1)^n$ , and it will still converge. Pretty robust.

On the other hand, we have explored the Harmonic series and its relatives quite a lot and we know it is much more delicate. While

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

diverges, we also know<sup>32</sup> the following two series converge:

$$\sum_{n=1}^{\infty} \frac{1}{n^{1.00000001}} \qquad \sum_{n=1}^{\infty} (-1)^n \frac{1}{n}.$$

32 The first is a  $p$ -series with  $p > 1$ ; the second is the alternating harmonic series, which we found to converge in Example 5.4.7.



This suggests that the divergence of the Harmonic series is much more delicate. In this section, we discuss one way to characterize this sort of delicate convergence — especially in the presence of changes of sign.

**Definition 5.6.1** (Absolute and conditional convergence).

- (a) A series  $\sum_{n=1}^{\infty} a_n$  is said to converge absolutely if the series  $\sum_{n=1}^{\infty} |a_n|$  converges.
- (b) If  $\sum_{n=1}^{\infty} a_n$  converges but  $\sum_{n=1}^{\infty} |a_n|$  diverges we say that  $\sum_{n=1}^{\infty} a_n$  is conditionally convergent.

If you consider these definitions for a moment, it should be clear that absolute convergence is a stronger condition than just simple convergence. All the terms in  $\sum_n |a_n|$  are forced to be positive (by the absolute value signs), so that  $\sum_n |a_n|$  must be bigger than  $\sum_n a_n$  — making it easier for  $\sum_n |a_n|$  to diverge. This is formalised by the following theorem, which is an immediate consequence of the comparison test, Theorem 5.4.1.a, with  $c_n = |a_n|$ .

**Theorem 5.6.2** (Absolute convergence implies convergence).

If the series  $\sum_{n=1}^{\infty} |a_n|$  converges then the series  $\sum_{n=1}^{\infty} a_n$  also converges. That is, absolute convergence implies convergence.

Recall that some of our convergence tests (for example, the integral test) may only be applied to series with positive terms. Theorem 5.6.2 opens up the possibility of applying “positive only” convergence tests to series whose terms are not all positive, by checking for “absolute convergence” rather than for plain “convergence”.

**Example 5.6.3**  $\left(\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n^2}\right)$

Because the series  $\sum_{n=1}^{\infty} |(-1)^{n-1} \frac{1}{n^2}| = \sum_{n=1}^{\infty} \frac{1}{n^2}$  of Example 5.3.7 converges (by the integral test), the series  $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n^2}$  converges absolutely, and hence converges.

Example 5.6.3

**Example 5.6.4** (random signs)

Imagine flipping a coin infinitely many times. Set  $\sigma_n = +1$  if the  $n^{\text{th}}$  flip comes up heads

and  $\sigma_n = -1$  if the  $n^{\text{th}}$  flip comes up tails. We know that the series  $\sum_{n=1}^{\infty} |(-1)^{\sigma_n} \frac{1}{n^2}| = \sum_{n=1}^{\infty} \frac{1}{n^2}$  converges. So  $\sum_{n=1}^{\infty} (-1)^{\sigma_n} \frac{1}{n^2}$  converges absolutely, and hence converges.

Example 5.6.4

With series that converge conditionally, arithmetic can get a little tricky. For some interesting examples of this trickiness, see Appendix [A.13](#).

# POWER SERIES

Let's return to the simple geometric series

$$\sum_{n=0}^{\infty} x^n$$

where  $x$  is some real number. As we have seen (back in Example 5.2.4), for  $|x| < 1$  this series converges to a limit, that varies with  $x$ , while for  $|x| \geq 1$  the series diverges. Consequently we can consider this series to be a function of  $x$

$$f(x) = \sum_{n=0}^{\infty} x^n \quad \text{on the domain } |x| < 1.$$

Furthermore (also from Example 5.2.4) we know what the function is.

$$f(x) = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}.$$

Hence we can consider the series  $\sum_{n=0}^{\infty} x^n$  as a new way of representing the function  $\frac{1}{1-x}$  when  $|x| < 1$ . This series is an example of a power series.

Of course, representing a function as simple as  $\frac{1}{1-x}$  by a series doesn't seem like it is going to make life easier. However the idea of representing a function by a series turns out to be extremely helpful. Power series turn out to be very robust mathematical objects and interact very nicely with not only standard arithmetic operations, but also with differentiation and integration (see Theorem 6.2.1). This means, for example, that

$$\begin{aligned} \frac{d}{dx} \left\{ \frac{1}{1-x} \right\} &= \frac{d}{dx} \sum_{n=0}^{\infty} x^n && \text{provided } |x| < 1 \\ &= \sum_{n=0}^{\infty} \frac{d}{dx} x^n && \text{just differentiate term by term} \\ &= \sum_{n=0}^{\infty} nx^{n-1} \end{aligned}$$

and in a very similar way

$$\begin{aligned} \int \frac{1}{1-x} dx &= \int \sum_{n=0}^{\infty} x^n dx && \text{provided } |x| < 1 \\ &= \sum_{n=0}^{\infty} \int x^n dx && \text{just integrate term by term} \\ &= C + \sum_{n=0}^{\infty} \frac{1}{n+1} x^{n+1} \end{aligned}$$

We are hiding some mathematics under the word “just” in the above, but you can see that once we have a power series representation of a function, differentiation and integration become very straightforward.

So we should set as our goal for this chapter, the development of machinery to define and understand power series. This will allow us to answer questions<sup>1</sup> like

$$\text{Is } e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} ?$$

## 6.1▲ Radius of Convergence

Our starting point (now that we have equipped ourselves with basic ideas about series), is the definition of power series.

### Definition 6.1.1.

A series of the form

$$A_0 + A_1(x-c) + A_2(x-c)^2 + A_3(x-c)^3 + \cdots = \sum_{n=0}^{\infty} A_n(x-c)^n$$

is called a *power series in*  $(x-c)$  or a *power series centered on*  $c$ . The numbers  $A_n$  are called the coefficients of the power series.

One often considers power series centered on  $c = 0$  and then the series reduces to

$$A_0 + A_1x + A_2x^2 + A_3x^3 + \cdots = \sum_{n=0}^{\infty} A_nx^n$$

For example  $\sum_{n=0}^{\infty} \frac{x^n}{n!}$  is the power series with  $c = 0$  and  $A_n = \frac{1}{n!}$ . Typically, as in that case, the coefficients  $A_n$  are given fixed numbers, but the “ $x$ ” is to be thought of as a variable. Thus each power series is really a whole family of series — a different series for each value of  $x$ .

<sup>1</sup> Recall that  $n! = 1 \times 2 \times 3 \times \cdots \times n$  is called “ $n$  factorial”. By convention  $0! = 1$ .

One possible value of  $x$  is  $x = c$  and then the series reduces<sup>2</sup> to

$$\begin{aligned} \sum_{n=0}^{\infty} A_n(x-c)^n \Big|_{x=c} &= \sum_{n=0}^{\infty} A_n(c-c)^n \\ &= \underbrace{A_0}_{n=0} + \underbrace{0}_{n=1} + \underbrace{0}_{n=2} + \underbrace{0}_{n=3} + \dots \end{aligned}$$

and so simply converges to  $A_0$ .

We now know that a power series converges when  $x = c$ . We can now use our convergence tests to determine for what other values of  $x$  the series converges. Perhaps most straightforward is the ratio test. The  $n^{\text{th}}$  term in the series  $\sum_{n=0}^{\infty} A_n(x-c)^n$  is  $a_n = A_n(x-c)^n$ . To apply the ratio test we need to compute the limit

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| &= \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}(x-c)^{n+1}}{A_n(x-c)^n} \right| \\ &= \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| \cdot |x-c| \\ &= |x-c| \cdot \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right|. \end{aligned}$$

When we do so there are several possible outcomes.

- If the limit of ratios exists and is non-zero

$$\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| = A \neq 0,$$

then the ratio test says that the series  $\sum_{n=0}^{\infty} A_n(x-c)^n$

- converges when  $A \cdot |x-c| < 1$ , i.e. when  $|x-c| < 1/A$ , and
- diverges when  $A \cdot |x-c| > 1$ , i.e. when  $|x-c| > 1/A$ .

Because of this, when the limit exists, the quantity

**Equation 6.1.2.**

$$R = \frac{1}{A} = \left[ \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| \right]^{-1}$$

is called the *radius of convergence* of the series<sup>3</sup>.

2 By convention, when the term  $(x-c)^0$  appears in a power series, it has value 1 for all values of  $x$ , even  $x = c$ .

3 The use of the word “radius” might seem a little odd here, since we are really describing the interval in the real line where the series converges. However, when one starts to consider power series over complex numbers, the radius of convergence does describe a circle inside the complex plane and so “radius” is a more natural descriptor.

- If the limit of ratios exists and is zero

$$\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| = 0$$

then  $\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| |x - c| = 0$  for every  $x$  and the ratio test tells us that the series  $\sum_{n=0}^{\infty} A_n(x - c)^n$  converges for every number  $x$ . In this case we say that the series has an infinite radius of convergence.

- If the limit of ratios diverges to  $+\infty$

$$\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| = +\infty$$

then  $\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| |x - c| = +\infty$  for every  $x \neq c$ . The ratio test then tells us that the series  $\sum_{n=0}^{\infty} A_n(x - c)^n$  diverges for every number  $x \neq c$ . As we have seen above, when  $x = c$ , the series reduces to  $A_0 + 0 + 0 + 0 + 0 + \dots$ , which of course converges. In this case we say that the series has radius of convergence zero.

- If  $\left| \frac{A_{n+1}}{A_n} \right|$  does not approach a limit as  $n \rightarrow \infty$ , then we learn nothing from the ratio test and we must use other tools to understand the convergence of the series.

All of these possibilities do happen. We give an example of each below. But first, the concept of “radius of convergence” is important enough to warrant a formal definition.

### Definition 6.1.3.

- (a) Let  $0 < R < \infty$ . If  $\sum_{n=0}^{\infty} A_n(x - c)^n$  converges for  $|x - c| < R$ , and diverges for  $|x - c| > R$ , then we say that the series has radius of convergence  $R$ .
- (b) If  $\sum_{n=0}^{\infty} A_n(x - c)^n$  converges for every number  $x$ , we say that the series has an infinite radius of convergence.
- (c) If  $\sum_{n=0}^{\infty} A_n(x - c)^n$  diverges for every  $x \neq c$ , we say that the series has radius of convergence zero.

### Example 6.1.4 (Finite nonzero radius of convergence)

We already know that, if  $a \neq 0$ , the geometric series  $\sum_{n=0}^{\infty} ax^n$  converges when  $|x| < 1$  and diverges when  $|x| \geq 1$ . So, in the terminology of Definition 6.1.3, the geometric series has radius of convergence  $R = 1$ . As a consistency check, we can also compute  $R$  using (6.1.2).

The series  $\sum_{n=0}^{\infty} ax^n$  has  $A_n = a$ . So

$$R = \left[ \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| \right]^{-1} = \left[ \lim_{n \rightarrow \infty} 1 \right]^{-1} = 1$$

as expected.

Example 6.1.4

Example 6.1.5 (Radius of convergence =  $+\infty$ )

The series  $\sum_{n=0}^{\infty} \frac{x^n}{n!}$  has  $A_n = \frac{1}{n!}$ . So

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| &= \lim_{n \rightarrow \infty} \frac{1/(n+1)!}{1/n!} = \lim_{n \rightarrow \infty} \frac{n!}{(n+1)!} = \lim_{n \rightarrow \infty} \frac{1 \times 2 \times 3 \times \cdots \times n}{1 \times 2 \times 3 \times \cdots \times n \times (n+1)} \\ &= \lim_{n \rightarrow \infty} \frac{1}{n+1} \\ &= 0 \end{aligned}$$

and  $\sum_{n=0}^{\infty} \frac{x^n}{n!}$  has radius of convergence  $\infty$ . It converges for every  $x$ .

Example 6.1.5

Example 6.1.6 (Radius of convergence = 0)

The series  $\sum_{n=0}^{\infty} n!x^n$  has  $A_n = n!$ . So

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right| &= \lim_{n \rightarrow \infty} \frac{(n+1)!}{n!} = \lim_{n \rightarrow \infty} \frac{1 \times 2 \times 3 \times 4 \times \cdots \times n \times (n+1)}{1 \times 2 \times 3 \times 4 \times \cdots \times n} \\ &= \lim_{n \rightarrow \infty} (n+1) \\ &= +\infty \end{aligned}$$

and  $\sum_{n=0}^{\infty} n!x^n$  has radius of convergence zero<sup>4</sup>. It converges only for  $x = 0$ , where it takes the value  $0! = 1$ .

Example 6.1.6

Example 6.1.7

Comparing the series

$$1 + 2x + x^2 + 2x^3 + x^4 + 2x^5 + \cdots$$

to

$$\sum_{n=1}^{\infty} A_n x^n = A_0 + A_1 x + A_2 x^2 + A_3 x^3 + A_4 x^4 + A_5 x^5 + \cdots$$

4 Because of this, it might seem that such a series is fairly pointless. However there are all sorts of mathematical games that can be played with them without worrying about their convergence. Such “formal” power series can still impart useful information and the interested reader is invited to look up “generating functions” with their preferred search engine.

we see that

$$A_0 = 1 \quad A_1 = 2 \quad A_2 = 1 \quad A_3 = 2 \quad A_4 = 1 \quad A_5 = 2 \quad \dots$$

so that

$$\frac{A_1}{A_0} = 2 \quad \frac{A_2}{A_1} = \frac{1}{2} \quad \frac{A_3}{A_2} = 2 \quad \frac{A_4}{A_3} = \frac{1}{2} \quad \frac{A_5}{A_4} = 2 \quad \dots$$

and  $\frac{A_{n+1}}{A_n}$  does not converge as  $n \rightarrow \infty$ . Since the limit of the ratios does not exist, we cannot tell anything from the ratio test. Nonetheless, we can still figure out for which  $x$ 's our power series converges.

- Because every coefficient  $A_n$  is either 1 or 2, the  $n^{\text{th}}$  term in our series obeys

$$|A_n x^n| \leq 2|x|^n$$

and so is smaller than the  $n^{\text{th}}$  term in the geometric series  $\sum_{n=0}^{\infty} 2|x|^n$ . This geometric series converges if  $|x| < 1$ . So, by the comparison test, our series converges for  $|x| < 1$  too.

- Since every  $A_n$  is at least one, the  $n^{\text{th}}$  term in our series obeys

$$|A_n x^n| \geq |x|^n$$

If  $|x| \geq 1$ , this  $a_n = A_n x^n$  cannot converge to zero as  $n \rightarrow \infty$ , and our series diverges by the divergence test.

In conclusion, our series converges if and only if  $|x| < 1$ , and so has radius of convergence 1.

Example 6.1.7

Example 6.1.8

Lets construct a series from the digits of  $\pi$ . Now to avoid dividing by zero, let us set

$$A_n = 1 + \text{the } n^{\text{th}} \text{ digit of } \pi$$

Since  $\pi = 3.141591\dots$

$$A_0 = 4 \quad A_1 = 2 \quad A_2 = 5 \quad A_3 = 2 \quad A_4 = 6 \quad A_5 = 10 \quad A_6 = 2 \quad \dots$$

Consequently every  $A_n$  is an integer between 1 and 10 and gives us the series

$$\sum_{n=0}^{\infty} A_n x^n = 4 + 2x + 5x^2 + 2x^3 + 6x^4 + 10x^5 + \dots$$

The number  $\pi$  is irrational and consequently the ratio  $\frac{A_{n+1}}{A_n}$  cannot have a limit as  $n \rightarrow \infty$ . If you do not understand why this is the case then don't worry too much about it<sup>5</sup>. As in

5 This is a little beyond the scope of the course. Roughly speaking, think about what would happen if the limit of the ratios did exist. If the limit were smaller than 1, then it would tell you that the terms of our series must be getting smaller and smaller and smaller — which is impossible because they are all integers between 1 and 10. Similarly if the limit existed and were bigger than 1 then the terms of the series would have to get bigger and bigger and bigger — also impossible. Hence if the ratio exists then it must be equal to 1 — but in that case because the terms are integers, they would have to be all equal when  $n$  became big enough. But that means that the expansion of  $\pi$  would be eventually periodic — something that only rational numbers do.



the last example, the limit of the ratios does not exist and we cannot tell anything from the ratio test. But we can still figure out for which  $x$ 's it converges.

- Because every coefficient  $A_n$  is no bigger (in magnitude) than 10, the  $n^{\text{th}}$  term in our series obeys

$$|A_n x^n| \leq 10|x|^n$$

and so is smaller than the  $n^{\text{th}}$  term in the geometric series  $\sum_{n=0}^{\infty} 10|x|^n$ . This geometric series converges if  $|x| < 1$ . So, by the comparison test, our series converges for  $|x| < 1$  too.

- Since every  $A_n$  is at least one, the  $n^{\text{th}}$  term in our series obeys

$$|A_n x^n| \geq |x|^n$$

If  $|x| \geq 1$ , this  $a_n = A_n x^n$  cannot converge to zero as  $n \rightarrow \infty$ , and our series diverges by the divergence test.

In conclusion, our series converges if and only if  $|x| < 1$ , and so has radius of convergence 1.

Example 6.1.8

Though we won't prove it, it is true that every power series has a radius of convergence, whether or not the limit  $\lim_{n \rightarrow \infty} \left| \frac{A_{n+1}}{A_n} \right|$  exists.

### Theorem 6.1.9.

Let  $\sum_{n=0}^{\infty} A_n (x - c)^n$  be a power series. Then one of the following alternatives must hold.

- The power series converges for every number  $x$ . In this case we say that the radius of convergence is  $\infty$ .
- There is a number  $0 < R < \infty$  such that the series converges for  $|x - c| < R$  and diverges for  $|x - c| > R$ . Then  $R$  is called the radius of convergence.
- The series converges for  $x = c$  and diverges for all  $x \neq c$ . In this case, we say that the radius of convergence is 0.

**Definition 6.1.10.**

Consider the power series

$$\sum_{n=0}^{\infty} A_n(x - c)^n.$$

The set of real  $x$ -values for which it converges is called the interval of convergence of the series.

Suppose that the power series  $\sum_{n=0}^{\infty} A_n(x - c)^n$  has radius of convergence  $R$ . Then from Theorem 6.1.9, we have that

- if  $R = \infty$ , then its interval of convergence is  $-\infty < x < \infty$ , which is also denoted  $(-\infty, \infty)$ , and
- if  $R = 0$ , then its interval of convergence is just the point  $x = c$ , and
- if  $0 < R < \infty$ , then we know that the series converges for any  $x$  which obeys

$$\begin{aligned} |x - c| < R \quad \text{or equivalently} \quad -R < x - c < R \\ \text{or equivalently} \quad c - R < x < c + R \end{aligned}$$

But we do not (yet) know whether or not the series converges at the two end points of that interval. We do know, however, that its interval of convergence must be one of

- $c - R < x < c + R$ , which is also denoted  $(c - R, c + R)$ , or
- $c - R \leq x < c + R$ , which is also denoted  $[c - R, c + R)$ , or
- $c - R < x \leq c + R$ , which is also denoted  $(c - R, c + R]$ , or
- $c - R \leq x \leq c + R$ , which is also denoted  $[c - R, c + R]$ .

To reiterate — while the radius convergence,  $R$  with  $0 < R < \infty$ , tells us that the series converges for  $|x - c| < R$  and diverges for  $|x - c| > R$ , it does not (by itself) tell us whether or not the series converges when  $|x - c| = R$ , i.e. when  $x = c \pm R$ . We will not generally concern ourselves with these final details. (Determining the endpoints of the interval of convergence often goes smoothest with the Alternating Series Test, which is available for your interest in Appendix A.12 but is not a part of our syllabus.)

**Example 6.1.11**

We are told that a certain power series with centre  $c = 3$ , converges at  $x = 4$  and diverges at  $x = 1$ . What else can we say about the convergence or divergence of the series for other values of  $x$ ?

We are told that the series is centred at 3, so its terms are all powers of  $(x - 3)$  and it is of the form

$$\sum_{n \geq 0} A_n(x - 3)^n.$$

A good way to summarise the convergence data we are given is with a figure like the one below. Green dots mark the values of  $x$  where the series is known to converge. (Recall that every power series converges at its centre.) The red dot marks the value of  $x$  where the series is known to diverge. The centre is at  $x = 3$ .



Can we say more about the convergence and/or divergence of the series for other values of  $x$ ? Yes!

Let us think about the radius of convergence,  $R$ , of the series. We know that it must exist and the information we have been given allows us to bound  $R$ . Recall that

- the series converges at  $x$  provided that  $|x - 3| < R$  and
- the series diverges at  $x$  if  $|x - 3| > R$ .

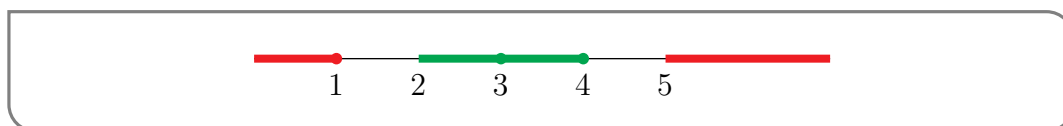
We have been told that

- the series converges when  $x = 4$ , which tells us that
  - $x = 4$  cannot obey  $|x - 3| > R$  so
  - $x = 4$  must obey  $|x - 3| \leq R$ , i.e.  $|4 - 3| \leq R$ , i.e.  $R \geq 1$
- the series diverges when  $x = 1$  so we also know that
  - $x = 1$  cannot obey  $|x - 3| < R$  so
  - $x = 1$  must obey  $|x - 3| \geq R$ , i.e.  $|1 - 3| \geq R$ , i.e.  $R \leq 2$

We still don't know  $R$  exactly. But we do know that  $1 \leq R \leq 2$ . Consequently,

- since 1 is the smallest that  $R$  could be, the series certainly converges at  $x$  if  $|x - 3| < 1$ , i.e. if  $2 < x < 4$  and
- since 2 is the largest that  $R$  could be, the series certainly diverges at  $x$  if  $|x - 3| > 2$ , i.e. if  $x > 5$  or if  $x < 1$ .

The following figure provides a resume of all of this convergence data — there is convergence at green  $x$ 's and divergence at red  $x$ 's.



Notice that from the data given we cannot say anything about the convergence or divergence of the series on the intervals  $(1, 2]$  and  $(4, 5]$ .

One lesson that we can derive from this example is that,

- if a series has centre  $c$  and converges at  $a$ ,
- then it also converges at all points between  $c$  and  $a$ , as well as at all points of distance strictly less than  $|a - c|$  from  $c$  on the other side of  $c$  from  $a$ .



Example 6.1.11

---

## 6.2▲ Working With Power Series

Just as we have done previously with limits, differentiation and integration, we can construct power series representations of more complicated functions by using those of simpler functions. Here is a theorem that helps us to do so.

**Theorem 6.2.1** (Operations on Power Series).

Assume that the functions  $f(x)$  and  $g(x)$  are given by the power series

$$f(x) = \sum_{n=0}^{\infty} A_n(x-c)^n \quad g(x) = \sum_{n=0}^{\infty} B_n(x-c)^n$$

for all  $x$  obeying  $|x-c| < R$ . In particular, we are assuming that both power series have radius of convergence at least  $R$ . Also let  $K$  be a constant. Then

$$f(x) + g(x) = \sum_{n=0}^{\infty} [A_n + B_n] (x-c)^n$$

$$Kf(x) = \sum_{n=0}^{\infty} K A_n (x-c)^n$$

$$(x-c)^N f(x) = \sum_{n=0}^{\infty} A_n (x-c)^{n+N} \quad \text{for any integer } N \geq 1$$

$$= \sum_{k=N}^{\infty} A_{k-N} (x-c)^k \quad \text{where } k = n + N$$

$$f'(x) = \sum_{n=0}^{\infty} A_n n (x-c)^{n-1} = \sum_{n=1}^{\infty} A_n n (x-c)^{n-1}$$

$$\int_c^x f(t) dt = \sum_{n=0}^{\infty} A_n \frac{(x-c)^{n+1}}{n+1}$$

$$\int f(x) dx = \left[ \sum_{n=0}^{\infty} A_n \frac{(x-c)^{n+1}}{n+1} \right] + C \quad \text{with } C \text{ an arbitrary constant}$$

for all  $x$  obeying  $|x-c| < R$ .

In particular the radius of convergence of each of the six power series on the right hand sides is at least  $R$ . In fact, if  $R$  is the radius of convergence of  $\sum_{n=0}^{\infty} A_n(x-c)^n$ , then  $R$  is also the radius of convergence of all of the above right hand sides, with the possible exceptions of  $\sum_{n=0}^{\infty} [A_n + B_n] (x-c)^n$  and  $\sum_{n=0}^{\infty} K A_n (x-c)^n$  when  $K = 0$ .

**Example 6.2.2**

The last statement of Theorem 6.2.1 might seem a little odd, but consider the following two power series centred at 0:

$$\sum_{n=0}^{\infty} 2^n x^n \quad \text{and} \quad \sum_{n=0}^{\infty} (1 - 2^n) x^n.$$

The ratio test tells us that they both have radius of convergence  $R = \frac{1}{2}$ . However their sum is

$$\sum_{n=0}^{\infty} 2^n x^n + \sum_{n=0}^{\infty} (1 - 2^n) x^n = \sum_{n=0}^{\infty} x^n$$

which has the larger radius of convergence 1.

A more extreme example of the same phenomenon is supplied by the two series

$$\sum_{n=0}^{\infty} 2^n x^n \text{ and } \sum_{n=0}^{\infty} (-2^n) x^n.$$

They are both geometric series with radius of convergence  $R = \frac{1}{2}$ . But their sum is

$$\sum_{n=0}^{\infty} 2^n x^n + \sum_{n=0}^{\infty} (-2^n) x^n = \sum_{n=0}^{\infty} (0) x^n$$

which has radius of convergence  $+\infty$ .

Example 6.2.2

We'll now use this theorem to build power series representations for a bunch of functions out of the one simple power series representation that we know — the geometric series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \quad \text{for all } |x| < 1$$

Example 6.2.3  $\left(\frac{1}{1-x^2}\right)$

Find a power series representation for  $\frac{1}{1-x^2}$ .

*Solution.* The secret to finding power series representations for a good many functions is to manipulate them into a form in which  $\frac{1}{1-y}$  appears and use the geometric series representation  $\frac{1}{1-y} = \sum_{n=0}^{\infty} y^n$ . We have deliberately renamed the variable to  $y$  here — it does not have to be  $x$ . We can use that strategy to find a power series expansion for  $\frac{1}{1-x^2}$  — we just have to recognize that  $\frac{1}{1-x^2}$  is the same as  $\frac{1}{1-y}$  if we set  $y$  to  $x^2$ .

$$\begin{aligned} \frac{1}{1-x^2} &= \frac{1}{1-y} \Big|_{y=x^2} = \left[ \sum_{n=0}^{\infty} y^n \right]_{y=x^2} \quad \text{if } |y| < 1, \text{ i.e. } |x| < 1 \\ &= \sum_{n=0}^{\infty} (x^2)^n = \sum_{n=0}^{\infty} x^{2n} \\ &= 1 + x^2 + x^4 + x^6 + \dots \end{aligned}$$

This is a perfectly good power series. There is nothing wrong with the power of  $x$  being  $2n$ . (This just means that the coefficients of all odd powers of  $x$  are zero.) In fact, you should

try to always write power series in forms that are as easy to understand as possible. The geometric series that we used at the end of the first line converges for

$$|y| < 1 \iff |x^2| < 1 \iff |x| < 1$$

So our power series has radius of convergence 1 and interval of convergence  $-1 < x < 1$ .

Example 6.2.3

Example 6.2.4  $\left(\frac{x}{2+x^2}\right)$

Find a power series representation for  $\frac{x}{2+x^2}$ .

*Solution.* This example is just a more algebraically involved variant of the last one. Again, the strategy is to manipulate  $\frac{x}{2+x^2}$  into a form in which  $\frac{1}{1-y}$  appears.

$$\begin{aligned} \frac{x}{2+x^2} &= \frac{x}{2} \frac{1}{1+x^2/2} = \frac{x}{2} \frac{1}{1-(-x^2/2)} \quad \text{set } -\frac{x^2}{2} = y \\ &= \frac{x}{2} \frac{1}{1-y} \Big|_{y=-x^2/2} = \frac{x}{2} \left[ \sum_{n=0}^{\infty} y^n \right]_{y=-x^2/2} \quad \text{if } |y| < 1 \\ &= \frac{x}{2} \sum_{n=0}^{\infty} \left(-\frac{x^2}{2}\right)^n = \frac{x}{2} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n} x^{2n} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2^{n+1}} x^{2n+1} \quad \text{by Theorem 6.2.1, twice} \\ &= \frac{x}{2} - \frac{x^3}{4} + \frac{x^5}{8} - \frac{x^7}{16} + \dots \end{aligned}$$

The geometric series that we used in the second line converges when

$$|y| < 1 \iff |-x^2/2| < 1 \iff |x|^2 < 2 \iff |x| < \sqrt{2}$$

So the given power series has radius of convergence  $\sqrt{2}$  and interval of convergence  $-\sqrt{2} < x < \sqrt{2}$ .

Example 6.2.4

Example 6.2.5 (Nonzero centre)

Find a power series representation for  $\frac{1}{5-x}$  with centre 3.

*Solution.* The new wrinkle in this example is the requirement that the centre be 3. That the centre is to be 3 means that we need a power series in powers of  $x - c$ , with  $c = 3$ . So we are looking for a power series of the form  $\sum_{n=0}^{\infty} A_n(x-3)^n$ . The easy way to find such a series is to force an  $x-3$  to appear by adding and subtracting a 3.

$$\frac{1}{5-x} = \frac{1}{5-(x-3)-3} = \frac{1}{2-(x-3)}$$

Now we continue, as in the last example, by manipulating  $\frac{1}{2-(x-3)}$  into a form in which  $\frac{1}{1-y}$  appears.

$$\begin{aligned} \frac{1}{5-x} &= \frac{1}{2-(x-3)} = \frac{1}{2} \frac{1}{1-\frac{x-3}{2}} && \text{set } \frac{x-3}{2} = y \\ &= \frac{1}{2} \frac{1}{1-y} \Big|_{y=\frac{x-3}{2}} = \frac{1}{2} \left[ \sum_{n=0}^{\infty} y^n \right]_{y=\frac{x-3}{2}} && \text{if } |y| < 1 \\ &= \frac{1}{2} \sum_{n=0}^{\infty} \left( \frac{x-3}{2} \right)^n = \sum_{n=0}^{\infty} \frac{(x-3)^n}{2^{n+1}} \\ &= \frac{x-3}{2} + \frac{(x-3)^2}{4} + \frac{(x-3)^3}{8} + \dots \end{aligned}$$

The geometric series that we used in the second line converges when

$$|y| < 1 \iff \left| \frac{x-3}{2} \right| < 1 \iff |x-3| < 2 \iff -2 < x-3 < 2 \iff 1 < x < 5$$

So the power series has radius of convergence 2 and interval of convergence  $1 < x < 5$ .

Example 6.2.5

In the previous two examples, to construct a new series from an existing series, we replaced  $x$  by a simple function. The following theorem gives us some more (but certainly not all) commonly used substitutions.

**Theorem 6.2.6** (Substituting in a Power Series).

Assume that the function  $f(x)$  is given by the power series

$$f(x) = \sum_{n=0}^{\infty} A_n x^n$$

for all  $x$  in the interval  $I$ . Also let  $K$  and  $k$  be real constants. Then

$$f(Kx^k) = \sum_{n=0}^{\infty} A_n K^n x^{kn}$$

whenever  $Kx^k$  is in  $I$ . In particular, if  $\sum_{n=0}^{\infty} A_n x^n$  has radius of convergence  $R$ ,  $K$  is nonzero and  $k$  is a natural number, then  $\sum_{n=0}^{\infty} A_n K^n x^{kn}$  has radius of convergence  $\sqrt[k]{R/|K|}$ .

Example 6.2.7  $\left( \frac{1}{(1-x)^2} \right)$

Find a power series representation for  $\frac{1}{(1-x)^2}$ .



*Solution.* Once again the trick is to express  $\frac{1}{(1-x)^2}$  in terms of  $\frac{1}{1-x}$ . Notice that

$$\begin{aligned}\frac{1}{(1-x)^2} &= \frac{d}{dx} \left\{ \frac{1}{1-x} \right\} \\ &= \frac{d}{dx} \left\{ \sum_{n=0}^{\infty} x^n \right\} \\ &= \sum_{n=1}^{\infty} nx^{n-1} \quad \text{by Theorem 6.2.1}\end{aligned}$$

Note that the  $n = 0$  term has disappeared because, for  $n = 0$ ,

$$\frac{d}{dx}x^n = \frac{d}{dx}x^0 = \frac{d}{dx}1 = 0$$

Also note that the radius of convergence of this series is one. We can see this via Theorem 6.2.1. That theorem tells us that the radius of convergence of a power series is not changed by differentiation — and since  $\sum_{n=0}^{\infty} x^n$  has radius of convergence one, so too does its derivative.

Without much more work we can determine the interval of convergence by testing at  $x = \pm 1$ . When  $x = \pm 1$  the terms of the series do not go to zero as  $n \rightarrow \infty$  and so, by the divergence test, the series does not converge there. Hence the interval of convergence for the series is  $-1 < x < 1$ .

Example 6.2.7

Notice that, in this last example, we differentiated a known series to get to our answer. As per Theorem 6.2.1, the radius of convergence didn't change. In addition, in this particular example, the interval of convergence didn't change. This is not always the case. Differentiation of some series causes the interval of convergence to shrink. In particular the differentiated series may no longer be convergent at the end points of the interval<sup>6</sup>. Similarly, when we integrate a power series the radius of convergence is unchanged, but the interval of convergence may expand to include one or both ends, as illustrated by the next example.

Example 6.2.8 ( $\ln(1+x)$ )

Find a power series representation for  $\ln(1+x)$ .

<sup>6</sup> Consider the power series  $\sum_{n=1}^{\infty} \frac{x^n}{n}$ . We know that its interval of convergence is  $-1 \leq x < 1$ . (Indeed see the next example.) When we differentiate the series we get the geometric series  $\sum_{n=0}^{\infty} x^n$  which has interval of convergence  $-1 < x < 1$ .

*Solution.* Recall that  $\frac{d}{dx} \ln(1+x) = \frac{1}{1+x}$  so that  $\ln(1+t)$  is an antiderivative of  $\frac{1}{1+t}$  and

$$\begin{aligned} \ln(1+x) &= \int_0^x \frac{dt}{1+t} = \int_0^x \left[ \sum_{n=0}^{\infty} (-t)^n \right] dt \\ &= \sum_{n=0}^{\infty} \int_0^x (-t)^n dt \quad \text{by Theorem 6.2.1} \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} \\ &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \end{aligned}$$

Theorem 6.2.1 guarantees that the radius of convergence is exactly one (the radius of convergence of the geometric series  $\sum_{n=0}^{\infty} (-t)^n$ ) and that

$$\ln(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} \quad \text{for all } -1 < x < 1$$

In general, we won't worry about the endpoints of the interval of convergence. So, in general, we wouldn't bother testing  $x = 1$  and  $x = -1$ . However, in this instance, both examples are pretty accessible. We include them below for interest.

When  $x = -1$  our series reduces to  $\sum_{n=0}^{\infty} \frac{-1}{n+1}$ , which is (minus) the harmonic series and so diverges. That's no surprise:  $\ln(1+(-1)) = \ln 0$  is undefined, with  $\lim_{x \rightarrow 0^+} \ln x = -\infty$ .

When  $x = 1$ , we get the alternating harmonic series, which converges. (It is possible to prove by continuity, though we won't do so here, that the sum is  $\ln 2$ .)

So the interval of convergence is  $-1 < x \leq 1$ .

Example 6.2.8

Example 6.2.9 (arctan  $x$ )

Find a power series representation for arctan  $x$ .

*Solution.* Recall that  $\frac{d}{dx} \arctan x = \frac{1}{1+x^2}$  so that  $\arctan t$  is an antiderivative of  $\frac{1}{1+t^2}$  and

$$\begin{aligned} \arctan x &= \int_0^x \frac{dt}{1+t^2} = \int_0^x \left[ \sum_{n=0}^{\infty} (-t^2)^n \right] dt = \sum_{n=0}^{\infty} \int_0^x (-1)^n t^{2n} dt \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots \end{aligned}$$

Theorem 6.2.1 guarantees that the radius of convergence is exactly one (the radius of convergence of the geometric series  $\sum_{n=0}^{\infty} (-t^2)^n$ ) and that

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} \quad \text{for all } -1 < x < 1$$

Since we're not generally concerned with the endpoints of the interval of convergence, we'll leave as a mystery whether the series converges at  $x = 1$  and  $x = -1$ .

Example 6.2.9

The operations on power series dealt with in Theorem 6.2.1 are fairly easy to apply. Unfortunately taking the product, ratio or composition of two power series is more involved and is beyond the scope of this course<sup>7</sup>. Unfortunately Theorem 6.2.1 alone will not get us power series representations of many of our standard functions (like  $e^x$  and  $\sin x$ ). Fortunately we can find such representations by extending Taylor polynomials<sup>8</sup> to Taylor series.

### 6.3▲ Extending Taylor Polynomials

Recall<sup>9</sup> that Taylor polynomials provide a hierarchy of approximations to a given function  $f(x)$  near a given point  $a$ . Typically, the quality of these approximations improves as we move up the hierarchy.

- The crudest approximation is the constant approximation  $f(x) \approx f(a)$ .
- Then comes the linear, or tangent line, approximation  $f(x) \approx f(a) + f'(a)(x - a)$ .
- Then comes the quadratic approximation

$$f(x) \approx f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2$$

- In general, the Taylor polynomial of degree  $n$ , for the function  $f(x)$ , about the expansion point  $a$ , is the polynomial,  $T_n(x)$ , determined by the requirements that  $f^{(k)}(a) = T_n^{(k)}(a)$  for all  $0 \leq k \leq n$ . That is,  $f$  and  $T_n$  have the same derivatives at  $a$ , up to order  $n$ . Explicitly,

$$\begin{aligned} f(x) \approx T_n(x) &= f(a) + f'(a)(x - a) + \frac{1}{2}f''(a)(x - a)^2 + \cdots + \frac{1}{n!}f^{(n)}(a)(x - a)^n \\ &= \sum_{k=0}^n \frac{1}{k!}f^{(k)}(a)(x - a)^k \end{aligned}$$

These are, of course, approximations — often very good approximations near  $x = a$  — but still just approximations. One might hope that if we let the degree,  $n$ , of the approximation go to infinity then the error in the approximation might go to zero. If that is the case then the “infinite” Taylor polynomial would be an exact representation of the function. Let's see how this might work.

Fix a real number  $a$  and suppose that all derivatives of the function  $f(x)$  exist. Then, for any natural number  $n$ ,

7 As always, a quick visit to your favourite search engine will direct the interested reader to more information.

8 Now is a good time to review your notes from last term, though we'll give you a whirlwind review over the next page or two.

9 Please review your notes from last term if this material is feeling a little unfamiliar.

## Equation 6.3.1.

$$f(x) = T_n(x) + E_n(x)$$

where  $T_n(x)$  is the Taylor polynomial of degree  $n$  for the function  $f(x)$  expanded about  $a$ , and  $E_n(x) = f(x) - T_n(x)$  is the error in our approximation. The Taylor polynomial<sup>10</sup> is given by the formula

## Equation 6.3.1-a

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

while the error satisfies

## Equation 6.3.1-b

$$E_n(x) = \frac{1}{(n+1)!}f^{(n+1)}(c)(x-a)^{n+1}$$

for some  $c$  strictly between  $a$  and  $x$ . Note that we typically do not know the value of  $c$  in the formula for the error. Instead we use the bounds on  $c$  to find bounds on  $f^{(n+1)}(c)$  and so bound the error<sup>11</sup>.

In order for our Taylor polynomial to be an exact representation of the function  $f(x)$  we need the error  $E_n(x)$  to be zero. This will not happen when  $n$  is finite unless  $f(x)$  is a polynomial. However it can happen in the limit as  $n \rightarrow \infty$ , and in that case we can write  $f(x)$  as the limit

$$f(x) = \lim_{n \rightarrow \infty} T_n(x) = \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{1}{k!}f^{(k)}(a)(x-a)^k$$

This is really a limit of partial sums, and so we can write

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}f^{(k)}(a)(x-a)^k$$

which is a power series representation of the function. Let us formalise this in a definition.

10 Did you take a quick look at your notes?

11 The discussion here is only supposed to jog your memory. If it is feeling insufficiently jogged, then please look at your notes from last term.

**Definition 6.3.2** (Taylor series).

The Taylor series for the function  $f(x)$  expanded around  $a$  is the power series

$$f(x) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(a) (x-a)^n$$

provided the series converges. When  $a = 0$  it is also called the Maclaurin series of  $f(x)$ .

This definition hides the discussion of whether or not  $E_n(x) \rightarrow 0$  as  $n \rightarrow \infty$  within the caveat “provided the series converges”. Demonstrating that for a given function can be difficult, but for many of the standard functions you are used to dealing with, it turns out to be pretty easy. Let’s compute a few Taylor series and see how we do it.

**Example 6.3.3** (Exponential Series)

Find the Maclaurin series<sup>12</sup> for  $f(x) = e^x$ .

*Solution.* Just as was the case for computing Taylor polynomials, we need to compute the derivatives of the function at the particular choice of  $a$ . Since we are asked for a Maclaurin series,  $a = 0$ . So now we just need to find  $f^{(k)}(0)$  for all integers  $k \geq 0$ .

We know that  $\frac{d}{dx}e^x = e^x$  and so

$$\begin{aligned} e^x &= f(x) = f'(x) = f''(x) = \dots = f^{(k)}(x) = \dots && \text{which gives} \\ 1 &= f(0) = f'(0) = f''(0) = \dots = f^{(k)}(0) = \dots \end{aligned}$$

Equations (6.3.1) and (6.3.1-a) then give us

$$e^x = f(x) = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + E_n(x)$$

We shall see, in the optional Example 6.3.6 below, that, for any fixed  $x$ ,  $\lim_{n \rightarrow \infty} E_n(x) = 0$ . Consequently, for all  $x$ ,

$$e^x = \lim_{n \rightarrow \infty} \left[ 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \dots + \frac{1}{n!}x^n \right] = \sum_{n=0}^{\infty} \frac{1}{n!}x^n$$

**Example 6.3.3**

We have now seen power series representations for the functions

$$\frac{1}{1-x} \quad \frac{1}{(1-x)^2} \quad \ln(1+x) \quad \arctan(x) \quad e^x.$$

12 Taylor series centred at  $a = 0$

We do not think that you, the reader, will be terribly surprised to see that we develop series for sine and cosine next.

Example 6.3.4 (Sine and Cosine Series)

The trigonometric functions  $\sin x$  and  $\cos x$  also have widely used Maclaurin series expansions (i.e. Taylor series expansions about  $a = 0$ ). To find them, we first compute all derivatives at general  $x$ .

$$\begin{aligned} f(x) &= \sin x & f'(x) &= \cos x & f''(x) &= -\sin x & f^{(3)}(x) &= -\cos x & f^{(4)}(x) &= \sin x & \dots \\ g(x) &= \cos x & g'(x) &= -\sin x & g''(x) &= -\cos x & g^{(3)}(x) &= \sin x & g^{(4)}(x) &= \cos x & \dots \end{aligned}$$

Now set  $x = a = 0$ .

$$\begin{aligned} f(x) &= \sin x & f(0) &= 0 & f'(0) &= 1 & f''(0) &= 0 & f^{(3)}(0) &= -1 & f^{(4)}(0) &= 0 & \dots \\ g(x) &= \cos x & g(0) &= 1 & g'(0) &= 0 & g''(0) &= -1 & g^{(3)}(0) &= 0 & g^{(4)}(0) &= 1 & \dots \end{aligned}$$

For  $\sin x$ , all even numbered derivatives (at  $x = 0$ ) are zero, while the odd numbered derivatives alternate between 1 and  $-1$ . Very similarly, for  $\cos x$ , all odd numbered derivatives (at  $x = 0$ ) are zero, while the even numbered derivatives alternate between 1 and  $-1$ . So, the Taylor polynomials that best approximate  $\sin x$  and  $\cos x$  near  $x = a = 0$  are

$$\begin{aligned} \sin x &\approx x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots \\ \cos x &\approx 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots \end{aligned}$$

We shall see, in the optional Example 6.3.8 below, that, for both  $\sin x$  and  $\cos x$ , we have  $\lim_{n \rightarrow \infty} E_n(x) = 0$  so that

$$\begin{aligned} f(x) &= \lim_{n \rightarrow \infty} \left[ f(0) + f'(0)x + \dots + \frac{1}{n!}f^{(n)}(0)x^n \right] \\ g(x) &= \lim_{n \rightarrow \infty} \left[ g(0) + g'(0)x + \dots + \frac{1}{n!}g^{(n)}(0)x^n \right] \end{aligned}$$

Reviewing the patterns we found in the derivatives, we conclude that, for all  $x$ ,

$$\begin{aligned} \sin x &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)!} x^{2n+1} \\ \cos x &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n)!} x^{2n} \end{aligned}$$

and, in particular, both of the series on the right hand sides converge for all  $x$ .

We could also test for convergence of the series using the ratio test. Computing the ratios of successive terms in these two series gives us

$$\begin{aligned} \left| \frac{A_{n+1}}{A_n} \right| &= \frac{|x|^{2n+3} / (2n+3)!}{|x|^{2n+1} / (2n+1)!} = \frac{|x|^2}{(2n+3)(2n+2)} \\ \left| \frac{A_{n+1}}{A_n} \right| &= \frac{|x|^{2n+2} / (2n+2)!}{|x|^{2n} / (2n)!} = \frac{|x|^2}{(2n+2)(2n+1)} \end{aligned}$$

for sine and cosine respectively. Hence as  $n \rightarrow \infty$  these ratios go to zero and consequently both series are convergent for all  $x$ . (This is very similar to what was observed in Example 6.1.5.)

Example 6.3.4

We have developed power series representations for a number of important functions<sup>13</sup>. Here is a theorem that summarizes them.

**Theorem 6.3.5.**

$$\begin{aligned}
 e^x &= \sum_{n=0}^{\infty} \frac{x^n}{n!} &&= 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots && \text{for all } -\infty < x < \infty \\
 \sin(x) &= \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)!} x^{2n+1} &&= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots && \text{for all } -\infty < x < \infty \\
 \cos(x) &= \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n)!} x^{2n} &&= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots && \text{for all } -\infty < x < \infty \\
 \frac{1}{1-x} &= \sum_{n=0}^{\infty} x^n &&= 1 + x + x^2 + x^3 + \dots && \text{for all } -1 < x < 1 \\
 \ln(1+x) &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} &&= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots && \text{for all } -1 < x \leq 1 \\
 \arctan x &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} &&= x - \frac{x^3}{3} + \frac{x^5}{5} - \dots && \text{for all } -1 \leq x \leq 1
 \end{aligned}$$

Notice that the series for sine and cosine sum to something that looks very similar to

13 The reader might ask whether or not we will give the series for other trigonometric functions or their inverses. While the tangent function has a perfectly well defined series, its coefficients are not as simple as those of the series we have seen — they form a sequence of numbers known (perhaps unsurprisingly) as the “tangent numbers”. They, and the related Bernoulli numbers, have many interesting properties, links to which the interested reader can find with their favourite search engine. The Maclaurin series for inverse sine is

$$\arcsin(x) = \sum_{n=0}^{\infty} \frac{4^{-n}}{2n+1} \frac{(2n)!}{(n!)^2} x^{2n+1}$$

which is quite tidy, but proving it is beyond the scope of the course.

the series for  $e^x$ :

$$\begin{aligned}\sin(x) + \cos(x) &= \left(x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots\right) + \left(1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots\right) \\ &= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \dots \\ e^x &= 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \dots\end{aligned}$$

So both series have coefficients with the same absolute value (namely  $\frac{1}{n!}$ ), but there are differences in sign<sup>14</sup>.

Example 6.3.6 (Optional — Why  $\sum_{n=0}^{\infty} \frac{1}{n!}x^n$  is  $e^x$ .)

We have already seen, in Example 6.3.3, that

$$e^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + E_n(x)$$

By (6.3.1-b)

$$E_n(x) = \frac{1}{(n+1)!}e^c x^{n+1}$$

for some (unknown)  $c$  between 0 and  $x$ . Fix any real number  $x$ . We'll now show that  $E_n(x)$  converges to zero as  $n \rightarrow \infty$ .

To do this we need get bound the size of  $e^c$ , and to do this, consider what happens if  $x$  is positive or negative.

- If  $x < 0$  then  $x \leq c \leq 0$  and hence  $e^x \leq e^c \leq e^0 = 1$ .
- On the other hand, if  $x \geq 0$  then  $0 \leq c \leq x$  and so  $1 = e^0 \leq e^c \leq e^x$ .

In either case we have that  $0 \leq e^c \leq 1 + e^x$ . Because of this the error term

$$|E_n(x)| = \left| \frac{e^c}{(n+1)!} x^{n+1} \right| \leq [e^x + 1] \frac{|x|^{n+1}}{(n+1)!}$$

We claim that this upper bound, and hence the error  $E_n(x)$ , quickly shrinks to zero as  $n \rightarrow \infty$ .

Call the upper bound (except for the factor  $e^x + 1$ , which is independent of  $n$ )  $e_n(x) = \frac{|x|^{n+1}}{(n+1)!}$ . To show that this shrinks to zero as  $n \rightarrow \infty$ , let's write it as follows.

$$e_n(x) = \frac{|x|^{n+1}}{(n+1)!} = \overbrace{\frac{|x|}{1} \cdot \frac{|x|}{2} \cdot \frac{|x|}{3} \cdots \frac{|x|}{n} \cdot \frac{|x|}{n+1}}^{n+1 \text{ factors}}$$

14 Warning: antique sign-sine pun. No doubt the reader first saw it many years syne.



Now let  $k$  be an integer bigger than  $|x|$ . We can split the product

$$\begin{aligned} e_n(x) &= \overbrace{\left( \frac{|x|}{1} \cdot \frac{|x|}{2} \cdot \frac{|x|}{3} \cdots \frac{|x|}{k} \right)}^{k \text{ factors}} \cdot \left( \frac{|x|}{k+1} \cdots \frac{|x|}{n+1} \right) \\ &\leq \underbrace{\left( \frac{|x|}{1} \cdot \frac{|x|}{2} \cdot \frac{|x|}{3} \cdots \frac{|x|}{k} \right)}_{=Q(x)} \cdot \left( \frac{|x|}{k+1} \right)^{n+1-k} \\ &= Q(x) \cdot \left( \frac{|x|}{k+1} \right)^{n+1-k} \end{aligned}$$

Since  $k$  does not depend on  $n$  (though it does depend on  $x$ ), the function  $Q(x)$  does not change as we increase  $n$ . Additionally, we know that  $|x| < k+1$  and so  $\frac{|x|}{k+1} < 1$ . Hence as we let  $n \rightarrow \infty$  the above bound must go to zero.

Alternatively, compare  $e_n(x)$  and  $e_{n+1}(x)$ .

$$\frac{e_{n+1}(x)}{e_n(x)} = \frac{\frac{|x|^{n+2}}{(n+2)!}}{\frac{|x|^{n+1}}{(n+1)!}} = \frac{|x|}{n+2}$$

When  $n$  is bigger than, for example  $2|x|$ , we have  $\frac{e_{n+1}(x)}{e_n(x)} < \frac{1}{2}$ . That is, increasing the index on  $e_n(x)$  by one decreases the size of  $e_n(x)$  by a factor of at least two. As a result  $e_n(x)$  must tend to zero as  $n \rightarrow \infty$ .

Consequently, for all  $x$ ,  $\lim_{n \rightarrow \infty} E_n(x) = 0$ , as claimed, and we really have

$$e^x = \lim_{n \rightarrow \infty} \left[ 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{n!}x^n \right] = \sum_{n=0}^{\infty} \frac{1}{n!}x^n$$

Example 6.3.6

There is another way to prove that the series  $\sum_{n=0}^{\infty} \frac{x^n}{n!}$  converges to the function  $e^x$ . Rather than looking at how the error term  $E_n(x)$  behaves as  $n \rightarrow \infty$ , we can show that the series satisfies the same simple differential equation<sup>15</sup> and the same initial condition as the function.

Example 6.3.7 (Optional — Another approach to showing that  $\sum_{n=0}^{\infty} \frac{1}{n!}x^n$  is  $e^x$ .)

We already know from Example 6.1.5, that the series  $\sum_{n=0}^{\infty} \frac{1}{n!}x^n$  converges to some function  $f(x)$  for all values of  $x$ . All that remains to do is to show that  $f(x)$  is really  $e^x$ . We will do this by showing that  $f(x)$  and  $e^x$  satisfy the same differential equation with the same

<sup>15</sup> Recall, you studied that differential equation in the section on separable differential equations (Theorem 3.9.10 in Section 3.9) as well as wayyyy back in the section on exponential growth and decay in differential calculus.

initial conditions<sup>16</sup>. We know that  $y = e^x$  satisfies

$$\frac{dy}{dx} = y \quad \text{and} \quad y(0) = 1$$

and by Theorem 3.9.10 (with  $a = 1$ ,  $b = 0$  and  $y(0) = 1$ ), this is the only solution. So it suffices to show that  $f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$  satisfies

$$\frac{df}{dx} = f(x) \quad \text{and} \quad f(0) = 1.$$

- By Theorem 6.2.1,

$$\begin{aligned} \frac{df}{dx} &= \frac{d}{dx} \left\{ \sum_{n=0}^{\infty} \frac{1}{n!} x^n \right\} = \sum_{n=1}^{\infty} \frac{n}{n!} x^{n-1} = \sum_{n=1}^{\infty} \frac{1}{(n-1)!} x^{n-1} \\ &= \underbrace{1}_{n=1} + \underbrace{x}_{n=2} + \underbrace{\frac{x^2}{2!}}_{n=3} + \underbrace{\frac{x^3}{3!}}_{n=4} + \cdots \\ &= f(x) \end{aligned}$$

- When we substitute  $x = 0$  into the series we get (see the discussion after Definition 6.1.1)

$$f(0) = 1 + \frac{0}{1!} + \frac{0}{2!} + \cdots = 1.$$

Hence  $f(x)$  solves the same initial value problem and we must have  $f(x) = e^x$ .

Example 6.3.7

We can show that the error terms in Maclaurin polynomials<sup>17</sup> for sine and cosine go to zero as  $n \rightarrow \infty$  using very much the same approach as in Example 6.3.6.

Example 6.3.8 (Optional — Why  $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = \sin x$  and  $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = \cos x$ )

Let  $f(x)$  be either  $\sin x$  or  $\cos x$ . We know that every derivative of  $f(x)$  will be one of  $\pm \sin(x)$  or  $\pm \cos(x)$ . Consequently, when we compute the error term using equation (6.3.1-b) we always have  $|f^{(n+1)}(c)| \leq 1$  and hence

$$|E_n(x)| \leq \frac{|x|^{n+1}}{(n+1)!}.$$

<sup>16</sup> Recall that when we solve of a separable differential equation our general solution will have an arbitrary constant in it. That constant cannot be determined from the differential equation alone and we need some extra data to find it. This extra information is often information about the system at its beginning (for example when position or time is zero) — hence “initial conditions”. Of course the reader is already familiar with this because it was covered back in Section 3.9.

<sup>17</sup> Taylor polynomials centred at  $a = 0$

In Example 6.3.3, we showed that  $\frac{|x|^{n+1}}{(n+1)!} \rightarrow 0$  as  $n \rightarrow \infty$  — so all the hard work is already done. Since the error term shrinks to zero for both  $f(x) = \sin x$  and  $f(x) = \cos x$ , and

$$f(x) = \lim_{n \rightarrow \infty} \left[ f(0) + f'(0)x + \cdots + \frac{1}{n!} f^{(n)}(0) x^n \right]$$

as required.

Example 6.3.8

## 6.4▲ Computing with Taylor Series

Taylor series have a great many applications. (Hence their place in this course.) One of the most immediate of these is that they give us an alternate way of computing many functions. For example, the first definition we see for the sine and cosine functions is in terms of triangles. Those definitions, however, do not lend themselves to computing sine and cosine except at very special angles. Armed with power series representations, however, we can compute them to very high precision at any angle. To illustrate this, consider the computation of  $\pi$  — a problem that dates back to the Babylonians.

Example 6.4.1 (Computing the number  $\pi$ )

There are numerous methods for computing  $\pi$  to any desired degree of accuracy<sup>18</sup>. Many of them use the Maclaurin expansion<sup>19</sup>

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$$

of Theorem 6.3.5. Since  $\arctan(1) = \frac{\pi}{4}$ , the series gives us a very pretty formula for  $\pi$ :

$$\begin{aligned} \frac{\pi}{4} &= \arctan 1 = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \\ \pi &= 4 \left( 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots \right) \end{aligned}$$

Unfortunately, this series is not very useful for computing  $\pi$  because it converges so slowly. If we approximate the series by its  $N^{\text{th}}$  partial sum, then the alternating series test (Theorem A.12.1 in the appendix) tells us that the error is bounded by the first term we drop. To guarantee that we have 2 decimal digits of  $\pi$  correct, we need to sum about the first 200 terms!

18 The computation of  $\pi$  has a very, very long history and your favourite search engine will turn up many sites that explore the topic. For a more comprehensive history one can turn to books such as “A history of Pi” by Petr Beckmann and “The joy of  $\pi$ ” by David Blatner.

19 Taylor expansion centred at  $a = 0$

A much better way to compute  $\pi$  using this series is to take advantage of the fact that  $\tan \frac{\pi}{6} = \frac{1}{\sqrt{3}}$ :

$$\begin{aligned}\pi &= 6 \arctan \left( \frac{1}{\sqrt{3}} \right) = 6 \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \frac{1}{(\sqrt{3})^{2n+1}} \\ &= 2\sqrt{3} \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \frac{1}{3^n} \\ &= 2\sqrt{3} \left( 1 - \frac{1}{3 \times 3} + \frac{1}{5 \times 9} - \frac{1}{7 \times 27} + \frac{1}{9 \times 81} - \frac{1}{11 \times 243} + \dots \right)\end{aligned}$$

Again, this is an alternating series and so (via Theorem A.12.1 in the appendix) the error we introduce by truncating it is bounded by the first term dropped. For example, if we keep ten terms, stopping at  $n = 9$ , we get  $\pi = 3.141591$  (to 6 decimal places) with an error between zero and

$$\frac{2\sqrt{3}}{21 \times 3^{10}} < 3 \times 10^{-6}$$

In 1699, the English astronomer/mathematician Abraham Sharp (1653–1742) used 150 terms of this series to compute 72 digits of  $\pi$  — by hand!

This is just one of very many ways to compute  $\pi$ . Another one, which still uses the Maclaurin expansion<sup>20</sup> of  $\arctan x$ , but is much more efficient, is

$$\pi = 16 \arctan \frac{1}{5} - 4 \arctan \frac{1}{239}$$

This formula was used by John Machin in 1706 to compute  $\pi$  to 100 decimal digits — again, by hand.

(You won't be asked to compute errors using Theorem A.12.1, but we include them here for interest.)

Example 6.4.1

Power series also give us access to new functions which might not be easily expressed in terms of the functions we have been introduced to so far. The following is a good example of this.

Example 6.4.2 (Error function)

The *error function*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

is used in computing “bell curve” probabilities. The indefinite integral of the integrand  $e^{-t^2}$  cannot be expressed in terms of standard functions. But we can still evaluate the integral to within any desired degree of accuracy by using the Taylor expansion of the

<sup>20</sup> Taylor expansion centred at  $a = 0$

exponential. Start with the Maclaurin series<sup>21</sup> for  $e^x$ :

$$e^x = \sum_{n=0}^{\infty} \frac{1}{n!} x^n$$

and then substitute  $x = -t^2$  into this:

$$e^{-t^2} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} t^{2n}$$

We can then apply Theorem 6.2.1 to integrate term-by-term:

$$\begin{aligned} \operatorname{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x \left[ \sum_{n=0}^{\infty} \frac{(-t^2)^n}{n!} \right] dt \\ &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)n!} \end{aligned}$$

For example, for the bell curve, the probability of being within one standard deviation of the mean<sup>22</sup>, is

$$\begin{aligned} \operatorname{erf}(1/\sqrt{2}) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{(1/\sqrt{2})^{2n+1}}{(2n+1)n!} = \frac{2}{\sqrt{2\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)2^n n!} \\ &= \sqrt{\frac{2}{\pi}} \left( 1 - \frac{1}{3 \times 2} + \frac{1}{5 \times 2^2 \times 2} - \frac{1}{7 \times 2^3 \times 3!} + \frac{1}{9 \times 2^4 \times 4!} - \dots \right) \end{aligned}$$

This is yet another alternating series. If we keep five terms, stopping at  $n = 4$ , we get 0.68271 (to 5 decimal places) with, by Theorem A.12.1 in the appendix again, an error between zero and the first dropped term, which is minus

$$\sqrt{\frac{2}{\pi}} \frac{1}{11 \times 2^5 \times 5!} < 2 \times 10^{-5}$$

(You won't be asked to compute such an error, but we include it for interest.)

Example 6.4.2

Example 6.4.3

Evaluate

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n3^n} \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{1}{n3^n}$$

*Solution.* There are not very many series that can be easily evaluated exactly. But occasionally one encounters a series that can be evaluated simply by realizing that it is exactly

<sup>21</sup> Taylor series centred at  $a = 0$

<sup>22</sup> If you don't know what this means (forgive the pun) don't worry, because it is not part of the course. Standard deviation a way of quantifying variation within a population.

one of the series in Theorem 6.3.5, just with a specific value of  $x$ . The left hand given series is

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \frac{1}{3^n} = \frac{1}{3} - \frac{1}{2} \frac{1}{3^2} + \frac{1}{3} \frac{1}{3^3} - \frac{1}{4} \frac{1}{3^4} + \dots$$

The series in Theorem 6.3.5 that this most closely resembles is

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - \dots$$

Indeed

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \frac{1}{3^n} &= \frac{1}{3} - \frac{1}{2} \frac{1}{3^2} + \frac{1}{3} \frac{1}{3^3} - \frac{1}{4} \frac{1}{3^4} + \dots \\ &= \left[ x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - \dots \right]_{x=\frac{1}{3}} \\ &= \left[ \ln(1+x) \right]_{x=\frac{1}{3}} \\ &= \ln \frac{4}{3} \end{aligned}$$

The right hand series above differs from the left hand series above only that the signs of the left hand series alternate while those of the right hand series do not. We can flip every second sign in a power series just by using a negative  $x$ .

$$\begin{aligned} \left[ \ln(1+x) \right]_{x=-\frac{1}{3}} &= \left[ x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - \dots \right]_{x=-\frac{1}{3}} \\ &= -\frac{1}{3} - \frac{1}{2} \frac{1}{3^2} - \frac{1}{3} \frac{1}{3^3} - \frac{1}{4} \frac{1}{3^4} + \dots \end{aligned}$$

which is exactly minus the desired right hand series. So

$$\sum_{n=1}^{\infty} \frac{1}{n3^n} = - \left[ \ln(1+x) \right]_{x=-\frac{1}{3}} = -\ln \frac{2}{3} = \ln \frac{3}{2}$$

Example 6.4.3

Example 6.4.4

Let  $f(x) = \sin(2x^3)$ . Find  $f^{(15)}(0)$ , the fifteenth derivative of  $f$  at  $x = 0$ .

*Solution.* This is a bit of a trick question. We could of course use the product and chain rules to directly apply fifteen derivatives and then set  $x = 0$ , but that would be extremely tedious<sup>23</sup>. There is a much more efficient approach that exploits two pieces of knowledge that we have.

<sup>23</sup> We could get a computer algebra system to do it for us without much difficulty — but we wouldn't learn much in the process. The point of this example is to illustrate that one can do more than just represent a function with Taylor series. More on this in the next section.

- From equation (6.3.1-a), we see that the coefficient of  $(x - a)^n$  in the Taylor series of  $f(x)$  with expansion point  $a$  is exactly  $\frac{1}{n!}f^{(n)}(a)$ . So  $f^{(n)}(a)$  is exactly  $n!$  times the coefficient of  $(x - a)^n$  in the Taylor series of  $f(x)$  with expansion point  $a$ .
- We know, or at least can easily find, the Taylor series for  $\sin(2x^3)$ .

Let's apply that strategy.

- First, we know that, for all  $y$ ,

$$\sin y = y - \frac{1}{3!}y^3 + \frac{1}{5!}y^5 - \dots$$

- Just substituting  $y = 2x^3$ , we have

$$\begin{aligned}\sin(2x^3) &= 2x^3 - \frac{1}{3!}(2x^3)^3 + \frac{1}{5!}(2x^3)^5 - \dots \\ &= 2x^3 - \frac{8}{3!}x^9 + \frac{2^5}{5!}x^{15} - \dots\end{aligned}$$

- So the coefficient of  $x^{15}$  in the Taylor series of  $f(x) = \sin(2x^3)$  with expansion point  $a = 0$  is  $\frac{2^5}{5!}$

and we have

$$f^{(15)}(0) = 15! \times \frac{2^5}{5!} = 348,713,164,800$$

Example 6.4.4

Example 6.4.5 (Optional — Computing the number  $e$ )

Back in Example 6.3.6, we saw that

$$e^x = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \frac{1}{(n+1)!}e^c x^{n+1}$$

for some (unknown)  $c$  between 0 and  $x$ . This can be used to approximate the number  $e$ , with any desired degree of accuracy. Setting  $x = 1$  in this equation gives

$$e = 1 + 1 + \frac{1}{2!} + \dots + \frac{1}{n!} + \frac{1}{(n+1)!}e^c$$

for some  $c$  between 0 and 1. Even though we don't know  $c$  exactly, we can bound that term quite readily. We do know that  $e^c$  is an increasing function<sup>24</sup> of  $c$ , and so  $1 = e^0 \leq e^c \leq e^1 = e$ . Thus we know that

$$\frac{1}{(n+1)!} \leq e - \left(1 + 1 + \frac{1}{2!} + \dots + \frac{1}{n!}\right) \leq \frac{e}{(n+1)!}$$

So we have a lower bound on the error, but our upper bound involves the  $e$  — precisely the quantity we are trying to get a handle on.

24 Check the derivative!

But all is not lost. Let's look a little more closely at the right-hand inequality when  $n = 1$ :

$$\begin{aligned} e - (1 + 1) &\leq \frac{e}{2} && \text{move the } e\text{'s to one side} \\ \frac{e}{2} &\leq 2 && \text{and clean it up} \\ e &\leq 4. \end{aligned}$$

Now this is a pretty crude bound<sup>25</sup> but it isn't hard to improve. Try this again with  $n = 1$ :

$$\begin{aligned} e - \left(1 + 1 + \frac{1}{2}\right) &\leq \frac{e}{6} && \text{move } e\text{'s to one side} \\ \frac{5e}{6} &\leq \frac{5}{2} \\ e &\leq 3. \end{aligned}$$

Better. Now we can rewrite our bound:

$$\frac{1}{(n+1)!} \leq e - \left(1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{n!}\right) \leq \frac{e}{(n+1)!} \leq \frac{3}{(n+1)!}$$

If we set  $n = 4$  in this we get

$$\frac{1}{120} = \frac{1}{5!} \leq e - \left(1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24}\right) \leq \frac{3}{120}$$

So the error is between  $\frac{1}{120}$  and  $\frac{3}{120} = \frac{1}{40}$  — this approximation isn't guaranteed to give us the first 2 decimal places. If we ramp  $n$  up to 9 however, we get

$$\frac{1}{10!} \leq e - \left(1 + 1 + \frac{1}{2} + \cdots + \frac{1}{9!}\right) \leq \frac{3}{10!}$$

Since  $10! = 3628800$ , the upper bound on the error is  $\frac{3}{3628800} < \frac{3}{3000000} = 10^{-6}$ , and we can approximate  $e$  by

$$\begin{aligned} &1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} + \frac{1}{8!} + \frac{1}{9!} \\ &= 1 + 1 + 0.5 + 0.1\dot{6} + 0.041\dot{6} + 0.008\dot{3} + 0.0013\dot{8} + 0.0001984 + 0.0000248 + 0.0000028 \\ &= 2.718282 \end{aligned}$$

and it is correct to six decimal places.

Example 6.4.5

<sup>25</sup> The authors hope that by now we all "know" that  $e$  is between 2 and 3, but maybe we don't know how to prove it.



## 6.5▲ Evaluating Limits using Taylor Expansions

Taylor polynomials provide a good way to understand the behaviour of a function near a specified point and so are useful for evaluating complicated limits. Here are some examples.

### Example 6.5.1

In this example, we'll start with a relatively simple limit, namely

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

The first thing to notice about this limit is that, as  $x$  tends to zero, both the numerator,  $\sin x$ , and the denominator,  $x$ , tend to 0. So we may not evaluate the limit of the ratio by simply dividing the limits of the numerator and denominator. To find the limit, or show that it does not exist, we are going to have to exhibit a cancellation between the numerator and the denominator. Let's start by taking a closer look at the numerator. By Example 6.3.4,

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots$$

Consequently<sup>26</sup>

$$\frac{\sin x}{x} = 1 - \frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \dots$$

Every term in this series, except for the very first term, is proportional to a strictly positive power of  $x$ . Consequently, as  $x$  tends to zero, all terms in this series, except for the very first term, tend to zero. In fact the sum of all terms, starting with the second term, also tends to zero. That is,

$$\lim_{x \rightarrow 0} \left[ -\frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \dots \right] = 0$$

We won't justify that statement here, but it will be justified in the following (optional) subsection. So

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\sin x}{x} &= \lim_{x \rightarrow 0} \left[ 1 - \frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \dots \right] \\ &= 1 + \lim_{x \rightarrow 0} \left[ -\frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \dots \right] \\ &= 1 \end{aligned}$$

26 We are hiding some mathematics behind this "consequently". What we are really using our knowledge of Taylor polynomials to write

$$f(x) = \sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + E_5(x)$$

where  $E_5(x) = \frac{f^{(6)}(c)}{6!}x^6$  and  $c$  is between 0 and  $x$ . We are effectively hiding " $E_5(x)$ " inside the "...". Now we can divide both sides by  $x$  (assuming  $x \neq 0$ ):

$$\frac{\sin(x)}{x} = 1 - \frac{1}{3!}x^2 + \frac{1}{5!}x^4 + \frac{E_5(x)}{x}.$$

and everything is fine provided the term  $\frac{E_5(x)}{x}$  stays well behaved.

## Example 6.5.1

The limit in the previous example can also be evaluated relatively easily using l'Hôpital's rule<sup>27</sup>. While the following limit can also, in principle, be evaluated using l'Hôpital's rule, it is much more efficient to use Taylor series<sup>28</sup>.

## Example 6.5.2

In this example we evaluate

$$\lim_{x \rightarrow 0} \frac{\arctan x - x}{\sin x - x}$$

Once again, the first thing to notice about this limit is that, as  $x$  tends to zero, the numerator tends to  $\arctan 0 - 0$ , which is 0, and the denominator tends to  $\sin 0 - 0$ , which is also 0. So we may not evaluate the limit of the ratio by simply dividing the limits of the numerator and denominator. Again, to find the limit, or show that it does not exist, we are going to have to exhibit a cancellation between the numerator and the denominator. To get a more detailed understanding of the behaviour of the numerator and denominator near  $x = 0$ , we find their Taylor expansions. By Example 6.2.9,

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \dots$$

so the numerator

$$\arctan x - x = -\frac{x^3}{3} + \frac{x^5}{5} - \dots$$

By Example 6.3.4,

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots$$

so the denominator

$$\sin x - x = -\frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots$$

and the ratio

$$\frac{\arctan x - x}{\sin x - x} = \frac{-\frac{x^3}{3} + \frac{x^5}{5} - \dots}{-\frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots}$$

Notice that every term in both the numerator and the denominator contains a common factor of  $x^3$ , which we can cancel out.

$$\frac{\arctan x - x}{\sin x - x} = \frac{-\frac{1}{3} + \frac{x^2}{5} - \dots}{-\frac{1}{3!} + \frac{1}{5!}x^2 - \dots}$$

As  $x$  tends to zero,

27 Many of you learned about l'Hôpital's rule in school and all of you should have seen it last term in your differential calculus course.

28 It takes 3 applications of l'Hôpital's rule and some careful cleaning up of the intermediate expressions. Oof!

- the numerator tends to  $-\frac{1}{3}$ , which is not 0, and
- the denominator tends to  $-\frac{1}{3!} = -\frac{1}{6}$ , which is also not 0.

so we may now legitimately evaluate the limit of the ratio by simply dividing the limits of the numerator and denominator.

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\arctan x - x}{\sin x - x} &= \lim_{x \rightarrow 0} \frac{-\frac{1}{3} + \frac{x^2}{5} - \dots}{-\frac{1}{3!} + \frac{1}{5!}x^2 - \dots} \\ &= \frac{\lim_{x \rightarrow 0} \left[ -\frac{1}{3} + \frac{x^2}{5} - \dots \right]}{\lim_{x \rightarrow 0} \left[ -\frac{1}{3!} + \frac{1}{5!}x^2 - \dots \right]} \\ &= \frac{-1/3}{-1/3!} \\ &= 2 \end{aligned}$$

Example 6.5.2

Chapter 5 of this work was adapted from Chapter 3 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

# PROOFS AND SUPPLEMENTS

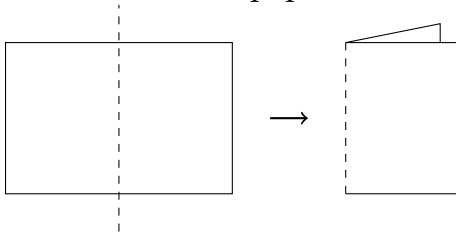
## A.1▲ Folding the First Octant of $\mathbb{R}^3$

This text, whether you're reading it on a computer screen or a printed page, exists in two dimensions. So, anything we draw in three dimensions is going to require a little bit of imagination. If you're struggling to understand the figures with three coordinates, it might help to make your own model of these axes.

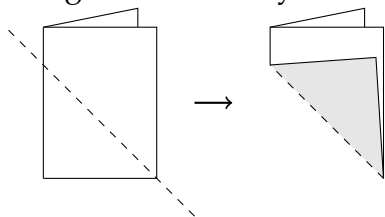
In the Cartesian plane, the first quadrant is the part of the plane where both  $x$  and  $y$  are positive.  $\mathbb{R}^3$  divides three-dimensional space into eight regions, called octants. The first octant is the region where all of  $x$ ,  $y$ , and  $z$  are positive.

Following the instructions below, you can fold a piece of paper into an octant.

1. Fold your paper in half "hamburger style" (so that the fold goes along the shorter dimension of the paper). Position it so that it opens like a book<sup>1</sup>.

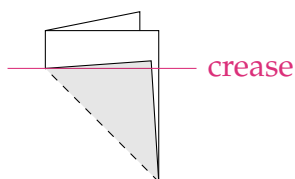


2. Bring the corner of your folded paper up to the side.

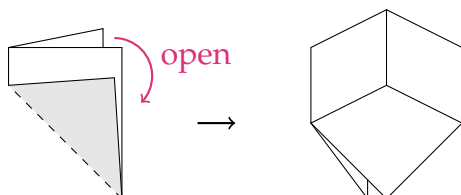


3. Your paper now has a triangle sitting on top of a rectangle. Where the triangle ends, make a crease in the underlying rectangle shapes.

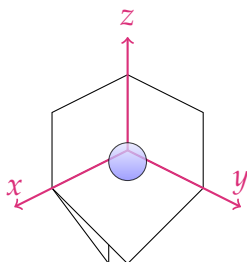
<sup>1</sup> in a language written left-to-right



4. Your paper has four layers, with the triangle shapes on top. Open the paper so that three layers are on top, and one is on the bottom. The result should look like the inside corner of a box.



Your octant is created! The vertical crease is the  $z$  axis, the crease to the left is the  $x$  axis, and the crease to the right is the  $y$  axis. In the picture below, the blue sphere indicates that the octant is open towards you: if you were to put a marble inside the paper structure, it would sit as shown.



To practice with your octant, label the following points directly on the paper:

- $(1, 1, 0)$
- $(0, 1, 1)$
- $(1, 0, 1)$

The next collection of points will exist out in space, not on any of the paper sides. Point to their positions relative to your octant:

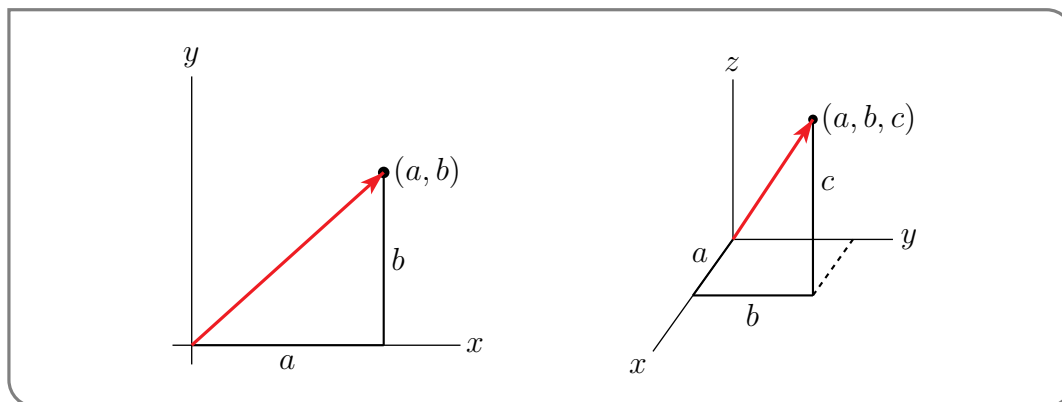
- $(1, 1, 1)$
- $(1, 2, 3)$
- $(1, -1, 1)$
- $(1, 1, -1)$

---

## A.2▲ Vectors

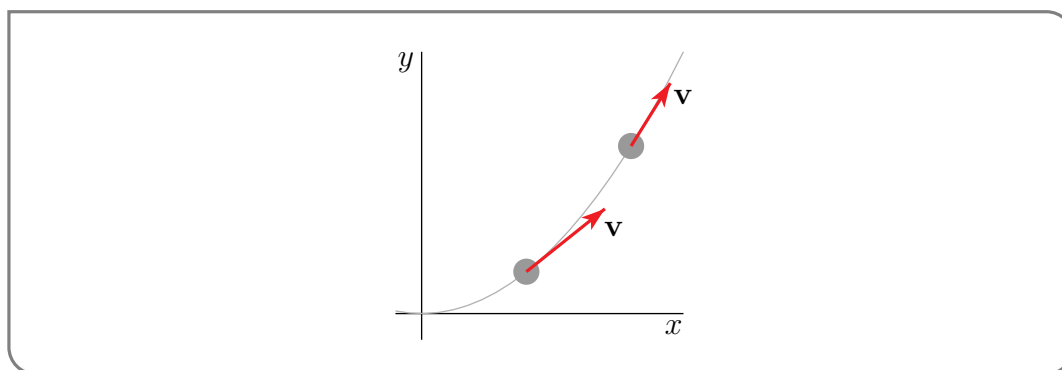
In many of our applications in 2d and 3d, we will encounter quantities that have both a magnitude (like a distance) and also a direction. Such quantities are called vectors. That is,

a *vector* is a quantity which has both a direction and a magnitude, like a velocity. If you are moving, the magnitude (length) of your velocity vector is your speed (distance travelled per unit time) and the direction of your velocity vector is your direction of motion. To specify a vector in three dimensions you have to give three components, just as for a point. To draw the vector with components  $a$ ,  $b$ ,  $c$  you can draw an arrow from the point  $(0,0,0)$  to the point  $(a,b,c)$ . Similarly, to specify a vector in two dimensions you have to

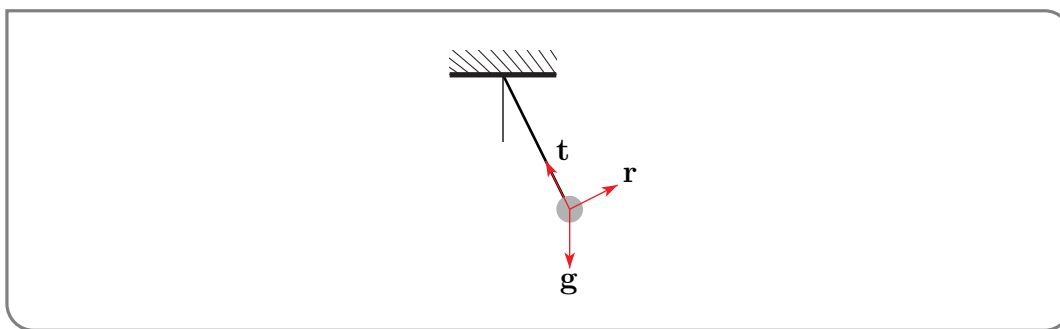


give two components. To draw the vector with components  $a$  and  $b$ , you can draw an arrow from the point  $(0,0)$  to the point  $(a,b)$ .

There are many situations in which it is preferable to draw a vector with its tail at some point other than the origin. For example, it is natural to draw the velocity vector of a moving particle with the tail of the velocity vector at the position of the particle, whether or not the particle is at the origin. The sketch below shows a moving particle and its velocity vector at two different times.

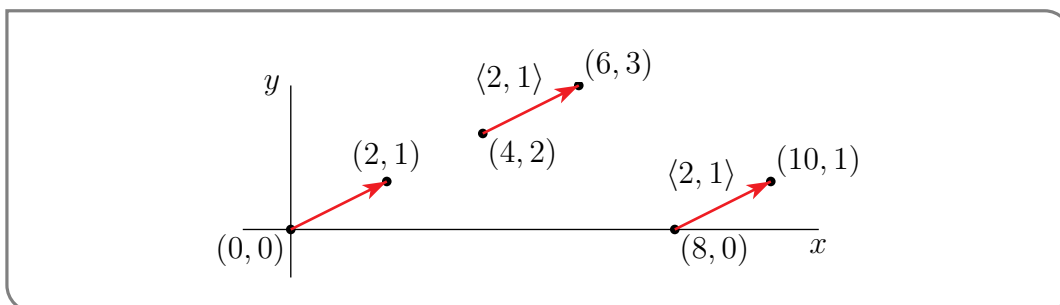


As a second example, suppose that you are analyzing the motion of a pendulum. There are three forces acting on the pendulum bob: gravity  $\mathbf{g}$ , which is pulling the bob straight down, tension  $\mathbf{t}$  in the rod, which is pulling the bob in the direction of the rod, and air resistance  $\mathbf{r}$ , which is pulling the bob in a direction opposite to its direction of motion. All three forces are acting on the bob. So it is natural to draw all three arrows representing the forces with their tails at the ball.



In this text, we will use bold faced letters, like  $\mathbf{v}$ ,  $\mathbf{t}$ ,  $\mathbf{g}$ , to designate vectors. In handwriting, it is clearer to use a small overhead arrow<sup>2</sup>, as in  $\vec{v}$ ,  $\vec{t}$ ,  $\vec{g}$ , instead. Also, when we want to emphasize that some quantity is a number, rather than a vector, we will call the number a *scalar*.

Both points and vectors in 2d are specified by two numbers. Until you get used to this, it might confuse you sometimes — does a given pair of numbers represent a point or a vector? To distinguish<sup>3</sup> between the components of a vector and the coordinates of the point at its head, when its tail is at some point other than the origin, we shall use angle brackets rather than round brackets around the components of a vector. For example, the figure below shows the two-dimensional vector  $\langle 2, 1 \rangle$  drawn in three different positions. In each case, when the tail is at the point  $(u, v)$  the head is at  $(2 + u, 1 + v)$ . We warn you that, out in the real world<sup>4</sup>, no one uses notation that distinguishes between components of a vector and the coordinates of its head — usually round brackets are used for both. It is up to you to keep straight which is being referred to.



By way of summary,

#### Notation A.2.1.

we use

- bold faced letters, like  $\mathbf{v}$ ,  $\mathbf{t}$ ,  $\mathbf{g}$ , to designate vectors, and
- angle brackets, like  $\langle 2, 1 \rangle$ , around the components of a vector, but use
- round brackets, like  $(2, 1)$ , around the coordinates of a point, and use
- “scalar” to emphasise that some quantity is a number, rather than a vector.

2 Some people use an underline, as in  $\underline{v}$ , rather than an arrow.

3 Or, in the Wikipedia jargon, disambiguate.

4 OK. OK. Out in that (admittedly very small) part of the real world that actually knows what a vector is.

### A.2.1 ► Addition of Vectors and Multiplication of a Vector by a Scalar

Just as we have done many times in the texts, when we define a new type of object, we want to understand how it interacts with the basic operations of addition and multiplication. Vectors are no different, and the following is a natural way to define addition of vectors. Multiplication will be more subtle, and we start with multiplication of a vector by a number (rather than with multiplication of a vector by another vector).

#### Definition A.2.2 (Adding Vectors and Multiplying a Vector by a Number).

These two operations have the obvious definitions

$$\mathbf{a} = \langle a_1, a_2 \rangle, \mathbf{b} = \langle b_1, b_2 \rangle \implies \mathbf{a} + \mathbf{b} = \langle a_1 + b_1, a_2 + b_2 \rangle$$

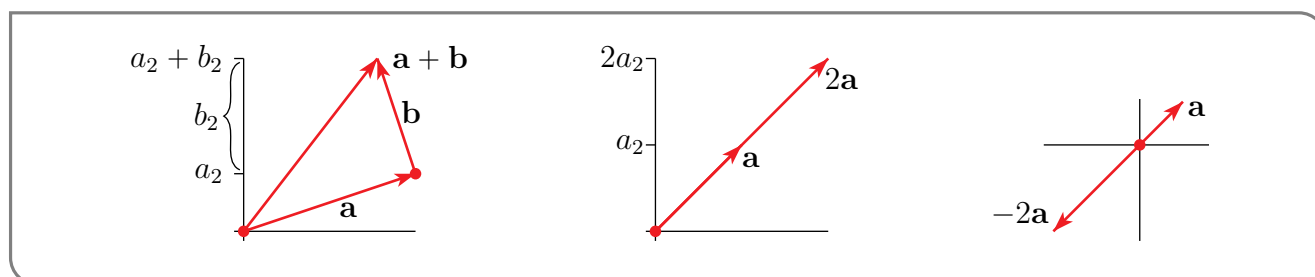
$$\mathbf{a} = \langle a_1, a_2 \rangle, s \text{ a number} \implies s\mathbf{a} = \langle sa_1, sa_2 \rangle$$

and similarly in three dimensions.

Pictorially, you add the vector  $\mathbf{b}$  to the vector  $\mathbf{a}$  by drawing  $\mathbf{b}$  with its tail at the head of  $\mathbf{a}$  and then drawing a vector from the tail of  $\mathbf{a}$  to the head of  $\mathbf{b}$ , as in the figure on the left below. For a number  $s$ , we can draw the vector  $s\mathbf{a}$ , by just

- changing the vector  $\mathbf{a}$ 's length by the factor  $|s|$ , and,
- if  $s < 0$ , reversing the arrow's direction,

as in the other two figures below.



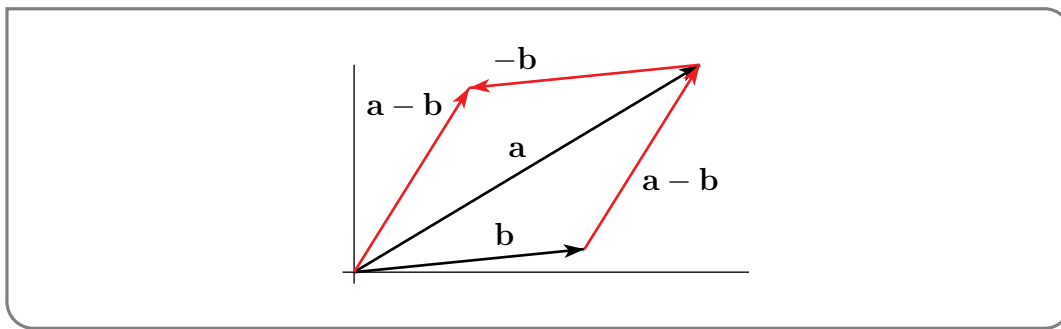
The special case of multiplication by  $s = -1$  appears so frequently that  $(-1)\mathbf{a}$  is given the shorter notation  $-\mathbf{a}$ . That is,

$$-\langle a_1, a_2 \rangle = \langle -a_1, -a_2 \rangle$$

Of course  $\mathbf{a} + (-\mathbf{a})$  is  $\mathbf{0}$ , the vector all of whose components are zero.

To subtract  $\mathbf{b}$  from  $\mathbf{a}$  pictorially, you may add  $-\mathbf{b}$  (which is drawn by reversing the direction of  $\mathbf{b}$ ) to  $\mathbf{a}$ . Alternatively, if you draw  $\mathbf{a}$  and  $\mathbf{b}$  with their tails at a common point, then  $\mathbf{a} - \mathbf{b}$  is the vector from the head of  $\mathbf{b}$  to the head of  $\mathbf{a}$ . That is,  $\mathbf{a} - \mathbf{b}$  is the vector you must add to  $\mathbf{b}$  in order to get  $\mathbf{a}$ .





The operations of addition and multiplication by a scalar that we have just defined are quite natural and rarely cause any problems, because they inherit from the real numbers the properties of addition and multiplication that you are used to.

**Theorem A.2.3** (Properties of Addition and Scalar Multiplication).

Let  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  be vectors and  $s$  and  $t$  be scalars. Then

- |  |   |
|--|---|
| (1) $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$      | (2) $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$ |
| (3) $\mathbf{a} + \mathbf{0} = \mathbf{a}$                   | (4) $\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$   |
| (5) $s(\mathbf{a} + \mathbf{b}) = s\mathbf{a} + s\mathbf{b}$ | (6) $(s + t)\mathbf{a} = s\mathbf{a} + t\mathbf{a}$                                   |
| (7) $(st)\mathbf{a} = s(t\mathbf{a})$                        | (8) $1\mathbf{a} = \mathbf{a}$  |

We have just been introduced to many definitions. Let's see some of them in action.

**Example A.2.4**

For example, if

$$\mathbf{a} = \langle 1, 2, 3 \rangle \quad \mathbf{b} = \langle 3, 2, 1 \rangle \quad \mathbf{c} = \langle 1, 0, 1 \rangle$$

then

$$2\mathbf{a} = 2 \langle 1, 2, 3 \rangle = \langle 2, 4, 6 \rangle$$

$$-\mathbf{b} = -\langle 3, 2, 1 \rangle = \langle -3, -2, -1 \rangle$$

$$3\mathbf{c} = 3 \langle 1, 0, 1 \rangle = \langle 3, 0, 3 \rangle$$

and

$$\begin{aligned} 2\mathbf{a} - \mathbf{b} + 3\mathbf{c} &= \langle 2, 4, 6 \rangle + \langle -3, -2, -1 \rangle + \langle 3, 0, 3 \rangle \\ &= \langle 2 - 3 + 3, 4 - 2 + 0, 6 - 1 + 3 \rangle \\ &= \langle 2, 2, 8 \rangle \end{aligned}$$

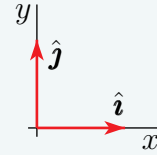
**Example A.2.4**

There are some vectors that occur sufficiently commonly that they are given special names. One is the vector  $\mathbf{0}$ . Some others are the “standard basis vectors”.

**Definition A.2.5.**

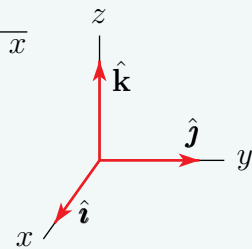
(a) The standard basis vectors in two dimensions are

$$\hat{\mathbf{i}} = \langle 1, 0 \rangle \quad \hat{\mathbf{j}} = \langle 0, 1 \rangle$$



(b) The standard basis vectors in three dimensions are

$$\hat{\mathbf{i}} = \langle 1, 0, 0 \rangle \quad \hat{\mathbf{j}} = \langle 0, 1, 0 \rangle \quad \hat{\mathbf{k}} = \langle 0, 0, 1 \rangle$$



We’ll explain the little hats in the notation  $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$  shortly. Some people rename  $\hat{\mathbf{i}}, \hat{\mathbf{j}}$  and  $\hat{\mathbf{k}}$  to  $\mathbf{e}_1, \mathbf{e}_2$  and  $\mathbf{e}_3$  respectively. Using the above properties we have, for all vectors,

$$\langle a_1, a_2 \rangle = a_1 \hat{\mathbf{i}} + a_2 \hat{\mathbf{j}} \quad \langle a_1, a_2, a_3 \rangle = a_1 \hat{\mathbf{i}} + a_2 \hat{\mathbf{j}} + a_3 \hat{\mathbf{k}}$$

A sum of numbers times vectors, like  $a_1 \hat{\mathbf{i}} + a_2 \hat{\mathbf{j}}$  is called a linear combination of the vectors. Thus all vectors can be expressed as linear combinations of the standard basis vectors. This makes basis vectors very helpful in computations. The standard basis vectors are unit vectors, meaning that they are of length one, where the length of a vector  $\mathbf{a}$  is denoted<sup>5</sup>  $|\mathbf{a}|$  and is defined by

**Definition A.2.6 (Length of a Vector).**

$$\mathbf{a} = \langle a_1, a_2 \rangle \quad \implies \quad |\mathbf{a}| = \sqrt{a_1^2 + a_2^2}$$

$$\mathbf{a} = \langle a_1, a_2, a_3 \rangle \quad \implies \quad |\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

A unit vector is a vector of length one. We’ll sometimes use the accent ^ to emphasise that the vector  $\hat{\mathbf{a}}$  is a unit vector. That is,  $|\hat{\mathbf{a}}| = 1$ .

**Example A.2.7**

Recall that multiplying a vector  $\mathbf{a}$  by a positive number  $s$ , changes the length of the vector by a factor  $s$  without changing the direction of the vector. So (assuming that  $|\mathbf{a}| \neq 0$ )  $\frac{\mathbf{a}}{|\mathbf{a}|}$  is a unit vector that has the same direction as  $\mathbf{a}$ . For example,  $\frac{\langle 1, 1, 1 \rangle}{\sqrt{3}}$  is a unit vector that points in the same direction as  $\langle 1, 1, 1 \rangle$ .

<sup>5</sup> The notation  $\|\mathbf{a}\|$  is also used for the length of  $\mathbf{a}$ .

## A.2.2 ▶ The Dot Product

Let's get back to the arithmetic operations of addition and multiplication. We will be using both scalars and vectors. So, for each operation there are three possibilities that we need to explore:

- “scalar plus scalar”, “scalar plus vector” and “vector plus vector”
- “scalar times scalar”, “scalar times vector” and “vector times vector”

We have been using “scalar plus scalar” and “scalar times scalar” since childhood. “Vector plus vector” and “scalar times vector” were just defined above. There is no sensible way to define “scalar plus vector”, so we won't. This leaves “vector times vector”. There are actually two widely used such products. The first is the *dot product*, which is the topic of this section, and which is used to easily determine the angle  $\theta$  (or more precisely,  $\cos \theta$ ) between two vectors. (The second widely-used product of two vectors, the *cross product*, is not a part of this course.)

### Definition A.2.8 (Dot Product).

The dot product of the vectors  $\mathbf{a}$  and  $\mathbf{b}$  is denoted  $\mathbf{a} \cdot \mathbf{b}$  and is defined by

$$\begin{aligned} \mathbf{a} = \langle a_1, a_2 \rangle, \quad \mathbf{b} = \langle b_1, b_2 \rangle &\implies \mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 \\ \mathbf{a} = \langle a_1, a_2, a_3 \rangle, \quad \mathbf{b} = \langle b_1, b_2, b_3 \rangle &\implies \mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 \end{aligned}$$

in two and three dimensions respectively.

The properties of the dot product are as follows:

### Theorem A.2.9 (Properties of the Dot Product).

Let  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  be vectors and let  $s$  be a scalar. Then

- (0)  $\mathbf{a}, \mathbf{b}$  are vectors and  $\mathbf{a} \cdot \mathbf{b}$  is a scalar
- (1)  $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$
- (2)  $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$
- (3)  $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}, \quad (\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c}$
- (4)  $(s\mathbf{a}) \cdot \mathbf{b} = s(\mathbf{a} \cdot \mathbf{b})$
- (5)  $\mathbf{0} \cdot \mathbf{a} = 0$
- (6)  $\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta$  where  $\theta$  is the angle between  $\mathbf{a}$  and  $\mathbf{b}$
- (7)  $\mathbf{a} \cdot \mathbf{b} = 0 \iff \mathbf{a} = \mathbf{0} \text{ or } \mathbf{b} = \mathbf{0} \text{ or } \mathbf{a} \perp \mathbf{b}$

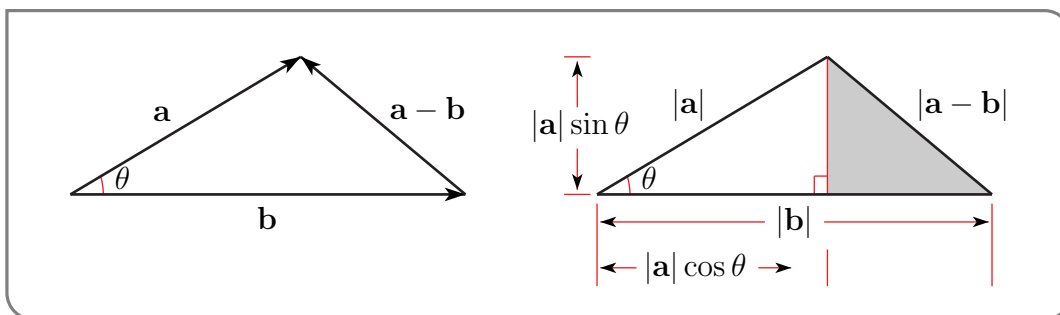
*Proof.* Properties 0 through 5 are almost immediate consequences of the definition. For example, for property 3 (which is called the distributive law) in dimension 2,

$$\begin{aligned} \mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) &= \langle a_1, a_2 \rangle \cdot \langle b_1 + c_1, b_2 + c_2 \rangle \\ &= a_1(b_1 + c_1) + a_2(b_2 + c_2) = a_1b_1 + a_1c_1 + a_2b_2 + a_2c_2 \\ \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} &= \langle a_1, a_2 \rangle \cdot \langle b_1, b_2 \rangle + \langle a_1, a_2 \rangle \cdot \langle c_1, c_2 \rangle \\ &= a_1b_1 + a_2b_2 + a_1c_1 + a_2c_2 \end{aligned}$$

Property 6 is sufficiently important that it is often used as the definition of dot product. It is not at all an obvious consequence of the definition. To verify it, we just write  $|\mathbf{a} - \mathbf{b}|^2$  in two different ways. The first expresses  $|\mathbf{a} - \mathbf{b}|^2$  in terms of  $\mathbf{a} \cdot \mathbf{b}$ . It is

$$\begin{aligned} |\mathbf{a} - \mathbf{b}|^2 &\stackrel{1}{=} (\mathbf{a} - \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) \\ &\stackrel{3}{=} \mathbf{a} \cdot \mathbf{a} - \mathbf{a} \cdot \mathbf{b} - \mathbf{b} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b} \\ &\stackrel{1,2}{=} |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2\mathbf{a} \cdot \mathbf{b} \end{aligned}$$

Here,  $\stackrel{1}{=}$ , for example, means that the equality is a consequence of property 1. The second way we write  $|\mathbf{a} - \mathbf{b}|^2$  involves  $\cos \theta$  and follows from the cosine law for triangles. Just in case you don't remember the cosine law, we'll derive it right now! Start by applying Pythagoras to the shaded triangle in the right hand figure of



That triangle is a right triangle whose hypotenuse has length  $|\mathbf{a} - \mathbf{b}|$  and whose other two sides have lengths  $(|\mathbf{b}| - |\mathbf{a}| \cos \theta)$  and  $|\mathbf{a}| \sin \theta$ . So Pythagoras gives

$$\begin{aligned} |\mathbf{a} - \mathbf{b}|^2 &= (|\mathbf{b}| - |\mathbf{a}| \cos \theta)^2 + (|\mathbf{a}| \sin \theta)^2 \\ &= |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}| \cos \theta + |\mathbf{a}|^2 \cos^2 \theta + |\mathbf{a}|^2 \sin^2 \theta \\ &= |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}| \cos \theta + |\mathbf{a}|^2 \end{aligned}$$

This is precisely the cosine law<sup>6</sup>. Observe that, when  $\theta = \frac{\pi}{2}$ , this reduces to, (surprise!) Pythagoras' Theorem.

Setting our two expressions for  $|\mathbf{a} - \mathbf{b}|^2$  equal to each other,

$$|\mathbf{a} - \mathbf{b}|^2 = |\mathbf{a}|^2 + |\mathbf{b}|^2 - 2\mathbf{a} \cdot \mathbf{b} = |\mathbf{b}|^2 - 2|\mathbf{a}||\mathbf{b}| \cos \theta + |\mathbf{a}|^2$$

<sup>6</sup> You may be used to seeing it written as  $c^2 = a^2 + b^2 - 2ab \cos C$ , where  $a$ ,  $b$  and  $c$  are the lengths of the three sides of the triangle and  $C$  is the angle opposite the side of length  $c$

cancelling the  $|\mathbf{a}|^2$  and  $|\mathbf{b}|^2$  common to both sides

$$-2\mathbf{a} \cdot \mathbf{b} = -2|\mathbf{a}||\mathbf{b}|\cos\theta$$

and dividing by  $-2$  gives

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$$

which is exactly property 6.

Property 7 follows directly from property 6. First note that the dot product  $\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta$  is zero if and only if at least one of the three factors  $|\mathbf{a}|$ ,  $|\mathbf{b}|$ ,  $\cos\theta$  is zero. The first factor is zero if and only if  $\mathbf{a} = \mathbf{0}$ . The second factor is zero if and only if  $\mathbf{b} = \mathbf{0}$ . The third factor is zero if and only if  $\theta = \pm\frac{\pi}{2} + 2k\pi$ , for some integer  $k$ , which in turn is true if and only if  $\mathbf{a}$  and  $\mathbf{b}$  are mutually perpendicular.  $\square$

Because of Property 7 of Theorem A.2.9, the dot product can be used to test whether or not two vectors are perpendicular to each other. That is, whether or not the angle between the two vectors is  $90^\circ$ . Another name<sup>7</sup> for “perpendicular” is “orthogonal”. Testing for orthogonality is one of the main uses of the dot product.

### Example A.2.10

Consider the three vectors

$$\mathbf{a} = \langle 1, 1, 0 \rangle \quad \mathbf{b} = \langle 1, 0, 1 \rangle \quad \mathbf{c} = \langle -1, 1, 1 \rangle$$

Their dot products

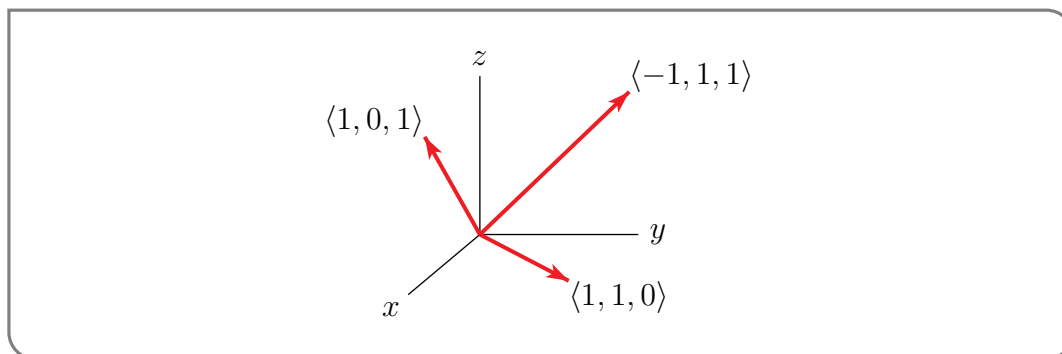
$$\mathbf{a} \cdot \mathbf{b} = \langle 1, 1, 0 \rangle \cdot \langle 1, 0, 1 \rangle = 1 \times 1 + 1 \times 0 + 0 \times 1 = 1$$

$$\mathbf{a} \cdot \mathbf{c} = \langle 1, 1, 0 \rangle \cdot \langle -1, 1, 1 \rangle = 1 \times (-1) + 1 \times 1 + 0 \times 1 = 0$$

$$\mathbf{b} \cdot \mathbf{c} = \langle 1, 0, 1 \rangle \cdot \langle -1, 1, 1 \rangle = 1 \times (-1) + 0 \times 1 + 1 \times 1 = 0$$

tell us that  $\mathbf{c}$  is perpendicular to both  $\mathbf{a}$  and  $\mathbf{b}$ . Since both  $|\mathbf{a}| = |\mathbf{b}| = \sqrt{1^2 + 1^2 + 0^2} = \sqrt{2}$  the first dot product tells us that the angle,  $\theta$ , between  $\mathbf{a}$  and  $\mathbf{b}$  obeys

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \frac{1}{2} \implies \theta = \frac{\pi}{3}$$

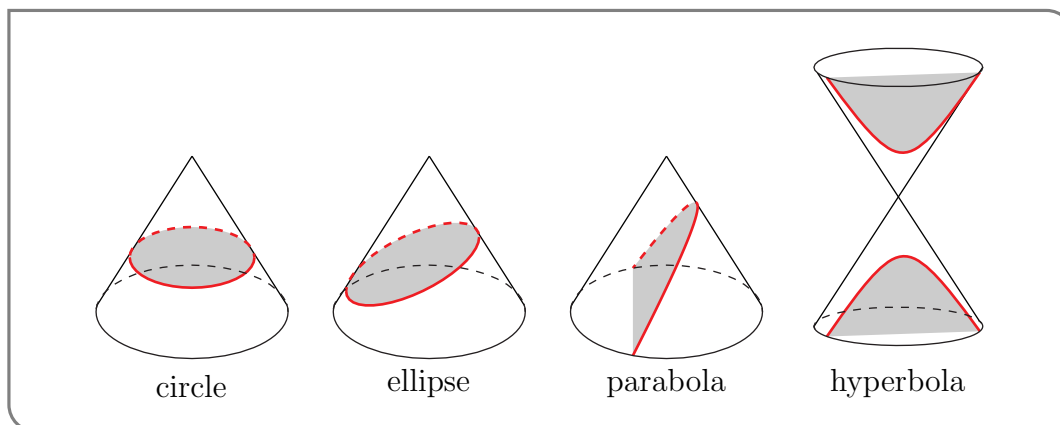


### Example A.2.10

<sup>7</sup> The concepts of the dot product and perpendicularity have been generalized a lot in mathematics (for example, from 2d and 3d vectors to functions). The generalization of the dot product is called the “inner product” and the generalization of perpendicularity is called “orthogonality”.

## A.3<sup>▲</sup> Conic Sections and Quadric Surfaces

A conic section is the curve of intersection of a cone and a plane that does not pass through the vertex of the cone. This is illustrated in the figures below. An equivalent<sup>8</sup> (and often



used) definition is that a conic section is the set of all points in the  $xy$ -plane that obey  $Q(x, y) = 0$  with

$$Q(x, y) = Ax^2 + By^2 + Cxy + Dx + Ey + F = 0$$

being a polynomial of degree two<sup>9</sup>. By rotating and translating our coordinate system the equation of the conic section can be brought into one of the forms<sup>10</sup>

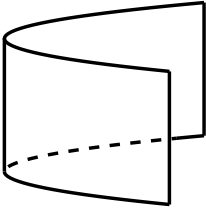
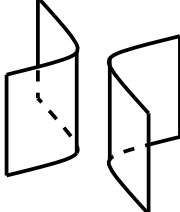
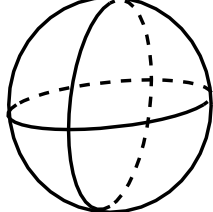
- $\alpha x^2 + \beta y^2 = \gamma$  with  $\alpha, \beta, \gamma > 0$ , which is an ellipse (or a circle),
- $\alpha x^2 - \beta y^2 = \gamma$  with  $\alpha, \beta > 0, \gamma \neq 0$ , which is a hyperbola,
- $x^2 = \delta y$ , with  $\delta \neq 0$  which is a parabola.

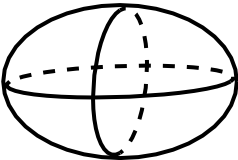
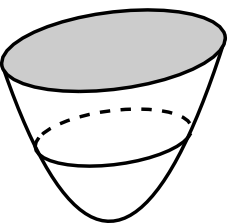
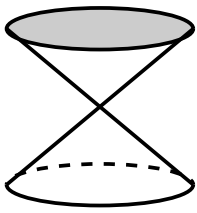
The three dimensional analogs of conic sections, surfaces in three dimensions given by quadratic equations, are called quadrics. An example is the sphere  $x^2 + y^2 + z^2 = 1$ . Here are some tables giving all of the quadric surfaces.

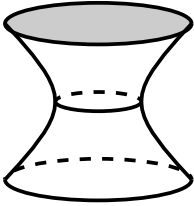
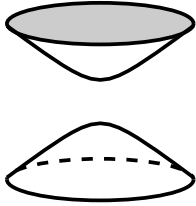
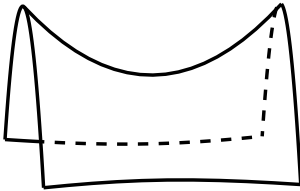
<sup>8</sup> It is outside our scope to prove this equivalence.

<sup>9</sup> Technically, we should also require that the constants  $A, B, C, D, E, F$ , are real numbers, that  $A, B, C$  are not all zero, that  $Q(x, y) = 0$  has more than one real solution, and that the polynomial can't be factored into the product of two polynomials of degree one.

<sup>10</sup> This statement can be justified using a linear algebra eigenvalue/eigenvector analysis. It is beyond what we can cover here, but is not too difficult for a standard linear algebra course.

name	elliptic cylinder	parabolic cylinder	hyperbolic cylinder	sphere
equation in standard form	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	$y = ax^2$	$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	$x^2 + y^2 + z^2 = r^2$
$x = \text{constant}$ cross-section	two lines	one line	two lines	circle
$y = \text{constant}$ cross-section	two lines	two lines	two lines	circle
$z = \text{constant}$ cross-section	ellipse	parabola	hyperbola	circle
sketch				

name	ellipsoid	elliptic paraboloid	elliptic cone
equation in standard form	$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z}{c}$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z^2}{c^2}$
$x = \text{constant}$ cross-section	ellipse	parabola	two lines if $x = 0$ hyperbola if $x \neq 0$
$y = \text{constant}$ cross-section	ellipse	parabola	two lines if $y = 0$ hyperbola if $y \neq 0$
$z = \text{constant}$ cross-section	ellipse	ellipse	ellipse
sketch			

name	hyperboloid of one sheet	hyperboloid of two sheets	hyperbolic paraboloid
equation in standard form	$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = -1$	$\frac{y^2}{b^2} - \frac{x^2}{a^2} = \frac{z}{c}$
$x = \text{constant}$ cross-section	hyperbola	hyperbola	parabola
$y = \text{constant}$ cross-section	hyperbola	hyperbola	ellipse
$z = \text{constant}$ cross-section	ellipse	ellipse	two lines if $z = 0$ hyperbola if $z \neq 0$
sketch			

Section A.3 of this work was adapted from Appendix G of [CLP 3 – Multivariable Calculus](#) by Feldman, Reznitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.4▲ Mixed Partial Derivatives

### A.4.1 ► Clairaut: The Proof of Theorem 2.2.5

#### ►►► Outline

Here is an outline of the proof of Theorem 2.2.5. The (numbered) details are in the subsection below.

Fix real numbers  $x_0$  and  $y_0$  and define

$$F(h, k) = \frac{1}{hk} [f(x_0 + h, y_0 + k) - f(x_0, y_0 + k) - f(x_0 + h, y_0) + f(x_0, y_0)].$$

We define  $F(h, k)$  in this way because both partial derivatives  $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$  and  $\frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$  are limits of  $F(h, k)$  as  $h, k \rightarrow 0$ . We show in item (1) in the details below that

$$\begin{aligned} \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0, y_0) &= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} F(h, k) \\ \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0, y_0) &= \lim_{h \rightarrow 0} \lim_{k \rightarrow 0} F(h, k) \end{aligned}$$

and therefore the partial derivatives  $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)$  and  $\frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$  are identical except for the order in which the limits are taken.



Now, by applying the Mean Value Theorem multiple times (see items (2) to (5) for more details) we get

$$\begin{aligned} F(h, k) &\stackrel{(2)}{=} \frac{1}{h} \left[ \frac{\partial f}{\partial y}(x_0 + h, y_0 + \theta_1 k) - \frac{\partial f}{\partial y}(x_0, y_0 + \theta_1 k) \right] \\ &\stackrel{(3)}{=} \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0 + \theta_2 h, y_0 + \theta_1 k) \\ F(h, k) &\stackrel{(4)}{=} \frac{1}{k} \left[ \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0 + k) - \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0) \right] \\ &\stackrel{(5)}{=} \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0 + \theta_4 k) \end{aligned}$$

for some numbers  $0 < \theta_1, \theta_2, \theta_3, \theta_4 < 1$ . All of the numbers  $\theta_1, \theta_2, \theta_3, \theta_4$  depend on  $x_0, y_0, h, k$ . Hence

$$\frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0 + \theta_2 h, y_0 + \theta_1 k) = \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0 + \theta_4 k)$$

for all  $h$  and  $k$ . Taking the limit  $(h, k) \rightarrow (0, 0)$  and using the assumed continuity of both partial derivatives at  $(x_0, y_0)$  gives

$$\lim_{(h,k) \rightarrow (0,0)} F(h, k) = \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0, y_0) = \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0, y_0)$$

as desired. To complete the proof we just have to justify the details (1), (2), (3), (4) and (5).

►►► The Details

(1) By definition,

$$\begin{aligned} \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0, y_0) &= \lim_{k \rightarrow 0} \frac{1}{k} \left[ \frac{\partial f}{\partial x}(x_0, y_0 + k) - \frac{\partial f}{\partial x}(x_0, y_0) \right] \\ &= \lim_{k \rightarrow 0} \frac{1}{k} \left[ \lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k)}{h} - \lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} \right] \\ &= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} \frac{f(x_0 + h, y_0 + k) - f(x_0, y_0 + k) - f(x_0 + h, y_0) + f(x_0, y_0)}{hk} \\ &= \lim_{k \rightarrow 0} \lim_{h \rightarrow 0} F(h, k) \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0, y_0) &= \lim_{h \rightarrow 0} \frac{1}{h} \left[ \frac{\partial f}{\partial y}(x_0 + h, y_0) - \frac{\partial f}{\partial y}(x_0, y_0) \right] \\ &= \lim_{h \rightarrow 0} \frac{1}{h} \left[ \lim_{k \rightarrow 0} \frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0)}{k} - \lim_{k \rightarrow 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k} \right] \\ &= \lim_{h \rightarrow 0} \lim_{k \rightarrow 0} \frac{f(x_0 + h, y_0 + k) - f(x_0 + h, y_0) - f(x_0, y_0 + k) + f(x_0, y_0)}{hk} \\ &= \lim_{h \rightarrow 0} \lim_{k \rightarrow 0} F(h, k) \end{aligned}$$

(2) The Mean Value Theorem (probably covered in your last calculus class) says that, for any differentiable function  $\varphi(x)$ ,

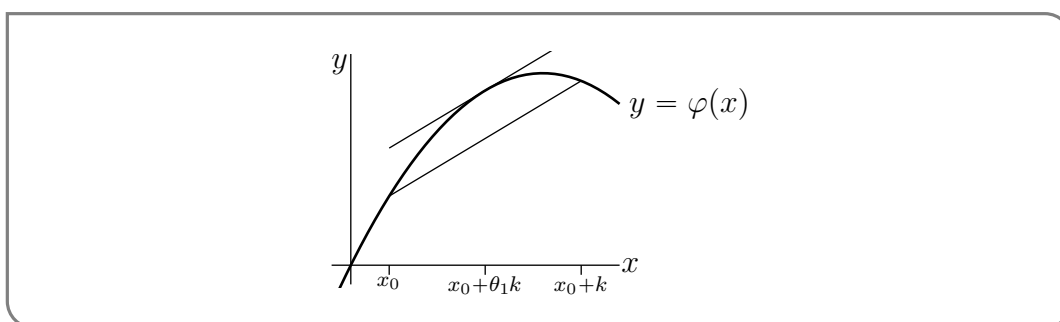
- the slope of the line joining the points  $(x_0, \varphi(x_0))$  and  $(x_0 + k, \varphi(x_0 + k))$  on the graph of  $\varphi$

is the same as

- the slope of the tangent to the graph at some point between  $x_0$  and  $x_0 + k$ .

That is, there is some  $0 < \theta_1 < 1$  such that

$$\frac{\varphi(x_0 + k) - \varphi(x_0)}{k} = \frac{d\varphi}{dx}(x_0 + \theta_1 k)$$



Applying this with  $x$  replaced by  $y$  and  $\varphi$  replaced by  $G(y) = f(x_0 + h, y) - f(x_0, y)$  gives

$$\begin{aligned} \frac{G(y_0 + k) - G(y_0)}{k} &= \frac{dG}{dy}(y_0 + \theta_1 k) \quad \text{for some } 0 < \theta_1 < 1 \\ &= \frac{\partial f}{\partial y}(x_0 + h, y_0 + \theta_1 k) - \frac{\partial f}{\partial y}(x_0, y_0 + \theta_1 k) \end{aligned}$$

Hence, for some  $0 < \theta_1 < 1$ ,

$$F(h, k) = \frac{1}{h} \left[ \frac{G(y_0 + k) - G(y_0)}{k} \right] = \frac{1}{h} \left[ \frac{\partial f}{\partial y}(x_0 + h, y_0 + \theta_1 k) - \frac{\partial f}{\partial y}(x_0, y_0 + \theta_1 k) \right]$$

(3) Define  $H(x) = \frac{\partial f}{\partial y}(x, y_0 + \theta_1 k)$ . By the Mean Value Theorem,

$$\begin{aligned} F(h, k) &= \frac{1}{h} [H(x_0 + h) - H(x_0)] \\ &= \frac{dH}{dx}(x_0 + \theta_2 h) \quad \text{for some } 0 < \theta_2 < 1 \\ &= \frac{\partial}{\partial x} \frac{\partial f}{\partial y}(x_0 + \theta_2 h, y_0 + \theta_1 k) \end{aligned}$$

(4) Define  $A(x) = f(x, y_0 + k) - f(x, y_0)$ . By the Mean Value Theorem,

$$\begin{aligned} F(h, k) &= \frac{1}{k} \left[ \frac{A(x_0 + h) - A(x_0)}{h} \right] \\ &= \frac{1}{k} \frac{dA}{dx}(x_0 + \theta_3 h) \quad \text{for some } 0 < \theta_3 < 1 \\ &= \frac{1}{k} \left[ \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0 + k) - \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0) \right] \end{aligned}$$

(5) Define  $B(y) = \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y)$ . By the Mean Value Theorem

$$\begin{aligned} F(h, k) &= \frac{1}{k} [B(y_0 + k) - B(y_0)] \\ &= \frac{dB}{dy}(y_0 + \theta_4 k) \quad \text{for some } 0 < \theta_4 < 1 \\ &= \frac{\partial}{\partial y} \frac{\partial f}{\partial x}(x_0 + \theta_3 h, y_0 + \theta_4 k) \end{aligned}$$

This completes the proof of Theorem 2.2.5.

Section A.4.1 of this work was adapted from Section 2.3.1 of [CLP 3 – Multivariable Calculus](#) by Feldman, Reznitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

#### A.4.2 ►► An Example of $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \neq \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$

In Theorem 2.2.5, we showed that  $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) = \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$  if the partial derivatives  $\frac{\partial^2 f}{\partial x \partial y}$  and  $\frac{\partial^2 f}{\partial y \partial x}$  exist and are continuous at  $(x_0, y_0)$ . Here is an example which shows that if the partial derivatives  $\frac{\partial^2 f}{\partial x \partial y}$  and  $\frac{\partial^2 f}{\partial y \partial x}$  are not continuous at  $(x_0, y_0)$ , then it is possible that  $\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \neq \frac{\partial^2 f}{\partial y \partial x}(x_0, y_0)$ .

Define

$$f(x, y) = \begin{cases} xy \frac{x^2 - y^2}{x^2 + y^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}$$

This function is continuous everywhere. Note that  $f(x, 0) = 0$  for all  $x$  and  $f(0, y) = 0$  for all  $y$ . We now compute the first order partial derivatives. For  $(x, y) \neq (0, 0)$ ,

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) &= y \frac{x^2 - y^2}{x^2 + y^2} + xy \frac{2x}{x^2 + y^2} - xy \frac{2x(x^2 - y^2)}{(x^2 + y^2)^2} = y \frac{x^2 - y^2}{x^2 + y^2} + xy \frac{4xy^2}{(x^2 + y^2)^2} \\ \frac{\partial f}{\partial y}(x, y) &= x \frac{x^2 - y^2}{x^2 + y^2} - xy \frac{2y}{x^2 + y^2} - xy \frac{2y(x^2 - y^2)}{(x^2 + y^2)^2} = x \frac{x^2 - y^2}{x^2 + y^2} - xy \frac{4yx^2}{(x^2 + y^2)^2} \end{aligned}$$

For  $(x, y) = (0, 0)$ ,

$$\begin{aligned}\frac{\partial f}{\partial x}(0, 0) &= \left[ \frac{d}{dx} f(x, 0) \right]_{x=0} = \left[ \frac{d}{dx} 0 \right]_{x=0} = 0 \\ \frac{\partial f}{\partial y}(0, 0) &= \left[ \frac{d}{dy} f(0, y) \right]_{y=0} = \left[ \frac{d}{dy} 0 \right]_{y=0} = 0\end{aligned}$$

By way of summary, the two first order partial derivatives are

$$\begin{aligned}f_x(x, y) &= \begin{cases} y \frac{x^2 - y^2}{x^2 + y^2} + \frac{4x^2 y^3}{(x^2 + y^2)^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases} \\ f_y(x, y) &= \begin{cases} x \frac{x^2 - y^2}{x^2 + y^2} - \frac{4x^3 y^2}{(x^2 + y^2)^2} & \text{if } (x, y) \neq (0, 0) \\ 0 & \text{if } (x, y) = (0, 0) \end{cases}\end{aligned}$$

Both  $\frac{\partial f}{\partial x}(x, y)$  and  $\frac{\partial f}{\partial y}(x, y)$  are continuous. Finally, we compute

$$\begin{aligned}\frac{\partial^2 f}{\partial x \partial y}(0, 0) &= \left[ \frac{d}{dx} f_y(x, 0) \right]_{x=0} = \lim_{h \rightarrow 0} \frac{1}{h} [f_y(h, 0) - f_y(0, 0)] = \lim_{h \rightarrow 0} \frac{1}{h} \left[ h \frac{h^2 - 0^2}{h^2 + 0^2} - 0 \right] = 1 \\ \frac{\partial^2 f}{\partial y \partial x}(0, 0) &= \left[ \frac{d}{dy} f_x(0, y) \right]_{y=0} = \lim_{k \rightarrow 0} \frac{1}{k} [f_x(0, k) - f_x(0, 0)] = \lim_{k \rightarrow 0} \frac{1}{k} \left[ k \frac{0^2 - k^2}{0^2 + k^2} - 0 \right] = -1\end{aligned}$$

Section A.4.2 of this work was adapted from Section 2.3.2 of [CLP 3 – Multivariable Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.5▲ The (multivariable) chain rule

You already routinely use the one dimensional chain rule

$$\frac{d}{dt} f(x(t)) = \frac{df}{dx}(x(t)) \frac{dx}{dt}(t)$$

in doing computations like

$$\frac{d}{dt} \sin(t^2) = \cos(t^2) 2t$$

In this example,  $f(x) = \sin(x)$  and  $x(t) = t^2$ .

We now generalize the chain rule to functions of more than one variable. For concreteness, we concentrate on the case in which all functions are functions of two variables. That is, we find the partial derivatives  $\frac{\partial F}{\partial s}$  and  $\frac{\partial F}{\partial t}$  of a function  $F(s, t)$  that is defined as a composition

$$F(s, t) = f(x(s, t), y(s, t))$$

We are using the name  $F$  for the new function  $F(s, t)$  as a reminder that it is closely related to, though not the same as, the function  $f(x, y)$ . The partial derivative  $\frac{\partial F}{\partial s}$  is the rate of

change of  $F$  when  $s$  is varied with  $t$  held constant. When  $s$  is varied, both the  $x$ -argument,  $x(s, t)$ , and the  $y$ -argument,  $y(s, t)$ , in  $f(x(s, t), y(s, t))$  vary. Consequently, the chain rule for  $f(x(s, t), y(s, t))$  is a sum of two terms — one resulting from the variation of the  $x$ -argument and the other resulting from the variation of the  $y$ -argument.

**Theorem A.5.1** (The Chain Rule).

Assume that all first order partial derivatives of  $f(x, y)$ ,  $x(s, t)$  and  $y(s, t)$  exist and are continuous. Then the same is true for  $F(s, t) = f(x(s, t), y(s, t))$  and

$$\begin{aligned}\frac{\partial F}{\partial s}(s, t) &= \frac{\partial f}{\partial x}(x(s, t), y(s, t)) \frac{\partial x}{\partial s}(s, t) + \frac{\partial f}{\partial y}(x(s, t), y(s, t)) \frac{\partial y}{\partial s}(s, t) \\ \frac{\partial F}{\partial t}(s, t) &= \frac{\partial f}{\partial x}(x(s, t), y(s, t)) \frac{\partial x}{\partial t}(s, t) + \frac{\partial f}{\partial y}(x(s, t), y(s, t)) \frac{\partial y}{\partial t}(s, t)\end{aligned}$$

We will give the proof of this theorem in §A.5.2, below. It is common to state this chain rule as

$$\begin{aligned}\frac{\partial F}{\partial s} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial s} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial s} \\ \frac{\partial F}{\partial t} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial t}\end{aligned}$$

That is, it is common to suppress the function arguments. But you should make sure that you understand what the arguments are before doing so.

Theorem A.5.1 is given for the case that  $F$  is the composition of a function of two variables,  $f(x, y)$ , with two functions,  $x(s, t)$  and  $y(s, t)$ , of two variables each. There is nothing magical about the number two. There are obvious variants for any numbers of variables. For example,

**Equation A.5.2.**

if  $F(t) = f(x(t), y(t), z(t))$ , then

$$\begin{aligned}\frac{dF}{dt}(t) &= \frac{\partial f}{\partial x}(x(t), y(t), z(t)) \frac{dx}{dt}(t) + \frac{\partial f}{\partial y}(x(t), y(t), z(t)) \frac{dy}{dt}(t) \\ &\quad + \frac{\partial f}{\partial z}(x(t), y(t), z(t)) \frac{dz}{dt}(t)\end{aligned}$$

and

**Equation A.5.3.**

if  $F(s, t) = f(x(s, t))$ , then

$$\frac{\partial F}{\partial t}(s, t) = \frac{df}{dx}(x(s, t)) \frac{\partial x}{\partial t}(s, t)$$

To give you an idea of how the proof of Theorem A.5.1 will go, we first review the proof of the familiar one dimensional chain rule.

### A.5.1 ▶ Review of the Proof of $\frac{d}{dt}f(x(t)) = \frac{df}{dx}(x(t)) \frac{dx}{dt}(t)$

As a warm up, let's review the proof of the one dimensional chain rule

$$\frac{d}{dt}f(x(t)) = \frac{df}{dx}(x(t)) \frac{dx}{dt}(t)$$

We wish to find the derivative of  $F(t) = f(x(t))$ . By definition

$$\begin{aligned} F'(t) &= \lim_{h \rightarrow 0} \frac{F(t+h) - F(t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x(t+h)) - f(x(t))}{h} \end{aligned}$$

Notice that the numerator is the difference of  $f(x)$  evaluated at two nearby values of  $x$ , namely  $x_1 = x(t+h)$  and  $x_0 = x(t)$ . The Mean Value Theorem is a good tool for studying the difference in the values of  $f(x)$  at two nearby points. Recall that the Mean Value Theorem says that, for any given  $x_0$  and  $x_1$ , there exists an (in general unknown)  $c$  between them so that

$$f(x_1) - f(x_0) = f'(c) (x_1 - x_0)$$

For this proof, we choose  $x_0 = x(t)$  and  $x_1 = x(t+h)$ . The Mean Value Theorem tells us that there exists a  $c_h$  so that

$$f(x(t+h)) - f(x(t)) = f(x_1) - f(x_0) = f'(c_h) [x(t+h) - x(t)]$$

We have put the subscript  $h$  on  $c_h$  to emphasise that  $c_h$ , which is between  $x_0 = x(t)$  and  $x_1 = x(t+h)$ , may depend on  $h$ . Now since  $c_h$  is trapped between  $x(t)$  and  $x(t+h)$  and since  $x(t+h) \rightarrow x(t)$  as  $h \rightarrow 0$ , we have that  $c_h$  must also tend to  $x(t)$  as  $h \rightarrow 0$ . Plugging this into the definition of  $F'(t)$ ,

$$\begin{aligned} F'(t) &= \lim_{h \rightarrow 0} \frac{f(x(t+h)) - f(x(t))}{h} \\ &= \lim_{h \rightarrow 0} \frac{f'(c_h) [x(t+h) - x(t)]}{h} \\ &= \lim_{h \rightarrow 0} f'(c_h) \lim_{h \rightarrow 0} \frac{x(t+h) - x(t)}{h} \\ &= f'(x(t)) x'(t) \end{aligned}$$

as desired.

### A.5.2 ▶ Proof of Theorem A.5.1

We'll now prove the formula for  $\frac{\partial}{\partial s}f(x(s,t), y(s,t))$  that is given in Theorem A.5.1. The proof uses the same ideas as the proof of the one variable chain rule, that we have just reviewed.

We wish to find the partial derivative with respect to  $s$  of  $F(s, t) = f(x(s, t), y(s, t))$ .  
By definition

$$\begin{aligned}\frac{\partial F}{\partial s}(s, t) &= \lim_{h \rightarrow 0} \frac{F(s+h, t) - F(s, t)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x(s+h, t), y(s+h, t)) - f(x(s, t), y(s, t))}{h}\end{aligned}$$

The numerator is the difference of  $f(x, y)$  evaluated at two nearby values of  $(x, y)$ , namely  $(x_1, y_1) = (x(s+h, t), y(s+h, t))$  and  $(x_0, y_0) = (x(s, t), y(s, t))$ . In going from  $(x_0, y_0)$  to  $(x_1, y_1)$ , both the  $x$  and  $y$ -coordinates change. By adding and subtracting we can separate the change in the  $x$ -coordinate from the change in the  $y$ -coordinate.

$$f(x_1, y_1) - f(x_0, y_0) = \{f(x_1, y_1) - f(x_0, y_1)\} + \{f(x_0, y_1) - f(x_0, y_0)\}$$

The first half,  $\{f(x_1, y_1) - f(x_0, y_1)\}$ , has the same  $y$  argument in both terms and so is the difference of the function of one variable  $g(x) = f(x, y_1)$  (viewing  $y_1$  just as a constant) evaluated at the two nearby values,  $x_0, x_1$ , of  $x$ . Consequently, we can make use of the Mean Value Theorem as we did in §A.5.1 above. There is a  $c_{x,h}$  between  $x_0 = x(s, t)$  and  $x_1 = x(s+h, t)$  such that

$$\begin{aligned}f(x_1, y_1) - f(x_0, y_1) &= g(x_1) - g(x_0) = g'(c_{x,h})[x_1 - x_0] = \frac{\partial f}{\partial x}(c_{x,h}, y_1) [x_1 - x_0] \\ &= \frac{\partial f}{\partial x}(c_{x,h}, y(s+h, t)) [x(s+h, t) - x(s, t)]\end{aligned}$$

We have introduced the two subscripts in  $c_{x,h}$  to remind ourselves that it may depend on  $h$  and that it lies between the two  $x$ -values  $x_0$  and  $x_1$ .

Similarly, the second half,  $\{f(x_0, y_1) - f(x_0, y_0)\}$ , is the difference of the function of one variable  $h(y) = f(x_0, y)$  (viewing  $x_0$  just as a constant) evaluated at the two nearby values,  $y_0, y_1$ , of  $y$ . So, by the mean value theorem,

$$\begin{aligned}f(x_0, y_1) - f(x_0, y_0) &= h(y_1) - h(y_0) = h'(c_{y,h})[y_1 - y_0] = \frac{\partial f}{\partial y}(x_0, c_{y,h}) [y_1 - y_0] \\ &= \frac{\partial f}{\partial y}(x(s, t), c_{y,h}) [y(s+h, t) - y(s, t)]\end{aligned}$$

for some (unknown)  $c_{y,h}$  between  $y_0 = y(s, t)$  and  $y_1 = y(s+h, t)$ . Again, the two subscripts in  $c_{y,h}$  remind ourselves that it may depend on  $h$  and that it lies between the two  $y$ -values  $y_0$  and  $y_1$ . So, noting that, as  $h$  tends to zero,  $c_{x,h}$ , which is trapped between  $x(s, t)$  and  $x(s+h, t)$ , must tend to  $x(s, t)$ , and  $c_{y,h}$ , which is trapped between  $y(s, t)$  and

$y(s + h, t)$ , must tend to  $y(s, t)$ ,

$$\begin{aligned} \frac{\partial F}{\partial s}(s, t) &= \lim_{h \rightarrow 0} \frac{f(x(s+h, t), y(s+h, t)) - f(x(s, t), y(s, t))}{h} \\ &= \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial x}(c_{x,h}, y(s+h, t)) [x(s+h, t) - x(s, t)]}{h} \\ &\quad + \lim_{h \rightarrow 0} \frac{\frac{\partial f}{\partial y}(x(s, t), c_{y,h}) [y(s+h, t) - y(s, t)]}{h} \\ &= \lim_{h \rightarrow 0} \frac{\partial f}{\partial x}(c_{x,h}, y(s+h, t)) \lim_{h \rightarrow 0} \frac{x(s+h, t) - x(s, t)}{h} \\ &\quad + \lim_{h \rightarrow 0} \frac{\partial f}{\partial y}(x(s, t), c_{y,h}) \lim_{h \rightarrow 0} \frac{y(s+h, t) - y(s, t)}{h} \\ &= \frac{\partial f}{\partial x}(x(s, t), y(s, t)) \frac{\partial x}{\partial s}(s, t) + \frac{\partial f}{\partial y}(x(s, t), y(s, t)) \frac{\partial y}{\partial s}(s, t) \end{aligned}$$

We can of course follow the same procedure to evaluate the partial derivative with respect to  $t$ . This concludes the proof of Theorem A.5.1.

Example A.5.4 (Implicit Differentiation on Level Curves)

Level curves of the surface  $z = f(x, y)$  are points  $(x, y)$  such that  $f(x, y) = z_0$ , where  $z_0$  is some fixed constant. We can think of level curves as existing in an  $xy$ -plane by ignoring the  $z$  coordinate (which, remember, is constant). Consider a point  $(a, b, z_0)$  on a level curve  $f(x, y) = z_0$ . In the two-dimensional view, near  $(a, b)$  we can think of  $y$  as a function of  $x$ : if  $x$  moves a little but, then  $y$  changes as well to “compensate” and maintain the constant  $z$ -value. So, we can write  $y(x)$  to remember that  $y$  depends on  $x$  in this situation.

$$z_0 = f(x, y(x))$$

Now we can think about the single-variable function  $g(x) = f(x, y(x))$ . Since this function is equal to the constant value  $z_0$ , its derivative is zero. Then, using the chain rule:

$$\begin{aligned} 0 = g'(x) &= \frac{d}{dx} [f(x, y(x))] = \frac{\partial f}{\partial x} \frac{dx}{dx} + \frac{\partial f}{\partial y} \frac{dy}{dx} \\ &= f_x \cdot 1 + f_y \cdot \frac{dy}{dx} \end{aligned}$$

If  $f_y \neq 0$ , then

$$\frac{dy}{dx} = -\frac{f_x}{f_y}$$

What we’ve proved is the following.



**Theorem A.5.5.**

The derivative of the curve in the  $xy$  plane that is implicitly defined by the equation

$$z_0 = f(x, y)$$

for some constant  $z_0$  and some differentiable function  $f(x, y)$  is

$$\frac{dy}{dx} = -\frac{f_x}{f_y}$$

as long as  $f_y \neq 0$ .

Example A.5.4

**Definition A.5.6.**

The vector  $\langle f_x(a, b), f_y(a, b) \rangle$  is denoted  $\nabla f(a, b)$  and is called “the **gradient** of the function  $f$  at the point  $(a, b)$ ”.

**Corollary A.5.7.**

Let  $f(x, y)$  be a function whose partial derivatives exist.  $\nabla f(a, b)$  is perpendicular to the level curve  $f(x, y) = f(a, b)$  at  $(a, b)$  as long as  $\nabla f(a, b) \neq \mathbf{0}$ .

*Proof.* First, suppose  $f_y(a, b) \neq 0$ . Using Theorem A.5.5, the line tangent to the level curve has slope (in the  $xy$  plane)  $-\frac{f_x}{f_y}$ . So, one vector in the direction tangent to the level curve is  $\langle -f_y, f_x \rangle$ . Then

$$\langle -f_y, f_x \rangle \cdot \langle f_x, f_y \rangle = 0$$

so  $\langle -f_y, f_x \rangle$  and  $\nabla f = \langle f_x, f_y \rangle$  are perpendicular.

Second, consider the case  $f_y(a, b) = 0$ . In this case, at  $(a, b)$  the level curve has a vertical tangent line. If  $\nabla f(a, b) \neq \mathbf{0}$ , then  $f_x(a, b) \neq 0$ , so the gradient  $\nabla f(a, b) = \langle f_x, 0 \rangle$  is horizontal.  $\square$

Section A.5 of this work was adapted from Section 2.4 of **CLP 3 – Multivariable Calculus** by Feldman, Reznitzer, and Yeager under a **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license**.

## A.6<sup>▲</sup> Lagrange Multipliers: Proof of Theorem 2.5.2

First, some intuition. When we talk about derivatives on a surface, we need to think about the derivatives in a particular direction.<sup>11</sup> Consider in particular the surface formed by all points  $(x, y)$  such that  $f(x, y) = z$ , for some function  $f(x, y)$ . The directions giving zero rate of increase are those that keep you on a level curve. By Corollary A.5.7, those directions are perpendicular to  $\nabla f(a, b)$ .

The corresponding statement in three dimensions is that  $\nabla F(a, b, c)$  is perpendicular to the level surface  $F(x, y, z) = F(a, b, c)$  at  $(a, b, c)$ . Hence a good way to find a vector normal to the surface  $F(x, y, z) = 0$  at the point  $(a, b, c)$  is to compute the gradient  $\nabla F(a, b, c)$ .

### Theorem A.6.1 (Lagrange Multipliers).

Let  $f(x, y, z)$  and  $g(x, y, z)$  have continuous first partial derivatives in a region of  $\mathbb{R}^3$  that contains the surface  $S$  given by the equation  $g(x, y, z) = 0$ . Further assume that  $\nabla g(x, y, z) \neq \mathbf{0}$  on  $S$ .

If  $f$ , restricted to the surface  $S$ , has a local extreme value at the point  $(a, b, c)$  on  $S$ , then there is a real number  $\lambda$  such that

$$\nabla f(a, b, c) = \lambda \nabla g(a, b, c)$$

that is

$$\begin{aligned} f_x(a, b, c) &= \lambda g_x(a, b, c) \\ f_y(a, b, c) &= \lambda g_y(a, b, c) \\ f_z(a, b, c) &= \lambda g_z(a, b, c) \end{aligned}$$

The number  $\lambda$  is called a *Lagrange multiplier*.

*Proof.* Suppose that  $(a, b, c)$  is a point of  $S$  and that  $f(x, y, z) \geq f(a, b, c)$  for all points  $(x, y, z)$  on  $S$  that are close to  $(a, b, c)$ . That is  $(a, b, c)$  is a local minimum for  $f$  on  $S$ . Of course the argument for a local maximum is virtually identical.

Imagine that we go for a walk on  $S$ , with the time  $t$  running, say, from  $t = -1$  to  $t = +1$  and that at time  $t = 0$  we happen to be exactly at  $(a, b, c)$ . Let's say that our position is  $(x(t), y(t), z(t))$  at time  $t$ . Write

$$F(t) = f(x(t), y(t), z(t))$$

So  $F(t)$  is the value of  $f$  that we see on our walk at time  $t$ . Then for all  $t$  close to 0,  $(x(t), y(t), z(t))$  is close to  $(x(0), y(0), z(0)) = (a, b, c)$  so that

$$F(0) = f(x(0), y(0), z(0)) = f(a, b, c) \leq f(x(t), y(t), z(t)) = F(t)$$

for all  $t$  close to zero. So  $F(t)$  has a local minimum at  $t = 0$  and consequently  $F'(0) = 0$ .

11 If you're walking along hilly terrain, changing direction can cause you to change from going uphill to downhill. Direction definitely matters!

By the chain rule, Theorem A.5.1,

$$\begin{aligned} F'(0) &= \left. \frac{d}{dt} f(x(t), y(t), z(t)) \right|_{t=0} \\ &= f_x(a, b, c)x'(0) + f_y(a, b, c)y'(0) + f_z(a, b, c)z'(0) = 0 \end{aligned} \quad (*)$$

We may rewrite this as a dot product:

$$\begin{aligned} 0 &= F'(0) = \nabla f(a, b, c) \cdot \langle x'(0), y'(0), z'(0) \rangle \\ \implies \nabla f(a, b, c) &\perp \langle x'(0), y'(0), z'(0) \rangle \end{aligned}$$

This is true for all paths on  $S$  that pass through  $(a, b, c)$  at time 0. In particular it is true for all vectors  $\langle x'(0), y'(0), z'(0) \rangle$  that are tangent to  $S$  at  $(a, b, c)$ . So  $\nabla f(a, b, c)$  is perpendicular to  $S$  at  $(a, b, c)$ .

But we already know, by the three-dimensional analogue to Corollary A.5.7, that  $\nabla g(a, b, c)$  is also perpendicular to  $S$  at  $(a, b, c)$ . So  $\nabla f(a, b, c)$  and  $\nabla g(a, b, c)$  have to be parallel vectors. That is,

$$\nabla f(a, b, c) = \lambda \nabla g(a, b, c)$$

for some number  $\lambda$ . That's the Lagrange multiplier rule of our theorem.  $\square$

Section A.6 of this work was adapted from Section 2.10 of [CLP 3 – Multivariable Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.7▲ A More Rigorous Area Computation

In Example 3.1.1 above we considered the area of the region  $\{ (x, y) \mid 0 \leq y \leq e^x, 0 \leq x \leq 1 \}$ . We approximated that area by the area of a union of  $n$  thin rectangles. We then claimed that upon taking the number of rectangles to infinity, the approximation of the area became the exact area. However we did not justify the claim. The purpose of this optional section is to make that calculation rigorous.

The broad set-up is the same. We divide the region up into  $n$  vertical strips, each of width  $1/n$  and we then approximate those strips by rectangles. However rather than an uncontrolled approximation, we construct two sets of rectangles — one set always smaller than the original area and one always larger. This then gives us lower and upper bounds on the area of the region. Finally we make use of the squeeze theorem<sup>12</sup> to establish the result.

- To find our upper and lower bounds we make use of the fact that  $e^x$  is an increasing function. We know this because the derivative  $\frac{d}{dx}e^x = e^x$  is always positive. Consequently, the smallest and largest values of  $e^x$  on the interval  $a \leq x \leq b$  are  $e^a$  and  $e^b$ , respectively.

12 Recall that if we have 3 functions  $f(x), g(x), h(x)$  that satisfy  $f(x) \leq g(x) \leq h(x)$  and we know that  $\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x) = L$  exists and is finite, then the *Squeeze Theorem* tells us that  $\lim_{x \rightarrow a} g(x) = L$ .

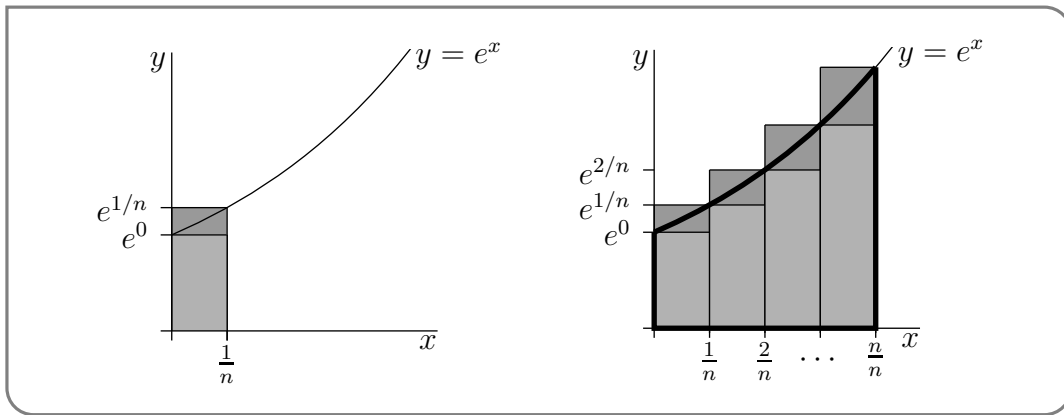
- In particular, for  $0 \leq x \leq 1/n$ ,  $e^x$  takes values only between  $e^0$  and  $e^{1/n}$ . As a result, the first strip

$$\{ (x, y) \mid 0 \leq x \leq 1/n, 0 \leq y \leq e^x \}$$

- contains the rectangle of  $0 \leq x \leq 1/n, 0 \leq y \leq e^0$  (the lighter rectangle in the figure on the left below) and
- is contained in the rectangle  $0 \leq x \leq 1/n, 0 \leq y \leq e^{1/n}$  (the largest rectangle in the figure on the left below).

Hence

$$\frac{1}{n}e^0 \leq \text{Area}\{ (x, y) \mid 0 \leq x \leq 1/n, 0 \leq y \leq e^x \} \leq \frac{1}{n}e^{1/n}$$



- Similarly, for the second, third, ..., last strips, as in the figure on the right above,

$$\begin{aligned} \frac{1}{n}e^{1/n} &\leq \text{Area}\{ (x, y) \mid 1/n \leq x \leq 2/n, 0 \leq y \leq e^x \} &&\leq \frac{1}{n}e^{2/n} \\ \frac{1}{n}e^{2/n} &\leq \text{Area}\{ (x, y) \mid 2/n \leq x \leq 3/n, 0 \leq y \leq e^x \} &&\leq \frac{1}{n}e^{3/n} \\ &\vdots &&\vdots \\ \frac{1}{n}e^{(n-1)/n} &\leq \text{Area}\{ (x, y) \mid (n-1)/n \leq x \leq n/n, 0 \leq y \leq e^x \} &&\leq \frac{1}{n}e^{n/n} \end{aligned}$$

- Adding these  $n$  inequalities together gives

$$\begin{aligned} \frac{1}{n} \left( 1 + e^{1/n} + \dots + e^{(n-1)/n} \right) &\leq \text{Area}\{ (x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq e^x \} \\ &\leq \frac{1}{n} \left( e^{1/n} + e^{2/n} + \dots + e^{n/n} \right) \end{aligned}$$

- We can then recycle equation (3.1.3) with  $r = e^{1/n}$ , so that  $r^n = (e^{1/n})^n = e$ . Thus we have

$$\frac{1}{n} \frac{e - 1}{e^{1/n} - 1} \leq \text{Area}\{ (x, y) \mid 0 \leq x \leq 1, 0 \leq y \leq e^x \} \leq \frac{1}{n} e^{1/n} \frac{e - 1}{e^{1/n} - 1}$$

where we have used the fact that the upper bound is a simple multiple of the lower bound:

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

- We now apply the Squeeze Theorem to the above inequalities. In particular, the limits of the lower and upper bounds are

$$\lim_{n \rightarrow \infty} \frac{1}{n} \frac{e-1}{e^{1/n}-1} = (e-1) \lim_{X=1/n \rightarrow 0} \frac{X}{e^X-1} = e-1$$

(by l'Hôpital's rule) and

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} e^{1/n} \frac{e-1}{e^{1/n}-1} &= (e-1) \lim_{X=1/n \rightarrow 0} \frac{Xe^X}{e^X-1} \\ &= (e-1) \lim_{X \rightarrow 0} e^X \cdot \lim_{X \rightarrow 0} \frac{X}{e^X-1} \\ &= (e-1) \cdot 1 \cdot 1 \end{aligned}$$

Thus, since the exact area is trapped between the lower and upper bounds, the squeeze theorem then implies that

$$\text{Exact area} = e - 1.$$

Section A.7 of this work was adapted from Section 1.1.1 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.8▲ Careful Definition of the Integral

In this optional section we give a more mathematically rigorous definition of the definite integral  $\int_a^b f(x)dx$ . Some textbooks use a sneakier, but equivalent, definition. The integral will be defined as the limit of a family of approximations to the area between the graph of  $y = f(x)$  and the  $x$ -axis, with  $x$  running from  $a$  to  $b$ . We will then show conditions under which this limit is guaranteed to exist. We should state up front that these conditions are more restrictive than is strictly necessary — this is done so as to keep the proof accessible.

The family of approximations needed is slightly more general than that used to define Riemann sums in the previous sections, though it is quite similar. The main difference is that we do not require that all the subintervals have the same size.

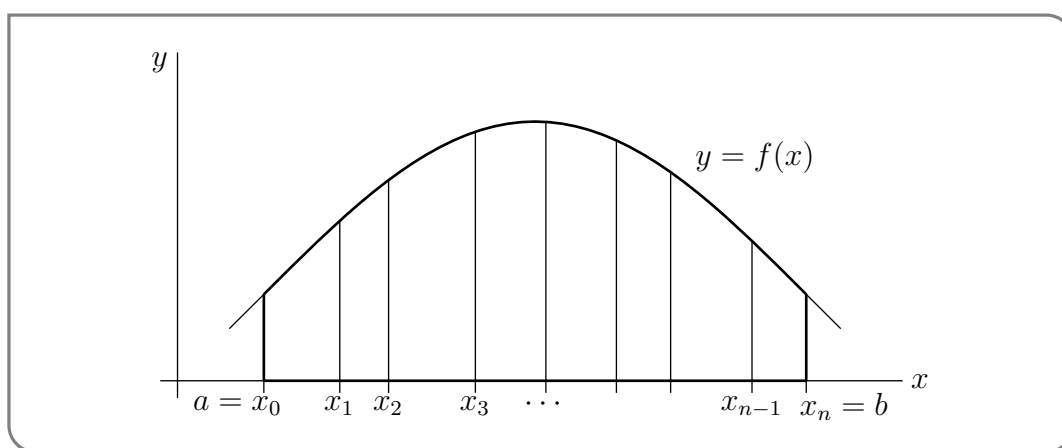
- We start by selecting a positive integer  $n$ . As was the case previously, this will be the number of subintervals used in the approximation and eventually we will take the limit as  $n \rightarrow \infty$ .

- Now subdivide the interval from  $a$  to  $b$  into  $n$  subintervals by selecting  $n + 1$  values of  $x$  that obey

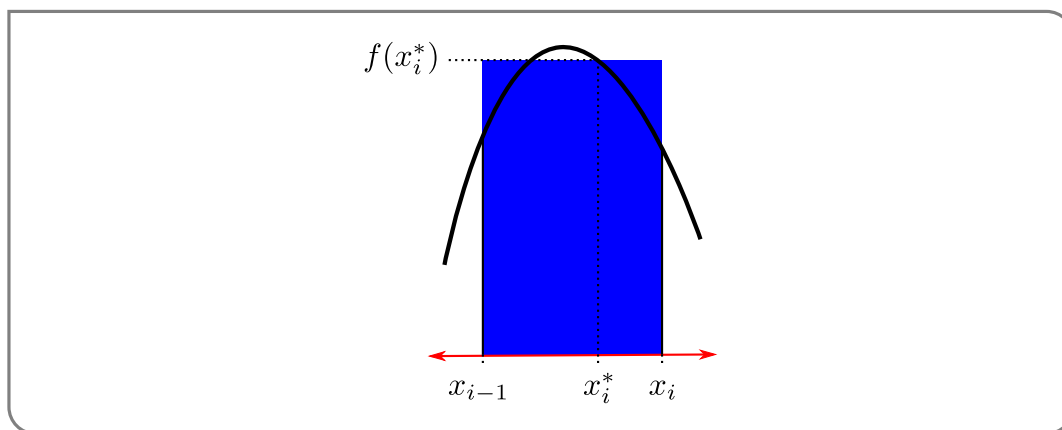
$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

The subinterval number  $i$  runs from  $x_{i-1}$  to  $x_i$ . This formulation does not require the subintervals to have the same size. However we will eventually require that the widths of the subintervals shrink towards zero as  $n \rightarrow \infty$ .

- Then for each subinterval we select a value of  $x$  in that interval. That is, for  $i = 1, 2, \dots, n$ , choose  $x_i^*$  satisfying  $x_{i-1} \leq x_i^* \leq x_i$ . We will use these values of  $x$  to help approximate  $f(x)$  on each subinterval.
- The area between the graph of  $y = f(x)$  and the  $x$ -axis, with  $x$  running from  $x_{i-1}$



to  $x_i$ , i.e. the contribution,  $\int_{x_{i-1}}^{x_i} f(x) dx$ , from interval number  $i$  to the integral, is approximated by the area of a rectangle. The rectangle has width  $x_i - x_{i-1}$  and height  $f(x_i^*)$ .



- Thus the approximation to the integral, using all  $n$  subintervals, is

$$\int_a^b f(x) dx \approx f(x_1^*)[x_1 - x_0] + f(x_2^*)[x_2 - x_1] + \cdots + f(x_n^*)[x_n - x_{n-1}]$$

- Of course every different choice of  $n$  and  $x_1, x_2, \dots, x_{n-1}$  and  $x_1^*, x_2^*, \dots, x_n^*$  gives a different approximation. So to simplify the discussion that follows, let us denote a particular choice of all these numbers by  $\mathbb{P}$ :

$$\mathbb{P} = (n, x_1, x_2, \dots, x_{n-1}, x_1^*, x_2^*, \dots, x_n^*).$$

Similarly let us denote the resulting approximation by  $\mathcal{I}(\mathbb{P})$ :

$$\mathcal{I}(\mathbb{P}) = f(x_1^*)[x_1 - x_0] + f(x_2^*)[x_2 - x_1] + \dots + f(x_n^*)[x_n - x_{n-1}]$$

- We claim that, for any reasonable<sup>13</sup> function  $f(x)$ , if you take any reasonable<sup>14</sup> sequence of these approximations you always get the exactly the same limiting value. We define  $\int_a^b f(x)dx$  to be this limiting value.
- Let's be more precise. We can take the limit of these approximations in two equivalent ways. Above we did this by taking the number of subintervals  $n$  to infinity. When we did this, the width of all the subintervals went to zero. With the formulation we are now using, simply taking the number of subintervals to be very large does not imply that they will all shrink in size. We could have one very large subinterval and a large number of tiny ones. Thus we take the limit we need by taking the width of the subintervals to zero. So for any choice  $\mathbb{P}$ , we define

$$M(\mathbb{P}) = \max \{x_1 - x_0, x_2 - x_1, \dots, x_n - x_{n-1}\}$$

that is the maximum width of the subintervals used in the approximation determined by  $\mathbb{P}$ . By forcing the maximum width to go to zero, the widths of all the subintervals go to zero.

- We then define the definite integral as the limit

$$\int_a^b f(x)dx = \lim_{M(\mathbb{P}) \rightarrow 0} \mathcal{I}(\mathbb{P}).$$

Of course, one is now left with the question of determining when the above limit exists. A proof of the very general conditions which guarantee existence of this limit is beyond the scope of this course, so we instead give a weaker result (with stronger conditions) which is far easier to prove.

For the rest of this section, assume

- that  $f(x)$  is continuous for  $a \leq x \leq b$ ,
- that  $f(x)$  is differentiable for  $a < x < b$ , and
- that  $f'(x)$  is bounded — ie  $|f'(x)| \leq F$  for some constant  $F$ .

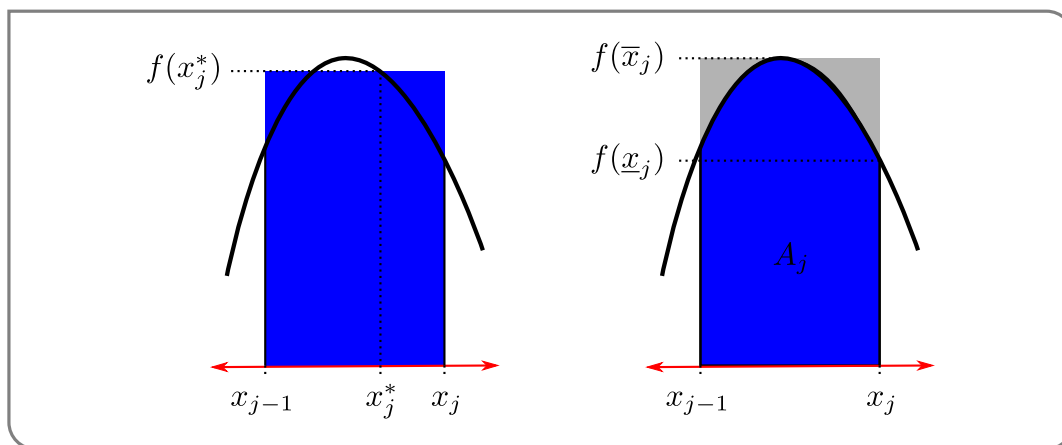
13 We'll be more precise about what "reasonable" means shortly.

14 Again, we'll explain this "reasonable" shortly

We will now show that, under these hypotheses, as  $M(\mathbb{P})$  approaches zero,  $\mathcal{I}(\mathbb{P})$  always approaches the area,  $A$ , between the graph of  $y = f(x)$  and the  $x$ -axis, with  $x$  running from  $a$  to  $b$ .

These assumptions are chosen to make the argument particularly transparent. With a little more work one can weaken the hypotheses considerably. We are cheating a little by implicitly assuming that the area  $A$  exists. In fact, one can adjust the argument below to remove this implicit assumption.

- Consider  $A_j$ , the part of the area coming from  $x_{j-1} \leq x \leq x_j$ .



We have approximated this area by  $f(x_j^*)[x_j - x_{j-1}]$  (see figure left).

- Let  $f(\bar{x}_j)$  and  $f(\underline{x}_j)$  be the largest and smallest values<sup>15</sup> of  $f(x)$  for  $x_{j-1} \leq x \leq x_j$ . Then the true area is bounded by

$$f(\underline{x}_j)[x_j - x_{j-1}] \leq A_j \leq f(\bar{x}_j)[x_j - x_{j-1}].$$

(see figure right).

- Now since  $f(\underline{x}_j) \leq f(x_j^*) \leq f(\bar{x}_j)$ , we also know that

$$f(\underline{x}_j)[x_j - x_{j-1}] \leq f(x_j^*)[x_j - x_{j-1}] \leq f(\bar{x}_j)[x_j - x_{j-1}].$$

- So both the true area,  $A_j$ , and our approximation of that area  $f(x_j^*)[x_j - x_{j-1}]$  have to lie between  $f(\bar{x}_j)[x_j - x_{j-1}]$  and  $f(\underline{x}_j)[x_j - x_{j-1}]$ . Combining these bounds we have that the difference between the true area and our approximation of that area is bounded by

$$|A_j - f(x_j^*)[x_j - x_{j-1}]| \leq [f(\bar{x}_j) - f(\underline{x}_j)] \cdot [x_j - x_{j-1}].$$

(To see this think about the smallest the true area can be and the largest our approximation can be and vice versa.)

<sup>15</sup> Here we are using the Extreme Value Theorem — its proof is beyond the scope of this course. The theorem says that any continuous function on a closed interval must attain a minimum and maximum at least once. In this situation this implies that for any continuous function  $f(x)$ , there are  $x_{j-1} \leq \bar{x}_j, \underline{x}_j \leq x_j$  such that  $f(\underline{x}_j) \leq f(x) \leq f(\bar{x}_j)$  for all  $x_{j-1} \leq x \leq x_j$ .



- Now since our function,  $f(x)$  is differentiable we can apply one of the main theorems we learned in first-semester calculus — the Mean Value Theorem<sup>16</sup>. The MVT implies that there exists a  $c$  between  $\underline{x}_j$  and  $\bar{x}_j$  such that

$$f(\bar{x}_j) - f(\underline{x}_j) = f'(c) \cdot [\bar{x}_j - \underline{x}_j]$$

- By the assumption that  $|f'(x)| \leq F$  for all  $x$  and the fact that  $\underline{x}_j$  and  $\bar{x}_j$  must both be between  $x_{j-1}$  and  $x_j$

$$|f(\bar{x}_j) - f(\underline{x}_j)| \leq F \cdot |\bar{x}_j - \underline{x}_j| \leq F \cdot [x_j - x_{j-1}]$$

Hence the error in this part of our approximation obeys

$$|A_j - f(x_j^*)[x_j - x_{j-1}]| \leq F \cdot [x_j - x_{j-1}]^2.$$

- That was just the error in approximating  $A_j$ . Now we bound the total error by combining the errors from approximating on all the subintervals. This gives

$$\begin{aligned} |A - \mathcal{I}(\mathbb{P})| &= \left| \sum_{j=1}^n A_j - \sum_{j=1}^n f(x_j^*)[x_j - x_{j-1}] \right| \\ &= \left| \sum_{j=1}^n \left( A_j - f(x_j^*)[x_j - x_{j-1}] \right) \right| && \text{triangle inequality} \\ &\leq \sum_{j=1}^n \left| A_j - f(x_j^*)[x_j - x_{j-1}] \right| \\ &\leq \sum_{j=1}^n F \cdot [x_j - x_{j-1}]^2 && \text{from above} \end{aligned}$$

Now do something a little sneaky. Replace one of these factors of  $[x_j - x_{j-1}]$  (which is just the width of the  $j^{\text{th}}$  subinterval) by the maximum width of the subintervals:

$$\begin{aligned} &\leq \sum_{j=1}^n F \cdot M(\mathbb{P}) \cdot [x_j - x_{j-1}] && F \text{ and } M(\mathbb{P}) \text{ are constant} \\ &\leq F \cdot M(\mathbb{P}) \cdot \sum_{j=1}^n [x_j - x_{j-1}] && \text{sum is total width} \\ &= F \cdot M(\mathbb{P}) \cdot (b - a). \end{aligned}$$

---

16 Recall that the Mean Value Theorem states that for a function continuous on  $[a, b]$  and differentiable on  $(a, b)$ , there exists a number  $c$  between  $a$  and  $b$  so that

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

- Since  $a$ ,  $b$  and  $F$  are fixed, this tends to zero as the maximum rectangle width  $M(\mathbb{P})$  tends to zero.

Thus, we have proven

**Theorem A.8.1.**

Assume that  $f(x)$  is continuous for  $a \leq x \leq b$ , and is differentiable for all  $a < x < b$  with  $|f'(x)| \leq F$ , for some constant  $F$ . Then, as the maximum rectangle width  $M(\mathbb{P})$  tends to zero,  $\mathcal{I}(\mathbb{P})$  always converges to  $A$ , the area between the graph of  $y = f(x)$  and the  $x$ -axis, with  $x$  running from  $a$  to  $b$ .

Section A.8 of this work was adapted from Section 1.1.6 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.9▲ Integrating $\sec x$ and $\csc x$

The antiderivatives of secant and (to a lesser extent) cosecant come up commonly enough that we include their computations here.

In a traditional integral calculus course, students may learn general methods to compute integrals of the form  $\int \sec^n x \tan^m x dx$ . As of 2021, Math 105 no longer includes this content in its syllabus.

Example A.9.1 ( $\int \sec x dx$  — by trickery)

*Solution.* There is a very sneaky trick to compute this integral.

- The standard trick for this integral is to multiply the integrand by  $1 = \frac{\sec x + \tan x}{\sec x + \tan x}$

$$\sec x = \sec x \frac{\sec x + \tan x}{\sec x + \tan x} = \frac{\sec^2 x + \sec x \tan x}{\sec x + \tan x}$$

- Notice now that the numerator of this expression is exactly the derivative its denominator. Hence we can substitute  $u = \sec x + \tan x$  and  $du = (\sec x \tan x + \sec^2 x) dx$ .
- Hence

$$\begin{aligned} \int \sec x dx &= \int \sec x \frac{\sec x + \tan x}{\sec x + \tan x} dx = \int \frac{\sec^2 x + \sec x \tan x}{\sec x + \tan x} dx \\ &= \int \frac{1}{u} du \\ &= \ln |u| + C \\ &= \ln |\sec x + \tan x| + C \end{aligned}$$

- The above trick appears both totally unguessable and very hard to remember. Fortunately, there is a simple way<sup>17</sup> to recover the trick. Here it is.
  - The goal is to guess a function whose derivative is  $\sec x$ .
  - So get out a table of derivatives and look for functions whose derivatives at least contain  $\sec x$ . There are two:

$$\begin{aligned}\frac{d}{dx} \tan x &= \sec^2 x \\ \frac{d}{dx} \sec x &= \tan x \sec x\end{aligned}$$

- Notice that if we add these together we get

$$\frac{d}{dx} (\sec x + \tan x) = (\sec x + \tan x) \sec x \implies \frac{\frac{d}{dx} (\sec x + \tan x)}{\sec x + \tan x} = \sec x$$

- We've done it! The right hand side is  $\sec x$  and the left hand side is the derivative of  $\ln |\sec x + \tan x|$ .

Example A.9.1

Example A.9.2 ( $\int \csc x dx$  — by the same trick)

*Solution.* The integral  $\int \csc x dx$  may also be evaluated by the method above. That is, by multiplying the integrand by a cleverly chosen  $1 = \frac{\cot x - \csc x}{\cot x - \csc x}$  and then substituting  $u = \cot x - \csc x$ ,  $du = (-\csc^2 x + \csc x \cot x) dx$ .

$$\begin{aligned}\int \csc x dx &= \int \csc x \frac{\cot x - \csc x}{\cot x - \csc x} dx = \int \frac{\csc x \cot x - \csc^2 x}{\cot x - \csc x} dx \\ &= \int \frac{1}{u} du \\ &= \ln |u| + C \\ &= \ln |\cot x - \csc x| + C\end{aligned}$$

Example A.9.2

Section A.9 of this work was adapted from Section 1.8.3 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

<sup>17</sup> We thank Serban Raianu for bringing this to our attention.

## A.10<sup>▲</sup> Further Reading on Numerical Integration

### A.10.1 ▶ The Midpoint Rule

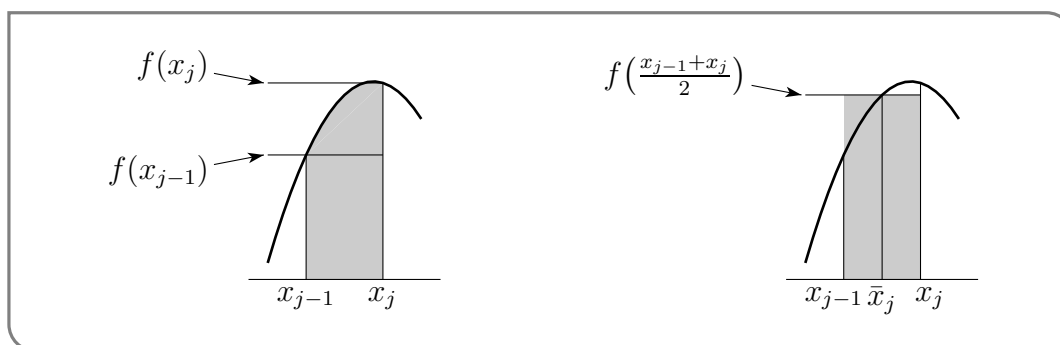
#### Notation A.10.1 (Midpoints).

In what follows we need to refer to the midpoint between  $x_{j-1}$  and  $x_j$  very frequently. To save on typing (and reading) we introduce the notation

$$\bar{x}_j = \frac{1}{2}(x_{j-1} + x_j).$$

The integral  $\int_{x_{j-1}}^{x_j} f(x) dx$  represents the area between the curve  $y = f(x)$  and the  $x$ -axis with  $x$  running from  $x_{j-1}$  to  $x_j$ . The width of this region is  $x_j - x_{j-1} = \Delta x$ . The height varies over the different values that  $f(x)$  takes as  $x$  runs from  $x_{j-1}$  to  $x_j$ .

The midpoint rule approximates this area by the area of a rectangle of width  $x_j - x_{j-1} = \Delta x$  and height  $f(\bar{x}_j)$  which is the exact height at the midpoint of the range covered by  $x$ .



The area of the approximating rectangle is  $f(\bar{x}_j)\Delta x$ , and the midpoint rule approximates each subintegral by

$$\int_{x_{j-1}}^{x_j} f(x) dx \approx f(\bar{x}_j)\Delta x$$

Applying this approximation to each subinterval and summing gives us the following approximation of the full integral:

$$\begin{aligned} \int_a^b f(x) dx &= \int_{x_0}^{x_1} f(x) dx + \int_{x_1}^{x_2} f(x) dx + \cdots + \int_{x_{n-1}}^{x_n} f(x) dx \\ &\approx f(\bar{x}_1)\Delta x + f(\bar{x}_2)\Delta x + \cdots + f(\bar{x}_n)\Delta x \end{aligned}$$

So notice that the approximation is the sum of the function evaluated at the midpoint of each interval and then multiplied by  $\Delta x$ . Our other approximations will have similar forms.

In summary:

**Equation A.10.2**(The midpoint rule).

The midpoint rule approximation is

$$\int_a^b f(x) \, dx \approx [f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)] \Delta x$$

where  $\Delta x = \frac{b-a}{n}$  and

$$\begin{aligned} x_0 = a & \quad x_1 = a + \Delta x & \quad x_2 = a + 2\Delta x & \quad \cdots & \quad x_{n-1} = b - \Delta x & \quad x_n = b \\ \bar{x}_1 = \frac{x_0+x_1}{2} & \quad \bar{x}_2 = \frac{x_1+x_2}{2} & \quad \cdots & \quad \bar{x}_{n-1} = \frac{x_{n-2}+x_{n-1}}{2} & \quad \bar{x}_n = \frac{x_{n-1}+x_n}{2} \end{aligned}$$

**Example A.10.3**  $\left(\int_0^1 \frac{4}{1+x^2} \, dx\right)$ 

We approximate the above integral using the midpoint rule with  $n = 8$  steps.

*Solution.*

- First we set up all the  $x$ -values that we will need. Note that  $a = 0$ ,  $b = 1$ ,  $\Delta x = \frac{1}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{1}{8} \quad x_2 = \frac{2}{8} \quad \cdots \quad x_7 = \frac{7}{8} \quad x_8 = \frac{8}{8} = 1$$

Consequently

$$\bar{x}_1 = \frac{1}{16} \quad \bar{x}_2 = \frac{3}{16} \quad \bar{x}_3 = \frac{5}{16} \quad \cdots \quad \bar{x}_8 = \frac{15}{16}$$

- We now apply Equation (A.10.2) to the integrand  $f(x) = \frac{4}{1+x^2}$ :

$$\begin{aligned} \int_0^1 \frac{4}{1+x^2} \, dx &\approx \left[ \overbrace{\frac{4}{1+\bar{x}_1^2}}^{f(\bar{x}_1)} + \overbrace{\frac{4}{1+\bar{x}_2^2}}^{f(\bar{x}_2)} + \cdots + \overbrace{\frac{4}{1+\bar{x}_7^2}}^{f(\bar{x}_{n-1})} + \overbrace{\frac{4}{1+\bar{x}_8^2}}^{f(\bar{x}_n)} \right] \Delta x \\ &= \left[ \frac{4}{1+\frac{1}{16^2}} + \frac{4}{1+\frac{3^2}{16^2}} + \frac{4}{1+\frac{5^2}{16^2}} + \frac{4}{1+\frac{7^2}{16^2}} + \frac{4}{1+\frac{9^2}{16^2}} + \frac{4}{1+\frac{11^2}{16^2}} + \frac{4}{1+\frac{13^2}{16^2}} + \frac{4}{1+\frac{15^2}{16^2}} \right] \frac{1}{8} \\ &= [3.98444 + 3.86415 + 3.64413 + 3.35738 + 3.03858 + 2.71618 + 2.40941 + 2.12890] \frac{1}{8} \\ &= 3.1429 \end{aligned}$$

where we have rounded to four decimal places.

- In this case we can compute the integral exactly (which is one of the reasons it was chosen as a first example):

$$\int_0^1 \frac{4}{1+x^2} \, dx = 4 \arctan x \Big|_0^1 = \pi$$

- So the error in the approximation generated by eight steps of the midpoint rule is

$$|3.1429 - \pi| = 0.0013$$

- The relative error is then

$$\frac{|\text{approximate} - \text{exact}|}{\text{exact}} = \frac{|3.1429 - \pi|}{\pi} = 0.0004$$

That is the error is 0.0004 times the actual value of the integral.

- We can write this as a percentage error by multiplying it by 100

$$\text{percentage error} = 100 \times \frac{|\text{approximate} - \text{exact}|}{\text{exact}} = 0.04\%$$

That is, the error is about 0.04% of the exact value.

Example A.10.3

The midpoint rule gives us quite good estimates of the integral without too much work — though it is perhaps a little tedious to do by hand<sup>18</sup>.

Example A.10.4 ( $\int_0^\pi \sin x \, dx$ )

As a second example, we apply the midpoint rule with  $n = 8$  steps to the above integral.

- We again start by setting up all the  $x$ -values that we will need. So  $a = 0$ ,  $b = \pi$ ,  $\Delta x = \frac{\pi}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{\pi}{8} \quad x_2 = \frac{2\pi}{8} \quad \dots \quad x_7 = \frac{7\pi}{8} \quad x_8 = \frac{8\pi}{8} = \pi$$

Consequently,

$$\bar{x}_1 = \frac{\pi}{16} \quad \bar{x}_2 = \frac{3\pi}{16} \quad \dots \quad \bar{x}_7 = \frac{13\pi}{16} \quad \bar{x}_8 = \frac{15\pi}{16}$$

- Now apply Equation (A.10.2) to the integrand  $f(x) = \sin x$ :

$$\begin{aligned} \int_0^\pi \sin x \, dx &\approx \left[ \sin(\bar{x}_1) + \sin(\bar{x}_2) + \dots + \sin(\bar{x}_8) \right] \Delta x \\ &= \left[ \sin\left(\frac{\pi}{16}\right) + \sin\left(\frac{3\pi}{16}\right) + \sin\left(\frac{5\pi}{16}\right) + \sin\left(\frac{7\pi}{16}\right) + \sin\left(\frac{9\pi}{16}\right) + \sin\left(\frac{11\pi}{16}\right) + \sin\left(\frac{13\pi}{16}\right) + \sin\left(\frac{15\pi}{16}\right) \right] \frac{\pi}{8} \\ &= \left[ 0.1951 + 0.5556 + 0.8315 + 0.9808 + 0.9808 + 0.8315 + 0.5556 + 0.1951 \right] \times 0.3927 \\ &= 5.1260 \times 0.3927 = 2.013 \end{aligned}$$

- Again, we have chosen this example so that we can compare it against the exact value:

$$\int_0^\pi \sin x \, dx = \left[ -\cos x \right]_0^\pi = -\cos \pi + \cos 0 = 2.$$

<sup>18</sup> Thankfully it is very easy to write a program to apply the midpoint rule.

- So with eight steps of the midpoint rule we achieved

$$\text{absolute error} = |2.013 - 2| = 0.013$$

$$\text{relative error} = \frac{|2.013 - 2|}{2} = 0.0065$$

$$\text{percentage error} = 100 \times \frac{|2.013 - 2|}{2} = 0.65\%$$

With little work we have managed to estimate the integral to within 1% of its true value.

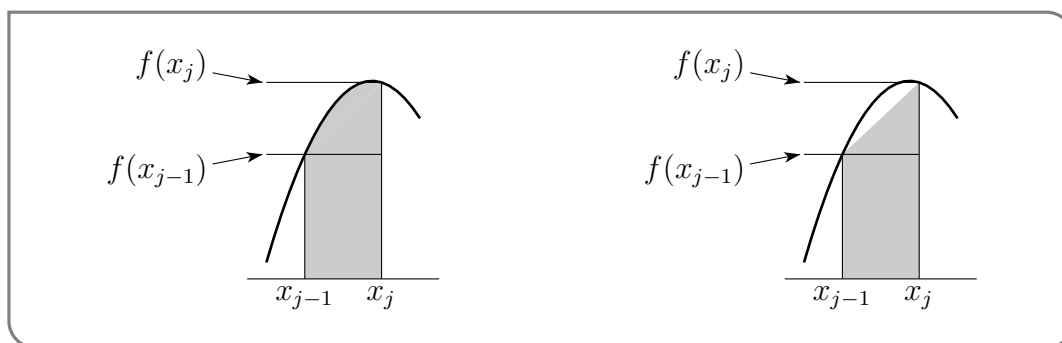
Example A.10.4

Subsection A.10.1 of this work is from Section 1.11.1 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.10.2 ▶ The Trapezoidal Rule

Consider again the area represented by the integral  $\int_{x_{j-1}}^{x_j} f(x) dx$ . The trapezoidal rule<sup>19</sup> (unsurprisingly) approximates this area by a trapezoid<sup>20</sup> whose vertices lie at

$$(x_{j-1}, 0), (x_{j-1}, f(x_{j-1})), (x_j, f(x_j)) \text{ and } (x_j, 0).$$

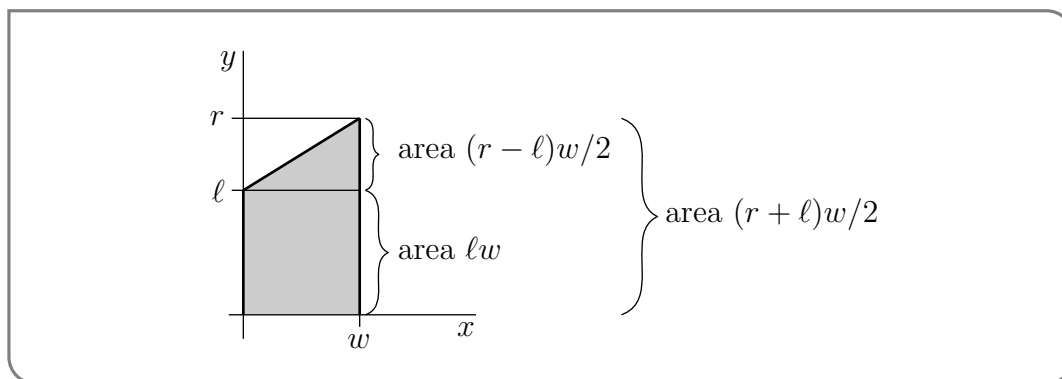


The trapezoidal approximation of the integral  $\int_{x_{j-1}}^{x_j} f(x) dx$  is the shaded region in the figure on the right above. It has width  $x_j - x_{j-1} = \Delta x$ . Its left hand side has height  $f(x_{j-1})$  and its right hand side has height  $f(x_j)$ .

As the figure below shows, the area of a trapezoid is its width times its average height.

<sup>19</sup> This method is also called the “trapezoid rule” and “trapezium rule”.

<sup>20</sup> A trapezoid is a four sided polygon, like a rectangle. But, unlike a rectangle, the top and bottom of a trapezoid need not be parallel.



So the trapezoidal rule approximates each subintegral by

$$\int_{x_{j-1}}^{x_j} f(x) \, dx \approx \frac{f(x_{j-1}) + f(x_j)}{2} \Delta x$$

Applying this approximation to each subinterval and then summing the result gives us the following approximation of the full integral

$$\begin{aligned} \int_a^b f(x) \, dx &= \int_{x_0}^{x_1} f(x) \, dx + \int_{x_1}^{x_2} f(x) \, dx + \cdots + \int_{x_{n-1}}^{x_n} f(x) \, dx \\ &\approx \frac{f(x_0) + f(x_1)}{2} \Delta x + \frac{f(x_1) + f(x_2)}{2} \Delta x + \cdots + \frac{f(x_{n-1}) + f(x_n)}{2} \Delta x \\ &= \left[ \frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n) \right] \Delta x \end{aligned}$$

So notice that the approximation has a very similar form to the midpoint rule, excepting that

- we evaluate the function at the  $x_j$ 's rather than at the midpoints, and
- we multiply the value of the function at the endpoints  $x_0, x_n$  by  $1/2$ .

In summary:

**Equation A.10.5**(The trapezoidal rule).

The trapezoidal rule approximation is

$$\int_a^b f(x) \, dx \approx \left[ \frac{1}{2}f(x_0) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(x_n) \right] \Delta x$$

where

$$\Delta x = \frac{b-a}{n}, \quad x_0 = a, \quad x_1 = a + \Delta x, \quad x_2 = a + 2\Delta x, \quad \cdots, \quad x_{n-1} = b - \Delta x, \quad x_n = b$$

To compare and contrast we apply the trapezoidal rule to the examples we did above with the midpoint rule.

Example A.10.6  $\left( \int_0^1 \frac{4}{1+x^2} \, dx \text{ — using the trapezoidal rule} \right)$

*Solution.* We proceed very similarly to Example A.10.3 and again use  $n = 8$  steps.



- We again have  $f(x) = \frac{4}{1+x^2}$ ,  $a = 0$ ,  $b = 1$ ,  $\Delta x = \frac{1}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{1}{8} \quad x_2 = \frac{2}{8} \quad \cdots \quad x_7 = \frac{7}{8} \quad x_8 = \frac{8}{8} = 1$$

- Applying the trapezoidal rule, Equation (A.10.5), gives

$$\begin{aligned} \int_0^1 \frac{4}{1+x^2} dx &\approx \left[ \frac{1}{2} \overbrace{\frac{4}{1+x_0^2}}^{f(x_0)} + \overbrace{\frac{4}{1+x_1^2}}^{f(x_1)} + \cdots + \overbrace{\frac{4}{1+x_7^2}}^{f(x_{n-1})} + \frac{1}{2} \overbrace{\frac{4}{1+x_8^2}}^{f(x_n)} \right] \Delta x \\ &= \left[ \frac{1}{2} \frac{4}{1+0^2} + \frac{4}{1+\frac{1}{8^2}} + \frac{4}{1+\frac{2^2}{8^2}} + \frac{4}{1+\frac{3^2}{8^2}} \right. \\ &\quad \left. + \frac{4}{1+\frac{4^2}{8^2}} + \frac{4}{1+\frac{5^2}{8^2}} + \frac{4}{1+\frac{6^2}{8^2}} + \frac{4}{1+\frac{7^2}{8^2}} + \frac{1}{2} \frac{4}{1+\frac{8^2}{8^2}} \right] \frac{1}{8} \\ &= \left[ \frac{1}{2} \times 4 + 3.939 + 3.765 + 3.507 \right. \\ &\quad \left. + 3.2 + 2.876 + 2.56 + 2.266 + \frac{1}{2} \times 2 \right] \frac{1}{8} \\ &= 3.139 \end{aligned}$$

to three decimal places.

- The exact value of the integral is still  $\pi$ . So the error in the approximation generated by eight steps of the trapezoidal rule is  $|3.139 - \pi| = 0.0026$ , which is  $100 \frac{|3.139 - \pi|}{\pi} \% = 0.08\%$  of the exact answer. Notice that this is roughly twice the error that we achieved using the midpoint rule in Example A.10.3.

Example A.10.6

Let us also redo Example A.10.4 using the trapezoidal rule.

Example A.10.7 ( $\int_0^\pi \sin x \, dx$  — using the trapezoidal rule)

*Solution.* We proceed very similarly to Example A.10.4 and again use  $n = 8$  steps.

- We again have  $a = 0$ ,  $b = \pi$ ,  $\Delta x = \frac{\pi}{8}$  and

$$x_0 = 0 \quad x_1 = \frac{\pi}{8} \quad x_2 = \frac{2\pi}{8} \quad \cdots \quad x_7 = \frac{7\pi}{8} \quad x_8 = \frac{8\pi}{8} = \pi$$

- Applying the trapezoidal rule, Equation (A.10.5), gives

$$\begin{aligned} \int_0^\pi \sin x \, dx &\approx \left[ \frac{1}{2} \sin(x_0) + \sin(x_1) + \cdots + \sin(x_7) + \frac{1}{2} \sin(x_8) \right] \Delta x \\ &= \left[ \frac{1}{2} \sin 0 + \sin \frac{\pi}{8} + \sin \frac{2\pi}{8} + \sin \frac{3\pi}{8} + \sin \frac{4\pi}{8} + \sin \frac{5\pi}{8} + \sin \frac{6\pi}{8} + \sin \frac{7\pi}{8} + \frac{1}{2} \sin \frac{8\pi}{8} \right] \frac{\pi}{8} \\ &= \left[ \frac{1}{2} \times 0 + 0.3827 + 0.7071 + 0.9239 + 1.0000 + 0.9239 + 0.7071 + 0.3827 + \frac{1}{2} \times 0 \right] \times 0.3927 \\ &= 5.0274 \times 0.3927 = 1.974 \end{aligned}$$

- The exact answer is  $\int_0^\pi \sin x \, dx = -\cos x \Big|_0^\pi = 2$ . So with eight steps of the trapezoidal rule we achieved  $100 \frac{|1.974-2|}{2} = 1.3\%$  accuracy. Again this is approximately twice the error we achieved in Example A.10.4 using the midpoint rule.

Example A.10.7

These two examples suggest that the midpoint rule is more accurate than the trapezoidal rule. Indeed, this observation is born out by a rigorous analysis of the error — see A.10.3.

Subsection A.10.2 of this work is included from Section 1.11.2 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

### A.10.3 ▶ Error Behaviour

As mentioned in Section 3.6.2,  $n$  steps of Simpson’s rule requires us to evaluate  $f(x)$  a total of  $n + 1$  times. For the trapezoid rule, it is the same; for the midpoint rule, we evaluate  $f(x)$  only  $n$  times. So in all three rules, the amount of “effort” needed is roughly equivalent in most circumstances.

To get a first impression of the error behaviour of these methods, we apply them to a problem whose answer we know exactly:

$$\int_0^\pi \sin x \, dx = -\cos x \Big|_0^\pi = 2.$$

To be a little more precise, we would like to understand how the errors of the three methods change as we increase the effort we put in (as measured by the number of steps  $n$ ). The following table lists the error in the approximate value for this number generated by our three rules applied with three different choices of  $n$ . It also lists the number of evaluations of  $f$  required to compute the approximation.

n	Midpoint		Trapezoidal		Simpson’s	
	error	# evals	error	# evals	error	# evals
10	$8.2 \times 10^{-3}$	10	$1.6 \times 10^{-2}$	11	$1.1 \times 10^{-4}$	11
100	$8.2 \times 10^{-5}$	100	$1.6 \times 10^{-4}$	101	$1.1 \times 10^{-8}$	101
1000	$8.2 \times 10^{-7}$	1000	$1.6 \times 10^{-6}$	1001	$1.1 \times 10^{-12}$	1001

Observe that

- Using 101 evaluations of  $f$  worth of Simpson’s rule gives an error 80 times smaller than 1000 evaluations of  $f$  worth of the midpoint rule. (See why we focus on Simpson’s over midpoint?)
- The trapezoidal rule error with  $n$  steps is about twice the midpoint rule error with  $n$  steps. (Hence its relegation to this appendix.)

- With the midpoint rule, increasing the number of steps by a factor of 10 appears to reduce the error by about a factor of  $100 = 10^2 = n^2$ .
- With the trapezoidal rule, increasing the number of steps by a factor of 10 appears to reduce the error by about a factor of  $10^2 = n^2$ .
- With Simpson's rule, increasing the number of steps by a factor of 10 appears to reduce the error by about a factor of  $10^4 = n^4$ .

So it looks like

$$\begin{aligned} \text{approx value of } \int_a^b f(x) \, dx \text{ given by } n \text{ midpoint steps} &\approx \int_a^b f(x) \, dx + K_M \cdot \frac{1}{n^2} \\ \text{approx value of } \int_a^b f(x) \, dx \text{ given by } n \text{ trapezoidal steps} &\approx \int_a^b f(x) \, dx + K_T \cdot \frac{1}{n^2} \\ \text{approx value of } \int_a^b f(x) \, dx \text{ given by } n \text{ Simpson's steps} &\approx \int_a^b f(x) \, dx + K_S \cdot \frac{1}{n^4} \end{aligned}$$

with some constants  $K_M$ ,  $K_T$  and  $K_S$ . It also seems that  $K_T \approx 2K_M$ . The intuition, about the error behaviour, that we have just developed is in fact correct — provided the integrand  $f(x)$  is reasonably smooth. To be more precise:

**Theorem A.10.8** (Numerical integration errors).

Assume that  $|f''(x)| \leq M$  for all  $a \leq x \leq b$ . Then

the total error introduced by the midpoint rule is bounded by  $\frac{M(b-a)^3}{24n^2}$

and

the total error introduced by the trapezoidal rule is bounded by  $\frac{M(b-a)^3}{12n^2}$

when approximating  $\int_a^b f(x) \, dx$ . Further, if  $|f^{(4)}(x)| \leq L$  for all  $a \leq x \leq b$ , then

the total error introduced by Simpson's rule is bounded by  $\frac{L(b-a)^5}{180n^4}$ .

Subsection A.10.3 of this work is adapted from Section 1.11.4 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

### A.10.4 ▶ An Error Bound for the Midpoint Rule

We now try develop some understanding as to why we got the experimental results of section A.10.3. We start with the error generated by a single step of the midpoint rule.

That is, the error introduced by the approximation

$$\int_{x_0}^{x_1} f(x) \, dx \approx f(\bar{x}_1)\Delta x \quad \text{where } \Delta x = x_1 - x_0, \bar{x}_1 = \frac{x_0+x_1}{2}$$

To do this we are going to need to apply integration by parts in a sneaky way. Let us start by considering<sup>21</sup> a subinterval  $\alpha \leq x \leq \beta$  and let's call the width of the subinterval  $2q$  so that  $\beta = \alpha + 2q$ . If we were to now apply the midpoint rule to this subinterval, then we would write

$$\int_{\alpha}^{\beta} f(x) \, dx \approx 2q \cdot f(\alpha + q) = qf(\alpha + q) + qf(\beta - q)$$

since the interval has width  $2q$  and the midpoint is  $\alpha + q = \beta - q$ .

The sneaky trick we will employ is to write

$$\int_{\alpha}^{\beta} f(x) \, dx = \int_{\alpha}^{\alpha+q} f(x) \, dx + \int_{\beta-q}^{\beta} f(x) \, dx$$

and then examine each of the integrals on the right-hand side (using integration by parts) and show that they are each of the form

$$\begin{aligned} \int_{\alpha}^{\alpha+q} f(x) \, dx &\approx qf(\alpha + q) + \text{small error term} \\ \int_{\beta-q}^{\beta} f(x) \, dx &\approx qf(\beta - q) + \text{small error term} \end{aligned}$$

Let us apply integration by parts to  $\int_{\alpha}^{\alpha+q} f(x) \, dx$  — with  $u = f(x)$ ,  $dv = dx$  so  $du = f'(x) \, dx$  and we will make the slightly non-standard choice of  $v = x - \alpha$ :

$$\begin{aligned} \int_{\alpha}^{\alpha+q} f(x) \, dx &= [(x - \alpha)f(x)]_{\alpha}^{\alpha+q} - \int_{\alpha}^{\alpha+q} (x - \alpha)f'(x) \, dx \\ &= qf(\alpha + q) - \int_{\alpha}^{\alpha+q} (x - \alpha)f'(x) \, dx \end{aligned}$$

Notice that the first term on the right-hand side is the term we need, and that our non-standard choice of  $v$  allowed us to avoid introducing an  $f(\alpha)$  term.

Now integrate by parts again using  $u = f'(x)$ ,  $dv = (x - \alpha) \, dx$ , so  $du = f''(x)$ ,  $v = \frac{(x - \alpha)^2}{2}$ :

$$\begin{aligned} \int_{\alpha}^{\alpha+q} f(x) \, dx &= qf(\alpha + q) - \int_{\alpha}^{\alpha+q} (x - \alpha)f'(x) \, dx \\ &= qf(\alpha + q) - \left[ \frac{(x - \alpha)^2}{2} f'(x) \right]_{\alpha}^{\alpha+q} + \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2} f''(x) \, dx \\ &= qf(\alpha + q) - \frac{q^2}{2} f'(\alpha + q) + \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2} f''(x) \, dx \end{aligned}$$

21 We chose this interval so that we didn't have lots of subscripts floating around in the algebra.

To obtain a similar expression for the other integral, we repeat the above steps and obtain:

$$\int_{\beta-q}^{\beta} f(x)dx = qf(\beta - q) + \frac{q^2}{2}f'(\beta - q) + \int_{\beta-q}^{\beta} \frac{(x - \beta)^2}{2}f''(x)dx$$

Now add together these two expressions

$$\begin{aligned} \int_{\alpha}^{\alpha+q} f(x)dx + \int_{\beta-q}^{\beta} f(x)dx &= qf(\alpha + q) + qf(\beta - q) + \frac{q^2}{2}(f'(\beta - q) - f'(\alpha + q)) \\ &\quad + \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2}f''(x)dx + \int_{\beta-q}^{\beta} \frac{(x - \beta)^2}{2}f''(x)dx \end{aligned}$$

Then since  $\alpha + q = \beta - q$  we can combine the integrals on the left-hand side and eliminate some terms from the right-hand side:

$$\int_{\alpha}^{\beta} f(x)dx = 2qf(\alpha + q) + \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2}f''(x)dx + \int_{\beta-q}^{\beta} \frac{(x - \beta)^2}{2}f''(x)dx$$

Rearrange this expression a little and take absolute values

$$\left| \int_{\alpha}^{\beta} f(x)dx - 2qf(\alpha + q) \right| \leq \left| \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2}f''(x)dx \right| + \left| \int_{\beta-q}^{\beta} \frac{(x - \beta)^2}{2}f''(x)dx \right|$$

where we have also made use of the triangle inequality<sup>22</sup>. By assumption  $|f''(x)| \leq M$  on the interval  $\alpha \leq x \leq \beta$ , so

$$\begin{aligned} \left| \int_{\alpha}^{\beta} f(x)dx - 2qf(\alpha + q) \right| &\leq M \int_{\alpha}^{\alpha+q} \frac{(x - \alpha)^2}{2}dx + M \int_{\beta-q}^{\beta} \frac{(x - \beta)^2}{2}dx \\ &= \frac{Mq^3}{3} = \frac{M(\beta - \alpha)^3}{24} \end{aligned}$$

where we have used  $q = \frac{\beta - \alpha}{2}$  in the last step.

Thus on any interval  $x_i \leq x \leq x_{i+1} = x_i + \Delta x$

$$\left| \int_{x_i}^{x_{i+1}} f(x)dx - \Delta x f\left(\frac{x_i + x_{i+1}}{2}\right) \right| \leq \frac{M}{24}(\Delta x)^3$$

Putting everything together we see that the error using the midpoint rule is bounded by

$$\begin{aligned} &\left| \int_a^b f(x)dx - [f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)] \Delta x \right| \\ &\leq \left| \int_{x_0}^{x_1} f(x)dx - \Delta x f(\bar{x}_1) \right| + \cdots + \left| \int_{x_{n-1}}^{x_n} f(x)dx - \Delta x f(\bar{x}_n) \right| \\ &\leq n \times \frac{M}{24}(\Delta x)^3 = n \times \frac{M}{24} \left( \frac{b - a}{n} \right)^3 = \frac{M(b - a)^3}{24n^2} \end{aligned}$$

22 The triangle inequality says that for any real numbers  $x, y$

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

as required.

A very similar analysis shows that, as was stated in Theorem 3.6.5 above,

- the total error introduced by the trapezoidal rule is bounded by  $\frac{M(b-a)^3}{12n^2}$ ,
- the total error introduced by Simpson's rule is bounded by  $\frac{M(b-a)^5}{180n^4}$

Subsection A.10.4 of this work was adapted from Section 1.11.5 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.11▲ Comparison Tests Proof

In this and the next optional section we provide proofs of two convergence tests. We shall repeatedly use the fact that any sequence  $a_1, a_2, a_3, \dots$ , of real numbers which is increasing (i.e.  $a_{n+1} \geq a_n$  for all  $n$ ) and bounded (i.e. there is a constant  $M$  such that  $a_n \leq M$  for all  $n$ ) converges. We shall not prove this fact<sup>23</sup>.

We start with the comparison test, and then move on to the alternating series test.

### Theorem A.11.1 (The Comparison Test).

Let  $N_0$  be a natural number and let  $K > 0$ .

- (a) If  $|a_n| \leq Kc_n$  for all  $n \geq N_0$  and  $\sum_{n=0}^{\infty} c_n$  converges, then  $\sum_{n=0}^{\infty} a_n$  converges.
- (b) If  $a_n \geq Kd_n \geq 0$  for all  $n \geq N_0$  and  $\sum_{n=0}^{\infty} d_n$  diverges, then  $\sum_{n=0}^{\infty} a_n$  diverges.

*Proof.* (a) By hypothesis  $\sum_{n=0}^{\infty} c_n$  converges. So it suffices to prove that  $\sum_{n=0}^{\infty} [Kc_n - a_n]$  converges, because then, by our Arithmetic of series Theorem 5.2.9,

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} Kc_n - \sum_{n=0}^{\infty} [Kc_n - a_n]$$

will converge too. But for all  $n \geq N_0$ ,  $Kc_n - a_n \geq 0$  so that, for all  $N \geq N_0$ , the partial sums

$$S_N = \sum_{n=0}^N [Kc_n - a_n]$$

increase with  $N$ , but never gets bigger than the finite number  $\sum_{n=0}^{N_0} [Kc_n - a_n] + K \sum_{n=N_0+1}^{\infty} c_n$ . So the partial sums  $S_N$  converge as  $N \rightarrow \infty$ .

<sup>23</sup> It is one way to state a property of the real number system called “completeness”. The interested reader should use their favourite search engine to look up “completeness of the real numbers”.

(b) For all  $N > N_0$ , the partial sum

$$S_N = \sum_{n=0}^N a_n \geq \sum_{n=0}^{N_0} a_n + K \sum_{n=N_0+1}^N d_n$$

By hypothesis,  $\sum_{n=N_0+1}^N d_n$ , and hence  $S_N$ , grows without bound as  $N \rightarrow \infty$ . So  $S_N \rightarrow \infty$  as  $N \rightarrow \infty$ .  $\square$

Section A.11 of this work was adapted from Section 3.3.10 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.12▲ Alternating Series

### A.12.1 ►► The Alternating Series Test

A common convergence test for series that is not included in our learning goals is the Alternating Series Test. First, we need to know what an alternating series is.

When the signs of successive terms in a series alternate between  $+$  and  $-$ , like for example in  $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$ , the series is called an *alternating series*. More generally, the series

$$A_1 - A_2 + A_3 - A_4 + \dots = \sum_{n=1}^{\infty} (-1)^{n-1} A_n$$

is alternating if every  $A_n \geq 0$ . Often (but not always) the terms in alternating series get successively smaller. That is, then  $A_1 \geq A_2 \geq A_3 \geq \dots$ . In this case:

- The first partial sum is  $S_1 = A_1$ .
- The second partial sum,  $S_2 = A_1 - A_2$ , is smaller than  $S_1$  by  $A_2$ .
- The third partial sum,  $S_3 = S_2 + A_3$ , is bigger than  $S_2$  by  $A_3$ , but because  $A_3 \leq A_2$ ,  $S_3$  remains smaller than  $S_1$ . See the figure below.
- The fourth partial sum,  $S_4 = S_3 - A_4$ , is smaller than  $S_3$  by  $A_4$ , but because  $A_4 \leq A_3$ ,  $S_4$  remains bigger than  $S_2$ . Again, see the figure below.
- And so on.

So the successive partial sums oscillate, but with ever decreasing amplitude. If, in addition,  $A_n$  tends to 0 as  $n$  tends to  $\infty$ , the amplitude of oscillation tends to zero and the sequence  $S_1, S_2, S_3, \dots$  converges to some limit  $S$ .

Here is a convergence test for alternating series that exploits this structure, and that is really easy to apply.

**Theorem A.12.1** (Alternating Series Test).

Let  $\{A_n\}_{n=1}^{\infty}$  be a sequence of real numbers that obeys

- (i)  $A_n \geq 0$  for all  $n \geq 1$  and
- (ii)  $A_{n+1} \leq A_n$  for all  $n \geq 1$  (i.e. the sequence is monotone decreasing) and
- (iii)  $\lim_{n \rightarrow \infty} A_n = 0$ .

Then

$$A_1 - A_2 + A_3 - A_4 + \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} A_n = S$$

converges and, for each natural number  $N$ ,  $S - S_N$  is between 0 and (the first dropped term)  $(-1)^N A_{N+1}$ . Here  $S_N$  is, as previously, the  $N^{\text{th}}$  partial sum  $\sum_{n=1}^N (-1)^{n-1} A_n$ .

“Proof”. We shall only give part of the proof here. For the rest of the proof see the appendix section A.11. We shall fix any natural number  $N$  and concentrate on the last statement, which gives a bound on the truncation error (which is the error introduced when you approximate the full series by the partial sum  $S_N$ )

$$E_N = S - S_N = \sum_{n=N+1}^{\infty} (-1)^{n-1} A_n = (-1)^N [A_{N+1} - A_{N+2} + A_{N+3} - A_{N+4} + \cdots]$$

This is of course another series. We’re going to study the partial sums

$$S_{N,\ell} = \sum_{n=N+1}^{\ell} (-1)^{n-1} A_n = (-1)^N \sum_{m=1}^{\ell-N} (-1)^{m-1} A_{N+m}$$

for that series.

- If  $\ell' > N + 1$ , with  $\ell' - N$  even,

$$(-1)^N S_{N,\ell'} = \overbrace{(A_{N+1} - A_{N+2})}^{\geq 0} + \overbrace{(A_{N+3} - A_{N+4})}^{\geq 0} + \cdots + \overbrace{(A_{\ell'-1} - A_{\ell'})}^{\geq 0} \geq 0 \quad \text{and}$$

$$(-1)^N S_{N,\ell'+1} = \overbrace{(-1)^N S_{N,\ell'}}^{\geq 0} + \overbrace{A_{\ell'+1}}^{\geq 0} \geq 0$$

This tells us that  $(-1)^N S_{N,\ell} \geq 0$  for all  $\ell > N + 1$ , both even and odd.

- Similarly, if  $\ell' > N + 1$ , with  $\ell' - N$  odd,

$$(-1)^N S_{N,\ell'} = A_{N+1} - \overbrace{(A_{N+2} - A_{N+3})}^{\geq 0} - \overbrace{(A_{N+4} - A_{N+5})}^{\geq 0} - \cdots - \overbrace{(A_{\ell'-1} - A_{\ell'})}^{\geq 0} \leq A_{N+1}$$

$$(-1)^N S_{N,\ell'+1} = \overbrace{(-1)^N S_{N,\ell'}}^{\leq A_{N+1}} - \overbrace{A_{\ell'+1}}^{\geq 0} \leq A_{N+1}$$

This tells us that  $(-1)^N S_{N,\ell} \leq A_{N+1}$  for all for all  $\ell > N + 1$ , both even and odd.



So we now know that  $S_{N,\ell}$  lies between its first term,  $(-1)^N A_{N+1}$ , and 0 for all  $\ell > N + 1$ . While we are not going to prove it here (see the optional section A.11), this implies that, since  $A_{N+1} \rightarrow 0$  as  $N \rightarrow \infty$ , the series converges and that

$$S - S_N = \lim_{\ell \rightarrow \infty} S_{N,\ell}$$

lies between  $(-1)^N A_{N+1}$  and 0. □

**Example A.12.2**

We have already seen, in Example 5.3.7, that the harmonic series  $\sum_{n=1}^{\infty} \frac{1}{n}$  diverges. On the other hand, the series  $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$  converges by the alternating series test with  $A_n = \frac{1}{n}$ . Note that

- (i)  $A_n = \frac{1}{n} \geq 0$  for all  $n \geq 1$ , so that  $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$  really is an alternating series, and
- (ii)  $A_n = \frac{1}{n}$  decreases as  $n$  increases, and
- (iii)  $\lim_{n \rightarrow \infty} A_n = \lim_{n \rightarrow \infty} \frac{1}{n} = 0$ .

so that all of the hypotheses of the alternating series test, i.e. of Theorem A.12.1, are satisfied. We shall see, in Example 6.2.8, that

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \ln 2.$$

**Example A.12.2**

**Example A.12.3 (e)**

You may already know that  $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ . In any event, we shall prove this in Example 6.3.3, below. In particular

$$\frac{1}{e} = e^{-1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \dots$$

is an alternating series and satisfies all of the conditions of the alternating series test, Theorem A.12.1a:

- (i) The terms in the series alternate in sign.
- (ii) The magnitude of the  $n^{\text{th}}$  term in the series decreases monotonically as  $n$  increases.
- (iii) The  $n^{\text{th}}$  term in the series converges to zero as  $n \rightarrow \infty$ .

So the alternating series test guarantees that, if we approximate, for example,

$$\frac{1}{e} \approx \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!}$$

then the error in this approximation lies between 0 and the next term in the series, which is  $\frac{1}{10!}$ . That is

$$\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} \leq \frac{1}{e} \leq \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} + \frac{1}{10!}$$

so that

$$\frac{1}{\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} + \frac{1}{10!}} \leq e \leq \frac{1}{\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!}}$$

which, to seven decimal places says

$$2.7182816 \leq e \leq 2.7182837$$

(To seven decimal places  $e = 2.7182818$ .)

The alternating series test tells us that, for any natural number  $N$ , the error that we make when we approximate  $\frac{1}{e}$  by the partial sum  $S_N = \sum_{n=0}^N \frac{(-1)^n}{n!}$  has magnitude no larger than  $\frac{1}{(N+1)!}$ . This tends to zero spectacularly quickly as  $N$  increases, simply because  $(N+1)!$  increases spectacularly quickly as  $N$  increases<sup>24</sup>. For example  $20! \approx 2.4 \times 10^{27}$ .

Example A.12.3

Example A.12.4

We will shortly see, in Example 6.2.8, that if  $-1 < x \leq 1$ , then

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n}$$

Suppose that we have to compute  $\ln \frac{11}{10}$  to within an accuracy of  $10^{-12}$ . Since  $\frac{11}{10} = 1 + \frac{1}{10}$ , we can get  $\ln \frac{11}{10}$  by evaluating  $\ln(1+x)$  at  $x = \frac{1}{10}$ , so that

$$\ln \frac{11}{10} = \ln \left(1 + \frac{1}{10}\right) = \frac{1}{10} - \frac{1}{2 \times 10^2} + \frac{1}{3 \times 10^3} - \frac{1}{4 \times 10^4} + \dots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n \times 10^n}$$

By the alternating series test, this series converges. Also by the alternating series test, approximating  $\ln \frac{11}{10}$  by throwing away all but the first  $N$  terms

$$\ln \frac{11}{10} \approx \frac{1}{10} - \frac{1}{2 \times 10^2} + \frac{1}{3 \times 10^3} - \frac{1}{4 \times 10^4} + \dots + (-1)^{N-1} \frac{1}{N \times 10^N} = \sum_{n=1}^N (-1)^{n-1} \frac{1}{n \times 10^n}$$

introduces an error whose magnitude is no more than the magnitude of the first term that we threw away.

$$\text{error} \leq \frac{1}{(N+1) \times 10^{N+1}}$$

24 The interested reader may wish to check out “Stirling’s approximation”, which says that  $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$ .

To achieve an error that is no more than  $10^{-12}$ , we have to choose  $N$  so that

$$\frac{1}{(N+1) \times 10^{N+1}} \leq 10^{-12}$$

The best way to do so is simply to guess — we are not going to be able to manipulate the inequality  $\frac{1}{(N+1) \times 10^{N+1}} \leq \frac{1}{10^{12}}$  into the form  $N \leq \dots$ , and even if we could, it would not be worth the effort. We need to choose  $N$  so that the denominator  $(N+1) \times 10^{N+1}$  is at least  $10^{12}$ . That is easy, because the denominator contains the factor  $10^{N+1}$  which is at least  $10^{12}$  whenever  $N+1 \geq 12$ , i.e. whenever  $N \geq 11$ . So we will achieve an error of less than  $10^{-12}$  if we choose  $N = 11$ .

$$\frac{1}{(N+1) \times 10^{N+1}} \Big|_{N=11} = \frac{1}{12 \times 10^{12}} < \frac{1}{10^{12}}$$

This is not the smallest possible choice of  $N$ , but in practice that just doesn't matter — your computer is not going to care whether or not you ask it to compute a few extra terms. If you really need the smallest  $N$  that obeys  $\frac{1}{(N+1) \times 10^{N+1}} \leq \frac{1}{10^{12}}$ , you can next just try  $N = 10$ , then  $N = 9$ , and so on.

$$\begin{aligned} \frac{1}{(N+1) \times 10^{N+1}} \Big|_{N=11} &= \frac{1}{12 \times 10^{12}} < \frac{1}{10^{12}} \\ \frac{1}{(N+1) \times 10^{N+1}} \Big|_{N=10} &= \frac{1}{11 \times 10^{11}} < \frac{1}{10 \times 10^{11}} = \frac{1}{10^{12}} \\ \frac{1}{(N+1) \times 10^{N+1}} \Big|_{N=9} &= \frac{1}{10 \times 10^{10}} = \frac{1}{10^{11}} > \frac{1}{10^{12}} \end{aligned}$$

So in this problem, the smallest acceptable  $N = 10$ .

Example A.12.4

### A.12.2 ▶ Alternating Series Test Proof

**Theorem A.12.5** (Alternating Series Test).

Let  $\{a_n\}_{n=1}^{\infty}$  be a sequence of real numbers that obeys

- (i)  $a_n \geq 0$  for all  $n \geq 1$  and
- (ii)  $a_{n+1} \leq a_n$  for all  $n \geq 1$  (i.e. the sequence is monotone decreasing) and
- (iii)  $\lim_{n \rightarrow \infty} a_n = 0$ .

Then

$$a_1 - a_2 + a_3 - a_4 + \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} a_n = S$$

converges and, for each natural number  $N$ ,  $S - S_N$  is between 0 and (the first dropped term)  $(-1)^N a_{N+1}$ . Here  $S_N$  is, as previously, the  $N^{\text{th}}$  partial sum  $\sum_{n=1}^N (-1)^{n-1} a_n$ .

*Proof.* Let  $2n$  be an even natural number. Then the  $2n^{\text{th}}$  partial sum obeys

$$\begin{aligned} S_{2n} &= \overbrace{(a_1 - a_2)}^{\geq 0} + \overbrace{(a_3 - a_4)}^{\geq 0} + \cdots + \overbrace{(a_{2n-1} - a_{2n})}^{\geq 0} \\ &\leq \overbrace{(a_1 - a_2)}^{\geq 0} + \overbrace{(a_3 - a_4)}^{\geq 0} + \cdots + \overbrace{(a_{2n-1} - a_{2n})}^{\geq 0} + \overbrace{(a_{2n+1} - a_{2n+2})}^{\geq 0} = S_{2(n+1)} \end{aligned}$$

and

$$\begin{aligned} S_{2n} &= a_1 - \overbrace{(a_2 - a_3)}^{\geq 0} - \overbrace{(a_4 - a_5)}^{\geq 0} - \cdots - \overbrace{(a_{2n-2} - a_{2n-1})}^{\geq 0} - \overbrace{a_{2n}}^{\geq 0} \\ &\leq a_1 \end{aligned}$$

So the sequence  $S_2, S_4, S_6, \dots$  of even partial sums is a bounded, increasing sequence and hence converges to some real number  $S$ . Since  $S_{2n+1} = S_{2n} + a_{2n+1}$  and  $a_{2n+1}$  converges zero as  $n \rightarrow \infty$ , the odd partial sums  $S_{2n+1}$  also converge to  $S$ . That  $S - S_N$  is between 0 and (the first dropped term)  $(-1)^N a_{N+1}$  was already proved in §A.12.1.  $\square$

Section A.12 of this work was adapted from Sections 3.3.4 and 3.3.10 of [CLP 2 – Integral Calculus](#) by Feldman, Reznitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## A.13▲ Delicacy of Conditional Convergence

Conditionally convergent series have to be treated with great care. For example, switching the order of the terms in a finite sum does not change its value.

$$1 + 2 + 3 + 4 + 5 + 6 = 6 + 3 + 5 + 2 + 4 + 1$$

The same is true for absolutely convergent series. But it is *not true* for conditionally convergent series. In fact by reordering *any* conditionally convergent series, you can make it add up to *any* number you like, including  $+\infty$  and  $-\infty$ . This very strange result is known as Riemann's rearrangement Theorem, named after Bernhard Riemann (1826–1866). The following example illustrates the phenomenon.

Example A.13.1

The alternating Harmonic series

$$\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$$

is a very good example of conditional convergence. We can show, quite explicitly, how we can rearrange the terms to make it add up to two different numbers. Later, in Example 6.2.8, we'll show that this series is equal to  $\ln 2$ . However, by rearranging the terms we can make it sum to  $\frac{1}{2} \ln 2$ . The usual order is

$$\frac{1}{1} - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots$$

For the moment think of the terms being paired as follows:

$$\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \dots$$

so the denominators go odd-even odd-even. Now rearrange the terms so the denominators are odd-even-even odd-even-even:

$$\left(1 - \frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{6} - \frac{1}{8}\right) + \left(\frac{1}{5} - \frac{1}{10} - \frac{1}{12}\right) + \dots$$

Now notice that the first term in each triple is exactly twice the second term. If we now combine those terms we get

$$\begin{aligned} & \left(\underbrace{1 - \frac{1}{2} - \frac{1}{4}}_{=1/2}\right) + \left(\underbrace{\frac{1}{3} - \frac{1}{6} - \frac{1}{8}}_{=1/6}\right) + \left(\underbrace{\frac{1}{5} - \frac{1}{10} - \frac{1}{12}}_{=1/10}\right) + \dots \\ &= \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{6} - \frac{1}{8}\right) + \left(\frac{1}{10} - \frac{1}{12}\right) + \dots \end{aligned}$$

We can now extract a factor of  $\frac{1}{2}$  from each term, so

$$\begin{aligned} &= \frac{1}{2} \left(\frac{1}{1} - \frac{1}{2}\right) + \frac{1}{2} \left(\frac{1}{3} - \frac{1}{4}\right) + \frac{1}{2} \left(\frac{1}{5} - \frac{1}{6}\right) + \dots \\ &= \frac{1}{2} \left[\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \dots\right] \end{aligned}$$

So by rearranging the terms, the sum of the series is now exactly half the original sum!

Example A.13.1

In fact, we can go even further, and show how we can rearrange the terms of the alternating harmonic series to add up to any given number<sup>25</sup>. For the purposes of the example we have chosen 1.234, but it could really be any number. The example below can actually be formalised to give a proof of the rearrangement Theorem.

Example A.13.2

We'll show how to reorder the conditionally convergent series  $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$  so that it adds up to exactly 1.234 (but the reader should keep in mind that any fixed number will work).

- First create two lists of numbers — the first list consisting of the positive terms of the series, in order, and the second consisting of the negative numbers of the series, in order.

$$1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \dots \quad \text{and} \quad -\frac{1}{2}, -\frac{1}{4}, -\frac{1}{6}, \dots$$

- Notice that that if we add together the numbers in the second list, we get

$$-\frac{1}{2} \left[ 1 + \frac{1}{2} + \frac{1}{3} + \dots \right]$$

which is just  $-\frac{1}{2}$  times the harmonic series. So the numbers in the second list add up to  $-\infty$ .

Also, if we add together the numbers in the first list, we get

$$1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} \dots \quad \text{which is greater than} \quad \frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \frac{1}{8} + \dots$$

That is, the sum of the first set of numbers must be bigger than the sum of the second set of numbers (which is just  $-1$  times the second list). So the numbers in the first list add up to  $+\infty$ .

- Now we build up our reordered series. Start by moving just enough numbers from the beginning of the first list into the reordered series to get a sum bigger than 1.234.

$$1 + \frac{1}{3} = 1.3333$$

We know that we can do this, because the sum of the terms in the first list diverges to  $+\infty$ .

25 This is reminiscent of the accounting trick of pushing all the company's debts off to next year so that this year's accounts look really good and you can collect your bonus.

- Next move just enough numbers from the beginning of the second list into the re-ordered series to get a number less than 1.234.

$$1 + \frac{1}{3} - \frac{1}{2} = 0.8333$$

Again, we know that we can do this because the sum of the numbers in the second list diverges to  $-\infty$ .

- Next move just enough numbers from the beginning of the remaining part of the first list into the reordered series to get a number bigger than 1.234.

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} + \frac{1}{9} = 1.2873$$

Again, this is possible because the sum of the numbers in the first list diverges. Even though we have already used the first few numbers, the sum of the rest of the list will still diverge.

- Next move just enough numbers from the beginning of the remaining part of the second list into the reordered series to get a number less than 1.234.

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} + \frac{1}{9} - \frac{1}{4} = 1.0373$$

- At this point the idea is clear, just keep going like this. At the end of each step, the difference between the sum and 1.234 is smaller than the magnitude of the first unused number in the lists. Since the numbers in both lists tend to zero as you go farther and farther up the list, this procedure will generate a series whose sum is exactly 1.234. Since in each step we remove at least one number from a list and we alternate between the two lists, the reordered series will contain all of the terms from  $\sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n}$ , with each term appearing exactly once.

Example A.13.2

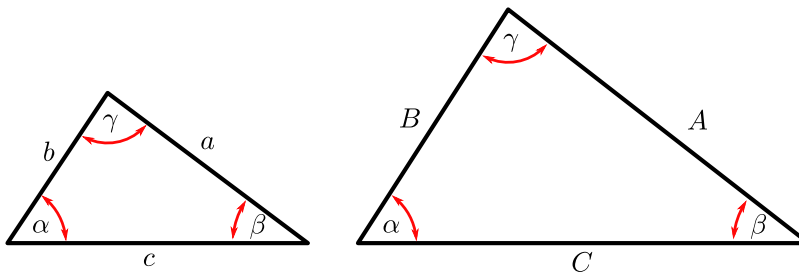
Section A.13 of this work was adapted from Section 3.4.2 of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).

## HIGH SCHOOL MATERIAL

This chapter is really split into three parts.

- Sections B.1 to B.11 contains results that we expect you to understand and know.
- Then Section B.14 contains results that we don't expect you to memorise, but that we think you should be able to quickly derive from other results you know.
- The remaining sections contain some material (that may be new to you) that is related to topics covered in the main body of these notes.

## B.1▲ Similar Triangles



Two triangles  $T_1, T_2$  are similar when

- (AAA — angle angle angle) The angles of  $T_1$  are the same as the angles of  $T_2$ .
- (SSS — side side side) The ratios of the side lengths are the same. That is

$$\frac{A}{a} = \frac{B}{b} = \frac{C}{c}$$

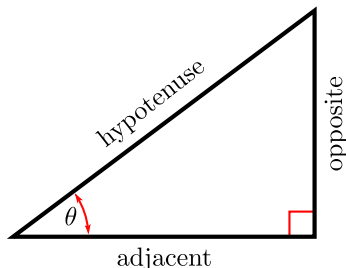
- (SAS — side angle side) Two sides have lengths in the same ratio and the angle between them is the same. For example

$$\frac{A}{a} = \frac{C}{c} \text{ and angle } \beta \text{ is same}$$



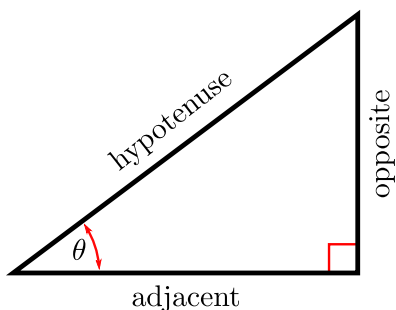
## B.2▲ Pythagoras

For a right-angled triangle the length of the hypotenuse is related to the lengths of the other two sides by



$$(\text{adjacent})^2 + (\text{opposite})^2 = (\text{hypotenuse})^2$$

## B.3▲ Trigonometry — Definitions

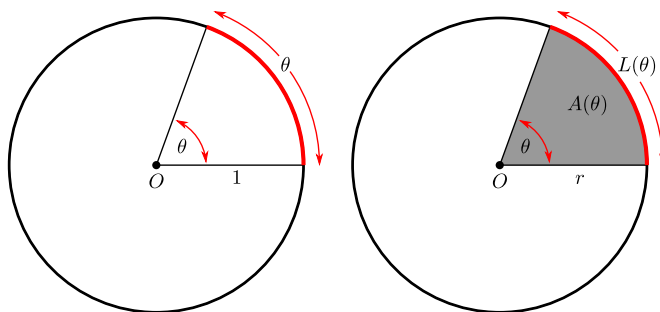


$$\sin \theta = \frac{\text{opposite}}{\text{hypotenuse}} \quad \csc \theta = \frac{1}{\sin \theta}$$

$$\cos \theta = \frac{\text{adjacent}}{\text{hypotenuse}} \quad \sec \theta = \frac{1}{\cos \theta}$$

$$\tan \theta = \frac{\text{opposite}}{\text{adjacent}} \quad \cot \theta = \frac{1}{\tan \theta}$$

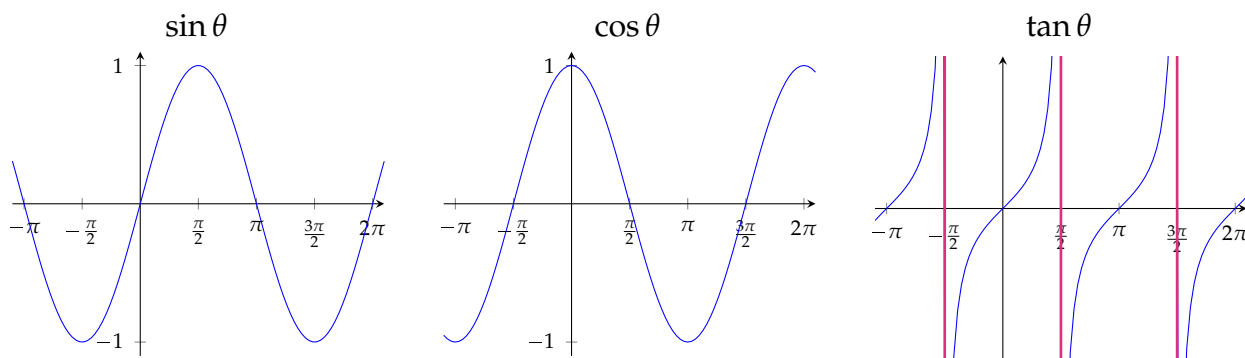
## B.4▲ Radians, Arcs and Sectors



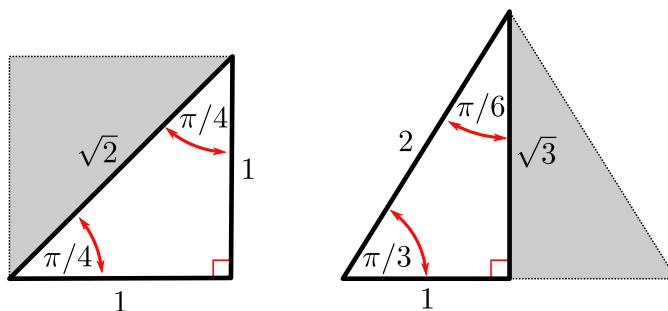
For a circle of radius  $r$  and angle of  $\theta$  radians:

- Arc length  $L(\theta) = r\theta$ .
- Area of sector  $A(\theta) = \frac{\theta}{2}r^2$ .

## B.5▲ Trigonometry — Graphs



## B.6▲ Trigonometry — Special Triangles



From the above pair of special triangles we have

$$\sin \frac{\pi}{4} = \frac{1}{\sqrt{2}}$$

$$\sin \frac{\pi}{6} = \frac{1}{2}$$

$$\sin \frac{\pi}{3} = \frac{\sqrt{3}}{2}$$

$$\cos \frac{\pi}{4} = \frac{1}{\sqrt{2}}$$

$$\cos \frac{\pi}{6} = \frac{\sqrt{3}}{2}$$

$$\cos \frac{\pi}{3} = \frac{1}{2}$$

$$\tan \frac{\pi}{4} = 1$$

$$\tan \frac{\pi}{6} = \frac{1}{\sqrt{3}}$$

$$\tan \frac{\pi}{3} = \sqrt{3}$$

## B.7▲ Trigonometry — Simple Identities

- Periodicity

$$\sin(\theta + 2\pi) = \sin(\theta)$$

$$\cos(\theta + 2\pi) = \cos(\theta)$$

- Reflection

$$\sin(-\theta) = -\sin(\theta)$$

$$\cos(-\theta) = \cos(\theta)$$

- Reflection around  $\pi/4$

$$\sin\left(\frac{\pi}{2} - \theta\right) = \cos \theta$$

$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin \theta$$

- Reflection around  $\pi/2$

$$\sin(\pi - \theta) = \sin \theta$$

$$\cos(\pi - \theta) = -\cos \theta$$

- Rotation by  $\pi$

$$\sin(\theta + \pi) = -\sin \theta$$

$$\cos(\theta + \pi) = -\cos \theta$$

- Pythagoras

$$\sin^2 \theta + \cos^2 \theta = 1$$

## B.8▲ Trigonometry — Add and Subtract Angles

- Sine

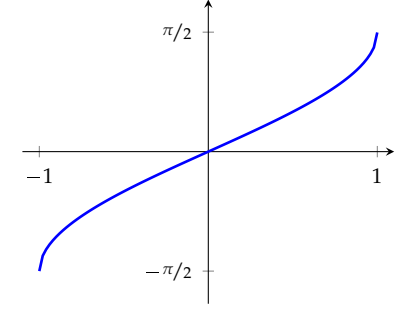
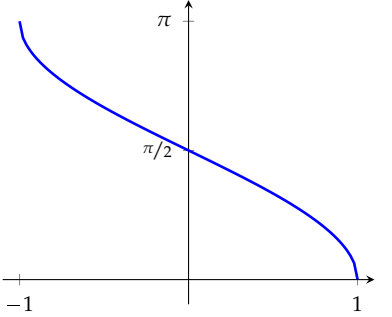
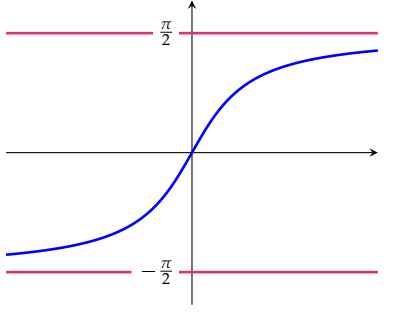
$$\sin(\alpha \pm \beta) = \sin(\alpha) \cos(\beta) \pm \cos(\alpha) \sin(\beta)$$

- Cosine

$$\cos(\alpha \pm \beta) = \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta)$$

## B.9▲ Inverse Trigonometric Functions

Some of you may not have studied inverse trigonometric functions in highschool, however we still expect you to know them by the end of the course.

$\arcsin x$	$\arccos x$	$\arctan x$
Domain: $-1 \leq x \leq 1$	Domain: $-1 \leq x \leq 1$	Domain: all real numbers
Range: $-\frac{\pi}{2} \leq \arcsin x \leq \frac{\pi}{2}$	Range: $0 \leq \arccos x \leq \pi$	Range: $-\frac{\pi}{2} < \arctan x < \frac{\pi}{2}$
		

Since these functions are inverses of each other we have

$$\arcsin(\sin \theta) = \theta$$

$$-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$

$$\arccos(\cos \theta) = \theta$$

$$0 \leq \theta \leq \pi$$

$$\arctan(\tan \theta) = \theta$$

$$-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}$$

and also

$$\sin(\arcsin x) = x$$

$$-1 \leq x \leq 1$$

$$\cos(\arccos x) = x$$

$$-1 \leq x \leq 1$$

$$\tan(\arctan x) = x$$

$$\text{any real } x$$

$\operatorname{arccsc} x$	$\operatorname{arcsec} x$	$\operatorname{arccot} x$
Domain: $ x  \geq 1$ Range: $-\frac{\pi}{2} \leq \operatorname{arccsc} x \leq \frac{\pi}{2}$ $\operatorname{arccsc} x \neq 0$	Domain: $ x  \geq 1$ Range: $0 \leq \operatorname{arcsec} x \leq \pi$ $\operatorname{arcsec} x \neq \frac{\pi}{2}$	Domain: all real numbers Range: $0 < \operatorname{arccot} x < \pi$

Again

$$\operatorname{arccsc}(\csc \theta) = \theta$$

$$-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}, \theta \neq 0$$

$$\operatorname{arcsec}(\sec \theta) = \theta$$

$$0 \leq \theta \leq \pi, \theta \neq \frac{\pi}{2}$$

$$\operatorname{arccot}(\cot \theta) = \theta$$

$$0 < \theta < \pi$$

and

$$\csc(\operatorname{arccsc} x) = x$$

$$|x| \geq 1$$

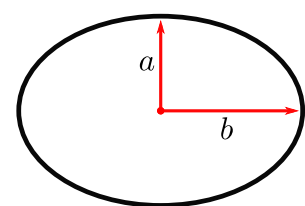
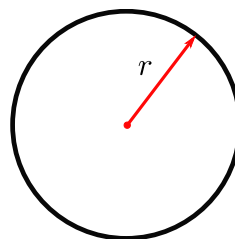
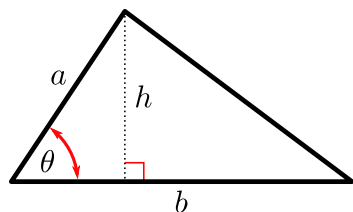
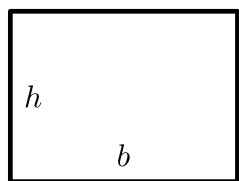
$$\sec(\operatorname{arcsec} x) = x$$

$$|x| \geq 1$$

$$\cot(\operatorname{arccot} x) = x$$

$$\text{any real } x$$

## B.10<sup>▲</sup> Areas



- Area of a rectangle

$$A = bh$$

- Area of a triangle

$$A = \frac{1}{2}bh = \frac{1}{2}ab \sin \theta$$

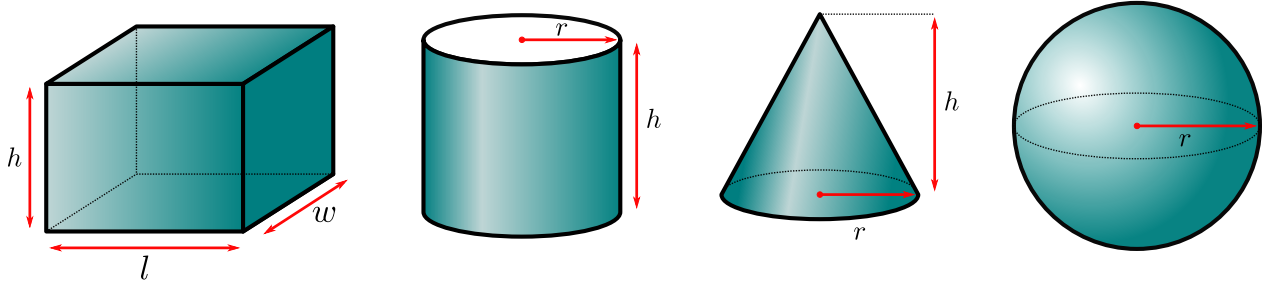
- Area of a circle

$$A = \pi r^2$$

- Area of an ellipse

$$A = \pi ab$$

## B.11▲ Volumes



- Volume of a rectangular prism

$$V = lwh$$

- Volume of a cylinder

$$V = \pi r^2 h$$

- Volume of a cone

$$V = \frac{1}{3}\pi r^2 h$$

- Volume of a sphere

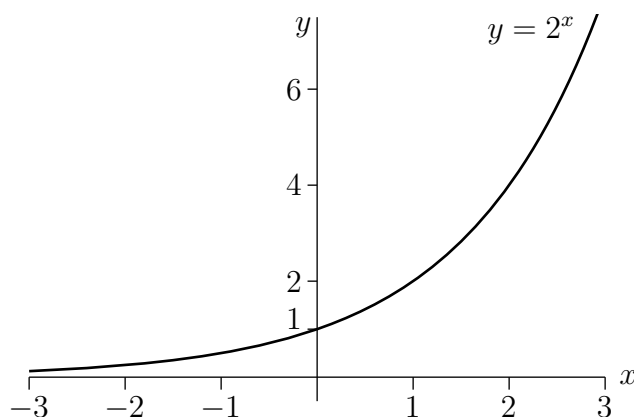
$$V = \frac{4}{3}\pi r^3$$

## B.12▲ Powers

In the following,  $x$  and  $y$  are arbitrary real numbers, and  $q$  is an arbitrary constant that is strictly bigger than zero.

- $q^0 = 1$

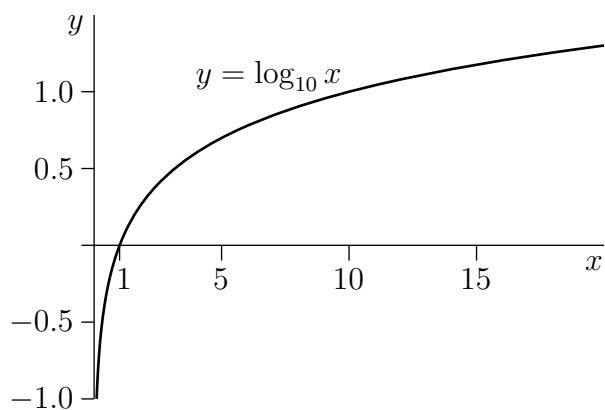
- $q^{x+y} = q^x q^y, q^{x-y} = \frac{q^x}{q^y}$
- $q^{-x} = \frac{1}{q^x}$
- $(q^x)^y = q^{xy}$
- $\lim_{x \rightarrow \infty} q^x = \infty, \lim_{x \rightarrow -\infty} q^x = 0$  if  $q > 1$
- $\lim_{x \rightarrow \infty} q^x = 0, \lim_{x \rightarrow -\infty} q^x = \infty$  if  $0 < q < 1$
- The graph of  $2^x$  is given below. The graph of  $q^x$ , for any  $q > 1$ , is similar.



## B.13▲ Logarithms

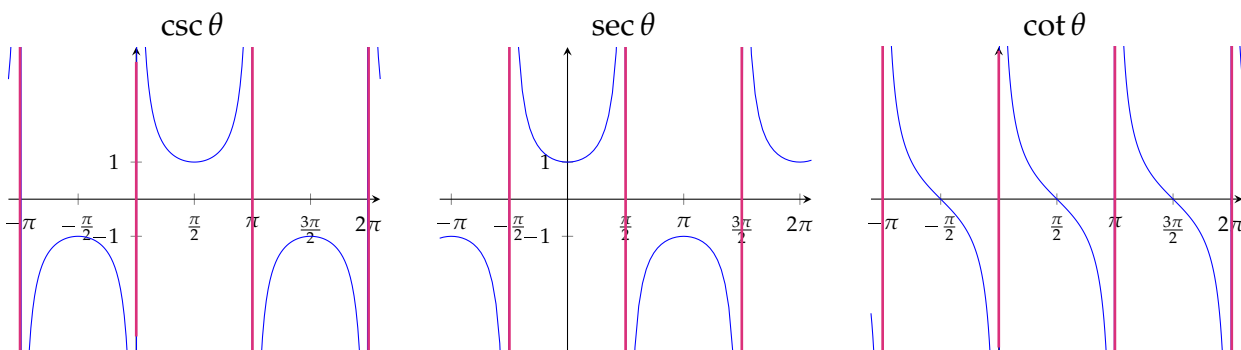
In the following,  $x$  and  $y$  are arbitrary real numbers that are strictly bigger than 0, and  $p$  and  $q$  are arbitrary constants that are strictly bigger than one.

- $q^{\log_q x} = x, \log_q (q^x) = x$
- $\log_q x = \frac{\log_p x}{\log_p q}$
- $\log_q 1 = 0, \log_q q = 1$
- $\log_q (xy) = \log_q x + \log_q y$
- $\log_q \left(\frac{x}{y}\right) = \log_q x - \log_q y$
- $\log_q \left(\frac{1}{y}\right) = -\log_q y,$
- $\log_q (x^y) = y \log_q x$
- $\lim_{x \rightarrow \infty} \log_q x = \infty, \lim_{x \rightarrow 0} \log_q x = -\infty$
- The graph of  $\log_{10} x$  is given below. The graph of  $\log_q x$ , for any  $q > 1$ , is similar.



## B.14▲ Highschool Material You Should be Able to Derive

- Graphs of  $\csc \theta$ ,  $\sec \theta$  and  $\cot \theta$ :



- More Pythagoras

$$\begin{array}{lcl} \sin^2 \theta + \cos^2 \theta = 1 & \xrightarrow{\text{divide by } \cos^2 \theta} & \tan^2 \theta + 1 = \sec^2 \theta \\ \sin^2 \theta + \cos^2 \theta = 1 & \xrightarrow{\text{divide by } \sin^2 \theta} & 1 + \cot^2 \theta = \csc^2 \theta \end{array}$$

- Sine — double angle (set  $\beta = \alpha$  in sine angle addition formula)

$$\sin(2\alpha) = 2 \sin(\alpha) \cos(\alpha)$$

- Cosine — double angle (set  $\beta = \alpha$  in cosine angle addition formula)

$$\begin{aligned} \cos(2\alpha) &= \cos^2(\alpha) - \sin^2(\alpha) \\ &= 2 \cos^2(\alpha) - 1 && \text{(use } \sin^2(\alpha) = 1 - \cos^2(\alpha)) \\ &= 1 - 2 \sin^2(\alpha) && \text{(use } \cos^2(\alpha) = 1 - \sin^2(\alpha)) \end{aligned}$$

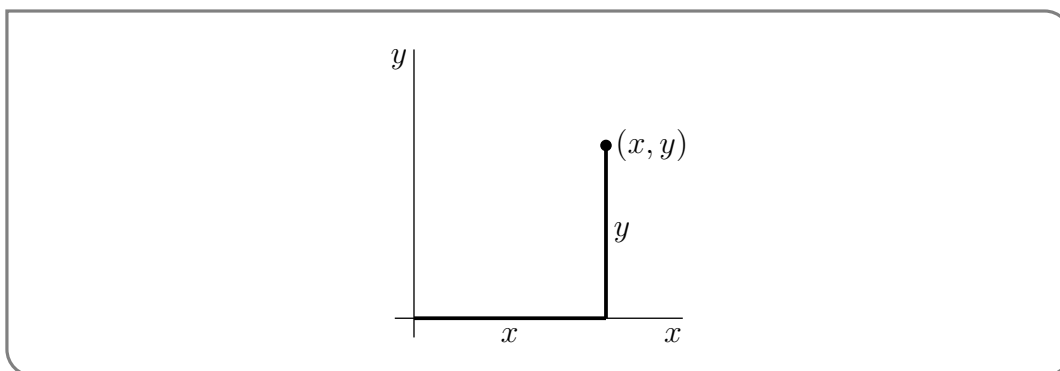
- Composition of trigonometric and inverse trigonometric functions:

$$\cos(\arcsin x) = \sqrt{1 - x^2} \qquad \sec(\arctan x) = \sqrt{1 + x^2}$$

and similar expressions.

## B.15▲ Cartesian Coordinates

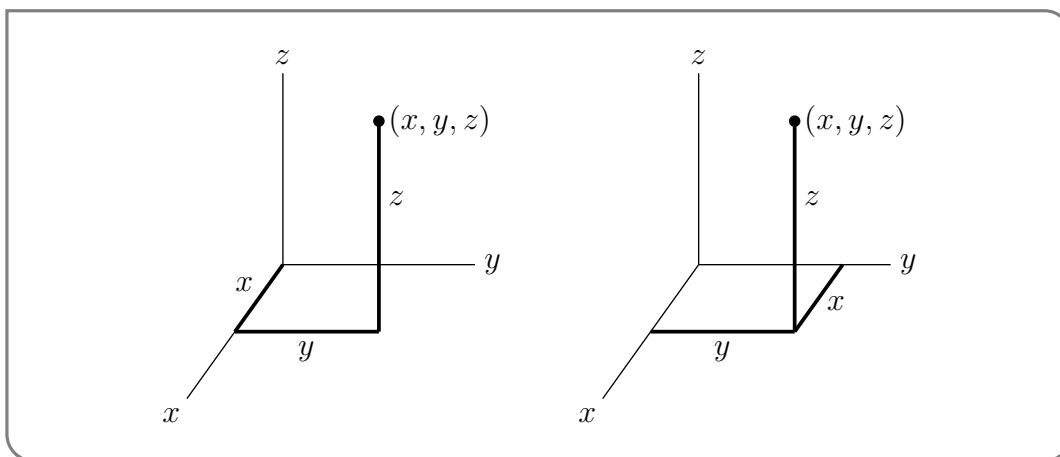
Each point in two dimensions may be labeled by two coordinates  $(x, y)$  which specify the position of the point in some units with respect to some axes as in the figure below.



The set of all points in two dimensions is denoted  $\mathbb{R}^2$ . Observe that

- the distance from the point  $(x, y)$  to the  $x$ -axis is  $|y|$
- the distance from the point  $(x, y)$  to the  $y$ -axis is  $|x|$
- the distance from the point  $(x, y)$  to the origin  $(0, 0)$  is  $\sqrt{x^2 + y^2}$

Similarly, each point in three dimensions may be labeled by three coordinates  $(x, y, z)$ , as in the two figures below.



The set of all points in three dimensions is denoted  $\mathbb{R}^3$ . The plane that contains, for example, the  $x$ - and  $y$ -axes is called the  $xy$ -plane.

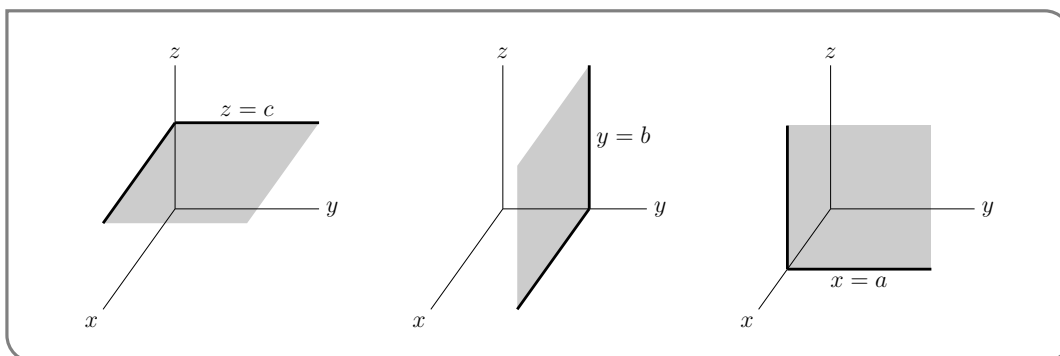
- The  $xy$ -plane is the set of all points  $(x, y, z)$  that obey  $z = 0$ .
- The  $xz$ -plane is the set of all points  $(x, y, z)$  that obey  $y = 0$ .
- The  $yz$ -plane is the set of all points  $(x, y, z)$  that obey  $x = 0$ .

More generally,

- The set of all points  $(x, y, z)$  that obey  $z = c$  is a plane that is parallel to the  $xy$ -plane and is a distance  $|c|$  from it. If  $c > 0$ , the plane  $z = c$  is above the  $xy$ -plane. If  $c < 0$ , the plane  $z = c$  is below the  $xy$ -plane. We say that the plane  $z = c$  is a signed distance  $c$  from the  $xy$ -plane.



- The set of all points  $(x, y, z)$  that obey  $y = b$  is a plane that is parallel to the  $xz$ -plane and is a signed distance  $b$  from it.
- The set of all points  $(x, y, z)$  that obey  $x = a$  is a plane that is parallel to the  $yz$ -plane and is a signed distance  $a$  from it.



Observe that

- the distance from the point  $(x, y, z)$  to the  $xy$ -plane is  $|z|$
- the distance from the point  $(x, y, z)$  to the  $xz$ -plane is  $|y|$
- the distance from the point  $(x, y, z)$  to the  $yz$ -plane is  $|x|$
- the distance from the point  $(x, y, z)$  to the origin  $(0, 0, 0)$  is  $\sqrt{x^2 + y^2 + z^2}$

The distance from the point  $(x, y, z)$  to the point  $(x', y', z')$  is

$$\sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}$$

so that the equation of the sphere centered on  $(1, 2, 3)$  with radius 4, that is, the set of all points  $(x, y, z)$  whose distance from  $(1, 2, 3)$  is 4, is

$$(x - 1)^2 + (y - 2)^2 + (z - 3)^2 = 16$$

## B.16<sup>▲</sup> Roots of Polynomials

Being able to factor polynomials is a very important part of many of the computations in this course. Related to this is the process of finding roots (or zeros) of polynomials. That is, given a polynomial  $P(x)$ , find all numbers  $r$  so that  $P(r) = 0$ .

In the case of a quadratic  $P(x) = ax^2 + bx + c$ , we can use the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

The corresponding formulas for cubics and quartics<sup>1</sup> are extremely cumbersome, and no such formula exists for polynomials of degree 5 and higher<sup>2</sup>.

1 The method for cubics was developed in the 15th century by del Ferro, Cardano and Ferrari (Cardano's student). Ferrari then went on to discover a formula for the roots of a quartic. His formula requires the solution of an associated cubic polynomial.

2 This is the famous Abel-Ruffini Theorem.

Despite this there are many tricks<sup>3</sup> for finding roots of polynomials that work well in some situations but not all. Here we describe approaches that will help you find integer and rational roots of polynomials that will work well on exams, quizzes and homework assignments.

Consider the quadratic equation  $x^2 - 5x + 6 = 0$ . We could<sup>4</sup> solve this using the quadratic formula

$$x = \frac{5 \pm \sqrt{25 - 4 \times 1 \times 6}}{2} = \frac{5 \pm 1}{2} = 2, 3.$$

Hence  $x^2 - 5x + 6$  has roots  $x = 2, 3$  and so it factors as  $(x - 3)(x - 2)$ . Notice<sup>5</sup> that the numbers 2 and 3 divide the constant term of the polynomial, 6. This happens in general and forms the basis of our first trick.

**Trick B.16.1** (A very useful trick).

If  $r$  or  $-r$  is an integer root of a polynomial  $P(x) = a_n x^n + \dots + a_1 x + a_0$  with integer coefficients, then  $r$  is a factor of the constant term  $a_0$ .

*Proof.* If  $r$  is a root of the polynomial we know that  $P(r) = 0$ . Hence

$$a_n \cdot r^n + \dots + a_1 \cdot r + a_0 = 0$$

If we isolate  $a_0$  in this expression we get

$$a_0 = -[a_n r^n + \dots + a_1 r]$$

We can see that  $r$  divides every term on the right-hand side. This means that the right-hand side is an integer times  $r$ . Thus the left-hand side, being  $a_0$ , is an integer times  $r$ , as required. The argument for when  $-r$  is a root is almost identical.  $\square$

Let us put this observation to work.

**Example B.16.1**

Find the integer roots of  $P(x) = x^3 - x^2 + 2$ .

*Solution.*

- The constant term in this polynomial is 2.
- The only divisors of 2 are 1, 2. So the only candidates for integer roots are  $\pm 1, \pm 2$ .

3 There is actually a large body of mathematics devoted to developing methods for factoring polynomials. Polynomial factorisation is a fundamental problem for most computer algebra systems. The interested reader should make use of their favourite search engine to find out more.

4 We probably shouldn't do it this way for such a simple polynomial, but for pedagogical purposes we do here.

5 Many of you may have been taught this approach in highschool.

- Trying each in turn

$$P(1) = 2$$

$$P(-1) = 0$$

$$P(2) = 6$$

$$P(-2) = -10$$

- Thus the only integer root is  $-1$ .

Example B.16.1

Example B.16.2

Find the integer roots of  $P(x) = 3x^3 + 8x^2 - 5x - 6$ .

*Solution.*

- The constant term is  $-6$ .
- The divisors of 6 are 1, 2, 3, 6. So the only candidates for integer roots are  $\pm 1, \pm 2, \pm 3, \pm 6$ .
- We try each in turn (it is tedious but not difficult):

$$P(1) = 0$$

$$P(-1) = 4$$

$$P(2) = 40$$

$$P(-2) = 12$$

$$P(3) = 132$$

$$P(-3) = 0$$

$$P(6) = 900$$

$$P(-6) = -336$$

- Thus the only integer roots are 1 and  $-3$ .

Example B.16.2

We can generalise this approach in order to find rational roots. Consider the polynomial  $6x^2 - x - 2$ . We can find its zeros using the quadratic formula:

$$x = \frac{1 \pm \sqrt{1 + 48}}{12} = \frac{1 \pm 7}{12} = -\frac{1}{2}, \frac{2}{3}.$$

Notice now that the numerators, 1 and 2, both divide the constant term of the polynomial (being 2). Similarly, the denominators, 2 and 3, both divide the coefficient of the highest power of  $x$  (being 6). This is quite general.

**Trick B.16.2** (Another nice trick).

If  $b/d$  or  $-b/d$  is a rational root in lowest terms (i.e.  $b$  and  $d$  are integers with no common factors) of a polynomial  $Q(x) = a_n x^n + \cdots + a_1 x + a_0$  with integer coefficients, then the numerator  $b$  is a factor of the constant term  $a_0$  and the denominator  $d$  is a factor of  $a_n$ .

*Proof.* Since  $b/d$  is a root of  $P(x)$  we know that

$$a_n(b/d)^n + \cdots + a_1(b/d) + a_0 = 0$$

Multiply this equation through by  $d^n$  to get

$$a_nb^n + \cdots + a_1bd^{n-1} + a_0d^n = 0$$

Move terms around to isolate  $a_0d^n$ :

$$a_0d^n = -[a_nb^n + \cdots + a_1bd^{n-1}]$$

Now every term on the right-hand side is some integer times  $b$ . Thus the left-hand side must also be an integer times  $b$ . We know that  $d$  does not contain any factors of  $b$ , hence  $a_0$  must be some integer times  $b$  (as required).

Similarly we can isolate the term  $a_nb^n$ :

$$a_nb^n = -[a_{n-1}b^{n-1}d + \cdots + a_1bd^{n-1} + a_0d^n]$$

Now every term on the right-hand side is some integer times  $d$ . Thus the left-hand side must also be an integer times  $d$ . We know that  $b$  does not contain any factors of  $d$ , hence  $a_n$  must be some integer times  $d$  (as required).

The argument when  $-b/d$  is a root is nearly identical. □

We should put this to work:

**Example B.16.3**

$$P(x) = 2x^2 - x - 3.$$

*Solution.*

- The constant term in this polynomial is  $3 = 1 \times 3$  and the coefficient of the highest power of  $x$  is  $2 = 1 \times 2$ .
- Thus the only candidates for integer roots are  $\pm 1, \pm 3$ .
- By our newest trick, the only candidates for fractional roots are  $\pm \frac{1}{2}, \pm \frac{3}{2}$ .
- We try each in turn<sup>6</sup>

$$\begin{array}{ll} P(1) = -2 & P(-1) = 0 \\ P(3) = 12 & P(-3) = 18 \\ P\left(\frac{1}{2}\right) = -3 & P\left(-\frac{1}{2}\right) = -2 \\ P\left(\frac{3}{2}\right) = 0 & P\left(-\frac{3}{2}\right) = 3 \end{array}$$

so the roots are  $-1$  and  $\frac{3}{2}$ .

<sup>6</sup> Again, this is a little tedious, but not difficult. It's actually pretty easy to code up for a computer to do. Modern polynomial factoring algorithms do more sophisticated things, but these are a pretty good way to start.

## Example B.16.3

The tricks above help us to find integer and rational roots of polynomials. With a little extra work we can extend those methods to help us factor polynomials. Say we have a polynomial  $P(x)$  of degree  $p$  and have established that  $r$  is one of its roots. That is, we know  $P(r) = 0$ . Then we can factor  $(x - r)$  out from  $P(x)$  — it is always possible to find a polynomial  $Q(x)$  of degree  $p - 1$  so that

$$P(x) = (x - r)Q(x)$$

In sufficiently simple cases, you can probably do this factoring by inspection. For example,  $P(x) = x^2 - 4$  has  $r = 2$  as a root because  $P(2) = 2^2 - 4 = 0$ . In this case,  $P(x) = (x - 2)(x + 2)$  so that  $Q(x) = (x + 2)$ . As another example,  $P(x) = x^2 - 2x - 3$  has  $r = -1$  as a root because  $P(-1) = (-1)^2 - 2(-1) - 3 = 1 + 2 - 3 = 0$ . In this case,  $P(x) = (x + 1)(x - 3)$  so that  $Q(x) = (x - 3)$ .

For higher degree polynomials we need to use something more systematic — long division.

**Trick B.16.3 (Long Division).**

Once you have found a root  $r$  of a polynomial, even if you cannot factor  $(x - r)$  out of the polynomial by inspection, you can find  $Q(x)$  by dividing  $P(x)$  by  $x - r$ , using the long division algorithm you learned<sup>7</sup> in school, but with 10 replaced by  $x$ .

## Example B.16.4

Factor  $P(x) = x^3 - x^2 + 2$ .

*Solution.*

- We can go hunting for integer roots of the polynomial by looking at the divisors of the constant term. This tells us to try  $x = \pm 1, \pm 2$ .
- A quick computation shows that  $P(-1) = 0$  while  $P(1), P(-2), P(2) \neq 0$ . Hence  $x = -1$  is a root of the polynomial and so  $x + 1$  must be a factor.
- So we divide  $\frac{x^3 - x^2 + 2}{x + 1}$ . The first term,  $x^2$ , in the quotient is chosen so that when you multiply it by the denominator,  $x^2(x + 1) = x^3 + x^2$ , the leading term,  $x^3$ , matches the leading term in the numerator,  $x^3 - x^2 + 2$ , exactly.

$$x + 1 \overline{) \begin{array}{r} x^3 - x^2 + 2 \\ x^3 + x^2 \phantom{+ 2} \\ \hline \phantom{x^3} - 2x^2 + 2 \end{array}} \quad \longleftarrow x^2(x + 1)$$

<sup>7</sup> This is a standard part of most highschool mathematics curricula, but perhaps not all. You should revise this carefully.

- When you subtract  $x^2(x + 1) = x^3 + x^2$  from the numerator  $x^3 - x^2 + 2$  you get the remainder  $-2x^2 + 2$ . Just like in public school, the 2 is not normally “brought down” until it is actually needed.

$$x + 1 \overline{\begin{array}{r} x^2 \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \end{array}} \longleftarrow x^2(x + 1)$$

- The next term,  $-2x$ , in the quotient is chosen so that when you multiply it by the denominator,  $-2x(x + 1) = -2x^2 - 2x$ , the leading term  $-2x^2$  matches the leading term in the remainder exactly.

$$x + 1 \overline{\begin{array}{r} x^2 - 2x \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \\ -2x^2 - 2x \\ \hline 2x + 2 \end{array}} \longleftarrow x^2(x + 1)$$

$$\phantom{x + 1} \phantom{\overline{\begin{array}{r} x^2 - 2x \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \\ -2x^2 - 2x \\ \hline 2x + 2 \end{array}}} \longleftarrow -2x(x + 1)$$

And so on.

$$x + 1 \overline{\begin{array}{r} x^2 - 2x + 2 \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \\ -2x^2 - 2x \\ \hline 2x + 2 \\ 2x + 2 \\ \hline 0 \end{array}} \longleftarrow x^2(x + 1)$$

$$\phantom{x + 1} \phantom{\overline{\begin{array}{r} x^2 - 2x + 2 \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \\ -2x^2 - 2x \\ \hline 2x + 2 \\ 2x + 2 \\ \hline 0 \end{array}}} \longleftarrow -2x(x + 1)$$

$$\phantom{x + 1} \phantom{\overline{\begin{array}{r} x^2 - 2x + 2 \\ x^3 - x^2 + 2 \\ x^3 + x^2 \\ \hline -2x^2 \\ -2x^2 - 2x \\ \hline 2x + 2 \\ 2x + 2 \\ \hline 0 \end{array}}} \longleftarrow 2(x + 1)$$

- Note that we finally end up with a remainder 0. A nonzero remainder would have signalled a computational error, since we know that the denominator  $x - (-1)$  must divide the numerator  $x^3 - x^2 + 2$  exactly.
- We conclude that

$$(x + 1)(x^2 - 2x + 2) = x^3 - x^2 + 2$$

To check this, just multiply out the left hand side explicitly.

- Applying the high school quadratic root formula  $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$  to  $x^2 - 2x + 2$  tells us that it has no real roots and that we cannot factor it further<sup>8</sup>.

Example B.16.4

<sup>8</sup> Because we are not permitted to use complex numbers.

We finish by describing an alternative to long division. The approach is roughly equivalent, but is perhaps more straightforward at the expense of requiring more algebra.

Example B.16.5

Factor  $P(x) = x^3 - x^2 + 2$ , again.

*Solution.* Let us do this again but avoid long division.

- From the previous example, we know that  $\frac{x^3 - x^2 + 2}{x + 1}$  must be a polynomial (since  $-1$  is a root of the numerator) of degree 2. So write

$$\frac{x^3 - x^2 + 2}{x + 1} = ax^2 + bx + c$$

for some, as yet unknown, coefficients  $a$ ,  $b$  and  $c$ .

- Cross multiplying and simplifying gives us

$$\begin{aligned} x^3 - x^2 + 2 &= (ax^2 + bx + c)(x + 1) \\ &= ax^3 + (a + b)x^2 + (b + c)x + c \end{aligned}$$

- Now matching coefficients of the various powers of  $x$  on the left and right hand sides

$$\text{coefficient of } x^3: \quad a = 1$$

$$\text{coefficient of } x^2: \quad a + b = -1$$

$$\text{coefficient of } x^1: \quad b + c = 0$$

$$\text{coefficient of } x^0: \quad c = 2$$

- This gives us a system of equations that we can solve quite directly. Indeed it tells us immediately that  $a = 1$  and  $c = 2$ . Subbing  $a = 1$  into  $a + b = -1$  tells us that  $1 + b = -1$  and hence  $b = -2$ .
- Thus

$$x^3 - x^2 + 2 = (x + 1)(x^2 - 2x + 2).$$

Example B.16.5

Appendix B of this work was taken from Appendix A of [CLP 2 – Integral Calculus](#) by Feldman, Rechnitzer, and Yeager under a [Create Commons Attribution-NonCommercial-ShareAlike 4.0 International license](#).