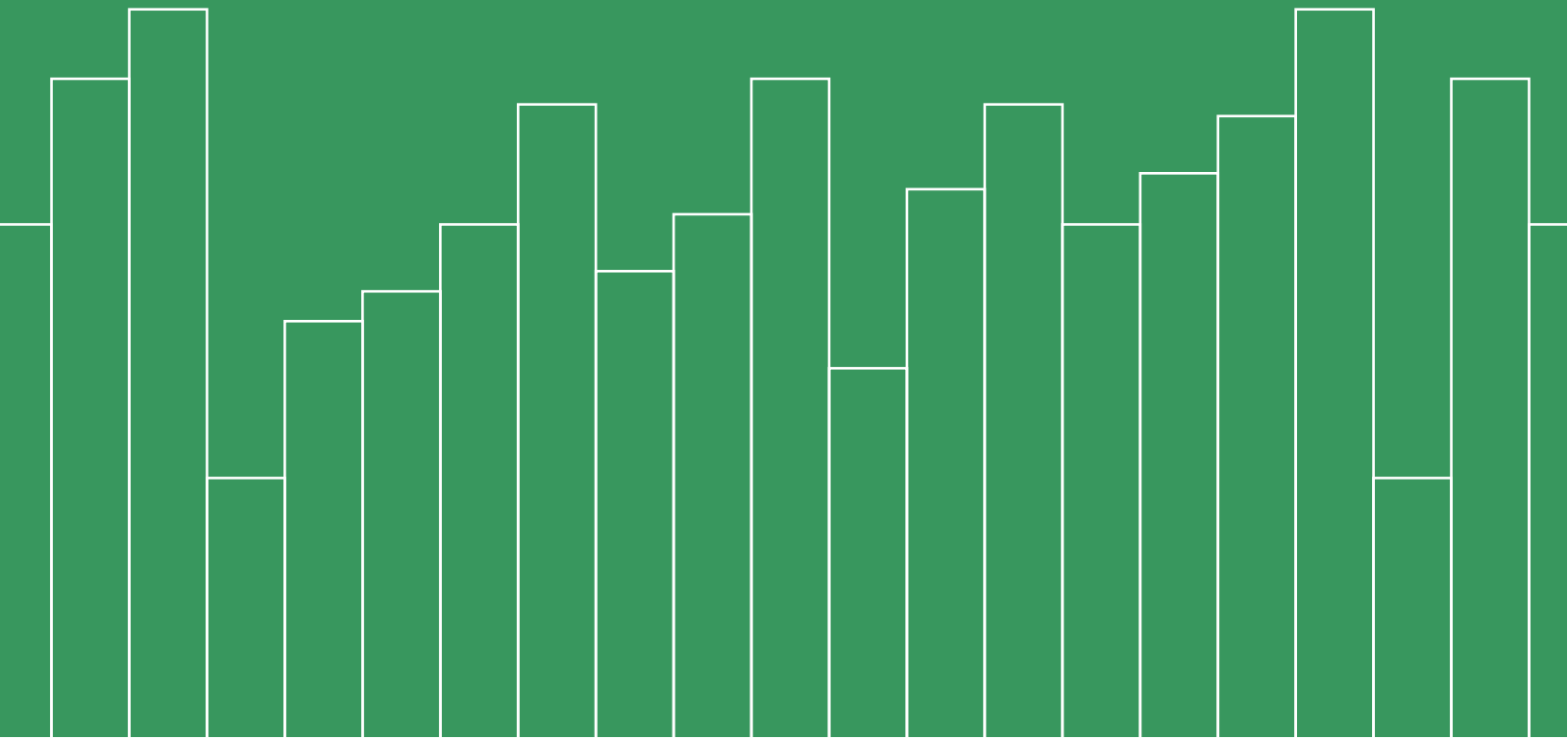
CLP 2

INTEGRAL CALCULUS

FELDMAN RECHNITZER YEAGER



CLP-2 INTEGRAL CALCULUS

Joel FELDMAN Andrew RECHNITZER Elyse YEAGER

THIS DOCUMENT WAS TYPESET ON FRIDAY 16TH AUGUST, 2024.

▸▸ Legal stuff

- Copyright © 2017–2024 Joel Feldman, Andrew Rechnitzer 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/.

- Links to the source files can be found at the text webpage

Contents

1 Integration **1**

1.1 Definition of the Integral . . . 1
1.1.1 A Motivating Example . . . 2
1.1.2 Optional — A More Rigorous Area Computation . . . 9
1.1.3 Summation Notation . . . 12
1.1.4 The Definition of the Definite Integral . . . 16
1.1.5 Using Known Areas to Evaluate Integrals . . . 23
1.1.6 Another Interpretation for Definite Integrals . . . 26
1.1.7 Optional — Careful Definition of the Integral . . . 27
1.2 Basic Properties of the Definite Integral . . . 32
1.2.1 More Properties of Integration: Even and Odd Functions . . . 41
1.2.2 Optional — More Properties of Integration: Inequalities for Integrals 44
1.3 The Fundamental Theorem of Calculus . . . 46
1.4 Substitution . . . 61
1.5 Area Between Curves . . . 75
1.6 Volumes . . . 87
1.7 Integration by Parts . . . 101
1.8 Trigonometric Integrals . . . 111
1.8.1 Integrating $\int \sin^m x \cos^n x \mathrm{d}x$ . . . 113
1.8.2 Integrating $\int \tan^m x \sec^n x \mathrm{d}x$ . . . 116
1.8.3 Optional — integrating $\sec x$, $\csc x$, $\sec^3 x$ and $\csc^3 x$ . . . 123
1.9 Trigonometric Substitution . . . 129
1.10 Partial Fractions . . . 139
1.10.1 Partial Fraction Decomposition Examples . . . 141
1.10.2 The Form of Partial Fraction Decompositions . . . 159
1.10.3 Optional — Justification of the Partial Fraction Decompositions . . . 164
1.11 Numerical Integration . . . 172
1.11.1 The Midpoint Rule . . . 175
1.11.2 The Trapezoidal Rule . . . 178
1.11.3 Simpson's Rule . . . 181
1.11.4 Three Simple Numerical Integrators – Error Behaviour . . . 185

1.11.5 Optional — An Error Bound for the Midpoint Rule . . . 192
1.12 Improper Integrals . . . 194
1.12.1 Definitions . . . 194
1.12.2 Examples . . . 200
1.12.3 Convergence Tests for Improper Integrals . . . 207

2 Applications of Integration 215
2.1 Work . . . 215
2.2 Averages . . . 222
2.3 Centre of Mass and Torque . . . 233
2.3.1 Centre of Mass . . . 233
2.3.2 Optional — Torque . . . 241
2.4 Separable Differential Equations . . . 245
2.4.1 Separate and Integrate . . . 246
2.4.2 Optional — Carbon Dating . . . 251
2.4.3 Optional — Newton's Law of Cooling . . . 254
2.4.4 Optional — Population Growth . . . 258
2.4.5 Optional — Mixing Problems . . . 263
2.4.6 Optional — Interest on Investments . . . 265

3 Sequences and Series 269
3.1 Sequences . . . 270
3.2 Series . . . 275
3.3 Convergence Tests . . . 288
3.3.1 The Divergence Test . . . 288
3.3.2 The Integral Test . . . 290
3.3.3 The Comparison Test . . . 296
3.3.4 The Alternating Series Test . . . 301
3.3.5 The Ratio Test . . . 306
3.3.6 Convergence Test List . . . 311
3.3.7 Optional — The Leaning Tower of Books . . . 311
3.3.8 Optional — The Root Test . . . 317
3.3.9 Optional — Harmonic and Basel Series . . . 320
3.3.10 Optional — Some Proofs . . . 322
3.4 Absolute and Conditional Convergence . . . 324
3.4.1 Definitions . . . 325
3.4.2 Optional — The Delicacy of Conditionally Convergent Series . . . 326
3.5 Power Series . . . 329
3.5.1 Radius and Interval of Convergence . . . 331
3.5.2 Working With Power Series . . . 339
3.6 Taylor Series . . . 346
3.6.1 Extending Taylor Polynomials . . . 346
3.6.2 Computing with Taylor Series . . . 362
3.6.3 Optional — Linking e^x with Trigonometric Functions . . . 368
3.6.4 Evaluating Limits using Taylor Expansions . . . 370
3.6.5 Optional — The Big O Notation . . . 372

3.6.6 Optional — Evaluating Limits Using Taylor Expansions — More Examples . . . 377
3.7 Optional — Rational and Irrational Numbers . . . 382

A High school material **392**
A.1 Similar Triangles . . . 392
A.2 Pythagoras . . . 393
A.3 Trigonometry — Definitions . . . 393
A.4 Radians, Arcs and Sectors . . . 393
A.5 Trigonometry — Graphs . . . 394
A.6 Trigonometry — Special Triangles . . . 394
A.7 Trigonometry — Simple Identities . . . 394
A.8 Trigonometry — Add and Subtract Angles . . . 395
A.9 Inverse Trigonometric Functions . . . 395
A.10 Areas . . . 396
A.11 Volumes . . . 397
A.12 Powers . . . 397
A.13 Logarithms . . . 398
A.14 Highschool Material You Should be Able to Derive . . . 399
A.15 Cartesian Coordinates . . . 400
A.16 Roots of Polynomials . . . 401

B Complex Numbers and Exponentials **408**
B.1 Definition and Basic Operations . . . 408
B.2 The Complex Exponential . . . 412
B.2.1 Definition and Basic Properties. . . . 412
B.2.2 Relationship with sin and cos. . . . 414
B.2.3 Polar Coordinates. . . . 415
B.2.4 Exploiting Complex Exponentials in Calculus Computations . . . 416
B.2.5 Exploiting Complex Exponentials in Differential Equation Computations . . . 418

C More About Numerical Integration **424**
C.1 Richardson Extrapolation . . . 424
C.2 Romberg Integration . . . 427
C.3 Adaptive Quadrature . . . 429

D Numerical Solution of ODE's **434**
D.1 Simple ODE Solvers — Derivation . . . 434
D.1.1 Euler's Method . . . 434
D.1.2 The Improved Euler's Method . . . 436
D.1.3 The Runge-Kutta Method . . . 439
D.2 Simple ODE Solvers — Error Behaviour . . . 441
D.2.1 Local Truncation Error for Euler's Method . . . 446
D.2.2 Global Truncation Error for Euler's Method . . . 448
D.3 Variable Step Size Methods . . . 450
D.3.1 Euler and Euler-2step (preliminary version) . . . 451

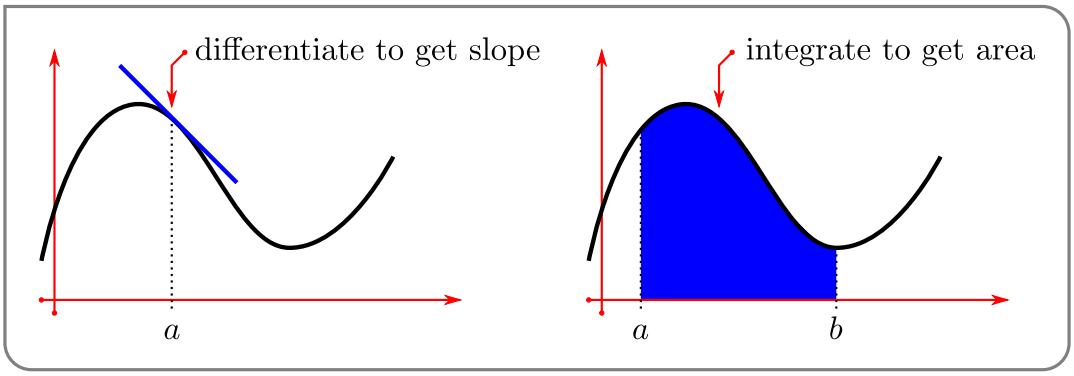
D.3.2 Euler and Euler-2step (final version) 454
D.3.3 Fehlberg's Method 457
D.3.4 The Kutta-Merson Process 458
D.3.5 The Local Truncation Error for Euler-2step 459

Chapter 1

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 and singular 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.



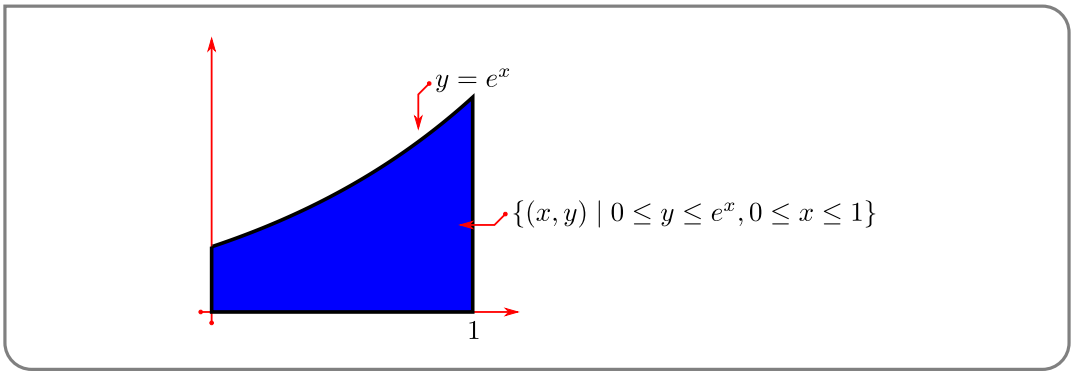
It is not immediately obvious that these two topics are related to each other. However, as we shall see, they are indeed intimately linked.

1.1▲ 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.

1.1.1 ▸▸ A Motivating Example

Let us find the area under the curve $y = e^x$ (and above the x–axis) for $0 \leqslant x \leqslant 1$. That is, the area of $\{\ (x,y)\ |\ 0 \leqslant y \leqslant e^x, 0 \leqslant x \leqslant 1\ \}$.



This area is equal to the "definite integral"

$$\text{Area} = \int_0^1 e^x \mathrm{d}x$$

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 1.1.18) 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 $\int_a^b f(x)\mathrm{d}x$.

Example 1.1.1

We wish to compute the area of $\{\ (x,y)\ |\ 0 \leqslant y \leqslant e^x, 0 \leqslant x \leqslant 1\ \}$. 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[1]. We choose[2] 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[3].

As a warm up exercise, we'll now just use four rectangles. In Example 1.1.2, below, we'll consider an arbitrary number of rectangles and then take the limit as the number of rectangles goes to infinity. So

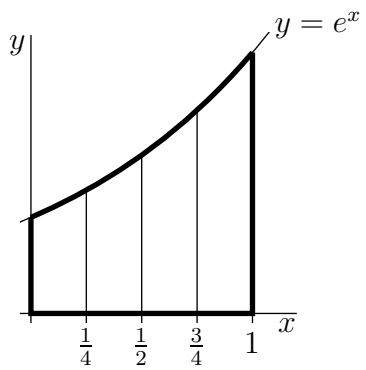
1 This should remind the reader of the approach taken to compute the slope of a tangent line way way back at the start of differential calculus.

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.

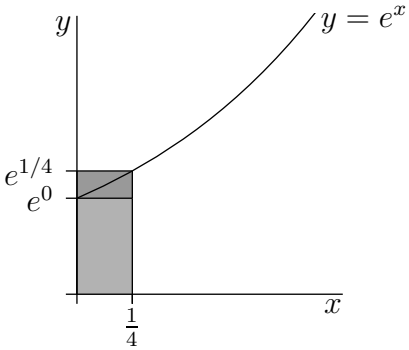
- subdivide the interval $0 \leqslant x \leqslant 1$ into 4 equal subintervals each of width $1/4$, and
- subdivide the area of interest into four corresponding vertical strips, as in the figure below.

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.

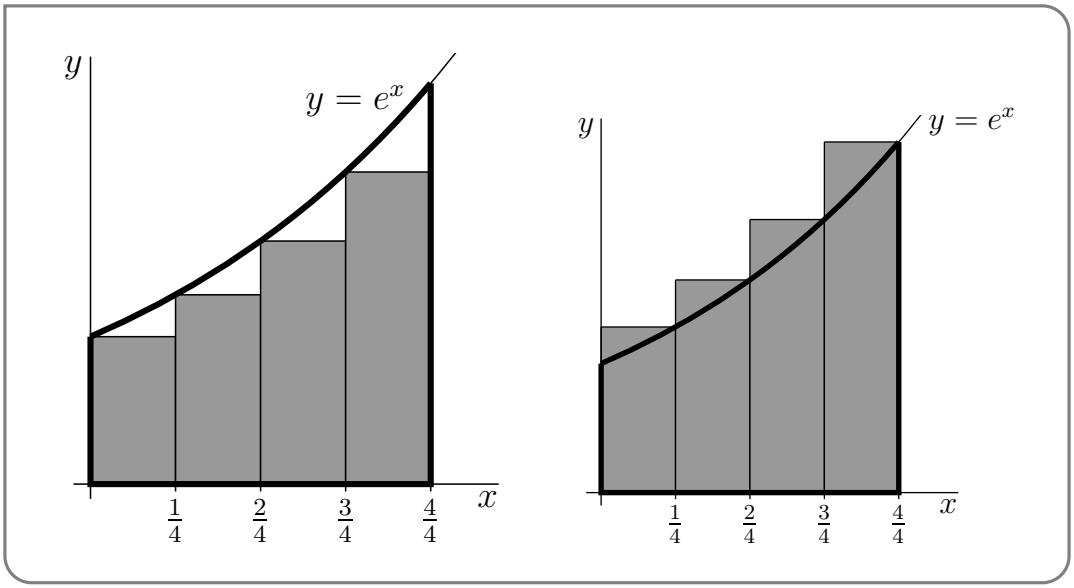
In this example we'll do two sample computations.

- 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 is called the "left Riemann sum approximation to $\int_0^1 e^x dx$ with 4 subintervals". We'll explain this terminology later.
- 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 is called the "right Riemann sum approximation to $\int_0^1 e^x dx$ with 4 subintervals".
- 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 1.1.1

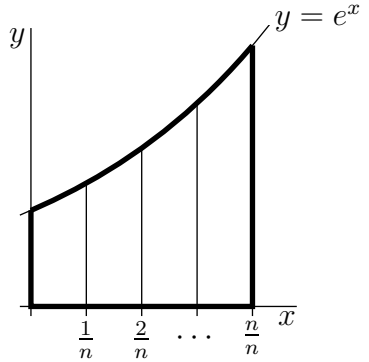
Now for the full computation that gives the exact area.

Example 1.1.2

Recall that we wish to compute the area of $\{\ (x,y)\ |\ 0 \leqslant y \leqslant e^x,\ 0 \leqslant x \leqslant 1\ \}$ 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 1.1.1, 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 $n \to \infty$. So

- pick a natural number n and
- subdivide the interval $0 \leqslant x \leqslant 1$ into n equal subintervals each of width $1/n$, and
- subdivide the area of interest into corresponding thin strips, as in the figure below.

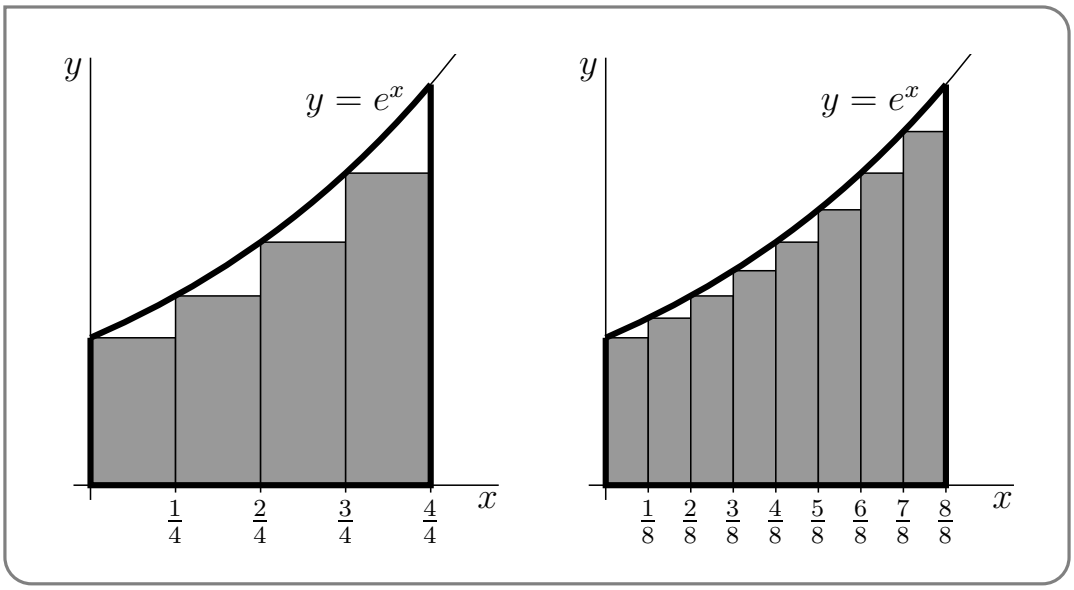
The area we want is exactly the sum of the areas of all of the thin strips.



Each of these strips is almost, but not quite, a rectangle. As in Example 1.1.1, 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 1.1.1, it doesn't matter. We shall shortly take the limit $n \to \infty$ 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 there will be an (optional) section shortly that 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[4]. 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 $\frac{1}{n}e^{(n-1)/n}$.

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 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

- 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

$$\text{Total approximating area} = \frac{1}{n}\left(1 + r + r^2 + \cdots + r^{n-1}\right)$$

The sum in brackets is known as a geometric sum and satisfies a nice simple formula:

Equation 1.1.3 (Geometric sum).

$$1 + r + r^2 + \cdots + r^{n-1} = \frac{r^n - 1}{r - 1} \qquad \text{provided } r \neq 1$$

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

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

Notice that if we multiply the whole sum by r we get back almost the same thing:

$$\begin{aligned}rS &= r\left(1 + r + r^2 + \cdots + r^{n-1}\right) \\ &= r + r^2 + r^3 + \cdots + r^n\end{aligned}$$

This right hand side differs from the original sum S only in that

- the right hand side, which starts with "$r+$", 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.

$$\begin{aligned}\text{Total approximating area} &= \frac{1}{n}\left(1 + r + r^2 + \cdots + r^{n-1}\right)\\ &= \frac{1}{n}\frac{r^n - 1}{r-1} && \text{remember that } r = e^{1/n}\\ &= \frac{1}{n}\frac{e^{n/n} - 1}{e^{1/n} - 1}\\ &= \frac{1}{n}\frac{e-1}{e^{1/n} - 1}\end{aligned}$$

To get the exact area[5] all we need to do is make the approximation better and better by taking the limit $n \to \infty$. The limit will look more familiar if we rename $1/n$ to X. As n tends to infinity, X tends to 0, so

$$\begin{aligned}\text{Area} &= \lim_{n\to\infty} \frac{1}{n}\,\frac{e-1}{e^{1/n} - 1}\\ &= (e-1)\lim_{n\to\infty} \frac{1/n}{e^{1/n} - 1}\\ &= (e-1)\lim_{X\to 0} \frac{X}{e^X - 1} && (\text{with } X = 1/n)\end{aligned}$$

Examining this limit we see that both numerator and denominator tend to zero as $X \to 0$, 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:

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.

- Perhaps the easiest way to compute the limit is by using l'Hôpital's rule[6]. Since both numerator and denominator go to zero, this is a $0/0$ indeterminate form. Thus

$$\lim_{X\to 0}\frac{X}{e^X-1}=\lim_{X\to 0}\frac{\frac{\mathrm{d}}{\mathrm{d}X}X}{\frac{\mathrm{d}}{\mathrm{d}X}(e^X-1)}=\lim_{X\to 0}\frac{1}{e^X}=1$$

- Another way[7] 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

$$\lim_{X\to 0}\frac{X}{e^X-1}=\left[\lim_{X\to 0}\frac{e^X-1}{X}\right]^{-1}$$

provided this second limit exists and is nonzero. This second limit should look a little familiar:

$$\lim_{X\to 0}\frac{e^X-1}{X}=\lim_{X\to 0}\frac{e^X-e^0}{X-0}$$

which is just the definition of the derivative of e^x at $x=0$. Hence we have

$$\begin{aligned}\lim_{X\to 0}\frac{X}{e^X-1}&=\left[\lim_{X\to 0}\frac{e^X-e^0}{X-0}\right]^{-1}\\&=\left[\frac{\mathrm{d}}{\mathrm{d}X}e^X\Big|_{X=0}\right]^{-1}\\&=\left[e^X\Big|_{X=0}\right]^{-1}\\&=1\end{aligned}$$

So, after this short aside into limits, we may now conclude that

$$\begin{aligned}\text{Area}&=(e-1)\lim_{X\to 0}\frac{X}{e^X-1}\\&=e-1\end{aligned}$$

Example 1.1.2

1.1.2 ▸▸ Optional — A More Rigorous Area Computation

In Example 1.1.1 above we considered the area of the region $\{\,(x,y)\mid 0\leqslant y\leqslant e^x,\ 0\leqslant x\leqslant 1\,\}$. 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

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.

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[8] 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 $\frac{\mathrm{d}}{\mathrm{d}x}e^x = e^x$ is always positive. Consequently, the smallest and largest values of e^x on the interval $a \leqslant x \leqslant b$ are e^a and e^b, respectively.

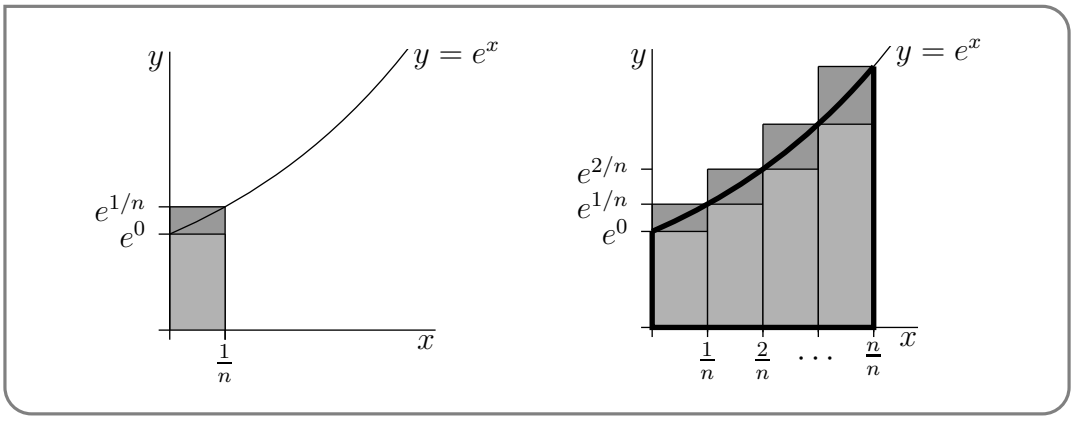
- In particular, for $0 \leqslant x \leqslant 1/n$, e^x takes values only between e^0 and $e^{1/n}$. As a result, the first strip

$$\left\{ (x,y) \,\middle|\, 0 \leqslant x \leqslant 1/n,\ 0 \leqslant y \leqslant e^x \right\}$$

  - contains the rectangle of $0 \leqslant x \leqslant 1/n$, $0 \leqslant y \leqslant e^0$ (the lighter rectangle in the figure on the left below) and
  - is contained in the rectangle $0 \leqslant x \leqslant 1/n$, $0 \leqslant y \leqslant e^{1/n}$ (the largest rectangle in the figure on the left below).

  Hence

$$\frac{1}{n}e^0 \leqslant \text{Area}\left\{ (x,y) \,\middle|\, 0 \leqslant x \leqslant 1/n,\ 0 \leqslant y \leqslant e^x \right\} \leqslant \frac{1}{n}e^{1/n}$$



8 Recall that if we have 3 functions $f(x), g(x), h(x)$ that satisfy $f(x) \leqslant g(x) \leqslant h(x)$ and we know that $\lim_{x\to a} f(x) = \lim_{x\to a} h(x) = L$ exists and is finite, then the squeeze theorem tells us that $\lim_{x\to a} g(x) = L$.

- Similarly, for the second, third, ..., last strips, as in the figure on the right above,

$$\begin{aligned}
\frac{1}{n}e^{1/n} &\leqslant \text{Area}\{\ (x,y)\ |\ 1/n \leqslant x \leqslant 2/n,\ 0 \leqslant y \leqslant e^x\ \} &&\leqslant \frac{1}{n}e^{2/n}\\
\frac{1}{n}e^{2/n} &\leqslant \text{Area}\{\ (x,y)\ |\ 2/n \leqslant x \leqslant 3/n,\ 0 \leqslant y \leqslant e^x\ \} &&\leqslant \frac{1}{n}e^{3/n}\\
&\ \ \vdots &&\ \ \vdots\\
\frac{1}{n}e^{(n-1)/n} &\leqslant \text{Area}\{\ (x,y)\ |\ (n-1)/n \leqslant x \leqslant n/n,\ 0 \leqslant y \leqslant e^x\ \} &&\leqslant \frac{1}{n}e^{n/n}
\end{aligned}$$

- Adding these n inequalities together gives

$$\begin{aligned}
&\frac{1}{n}\left(1 + e^{1/n} + \cdots + e^{(n-1)/n}\right)\\
&\qquad\leqslant \text{Area}\{\ (x,y)\ |\ 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant e^x\ \}\\
&\qquad\qquad\leqslant \frac{1}{n}\left(e^{1/n} + e^{2/n} + \cdots + e^{n/n}\right)
\end{aligned}$$

- We can then recycle equation (1.1.3) with $r = e^{1/n}$, so that $r^n = \left(e^{1/n}\right)^n = e$. Thus we have

$$\frac{1}{n}\frac{e-1}{e^{1/n}-1} \leqslant \text{Area}\{\ (x,y)\ |\ 0 \leqslant x \leqslant 1,\ 0 \leqslant y \leqslant e^x\ \} \leqslant \frac{1}{n}e^{1/n}\frac{e-1}{e^{1/n}-1}$$

  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 $\lim_{n\to\infty}\frac{1}{n}\frac{e-1}{e^{1/n}-1}$ and $\lim_{n\to\infty}\frac{1}{n}e^{1/n}\frac{e-1}{e^{1/n}-1}$, respectively. As we did near the end of Example 1.1.2, we make these limits look more familiar by renaming $1/n$ to X. As n tends to infinity, X tends to 0, so the limits of the lower and upper bounds are

$$\lim_{n\to\infty}\frac{1}{n}\frac{e-1}{e^{1/n}-1} = (e-1)\lim_{X=1/n\to 0}\frac{X}{e^X-1} = e-1$$

  (by l'Hôpital's rule) and

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}e^{1/n}\frac{e-1}{e^{1/n}-1} &= (e-1)\lim_{X=1/n\to 0}\cdot\frac{Xe^X}{e^X-1}\\
&= (e-1)\lim_{X\to 0}e^X\cdot\lim_{X=\to 0}\frac{X}{e^X-1}\\
&= (e-1)\cdot 1\cdot 1
\end{aligned}$$

  Thus, since the exact area is trapped between the lower and upper bounds, the squeeze theorem then implies that

$$\text{Exact area} = e-1.$$

1.1.3 ▸▸ 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 Σ (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 1.1.4.

Let $m \leqslant n$ 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 $\{a_m, \ldots, a_n\}$.

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[9]; 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.

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 1.1.5 (Arithmetic of Summation Notation).

Let $n \geqslant m$ be integers. Then for all real numbers c and a_i, b_i, $m \leqslant i \leqslant n$.

(a) $\displaystyle\sum_{i=m}^{n} ca_i = c\left(\sum_{i=m}^{n} a_i\right)$

(b) $\displaystyle\sum_{i=m}^{n} (a_i + b_i) = \left(\sum_{i=m}^{n} a_i\right) + \left(\sum_{i=m}^{n} b_i\right)$

(c) $\displaystyle\sum_{i=m}^{n} (a_i - b_i) = \left(\sum_{i=m}^{n} a_i\right) - \left(\sum_{i=m}^{n} b_i\right)$

9 Some careful addition shows it is $\frac{46181}{176400}$.

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[10]. 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[11]. Here are some that can. The first few are used a lot.

Theorem 1.1.6.

(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 $n \geqslant 0$.

(b) $\sum_{i=1}^{n} 1 = n$, for all integers $n \geqslant 1$.

(c) $\sum_{i=1}^{n} i = \frac{1}{2}n(n+1)$, for all integers $n \geqslant 1$.

(d) $\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$, for all integers $n \geqslant 1$.

(e) $\sum_{i=1}^{n} i^3 = \left[\frac{1}{2}n(n+1)\right]^2$, for all integers $n \geqslant 1$.

10 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 3.

11 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

$$\sum_{k=m}^{n} r^k = \frac{r^{n+1} - r^m}{r-1} \qquad \text{provided } r \neq 1$$

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.

▸▸▸ Proof of Theorem 1.1.6 (Optional)

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 (1.1.3), 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 third and fourth sums are discussed in the appendix of the CLP-1 text. In that discussion certain "tricks" are used to compute the sums with only simple arithmetic. Those tricks do not easily generalise to the fifth sum.

(c') Instead of repeating that appendix, we'll derive the third sum using a trick that generalises to the fourth and fifth sums (and also to higher powers). The trick uses the generating function[12] $S(x)$:

Equation 1.1.7.

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

Notice that this is just the geometric sum given by equation 1.1.3 with n replaced by $n+1$.

Now, consider the limit

$$\begin{aligned}
\lim_{x\to 1} S(x) &= \lim_{x\to 1} \left(1 + x + x^2 + \cdots + x^n\right) = n+1 && \text{but also} \\
&= \lim_{x\to 1} \frac{x^{n+1} - 1}{x - 1} && \text{now use l'Hôpital's rule} \\
&= \lim_{x\to 1} \frac{(n+1)x^n}{1} = n+1.
\end{aligned}$$

This is not so hard (or useful). But now consider the derivative of $S(x)$:

$$\begin{aligned}
S'(x) &= 1 + 2x + 3x^2 + \cdots + nx^{n-1} \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{x^{n+1} - 1}{x - 1}\right] && \text{use the quotient rule} \\
&= \frac{(x-1)\cdot(n+1)x^n - (x^{n+1} - 1)\cdot 1}{(x-1)^2} && \text{now clean it up} \\
&= \frac{nx^{n+1} - (n+1)x^n + 1}{(x-1)^2}.
\end{aligned}$$

12 Generating functions are frequently used in mathematics to analyse sequences and series, but are beyond the scope of the course. The interested reader should take a look at "Generatingfunctionology" by Herb Wilf. It is an excellent book and is also free to download.

Hence if we take the limit of the above expression as $x \to 1$ we recover

$$\begin{aligned}
\lim_{x\to 1} S'(x) &= 1+2+3+\cdots+n \\
&= \lim_{x\to 1} \frac{nx^{n+1}-(n+1)x^n+1}{(x-1)^2} && \text{now use l'Hôpital's rule} \\
&= \lim_{x\to 1} \frac{n(n+1)x^n-n(n+1)x^{n-1}}{2(x-1)} && \text{l'Hôpital's rule again} \\
&= \lim_{x\to 1} \frac{n^2(n+1)x^{n-1}-n(n+1)(n-1)x^{n-2}}{2} \\
&= \frac{n^2(n+1)-n(n-1)(n+1)}{2} = \frac{n(n+1)}{2}
\end{aligned}$$

as required. This computation can be done without l'Hôpital's rule, but the manipulations required are a fair bit messier.

(d) The derivation of the fourth and fifth sums is similar to, but even more tedious than, that of the third sum. One takes two or three derivatives of the generating function.

□

1.1.4 ▸▸ The Definition of the Definite Integral

In this section we give a definition of the definite integral $\int_a^b f(x)\mathrm{d}x$ generalising the machinery we used in Example 1.1.1. But first some terminology and a couple of remarks to better motivate the definition.

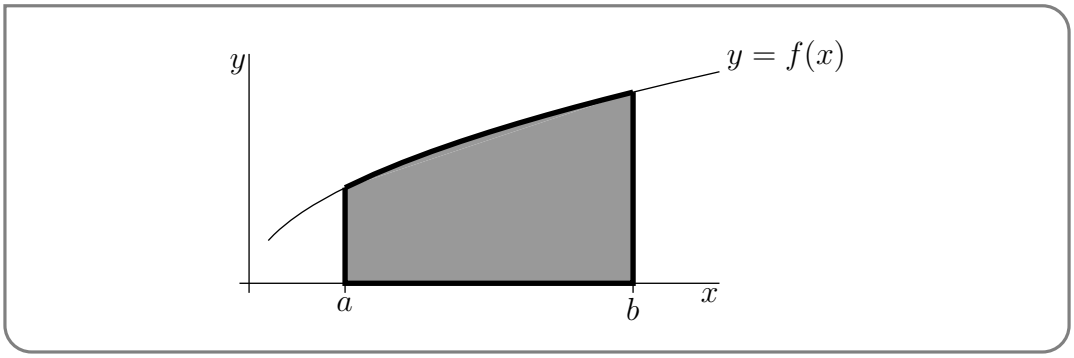
Notation 1.1.8.

The symbol $\int_a^b f(x)\mathrm{d}x$ is read "the definite integral of the function $f(x)$ from a to b". The function $f(x)$ is called the integrand of $\int_a^b f(x)\mathrm{d}x$ and a and b are called[13] the limits of integration. The interval $a \leqslant x \leqslant b$ 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 $f(x) \geqslant 0$ and $a \leqslant b$, one interpretation of the symbol $\int_a^b f(x)\mathrm{d}x$ is "the area of the region $\{\ (x,y) \mid a \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\ \}$".

13 a and b are also called the bounds of integration.



In this way we can rewrite the area in Example 1.1.1 as the definite integral $\int_0^1 e^x \mathrm{d}x$.

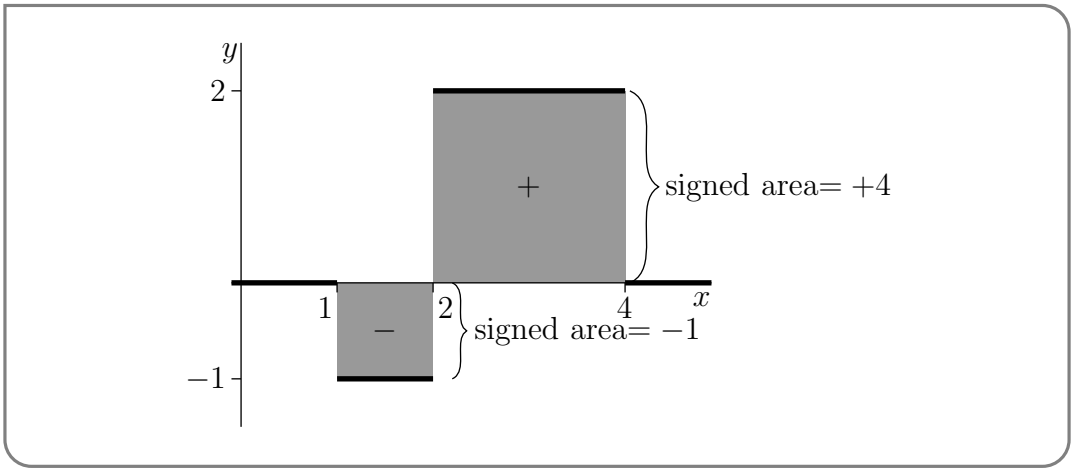
- This interpretation breaks down when either $a > b$ or $f(x)$ is not always positive, but it can be repaired by considering "signed areas".

- If $a \leqslant b$, but $f(x)$ is not always positive, one interpretation of $\int_a^b f(x)\mathrm{d}x$ is "the signed area between $y = f(x)$ and the x–axis for $a \leqslant x \leqslant b$". 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

$$f(x) = \begin{cases} -1 & \text{if } 1 \leqslant x \leqslant 2 \\ 2 & \text{if } 2 < x \leqslant 4 \\ 0 & \text{otherwise} \end{cases}$$

The 2×2 shaded square above the x–axis has signed area $+2 \times 2 = +4$. The 1×1 shaded square below the x–axis has signed area $-1 \times 1 = -1$. So, for this $f(x)$,

$$\int_0^5 f(x)\mathrm{d}x = +4 - 1 = 3$$



- We'll come back to the case $b < a$ later.

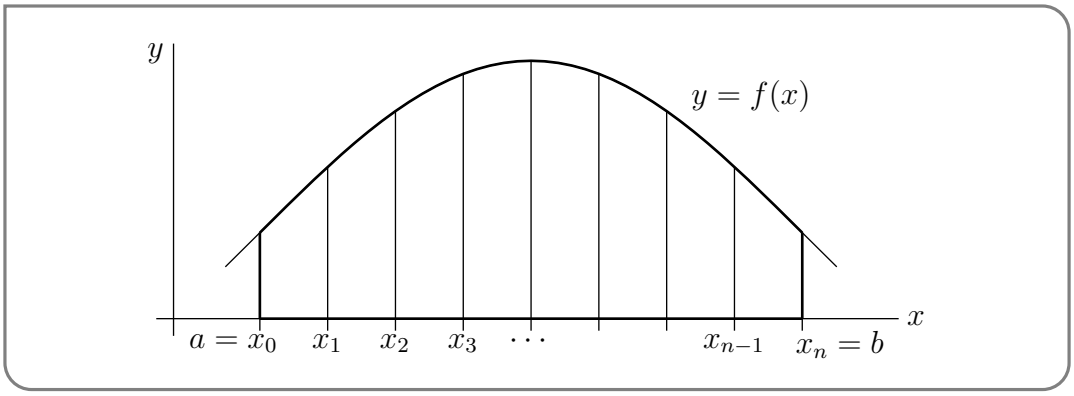
We're now ready to define $\int_a^b f(x)\mathrm{d}x$. The definition is a little involved, but essentially mimics what we did in Example 1.1.1 (which is why we did the example before the definition). The main differences are that we replace the function e^x by a generic function $f(x)$ and we replace the interval from 0 to 1 by the generic interval[14] 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 1.1.1 we will eventually take the limit as $n \to \infty$, which squeezes the width of each subinterval down to zero.

- For each integer $0 \leqslant i \leqslant n$, 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 $x_{i-1} \leqslant x \leqslant x_i$. 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 $x_2 = a + 2\frac{b-a}{n}$.



- On each subinterval we now pick $x_{i,n}^*$ between x_{i-1} and x_i. We then approximate $f(x)$ on the i^{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.

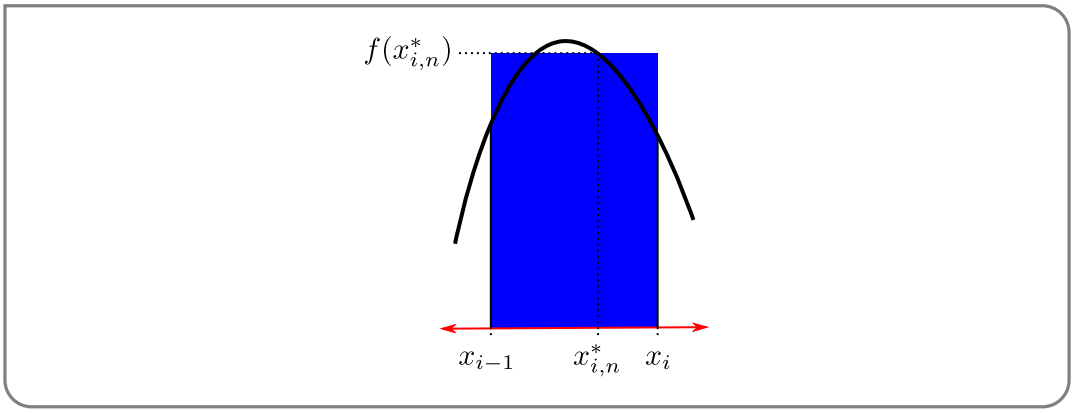
  Geometrically, we're approximating the region

$$\left\{ (x,y) \;\middle|\; x \text{ is between } x_{i-1} \text{ and } x_i, \text{ and } y \text{ is between 0 and } f(x) \right\}$$

  by the rectangle

$$\left\{ (x,y) \;\middle|\; x \text{ is between } x_{i-1} \text{ and } x_i, \text{ and } y \text{ is between 0 and } f(x_{i,n}^*) \right\}$$

14 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.



In Example 1.1.1 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 $n \to \infty$.

Oof! This is quite an involved process, but we can now write down the definition we need.

Definition 1.1.9.

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

$$\int_a^b f(x)\mathrm{d}x = \lim_{n\to\infty} \sum_{i=1}^{n} f(x^*_{i,n}) \cdot \frac{b-a}{n}$$

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 1.1.10.

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 1.1.7. 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 1.1.9, 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 $\int_a^b f(x)\mathrm{d}x$ is still defined by the formula of Definition 1.1.9. We'll get an interpretation for $\int_a^b f(x)\mathrm{d}x$, when $a > b$, later.

- It is important to note that the definite integral $\int_a^b f(x)\mathrm{d}x$ 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 1.1.3). The integration variable does not have to be called x. For example

$$\int_a^b f(x)\mathrm{d}x = \int_a^b f(t)\mathrm{d}t = \int_a^b f(u)\mathrm{d}u$$

  Just as with summation variables, the integration variable x has no meaning outside of $f(x)\mathrm{d}x$. For example

$$x\int_0^1 e^x\mathrm{d}x \qquad \text{and} \qquad \int_0^x e^x\mathrm{d}x$$

  are both gibberish.

The sum inside definition 1.1.9 is named after Bernhard Riemann[15] who made the first rigorous definition of the definite integral and so placed integral calculus on rigorous footings.

15 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).

Definition 1.1.11.

The sum inside definition 1.1.9

$$\sum_{i=1}^{n} f(x_{i,n}^*)\,\frac{b-a}{n}$$

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 each $x_{i,n}^* = x_{i-1} = a + (i-1)\frac{b-a}{n}$ to be the left hand end point of the i^{th} interval, $[x_{i-1}, x_i]$, we get the approximation

$$\sum_{i=1}^{n} f\left(a + (i-1)\frac{b-a}{n}\right)\,\frac{b-a}{n}$$

  which is called the "left Riemann sum approximation to $\int_a^b f(x)\mathrm{d}x$ with n subintervals". This is the approximation used in Example 1.1.1.

- In the same way, 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 $\int_a^b f(x)\mathrm{d}x$ 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.

- A third commonly used approximation is

$$\sum_{i=1}^{n} f\left(a + (i-1/2)\frac{b-a}{n}\right)\,\frac{b-a}{n}$$

  which is called the "midpoint Riemann sum approximation to $\int_a^b f(x)\mathrm{d}x$ with n subintervals". The word "midpoint" signifies that, on each subinterval $[x_{i-1}, x_i]$ we approximate f by its value at the midpoint, $\frac{x_{i-1}+x_i}{2} = a + (i-1/2)\frac{b-a}{n}$, 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 1.1.1.

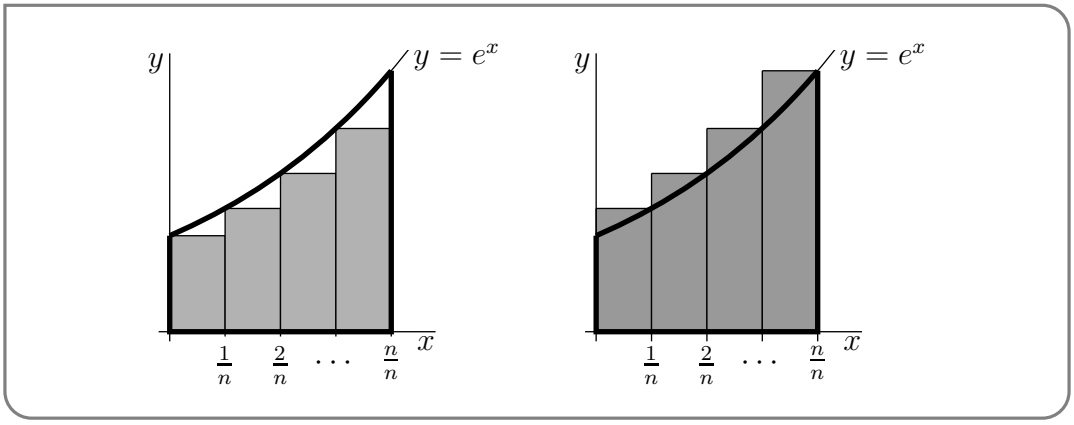
Example 1.1.12

Let's say we are again interested in the integral $\int_0^1 e^x \mathrm{d}x$. 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 $x_{i-1} \leqslant x \leqslant x_i$

- by its smallest value on the subinterval, namely $f(x_{i-1})$, when we compute the left Riemann sum approximation 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 left Riemann sum approximation 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 left Riemann sum describes an area smaller than the definite integral while the right Riemann sum gives an area larger[16] than the integral.

When we approximate the integral $\int_0^1 e^x \mathrm{d}x$ using n subintervals, then, on interval number i,

- x runs from $\frac{i-1}{n}$ to $\frac{i}{n}$ and

16 When a function is decreasing the situation is reversed — the left Riemann sum 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.

- $y = e^x$ 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 left Riemann sum approximation to $\int_0^1 e^x \mathrm{d}x$ 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

$$\sum_{i=1}^{n} e^{(i-1)/n} \frac{1}{n} \leqslant \int_0^1 e^x \mathrm{d}x \leqslant \sum_{i=1}^{n} e^{i/n} \cdot \frac{1}{n}$$

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 $\int_0^1 e^x \mathrm{d}x$ 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 $\int_0^1 e^x \mathrm{d}x$. For example, when $n = 1000$, $L_n = 1.7174$ and $R_n = 1.7191$ (both to four decimal places) so that, again to four decimal places,

$$1.7174 \leqslant \int_0^1 e^x \mathrm{d}x \leqslant 1.7191$$

Recall that the exact value is $e - 1 = 1.718281828\ldots$.

Example 1.1.12

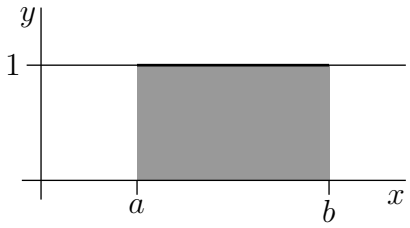
1.1.5 ▸▸ Using Known Areas to Evaluate Integrals

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 1.1.1. 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 1.1.13

The integral $\int_a^b 1 \mathrm{d}x$ (which is also written as just $\int_a^b \mathrm{d}x$) is the area of the shaded rectangle (of width $b - a$ and height 1) in the figure on the right below. So

$$\int_a^b \mathrm{d}x = (b - a) \times (1) = b - a$$

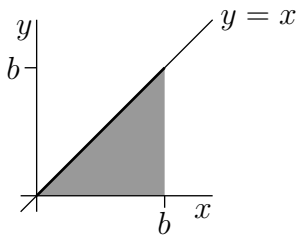


Example 1.1.13

Example 1.1.14

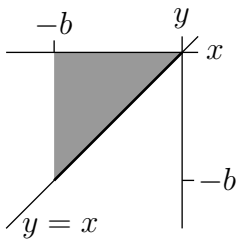
Let $b > 0$. The integral $\int_0^b x\mathrm{d}x$ is the area of the shaded triangle (of base b and of height b) in the figure on the right below. So

$$\int_0^b x\mathrm{d}x = \frac{1}{2}b \times b = \frac{b^2}{2}$$



The integral $\int_{-b}^0 x\mathrm{d}x$ is the signed area of the shaded triangle (again of base b and of height b) in the figure on the right below. So

$$\int_{-b}^0 x\mathrm{d}x = -\frac{b^2}{2}$$



Example 1.1.14

Notice that it is very easy to extend this example to the integral $\int_0^b cx\mathrm{d}x$ for any real numbers $b, c > 0$ and find

$$\int_0^b cx\mathrm{d}x = \frac{c}{2}b^2.$$

Example 1.1.15

In this example, we shall evaluate $\int_{-1}^1 (1 - |x|)\,\mathrm{d}x$. Recall that

$$|x| = \begin{cases} -x & \text{if } x \leqslant 0 \\ x & \text{if } x \geqslant 0 \end{cases}$$

so that

$$1 - |x| = \begin{cases} 1 + x & \text{if } x \leqslant 0 \\ 1 - x & \text{if } x \geqslant 0 \end{cases}$$

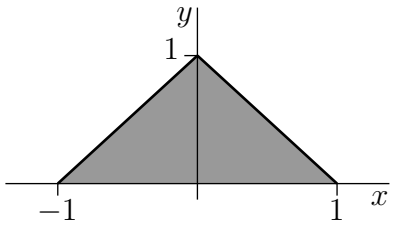
To picture the geometric figure whose area the integral represents observe that

- 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 $\int_{-1}^{1} (1 - |x|)\,\mathrm{d}x$ is the area of the shaded triangle (of base 2 and of height 1) in the figure on the right below and

$$\int_{-1}^{1} (1 - |x|)\,\mathrm{d}x = \frac{1}{2} \times 2 \times 1 = 1$$



Example 1.1.15

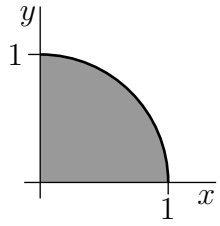
Example 1.1.16

The integral $\int_0^1 \sqrt{1 - x^2}\mathrm{d}x$ 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

$$y = \sqrt{1 - x^2} \qquad \text{as} \qquad x^2 + y^2 = 1, y \geqslant 0$$

But this is the (implicit) equation for a circle — the extra condition that $y \geqslant 0$ 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

$$\int_0^1 \sqrt{1 - x^2}\mathrm{d}x = \frac{1}{4}\pi(1)^2 = \frac{\pi}{4}$$



Example 1.1.16

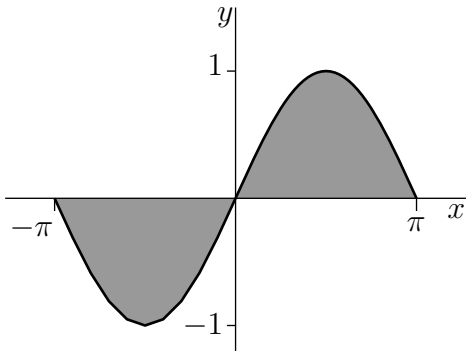
This next one is a little trickier and relies on us knowing the symmetries of the sine function.

Example 1.1.17

The integral $\int_{-\pi}^{\pi} \sin x \mathrm{d}x$ 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

$$\int_{-\pi}^{\pi} \sin x \mathrm{d}x = 0$$



Example 1.1.17

1.1.6 ▸▸ Another Interpretation for Definite Integrals

So far, we have only a single interpretation[17] for definite integrals — namely areas under graphs. In the following example, we develop a second interpretation.

Example 1.1.18

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 $n \to \infty$.

- 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 start of the time interval — $v(a)$. Over the subinterval the x-coordinate changes by velocity times time, namely $v(a) \cdot \frac{b-a}{n}$.

- 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 start of the subinterval

17 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.

— namely $v\left(a+\frac{b-a}{n}\right)$. So during the second subinterval the particle's x–coordinate changes by approximately $v\left(a+\frac{b-a}{n}\right)\frac{b-a}{n}$.

- 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

$$v\left(a+(i-1)\frac{b-a}{n}\right)\frac{b-a}{n}.$$

So the net change in x–coordinate from time a to time b is approximately

$$\begin{aligned} &v(a)\,\frac{b-a}{n}+v\Big(a+\frac{b-a}{n}\Big)\,\frac{b-a}{n}+\cdots+v\Big(a+(i-1)\frac{b-a}{n}\Big)\,\frac{b-a}{n}+\cdots \\ &\qquad\qquad +v\Big(a+(n-1)\frac{b-a}{n}\Big)\,\frac{b-a}{n} \\ &=\sum_{i=1}^{n} v\Big(a+(i-1)\frac{b-a}{n}\Big)\,\frac{b-a}{n} \end{aligned}$$

This is exactly the left Riemann sum approximation to the integral of v from a to b with n subintervals. The limit as $n\to\infty$ is exactly the definite integral $\int_a^b v(t)\mathrm{d}t$. 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 $b>a$,

$$x(b)-x(a)=\int_a^b v(t)\mathrm{d}t.$$

Example 1.1.18

1.1.7 ▸▸ Optional — Careful Definition of the Integral

In this optional section we give a more mathematically rigorous definition of the definite integral $\int_a^b f(x)\mathrm{d}x$. 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 $n\to\infty$.

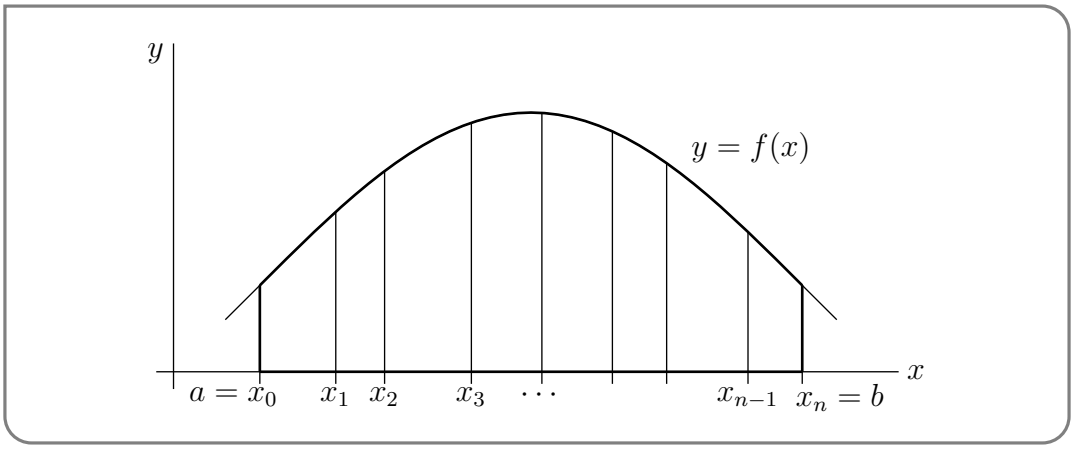
- Now subdivide the interval from a to b into n subintervals by selecting $n+1$ values of x that obey

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

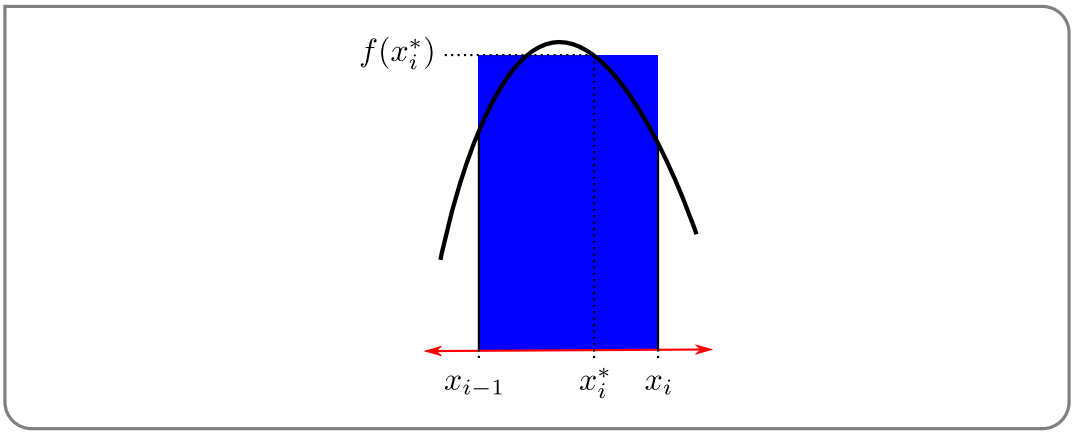
The subinterval number i runs from x_{i-1} to x_i. This formulation does not require the subintervals to have the same size. However we will eventually require that the widths of the subintervals shrink towards zero as $n \to \infty$.

- Then for each subinterval we select a value of x in that interval. That is, for $i = 1, 2, \ldots, n$, choose x_i^* satisfying $x_{i-1} \leqslant x_i^* \leqslant x_i$. 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, $\int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x$, 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

$$\int_a^b f(x)\mathrm{d}x \approx f(x_1^*)[x_1 - x_0] + f(x_2^*)[x_2 - x_1] + \cdots + f(x_n^*)[x_n - x_{n-1}]$$

- Of course every different choice of n and $x_1, x_2, \ldots, x_{n-1}$ and $x_1^*, x_2^*, \ldots, x_n^*$ gives a different approximation. So to simplify the discussion that follows, let us denote a particular choice of all these numbers by $\mathbb{P}$:

$$\mathbb{P} = (n, x_1, x_2, \cdots, x_{n-1}, x_1^*, x_2^*, \cdots, x_n^*)\,.$$

  Similarly let us denote the resulting approximation by $\mathcal{I}(\mathbb{P})$:

$$\mathcal{I}(\mathbb{P}) = f(x_1^*)[x_1 - x_0] + f(x_2^*)[x_2 - x_1] + \cdots + f(x_n^*)[x_n - x_{n-1}]$$

- We claim that, for any reasonable[18] function $f(x)$, if you take any reasonable[19] sequence of these approximations you always get the exactly the same limiting value. We define $\int_a^b f(x)\mathrm{d}x$ 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

$$M(\mathbb{P}) = \max\left\{x_1 - x_0\,,\ x_2 - x_1\,,\ \cdots\,,\ x_n - x_{n-1}\right\}$$

  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

$$\int_a^b f(x)\mathrm{d}x = \lim_{M(\mathbb{P})\to 0} \mathcal{I}(\mathbb{P}).$$

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

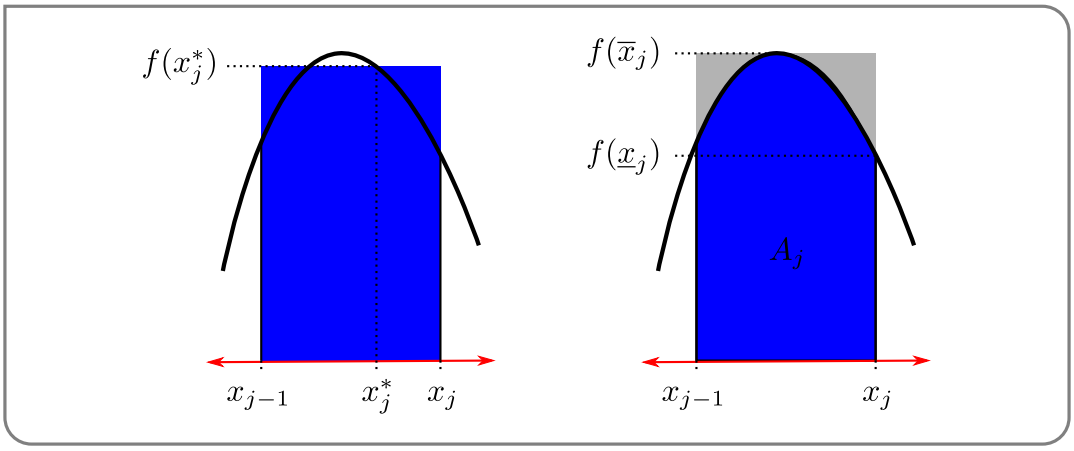
- that $f(x)$ is continuous for $a \leqslant x \leqslant b$,
- that $f(x)$ is differentiable for $a < x < b$, and
- that $f'(x)$ is bounded — ie $|f'(x)| \leqslant F$ for some constant F.

18 We'll be more precise about what "reasonable" means shortly.
19 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 $x_{j-1} \leqslant x \leqslant x_j$.



  We have approximated this area by $f(x_j^*)[x_j - x_{j-1}]$ (see figure left).

- Let $f(\overline{x}_j)$ and $f(\underline{x}_j)$ be the largest and smallest values[20] of $f(x)$ for $x_{j-1} \leqslant x \leqslant x_j$. Then the true area is bounded by

$$f(\underline{x}_j)[x_j - x_{j-1}] \leqslant A_j \leqslant f(\overline{x}_j)[x_j - x_{j-1}].$$

  (see figure right).

- Now since $f(\underline{x}_j) \leqslant f(x_j^*) \leqslant f(\overline{x}_j)$, we also know that

$$f(\underline{x}_j)[x_j - x_{j-1}] \leqslant f(x_j^*)[x_j - x_{j-1}] \leqslant f(\overline{x}_j)[x_j - x_{j-1}].$$

- 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 $f(\overline{x}_j)[x_j - x_{j-1}]$ 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

$$\left|A_j - f(x_j^*)[x_j - x_{j-1}]\right| \leqslant [f(\overline{x}_j) - f(\underline{x}_j)] \cdot [x_j - x_{j-1}].$$

  (To see this think about the smallest the true area can be and the largest our approximation can be and vice versa.)

20 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 $x_{j-1} \leqslant \overline{x}_j, \underline{x}_j \leqslant x_j$ such that $f(\underline{x}_j) \leqslant f(x) \leqslant f(\overline{x}_j)$ for all $x_{j-1} \leqslant x \leqslant x_j$.

- Now since our function, $f(x)$ is differentiable we can apply one of the main theorems we learned in CLP-1 — the Mean Value Theorem[21]. The MVT implies that there exists a c between $\underline{x}_j$ and $\overline{x}_j$ such that

$$f(\overline{x}_j) - f(\underline{x}_j) = f'(c) \cdot [\overline{x}_j - \underline{x}_j]$$

- By the assumption that $|f'(x)| \leqslant F$ for all x and the fact that $\underline{x}_j$ and $\overline{x}_j$ must both be between x_{j-1} and x_j

$$\left|f(\overline{x}_j) - f(\underline{x}_j)\right| \leqslant F \cdot \left|\overline{x}_j - \underline{x}_j\right| \leqslant F \cdot [x_j - x_{j-1}]$$

  Hence the error in this part of our approximation obeys

$$\left|A_j - f(x_j^*)[x_j - x_{j-1}]\right| \leqslant F \cdot [x_j - x_{j-1}]^2.$$

- 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

$$\begin{aligned}
|A - \mathcal{I}(\mathbb{P})| &= \left|\sum_{j=1}^{n} A_j - \sum_{j=1}^{n} f(x_j^*)[x_j - x_{j-1}]\right| \\
&= \left|\sum_{j=1}^{n} \Big(A_j - f(x_j^*)[x_j - x_{j-1}]\Big)\right| && \text{triangle inequality} \\
&\leqslant \sum_{j=1}^{n} \left|A_j - f(x_j^*)[x_j - x_{j-1}]\right| \\
&\leqslant \sum_{j=1}^{n} F \cdot [x_j - x_{j-1}]^2 && \text{from above}
\end{aligned}$$

  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^{th} subinterval) by the maximum width of the subintervals:

$$\begin{aligned}
&\leqslant \sum_{j=1}^{n} F \cdot M(\mathbb{P}) \cdot [x_j - x_{j-1}] && F \text{ and } M(\mathbb{P}) \text{ are constant} \\
&\leqslant F \cdot M(\mathbb{P}) \cdot \sum_{j=1}^{n} [x_j - x_{j-1}] && \text{sum is total width} \\
&= F \cdot M(\mathbb{P}) \cdot (b - a).
\end{aligned}$$

21 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 1.1.19.

Assume that $f(x)$ is continuous for $a \leqslant x \leqslant b$, and is differentiable for all $a < x < b$ with $|f'(x)| \leqslant F$, 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.

1.2▲ 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 of computing derivatives of simpler functions like x^2 and $\sin(x)$. Along the way we established simple rules such as

$$\lim_{x\to a}(f(x)+g(x)) = \lim_{x\to a} f(x) + \lim_{x\to a} g(x) \qquad \text{and} \qquad \frac{\mathrm{d}}{\mathrm{d}x}(f(x)+g(x)) = \frac{\mathrm{d}f}{\mathrm{d}x} + \frac{\mathrm{d}g}{\mathrm{d}x}$$

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 1.2.1 (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

(a) $$\int_a^b (f(x)+g(x))\,\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b g(x)\mathrm{d}x$$

(b) $$\int_a^b (f(x)-g(x))\,\mathrm{d}x = \int_a^b f(x)\mathrm{d}x - \int_a^b g(x)\mathrm{d}x$$

(c) $$\int_a^b Cf(x)\mathrm{d}x = C\cdot\int_a^b f(x)\mathrm{d}x$$

Combining these three rules we have

(d) $$\int_a^b (Af(x)+Bg(x))\,\mathrm{d}x = A\int_a^b f(x)\mathrm{d}x + B\int_a^b g(x)\mathrm{d}x$$

That is, integrals depend linearly on the integrand.

(e) $$\int_a^b \mathrm{d}x = \int_a^b 1\cdot\mathrm{d}x = b-a$$

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), $\int_a^b 1\mathrm{d}x$, 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 1.1.9):

$$\begin{aligned}
\int_a^b \mathrm{d}x &= \lim_{n\to\infty}\sum_{i=1}^{n} f(x_{i,n}^*)\,\frac{b-a}{n} && \text{by definition}\\
&= \lim_{n\to\infty}\sum_{i=1}^{n} 1\,\frac{b-a}{n} && \text{since } f(x)=1\\
&= \lim_{n\to\infty}(b-a)\sum_{i=1}^{n}\frac{1}{n} && \text{since } a,b \text{ are constants}\\
&= \lim_{n\to\infty}(b-a)\\
&= b-a
\end{aligned}$$

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 1.1.9 and some algebraic manipulations[22] to arrive at the result.

$$\begin{aligned}
\int_a^b h(x)\mathrm{d}x &= \sum_{i=1}^{n} h(x_{i,n}^*)\cdot\frac{b-a}{n} && \text{by Definition 1.1.9}\\
&= \sum_{i=1}^{n}\left(Af(x_{i,n}^*)+Bg(x_{i,n}^*)\right)\cdot\frac{b-a}{n}\\
&= \sum_{i=1}^{n}\left(Af(x_{i,n}^*)\cdot\frac{b-a}{n}+Bg(x_{i,n}^*)\cdot\frac{b-a}{n}\right)\\
&= \left(\sum_{i=1}^{n} Af(x_{i,n}^*)\cdot\frac{b-a}{n}\right)+\left(\sum_{i=1}^{n} Bg(x_{i,n}^*)\cdot\frac{b-a}{n}\right) && \text{by Theorem 1.1.5(b)}\\
&= A\left(\sum_{i=1}^{n} f(x_{i,n}^*)\cdot\frac{b-a}{n}\right)+B\left(\sum_{i=1}^{n} g(x_{i,n}^*)\cdot\frac{b-a}{n}\right) && \text{by Theorem 1.1.5(a)}\\
&= A\int_a^b f(x)\mathrm{d}x + B\int_a^b g(x)\mathrm{d}x && \text{by Definition 1.1.9}
\end{aligned}$$

as required.

□

Using this Theorem we can integrate sums, differences and constant multiples of functions we know how to integrate. For example:

Example 1.2.2

In Example 1.1.1 we saw that $\int_0^1 e^x\mathrm{d}x = e-1$. So

$$\begin{aligned}
\int_0^1 (e^x+7)\mathrm{d}x &= \int_0^1 e^x\mathrm{d}x + 7\int_0^1 1\mathrm{d}x\\
&\qquad \text{by Theorem 1.2.1(d) with } A=1, f(x)=e^x, B=7, g(x)=1\\
&= (e-1)+7\times(1-0)\\
&\qquad \text{by Example 1.1.1 and Theorem 1.2.1(e)}\\
&= e+6
\end{aligned}$$

Example 1.2.2

When we gave the formal definition of $\int_a^b f(x)\mathrm{d}x$ in Definition 1.1.9 we explained that the integral could be interpreted as the signed area between the curve $y = f(x)$ and the

22 Now is a good time to look back at Theorem 1.1.5.

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 $\int_a^b f(x)\mathrm{d}x$ in terms of $\int_b^a f(x)\mathrm{d}x$ — making it always possible to write an integral so the lower limit of integration is less than the upper limit of integration. Theorem 1.2.3, below, tell us that, for example, $\int_7^3 e^x\mathrm{d}x = -\int_3^7 e^x\mathrm{d}x$. The same theorem also provides us with two other simple manipulations of the limits of integration.

Theorem 1.2.3 (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

(a) $$\int_a^a f(x)\mathrm{d}x = 0$$

(b) $$\int_b^a f(x)\mathrm{d}x = -\int_a^b f(x)\mathrm{d}x$$

(c) $$\int_a^b f(x)\mathrm{d}x = \int_a^c f(x)\mathrm{d}x + \int_c^b f(x)\mathrm{d}x$$

The proof of this statement is not too difficult.

Proof. Let us prove the statements in order.

- Consider the definition of the definite integral

$$\int_a^b f(x)\mathrm{d}x = \lim_{n\to\infty}\sum_{i=1}^{n} f(x_{i,n}^*)\cdot\frac{b-a}{n}$$

If we now substitute $b = a$ in this expression we have

$$\begin{aligned}\int_a^a f(x)\mathrm{d}x &= \lim_{n\to\infty}\sum_{i=1}^{n} f(x_{i,n}^*)\cdot\underbrace{\frac{a-a}{n}}_{=0}\\ &= \lim_{n\to\infty}\sum_{i=1}^{n}\underbrace{f(x_{i,n}^*)\cdot 0}_{=0}\\ &= \lim_{n\to\infty} 0\\ &= 0\end{aligned}$$

as required.

- Consider now the definite integral $\int_a^b f(x)\mathrm{d}x$. We will sneak up on the proof by first examining Riemann sum approximations to both this and $\int_b^a f(x)\mathrm{d}x$. The midpoint

Riemann sum approximation to $\int_a^b f(x)\mathrm{d}x$ with 4 subintervals (so that each subinterval has width $\frac{b-a}{4}$) is

$$\left\{f\Big(a+\frac{1}{2}\frac{b-a}{4}\Big)+f\Big(a+\frac{3}{2}\frac{b-a}{4}\Big)+f\Big(a+\frac{5}{2}\frac{b-a}{4}\Big)+f\Big(a+\frac{7}{2}\frac{b-a}{4}\Big)\right\}\cdot\frac{b-a}{4}$$
$$=\left\{f\Big(\frac{7}{8}a+\frac{1}{8}b\Big)+f\Big(\frac{5}{8}a+\frac{3}{8}b\Big)+f\Big(\frac{3}{8}a+\frac{5}{8}b\Big)+f\Big(\frac{1}{8}a+\frac{7}{8}b\Big)\right\}\cdot\frac{b-a}{4}$$

Now we do the same for $\int_b^a f(x)\mathrm{d}x$ with 4 subintervals. Note that b is now the lower limit on the integral and a is now the upper limit on the integral. This is likely to cause confusion when we write out the Riemann sum, so we'll temporarily rename b to A and a to B. The midpoint Riemann sum approximation to $\int_A^B f(x)\mathrm{d}x$ with 4 subintervals is

$$\left\{f\Big(A+\frac{1}{2}\frac{B-A}{4}\Big)+f\Big(A+\frac{3}{2}\frac{B-A}{4}\Big)+f\Big(A+\frac{5}{2}\frac{B-A}{4}\Big)+f\Big(A+\frac{7}{2}\frac{B-A}{4}\Big)\right\}\cdot\frac{B-A}{4}$$
$$=\left\{f\Big(\frac{7}{8}A+\frac{1}{8}B\Big)+f\Big(\frac{5}{8}A+\frac{3}{8}B\Big)+f\Big(\frac{3}{8}A+\frac{5}{8}B\Big)+f\Big(\frac{1}{8}A+\frac{7}{8}B\Big)\right\}\cdot\frac{B-A}{4}$$

Now recalling that $A=b$ and $B=a$, we have that the midpoint Riemann sum approximation to $\int_b^a f(x)\mathrm{d}x$ with 4 subintervals is

$$\left\{f\Big(\frac{7}{8}b+\frac{1}{8}a\Big)+f\Big(\frac{5}{8}b+\frac{3}{8}a\Big)+f\Big(\frac{3}{8}b+\frac{5}{8}a\Big)+f\Big(\frac{1}{8}b+\frac{7}{8}a\Big)\right\}\cdot\frac{a-b}{4}$$

Thus we see that the Riemann sums for the two integrals are nearly identical — the only difference being the factor of $\frac{b-a}{4}$ versus $\frac{a-b}{4}$. Hence the two Riemann sums are negatives of each other.

The same computation with n subintervals shows that the midpoint Riemann sum approximations to $\int_b^a f(x)\mathrm{d}x$ and $\int_a^b f(x)\mathrm{d}x$ with n subintervals are negatives of each other. Taking the limit $n\to\infty$ gives $\int_b^a f(x)\mathrm{d}x=-\int_a^b f(x)\mathrm{d}x$.

- 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:

$$\int_a^a f(x)\mathrm{d}x=0 \qquad \text{and} \qquad \int_a^b f(x)\mathrm{d}x=\int_a^c f(x)\mathrm{d}x+\int_c^b f(x)\mathrm{d}x$$

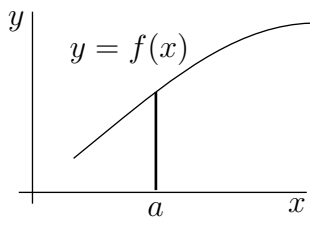
are

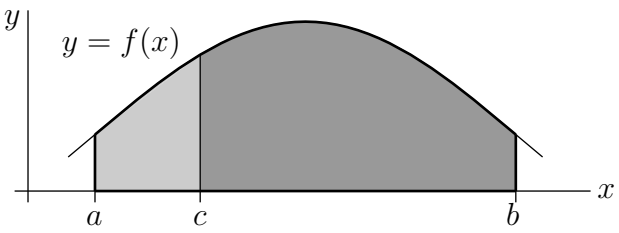
$$\text{Area}\big\{\ (x,y)\ \big|\ a\leqslant x\leqslant a,\ 0\leqslant y\leqslant f(x)\ \big\}=0$$

and

$$\text{Area}\big\{\ (x,y)\ \big|\ a\leqslant x\leqslant b,\ 0\leqslant y\leqslant f(x)\ \big\}=\text{Area}\big\{\ (x,y)\ \big|\ a\leqslant x\leqslant c,\ 0\leqslant y\leqslant f(x)\ \big\}$$
$$+\text{Area}\big\{\ (x,y)\ \big|\ c\leqslant x\leqslant b,\ 0\leqslant y\leqslant f(x)\ \big\}$$

respectively. Both of these geometric statements are intuitively obvious. See the figures below.





Note that we have assumed that $a \leqslant c \leqslant b$ and that $f(x) \geqslant 0$. One can remove these restrictions and also make the proof more formal, but it becomes quite tedious and less intuitive.

□

Remark 1.2.4. For notational simplicity, let's assume that $a \leqslant c \leqslant b$ and $f(x) \geqslant 0$ for all $a \leqslant x \leqslant b$. The geometric interpretations of the identities

$$\int_a^a f(x)\mathrm{d}x = 0 \qquad \text{and} \qquad \int_a^b f(x)\mathrm{d}x = \int_a^c f(x)\mathrm{d}x + \int_c^b f(x)\mathrm{d}x$$

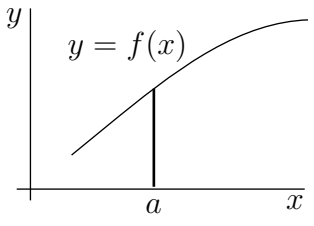
are

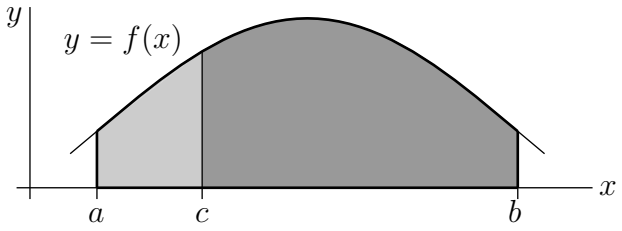
$$\text{Area}\{\ (x,y)\ |\ a \leqslant x \leqslant a,\ 0 \leqslant y \leqslant f(x)\ \} = 0$$

and

$$\begin{aligned}\text{Area}\{\ (x,y)\ |\ a \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\ \} = \text{Area}\{\ (x,y)\ |\ a \leqslant x \leqslant c,\ 0 \leqslant y \leqslant f(x)\ \}\\ + \text{Area}\{\ (x,y)\ |\ c \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\ \}\end{aligned}$$

respectively. Both of these geometric statements are intuitively obvious. See the figures below.





Example 1.2.5

Back in Example 1.1.14 we saw that when $b > 0$ $\int_0^b x\mathrm{d}x = \frac{b^2}{2}$. We'll now verify that $\int_0^b x\mathrm{d}x = \frac{b^2}{2}$ is still true when $b = 0$ and also when $b < 0$.

- First consider $b = 0$. Then the statement $\int_0^b x\mathrm{d}x = \frac{b^2}{2}$ becomes

$$\int_0^0 x\mathrm{d}x = 0$$

This is an immediate consequence of Theorem 1.2.3(a).

- Now consider $b < 0$. Let us write $B = -b$, so that $B > 0$. In Example 1.1.14 we saw that

$$\int_{-B}^0 x\mathrm{d}x = -\frac{B^2}{2}.$$

So we have

$$\begin{aligned}
\int_0^b x\mathrm{d}x = \int_0^{-B} x\mathrm{d}x = -\int_{-B}^0 x\mathrm{d}x & \qquad \text{by Theorem 1.2.3(b)} \\
= -\left(-\frac{B^2}{2}\right) & \qquad \text{by Example 1.1.14} \\
= \frac{B^2}{2} = \frac{b^2}{2} &
\end{aligned}$$

We have now shown that

$$\int_0^b x\mathrm{d}x = \frac{b^2}{2} \qquad \text{for all real numbers } b$$

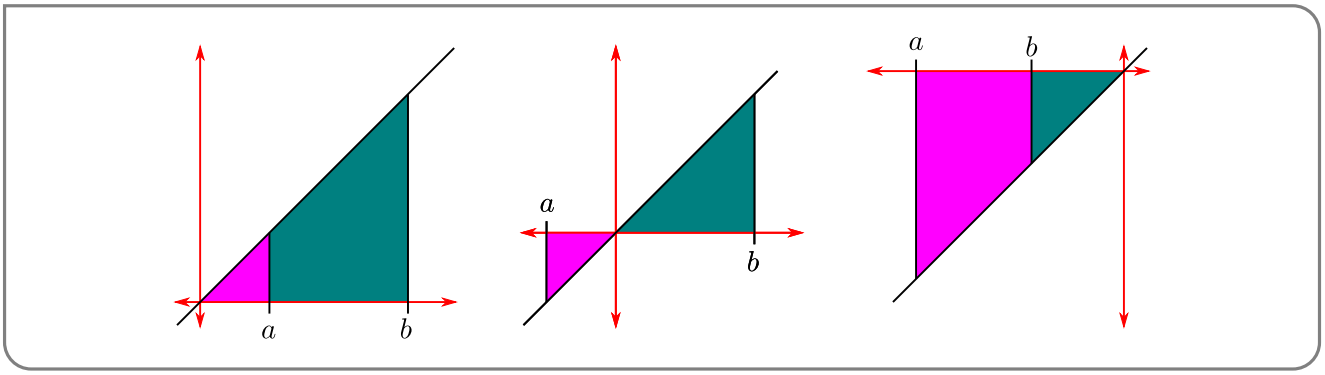
Example 1.2.5

Example 1.2.6

Applying Theorem 1.2.3 yet again, we have, for all real numbers a and b,

$$\begin{aligned}
\int_a^b x\mathrm{d}x = \int_a^0 x\mathrm{d}x + \int_0^b x\mathrm{d}x & \qquad \text{by Theorem 1.2.3(c) with } c = 0 \\
= \int_0^b x\mathrm{d}x - \int_0^a x\mathrm{d}x & \qquad \text{by Theorem 1.2.3(b)} \\
= \frac{b^2 - a^2}{2} & \qquad \text{by Example 1.2.5, twice}
\end{aligned}$$

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 1.2.6

Theorem 1.2.3(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 1.2.7

Using Theorem 1.2.3, we can readily evaluate integrals involving $|x|$. First, recall that

$$|x| = \begin{cases} x & \text{if } x \geqslant 0 \\ -x & \text{if } x < 0 \end{cases}$$

Now consider (for example) $\int_{-2}^{3} |x| \mathrm{d}x$. Since the integrand changes at $x = 0$, it makes sense to split the interval of integration at that point:

$$\begin{aligned}
\int_{-2}^{3} |x| \mathrm{d}x &= \int_{-2}^{0} |x| \mathrm{d}x + \int_{0}^{3} |x| \mathrm{d}x && \text{by Theorem 1.2.3} \\
&= \int_{-2}^{0} (-x) \mathrm{d}x + \int_{0}^{3} x \mathrm{d}x && \text{by definition of } |x| \\
&= -\int_{-2}^{0} x \mathrm{d}x + \int_{0}^{3} x \mathrm{d}x && \text{by Theorem 1.2.1(c)} \\
&= -(-2^2/2) + (3^2/2) = (4+9)/2 \\
&= 13/2
\end{aligned}$$

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 $f(-x)$.

$$\begin{aligned}\int_{-1}^{1} f\big(|x|\big)\mathrm{d}x &= \int_{-1}^{0} f\big(|x|\big)\mathrm{d}x + \int_{0}^{1} f\big(|x|\big)\mathrm{d}x \\ &= \int_{-1}^{0} f(-x)\mathrm{d}x + \int_{0}^{1} f(x)\mathrm{d}x\end{aligned}$$

Example 1.2.7

Here is a more concrete example.

Example 1.2.8

Let us compute $\int_{-1}^{1}\big(1-|x|\big)\mathrm{d}x$ again. In Example 1.1.15 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 1.2.1 and 1.2.3 and the facts that

$$\int_a^b \mathrm{d}x = b-a \qquad \text{and} \qquad \int_a^b x\mathrm{d}x = \frac{b^2-a^2}{2}$$

That $\int_a^b \mathrm{d}x = b-a$ is part (e) of Theorem 1.2.1. We saw that $\int_a^b x\mathrm{d}x = \frac{b^2-a^2}{2}$ in Example 1.2.6.

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 $x \geqslant 0$ and $|x| = -x$ whenever $x \leqslant 0$, we split the interval by Theorem 1.2.3(c),

$$\begin{aligned}\int_{-1}^{1}\big(1-|x|\big)\mathrm{d}x &= \int_{-1}^{0}\big(1-|x|\big)\mathrm{d}x + \int_{0}^{1}\big(1-|x|\big)\mathrm{d}x \\ &= \int_{-1}^{0}\big(1-(-x)\big)\mathrm{d}x + \int_{0}^{1}\big(1-x\big)\mathrm{d}x \\ &= \int_{-1}^{0}\big(1+x\big)\mathrm{d}x + \int_{0}^{1}\big(1-x\big)\mathrm{d}x\end{aligned}$$

Now we apply parts (a) and (b) of Theorem 1.2.1, and then

$$\begin{aligned}\int_{-1}^{1}\big[1-|x|\big]\mathrm{d}x &= \int_{-1}^{0} 1\mathrm{d}x + \int_{-1}^{0} x\mathrm{d}x + \int_{0}^{1} 1\mathrm{d}x - \int_{0}^{1} x\mathrm{d}x \\ &= [0-(-1)] + \frac{0^2-(-1)^2}{2} + [1-0] - \frac{1^2-0^2}{2} \\ &= 1\end{aligned}$$

Example 1.2.8

1.2.1 ▸▸ More Properties of Integration: Even and Odd Functions

Recall[23] the following definition

Definition 1.2.9.

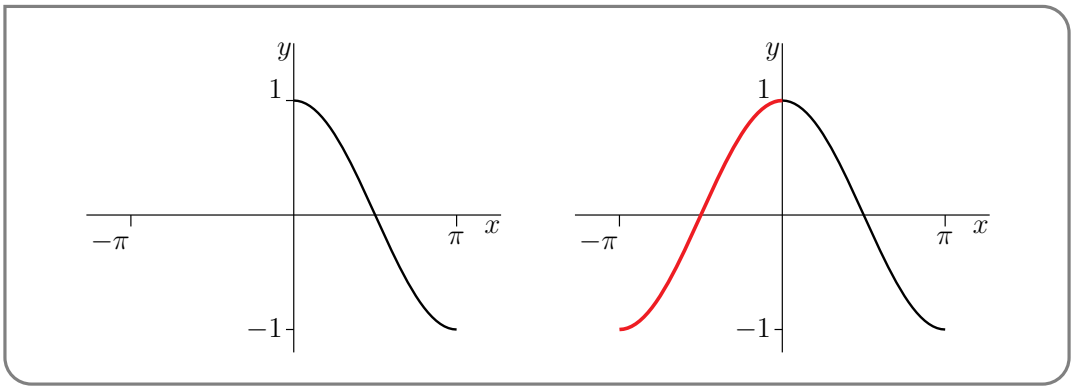
Let $f(x)$ be a function. Then,

- 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 1.2.10 (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.
- The part of the graph $y = f(x)$ with $x \leqslant 0$, may be constructed by drawing the part of the graph with $x \geqslant 0$ (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

$$\{\,(x,y)\;|\;0 \leqslant x \leqslant a \text{ and } y \text{ is between } 0 \text{ and } f(x)\,\}$$
$$\{\,(x,y)\;|\;-a \leqslant x \leqslant 0 \text{ and } y \text{ is between } 0 \text{ and } f(x)\,\}$$

are reflections of each other in the y–axis and so have the same signed area. That is

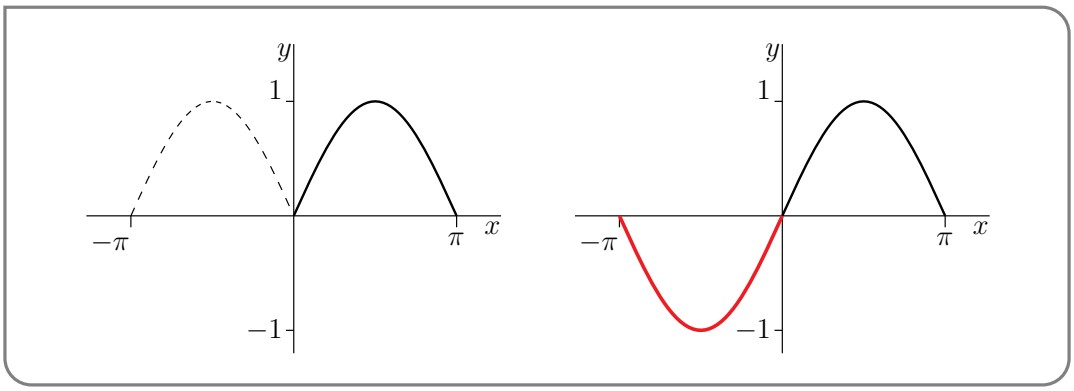
$$\int_0^a f(x)\mathrm{d}x = \int_{-a}^0 f(x)\mathrm{d}x$$

23 We haven't done this in this course, but you should have seen it in your differential calculus course or perhaps even earlier.

Example 1.2.10

Example 1.2.11 (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 $x \leqslant 0$, may be constructed by drawing the part of the graph with $x \geqslant 0$ (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

$$\{\ (x,y)\ |\ 0 \leqslant x \leqslant a \text{ and } y \text{ is between 0 and } f(x)\ \}$$
$$\{\ (x,y)\ |\ -a \leqslant x \leqslant 0 \text{ and } y \text{ is between 0 and } f(x)\ \}$$

  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 y–axis. That is

$$\int_0^a f(x)\mathrm{d}x = -\int_{-a}^0 f(x)\mathrm{d}x$$

Example 1.2.11

We can exploit the symmetries noted in the examples above, namely

$$\int_0^a f(x)\mathrm{d}x = \int_{-a}^0 f(x)\mathrm{d}x \qquad \text{for } f \text{ even}$$
$$\int_0^a f(x)\mathrm{d}x = -\int_{-a}^0 f(x)\mathrm{d}x \qquad \text{for } f \text{ odd}$$

together with Theorem 1.2.3

$$\int_{-a}^{a} f(x)\mathrm{d}x = \int_{-a}^{0} f(x)\mathrm{d}x + \int_{0}^{a} f(x)\mathrm{d}x$$

in order to simplify the integration of even and odd functions over intervals of the form $[-a, a]$.

Theorem 1.2.12 (Even and Odd)**.**

Let $a > 0$.

(a) If $f(x)$ is an even function, then

$$\int_{-a}^{a} f(x)\mathrm{d}x = 2\int_{0}^{a} f(x)\mathrm{d}x$$

(b) If $f(x)$ is an odd function, then

$$\int_{-a}^{a} f(x)\mathrm{d}x = 0$$

Proof. For any function

$$\int_{-a}^{a} f(x)\mathrm{d}x = \int_{0}^{a} f(x)\mathrm{d}x + \int_{-a}^{0} f(x)\mathrm{d}x$$

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. □

1.2.2 ▸▸ Optional — 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 1.2.13 (Inequalities for Integrals).

Let $a \leqslant b$ be real numbers and let the functions $f(x)$ and $g(x)$ be integrable on the interval $a \leqslant x \leqslant b$.

(a) If $f(x) \geqslant 0$ for all $a \leqslant x \leqslant b$, then

$$\int_a^b f(x)\,\mathrm{d}x \geqslant 0$$

(b) If $f(x) \leqslant g(x)$ for all $a \leqslant x \leqslant b$, then

$$\int_a^b f(x)\,\mathrm{d}x \leqslant \int_a^b g(x)\,\mathrm{d}x$$

(c) If there are constants m and M such that $m \leqslant f(x) \leqslant M$ for all $a \leqslant x \leqslant b$, then

$$m(b-a) \leqslant \int_a^b f(x)\,\mathrm{d}x \leqslant M(b-a)$$

(d) We have

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leqslant \int_a^b |f(x)|\,\mathrm{d}x$$

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 $a \leqslant b$, then the signed area of the set $\{\,(x,y) \mid a \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\,\}$ is at least zero. This is quite clear. Alternatively, we could argue more algebraically from Definition 1.1.9. We observe that when we define $\int_a^b f(x)\mathrm{d}x$ via Riemann sums, every summand, $f(x_{i,n}^*)\,\frac{b-a}{n} \geqslant 0$. Thus the whole sum is nonnegative and consequently, so is the limit, and thus so is the integral.

(b) We are assuming that $g(x) - f(x) \geqslant 0$, so part (a) gives

$$\begin{aligned}\int_a^b \big[g(x) - f(x)\big]\mathrm{d}x \geqslant 0 \implies \int_a^b g(x)\,\mathrm{d}x - \int_a^b f(x)\,\mathrm{d}x \geqslant 0\\ \implies \int_a^b f(x)\,\mathrm{d}x \leqslant \int_a^b g(x)\,\mathrm{d}x\end{aligned}$$

(c) Applying part (b) with $g(x) = M$ for all $a \leqslant x \leqslant b$ gives

$$\int_a^b f(x)\,\mathrm{d}x \leqslant \int_a^b M\,\mathrm{d}x = M(b-a)$$

Similarly, viewing m as a (constant) function, and applying part (b) gives

$$m \leqslant f(x) \implies \overbrace{\int_a^b m\,\mathrm{d}x}^{=m(b-a)} \leqslant \int_a^b f(x)\,\mathrm{d}x$$

(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

$$f(x) \leqslant |f(x)| \qquad \text{and} \qquad -f(x) \leqslant |f(x)|$$

Applying part (c) to each of those inequalities gives

$$\int_a^b f(x)\,\mathrm{d}x \leqslant \int_a^b |f(x)|\,\mathrm{d}x \qquad \text{and} \qquad -\int_a^b f(x)\,\mathrm{d}x \leqslant \int_a^b |f(x)|\,\mathrm{d}x$$

Now $\left|\int_a^b f(x)\,\mathrm{d}x\right|$ is either equal to $\int_a^b f(x)\,\mathrm{d}x$ or $-\int_a^b f(x)\,\mathrm{d}x$ (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

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leqslant \int_a^b |f(x)|\,\mathrm{d}x.$$

□

Example 1.2.14 $\left(\int_0^{\pi/3} \sqrt{\cos x}\mathrm{d}x\right)$

Consider the integral

$$\int_0^{\pi/3} \sqrt{\cos x}\mathrm{d}x$$

This is not so easy to compute exactly[24], but we can bound it quite quickly.

For x between 0 and $\frac{\pi}{3}$, the function $\cos x$ takes values[25] between 1 and $\frac{1}{2}$. Thus the function $\sqrt{\cos x}$ takes values between 1 and $\frac{1}{\sqrt{2}}$. That is

$$\frac{1}{\sqrt{2}} \leqslant \sqrt{\cos x} \leqslant 1 \qquad \text{for } 0 \leqslant x \leqslant \frac{\pi}{3}.$$

24 It is not too hard to use Riemann sums and a computer to evaluate it numerically: 0.948025319 . . . .

25 You know the graphs of sine and cosine, so you should be able to work this out without too much difficulty.

Consequently, by Theorem 1.2.13(c) with $a = 0$, $b = \frac{\pi}{3}$, $m = \frac{1}{\sqrt{2}}$ and $M = 1$,

$$\frac{\pi}{3\sqrt{2}} \leqslant \int_0^{\pi/3} \sqrt{\cos x}\,dx \leqslant \frac{\pi}{3}$$

Plugging these expressions into a calculator gives us

$$0.7404804898 \leqslant \int_0^{\pi/3} \sqrt{\cos x}\,dx \leqslant 1.047197551$$

Example 1.2.14

1.3▲ The Fundamental Theorem of Calculus

We have spent quite a few pages (and lectures) talking about definite integrals, what they are (Definition 1.1.9), when they exist (Theorem 1.1.10), how to compute some special cases (Section 1.1.5), some ways to manipulate them (Theorem 1.2.1 and 1.2.3) and how to bound them (Theorem 1.2.13). 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[26]. Of course "very rough" is not enough — let's be precise.

Theorem 1.3.1 (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 $x \in [a, b]$. 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

$$\int_a^b f(x)\,dx = G(b) - G(a) \quad \text{or equivalently} \quad \int_a^b G'(x)\,dx = G(b) - G(a)$$

26 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.

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

$$\int_a^b f(x)\mathrm{d}x = G(b) - G(a)$$

The form $\int_a^b G'(x)\,\mathrm{d}x = G(b) - G(a)$ of the fundamental theorem relates the rate of change of $G(x)$ over the interval $a \leqslant x \leqslant b$ 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 1.3.2 (A first example)

Consider the integral $\int_a^b x\mathrm{d}x$ which we have explored previously in Example 1.2.6.

- 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 1.3.1 then tells us that

$$\int_a^b f(x)\mathrm{d}x = G(b) - G(a)$$
$$\int_a^b x\mathrm{d}x = \frac{b^2}{2} - \frac{a^2}{2}$$

which is precisely the result we obtained (with more work) in Example 1.2.6.

Example 1.3.2

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

$$F(x) = \int_a^x f(t)\mathrm{d}t \qquad \text{then} \qquad F'(x) = f(x)$$

- Assume that F is the above integral and then consider $F'(x)$. By definition

$$F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h}$$

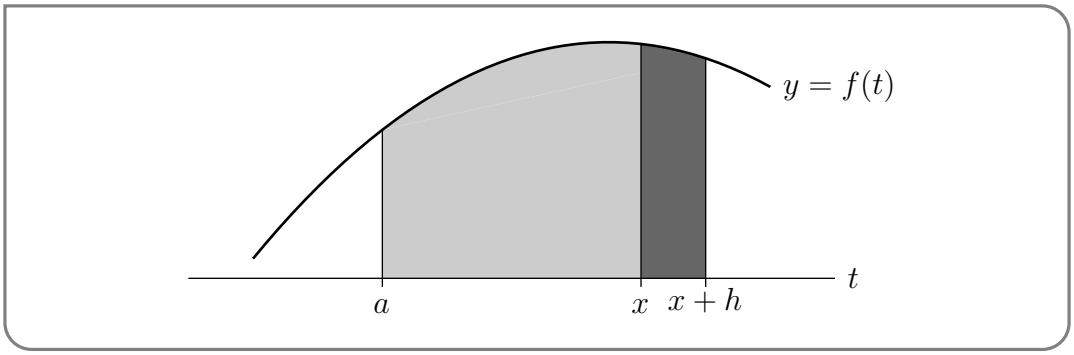
- To understand this limit, we interpret the terms $F(x), F(x+h)$ as signed areas. To simplify this further, let's only consider the case that f is always nonnegative and that $h > 0$. These restrictions are not hard to remove, but the proof ideas are a bit cleaner if we keep them in place. Then we have

$$\begin{aligned} F(x+h) &= \text{the area of the region } \left\{ (t,y) \,\middle|\, a \leqslant t \leqslant x+h,\ 0 \leqslant y \leqslant f(t) \right\} \\ F(x) &= \text{the area of the region } \left\{ (t,y) \,\middle|\, a \leqslant t \leqslant x, \qquad 0 \leqslant y \leqslant f(t) \right\} \end{aligned}$$

- Then the numerator

$$F(x+h) - F(x) = \text{the area of the region } \left\{ (t,y) \,\middle|\, x \leqslant t \leqslant x+h,\ 0 \leqslant y \leqslant f(t) \right\}$$

  This is just the more darkly shaded region in the figure



- We will be taking the limit $h \to 0$. 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[27], 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 $f(x)$.
- In the limit $h \to 0$, $\frac{F(x+h)-F(x)}{h}$ becomes exactly $f(x)$, which is precisely what we want.

□

Part 2. We want to show that $\int_a^b f(t)\mathrm{d}t = G(b) - G(a)$. To do this we exploit the fact that the derivative of a constant is zero.

- Let

$$H(x) = \int_a^x f(t)\mathrm{d}t - G(x) + G(a)$$

  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.

27 Notice that if f were discontinuous, then this might be false.

- We first show that $H(x)$ is constant by computing its derivative:

$$H'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\mathrm{d}t - \frac{\mathrm{d}}{\mathrm{d}x}\left(G(x)\right) + \frac{\mathrm{d}}{\mathrm{d}x}\left(G(a)\right)$$

  Since $G(a)$ is a constant, its derivative is 0 and by assumption the derivative of $G(x)$ is just $f(x)$, so

$$= \frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\mathrm{d}t - f(x)$$

  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 $H(a)$:

$$\begin{aligned} H(a) &= \int_a^a f(t)\mathrm{d}t - G(a) + G(a) \\ &= \int_a^a f(t)\mathrm{d}t && \text{by Theorem 1.2.3(a)} \\ &= 0 \end{aligned}$$

  as required.

□

The simple example we did above (Example 1.3.2), 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 1.3.3 $\left(\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x t\mathrm{d}t\right)$

Consider the integral $\int_0^x t\,\mathrm{d}t$. We know how to evaluate this — it is just Example 1.3.2 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 1.3.2 gives

$$F(x) = \int_0^x t\,\mathrm{d}t = \frac{x^2}{2}$$

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 1.3.3

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 1.3.4 $\left(\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\mathrm{d}t\right)$

We would like to find $\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\mathrm{d}t$. 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[28] the integral $\int_0^x e^{-t^2}\mathrm{d}t$ 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 fundamental theorem of calculus with $f(t) = e^{-t^2}$ and $a = 0$. That gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}x}\int_0^x f(t)\mathrm{d}t \\ &= f(x) = e^{-x^2}\end{aligned}$$

Example 1.3.4

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 1.3.5 $\left(\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\mathrm{d}t\right)$

Consider the integral $\int_0^{x^2} e^{-t^2}\mathrm{d}t$. 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 $\int_a^x f(t)\mathrm{d}t$ 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

$$E(x) = \int_0^x e^{-t^2}\mathrm{d}t$$

28 The integral $\int_0^x e^{-t^2}\mathrm{d}t$ 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

$$\int_0^x e^{-t^2}\mathrm{d}t = x - \frac{x^3}{3\cdot 1} + \frac{x^5}{5\cdot 2} - \frac{x^7}{7\cdot 3!} + \frac{x^9}{9\cdot 4!} + \cdots + (-1)^k\frac{x^{2k+1}}{(2k+1)\cdot k!} + \cdots$$

But more on this in Chapter 3.

Then the integral we want to work with is

$$E(x^2) = \int_0^{x^2} e^{-t^2} \mathrm{d}t$$

The derivative $E'(x)$ can be found via part 1 of the fundamental theorem of calculus (as we did in Example 1.3.4) and is $E'(x) = e^{-x^2}$. We can then use this fact with the chain rule to compute the derivative we need:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} \int_0^{x^2} e^{-t^2} \mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}x} E(x^2) && \text{use the chain rule} \\ &= 2xE'(x^2) \\ &= 2xe^{-x^4} \end{aligned}$$

Example 1.3.5

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 1.2.3.

Example 1.3.6 $\left(\frac{\mathrm{d}}{\mathrm{d}x} \int_x^{x^2} e^{-t^2} \mathrm{d}t\right)$

Consider the integral

$$\int_x^{x^2} e^{-t^2} \mathrm{d}t$$

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 $\int_a^x f(t)\mathrm{d}t$ in the fundamental theorem where the *lower* limit of integration is a constant.

Fortunately we can use the basic properties of integrals (Theorem 1.2.3(b) and (c)) to split $\int_x^{x^2} e^{-t^2} \mathrm{d}t$ into pieces whose derivatives we already know.

$$\begin{aligned} \int_x^{x^2} e^{-t^2} \mathrm{d}t &= \int_x^0 e^{-t^2} \mathrm{d}t + \int_0^{x^2} e^{-t^2} \mathrm{d}t && \text{by Theorem 1.2.3(c)} \\ &= -\int_0^x e^{-t^2} \mathrm{d}t + \int_0^{x^2} e^{-t^2} \mathrm{d}t && \text{by Theorem 1.2.3(b)} \end{aligned}$$

With this pre-processing, both integrals are of the right form. Using what we have learned

in the previous two examples,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\int_x^{x^2} e^{-t^2}\mathrm{d}t &= \frac{\mathrm{d}}{\mathrm{d}x}\left(-\int_0^x e^{-t^2}\mathrm{d}t + \int_0^{x^2} e^{-t^2}\mathrm{d}t\right)\\ &= -\frac{\mathrm{d}}{\mathrm{d}x}\int_0^x e^{-t^2}\mathrm{d}t + \frac{\mathrm{d}}{\mathrm{d}x}\int_0^{x^2} e^{-t^2}\mathrm{d}t\\ &= -e^{-x^2} + 2xe^{-x^4}\end{aligned}$$

Example 1.3.6

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 1.3.7 (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 1.3.8.

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

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(F(x) + C\right) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(F(x)\right) + \frac{\mathrm{d}}{\mathrm{d}x}\left(C\right)\\ &= f(x) + 0\end{aligned}$$

  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

$$\frac{\mathrm{d}}{\mathrm{d}x}H(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left(G(x) - F(x)\right) = \frac{\mathrm{d}}{\mathrm{d}x}G(x) - \frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x) - f(x) = 0$$

Since the derivative of $H(x)$ is zero, $H(x)$ must be a constant function[29]. 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.

Definition 1.3.9.

The "indefinite integral of $f(x)$" is denoted by $\int f(x)\mathrm{d}x$ and should be regarded as the general antiderivative of $f(x)$. In particular, if $F(x)$ is an antiderivative of $f(x)$ then

$$\int f(x)\mathrm{d}x = F(x) + C$$

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 1.3.10.

The symbol

$$\int f(x)\mathrm{d}x\Big|_a^b$$

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

$$\int f(x)\mathrm{d}x\Big|_a^b = F(x)|_a^b = F(b) - F(a)$$

Notice that this notation allows us to write part 2 of the fundamental theorem as

$$\begin{aligned}\int_a^b f(x)\mathrm{d}x &= \int f(x)\mathrm{d}x\Big|_a^b \\ &= F(x)|_a^b = F(b) - F(a)\end{aligned}$$

29 This follows from the Mean Value Theorem. Indeed, fix any number x_0. Then, for each $x \neq x_0$, the MVT gives us a number c between x_0 and x with

$$H(x) - H(x_0) = H'(c)\,(x - x_0) = 0$$

since the derivative of H is zero everywhere. Thus $H(x) = H(x_0)$ for all x and $H(x)$ is a constant function.

Some texts also use an equivalent notation using square brackets:

$$\int_a^b f(x)\mathrm{d}x = \Big[F(x)\Big]_a^b = F(b) - F(a).$$

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:

$$\frac{\mathrm{d}}{\mathrm{d}x}\int f(x)\mathrm{d}x = f(x).$$

We do these examples in some detail to help us become comfortable finding indefinite integrals.

Example 1.3.11

Compute the definite integral $\int_1^2 x\mathrm{d}x$.

Solution. We have already seen, in Example 1.2.6, that $\int_1^2 x\mathrm{d}x = \frac{2^2-1^2}{2} = \frac{3}{2}$. 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[30] 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

  $$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1}$$

  So if we want to end up with just $x = x^1$, we need to take $n = 2$. However this gives us

  $$\frac{\mathrm{d}}{\mathrm{d}x}x^2 = 2x$$

- 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:

  $$\frac{1}{2}\cdot\frac{\mathrm{d}}{\mathrm{d}x}x^2 = \frac{1}{2}\cdot 2x \qquad \text{which becomes}$$
  $$\cdot\frac{\mathrm{d}}{\mathrm{d}x}\frac{x^2}{2} = x$$

  which is exactly what we need. It tells us that $x^2/2$ is an antiderivative of x.

30 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.

- Once one has an antiderivative, it is easy to compute the indefinite integral

$$\int x\mathrm{d}x = \frac{1}{2}x^2 + C$$

as well as the definite integral:

$$\begin{aligned}\int_1^2 x\mathrm{d}x &= \frac{1}{2}x^2\Big|_1^2 && \text{since } x^2/2 \text{ is the antiderivative of } x\\ &= \frac{1}{2}2^2 - \frac{1}{2}1^2 = \frac{3}{2}\end{aligned}$$

Example 1.3.11

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 1.3.12

Compute $\int_0^{\pi/2} \sin x\mathrm{d}x$.

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

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x$$

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:

$$\begin{aligned}(-1)\cdot\frac{\mathrm{d}}{\mathrm{d}x}\cos x &= (-1)\cdot(-\sin x) && \text{giving us}\\ \frac{\mathrm{d}}{\mathrm{d}x}\big(-\cos x\big) &= \sin x\end{aligned}$$

This tells us that $-\cos x$ is an antiderivative of $\sin x$.

- Now it is straightforward to compute the integral:

$$\begin{aligned}\int_0^{\pi/2} \sin x\mathrm{d}x &= -\cos x|_0^{\pi/2} && \text{since } -\cos x \text{ is the antiderivative of } \sin x\\ &= -\cos\frac{\pi}{2} + \cos 0\\ &= 0 + 1 = 1\end{aligned}$$

Example 1.3.12

Example 1.3.13

Find $\int_1^2 \frac{1}{x}\mathrm{d}x$.

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

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = nx^{n-1},$$

doesn't work in this case — since it would require us to pick $n = 0$ and this would give

$$\frac{\mathrm{d}}{\mathrm{d}x}x^0 = \frac{\mathrm{d}}{\mathrm{d}x}1 = 0.$$

- Fortunately, we also know[31] that

$$\frac{\mathrm{d}}{\mathrm{d}x}\log x = \frac{1}{x}$$

which is exactly the derivative we want.

- We're now ready to compute the prescribed integral.

$$\begin{aligned}\int_1^2 \frac{1}{x}\mathrm{d}x &= \log x|_1^2 && \text{since } \log x \text{ is an antiderivative of } 1/x\\ &= \log 2 - \log 1 && \text{since } \log 1 = 0\\ &= \log 2\end{aligned}$$

Example 1.3.13

Example 1.3.14

Find $\int_{-2}^{-1} \frac{1}{x}\mathrm{d}x$.

Solution.

31 Recall that in most mathematics courses (especially this one) we use $\log x$ without any indicated base to denote the natural logarithm — the logarithm base e. Many widely used computer languages, like Java, C, Python, MATLAB, $\cdots$, use $\log(x)$ to denote the logarithm base e too. But many texts also use $\ln x$ to denote the natural logarithm

$$\log x = \log_e x = \ln x.$$

The reader should be comfortable with all three notations for this function. They should also be aware that in different contexts — such as in chemistry or physics — it is common to use $\log x$ to denote the logarithm base 10, while in computer science often $\log x$ denotes the logarithm base 2. Context is key.

- As we saw in the last example,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log x = \frac{1}{x}$$

and if we naively use this here, then we will obtain

$$\int_{-2}^{-1} \frac{1}{x}\mathrm{d}x = \log(-1) - \log(-2)$$

which makes no sense since the logarithm is only defined for positive numbers[32].

- We can work around this problem using a slight variation of the logarithm — $\log|x|$.
  - When $x > 0$, we know that $|x| = x$ and so we have

$$\begin{aligned} \log|x| &= \log x && \text{differentiating gives us} \\ \frac{\mathrm{d}}{\mathrm{d}x}\log|x| &= \frac{\mathrm{d}}{\mathrm{d}x}\log x = \frac{1}{x}. \end{aligned}$$

  - When $x < 0$ we have that $|x| = -x$ and so

$$\begin{aligned} \log|x| &= \log(-x) && \text{differentiating with the chain rule gives} \\ \frac{\mathrm{d}}{\mathrm{d}x}\log|x| &= \frac{\mathrm{d}}{\mathrm{d}x}\log(-x) \\ &= \frac{1}{(-x)}\cdot(-1) = \frac{1}{x} \end{aligned}$$

  - Indeed, more generally we should write the indefinite integral of $1/x$ as

$$\int \frac{1}{x}\mathrm{d}x = \log|x| + C$$

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.

$$\begin{aligned} \int_{-2}^{-1} \frac{1}{x}\mathrm{d}x &= \log|x|\Big|_{-2}^{-1} && \text{since } \log|x| \text{ is an antiderivative of } 1/x \\ &= \log|-1| - \log|-2| = \log 1 - \log 2 \\ &= -\log 2 = \log 1/2. \end{aligned}$$

Example 1.3.14

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

32 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.

$x = 0$ 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 1.3.15

Find $\int_{-1}^{1} \frac{1}{x^2} \mathrm{d}x$.

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

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{x} = -\frac{1}{x^2}$$

  which means that $-x^{-1}$ is an antiderivative of x^{-2}.

- This suggests (if we proceed naively) that

$$\begin{aligned}\int_{-1}^{1} x^{-2}\mathrm{d}x &= -\frac{1}{x}\Big|_{-1}^{1} \qquad \text{since } -1/x \text{ is an antiderivative of } 1/x^2\\ &= -\frac{1}{1} - \Big(-\frac{1}{-1}\Big)\\ &= -2\end{aligned}$$

  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

$$\text{if } F'(x) = f(x) \text{ then } \int_a^b f(x)\mathrm{d}x = F(b) - F(a),$$

  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 $\int_{-1}^{1} \frac{1}{x^2}\mathrm{d}x$, 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 $\int_{-1}^{1} \frac{1}{x^2}\mathrm{d}x$ not negative, it is infinite.

Example 1.3.15

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	$\log_e \lvert x \rvert$	$\arcsin x$	$\arctan x$
$f(x) = F'(x)$	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 1.3.16 (Important indefinite integrals).

$f(x)$	$F(x) = \int f(x)\mathrm{d}x$
1	$x + C$
x^n	$\frac{1}{n+1}x^{n+1} + C$ provided that $n \neq -1$
$\frac{1}{x}$	$\log_e \lvert x \rvert + C$
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 1.3.17

Find the following integrals

(i) $\int_2^7 e^x \mathrm{d}x$

(ii) $\int_{-2}^{2} \frac{1}{1+x^2} \mathrm{d}x$

(iii) $\int_0^3 (2x^3 + 7x - 2)\mathrm{d}x$

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 1.2.1 and do each separately.

(i) An antiderivative of e^x is just e^x, so

$$\begin{aligned}\int_2^7 e^x \mathrm{d}x &= e^x \Big|_2^7 \\ &= e^7 - e^2 = e^2(e^5 - 1).\end{aligned}$$

(ii) An antiderivative of $\frac{1}{1+x^2}$ is $\arctan(x)$, so

$$\begin{aligned}\int_{-2}^{2} \frac{1}{1+x^2} \mathrm{d}x &= \arctan(x) \Big|_{-2}^{2} \\ &= \arctan(2) - \arctan(-2)\end{aligned}$$

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 1.2.1(d)

$$\begin{aligned}\int_0^3 (2x^3 + 7x - 2)\mathrm{d}x &= \int_0^3 2x^3 \mathrm{d}x + \int_0^3 7x \mathrm{d}x - \int_0^3 2\mathrm{d}x \\ &= 2\int_0^3 x^3 \mathrm{d}x + 7\int_0^3 x \mathrm{d}x - 2\int_0^3 \mathrm{d}x\end{aligned}$$

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:

$$\begin{aligned}\int_0^3 (2x^3 + 7x - 2)\mathrm{d}x &= \left[\frac{x^4}{2} + \frac{7x^2}{2} - 2x\right]_0^3 \\ &= \frac{81}{2} + \frac{7 \cdot 9}{2} - 6 = 66.\end{aligned}$$

Example 1.3.17

1.4▲ 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 1.2.1 — that integrals interact very nicely with addition, subtraction and multiplication by constants:

$$\int_a^b (Af(x) + Bg(x))\,\mathrm{d}x = A\int_a^b f(x)\mathrm{d}x + B\int_a^b g(x)\mathrm{d}x$$

for A, B constants. By combining this with the list of indefinite integrals in Theorem 1.3.16, we can compute integrals of linear combinations of simple functions. For example

$$\begin{aligned}\int_1^4 \left(e^x - 2\sin x + 3x^2\right)\mathrm{d}x &= \int_1^4 e^x\mathrm{d}x - 2\int_1^4 \sin x\mathrm{d}x + 3\int_1^4 x^2\mathrm{d}x \\ &= \left(e^x + (-2)\cdot(-\cos x) + 3\frac{x^3}{3}\right)\Big|_1^4 \qquad \text{and so on}\end{aligned}$$

Of course there are a great many functions that can be approached in this way, however there are some very simple examples that cannot.

$$\int \sin(\pi x)\mathrm{d}x \qquad \int xe^x\mathrm{d}x \qquad \int \frac{x}{x^2-5x+6}\mathrm{d}x$$

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,

$$\frac{\mathrm{d}}{\mathrm{d}x}(Af(x) + Bg(x)) = A\frac{\mathrm{d}f}{\mathrm{d}x} + B\frac{\mathrm{d}g}{\mathrm{d}x},$$

but had not yet seen the chain, product and quotient rules[33]. 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,

33 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 1.3.4, is the composition of e^x and $-x^2$. Even though we know

$$\int e^x \mathrm{d}x = e^x + C \qquad \text{and} \qquad \int -x^2 \mathrm{d}x = -\frac{1}{3}x^3 + C$$

there is no simple function that is equal to the indefinite integral

$$\int e^{-x^2} \mathrm{d}x.$$

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 1.4.1 (The chain rule).

Let $F(u)$ and $u(x)$ be differentiable functions and form their composition $F(u(x))$. Then

$$\frac{\mathrm{d}}{\mathrm{d}x} F\big(u(x)\big) = F'\big(u(x)\big) \cdot u'(x)$$

Equivalently, if $y(x) = F(u(x))$, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}F}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}.$$

Consider a function $f(u)$, which has antiderivative $F(u)$. Then we know that

$$\int f(u)\mathrm{d}u = \int F'(u)\mathrm{d}u = F(u) + C$$

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

$$\int f(u)\mathrm{d}u\Big|_{u=u(x)} = F(u(x)) + C$$

But now the right-hand side is a function of x, so we can differentiate it with respect to x to get

$$\frac{\mathrm{d}}{\mathrm{d}x} F(u(x)) = F'(u(x)) \cdot u'(x)$$

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

$$\int f\big(u(x)\big) \cdot u'(x)\,\mathrm{d}x = F\big(u(x)\big) + C = \int f(u)\,\mathrm{d}u\Big|_{u=u(x)}$$

This is the substitution rule for indefinite integrals.

Theorem 1.4.2 (The substitution rule — indefinite integral version)**.**

For any differentiable function $u(x)$:

$$\int f(u(x))u'(x)\mathrm{d}x = \int f(u)\mathrm{d}u\Big|_{u=u(x)}$$

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[34]. 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 $f\big(u(x)\big)$ Choose $u(x)$ to be the complicated argument.

Here are two examples which illustrate each of those strategies in turn.

Example 1.4.3

Consider the integral

$$\int 9\sin^8(x)\cos(x)\mathrm{d}x$$

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 1.3.16) 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.

34 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

$$\begin{aligned}\int 9\sin^8(x)\cos(x)\mathrm{d}x &= \int 9u(x)^8 \cdot u'(x)\mathrm{d}x \\ &= \int 9u^8\mathrm{d}u\Big|_{u=\sin(x)} \qquad \text{by the substitution rule}\end{aligned}$$

We are now left with the problem of antidifferentiating a monomial; this we can do with Theorem 1.3.16.

$$\begin{aligned}&= \left(u^9 + C\right)\Big|_{u=\sin(x)} \\ &= \sin^9(x) + C\end{aligned}$$

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 1.4.3

Example 1.4.4

Evaluate the integral

$$\int 3x^2\cos(x^3)\mathrm{d}x$$

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, $\cos\left(x^3\right)$ 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

$$\begin{aligned}\int 3x^2\cos(x^3)\mathrm{d}x &= \int u'(x)\cos\left(u(x)\right)\mathrm{d}x \qquad \text{just swap order of factors} \\ &= \int \cos\left(u(x)\right)u'(x)\mathrm{d}x \qquad \text{by the substitution rule} \\ &= \int \cos(u)\mathrm{d}u\Big|_{u=x^3} \\ &= (\sin(u) + C)\Big|_{u=x^3} \qquad \text{using Theorem 1.3.16)} \\ &= \sin(x^3) + C\end{aligned}$$

Example 1.4.4

One more — we'll use this to show how to use the substitution rule with definite integrals.

Example 1.4.5 $\left(\int_0^1 e^x \sin(e^x)\mathrm{d}x\right)$

Compute

$$\int_0^1 e^x \sin\left(e^x\right)\mathrm{d}x.$$

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)

$$\begin{aligned}\int e^x \sin\left(e^x\right)\mathrm{d}x &= \int \sin\left(u(x)\right)u'(x)\mathrm{d}x && \text{apply the substitution rule}\\ &= \int \sin(u)\mathrm{d}u\Big|_{u=e^x}\\ &= \left(-\cos(u)+C\right)\Big|_{u=e^x}\\ &= -\cos\left(e^x\right)+C\end{aligned}$$

- 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

$$\begin{aligned}\int_0^1 e^x \sin\left(e^x\right)\mathrm{d}x &= \left[-\cos\left(e^x\right)\right]_0^1\\ &= -\cos(e^1)-(-\cos(e^0))\\ &= -\cos(e)+\cos(1)\end{aligned}$$

Example 1.4.5

Theorem 1.4.2, the substitution rule for indefinite integrals, tells us that if $F(u)$ is any antiderivative for $f(u)$, then $F\big(u(x)\big)$ is an antiderivative for $f\big(u(x)\big)u'(x)$. So the funda-

mental theorem of calculus gives us

$$\begin{aligned}\int_a^b f\big(u(x)\big)u'(x)\,\mathrm{d}x &= F\big(u(x)\big)\Big|_{x=a}^{x=b}\\ &= F\big(u(b)\big) - F\big(u(a)\big)\\ &= \int_{u(a)}^{u(b)} f(u)\,\mathrm{d}u \qquad \text{since } F(u) \text{ is an antiderivative for } f(u)\end{aligned}$$

and we have just found

Theorem 1.4.6 (The substitution rule — definite integral version).

For any differentiable function $u(x)$:

$$\int_a^b f(u(x))u'(x)\mathrm{d}x = \int_{u(a)}^{u(b)} f(u)\mathrm{d}u$$

Notice that to get from the integral on the left hand side to the integral on the right hand side you

- substitute[35] $u(x) \to u$ and $u'(x)\mathrm{d}x \to \mathrm{d}u$,
- 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
- 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 $\int_a^b f\big(u(x)\big)u'(x)\,\mathrm{d}x$.

- We can find the indefinite integral $\int f\big(u(x)\big)u'(x)\,\mathrm{d}x$, using Theorem 1.4.2, and then evaluate the result between $x = a$ and $x = b$. This is what was done in Example 1.4.5.
- Or we can apply Theorem 1.4.2. This entails finding the indefinite integral $\int f(u)\,\mathrm{d}u$ and evaluating the result between $u = u(a)$ and $u = u(b)$. This is what we will do in the following example.

Example 1.4.7 $\left(\int_0^1 x^2\sin(x^3+1)\mathrm{d}x\right)$

Compute

$$\int_0^1 x^2 \sin\big(x^3+1\big)\mathrm{d}x$$

Solution.

35 A good way to remember this last step is that we replace $\frac{\mathrm{d}u}{\mathrm{d}x}\mathrm{d}x$ by just $\mathrm{d}u$ — which looks like we cancelled out the $\mathrm{d}x$ terms: $\frac{\mathrm{d}u}{\cancel{\mathrm{d}x}}\cancel{\mathrm{d}x} = \mathrm{d}u$. While using "cancel the $\mathrm{d}x$" is a good mnemonic (memory aid), you should not think of the derivative $\frac{\mathrm{d}u}{\mathrm{d}x}$ as a fraction — you are not dividing $\mathrm{d}u$ by $\mathrm{d}x$.

- 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 $\sin\left(x^3+1\right)$, 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:

$$x^2\sin\left(x^3+1\right)=\frac{1}{3}\cdot 3x^2\cdot\sin\left(x^3+1\right).$$

- We want this in the form of the substitution rule, so we do a little massaging:

$$\begin{aligned}\int_0^1 x^2\sin\left(x^3+1\right)\mathrm{d}x &= \int_0^1 \frac{1}{3}\cdot 3x^2\cdot\sin\left(x^3+1\right)\mathrm{d}x\\ &= \frac{1}{3}\int_0^1 \sin\left(x^3+1\right)\cdot 3x^2\mathrm{d}x \qquad \text{by Theorem 1.2.1(c)}\end{aligned}$$

- Now we are ready for the substitution rule:

$$\begin{aligned}\frac{1}{3}\int_0^1 \sin\left(x^3+1\right)\cdot 3x^2\mathrm{d}x &= \frac{1}{3}\int_0^1 \underbrace{\sin\left(x^3+1\right)}_{=f(u(x))}\cdot\underbrace{3x^2}_{=u'(x)}\mathrm{d}x\\ &= \frac{1}{3}\int_0^1 f(u(x))u'(x)\mathrm{d}x \qquad \text{with } u(x)=x^3+1 \text{ and } f(u)=\sin(u)\\ &= \frac{1}{3}\int_{u(0)}^{u(1)} f(u)\mathrm{d}u \qquad \text{by the substitution rule}\\ &= \frac{1}{3}\int_1^2 \sin(u)\mathrm{d}u \qquad \text{since } u(0)=1 \text{ and } u(1)=2\\ &= \frac{1}{3}\big[-\cos(u)\big]_1^2\\ &= \frac{1}{3}\big(-\cos(2)-(-\cos(1))\big)\\ &= \frac{\cos(1)-\cos(2)}{3}.\end{aligned}$$

Example 1.4.7

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 $u(x)\to u$,
- replace $\mathrm{d}x\to\dfrac{1}{u'(x)}\mathrm{d}u$.

In so doing, we take the integral

$$\begin{aligned}\int_a^b f(u(x))\cdot u'(x)\mathrm{d}x &= \int_{u(a)}^{u(b)} f(u)\cdot u'(x)\cdot\frac{1}{u'(x)}\mathrm{d}u \\ &= \int_{u(a)}^{u(b)} f(u)\mathrm{d}u \qquad \text{exactly the substitution rule}\end{aligned}$$

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 1.4.8 (*Example 1.4.7 revisited*)

Compute the integral

$$\int_0^1 x^2\sin\left(x^3+1\right)\mathrm{d}x$$

Solution.

- We have already observed that one factor of the integrand is $\sin\left(x^3+1\right)$, 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 $u(x)=x^3+1\to u$ and $\mathrm{d}x\to\frac{1}{u'(x)}\mathrm{d}u=\frac{1}{3x^2}\mathrm{d}u$:

$$\begin{aligned}\int_0^1 x^2\sin\left(x^3+1\right)\mathrm{d}x &= \int_{u(0)}^{u(1)} x^2\underbrace{\sin\left(x^3+1\right)}_{=\sin(u)}\frac{1}{3x^2}\mathrm{d}u \\ &= \int_1^2 \sin(u)\frac{x^2}{3x^2}\mathrm{d}u \\ &= \int_1^2 \frac{1}{3}\sin(u)\mathrm{d}u \\ &= \frac{1}{3}\int_1^2 \sin(u)\mathrm{d}u\end{aligned}$$

  which is precisely the integral we found in Example 1.4.7.

Example 1.4.8

Example 1.4.9

Compute the indefinite integrals

$$\int\sqrt{2x+1}\mathrm{d}x \qquad \text{and} \qquad \int e^{3x-2}\mathrm{d}x$$

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 $2x + 1 \to u$ and $\mathrm{d}x \to \frac{1}{u'(x)}\mathrm{d}u = \frac{1}{2}\mathrm{d}u$:

$$\begin{aligned}\int \sqrt{2x+1}\mathrm{d}x &= \int \sqrt{u}\frac{1}{2}\mathrm{d}u \\ &= \int u^{1/2}\frac{1}{2}\mathrm{d}u \\ &= \left(\frac{2}{3}u^{3/2}\cdot\frac{1}{2} + C\right)\Big|_{u=2x+1} \\ &= \frac{1}{3}(2x+1)^{3/2} + C\end{aligned}$$

- We can evaluate the second integral in much the same way. Set $u(x) = 3x - 2$ and replace $\mathrm{d}x$ by $\frac{1}{u'(x)}\mathrm{d}u = \frac{1}{3}\mathrm{d}u$:

$$\begin{aligned}\int e^{3x-2}\mathrm{d}x &= \int e^{u}\frac{1}{3}\mathrm{d}u \\ &= \left(\frac{1}{3}e^{u} + C\right)\Big|_{u=3x-2} \\ &= \frac{1}{3}e^{3x-2} + C\end{aligned}$$

Example 1.4.9

This last example illustrates that substitution can be used to easily deal with arguments of the form $ax + b$ (with a, b constants and $a \neq 0$), i.e. that are linear functions of x, and suggests the following theorem.

Theorem 1.4.10.

Let $F(u)$ be an antiderivative of $f(u)$ and let a, b be constants with $a \neq 0$. Then

$$\int f(ax+b)\mathrm{d}x = \frac{1}{a}F(ax+b) + C$$

Proof. We can show this using the substitution rule. Let $u(x) = ax + b$ so $u'(x) = a$, then

$$\begin{aligned}
\int f(ax+b)\mathrm{d}x &= \int f(u) \cdot \frac{1}{u'(x)}\mathrm{d}u \\
&= \int \frac{1}{a} f(u)\mathrm{d}u \\
&= \frac{1}{a}\int f(u)\mathrm{d}u && \text{since } a \text{ is a constant} \\
&= \frac{1}{a}F(u)\Big|_{u=ax+b} + C && \text{since } F(u) \text{ is an antiderivative of } f(u) \\
&= \frac{1}{a}F(ax+b) + C.
\end{aligned}$$

□

Now we can do the following example using the substitution rule or the above theorem:

Example 1.4.11 $\left(\int_0^{\pi/2} \cos(3x)\mathrm{d}x\right)$

Compute $\int_0^{\pi/2} \cos(3x)\mathrm{d}x$.

- In this example we should set $u = 3x$, and substitute $\mathrm{d}x \to \frac{1}{u'(x)}\mathrm{d}u = \frac{1}{3}\mathrm{d}u$. 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

$$\begin{aligned}
\int_0^{\pi/2} \cos(3x)\mathrm{d}x &= \int_0^{3\pi/2} \cos(u)\frac{1}{3}\mathrm{d}u \\
&= \left[\frac{1}{3}\sin(u)\right]_0^{3\pi/2} \\
&= \frac{\sin(3\pi/2) - \sin(0)}{3} \\
&= \frac{-1-0}{3} = -\frac{1}{3}.
\end{aligned}$$

- We can also do this example more directly using the above theorem. Since $\sin(x)$ is an antiderivative of $\cos(x)$, Theorem 1.4.10 tells us that $\frac{\sin(3x)}{3}$ is an antiderivative of $\cos(3x)$. Hence

$$\begin{aligned}
\int_0^{\pi/2} \cos(3x)\mathrm{d}x &= \left[\frac{\sin(3x)}{3}\right]_0^{\pi/2} \\
&= \frac{\sin(3\pi/2) - \sin(0)}{3} \\
&= -\frac{1}{3}.
\end{aligned}$$

Example 1.4.11

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.

Example 1.4.12 $\left(\int_0^1 x^2 \sin(1-x^3)\mathrm{d}x\right)$

This integral looks a lot like that of Example 1.4.7. 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 $\mathrm{d}x$ with $\frac{1}{u'(x)}\mathrm{d}u = \frac{1}{-3x^2}\mathrm{d}u$,
- when $x = 0$, we have $u = 1 - 0^3 = 1$ and
- when $x = 1$, we have $u = 1 - 1^3 = 0$.

So

$$\begin{aligned}\int_0^1 x^2 \sin\left(1-x^3\right)\cdot \mathrm{d}x &= \int_1^0 x^2 \sin(u)\cdot \frac{1}{-3x^2}\mathrm{d}u \\ &= \int_1^0 -\frac{1}{3}\sin(u)\mathrm{d}u.\end{aligned}$$

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 1.4.12

Example 1.4.13 $\left(\int_0^1 \frac{1}{(2x+1)^3}\mathrm{d}x\right)$

Compute $\int_0^1 \frac{1}{(2x+1)^3}\mathrm{d}x$.

We could do this one using Theorem 1.4.10, but its 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 $\mathrm{d}x \to \frac{1}{u'(x)}\mathrm{d}u = \frac{1}{2}\mathrm{d}u$.
- When $x = 0$, we have $u = 2 \times 0 + 1 = 1$ and
- when $x = 1$, we have $u = 2 \times 1 + 1 = 3$.

So

$$\begin{aligned}\int_0^1 \frac{1}{(2x+1)^3}\mathrm{d}x &= \int_1^3 \frac{1}{u^3}\cdot\frac{1}{2}\mathrm{d}u \\ &= \frac{1}{2}\int_1^3 u^{-3}\mathrm{d}u \\ &= \frac{1}{2}\left[\frac{u^{-2}}{-2}\right]_1^3 \\ &= \frac{1}{2}\left(\frac{1}{-2}\cdot\frac{1}{9} - \frac{1}{-2}\cdot\frac{1}{1}\right) \\ &= \frac{1}{2}\left(\frac{1}{2} - \frac{1}{18}\right) = \frac{1}{2}\cdot\frac{8}{18} \\ &= \frac{2}{9}\end{aligned}$$

Example 1.4.13

Example 1.4.14 $\left(\int_0^1 \frac{x}{1+x^2}\mathrm{d}x\right)$

Evaluate $\int_0^1 \frac{x}{1+x^2}\mathrm{d}x$.

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 $\mathrm{d}x \to \frac{1}{2x}\mathrm{d}u$, 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

$$\begin{aligned}\int_0^1 \frac{x}{1+x^2}\mathrm{d}x &= \int_1^2 \frac{x}{u}\frac{1}{2x}\mathrm{d}u \\ &= \int_1^2 \frac{1}{2u}\mathrm{d}u \\ &= \frac{1}{2}\big[\log|u|\big]_1^2 \\ &= \frac{\log 2 - \log 1}{2} = \frac{\log 2}{2}.\end{aligned}$$

Remember that we are using the notation "log" for the natural logarithm, i.e. the logarithm with base e. You might also see it written as "$\ln x$", or with the base made explicit as "$\log_e x$".

Example 1.4.14

Example 1.4.15 $\left(\int x^3\cos\left(x^4+2\right)\mathrm{d}x\right)$

Compute the integral $\int x^3\cos\left(x^4+2\right)\mathrm{d}x$.

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 $\mathrm{d}x \to \frac{1}{u'(x)}\mathrm{d}u = \frac{1}{4x^3}\mathrm{d}u$.
- 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

$$\begin{aligned}\int x^3\cos\left(x^4+2\right)\mathrm{d}x &= \int x^3\cos(u)\frac{1}{4x^3}\mathrm{d}u\Big|_{u=x^4+2} \\ &= \int \frac{1}{4}\cos(u)\mathrm{d}u\Big|_{u=x^4+2} \\ &= \left(\frac{1}{4}\sin(u)+C\right)\Big|_{u=x^4+2} \\ &= \frac{1}{4}\sin(x^4+2)+C.\end{aligned}$$

Example 1.4.15

The next two examples are more involved and require more careful thinking.

Example 1.4.16 $\left(\int\sqrt{1+x^2}\,x^3\mathrm{d}x\right)$

Compute $\int\sqrt{1+x^2}\,x^3\mathrm{d}x$.

- An obvious choice of u is the argument inside the square root. So substitute $u = 1 + x^2$ and $\mathrm{d}x \to \frac{1}{2x}\mathrm{d}u$.
- When we do this we obtain

$$\begin{aligned}\int \sqrt{1+x^2}\cdot x^3 \mathrm{d}x &= \int \sqrt{u}\cdot x^3 \cdot \frac{1}{2x}\mathrm{d}u \\ &= \int \frac{1}{2}\sqrt{u}\cdot x^2 \mathrm{d}u\end{aligned}$$

  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 1.4.2.)

- 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

$$\begin{aligned}\int \sqrt{1+x^2}\cdot x^3 \mathrm{d}x &= \int \frac{1}{2}\sqrt{u}\cdot x^2 \mathrm{d}u \\ &= \int \frac{1}{2}\sqrt{u}\cdot (u-1)\mathrm{d}u \\ &= \frac{1}{2}\int \left(u^{3/2} - u^{1/2}\right)\mathrm{d}u \\ &= \frac{1}{2}\left(\frac{2}{5}u^{5/2} - \frac{2}{3}u^{3/2}\right)\Big|_{u=x^2+1} + C \\ &= \left(\frac{1}{5}u^{5/2} - \frac{1}{3}u^{3/2}\right)\Big|_{u=x^2+1} + C \\ &= \frac{1}{5}(x^2+1)^{5/2} - \frac{1}{3}(x^2+1)^{3/2} + C.\end{aligned}$$

  Oof!

- Don't forget that you can always check the answer by differentiating:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{5}(x^2+1)^{5/2} - \frac{1}{3}(x^2+1)^{3/2} + C\right) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{5}(x^2+1)^{5/2}\right) - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{3}(x^2+1)^{3/2}\right) \\ &= \frac{1}{5}\cdot 2x \cdot \frac{5}{2}\cdot (x^2+1)^{3/2} - \frac{1}{3}\cdot 2x\cdot \frac{3}{2}\cdot (x^2+1)^{1/2} \\ &= x(x^2+1)^{3/2} - x(x^2+1)^{1/2} \\ &= x\left[(x^2+1)-1\right]\cdot\sqrt{x^2+1} \\ &= x^3\sqrt{x^2+1}.\end{aligned}$$

  which is the original integrand ✓.

Example 1.4.16

Example 1.4.17 ($\int \tan x \mathrm{d}x$)

Evaluate the indefinite integral $\int \tan(x)\mathrm{d}x$.

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 $\int \frac{\sin x}{\cos x}\mathrm{d}x$. 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 $\mathrm{d}x \to \frac{1}{-\sin x}\mathrm{d}u$ to give:

$$\begin{aligned}
\int \frac{\sin x}{\cos x}\mathrm{d}x &= \int \frac{\sin x}{u}\frac{1}{-\sin x}\mathrm{d}u\Big|_{u=\cos x} \\
&= \int -\frac{1}{u}\mathrm{d}u\Big|_{u=\cos x} \\
&= -\log|\cos x| + C \qquad \text{and if we want to go further} \\
&= \log\left|\frac{1}{\cos x}\right| + C \\
&= \log|\sec x| + C.
\end{aligned}$$

Example 1.4.17

In all of the above substitution examples we expressed the new integration variable, u, as a function, $u(x)$, of the old integration variable x. It is also possible to express the old integration variable, x, as a function, $x(u)$, of the new integration variable u. We shall see examples of this in §1.9.

1.5▲ Area Between Curves

Before we continue our exploration of different methods for integrating functions, we have now have sufficient tools to examine some simple applications of definite integrals. One of the motivations for our definition of "integral" was the problem of finding the area between some curve and the x–axis for x running between two specified values. More precisely

$$\int_a^b f(x)\mathrm{d}x$$

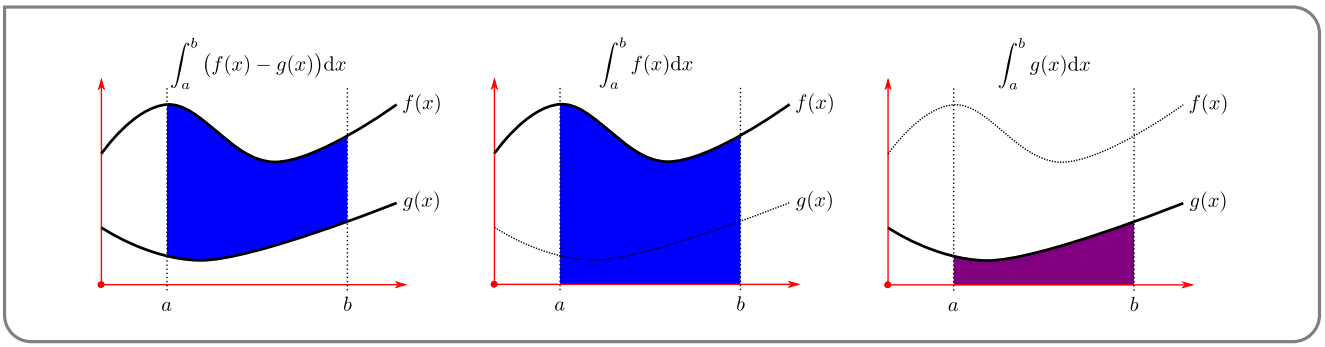
is equal to the signed area between the curve $y = f(x)$, the x-axis, and the vertical lines $x = a$ and $x = b$.

We found the area of this region by approximating it by the union of tall thin rectangles, and then found the exact area by taking the limit as the width of the approximating rectangles went to zero. We can use the same strategy to find areas of more complicated regions in the xy-plane.

As a preview of the material to come, let $f(x) > g(x) > 0$ and $a < b$ and suppose that we are interested in the area of the region

$$S_1 = \{\ (x,y)\ |\ a \leqslant x \leqslant b\,,\ g(x) \leqslant y \leqslant f(x)\ \}$$

that is sketched in the left hand figure below.



We already know that $\int_a^b f(x)\,\mathrm{d}x$ is the area of the region

$$S_2 = \{\ (x,y)\ |\ a \leqslant x \leqslant b\,,\ 0 \leqslant y \leqslant f(x)\ \}$$

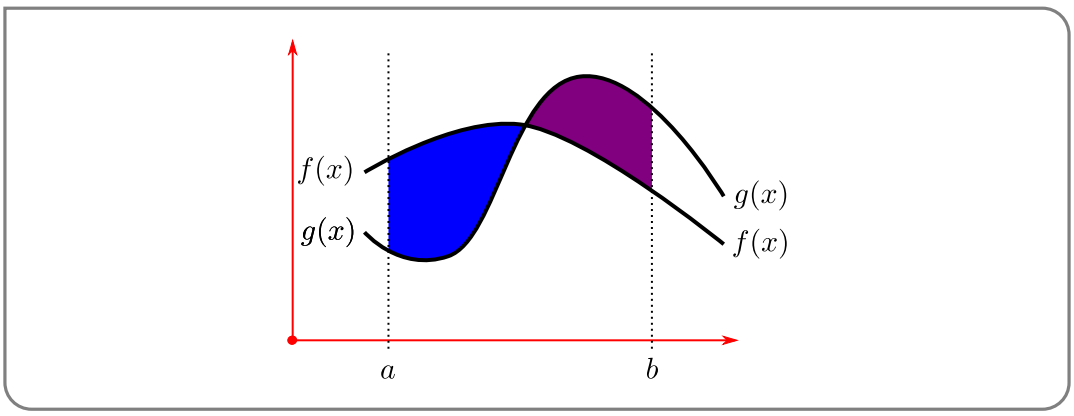
sketched in the middle figure above and that $\int_a^b g(x)\,\mathrm{d}x$ is the area of the region

$$S_3 = \{\ (x,y)\ |\ a \leqslant x \leqslant b\,,\ 0 \leqslant y \leqslant g(x)\ \}$$

sketched in the right hand figure above. Now the region S_1 of the left hand figure can be constructed by taking the region S_2 of center figure and removing from it the region S_3 of the right hand figure. So the area of S_1 is exactly

$$\int_a^b f(x)\,\mathrm{d}x - \int_a^b g(x)\,\mathrm{d}x = \int_a^b \big(f(x)-g(x)\big)\,\mathrm{d}x$$

This computation depended on the assumption that $f(x) > g(x)$ and, in particular, that the curves $y = g(x)$ and $y = f(x)$ did not cross. If they do cross, as in this figure



then we have to be a lot more careful. The idea is to separate the domain of integration depending on where $f(x) - g(x)$ changes sign — i.e. where the curves intersect. We will illustrate this in Example 1.5.5 below.

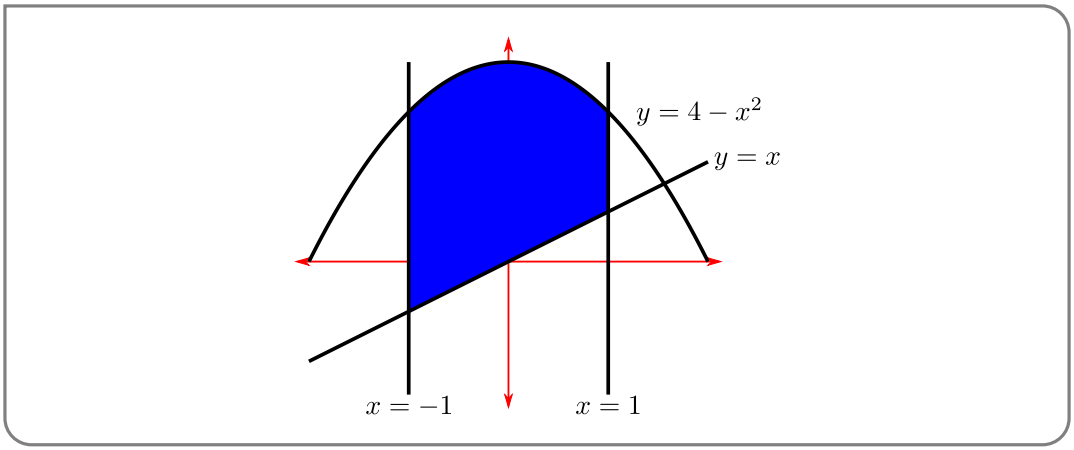
Let us start with an example that makes the link to Riemann sums and definite integrals quite explicit.

Example 1.5.1

Find the area bounded by the curves $y = 4 - x^2$, $y = x$, $x = -1$ and $x = 1$.

Solution.

- Before we do any calculus, it is a very good idea to make a sketch of the area in question. The curves $y = x$, $x = -1$ and $x = 1$ are all straight lines, while the curve $y = 4 - x^2$ is a parabola whose apex is at $(0, 4)$ and then curves down (because of the minus sign in $-x^2$) with x-intercepts at $(\pm 2, 0)$. Putting these together gives



  Notice that the curves $y = 4 - x^2$ and $y = x$ intersect when $4 - x^2 = x$, namely when $x = \frac{1}{2}\left(-1 \pm \sqrt{17}\right) \approx 1.56, -2.56$. Hence the curve $y = 4 - x^2$ lies above the line $y = x$ for all $-1 \leqslant x \leqslant 1$.

- We are to find the area of the shaded region. Each point (x, y) in this shaded region has $-1 \leqslant x \leqslant 1$ and $x \leqslant y \leqslant 4 - x^2$. When we were defining the integral (way back in Definition 1.1.9) we used a and b to denote the smallest and largest allowed values of x; let's do that here too. Let's also use $B(x)$ to denote the bottom curve (i.e. to denote the smallest allowed value of y for a given x) and use $T(x)$ to denote the top curve (i.e. to denote the largest allowed value of y for a given x). So in this example

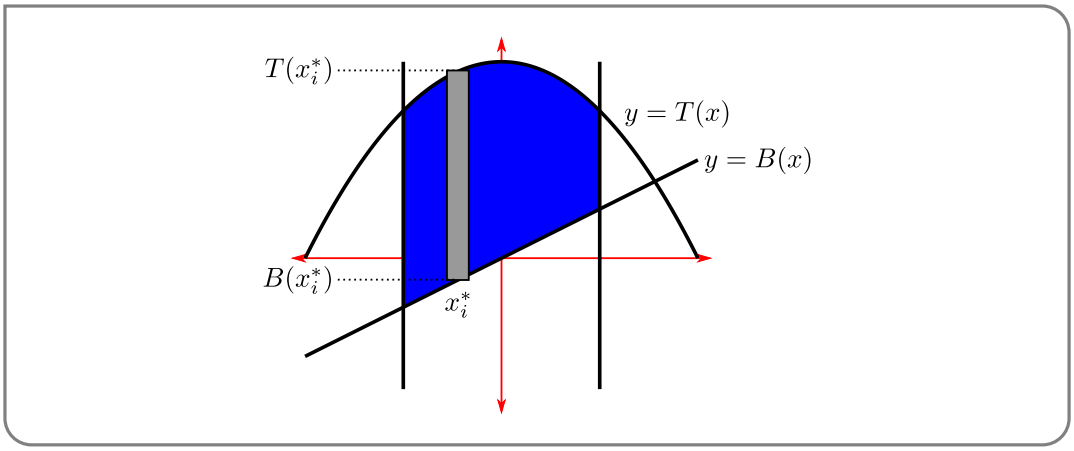
$$a = -1 \qquad b = 1 \qquad B(x) = x \qquad T(x) = 4 - x^2$$

  and the shaded region is

$$\left\{ (x, y) \,\middle|\, a \leqslant x \leqslant b,\ B(x) \leqslant y \leqslant T(x) \right\}$$

- We use the same strategy as we used when defining the integral in Section 1.1.4:

- Pick a natural number n (that we will later send to infinity), then
- subdivide the region into n narrow slices, each of width $\Delta x = \frac{b-a}{n}$.
- For each $i = 1, 2, \ldots, n$, slice number i runs from $x = x_{i-1}$ to $x = x_i$, and we approximate its area by the area of a rectangle. We pick a number x_i^* between x_{i-1} and x_i and approximate the slice by a rectangle whose top is at $y = T(x_i^*)$ and whose bottom is at $y = B(x_i^*)$.
- Thus the area of slice i is approximately $\left[T(x_i^*) - B(x_i^*)\right]\Delta x$ (as shown in the figure below).



- So the Riemann sum approximation of the area is

$$\text{Area} \approx \sum_{i=1}^{n} \left[T(x_i^*) - B(x_i^*)\right]\Delta x$$

- By taking the limit as $n \to \infty$ (i.e. taking the limit as the width of the rectangles goes to zero), we convert the Riemann sum into a definite integral (see Definition 1.1.9)

and at the same time our approximation of the area becomes the exact area:

$$\begin{aligned}
\lim_{n\to\infty}\sum_{i=1}^{n}\big[T(x_i^*)-B(x_i^*)\big]\Delta x &= \int_a^b \big[T(x)-B(x)\big]\,\mathrm{d}x && \text{Riemann sum} \to \text{integral}\\
&= \int_{-1}^{1}\big[(4-x^2)-x\big]\,\mathrm{d}x\\
&= \int_{-1}^{1}\big[4-x-x^2\big]\,\mathrm{d}x\\
&= \left[4x-\frac{x^2}{2}-\frac{x^3}{3}\right]_{-1}^{1}\\
&= \left(4-\frac{1}{2}-\frac{1}{3}\right)-\left(-4-\frac{1}{2}+\frac{1}{3}\right)\\
&= \frac{24-3-2}{6}-\frac{-24-3+2}{6}\\
&= \frac{19}{6}+\frac{25}{6}\\
&= \frac{44}{6}=\frac{22}{3}.
\end{aligned}$$

Example 1.5.1

Oof! Thankfully we generally do not need to go through the Riemann sum steps to get to the answer. Usually, provided we are careful to check where curves intersect and which curve lies above which, we can just jump straight to the integral

$$\text{Area} = \int_a^b \big[T(x)-B(x)\big]\,\mathrm{d}x. \tag{1.5.1}$$

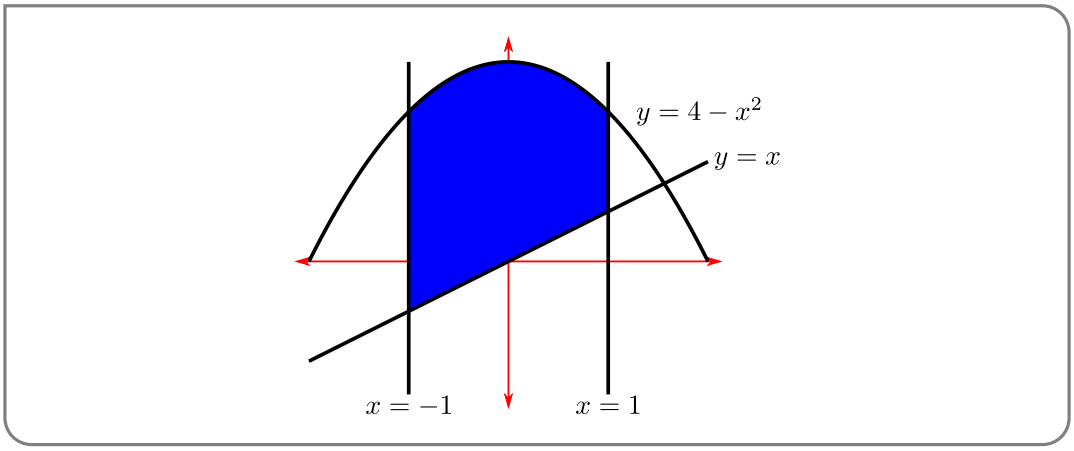
So let us redo the above example.

Example 1.5.2 (Example 1.5.1 revisited)

Find the area bounded by the curves $y = 4-x^2$, $y = x$, $x = -1$ and $x = 1$.

Solution.

- We first sketch the region



and verify[36] that $y = T(x) = 4 - x^2$ lies above the curve $y = B(x) = x$ on the region $-1 \leqslant x \leqslant 1$.

- The area between the curves is then

$$\begin{aligned}\text{Area} &= \int_a^b \big[T(x) - B(x)\big]\,dx \\ &= \int_{-1}^{1} \big[4 - x - x^2\big]\,dx \\ &= \left[4x - \frac{x^2}{2} - \frac{x^3}{3}\right]_{-1}^{1} \\ &= \frac{19}{6} + \frac{25}{6} = \frac{44}{6} = \frac{22}{3}.\end{aligned}$$

Example 1.5.2

Example 1.5.3

Find the area of the finite region bounded by $y = x^2$ and $y = 6x - 2x^2$.

Solution. This is a little different from the previous question, since we are not given bounding lines $x = a$ and $x = b$ — instead we have to determine the minimum and maximum allowed values of x by determining where the curves intersect. Hence our very first task is to get a good idea of what the region looks like by sketching it.

- Start by sketching the region:
  - The curve $y = x^2$ is a parabola. The point on this parabola with the smallest y–coordinate is $(0, 0)$. As $|x|$ increases, y increases so the parabola opens upward.
  - The curve $y = 6x - 2x^2 = -2(x^2 - 3x) = -2(x - \frac{3}{2})^2 + \frac{9}{2}$ is also a parabola. The point on this parabola with the largest value of y has $x = 3/2$ (so that the

36 We should do this by checking where the curves intersect; that is by solving $T(x) = B(x)$ and seeing if any of the solutions lie in the range $-1 \leqslant x \leqslant 1$.

negative term in $-2(x-\frac{3}{2})^2+\frac{9}{2}$ is zero). So the point with the largest value of y is is $(3/2, 9/2)$. As x moves away from $3/2$, either to the right or to the left, y decreases. So the parabola opens downward. The parabola crosses the x–axis when $0 = 6x - 2x^2 = 2x(3-x)$. That is, when $x = 0$ and $x = 3$.

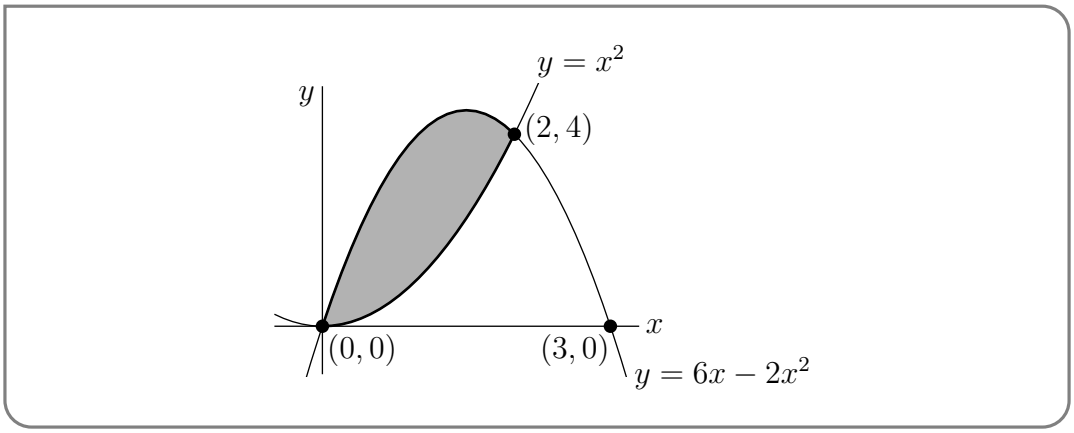
– The two parabolas intersect when $x^2 = 6x - 2x^2$, or

$$\begin{aligned} 3x^2 - 6x &= 0 \\ 3x(x-2) &= 0 \end{aligned}$$

So there are two points of intersection, one being $x = 0$, $y = 0^2 = 0$ and the other being $x = 2$, $y = 2^2 = 4$.

– The finite region between the curves lies between these two points of intersection.

This leads us to the sketch



- So on this region we have $0 \leqslant x \leqslant 2$, the top curve is $T(x) = 6x - 2x^2$ and the bottom curve is $B(x) = x^2$. Hence the area is given by

$$\begin{aligned} \text{Area} &= \int_a^b \big[T(x) - B(x)\big]\,\mathrm{d}x \\ &= \int_0^2 \big[(6x - 2x^2) - (x^2)\big]\,\mathrm{d}x \\ &= \int_0^2 \big[6x - 3x^2\big]\,\mathrm{d}x \\ &= \left[6\frac{x^2}{2} - 3\frac{x^3}{3}\right]_0^2 \\ &= 3(2)^2 - 2^3 = 4 \end{aligned}$$

Example 1.5.3

Example 1.5.4

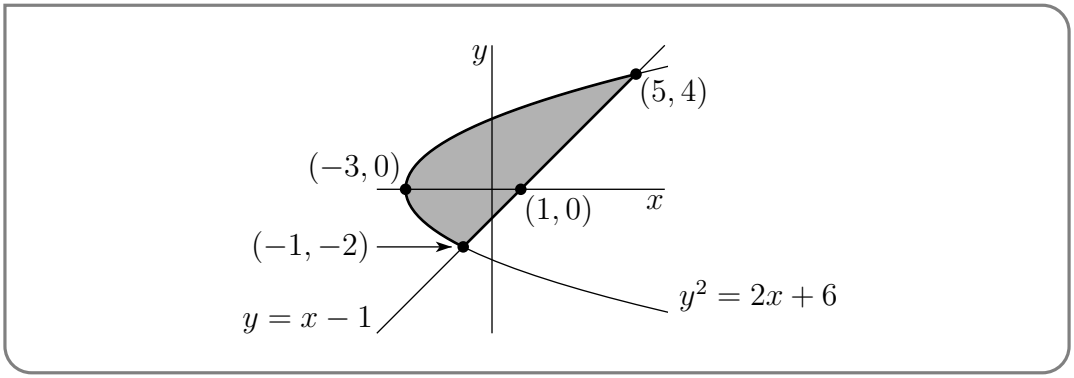
Find the area of the finite region bounded by $y^2 = 2x + 6$ and $y = x - 1$.

Solution. We show two different solutions to this problem. The first takes the approach we have in Example 1.5.3 but leads to messy algebra. The second requires a little bit of thinking at the beginning but then is quite straightforward. Before we get to that we should start by by sketching the region.

- The curve $y^2 = 2x + 6$, or equivalently $x = \frac{1}{2}y^2 - 3$ is a parabola. The point on this parabola with the smallest x–coordinate has $y = 0$ (so that the positive term in $\frac{1}{2}y^2 - 3$ is zero). So the point on this parabola with the smallest x–coordinate is $(-3, 0)$. As $|y|$ increases, x increases so the parabola opens to the right.
- The curve $y = x - 1$ is a straight line of slope 1 that passes through $x = 1$, $y = 0$.
- The two curves intersect when $\frac{y^2}{2} - 3 = y + 1$, or

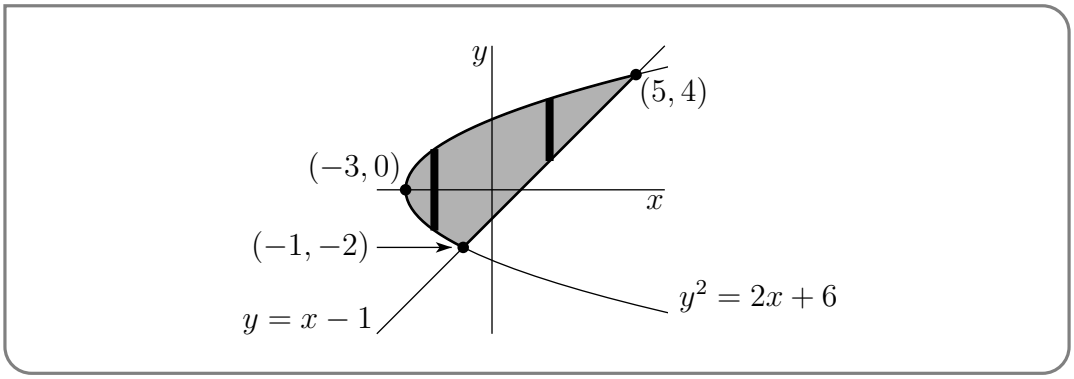
$$\begin{aligned} y^2 - 6 &= 2y + 2 \\ y^2 - 2y - 8 &= 0 \\ (y + 2)(y - 4) &= 0 \end{aligned}$$

  So there are two points of intersection, one being $y = 4$, $x = 4 + 1 = 5$ and the other being $y = -2$, $x = -2 + 1 = -1$.
- Putting this all together gives us the sketch



As noted above, we can find the area of this region by approximating it by a union of narrow vertical rectangles, as we did in Example 1.5.3 — though it is a little harder. The easy way is to approximate it by a union of narrow horizontal rectangles. Just for practice, here is the hard solution. The easy solution is after it.
Harder solution:

- As we have done previously, we approximate the region by a union of narrow vertical rectangles, each of width Δx. Two of those rectangles are illustrated in the sketch



- In this region, x runs from $a = -3$ to $b = 5$. The curve at the top of the region is
$$y = T(x) = \sqrt{2x+6}$$
The curve at the bottom of the region is more complicated. To the left of $(-1,-2)$ the lower half of the parabola gives the bottom of the region while to the right of $(-1,-2)$ the straight line gives the bottom of the region. So
$$B(x) = \begin{cases} -\sqrt{2x+6} & \text{if } -3 \leqslant x \leqslant -1 \\ x-1 & \text{if } -1 \leqslant x \leqslant 5 \end{cases}$$

- Just as before, the area is still given by the formula $\int_a^b \big[T(x) - B(x)\big]\mathrm{d}x$, but to accommodate our $B(x)$, we have to split up the domain of integration when we evaluate the integral.
$$\begin{aligned} \int_a^b \big[T(x) - B(x)\big]\mathrm{d}x &= \int_{-3}^{-1} \big[T(x) - B(x)\big]\mathrm{d}x + \int_{-1}^{5} \big[T(x) - B(x)\big]\mathrm{d}x \\ &= \int_{-3}^{-1} \big[\sqrt{2x+6} - (-\sqrt{2x+6})\big]\mathrm{d}x + \int_{-1}^{5} \big[\sqrt{2x+6} - (x-1)\big]\mathrm{d}x \\ &= 2\int_{-3}^{-1} \sqrt{2x+6}\mathrm{d}x + \int_{-1}^{5} \sqrt{2x+6} - \int_{-1}^{5} (x-1)\mathrm{d}x \end{aligned}$$

- The third integral is straightforward, while we evaluate the first two via the substitution rule. In particular, set $u = 2x+6$ and replace $\mathrm{d}x \to \frac{1}{2}\mathrm{d}u$. Also $u(-3) = 0, u(-1) = 4, u(5) = 16$. Hence
$$\begin{aligned} \text{Area} &= 2\int_0^4 \sqrt{u}\,\frac{\mathrm{d}u}{2} + \int_4^{16} \sqrt{u}\,\frac{\mathrm{d}u}{2} - \int_{-1}^{5} (x-1)\mathrm{d}x \\ &= 2\left[\frac{u^{3/2}}{3/2}\frac{1}{2}\right]_0^4 + \left[\frac{u^{3/2}}{3/2}\frac{1}{2}\right]_4^{16} - \left[\frac{x^2}{2} - x\right]_{-1}^{5} \\ &= \frac{2}{3}[8-0] + \frac{1}{3}[64-8] - \left[\left(\frac{25}{2} - 5\right) - \left(\frac{1}{2} + 1\right)\right] \\ &= \frac{72}{3} - \frac{24}{2} + 6 \\ &= 18 \end{aligned}$$
Oof!

Easier solution:
The easy way to determine the area of our region is to approximate by narrow horizontal rectangles, rather than narrow vertical rectangles. (Really we are just swapping the roles of x and y in this problem)

- Look at our sketch of the region again — each point (x, y) in our region has $-2 \leqslant y \leqslant 4$ and $\frac{1}{2}(y^2 - 6) \leqslant x \leqslant y + 1$.

- Let's use

  - c to denote the smallest allowed value of y,
  - d to denote the largest allowed value of y
  - $L(y)$ ("L" stands for "left") to denote the smallest allowed value of x, when the y–coordinate is y, and
  - $R(y)$ ("R" stands for "right") to denote the largest allowed value of x, when the y–coordinate is y.

  So, in this example,

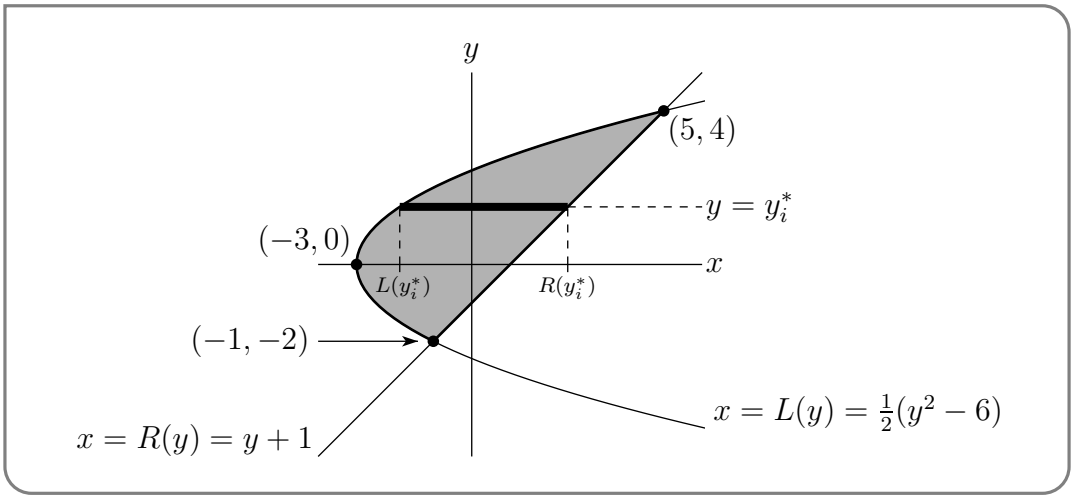
$$c = -2 \qquad d = 4 \qquad L(y) = \frac{1}{2}(y^2 - 6) \qquad R(y) = y + 1$$

  and the shaded region is

$$\left\{ (x, y) \,\middle|\, c \leqslant y \leqslant d,\ L(y) \leqslant x \leqslant R(y) \right\}$$

- Our strategy is now nearly the same as that used in Example 1.5.1:

  - Pick a natural number n (that we will later send to infinity), then
  - subdivide the interval $c \leqslant y \leqslant d$ into n narrow subintervals, each of width $\Delta y = \frac{d-c}{n}$. Each subinterval cuts a thin horizontal slice from the region (see the figure below).
  - We approximate the area of slice number i by the area of a thin horizontal rectangle (indicated by the dark rectangle in the figure below). On this slice, the y–coordinate runs over a very narrow range. We pick a number y_i^*, somewhere in that range. We approximate slice i by a rectangle whose left side is at $x = L(y_i^*)$ and whose right side is at $x = R(y_i^*)$.
  - Thus the area of slice i is approximately $\left[R(x_i^*) - L(x_i^*)\right]\Delta y$.



- The desired area is

$$\begin{aligned}
\lim_{n\to\infty}\sum_{i=1}^{n}\big[R(y_i^*)-L(y_i^*)\big]\Delta y &= \int_c^d \big[R(y)-L(y)\big]\,\mathrm{d}y && \text{Riemann sum} \to \text{integral}\\
&= \int_{-2}^{4}\big[(y+1)-\tfrac{1}{2}(y^2-6)\big]\,\mathrm{d}y\\
&= \int_{-2}^{4}\big[-\tfrac{1}{2}y^2+y+4\big]\,\mathrm{d}y\\
&= \Big[-\tfrac{1}{6}y^3+\tfrac{1}{2}y^2+4y\Big]_{-2}^{4}\\
&= -\tfrac{1}{6}\big(64-(-8)\big)+\tfrac{1}{2}(16-4)+4(4+2)\\
&= -12+6+24\\
&= 18
\end{aligned}$$

Example 1.5.4

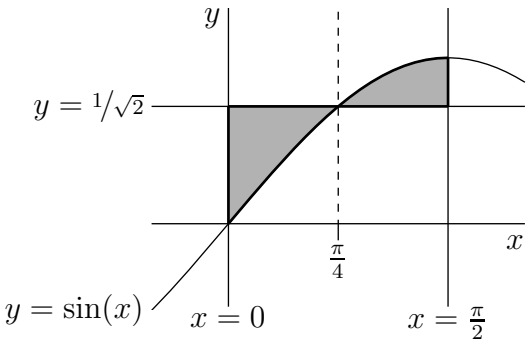
One last example.

Example 1.5.5

Find the area between the curves $y = \dfrac{1}{\sqrt{2}}$ and $y = \sin(x)$ with x running from 0 to $\pi/2$.

Solution. This one is a little trickier since (as we shall see) the region is split into two pieces and we need to treat them separately.

- Again we start by sketching the region.



We want the shaded area.

- Unlike our previous examples, the bounding curves $y = 1/\sqrt{2}$ and $y = \sin(x)$ cross in the middle of the region of interest. They cross when $y = 1/\sqrt{2}$ and $\sin(x) = y = 1/\sqrt{2}$, i.e. when $x = \pi/4$. So
  - to the left of $x = \pi/4$, the top boundary is part of the straight line $y = 1/\sqrt{2}$ and the bottom boundary is part of the curve $y = \sin(x)$
  - while to the right of $x = \pi/4$, the top boundary is part of the curve $y = \sin(x)$ and the bottom boundary is part of the straight line $y = 1/\sqrt{2}$.
- Thus the formulae for the top and bottom boundaries are

$$T(x) = \begin{Bmatrix} 1/\sqrt{2} & \text{if } 0 \leqslant x \leqslant \pi/4 \\ \sin(x) & \text{if } \pi/4 \leqslant x \leqslant \pi/2 \end{Bmatrix} \qquad B(x) = \begin{Bmatrix} \sin(x) & \text{if } 0 \leqslant x \leqslant \pi/4 \\ 1/\sqrt{2} & \text{if } \pi/4 \leqslant x \leqslant \pi/2 \end{Bmatrix}$$

We may compute the area of interest using our canned formula

$$\text{Area} = \int_a^b \big[T(x) - B(x)\big]\,\mathrm{d}x$$

but since the formulas for $T(x)$ and $B(x)$ change at the point $x = \pi/4$, we must split the domain of the integral in two at that point[37]

- Our integral over the domain $0 \leqslant x \leqslant \pi/2$ is split into an integral over $0 \leqslant x \leqslant \pi/4$

37 We are effectively computing the area of the region by computing the area of the two disjoint pieces separately. Alternatively, if we set $f(x) = \sin(x)$ and $g(x) = 1/\sqrt{2}$, we can rewrite the integral $\int_a^b \big[T(x) - B(x)\big]\,\mathrm{d}x$ as $\int_a^b \big|f(x) - g(x)\big|\,\mathrm{d}x$. To see that the two integrals are the same, split the domain of integration where $f(x) - g(x)$ changes sign.

and one over $\pi/4 \leqslant x \leqslant \pi/2$:

$$\begin{aligned}
\text{Area} &= \int_0^{\pi/2} \big[T(x) - B(x)\big]\,dx \\
&= \int_0^{\pi/4} \big[T(x) - B(x)\big]\,dx + \int_{\pi/4}^{\pi/2} \big[T(x) - B(x)\big]\,dx \\
&= \int_0^{\pi/4} \Big[\frac{1}{\sqrt{2}} - \sin(x)\Big]\,dx + \int_{\pi/4}^{\pi/2} \Big[\sin(x) - \frac{1}{\sqrt{2}}\Big]\,dx \\
&= \Big[\frac{x}{\sqrt{2}} + \cos(x)\Big]_0^{\pi/4} + \Big[-\cos(x) - \frac{x}{\sqrt{2}}\Big]_{\pi/4}^{\pi/2} \\
&= \Big[\frac{1}{\sqrt{2}}\frac{\pi}{4} + \frac{1}{\sqrt{2}} - 1\Big] + \Big[\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}\frac{\pi}{4}\Big] \\
&= \frac{2}{\sqrt{2}} - 1 \\
&= \sqrt{2} - 1
\end{aligned}$$

Example 1.5.5

1.6▲ Volumes

Another simple[38] application of integration is computing volumes. We use the same strategy as we used to express areas of regions in two dimensions as integrals — approximate the region by a union of small, simple pieces whose volume we can compute and then then take the limit as the "piece size" tends to zero.

In many cases this will lead to "multivariable integrals" that are beyond our present scope[39]. But there are some special cases in which this leads to integrals that we can handle. Here are some examples.

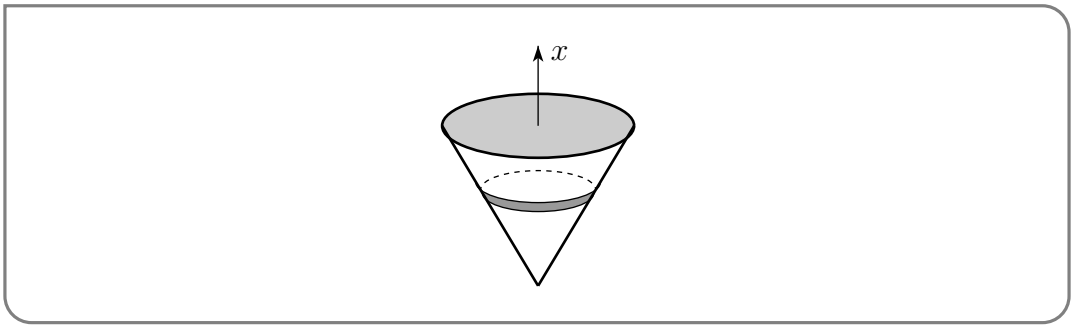
Example 1.6.1 (Cone)

Find the volume of the circular cone of height h and radius r.

Solution. Here is a sketch of the cone. We have called the vertical axis x, just so that we

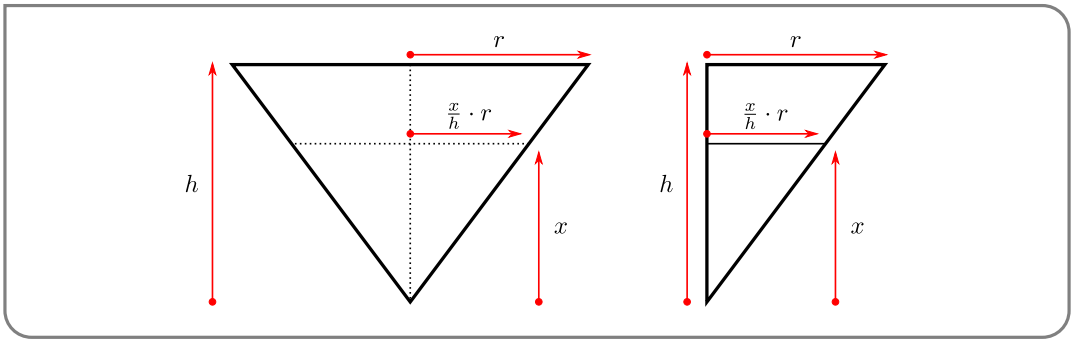
38 Well — arguably the idea isn't too complicated and is a continuation of the idea used to compute areas in the previous section. In practice this can be quite tricky as we shall see.

39 Typically such integrals (and more) are covered in a third calculus course.



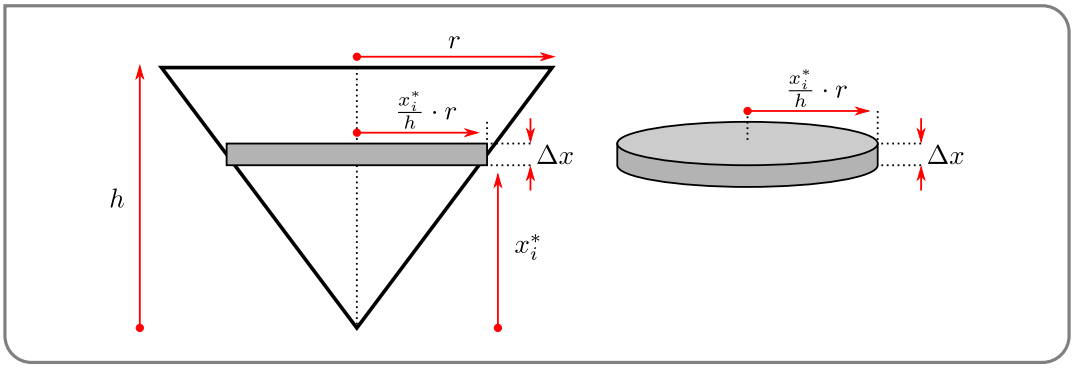
end up with a "dx" integral.

- In what follows we will slice the cone into thin horizontal "pancakes". In order to approximate the volume of those slices, we need to know the radius of the cone at a height x above its point. Consider the cross sections shown in the following figure.



  At full height h, the cone has radius r. If we cut the cone at height x, then by similar triangles (see the figure on the right) the radius will be $\frac{x}{h} \cdot r$.

- Now think of cutting the cone into n thin horizontal "pancakes". Each such pancake is approximately a squat cylinder of height $\Delta x = h/n$. This is very similar to how we approximated the area under a curve by n tall thin rectangles. Just as we approximated the area under the curve by summing these rectangles, we can approximate the volume of the cone by summing the volumes of these cylinders. Here is a side view of the cone and one of the cylinders.



- We follow the method we used in Example 1.5.1, except that our slices are now pancakes instead of rectangles.

  - Pick a natural number n (that we will later send to infinity), then
  - subdivide the cone into n thin pancakes, each of width $\Delta x = \frac{h}{n}$.
  - For each $i = 1, 2, \ldots, n$, pancake number i runs from $x = x_{i-1} = (i-1) \cdot \Delta x$ to $x = x_i = i \cdot \Delta x$, and we approximate its volume by the volume of a squat cone. We pick a number x_i^* between x_{i-1} and x_i and approximate the pancake by a cylinder of height Δx and radius $\frac{x_i^*}{h} r$.
  - Thus the volume of pancake i is approximately $\pi \left(\frac{x_i^*}{h} r\right)^2 \Delta x$ (as shown in the figure above).

- So the Riemann sum approximation of the volume is

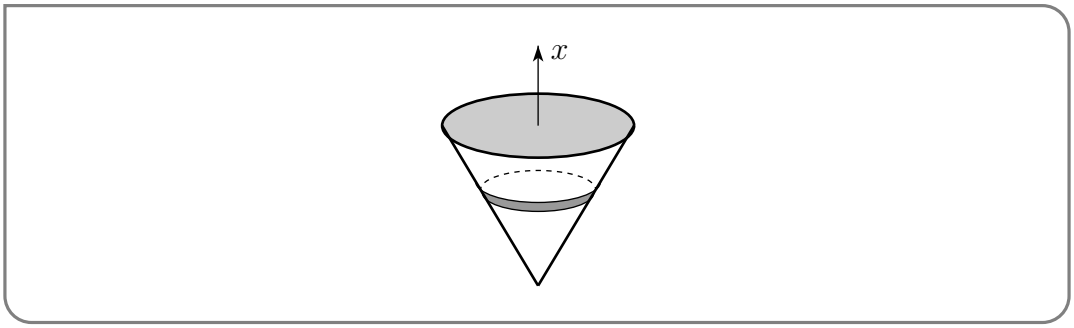
$$\text{Volume} \approx \sum_{i=1}^{n} \pi \left(\frac{x_i^*}{h} r\right)^2 \Delta x$$

- By taking the limit as $n \to \infty$ (i.e. taking the limit as the thickness of the pancakes goes to zero), we convert the Riemann sum into a definite integral (see Definition 1.1.9) and at the same time our approximation of the volume becomes the exact volume:

$$\int_0^h \pi \left(\frac{x}{h} r\right)^2 \mathrm{d}x$$

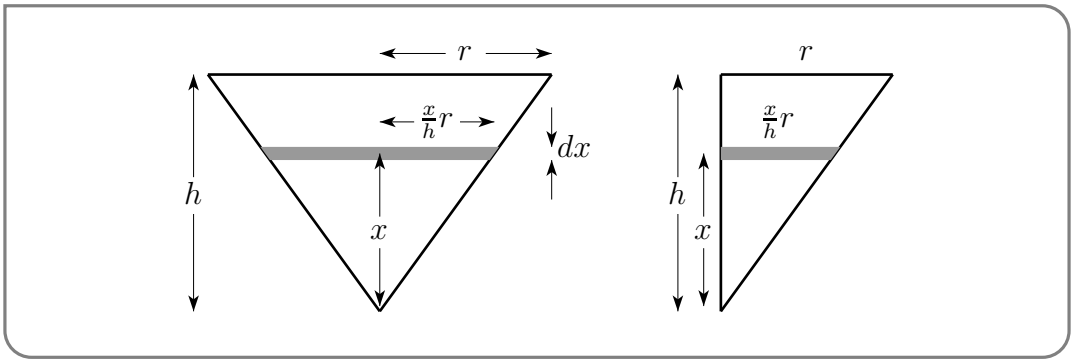
Our life[40] would be easier if we could avoid all this formal work with Riemann sums every time we encounter a new volume. So before we compute the above integral, let us redo the above calculation in a less formal manner.

- Start again from the picture of the cone and think of slicing it into thin pancakes,



each of width $\mathrm{d}x$.

40 At least the bits of it involving integrals.



- The pancake at height x above the point of the cone (which is the fraction $\frac{x}{h}$ of the total height of the cone) has
  - radius $\frac{x}{h} \cdot r$ (the fraction $\frac{x}{h}$ of the full radius, r) and so
  - cross–sectional area $\pi\left(\frac{x}{h}r\right)^2$,
  - thickness $\mathrm{d}x$ — we have done something a little sneaky here, see the discussion below.
  - volume $\pi\left(\frac{x}{h}r\right)^2\mathrm{d}x$

  As x runs from 0 to h, the total volume is

$$\begin{aligned}\int_0^h \pi\left(\frac{x}{h}r\right)^2\mathrm{d}x &= \frac{\pi r^2}{h^2}\int_0^h x^2\mathrm{d}x \\ &= \frac{\pi r^2}{h^2}\left[\frac{x^3}{3}\right]_0^h \\ &= \frac{1}{3}\pi r^2 h\end{aligned}$$

In this second computation we are using a time-saving trick. As we saw in the formal computation above, what we really need to do is pick a natural number n, slice the cone into n pancakes each of thickness $\Delta x = h/n$ and then take the limit as $n \to \infty$. This led to the Riemann sum

$$\sum_{i=1}^{n} \pi\left(\frac{x_i^*}{h}r\right)^2 \Delta x \qquad \text{which becomes} \int_0^h \pi\left(\frac{x}{h}r\right)^2 \mathrm{d}x$$

So knowing that we will replace

$$\begin{aligned}\sum_{i=1}^{n} &\longrightarrow \int_0^h \\ x_i^* &\longrightarrow x \\ \Delta x &\longrightarrow \mathrm{d}x\end{aligned}$$

when we take the limit, we have just skipped the intermediate steps. While this is not entirely rigorous, it can be made so, and does save us a lot of algebra.

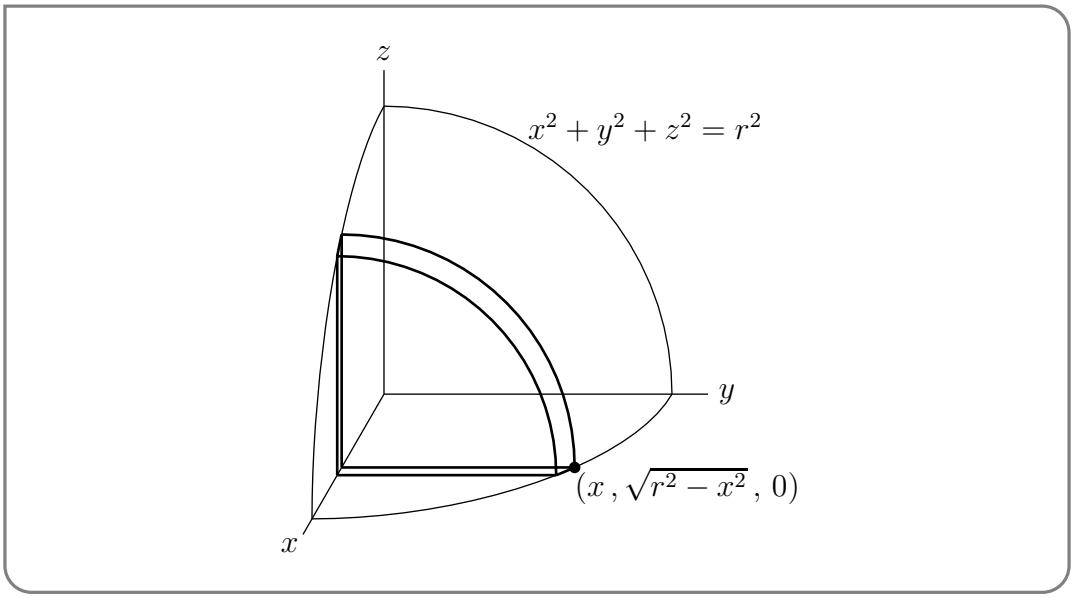
Example 1.6.1

Example 1.6.2 (Sphere)

Find the volume of the sphere of radius r.

Solution. We'll find the volume of the part of the sphere in the first octant[41], sketched below. Then we'll multiply by 8.

- To compute the volume, we slice it up into thin vertical "pancakes" (just as we did



  in the previous example).

- Each pancake is one quarter of a thin circular disk. The pancake a distance x from the yz–plane is shown in the sketch above. The radius of that pancake is the distance from the dot shown in the figure to the x–axis, i.e. the y–coordinate of the dot. To get the coordinates of the dot, observe that
  - it lies the xy–plane, and so has z–coordinate zero, and that
  - it also lies on the sphere, so that its coordinates obey $x^2 + y^2 + z^2 = r^2$. Since $z = 0$ and $y > 0$, $y = \sqrt{r^2 - x^2}$.
- So the pancake at distance x from the yz–plane has
  - thickness[42] $\mathrm{d}x$ and
  - radius $\sqrt{r^2 - x^2}$
  - cross–sectional area $\frac{1}{4}\pi\big(\sqrt{r^2 - x^2}\,\big)^2$ and hence
  - volume $\frac{\pi}{4}\big(r^2 - x^2\big)\mathrm{d}x$

41 The first octant is the set of all points (x, y, z) with $x \geqslant 0$, $y \geqslant 0$ and $z \geqslant 0$.

42 Yet again what we really do is pick a natural number n, slice the octant of the sphere into n pancakes each of thickness $\Delta x = \frac{r}{n}$ and then take the limit $n \to \infty$. In the integral Δx is replaced by $\mathrm{d}x$. Knowing that this is what is going to happen, we again just skip a few steps.

- As x runs from 0 to r, the total volume of the part of the sphere in the first octant is

$$\int_0^r \frac{\pi}{4}(r^2 - x^2)\,\mathrm{d}x = \frac{\pi}{4}\left[r^2 x - \frac{x^3}{3}\right]_0^r = \frac{1}{6}\pi r^3$$

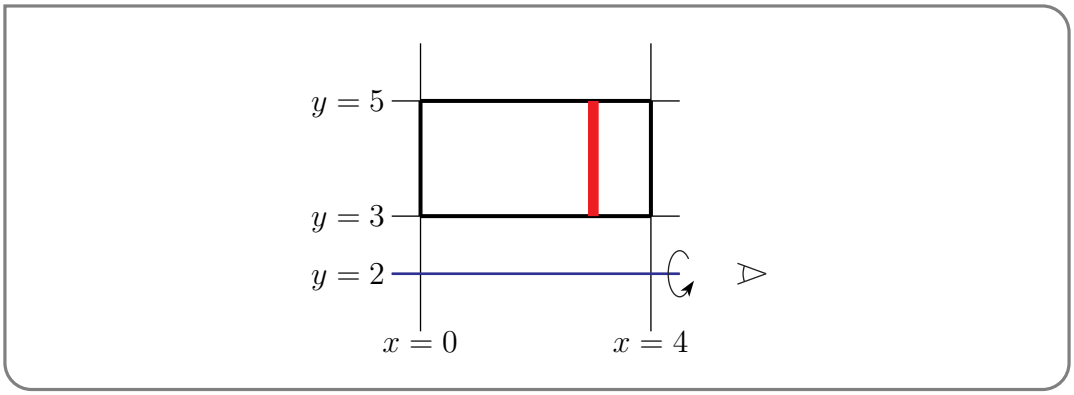
  and the total volume of the whole sphere is eight times that, which is $\frac{4}{3}\pi r^3$, as expected.

Example 1.6.2

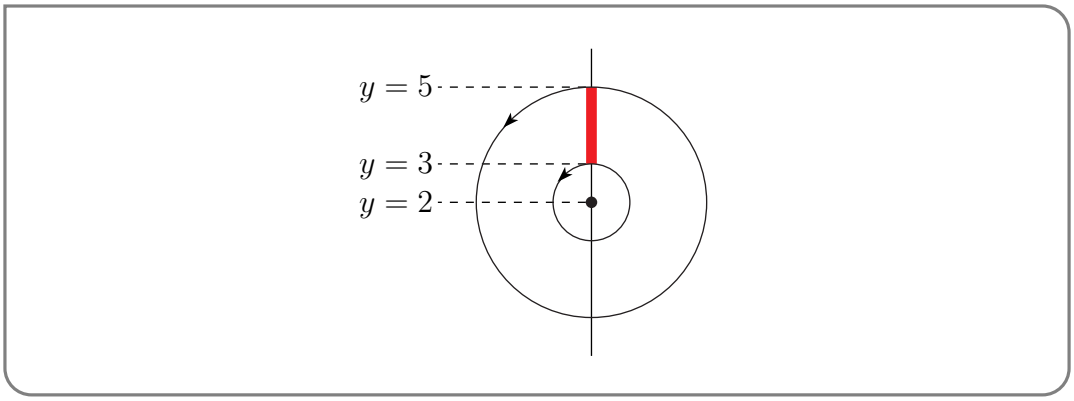
Example 1.6.3 (Revolving a region)

The region between the lines $y = 3$, $y = 5$, $x = 0$ and $x = 4$ is rotated around the line $y = 2$. Find the volume of the region swept out.

Solution. As with most of these problems, we should start by sketching the problem.



- Consider the region and slice it into thin vertical strips of width $\mathrm{d}x$.
- Now we are to rotate this region about the line $y = 2$. Imagine looking straight down the axis of rotation, $y = 2$, end on. The symbol in the figure above just to the right of the end the line $y = 2$ is supposed to represent your eye[43]. Here is what you see as the rotation takes place.



43 Okay okay... We missed the pupil. I'm sure there is a pun in there somewhere.

- Upon rotation about the line $y = 2$ our strip sweeps out a "washer"
  - whose cross–section is a disk of radius $5 - 2 = 3$ from which a disk of radius $3 - 2 = 1$ has been removed so that it has a
  - cross–sectional area of $\pi 3^2 - \pi 1^2 = 8\pi$ and a
  - thickness $\mathrm{d}x$ and hence a
  - volume $8\pi\,\mathrm{d}x$.
- As our leftmost strip is at $x = 0$ and our rightmost strip is at $x = 4$, the total

$$\text{Volume} = \int_0^4 8\pi\,\mathrm{d}x = (8\pi)(4) = 32\pi$$

Notice that we could also reach this answer by writing the volume as the difference of two cylinders.

- The outer cylinder has radius $(5 - 2)$ and length 4. This has volume

$$V_{outer} = \pi r^2 \ell = \pi \cdot 3^2 \cdot 4 = 36\pi.$$

- The inner cylinder has radius $(3 - 2)$ and length 4. This has volume

$$V_{inner} = \pi r^2 \ell = \pi \cdot 1^2 \cdot 4 = 4\pi.$$

- The volume we want is the difference of these two, namely

$$V = V_{outer} - V_{inner} = 32\pi.$$

Example 1.6.3

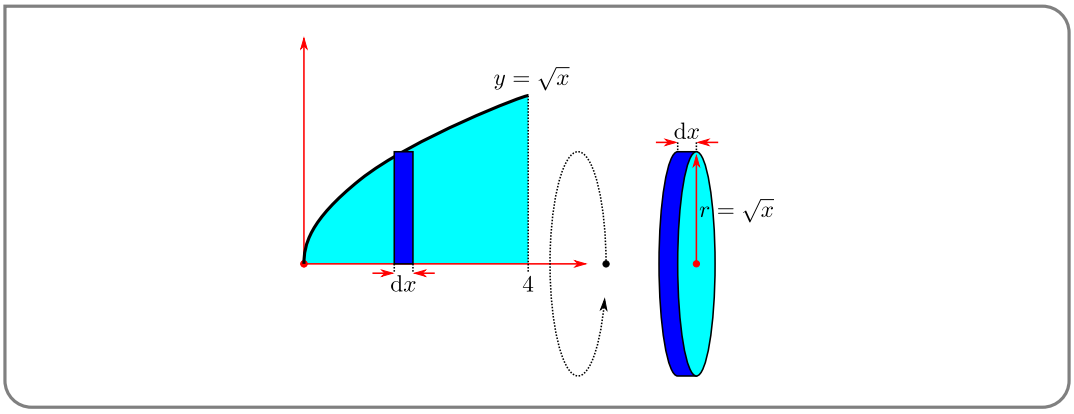
Let us turn up the difficulty a little on this last example.

Example 1.6.4 (Revolving again)

The region between the curve $y = \sqrt{x}$, and the lines $y = 0$, $x = 0$ and $x = 4$ is rotated around the line $y = 0$. Find the volume of the region swept out.

Solution. We can approach this in much the same way as the previous example.

- Consider the region and cut it into thin vertical strips of width $\mathrm{d}x$.



- When we rotate the region about the line $y = 0$, each strip sweeps out a thin pancake
  - whose cross-section is a disk of radius $\sqrt{x}$ with a
  - cross-sectional area of $\pi(\sqrt{x})^2 = \pi x$ and a
  - thickness $\mathrm{d}x$ and hence a
  - volume $\pi x \mathrm{d}x$.
- As our leftmost strip is at $x = 0$ and our rightmost strip is at $x = 4$, the total

$$\text{Volume} = \int_0^4 \pi x \mathrm{d}x = \left[\frac{\pi}{2}x^2\right]_0^4 = 8\pi$$

Example 1.6.4

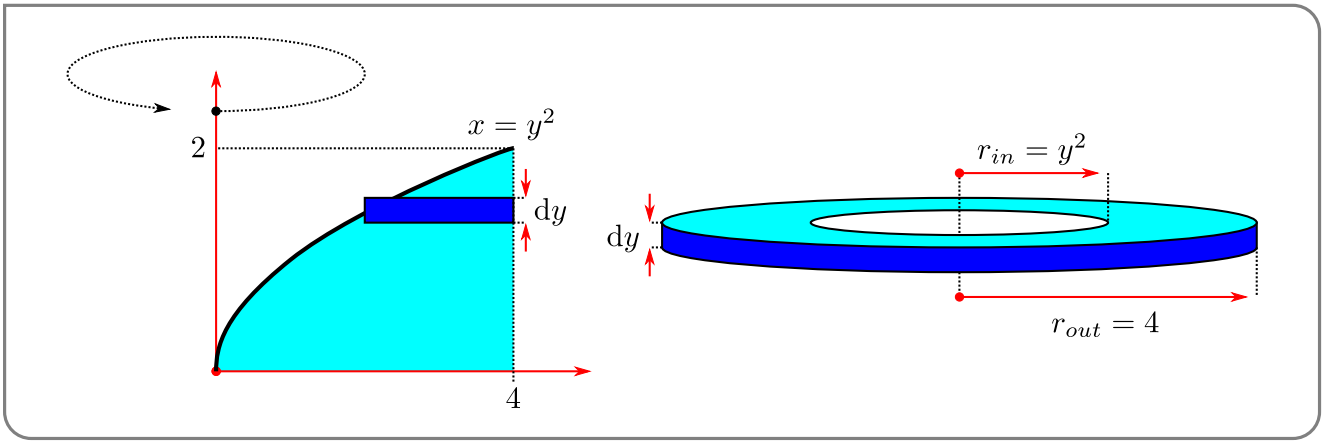
In the last example we considered rotating a region around the x-axis. Let us do the same but rotating around the y-axis.

Example 1.6.5 (Revolving yet again)

The region between the curve $y = \sqrt{x}$, and the lines $y = 0$, $x = 0$ and $x = 4$ is rotated around the line $x = 0$. Find the volume of the region swept out.

Solution.

- We will cut the region into horizontal slices, so we should write x as a function of y. That is, the region is bounded by $x = y^2$, $x = 4$, $y = 0$ and $y = 2$.
- Now slice the region into thin horizontal strips of width $\mathrm{d}y$.



- When we rotate the region about y-axis, each strip sweeps out a thin washer
  - whose inner radius is y^2 and outer radius is 4, and
  - thickness is $\mathrm{d}y$ and hence
  - has volume $\pi(r_{out}^2 - r_{in}^2)\mathrm{d}y = \pi(16 - y^4)\mathrm{d}y$.

- As our bottommost strip is at $y = 0$ and our topmost strip is at $y = 2$, the total

$$\text{Volume} = \int_0^2 \pi(16 - y^4)\mathrm{d}y = \left[16\pi y - \frac{\pi}{5}y^5\right]_0^2 = 32\pi - \frac{32\pi}{5} = \frac{128\pi}{5}.$$

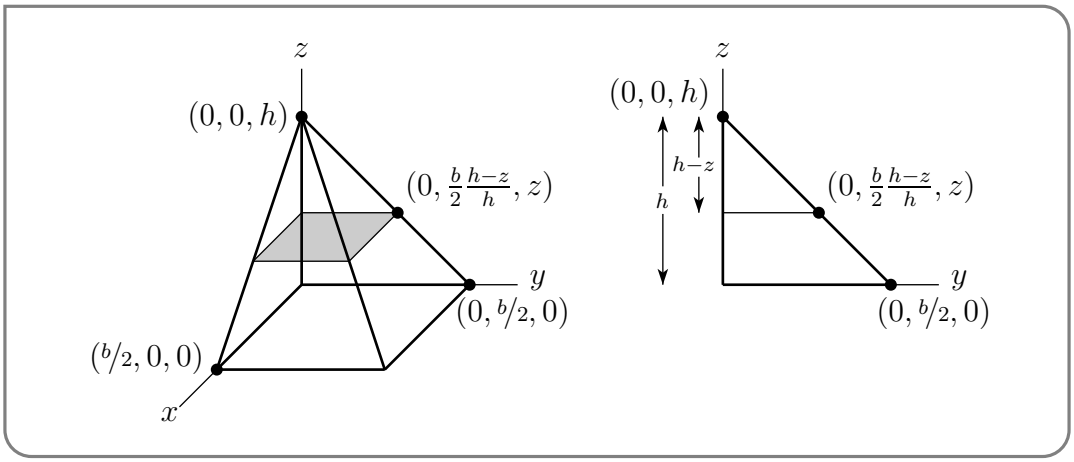
Example 1.6.5

There is another way[44] to do this one which we show at the end of this section.

Example 1.6.6 (Pyramid)

Find the volume of the pyramid which has height h and whose base is a square of side b.

Solution. Here is a sketch of the part of the pyramid that is in the first octant; we display only this portion to make the diagrams simpler. Note that this diagram shows only 1



quarter of the whole pyramid.

- To compute its volume, we slice it up into thin horizontal "square pancakes". A typical pancake also appears in the sketch above.
  - The pancake at height z is the fraction $\frac{h-z}{h}$ of the distance from the peak of the pyramid to its base.
  - So the *full* pancake[45] at height z is a square of side $\frac{h-z}{h}b$. As a check, note that when $z = h$ the pancake has side $\frac{h-h}{h}b = 0$, and when $z = 0$ the pancake has side $\frac{h-0}{h}b = b$.
  - So the pancake has cross-sectional area $\left(\frac{h-z}{h}b\right)^2$ and thickness[46] $\mathrm{d}z$ and hence
  - volume $\left(\frac{h-z}{h}b\right)^2\mathrm{d}z$.

44 The method is not a core part of the course and should be considered optional.
45 Note that this is the full pancake, not just the part in the first octant.
46 We are again using our Riemann sum avoiding trick.

- The volume of the whole pyramid (not just the part of the pyramid in the first octant) is

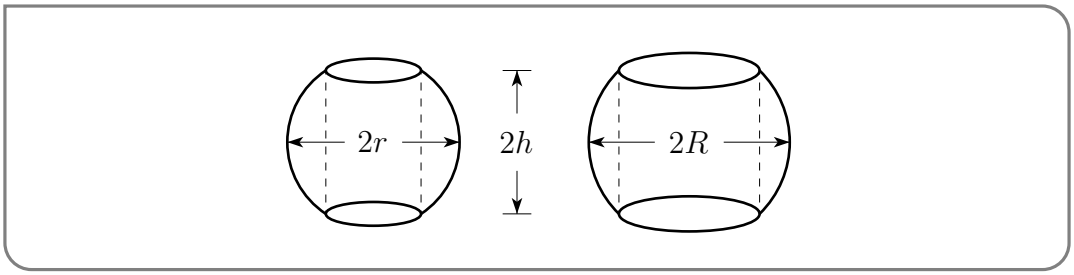
$$\begin{aligned}\int_0^h \left(\frac{h-z}{h}b\right)^2 \mathrm{d}z &= \frac{b^2}{h^2}\int_0^h (h-z)^2 \mathrm{d}z \\ &= \frac{b^2}{h^2}\int_h^0 -t^2 \mathrm{d}t \qquad \text{substitution rule with } t=(h-z), \mathrm{d}z \to -\mathrm{d}t \\ &= -\frac{b^2}{h^2}\left[\frac{t^3}{3}\right]_h^0 \\ &= -\frac{b^2}{h^2}\left[-\frac{h^3}{3}\right] \\ &= \frac{1}{3}b^2h\end{aligned}$$

Example 1.6.6

Let's ramp up the difficulty a little.

Example 1.6.7 (Napkin Ring)

Suppose you make two napkin rings[47] by drilling holes with different diameters through two wooden balls. One ball has radius r and the other radius R with $r < R$. You choose the diameter of the holes so that both napkin rings have the same height, $2h$. See the figure below.



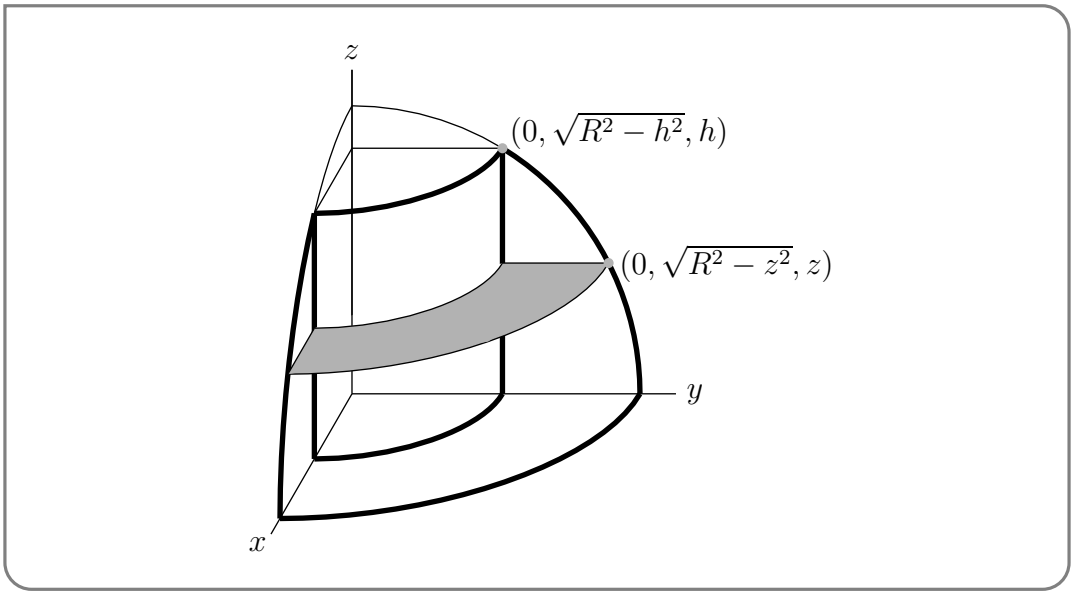
Which[48] ring has more wood in it?

Solution. We'll compute the volume of the napkin ring with radius R. We can then obtain the volume of the napkin ring of radius r, by just replacing $R \mapsto r$ in the result.

- To compute the volume of the napkin ring of radius R, we slice it up into thin horizontal "pancakes". Here is a sketch of the part of the napkin ring in the first octant showing a typical pancake.

47 Handy things to have (when combined with cloth napkins) if your parents are coming to dinner and you want to convince them that you are "taking care of yourself".

48 A good question to ask to distract your parents from the fact you are serving frozen burritos.



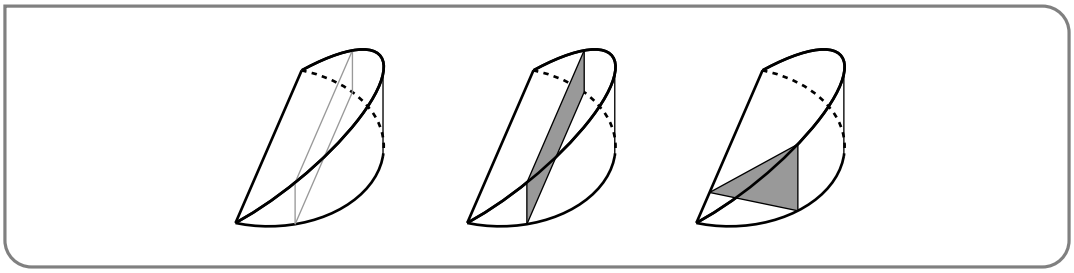
- The coordinates of the two points marked in the yz–plane of that figure are found by remembering that
  - the equation of the sphere is $x^2 + y^2 + z^2 = R^2$.
  - The two points have $y > 0$ and are in the yz–plane, so that $x = 0$ for them. So $y = \sqrt{R^2 - z^2}$.
  - In particular, at the top of the napkin ring $z = h$, so that $y = \sqrt{R^2 - h^2}$.
- The pancake at height z, shown in the sketch, is a "washer" — a circular disk with a circular hole cut in its center.
  - The outer radius of the washer is $\sqrt{R^2 - z^2}$ and
  - the inner radius of the washer is $\sqrt{R^2 - h^2}$. So the
  - cross–sectional area of the washer is
$$\pi\big(\sqrt{R^2 - z^2}\,\big)^2 - \pi\big(\sqrt{R^2 - h^2}\,\big)^2 = \pi(h^2 - z^2)$$
- The pancake at height z
  - has thickness dz and
  - cross–sectional area $\pi(h^2 - z^2)$ and hence
  - volume $\pi(h^2 - z^2)\mathrm{d}z$.
- Since z runs from $-h$ to $+h$, the total volume of wood in the napkin ring of radius R is
$$\begin{aligned}\int_{-h}^{h} \pi(h^2 - z^2)\mathrm{d}z &= \pi\Big[h^2 z - \frac{z^3}{3}\Big]_{-h}^{h} \\ &= \pi\Big[\Big(h^3 - \frac{h^3}{3}\Big) - \Big((-h)^3 - \frac{(-h)^3}{3}\Big)\Big] \\ &= \pi\Big[\frac{2}{3}h^3 - \frac{2}{3}\big(-h\big)^3\Big] \\ &= \frac{4\pi}{3}h^3\end{aligned}$$

This volume is independent of R. Hence the napkin ring of radius r contains precisely the same volume of wood as the napkin ring of radius R!

Example 1.6.7

Example 1.6.8 (Notch)

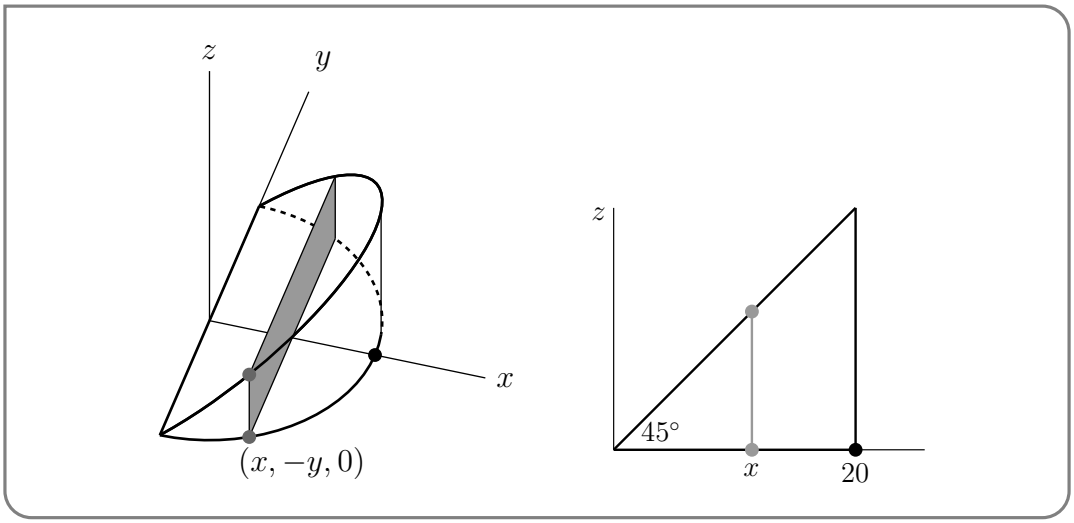
A 45° notch is cut to the centre of a cylindrical log having radius 20cm. One plane face of the notch is perpendicular to the axis of the log. See the sketch below. What volume of wood was removed?



Solution. We show two solutions to this problem which are of comparable difficulty. The difference lies in the shape of the pancakes we use to slice up the volume. In solution 1 we cut rectangular pancakes parallel to the yz–plane and in solution 2 we slice triangular pancakes parallel to the xz–plane.

Solution 1:

- Concentrate on the notch. Rotate it around so that the plane face lies in the xy–plane.
- Then slice the notch into vertical rectangles (parallel to the yz–plane) as in the figure on the left below.



- The cylindrical log had radius 20cm. So the circular part of the boundary of the base of the notch has equation $x^2 + y^2 = 20^2$. (We're putting the origin of the xy–plane at the centre of the circle.) If our coordinate system is such that x is constant on each slice, then

  - the base of the slice is the line segment from $(x, -y, 0)$ to $(x, +y, 0)$ where $y = \sqrt{20^2 - x^2}$ so that
  - the slice has width $2y = 2\sqrt{20^2 - x^2}$ and
  - height x (since the upper face of the notch is at 45° to the base — see the side view sketched in the figure on the right above).
  - So the slice has cross–sectional area $2x\sqrt{20^2 - x^2}$.

- On the base of the notch x runs from 0 to 20 so the volume of the notch is

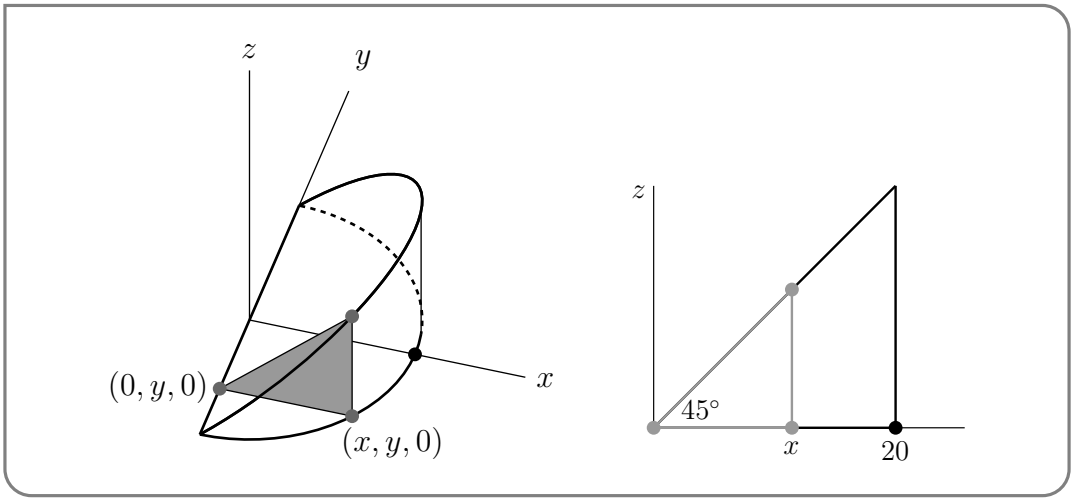
$$V = \int_0^{20} 2x\sqrt{20^2 - x^2}\,\mathrm{d}x$$

Make the change of variables $u = 20^2 - x^2$ (don't forget to change $\mathrm{d}x \to -\frac{1}{2x}\mathrm{d}u$):

$$\begin{aligned} V &= \int_{20^2}^{0} -\sqrt{u}\,\mathrm{d}u \\ &= \left[-\frac{u^{3/2}}{3/2}\right]_{20^2}^{0} \\ &= \frac{2}{3}20^3 = \frac{16,000}{3} \end{aligned}$$

Solution 2:

- Concentrate of the notch. Rotate it around so that its base lies in the xy–plane with the skinny edge along the y–axis.
- Slice the notch into triangles parallel to the xz–plane as in the figure on the left below. In the figure below, the triangle happens to lie in a plane where y is negative.



- The cylindrical log had radius 20cm. So the circular part of the boundary of the base of the notch has equation $x^2 + y^2 = 20^2$. Our coordinate system is such that y is constant on each slice, so that

  - the base of the triangle is the line segment from $(0, y, 0)$ to $(x, y, 0)$ where $x = \sqrt{20^2 - y^2}$ so that
  - the triangle has base $x = \sqrt{20^2 - y^2}$ and
  - height $x = \sqrt{20^2 - y^2}$ (since the upper face of the notch is at 45° to the base — see the side view sketched in the figure on the right above).
  - So the slice has cross–sectional area $\frac{1}{2}\left(\sqrt{20^2 - y^2}\right)^2$.

- On the base of the notch y runs from -20 to 20, so the volume of the notch is

$$\begin{aligned} V &= \tfrac{1}{2}\int_{-20}^{20} (20^2 - y^2)\,\mathrm{d}y \\ &= \int_{0}^{20} (20^2 - y^2)\,\mathrm{d}y \\ &= \left[20^2 y - \frac{y^3}{3}\right]_0^{20} \\ &= \frac{2}{3}20^3 = \frac{16,000}{3} \end{aligned}$$

Example 1.6.8

▸▸ Optional — Cylindrical Shells

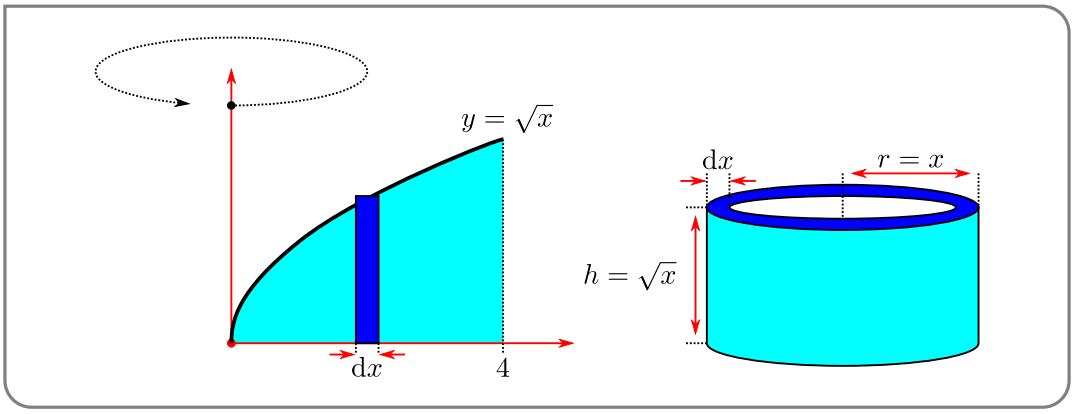
Let us return to Example 1.6.5 in which we rotate a region around the y-axis. Here we show another solution to this problem which is obtained by slicing the region into vertical strips. When rotated about the y-axis, each such strip sweeps out a thin cylindrical shell. Hence the name of this approach (and this subsection).

Example 1.6.9 (Revolving yet again)

The region between the curve $y = \sqrt{x}$, and the lines $y = 0$, $x = 0$ and $x = 4$ is rotated around the line $x = 0$. Find the volume of the region swept out.

Solution.

- Consider the region and cut it into thin vertical strips of width $\mathrm{d}x$.



- When we rotate the region about the y-axis, each strip sweeps out a thin cylindrical shell
  - whose radius is x,
  - height is $\sqrt{x}$, and
  - thickness is $\mathrm{d}x$ and hence
  - has volume $2\pi \times$ radius $\times$ height $\times$ thickness $= 2\pi x^{3/2}\mathrm{d}x$.
- As our leftmost strip is at $x = 0$ and our rightmost strip is at $x = 4$, the total

$$\text{Volume} = \int_0^4 2\pi x^{3/2}\mathrm{d}x = \left[\frac{4\pi}{5}x^{5/2}\right]_0^4 = \frac{4\pi}{5}\cdot 32 = \frac{128\pi}{5}$$

  which (thankfully) agrees with our previous computation.

Example 1.6.9

1.7▲ Integration by Parts

The fundamental theorem of calculus tells us that it is very easy to integrate a derivative. In particular, we know that

$$\int \frac{\mathrm{d}}{\mathrm{d}x}\left(F(x)\right)\mathrm{d}x = F(x) + C$$

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

$$\int xe^x\mathrm{d}x$$

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

$$\frac{\mathrm{d}}{\mathrm{d}x}u(x)v(x) = u'(x)\,v(x) + u(x)\,v'(x)$$

Integrating this gives

$$\begin{aligned}\int \left[u'(x)\,v(x) + u(x)\,v'(x)\right]\mathrm{d}x &= \left[\text{a function whose derivative is } u'v + uv'\right] + C\\ &= u(x)v(x) + C\end{aligned}$$

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:

$$\int \left[u'(x)\,v(x) + u(x)\,v'(x)\right]\mathrm{d}x = \int u'(x)\,v(x)\,\mathrm{d}x + \int u(x)\,v'(x)\,\mathrm{d}x$$

Bring one of the integrals to the left-hand side

$$u(x)v(x) - \int u'(x)\,v(x)\mathrm{d}x = \int u(x)\,v'(x)\mathrm{d}x$$

Swap left and right sides

$$\int u(x)\,v'(x)\mathrm{d}x = u(x)v(x) - \int u'(x)\,v(x)\mathrm{d}x$$

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 1.7.1 ($\int xe^x\mathrm{d}x$)

Compute the integral $\int xe^x\mathrm{d}x$.

Solution.

- We start by taking the equation above

$$\int u(x)\,v'(x)\mathrm{d}x = u(x)v(x) - \int u'(x)\,v(x)\mathrm{d}x$$

- 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:

$$\int xe^x\mathrm{d}x = xe^x - \int e^x\mathrm{d}x$$

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:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(xe^x - e^x + C\right) = \underbrace{xe^x + 1\cdot e^x}_{\text{by product rule}} - e^x + 0$$
$$= xe^x \qquad \text{as required.}$$

Example 1.7.1

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 1.7.2 (Integration by parts)**.**

Let $u(x)$ and $v(x)$ be continuously differentiable. Then

$$\int u(x)\, v'(x)\mathrm{d}x = u(x)\, v(x) - \int v(x)\, u'(x)\mathrm{d}x$$

If we write $\mathrm{d}v$ for $v'(x)\mathrm{d}x$ and $\mathrm{d}u$ for $u'(x)\mathrm{d}x$ (as the substitution rule suggests), then the formula becomes

$$\int u\mathrm{d}v = u\, v - \int v\mathrm{d}u$$

The application of this formula is known as integration by parts.
The corresponding statement for definite integrals is

$$\int_a^b u(x)\, v'(x)\mathrm{d}x = u(b)\, v(b) - u(a)\, v(a) - \int_a^b v(x)\, u'(x)\mathrm{d}x$$

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 $\int u'(x)v(x)\mathrm{d}x$ 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

$$\begin{aligned}\int u(x)v'(x)\mathrm{d}x &= u(x)\,(v(x)+A) - \int u'(x)\,(v(x)+A)\,\mathrm{d}x \\ &= u(x)v(x) + Au(x) - \int u'(x)v(x)\mathrm{d}x - \underbrace{A\int u'(x)\mathrm{d}x}_{=Au(x)+C} \\ &= u(x)v(x) - \int u'(x)v(x)\mathrm{d}x + C\end{aligned}$$

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 1.7.3 ($\int xe^x \mathrm{d}x$ — 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':

$$u(x) = x \qquad v'(x) = e^x$$
$$\text{or}$$
$$u(x) = e^x \qquad v'(x) = x$$

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)

   $$\int u'(x)v(x)\mathrm{d}x = \int 1 \cdot e^x \mathrm{d}x$$

   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

   $$\int u'(x)v(x)\mathrm{d}x = \int \frac{1}{2}x^2 \cdot e^x \mathrm{d}x.$$

   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

$$\begin{aligned}\int xe^x \mathrm{d}x &= xe^x - \int e^x \mathrm{d}x \\ &= xe^x - e^x + C.\end{aligned}$$

The above reasoning is a very typical workflow when using integration by parts.

Example 1.7.3

Integration by parts is often used

- to eliminate factors of x from an integrand like xe^x by using that $\frac{\mathrm{d}}{\mathrm{d}x}x = 1$ and
- to eliminate a $\log x$ from an integrand by using that $\frac{\mathrm{d}}{\mathrm{d}x}\log x = \frac{1}{x}$ and
- to eliminate inverse trig functions, like $\arctan x$, from an integrand by using that, for example, $\frac{\mathrm{d}}{\mathrm{d}x}\arctan x = \frac{1}{1+x^2}$.

Example 1.7.4 ($\int x\sin x\mathrm{d}x$)

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

  $$u'(x) = 1 \qquad \text{and} \qquad v(x) = -\cos x$$

  which is looking promising.

  (2) On the other hand if we choose $u(x) = \sin x$ and $v'(x) = x$, then we have

  $$u'(x) = \cos x \qquad \text{and} \qquad v(x) = \frac{1}{2}x^2$$

  which is looking worse — we'd need to integrate $\int \frac{1}{2}x^2\cos x\mathrm{d}x$.

- So we stick with the first choice. Plugging $u(x) = x$, $v(x) = -\cos x$ into integration by parts gives us

$$\begin{aligned}\int x\sin x\mathrm{d}x &= -x\cos x - \int 1\cdot(-\cos x)\mathrm{d}x\\ &= -x\cos x + \sin x + C\end{aligned}$$

- Again we can check our answer by differentiating:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(-x\cos x + \sin x + C\right) &= -\cos x + x\sin x + \cos x + 0\\ &= x\sin x\checkmark\end{aligned}$$

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

$$\begin{aligned}\int x\sin x\mathrm{d}x &= -x\cos x + \int \cos x\mathrm{d}x\\ &= -x\cos x + \sin x + C.\end{aligned}$$

Example 1.7.4

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

$$\int u\mathrm{d}v = uv - \int v\mathrm{d}u$$

where $\mathrm{d}u, \mathrm{d}v$ are shorthand for $u'(x)\mathrm{d}x, v'(x)\mathrm{d}x$. Let us write up the previous example using this notation.

Example 1.7.5 ($\int x \sin x \mathrm{d}x$ yet again)

Solution. Using integration by parts, we set $u = x$ and $\mathrm{d}v = \sin x \mathrm{d}x$. This makes $\mathrm{d}u = 1\mathrm{d}x$ and $v = -\cos x$. Consequently

$$\begin{aligned}\int x \sin x \mathrm{d}x &= \int u\mathrm{d}v \\ &= uv - \int v\mathrm{d}u \\ &= -x\cos x + \int \cos x \mathrm{d}x \\ &= -x\cos x + \sin x + C\end{aligned}$$

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 1.7.5

We can also use integration by parts to eliminate higher powers of x. We just need to apply the method more than once.

Example 1.7.6 ($\int x^2 e^x \mathrm{d}x$)

Solution.

- Let $u = x^2$ and $\mathrm{d}v = e^x\mathrm{d}x$. This then gives $\mathrm{d}u = 2x\mathrm{d}x$ and $v = e^x$, and

$$\int x^2 e^x \mathrm{d}x = x^2 e^x - \int 2xe^x \mathrm{d}x$$

- So we have reduced the problem of computing the original integral to one of integrating $2xe^x$. We know how to do this — just integrate by parts again:

$$\begin{aligned}\int x^2 e^x \mathrm{d}x &= x^2 e^x - \int 2xe^x \mathrm{d}x && \text{set } u = 2x, \mathrm{d}v = e^x\mathrm{d}x \\ &= x^2 e^x - \left(2xe^x - \int 2e^x \mathrm{d}x\right) && \text{since } \mathrm{d}u = 2\mathrm{d}x, v = e^x \\ &= x^2 e^x - 2xe^x + 2e^x + C\end{aligned}$$

- We can, if needed, check our answer by differentiating:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^2e^x - 2xe^x + 2e^x + C\right) &= \left(x^2e^x + 2xe^x\right) - (2xe^x + 2e^x) + 2e^x + 0\\ &= x^2e^x\checkmark\end{aligned}$$

A similar iterated application of integration by parts will work for integrals

$$\int P(x)\left(Ae^{ax} + B\sin(bx) + C\cos(cx)\right)\mathrm{d}x$$

where $P(x)$ is a polynomial and A, B, C, a, b, c are constants.

Example 1.7.6

Now let us look at integrands containing logarithms. We don't know the antiderivative of $\log x$, but we can eliminate $\log x$ from an integrand by using integration by parts with $u = \log x$. Remember $\log x = \log_e x = \ln x$.

Example 1.7.7 ($\int x \log x \mathrm{d}x$)

Solution.

- We have two choices for u and $\mathrm{d}v$.
  (1) Set $u = x$ and $\mathrm{d}v = \log x \mathrm{d}x$. This gives $\mathrm{d}u = \mathrm{d}x$ but v is hard to compute — we haven't done it yet[49]. Before we go further along this path, we should look to see what happens with the other choice.
  (2) Set $u = \log x$ and $\mathrm{d}v = x\mathrm{d}x$. This gives $\mathrm{d}u = \frac{1}{x}\mathrm{d}x$ and $v = \frac{1}{2}x^2$, and we have to integrate

$$\int v\,\mathrm{d}u = \int \frac{1}{x}\cdot\frac{1}{2}x^2\mathrm{d}x$$

  which is easy.

- So we proceed with the second choice.

$$\begin{aligned}\int x\log x\mathrm{d}x &= \frac{1}{2}x^2\log x - \int\frac{1}{2}x\mathrm{d}x\\ &= \frac{1}{2}x^2\log x - \frac{1}{4}x^2 + C\end{aligned}$$

- We can check our answer quickly:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^2}{2}\ln x - \frac{x^2}{4} + C\right) = x\,\ln x + \frac{x^2}{2}\frac{1}{x} - \frac{x}{2} + 0 = x\,\ln x$$

49 We will soon.

Example 1.7.7

Example 1.7.8 ($\int \log x \mathrm{d}x$)

It is not immediately obvious that one should use integration by parts to compute the integral

$$\int \log x \mathrm{d}x$$

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 $\log x$ by setting $u = \log x$ and using integration by parts. Let us try repeating this. When we make this choice, we are then forced to take $\mathrm{d}v = \mathrm{d}x$ — that is we choose $v'(x) = 1$. Once we have made this sneaky move everything follows quite directly.
- We then have $\mathrm{d}u = \frac{1}{x}\mathrm{d}x$ and $v = x$, and the integration by parts formula gives us

$$\begin{aligned}\int \log x \mathrm{d}x &= x \log x - \int \frac{1}{x} \cdot x \mathrm{d}x \\ &= x \log x - \int 1 \mathrm{d}x \\ &= x \log x - x + C\end{aligned}$$

- As always, it is a good idea to check our result by verifying that the derivative of the answer really is the integrand.

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x \ln x - x + C\right) = \ln x + x\,\frac{1}{x} - 1 + 0 = \ln x$$

Example 1.7.8

The same method works almost exactly to compute the antiderivatives of $\arcsin(x)$ and $\arctan(x)$:

Example 1.7.9 ($\int \arctan(x)\mathrm{d}x$ and $\int \arcsin(x)\mathrm{d}x$)

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 $\mathrm{d}v = \mathrm{d}x$ (ie $v'(x) = 1$) and choose $v(x) = x$.

- For inverse tan we choose $u = \arctan(x)$, so $\mathrm{d}u = \frac{1}{1+x^2}\mathrm{d}x$:

$$\begin{aligned}
\int \arctan(x)\mathrm{d}x &= x\arctan(x) - \int x \cdot \frac{1}{1+x^2}\mathrm{d}x && \text{now use substitution rule}\\
&= x\arctan(x) - \int \frac{w'(x)}{2}\cdot\frac{1}{w}\mathrm{d}x && \text{with } w(x) = 1+x^2, w'(x) = 2x\\
&= x\arctan(x) - \frac{1}{2}\int \frac{1}{w}\mathrm{d}w\\
&= x\arctan(x) - \frac{1}{2}\log|w| + C\\
&= x\arctan(x) - \frac{1}{2}\log|1+x^2| + C && \text{but } 1+x^2 > 0, \text{ so}\\
&= x\arctan(x) - \frac{1}{2}\log(1+x^2) + C
\end{aligned}$$

- Similarly for inverse sine we choose $u = \arcsin(x)$ so $\mathrm{d}u = \frac{1}{\sqrt{1-x^2}}\mathrm{d}x$:

$$\begin{aligned}
\int \arcsin(x)\mathrm{d}x &= x\arcsin(x) - \int \frac{x}{\sqrt{1-x^2}}\mathrm{d}x && \text{now use substitution rule}\\
&= x\arcsin(x) - \int \frac{-w'(x)}{2}\cdot w^{-1/2}\mathrm{d}x && \text{with } w(x) = 1-x^2, w'(x) = -2x\\
&= x\arcsin(x) + \frac{1}{2}\int w^{-1/2}\mathrm{d}w\\
&= x\arcsin(x) + \frac{1}{2}\cdot 2w^{1/2} + C\\
&= x\arcsin(x) + \sqrt{1-x^2} + C
\end{aligned}$$

- Both can be checked quite quickly by differentiating — but we leave that as an exercise for the reader.

Example 1.7.9

There are many other examples we could do, but we'll finish with a tricky one.

Example 1.7.10 ($\int e^x \sin x\mathrm{d}x$)

Solution. Let us attempt this one a little naively and then we'll come back and do it more carefully (and successfully).

- We can choose either $u = e^x, \mathrm{d}v = \sin x\mathrm{d}x$ or the other way around.

  1. Let $u = e^x, \mathrm{d}v = \sin x\mathrm{d}x$. Then $\mathrm{d}u = e^x\mathrm{d}x$ and $v = -\cos x$. This gives

$$\int e^x \sin x = -e^x\cos x + \int e^x\cos x\mathrm{d}x$$

  So we are left with an integrand that is very similar to the one we started with. What about the other choice?

2. Let $u = \sin x, \mathrm{d}v = e^x \mathrm{d}x$. Then $\mathrm{d}u = \cos x \mathrm{d}x$ and $v = e^x$. This gives

$$\int e^x \sin x = e^x \sin x - \int e^x \cos x \mathrm{d}x$$

So we are again left with an integrand that is very similar to the one we started with.

- How do we proceed? — It turns out to be easier if you do both $\int e^x \sin x \mathrm{d}x$ and $\int e^x \cos x \mathrm{d}x$ simultaneously. We do so in the next example.

Example 1.7.10

Example 1.7.11 $\left(\int_a^b e^x \sin x \mathrm{d}x \text{ and } \int_a^b e^x \cos x \mathrm{d}x\right)$

This time we're going to do the two integrals

$$I_1 = \int_a^b e^x \sin x \mathrm{d}x \qquad I_2 = \int_a^b e^x \cos x \mathrm{d}x$$

at more or less the same time.

- First

$$I_1 = \int_a^b e^x \sin x \mathrm{d}x = \int_a^b u \mathrm{d}v \qquad \text{with } u = e^x, \mathrm{d}v = \sin x \mathrm{d}x$$
$$\text{so } v = -\cos x, \mathrm{d}u = e^x \mathrm{d}x$$
$$= \Big[-e^x \cos x \Big]_a^b + \int_a^b e^x \cos x \mathrm{d}x$$

We have not found I_1 but we have related it to I_2.

$$I_1 = \Big[-e^x \cos x \Big]_a^b + I_2$$

- Now start over with I_2.

$$I_2 = \int_a^b e^x \cos x \mathrm{d}x = \int_a^b u \mathrm{d}v \qquad \text{with } u = e^x, \mathrm{d}v = \cos x \mathrm{d}x$$
$$\text{so } v = \sin x, \mathrm{d}u = e^x \mathrm{d}x$$
$$= \Big[e^x \sin x \Big]_a^b - \int_a^b e^x \sin x \mathrm{d}x$$

Once again, we have not found I_2 but we have related it back to I_1.

$$I_2 = \Big[e^x \sin x \Big]_a^b - I_1$$

- So summarising, we have

$$I_1 = \Big[-e^x \cos x \Big]_a^b + I_2 \qquad\qquad I_2 = \Big[e^x \sin x \Big]_a^b - I_1$$

- So now, substitute the expression for I_2 from the second equation into the first equation to get

$$I_1 = \Big[-e^x \cos x + e^x \sin x \Big]_a^b - I_1 \quad \text{which implies} \quad I_1 = \frac{1}{2}\Big[e^x \big(\sin x - \cos x \big) \Big]_a^b$$

If we substitute the other way around we get

$$I_2 = \Big[e^x \sin x + e^x \cos x \Big]_a^b - I_2 \quad \text{which implies} \quad I_2 = \frac{1}{2}\Big[e^x \big(\sin x + \cos x \big) \Big]_a^b$$

That is,

$$\int_a^b e^x \sin x \mathrm{d}x = \frac{1}{2}\Big[e^x \big(\sin x - \cos x \big) \Big]_a^b \qquad \int_a^b e^x \cos x \mathrm{d}x = \frac{1}{2}\Big[e^x \big(\sin x + \cos x \big) \Big]_a^b$$

- This also says, for example, that $\frac{1}{2}e^x \big(\sin x - \cos x \big)$ is an antiderivative of $e^x \sin x$ so that

$$\int e^x \sin x \mathrm{d}x = \frac{1}{2} e^x \big(\sin x - \cos x \big) + C$$

- Note that we can always check whether or not this is correct. It is correct if and only if the derivative of the right hand side is $e^x \sin x$. Here goes. By the product rule

$$\frac{\mathrm{d}}{\mathrm{d}x}\Big[\frac{1}{2} e^x \big(\sin x - \cos x \big) + C \Big] = \frac{1}{2}\Big[e^x \big(\sin x - \cos x \big) + e^x \big(\cos x + \sin x \big) \Big] = e^x \sin x$$

which is the desired derivative.

- There is another way to find $\int e^x \sin x \mathrm{d}x$ and $\int e^x \cos x \mathrm{d}x$ that, in contrast to the above computations, doesn't involve any trickery. But it does require the use of complex numbers and so is beyond the scope of this course. The secret is to use that $\sin x = \frac{e^{ix} - e^{-ix}}{2i}$ and $\cos x = \frac{e^{ix} + e^{-ix}}{2}$, where i is the square root of -1 of the complex number system. See Example B.2.6.

Example 1.7.11

1.8▲ Trigonometric Integrals

Integrals of polynomials of the trigonometric functions $\sin x$, $\cos x$, $\tan x$ and so on, are generally evaluated by using a combination of simple substitutions and trigonometric identities. There are of course a very large number[50] of trigonometric identities, but usually we use only a handful of them. The most important three are:

50 The more pedantic reader could construct an infinite list of them.

Equation 1.8.1.

$$\sin^2 x + \cos^2 x = 1$$

Equation 1.8.2.

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

Equation 1.8.3.

$$\begin{aligned}\cos(2x) &= \cos^2 x - \sin^2 x \\ &= 2\cos^2 x - 1 \\ &= 1 - 2\sin^2 x\end{aligned}$$

Notice that the last two lines of Equation (1.8.3) follow from the first line by replacing either $\sin^2 x$ or $\cos^2 x$ using Equation (1.8.1). It is also useful to rewrite these last two lines:

Equation 1.8.4.

$$\sin^2 x = \frac{1 - \cos(2x)}{2}$$

Equation 1.8.5.

$$\cos^2 x = \frac{1 + \cos(2x)}{2}$$

These last two are particularly useful since they allow us to rewrite higher powers of sine and cosine in terms of lower powers. For example:

$$\begin{aligned}\sin^4(x) &= \left[\frac{1-\cos(2x)}{2}\right]^2 && \text{by Equation (1.8.4)} \\ &= \frac{1}{4} - \frac{1}{2}\cos(2x) + \frac{1}{4}\underbrace{\cos^2(2x)}_{\text{do it again}} && \text{use Equation (1.8.5)} \\ &= \frac{1}{4} - \frac{1}{2}\cos(2x) + \frac{1}{8}(1 + \cos(4x)) \\ &= \frac{3}{8} - \frac{1}{2}\cos(2x) + \frac{1}{8}\cos(4x)\end{aligned}$$

So while it was hard to integrate $\sin^4(x)$ directly, the final expression is quite straightforward (with a little substitution rule).

There are many such tricks for integrating powers of trigonometric functions. Here we concentrate on two families

$$\int \sin^m x \cos^n x \mathrm{d}x \qquad \text{and} \qquad \int \tan^m x \sec^n x \mathrm{d}x$$

for integer n, m. The details of the technique depend on the parity of n and m — that is, whether n and m are even or odd numbers.

1.8.1 ▸▸ Integrating $\int \sin^m x \cos^n x \mathrm{d}x$

▸▸▸ One of n and m is Odd

Consider the integral $\int \sin^2 x \cos x \mathrm{d}x$. We can integrate this by substituting $u = \sin x$ and $\mathrm{d}u = \cos x \mathrm{d}x$. This gives

$$\begin{aligned}\int \sin^2 x \cos x \mathrm{d}x &= \int u^2 \mathrm{d}u \\ &= \frac{1}{3}u^3 + C = \frac{1}{3}\sin^3 x + C\end{aligned}$$

This method can be used whenever n is an odd integer.

- Substitute $u = \sin x$ and $\mathrm{d}u = \cos x \mathrm{d}x$.
- This leaves an even power of cosines — convert them using $\cos^2 x = 1 - \sin^2 x = 1 - u^2$.

Here is an example.

Example 1.8.6 ($\int \sin^2 x \cos^3 x \mathrm{d}x$)

Start by factoring off one power of $\cos x$ to combine with $\mathrm{d}x$ to get $\cos x \mathrm{d}x = \mathrm{d}u$.

$$\begin{aligned}\int \sin^2 x \cos^3 x \mathrm{d}x &= \int \underbrace{\sin^2 x}_{=u^2} \underbrace{\cos^2 x}_{=1-u^2} \underbrace{\cos x \mathrm{d}x}_{=\mathrm{d}u} && \text{set } u = \sin x \\ &= \int u^2 \,(1-u^2)\mathrm{d}u \\ &= \frac{u^3}{3} - \frac{u^5}{5} + C \\ &= \frac{\sin^3 x}{3} - \frac{\sin^5 x}{5} + C\end{aligned}$$

Example 1.8.6

Of course if m is an odd integer we can use the same strategy with the roles of $\sin x$ and $\cos x$ exchanged. That is, we substitute $u = \cos x$, $\mathrm{d}u = -\sin x \mathrm{d}x$ and $\sin^2 x = 1 - \cos^2 x = 1 - u^2$.

▸▸▸ Both n and m are Even

If m and n are both even, the strategy is to use the trig identities (1.8.4) and (1.8.5) to get back to the m or n odd case. This is typically more laborious than the previous case we studied. Here are a couple of examples that arise quite commonly in applications.

Example 1.8.7 $\left(\int \cos^2 x \mathrm{d}x\right)$

By (1.8.5)

$$\int \cos^2 x \mathrm{d}x = \frac{1}{2}\int \big[1+\cos(2x)\big]\mathrm{d}x = \frac{1}{2}\Big[x+\frac{1}{2}\sin(2x)\Big]+C$$

Example 1.8.7

Example 1.8.8 $\left(\int \cos^4 x \mathrm{d}x\right)$

First we'll prepare the integrand $\cos^4 x$ for easy integration by applying (1.8.5) a couple times. We have already used (1.8.5) once to get

$$\cos^2 x = \frac{1}{2}\big[1+\cos(2x)\big]$$

Squaring it gives

$$\cos^4 x = \frac{1}{4}\big[1+\cos(2x)\big]^2 = \frac{1}{4}+\frac{1}{2}\cos(2x)+\frac{1}{4}\cos^2(2x)$$

Now by (1.8.5) a second time

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

Now it's easy to integrate

$$\begin{aligned}\int \cos^4 x \mathrm{d}x &= \frac{3}{8}\int \mathrm{d}x+\frac{1}{2}\int \cos(2x)\mathrm{d}x+\frac{1}{8}\int \cos(4x)\mathrm{d}x\\ &= \frac{3}{8}x+\frac{1}{4}\sin(2x)+\frac{1}{32}\sin(4x)+C\end{aligned}$$

Example 1.8.8

Example 1.8.9 $\left(\int \cos^2 x \sin^2 x \mathrm{d}x\right)$

Here we apply both (1.8.4) and (1.8.5).

$$\begin{aligned}\int \cos^2 x \sin^2 x \mathrm{d}x &= \frac{1}{4}\int \big[1+\cos(2x)\big]\big[1-\cos(2x)\big]\mathrm{d}x\\ &= \frac{1}{4}\int \big[1-\cos^2(2x)\big]\mathrm{d}x\end{aligned}$$

We can then apply (1.8.5) again

$$\begin{aligned}
&= \frac{1}{4}\int \big[1 - \frac{1}{2}\left(1+\cos(4x)\right)\big]\,\mathrm{d}x \\
&= \frac{1}{8}\int \big[1-\cos(4x)\big]\,\mathrm{d}x \\
&= \frac{1}{8}x - \frac{1}{32}\sin(4x) + C
\end{aligned}$$

Oof! We could also have done this one using (1.8.2) to write the integrand as $\sin^2(2x)$ and then used (1.8.4) to write it in terms of $\cos(4x)$.

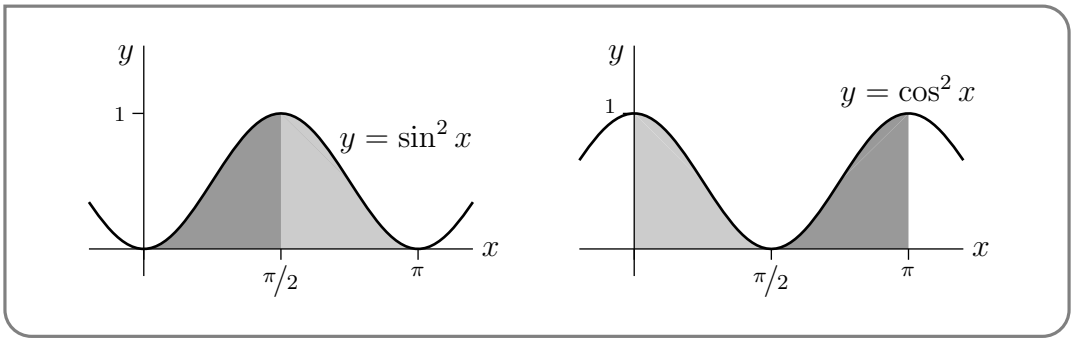
Example 1.8.9

Example 1.8.10 ($\int_0^\pi \cos^2 x\mathrm{d}x$ and $\int_0^\pi \sin^2 x\mathrm{d}x$)

Of course we can compute the definite integral $\int_0^\pi \cos^2 x\mathrm{d}x$ by using the antiderivative for $\cos^2 x$ that we found in Example 1.8.7. But here is a trickier way to evaluate that integral, and also the integral $\int_0^\pi \sin^2 x\mathrm{d}x$ at the same time, very quickly without needing the antiderivative of Example 1.8.7.

Solution.

- Observe that $\int_0^\pi \cos^2 x\mathrm{d}x$ and $\int_0^\pi \sin^2 x\mathrm{d}x$ are equal because they represent the same area — look at the graphs below — the darkly shaded regions in the two graphs have the same area and the lightly shaded regions in the two graphs have the same area.



- Consequently,

$$\begin{aligned}
\int_0^\pi \cos^2 x\mathrm{d}x = \int_0^\pi \sin^2 x\mathrm{d}x &= \frac{1}{2}\left[\int_0^\pi \sin^2 x\mathrm{d}x + \int_0^\pi \cos^2 x\mathrm{d}x\right] \\
&= \frac{1}{2}\int_0^\pi \big[\sin^2 x + \cos^2 x\big]\mathrm{d}x \\
&= \frac{1}{2}\int_0^\pi \mathrm{d}x \\
&= \frac{\pi}{2}
\end{aligned}$$

Example 1.8.10

1.8.2 ▸▸ Integrating $\int \tan^m x \sec^n x \mathrm{d}x$

The strategy for dealing with these integrals is similar to the strategy that we used to evaluate integrals of the form $\int \sin^m x \cos^n x \mathrm{d}x$ and again depends on the parity of the exponents n and m. It uses[51]

$$\frac{\mathrm{d}}{\mathrm{d}x}\tan x = \sec^2 x \qquad \frac{\mathrm{d}}{\mathrm{d}x}\sec x = \sec x\,\tan x \qquad 1+\tan^2 x = \sec^2 x$$

We split the methods for integrating $\int \tan^m x \sec^n x \mathrm{d}x$ into 5 cases which we list below. These will become much more clear after an example (or two).

(1) When m is odd and any n — rewrite the integrand in terms of $\sin x$ and $\cos x$:

$$\begin{aligned}\tan^m x\,\sec^n x \mathrm{d}x &= \left(\frac{\sin x}{\cos x}\right)^m \left(\frac{1}{\cos x}\right)^n \mathrm{d}x \\ &= \frac{\sin^{m-1} x}{\cos^{n+m} x}\,\sin x \mathrm{d}x\end{aligned}$$

and then substitute $u = \cos x$, $\mathrm{d}u = -\sin x \mathrm{d}x$, $\sin^2 x = 1-\cos^2 x = 1-u^2$. See Examples 1.8.11 and 1.8.12.

(2) Alternatively, if m is odd and $n \geqslant 1$ move one factor of $\sec x \tan x$ to the side so that you can see $\sec x \tan x \mathrm{d}x$ in the integral, and substitute $u = \sec x$, $\mathrm{d}u = \sec x \tan x\,\mathrm{d}x$ and $\tan^2 x = \sec^2 x - 1 = u^2 - 1$. See Example 1.8.13.

(3) If n is even with $n \geqslant 2$, move one factor of $\sec^2 x$ to the side so that you can see $\sec^2 x \mathrm{d}x$ in the integral, and substitute $u = \tan x$, $\mathrm{d}u = \sec^2 x\,\mathrm{d}x$ and $\sec^2 x = 1+\tan^2 x = 1+u^2$. See Example 1.8.14.

(4) When m is even and $n = 0$ — that is the integrand is just an even power of tangent — we can still use the $u = \tan x$ substitution, after using $\tan^2 x = \sec^2 x - 1$ (possibly more than once) to create a $\sec^2 x$. See Example 1.8.16.

(5) This leaves the case n odd and m even. There are strategies like those above for treating this case. But they are more complicated and also involve more tricks (that basically have to be memorized). Examples using them are provided in the optional section entitled "Integrating $\sec x$, $\csc x$, $\sec^3 x$ and $\csc^3 x$", below. A more straightforward strategy involves

- multiplying the numerator $\sin^m x$ and the denominator $\cos^{m+n} x$ by $\cos x$, and then
- expressing the denominator, which is now an even power of $\cos x$, in terms of $\sin x$, and then

51 You will need to memorise the derivatives of tangent and secant. However there is no need to memorise $1+\tan^2 x = \sec^2 x$. To derive it very quickly just divide $\sin^2 x + \cos^2 x = 1$ by $\cos^2 x$.

- substituting $u = \sin x$. This converts the integrand into a rational function (i.e. a ratio of two poynomials) of u.
- That rational function of u may be integrated using a technique called "partial fractions", that we will cover later.

We shall return to this strategy after we have learned about partial fractions. See Examples 1.10.5 and 1.10.6 in Section 1.10.

▸▸▸ m is Odd — Odd Power of Tangent

In this case we rewrite the integrand in terms of sine and cosine and then substitute $u = \cos x, \mathrm{d}u = -\sin x \mathrm{d}x$.

Example 1.8.11 $\left(\int \tan x \mathrm{d}x\right)$

Solution.

- Write the integrand $\tan x = \frac{1}{\cos x} \sin x$.
- Now substitute $u = \cos x$, $\mathrm{d}u = -\sin x \, \mathrm{d}x$ just as we did in treating integrands of the form $\sin^m x \cos^n x$ with m odd.

$$\begin{aligned}
\int \tan x \, \mathrm{d}x &= \int \frac{1}{\cos x} \sin x \, \mathrm{d}x && \text{substitute } u = \cos x \\
&= \int \frac{1}{u} \cdot (-1) \mathrm{d}u \\
&= -\log|u| + C \\
&= -\log|\cos x| + C && \text{can also write in terms of secant} \\
&= \log|\cos x|^{-1} + C = \log|\sec x| + C
\end{aligned}$$

Example 1.8.11

Example 1.8.12 $\left(\int \tan^3 x \mathrm{d}x\right)$

Solution.

- Write the integrand $\tan^3 x = \frac{\sin^2 x}{\cos^3 x} \sin x$.
- Again substitute $u = \cos x$, $\mathrm{d}u = -\sin x \, \mathrm{d}x$. We rewrite the remaining even powers of $\sin x$ using $\sin^2 x = 1 - \cos^2 x = 1 - u^2$.

- Hence

$$\begin{aligned}
\int \tan^3 x \, \mathrm{d}x &= \int \frac{\sin^2 x}{\cos^3 x} \sin x \, \mathrm{d}x && \text{substitute } u = \cos x \\
&= \int \frac{1-u^2}{u^3}(-1)\mathrm{d}u \\
&= \frac{u^{-2}}{2} + \log|u| + C \\
&= \frac{1}{2\cos^2 x} + \log|\cos x| + C && \text{can rewrite in terms of secant} \\
&= \frac{1}{2}\sec^2 x - \log|\sec x| + C
\end{aligned}$$

Example 1.8.12

▸▸▸ m **is Odd and** $n \geqslant 1$ **— Odd Power of Tangent and at Least One Secant**

Here we collect a factor of $\tan x \sec x$ and then substitute $u = \sec x$ and $\mathrm{d}u = \sec x \tan x \mathrm{d}x$. We can then rewrite any remaining even powers of $tanx$ in terms of $\sec x$ using $\tan^2 x = \sec^2 x - 1 = u^2 - 1$.

Example 1.8.13 $\left(\int \tan^3 x \sec^4 x \mathrm{d}x\right)$

Solution.

- Start by factoring off one copy of $\sec x \tan x$ and combine it with $\mathrm{d}x$ to form $\sec x \tan x \mathrm{d}x$, which will be $\mathrm{d}u$.
- Now substitute $u = \sec x$, $\mathrm{d}u = \sec x \tan x \mathrm{d}x$ and $\tan^2 x = \sec^2 x - 1 = u^2 - 1$.
- This gives

$$\begin{aligned}
\int \tan^3 x \sec^4 x \mathrm{d}x &= \int \underbrace{\tan^2 x}_{u^2-1} \underbrace{\sec^3 x}_{u^3} \underbrace{\sec x \tan x \mathrm{d}x}_{\mathrm{d}u} \\
&= \int \left[u^2 - 1\right] u^3 \mathrm{d}u \\
&= \frac{u^6}{6} - \frac{u^4}{4} + C \\
&= \frac{1}{6}\sec^6 x - \frac{1}{4}\sec^4 x + C
\end{aligned}$$

Example 1.8.13

▸▸▸ $n \geqslant 2$ is Even — a Positive Even Power of Secant

In the previous case we substituted $u = \sec x$, while in this case we substitute $u = \tan x$. When we do this we write $\mathrm{d}u = \sec^2 x \mathrm{d}x$ and then rewrite any remaining even powers of $\sec x$ as powers of $\tan x$ using $\sec^2 x = 1 + \tan^2 x = 1 + u^2$.

Example 1.8.14 $\left(\int \sec^4 x \mathrm{d}x\right)$

Solution.

- Factor off one copy of $\sec^2 x$ and combine it with $\mathrm{d}x$ to form $\sec^2 x \mathrm{d}x$, which will be $\mathrm{d}u$.
- Then substitute $u = \tan x$, $\mathrm{d}u = \sec^2 x \mathrm{d}x$ and rewrite any remaining even powers of $\sec x$ as powers of $\tan x = u$ using $\sec^2 x = 1 + \tan^2 x = 1 + u^2$.
- This gives

$$\begin{aligned}\int \sec^4 x \mathrm{d}x &= \int \underbrace{\sec^2 x}_{1+u^2} \underbrace{\sec^2 x \mathrm{d}x}_{\mathrm{d}u} \\ &= \int \left[1 + u^2\right] \mathrm{d}u \\ &= u + \frac{u^3}{3} + C \\ &= \tan x + \frac{1}{3}\tan^3 x + C\end{aligned}$$

Example 1.8.14

Example 1.8.15 $\left(\int \tan^3 x \sec^4 x \mathrm{d}x \text{ — redux}\right)$

Solution. Let us revisit this example using this slightly different approach.

- Factor off one copy of $\sec^2 x$ and combine it with $\mathrm{d}x$ to form $\sec^2 x \mathrm{d}x$, which will be $\mathrm{d}u$.
- Then substitute $u = \tan x$, $\mathrm{d}u = \sec^2 x \mathrm{d}x$ and rewrite any remaining even powers of $\sec x$ as powers of $\tan x = u$ using $\sec^2 x = 1 + \tan^2 x = 1 + u^2$.
- This gives

$$\begin{aligned}\int \tan^3 x \sec^4 x \mathrm{d}x &= \int \underbrace{\tan^3 x}_{u^3} \underbrace{\sec^2 x}_{1+u^2} \underbrace{\sec^2 x \mathrm{d}x}_{\mathrm{d}u} \\ &= \int \left[u^3 + u^5\right] \mathrm{d}u \\ &= \frac{u^4}{4} + \frac{u^6}{6} + C \\ &= \frac{1}{4}\tan^4 x + \frac{1}{6}\tan^6 x + C\end{aligned}$$

- This is not quite the same as the answer we got above in Example 1.8.13. However we can show they are (nearly) equivalent. To do so we substitute $v = \sec x$ and $\tan^2 x = \sec^2 x - 1 = v^2 - 1$:

$$\begin{aligned}\frac{1}{6}\tan^6 x + \frac{1}{4}\tan^4 x &= \frac{1}{6}(v^2-1)^3 + \frac{1}{4}(v^2-1)^2 \\ &= \frac{1}{6}(v^6 - 3v^4 + 3v^2 - 1) + \frac{1}{4}(v^4 - 2v^2 + 1) \\ &= \frac{v^6}{6} - \frac{v^4}{2} + \frac{v^2}{2} - \frac{1}{6} + \frac{v^4}{4} - \frac{v^2}{2} + \frac{1}{4} \\ &= \frac{v^6}{6} - \frac{v^4}{4} + 0\cdot v^2 + \left(\frac{1}{4} - \frac{1}{6}\right) \\ &= \frac{1}{6}\sec^6 x - \frac{1}{4}\sec^4 x + \frac{1}{12}.\end{aligned}$$

So while $\frac{1}{6}\tan^6 x + \frac{1}{4}\tan^4 x \neq \frac{1}{6}\sec^6 x - \frac{1}{4}\sec^4 x$, they only differ by a constant. Hence both are valid antiderivatives of $\tan^3 x \sec^4 x$.

Example 1.8.15

▸▸▸ m **is Even and** $n = 0$ **— Even Powers of Tangent**

We integrate this by setting $u = \tan x$. For this to work we need to pull one factor of $\sec^2 x$ to one side to form $du = \sec^2 x dx$. To find this factor of $\sec^2 x$ we (perhaps repeatedly) apply the identity $\tan^2 x = \sec^2 x - 1$.

Example 1.8.16 ($\int \tan^4 x dx$)

Solution.

- There is no $\sec^2 x$ term present, so we try to create it from $\tan^4 x$ by using $\tan^2 x = \sec^2 x - 1$.

$$\begin{aligned}\tan^4 x &= \tan^2 x \cdot \tan^2 x \\ &= \tan^2 x\left[\sec^2 x - 1\right] \\ &= \tan^2 x \sec^2 x - \underbrace{\tan^2 x}_{\sec^2 x - 1} \\ &= \tan^2 x \sec^2 x - \sec^2 x + 1\end{aligned}$$

- Now we can substitute $u = \tan x$, $\mathrm{d}u = \sec^2 x \mathrm{d}x$.

$$\begin{aligned}\int \tan^4 x \mathrm{d}x &= \int \underbrace{\tan^2 x}_{u^2}\ \underbrace{\sec^2 x \mathrm{d}x}_{\mathrm{d}u} - \int \underbrace{\sec^2 x \mathrm{d}x}_{\mathrm{d}u} + \int \mathrm{d}x \\ &= \int u^2 \mathrm{d}u - \int \mathrm{d}u + \int \mathrm{d}x \\ &= \frac{u^3}{3} - u + x + C \\ &= \frac{\tan^3 x}{3} - \tan x + x + C\end{aligned}$$

Example 1.8.16

Example 1.8.17 $\left(\int \tan^8 x \mathrm{d}x\right)$

Solution. Let us try the same approach.

- First pull out a factor of $\tan^2 x$ to create a $\sec^2 x$ factor:

$$\begin{aligned}\tan^8 x &= \tan^6 x \cdot \tan^2 x \\ &= \tan^6 x \cdot \left[\sec^2 x - 1\right] \\ &= \tan^6 x \sec^2 x - \tan^6 x\end{aligned}$$

The first term is now ready to be integrated, but we need to reapply the method to the second term:

$$\begin{aligned}&= \tan^6 x \sec^2 x - \tan^4 x \cdot \left[\sec^2 x - 1\right] \\ &= \tan^6 x \sec^2 x - \tan^4 x \sec^2 x + \tan^4 x && \text{do it again} \\ &= \tan^6 x \sec^2 x - \tan^4 x \sec^2 x + \tan^2 x \cdot \left[\sec^2 x - 1\right] \\ &= \tan^6 x \sec^2 x - \tan^4 x \sec^2 x + \tan^2 x \sec^2 x - \tan^2 x && \text{and again} \\ &= \tan^6 x \sec^2 x - \tan^4 x \sec^2 x + \tan^2 x \sec^2 x - \left[\sec^2 x - 1\right]\end{aligned}$$

- Hence

$$\begin{aligned}\int \tan^8 x \mathrm{d}x &= \int \left[\tan^6 x \sec^2 x - \tan^4 x \sec^2 x + \tan^2 x \sec^2 x - \sec^2 x + 1\right] \mathrm{d}x \\ &= \int \left[\tan^6 x - \tan^4 x + \tan^2 x - 1\right] \sec^2 x \mathrm{d}x + \int \mathrm{d}x \\ &= \int \left[u^6 - u^4 + u^2 - 1\right] \mathrm{d}u + x + C \\ &= \frac{u^7}{7} - \frac{u^5}{5} + \frac{u^3}{3} - u + x + C \\ &= \frac{1}{7}\tan^7 x - \frac{1}{5}\tan^5 x + \frac{1}{3}\tan^3 x - \tan x + x + C\end{aligned}$$

Indeed this example suggests that for integer $k \geqslant 0$:

$$\int \tan^{2k} x \mathrm{d}x = \frac{1}{2k-1}\tan^{2k-1}(x) - \frac{1}{2k-3}\tan^{2k-3} x + \cdots - (-1)^k \tan x + (-1)^k x + C$$

Example 1.8.17

This last example also shows how we might integrate an odd power of tangent:

Example 1.8.18 $\left(\int \tan^7 x\right)$

Solution. We follow the same steps

- Pull out a factor of $\tan^2 x$ to create a factor of $\sec^2 x$:

$$\begin{aligned}
\tan^7 x &= \tan^5 x \cdot \tan^2 x \\
&= \tan^5 x \cdot \left[\sec^2 x - 1\right] \\
&= \tan^5 x \sec^2 x - \tan^5 x && \text{do it again} \\
&= \tan^5 x \sec^2 x - \tan^3 x \cdot \left[\sec^2 x - 1\right] \\
&= \tan^5 x \sec^2 x - \tan^3 x \sec^2 x + \tan^3 x && \text{and again} \\
&= \tan^5 x \sec^2 x - \tan^3 x \sec^2 x + \tan x \left[\sec^2 x - 1\right] \\
&= \tan^5 x \sec^2 x - \tan^3 x \sec^2 x + \tan x \sec^2 x - \tan x
\end{aligned}$$

- Now we can substitute $u = \tan x$ and $\mathrm{d}u = \sec^2 x \mathrm{d}x$ and also use the result from Example 1.8.11 to take care of the last term:

$$\int \tan^7 x \mathrm{d}x = \int \left[\tan^5 x \sec^2 x - \tan^3 x \sec^2 x + \tan x \sec^2 x\right] \mathrm{d}x - \int \tan x \mathrm{d}x$$

Now factor out the common $\sec^2 x$ term and integrate $\tan x$ via Example 1.8.11

$$\begin{aligned}
&= \int \left[\tan^5 x - \tan^3 x + \tan x\right] \sec^2 x \, \mathrm{d}x - \log|\sec x| + C \\
&= \int \left[u^5 - u^3 + u\right] \mathrm{d}u - \log|\sec x| + C \\
&= \frac{u^6}{6} - \frac{u^4}{4} + \frac{u^2}{2} - \log|\sec x| + C \\
&= \frac{1}{6}\tan^6 x - \frac{1}{4}\tan^4 x + \frac{1}{2}\tan^2 x - \log|\sec x| + C
\end{aligned}$$

This example suggests that for integer $k \geqslant 0$:

$$\int \tan^{2k+1} x \mathrm{d}x = \frac{1}{2k}\tan^{2k}(x) - \frac{1}{2k-2}\tan^{2k-2} x + \cdots - (-1)^k \frac{1}{2}\tan^2 x + (-1)^k \log|\sec x| + C$$

Example 1.8.18

Of course we have not considered integrals involving powers of $\cot x$ and $\csc x$. But they can be treated in much the same way as $\tan x$ and $\sec x$ were.

1.8.3 ▸▸ Optional — integrating $\sec x$, $\csc x$, $\sec^3 x$ and $\csc^3 x$

As noted above, when n is odd and m is even, one can use similar strategies as to the previous cases. However the computations are often more involved and more tricks need to be deployed. For this reason we make this section optional — the computations are definitely non-trivial. Rather than trying to construct a coherent "method" for this case, we instead give some examples to give the idea of what to expect.

Example 1.8.19 ($\int \sec x \mathrm{d}x$ — by trickery)

Solution. There is a very sneaky trick to compute this integral[52].

- 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 $\mathrm{d}u = (\sec x \tan x + \sec^2 x)\, \mathrm{d}x$.
- Hence

$$\begin{aligned} \int \sec x \mathrm{d}x &= \int \sec x \, \frac{\sec x + \tan x}{\sec x + \tan x} \mathrm{d}x = \int \frac{\sec^2 x + \sec x \tan x}{\sec x + \tan x} \mathrm{d}x \\ &= \int \frac{1}{u} \mathrm{d}u \\ &= \log |u| + C \\ &= \log |\sec x + \tan x| + C \end{aligned}$$

- The above trick appears both totally unguessable and very hard to remember. Fortunately, there is a simple way[53] 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:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} \tan x &= \sec^2 x \\ \frac{\mathrm{d}}{\mathrm{d}x} \sec x &= \tan x \, \sec x \end{aligned}$$

  - Notice that if we add these together we get

$$\frac{\mathrm{d}}{\mathrm{d}x} \big(\sec x + \tan x \big) = (\sec x + \tan x) \sec x \implies \frac{\frac{\mathrm{d}}{\mathrm{d}x} \big(\sec x + \tan x \big)}{\sec x + \tan x} = \sec x$$

52 The integral of secant played an important role in constructing accurate Mercator projection maps of the earth. See https://en.wikipedia.org/wiki/Integral_of_the_secant_function and https://arxiv.org/pdf/2204.11187.pdf. The method for evaluating the integral that is being presented in this example was invented by the Scottish mathematician James Gregory in 1668. There are also a number of other methods. See the previous two references.

53 We thank Serban Raianu for bringing this to our attention.

– We've done it! The right hand side is $\sec x$ and the left hand side is the derivative of $\log|\sec x + \tan x|$.

Example 1.8.19

There is another method for integrating $\int \sec x \mathrm{d}x$, that is more tedious, but more straight forward. In particular, it does not involve a memorized trick. We first use the substitution $u = \sin x$, $\mathrm{d}u = \cos x \, \mathrm{d}x$, together with $\cos^2 x = 1 - \sin^2 x = 1 - u^2$. This converts the integral into

$$\begin{aligned}\int \sec x \mathrm{d}x &= \int \frac{1}{\cos x} \mathrm{d}x = \int \frac{\cos x \, \mathrm{d}x}{\cos^2 x} \\ &= \int \frac{\mathrm{d}u}{1-u^2}\Big|_{u=\sin x}\end{aligned}$$

The integrand $\frac{1}{1-u^2}$ is a rational function, i.e. a ratio of two polynomials. There is a procedure, called the method of partial fractions, that may be used to integrate any rational function. We shall learn about it in Section 1.10 "Partial Fractions". The detailed evaluation of the integral $\int \sec x \, \mathrm{d}x = \int \frac{\mathrm{d}u}{1-u^2}$ by the method of partial fractions is presented in Example 1.10.5 below.

In addition, there is a standard trick for evaluating $\int \frac{\mathrm{d}u}{1-u^2}$ that allows us to avoid going through the whole partial fractions algorithm.

Example 1.8.20 ($\int \sec x \mathrm{d}x$ — by more trickery)

Solution. We have already seen that

$$\int \sec x \mathrm{d}x = \int \frac{\mathrm{d}u}{1-u^2}\Big|_{u=\sin x}$$

The trick uses the obervations that

- $\frac{1}{1-u^2} = \frac{1+u-u}{1-u^2} = \frac{1}{1-u} - \frac{u}{1-u^2}$
- $\frac{1}{1-u}$ has antiderivative $-\log(1-u)$ (for $u < 1$)
- The derivative $\frac{\mathrm{d}}{\mathrm{d}u}(1-u^2) = -2u$ of the denominator of $\frac{u}{1-u^2}$ is the same, up to a factor of -2, as the numerator of $\frac{u}{1-u^2}$. So we can easily evaluate the integral of $\frac{u}{1-u^2}$ by substituting $v = 1-u^2$, $\mathrm{d}v = -2u \, \mathrm{d}u$.

$$\int \frac{u \, \mathrm{d}u}{1-u^2} = \int \frac{\frac{\mathrm{d}v}{-2}}{v}\Big|_{v=1-u^2} = -\frac{1}{2}\log(1-u^2) + C$$

Combining these observations gives

$$\begin{aligned}
\int \sec x \mathrm{d}x &= \left[\int \frac{\mathrm{d}u}{1-u^2}\right]_{u=\sin x} = \left[\int \frac{1}{1-u}\mathrm{d}u - \int \frac{u}{1-u^2}\mathrm{d}u\right]_{u=\sin x} \\
&= \left[-\log(1-u) + \frac{1}{2}\log(1-u^2) + C\right]_{u=\sin x} \\
&= -\log(1-\sin x) + \frac{1}{2}\log(1-\sin^2 x) + C \\
&= -\log(1-\sin x) + \frac{1}{2}\log(1-\sin x) + \frac{1}{2}\log(1+\sin x) + C \\
&= \frac{1}{2}\log\frac{1+\sin x}{1-\sin x} + C.
\end{aligned}$$

Example 1.8.20

Example 1.8.20 has given the answer

$$\int \sec x \mathrm{d}x = \frac{1}{2}\log\frac{1+\sin x}{1-\sin x} + C$$

which appears to be different than the answer in Example 1.8.19. But they really are the same since

$$\begin{aligned}
&\frac{1+\sin x}{1-\sin x} = \frac{(1+\sin x)^2}{1-\sin^2 x} = \frac{(1+\sin x)^2}{\cos^2 x} \\
\implies &\frac{1}{2}\log\frac{1+\sin x}{1-\sin x} = \frac{1}{2}\log\frac{(1+\sin x)^2}{\cos^2 x} = \log\left|\frac{\sin x+1}{\cos x}\right| = \log|\tan x + \sec x|
\end{aligned}$$

Oof!

Example 1.8.21 ($\int \csc x \mathrm{d}x$ — by the $u = \tan\frac{x}{2}$ substitution)

Solution. The integral $\int \csc x \mathrm{d}x$ may also be evaluated by both the methods above. That is either

- by multiplying the integrand by a cleverly chosen $1 = \frac{\cot x - \csc x}{\cot x - \csc x}$ and then substituting $u = \cot x - \csc x$, $\mathrm{d}u = (-\csc^2 x + \csc x \cot x)\,\mathrm{d}x$, or
- by substituting $u = \cos x$, $\mathrm{d}u = -\sin x\,\mathrm{d}x$ to give $\int \csc x \mathrm{d}x = -\int \frac{\mathrm{d}u}{1-u^2}$ and then using the method of partial fractions.

These two methods give the answers

$$\int \csc x \mathrm{d}x = \log|\cot x - \csc x| + C = -\frac{1}{2}\log\frac{1+\cos x}{1-\cos x} + C \tag{1.8.1}$$

In this example, we shall evaluate $\int \csc x \mathrm{d}x$ by yet a third method, which can be used to integrate rational functions[54] of $\sin x$ and $\cos x$.

54 A rational function of $\sin x$ and $\cos x$ is a ratio with both the numerator and denominator being finite sums of terms of the form $a \sin^m x \cos^n x$, where a is a constant and m and n are positive integers.

- This method uses the substitution

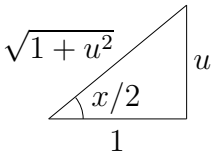
$$x = 2\arctan u \qquad \text{i.e. } u = \tan\frac{x}{2} \qquad \text{and } \mathrm{d}x = \frac{2}{1+u^2}\mathrm{d}u$$

— a half-angle substitution.

- To express $\sin x$ and $\cos x$ in terms of u, we first use the double angle trig identities (Equations 1.8.2 and 1.8.3 with $x \mapsto x/2$) to express $\sin x$ and $\cos x$ in terms of $\sin\frac{x}{2}$ and $\cos\frac{x}{2}$:

$$\sin x = 2\sin\frac{x}{2}\cos\frac{x}{2}$$
$$\cos x = \cos^2\frac{x}{2} - \sin^2\frac{x}{2}$$

- We then use the triangle



to express $\sin\frac{x}{2}$ and $\cos\frac{x}{2}$ in terms of u. The bottom and right hand sides of the triangle have been chosen so that $\tan\frac{x}{2} = u$. This tells us that

$$\sin\frac{x}{2} = \frac{u}{\sqrt{1+u^2}} \qquad \cos\frac{x}{2} = \frac{1}{\sqrt{1+u^2}}$$

- This in turn implies that:

$$\sin x = 2\sin\frac{x}{2}\cos\frac{x}{2} = 2\frac{u}{\sqrt{1+u^2}}\frac{1}{\sqrt{1+u^2}} = \frac{2u}{1+u^2}$$
$$\cos x = \cos^2\frac{x}{2} - \sin^2\frac{x}{2} = \frac{1}{1+u^2} - \frac{u^2}{1+u^2} = \frac{1-u^2}{1+u^2}$$

Oof!

- Let's use this substitution to evaluate $\int \csc x\,\mathrm{d}x$.

$$\int \csc x \mathrm{d}x = \int \frac{1}{\sin x}\mathrm{d}x = \int \frac{1+u^2}{2u}\,\frac{2}{1+u^2}\mathrm{d}u = \int \frac{1}{u}\mathrm{d}u = \log|u| + C$$
$$= \log\left|\tan\frac{x}{2}\right| + C$$

To see that this answer is really the same as that in (1.8.1), note that

$$\cot x - \csc x = \frac{\cos x - 1}{\sin x} = \frac{-2\sin^2(x/2)}{2\sin(x/2)\cos(x/2)} = -\tan\frac{x}{2}$$

Example 1.8.21

Example 1.8.22 ($\int \sec^3 x\,dx$ — by trickery)

Solution. The standard trick used to evaluate $\int \sec^3 x\,dx$ is integration by parts.

- Set $u = \sec x$, $dv = \sec^2 x\,dx$. Hence $du = \sec x \tan x\,dx$, $v = \tan x$ and

$$\begin{aligned}\int \sec^3 x\,dx &= \int \underbrace{\sec x}_{u}\ \underbrace{\sec^2 x\,dx}_{dv} \\ &= \underbrace{\sec x}_{u}\ \underbrace{\tan x}_{v} - \int \underbrace{\tan x}_{v}\ \underbrace{\sec x \tan x\,dx}_{du}\end{aligned}$$

- Since $\tan^2 x + 1 = \sec^2 x$, we have $\tan^2 x = \sec^2 x - 1$ and

$$\begin{aligned}\int \sec^3 x\,dx &= \sec x\ \tan x - \int [\sec^3 x - \sec x]\,dx \\ &= \sec x\ \tan x + \log|\sec x + \tan x| + C - \int \sec^3 x\,dx\end{aligned}$$

  where we used $\int \sec x\,dx = \log|\sec x + \tan x| + C$, which we saw in Example 1.8.19.

- Now moving the $\int \sec^3 x\,dx$ from the right hand side to the left hand side

$$\begin{aligned}2\int \sec^3 x\,dx &= \sec x \tan x + \log|\sec x + \tan x| + C \qquad \text{and so} \\ \int \sec^3 x\,dx &= \frac{1}{2}\sec x \tan x + \frac{1}{2}\log|\sec x + \tan x| + C\end{aligned}$$

  for a new arbitrary constant C (which is just one half the old one).

Example 1.8.22

The integral $\int \sec^3 dx$ can also be evaluated by two other methods.

- Substitute $u = \sin x$, $du = \cos x\,dx$ to convert $\int \sec^3 x\,dx$ into $\int \frac{du}{[1-u^2]^2}$ and evaluate the latter using the method of partial fractions. This is done in Example 1.10.6 in Section 1.10.
- Use the $u = \tan \frac{x}{2}$ substitution. We use this method to evaluate $\int \csc^3 x\,dx$ in Example 1.8.23, below.

Example 1.8.23 ($\int \csc^3 x\,dx$ – by the $u = \tan\frac{x}{2}$ substitution)

Solution. Let us use the half-angle substitution that we introduced in Example 1.8.21.

- In this method we set

$$u = \tan\frac{x}{2} \qquad dx = \frac{2}{1+u^2}du \qquad \sin x = \frac{2u}{1+u^2} \qquad \cos x = \frac{1-u^2}{1+u^2}$$

- The integral then becomes

$$\begin{aligned}\int \csc^3 x\,dx &= \int \frac{1}{\sin^3 x}dx \\ &= \int \Big(\frac{1+u^2}{2u}\Big)^3 \frac{2}{1+u^2}du \\ &= \frac{1}{4}\int \frac{1+2u^2+u^4}{u^3}du \\ &= \frac{1}{4}\Big\{\frac{u^{-2}}{-2} + 2\log|u| + \frac{u^2}{2}\Big\} + C \\ &= \frac{1}{8}\Big\{-\cot^2\frac{x}{2} + 4\log\Big|\tan\frac{x}{2}\Big| + \tan^2\frac{x}{2}\Big\} + C\end{aligned}$$

  Oof!

- This is a perfectly acceptable answer. But if you don't like the $\frac{x}{2}$'s, they may be eliminated by using

$$\begin{aligned}\tan^2\frac{x}{2} - \cot^2\frac{x}{2} &= \frac{\sin^2\frac{x}{2}}{\cos^2\frac{x}{2}} - \frac{\cos^2\frac{x}{2}}{\sin^2\frac{x}{2}} \\ &= \frac{\sin^4\frac{x}{2} - \cos^4\frac{x}{2}}{\sin^2\frac{x}{2}\cos^2\frac{x}{2}} \\ &= \frac{\big(\sin^2\frac{x}{2} - \cos^2\frac{x}{2}\big)\big(\sin^2\frac{x}{2} + \cos^2\frac{x}{2}\big)}{\sin^2\frac{x}{2}\cos^2\frac{x}{2}} \\ &= \frac{\sin^2\frac{x}{2} - \cos^2\frac{x}{2}}{\sin^2\frac{x}{2}\cos^2\frac{x}{2}} && \text{since } \sin^2\frac{x}{2} + \cos^2\frac{x}{2} = 1 \\ &= \frac{-\cos x}{\frac{1}{4}\sin^2 x} && \text{by (1.8.2) and (1.8.3)}\end{aligned}$$

  and

$$\begin{aligned}\tan\frac{x}{2} &= \frac{\sin\frac{x}{2}}{\cos\frac{x}{2}} = \frac{\sin^2\frac{x}{2}}{\sin\frac{x}{2}\cos\frac{x}{2}} \\ &= \frac{\frac{1}{2}[1-\cos x]}{\frac{1}{2}\sin x} \qquad \text{by (1.8.2) and (1.8.3)}\end{aligned}$$

  So we may also write

$$\int \csc^3 x\,dx = -\frac{1}{2}\cot x\csc x + \frac{1}{2}\log|\csc x - \cot x| + C$$

Example 1.8.23

That last optional section was a little scary — let's get back to something a little easier.

1.9▲ Trigonometric Substitution

In this section we discuss substitutions that simplify integrals containing square roots of the form

$$\sqrt{a^2-x^2} \qquad \sqrt{a^2+x^2} \qquad \sqrt{x^2-a^2}.$$

When the integrand contains one of these square roots, then we can use trigonometric substitutions to eliminate them. That is, we substitute

$$x = a\sin u \qquad \text{or} \qquad x = a\tan u \qquad \text{or} \qquad x = a\sec u$$

and then use trigonometric identities

$$\sin^2\theta + \cos^2\theta = 1 \quad \text{and} \quad 1+\tan^2\theta = \sec^2\theta$$

to simplify the result. To be more precise, we can

- eliminate $\sqrt{a^2-x^2}$ from an integrand by substituting $x = a\sin u$ to give

$$\sqrt{a^2-x^2} = \sqrt{a^2-a^2\sin^2 u} = \sqrt{a^2\cos^2 u} = |a\cos u|$$

- eliminate $\sqrt{a^2+x^2}$ from an integrand by substituting $x = a\tan u$ to give

$$\sqrt{a^2+x^2} = \sqrt{a^2+a^2\tan^2 u} = \sqrt{a^2\sec^2 u} = |a\sec u|$$

- eliminate $\sqrt{x^2-a^2}$ from an integrand by substituting $x = a\sec u$ to give

$$\sqrt{x^2-a^2} = \sqrt{a^2\sec^2 u - a^2} = \sqrt{a^2\tan^2 u} = |a\tan u|$$

  Be very careful with signs and absolute values when using this substitution. See Example 1.9.6.

When we have used substitutions before, we usually gave the new integration variable, u, as a function of the old integration variable x. Here we are doing the reverse — we are giving the old integration variable, x, in terms of the new integration variable u. We may do so, as long as we may invert to get u as a function of x. For example, with $x = a\sin u$, we may take $u = \arcsin\frac{x}{a}$. This is a good time for you to review the definitions of $\arcsin\theta$, $\arctan\theta$ and $\operatorname{arcsec}\theta$. See Section 2.12, "Inverse Functions", of the CLP-1 text.

As a warm-up, consider the area of a quarter of the unit circle.

Example 1.9.1 (Quarter of the unit circle)

Compute the area of the unit circle lying in the first quadrant.

Solution. We know that the answer is $\pi/4$, but we can also compute this as an integral — we saw this way back in Example 1.1.16:

$$\text{area} = \int_0^1 \sqrt{1-x^2}\,\mathrm{d}x$$

- To simplify the integrand we substitute $x = \sin u$. With this choice $\frac{\mathrm{d}x}{\mathrm{d}u} = \cos u$ and so $\mathrm{d}x = \cos u\,\mathrm{d}u$.
- We also need to translate the limits of integration and it is perhaps easiest to do this by writing u as a function of x — namely $u(x) = \arcsin x$. Hence $u(0) = 0$ and $u(1) = \pi/2$.
- Hence the integral becomes

$$\begin{aligned}\int_0^1 \sqrt{1-x^2}\,\mathrm{d}x &= \int_0^{\pi/2} \sqrt{1-\sin^2 u}\cdot\cos u\,\mathrm{d}u \\ &= \int_0^{\pi/2} \sqrt{\cos^2 u}\cdot\cos u\,\mathrm{d}u \\ &= \int_0^{\pi/2} \cos^2 u\,\mathrm{d}u\end{aligned}$$

  Notice that here we have used that the *positive* square root $\sqrt{\cos^2 u} = |\cos u| = \cos u$ because $\cos(u) \geqslant 0$ for $0 \leqslant u \leqslant \pi/2$.
- To go further we use the techniques of Section 1.8.

$$\begin{aligned}\int_0^1 \sqrt{1-x^2}\,\mathrm{d}x &= \int_0^{\pi/2} \cos^2 u\,\mathrm{d}u \qquad \text{and since } \cos^2 u = \frac{1+\cos 2u}{2} \\ &= \frac{1}{2}\int_0^{\pi/2} (1+\cos(2u))\,\mathrm{d}u \\ &= \frac{1}{2}\left[u + \frac{1}{2}\sin(2u)\right]_0^{\pi/2} \\ &= \frac{1}{2}\left(\frac{\pi}{2} - 0 + \frac{\sin\pi}{2} - \frac{\sin 0}{2}\right) \\ &= \frac{\pi}{4}\checkmark\end{aligned}$$

Example 1.9.1

Example 1.9.2 $\left(\int \frac{x^2}{\sqrt{1-x^2}}\mathrm{d}x\right)$

Solution. We proceed much as we did in the previous example.

- To simplify the integrand we substitute $x = \sin u$. With this choice $\frac{dx}{du} = \cos u$ and so $dx = \cos u du$. Also note that $u = \arcsin x$.
- The integral becomes

$$\begin{aligned}\int \frac{x^2}{\sqrt{1-x^2}} dx &= \int \frac{\sin^2 u}{\sqrt{1-\sin^2 u}} \cdot \cos u du \\ &= \int \frac{\sin^2 u}{\sqrt{\cos^2 u}} \cdot \cos u du\end{aligned}$$

- To proceed further we need to get rid of the square-root. Since $u = \arcsin x$ has domain $-1 \leqslant x \leqslant 1$ and range $-\pi/2 \leqslant u \leqslant \pi/2$, it follows that $\cos u \geqslant 0$ (since cosine is non-negative on these inputs). Hence

$$\sqrt{\cos^2 u} = \cos u \qquad \text{when } -\pi/2 \leqslant u \leqslant \pi/2$$

- So our integral now becomes

$$\begin{aligned}\int \frac{x^2}{\sqrt{1-x^2}} dx &= \int \frac{\sin^2 u}{\sqrt{\cos^2 u}} \cdot \cos u du \\ &= \int \frac{\sin^2 u}{\cos u} \cdot \cos u du \\ &= \int \sin^2 u du \\ &= \frac{1}{2}\int (1 - \cos 2u) du \qquad \text{by Equation (1.8.4)} \\ &= \frac{u}{2} - \frac{1}{4}\sin 2u + C \\ &= \frac{1}{2}\arcsin x - \frac{1}{4}\sin(2\arcsin x) + C\end{aligned}$$

- We can simplify this further using a double-angle identity. Recall that $u = \arcsin x$ and that $x = \sin u$. Then

$$\sin 2u = 2 \sin u \cos u$$

We can replace $\cos u$ using $\cos^2 u = 1 - \sin^2 u$. Taking a square-root of this formula gives $\cos u = \pm\sqrt{1-\sin^2 u}$. We need the positive branch here since $\cos u \geqslant 0$ when $-\pi/2 \leqslant u \leqslant \pi/2$ (which is exactly the range of $\arcsin x$). Continuing along:

$$\begin{aligned}\sin 2u &= 2 \sin u \cdot \sqrt{1-\sin^2 u} \\ &= 2x\sqrt{1-x^2}\end{aligned}$$

Thus our solution is

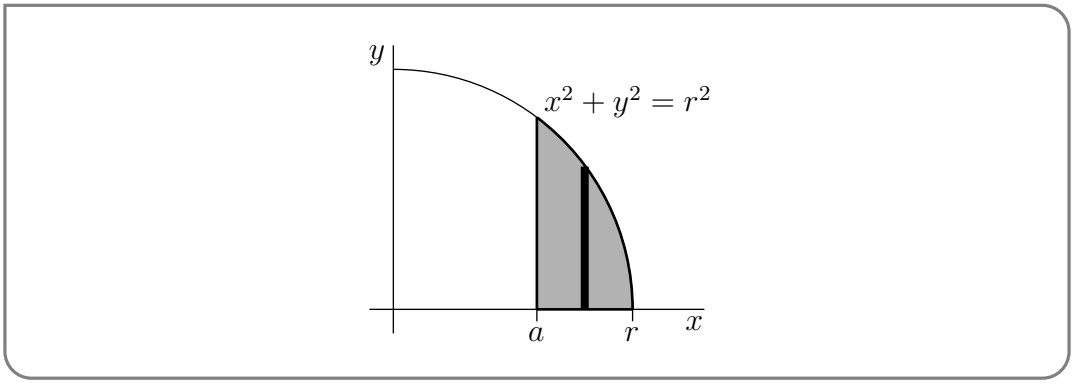
$$\begin{aligned}\int \frac{x^2}{\sqrt{1-x^2}} dx &= \frac{1}{2}\arcsin x - \frac{1}{4}\sin(2\arcsin x) + C \\ &= \frac{1}{2}\arcsin x - \frac{1}{2}x\sqrt{1-x^2} + C\end{aligned}$$

Example 1.9.2

The above two example illustrate the main steps of the approach. The next example is similar, but with more complicated limits of integration.

Example 1.9.3 $\left(\int_a^r \sqrt{r^2-x^2}\,\mathrm{d}x\right)$

Let's find the area of the shaded region in the sketch below.



We'll set up the integral using vertical strips. The strip in the figure has width $\mathrm{d}x$ and height $\sqrt{r^2-x^2}$. So the area is given by the integral

$$\text{area} = \int_a^r \sqrt{r^2-x^2}\,\mathrm{d}x$$

Which is very similar to the previous example.

Solution.

- To evaluate the integral we substitute

$$x = x(u) = r\sin u \qquad\qquad \mathrm{d}x = \frac{\mathrm{d}x}{\mathrm{d}u}\mathrm{d}u = r\cos u\,\mathrm{d}u$$

  It is also helpful to write u as a function of x — namely $u = \arcsin\frac{x}{r}$.

- The integral runs from $x = a$ to $x = r$. These correspond to

$$u(r) = \arcsin\frac{r}{r} = \arcsin 1 = \frac{\pi}{2}$$
$$u(a) = \arcsin\frac{a}{r} \quad \text{which does not simplify further}$$

- The integral then becomes

$$\begin{aligned}\int_a^r \sqrt{r^2-x^2}\mathrm{d}x &= \int_{\arcsin(a/r)}^{\pi/2} \sqrt{r^2-r^2\sin^2 u}\cdot r\cos u\mathrm{d}u\\ &= \int_{\arcsin(a/r)}^{\pi/2} r^2\sqrt{1-\sin^2 u}\cdot\cos u\mathrm{d}u\\ &= r^2\int_{\arcsin(a/r)}^{\pi/2} \sqrt{\cos^2 u}\cdot\cos u\mathrm{d}u\end{aligned}$$

To proceed further (as we did in Examples 1.9.1 and 1.9.2) we need to think about whether $\cos u$ is positive or negative.

- Since a (as shown in the diagram) satisfies $0 \leqslant a \leqslant r$, we know that $u(a)$ lies between $\arcsin(0) = 0$ and $\arcsin(1) = \pi/2$. Hence the variable u lies between 0 and $\pi/2$, and on this range $\cos u \geqslant 0$. This allows us get rid of the square-root:

$$\sqrt{\cos^2 u} = |\cos u| = \cos u$$

- Putting this fact into our integral we get

$$\begin{aligned}\int_a^r \sqrt{r^2 - x^2}\,\mathrm{d}x &= r^2 \int_{\arcsin(a/r)}^{\pi/2} \sqrt{\cos^2 u} \cdot \cos u \,\mathrm{d}u \\ &= r^2 \int_{\arcsin(a/r)}^{\pi/2} \cos^2 u \,\mathrm{d}u\end{aligned}$$

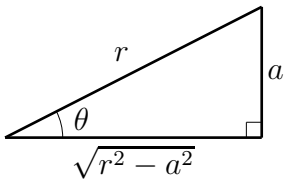
Recall the identity $\cos^2 u = \frac{1+\cos 2u}{2}$ from Section 1.8

$$\begin{aligned}&= \frac{r^2}{2} \int_{\arcsin(a/r)}^{\pi/2} (1 + \cos 2u)\mathrm{d}u \\ &= \frac{r^2}{2}\left[u + \frac{1}{2}\sin(2u)\right]_{\arcsin(a/r)}^{\pi/2} \\ &= \frac{r^2}{2}\left(\frac{\pi}{2} + \frac{1}{2}\sin\pi - \arcsin(a/r) - \frac{1}{2}\sin(2\arcsin(a/r))\right) \\ &= \frac{r^2}{2}\left(\frac{\pi}{2} - \arcsin(a/r) - \frac{1}{2}\sin(2\arcsin(a/r))\right)\end{aligned}$$

Oof! But there is a little further to go before we are done.

- We can again simplify the term $\sin(2\arcsin(a/r))$ using a double angle identity. Set $\theta = \arcsin(a/r)$. Then θ is the angle in the triangle on the right below. By the double angle formula for $\sin(2\theta)$ (Equation (1.8.2))

$$\begin{aligned}\sin(2\theta) &= 2\sin\theta\ \cos\theta \\ &= 2\,\frac{a}{r}\,\frac{\sqrt{r^2 - a^2}}{r}.\end{aligned}$$



- So finally the area is

$$\begin{aligned}\text{area} &= \int_a^r \sqrt{r^2 - x^2}\,\mathrm{d}x \\ &= \frac{r^2}{2}\left(\frac{\pi}{2} - \arcsin(a/r) - \frac{1}{2}\sin(2\arcsin(a/r))\right) \\ &= \frac{\pi r^2}{4} - \frac{r^2}{2}\arcsin(a/r) - \frac{a}{2}\sqrt{r^2 - a^2}\end{aligned}$$

- This is a relatively complicated formula, but we can make some "reasonableness" checks, by looking at special values of a.
  - If $a = 0$ the shaded region, in the figure at the beginning of this example, is exactly one quarter of a disk of radius r and so has area $\frac{1}{4}\pi r^2$. Substituting $a = 0$ into our answer does indeed give $\frac{1}{4}\pi r^2$.
  - At the other extreme, if $a = r$, the shaded region disappears completely and so has area 0. Subbing $a = r$ into our answer does indeed give 0, since $\arcsin 1 = \frac{\pi}{2}$.

Example 1.9.3

Example 1.9.4 $\left(\int_a^r x\sqrt{r^2-x^2}\,\mathrm{d}x\right)$

The integral $\int_a^r x\sqrt{r^2-x^2}\,\mathrm{d}x$ looks a lot like the integral we just did in the previous 3 examples. It can also be evaluated using the trigonometric substitution $x = r\sin u$ — but that is unnecessarily complicated. Just because you have now learned how to use trigonometric substitution[55] doesn't mean that you should forget everything you learned before.

Solution. This integral is *much* more easily evaluated using the simple substitution $u = r^2 - x^2$.

- Set $u = r^2 - x^2$. Then $\mathrm{d}u = -2x\mathrm{d}x$, and so

$$\begin{aligned}\int_a^r x\sqrt{r^2-x^2}\,\mathrm{d}x &= \int_{r^2-a^2}^0 \sqrt{u}\,\frac{\mathrm{d}u}{-2}\\ &= -\frac{1}{2}\left[\frac{u^{3/2}}{3/2}\right]_{r^2-a^2}^0\\ &= \frac{1}{3}\left[r^2-a^2\right]^{3/2}\end{aligned}$$

Example 1.9.4

Enough sines and cosines — let us try a tangent substitution.

Example 1.9.5 $\left(\int \frac{\mathrm{d}x}{x^2\sqrt{9+x^2}}\right)$

Solution. As per our guidelines at the start of this section, the presence of the square root term $\sqrt{3^2+x^2}$ tells us to substitute $x = 3\tan u$.

- Substitute

$$x = 3\tan u \qquad\qquad \mathrm{d}x = 3\sec^2 u\,\mathrm{d}u$$

55 To paraphrase the Law of the Instrument, possibly Mark Twain and definitely some psychologists, when you have a shiny new hammer, everything looks like a nail.

This allows us to remove the square root:

$$\sqrt{9+x^2} = \sqrt{9+9\tan^2 u} = 3\sqrt{1+\tan^2 u} = 3\sqrt{\sec^2 u} = 3|\sec u|$$

- Hence our integral becomes

$$\int \frac{\mathrm{d}x}{x^2\sqrt{9+x^2}} = \int \frac{3\sec^2 u}{9\tan^2 u \cdot 3|\sec u|}\mathrm{d}u$$

- To remove the absolute value we must consider the range of values of u in the integral. Since $x = 3\tan u$ we have $u = \arctan(x/3)$. The range[56] of arctangent is $-\pi/2 \leqslant \arctan \leqslant \pi/2$ and so $u = \arctan(x/3)$ will always like between $-\pi/2$ and $+\pi/2$. Hence $\cos u$ will always be positive, which in turn implies that $|\sec u| = \sec u$.

- Using this fact our integral becomes:

$$\begin{aligned}\int \frac{\mathrm{d}x}{x^2\sqrt{9+x^2}} &= \int \frac{3\sec^2 u}{27\tan^2 u|\sec u|}\mathrm{d}u \\ &= \frac{1}{9}\int \frac{\sec u}{\tan^2 u}\mathrm{d}u \qquad \text{since } \sec u > 0\end{aligned}$$

- Rewrite this in terms of sine and cosine

$$\int \frac{\mathrm{d}x}{x^2\sqrt{9+x^2}} = \frac{1}{9}\int \frac{\sec u}{\tan^2 u}\mathrm{d}u \tag{1.9.1}$$

$$= \frac{1}{9}\int \frac{1}{\cos u}\cdot\frac{\cos^2 u}{\sin^2 u}\mathrm{d}u = \frac{1}{9}\int \frac{\cos u}{\sin^2 u}\mathrm{d}u \tag{1.9.2}$$

Now we can use the substitution rule with $y = \sin u$ and $\mathrm{d}y = \cos u\mathrm{d}u$

$$= \frac{1}{9}\int \frac{\mathrm{d}y}{y^2} \tag{1.9.3}$$

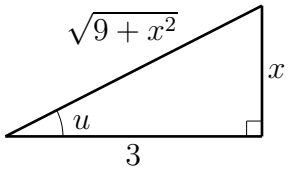
$$= -\frac{1}{9y} + C \tag{1.9.4}$$

$$= -\frac{1}{9\sin u} + C \tag{1.9.5}$$

- The original integral was a function of x, so we still have to rewrite $\sin u$ in terms of x. Remember that $x = 3\tan u$ or $u = \arctan(x/3)$. So u is the angle shown in the triangle below and we can read off the triangle that

$$\sin u = \frac{x}{\sqrt{9+x^2}}$$

$$\implies \int \frac{\mathrm{d}x}{x^2\sqrt{9+x^2}} = -\frac{\sqrt{9+x^2}}{9x} + C$$



56 To be pedantic, we mean the range of the "standard" arctangent function or its "principal value". One can define other arctangent functions with different ranges.

Example 1.9.5

Example 1.9.6 $\left(\int \frac{x^2}{\sqrt{x^2-1}}\mathrm{d}x\right)$

Solution. This one requires a secant substitution, but otherwise is very similar to those above.

- Set $x = \sec u$ and $\mathrm{d}x = \sec u \tan u \,\mathrm{d}u$. Then

$$\begin{aligned}\int \frac{x^2}{\sqrt{x^2-1}}\mathrm{d}x &= \int \frac{\sec^2 u}{\sqrt{\sec^2 u - 1}} \sec u \tan u \,\mathrm{d}u \\ &= \int \sec^3 u \cdot \frac{\tan u}{\sqrt{\tan^2 u}}\mathrm{d}u \qquad \text{since } \tan^2 u = \sec^2 u - 1 \\ &= \int \sec^3 u \cdot \frac{\tan u}{|\tan u|}\mathrm{d}u\end{aligned}$$

- As before we need to consider the range of u values in order to determine the sign of $\tan u$. Notice that the integrand is only defined when either $x < -1$ or $x > 1$; thus we should treat the cases $x < -1$ and $x > 1$ separately. Let us assume that $x > 1$ and we will come back to the case $x < -1$ at the end of the example.

  When $x > 1$, our $u = \operatorname{arcsec} x$ takes values in $(0, \pi/2)$. This follows since when $0 < u < \pi/2$, we have $0 < \cos u < 1$ and so $\sec u > 1$. Further, when $0 < u < \pi/2$, we have $\tan u > 0$. Thus $|\tan u| = \tan u$.

- Back to our integral, when $x > 1$:

$$\begin{aligned}\int \frac{x^2}{\sqrt{x^2-1}}\mathrm{d}x &= \int \sec^3 u \cdot \frac{\tan u}{|\tan u|}\mathrm{d}u \\ &= \int \sec^3 u \mathrm{d}u \qquad \text{since } \tan u \geqslant 0\end{aligned}$$

  This is exactly Example 1.8.22

$$= \frac{1}{2}\sec u \tan u + \frac{1}{2}\log|\sec u + \tan u| + C$$

- Since we started with a function of x we need to finish with one. We know that $\sec u = x$ and then we can use trig identities

  $\tan^2 u = \sec^2 u - 1 = x^2 - 1$ so $\tan u = \pm\sqrt{x^2-1}$, but we know $\tan u \geqslant 0$, so $\tan u = \sqrt{x^2-1}$

  Thus

$$\int \frac{x^2}{\sqrt{x^2-1}}\mathrm{d}x = \frac{1}{2}x\sqrt{x^2-1} + \frac{1}{2}\log|x + \sqrt{x^2-1}| + C$$

- The above holds when $x > 1$. We can confirm that it is also true when $x < -1$ by showing the right-hand side is a valid antiderivative of the integrand. To do so we must differentiate our answer. Notice that we do not need to consider the sign of $x + \sqrt{x^2-1}$ when we differentiate since we have already seen that

$$\frac{\mathrm{d}}{\mathrm{d}x} \log|x| = \frac{1}{x}$$

when either $x < 0$ or $x > 0$. So the following computation applies to both $x > 1$ and $x < -1$. The expressions become quite long so we differentiate each term separately:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left[x\sqrt{x^2-1}\right] &= \left[\sqrt{x^2-1} + \frac{x^2}{\sqrt{x^2-1}}\right] \\
&= \frac{1}{\sqrt{x^2-1}}\left[(x^2-1)+x^2\right] \\
\frac{\mathrm{d}}{\mathrm{d}x}\log\left|x+\sqrt{x^2-1}\right| &= \frac{1}{x+\sqrt{x^2-1}} \cdot \left[1 + \frac{x}{\sqrt{x^2-1}}\right] \\
&= \frac{1}{x+\sqrt{x^2-1}} \cdot \frac{x+\sqrt{x^2-1}}{\sqrt{x^2-1}} \\
&= \frac{1}{\sqrt{x^2-1}}
\end{aligned}$$

Putting things together then gives us

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1}{2}x\sqrt{x^2-1} + \frac{1}{2}\log|x+\sqrt{x^2-1}| + C\right] &= \frac{1}{2\sqrt{x^2-1}}\left[(x^2-1)+x^2+1\right] + 0 \\
&= \frac{x^2}{\sqrt{x^2-1}}
\end{aligned}$$

This tells us that our answer for $x > 1$ is also valid when $x < -1$ and so

$$\int \frac{x^2}{\sqrt{x^2-1}}\mathrm{d}x = \frac{1}{2}x\sqrt{x^2-1} + \frac{1}{2}\log|x+\sqrt{x^2-1}| + C$$

when $x < -1$ and when $x > 1$.

In this example, we were lucky. The answer that we derived for $x > 1$ happened to be also valid for $x < -1$. This does not always happen with the $x = a\sec u$ substitution. When it doesn't, we have to apply separate $x > a$ and $x < -a$ analyses that are very similar to our $x > 1$ analysis above. Of course that doubles the tedium. So in the CLP-2 problem book, we will not pose questions that require separate $x > a$ and $x < -a$ computations.

Example 1.9.6

The method, as we have demonstrated it above, works when our integrand contains the square root of very specific families of quadratic polynomials. In fact, the same method works for more general quadratic polynomials — all we need to do is complete the square[57].

57 If you have not heard of "completing the square" don't worry. It is not a difficult method and it will only take you a few moments to learn. It refers to rewriting a quadratic polynomial

$$P(x) = ax^2 + bx + c \qquad \text{as} \qquad P(x) = a(x+h)^2 + k$$

Example 1.9.7 $\left(\int_3^5 \frac{\sqrt{x^2-2x-3}}{x-1}\,\mathrm{d}x\right)$

This time we have an integral with a square root in the integrand, but the argument of the square root, while a quadratic function of x, is not in one of the standard forms $\sqrt{a^2-x^2}$, $\sqrt{a^2+x^2}$, $\sqrt{x^2-a^2}$. The reason that it is not in one of those forms is that the argument, x^2-2x-3, contains a term , namely $-2x$ that is of degree one in x. So we try to manipulate it into one of the standard forms by completing the square.

Solution.

- We first rewrite the quadratic polynomial x^2-2x-3 in the form $(x-a)^2+b$ for some constants a, b. The easiest way to do this is to expand both expressions and compare coefficients of x:

$$x^2-2x-3 = (x-a)^2+b = (x^2-2ax+a^2)+b$$

  So — if we choose $-2a=-2$ (so the coefficients of x^1 match) and $a^2+b=-3$ (so the coefficients of x^0 match), then both expressions are equal. Hence we set $a=1$ and $b=-4$. That is

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

  Many of you may have seen this method when learning to sketch parabolas.

- Once this is done we can convert the square root of the integrand into a standard form by making the simple substitution $y=x-1$. Here goes

$$\begin{aligned}
\int_3^5 \frac{\sqrt{x^2-2x-3}}{x-1}\,\mathrm{d}x &= \int_3^5 \frac{\sqrt{(x-1)^2-4}}{x-1}\,\mathrm{d}x \\
&= \int_2^4 \frac{\sqrt{y^2-4}}{y}\mathrm{d}y && \text{with } y=x-1, \mathrm{d}y=\mathrm{d}x \\
&= \int_0^{\pi/3} \frac{\sqrt{4\sec^2 u-4}}{2\sec u}\,2\sec u\tan u\,\mathrm{d}u && \text{with } y=2\sec u \\
& && \text{and } \mathrm{d}y=2\sec u\tan u\,\mathrm{d}u
\end{aligned}$$

  Notice that we could also do this in fewer steps by setting $(x-1)=2\sec u, \mathrm{d}x = 2\sec u\tan u\mathrm{d}u$.

- To get the limits of integration we used that
  - the value of u that corresponds to $y=2$ obeys $2=y=2\sec u=\frac{2}{\cos u}$ or $\cos u=1$, so that $u=0$ works and
  - the value of u that corresponds to $y=4$ obeys $4=y=2\sec u=\frac{2}{\cos u}$ or $\cos u=\frac{1}{2}$, so that $u=\pi/3$ works.

for new constants h, k.

- Now returning to the evaluation of the integral, we simplify and continue.

$$\begin{aligned}\int_3^5 \frac{\sqrt{x^2-2x-3}}{x-1}\,\mathrm{d}x &= \int_0^{\pi/3} 2\sqrt{\sec^2 u - 1}\,\tan u\,\mathrm{d}u \\ &= 2\int_0^{\pi/3} \tan^2 u\,\mathrm{d}u \qquad \text{since } \sec^2 u = 1+\tan^2 u\end{aligned}$$

In taking the square root of $\sec^2 u - 1 = \tan^2 u$ we used that $\tan u \geqslant 0$ on the range $0 \leqslant u \leqslant \frac{\pi}{3}$.

$$\begin{aligned}&= 2\int_0^{\pi/3} \left[\sec^2 u - 1\right]\mathrm{d}u \qquad \text{since } \sec^2 u = 1+\tan^2 u, \text{ again} \\ &= 2\Big[\tan u - u\Big]_0^{\pi/3} \qquad \text{since } \frac{\mathrm{d}}{\mathrm{d}u}\tan u = \sec^2 u \\ &= 2\big[\sqrt{3} - \pi/3\big]\end{aligned}$$

Example 1.9.7

1.10▲ Partial Fractions

Partial fractions is the name given to a technique of integration that may be used to integrate any rational function[58]. We already know how to integrate some simple rational functions

$$\int \frac{1}{x}\mathrm{d}x = \log|x| + C \qquad \int \frac{1}{1+x^2}\mathrm{d}x = \arctan(x) + C$$

Combining these with the substitution rule, we can integrate similar but more complicated rational functions:

$$\int \frac{1}{2x+3}\mathrm{d}x = \frac{1}{2}\log|2x+3| + C \qquad \int \frac{1}{3+4x^2}\mathrm{d}x = \frac{1}{2\sqrt{3}}\arctan\left(\frac{2x}{\sqrt{3}}\right) + C$$

By summing such terms together we can integrate yet more complicated forms

$$\int \left[x + \frac{1}{x+1} + \frac{1}{x-1}\right]\mathrm{d}x = \frac{x^2}{2} + \log|x+1| + \log|x-1| + C$$

However we are not (typically) presented with a rational function nicely decomposed into neat little pieces. It is far more likely that the rational function will be written as the ratio of two polynomials. For example:

$$\int \frac{x^3+x}{x^2-1}\mathrm{d}x$$

58 Recall that a rational function is the ratio of two polynomials.

In this specific example it is not hard to confirm that

$$x+\frac{1}{x+1}+\frac{1}{x-1}=\frac{x(x+1)(x-1)+(x-1)+(x+1)}{(x+1)(x-1)}=\frac{x^3+x}{x^2-1}$$

and hence

$$\begin{aligned}\int\frac{x^3+x}{x^2-1}\mathrm{d}x&=\int\left[x+\frac{1}{x+1}+\frac{1}{x-1}\right]\mathrm{d}x\\&=\frac{x^2}{2}+\log|x+1|+\log|x-1|+C\end{aligned}$$

Of course going in this direction (from a sum of terms to a single rational function) is straightforward. To be useful we need to understand how to do this in reverse: decompose a given rational function into a sum of simpler pieces that we can integrate.

Suppose that $N(x)$ and $D(x)$ are polynomials. The basic strategy is to write $\frac{N(x)}{D(x)}$ as a sum of very simple, easy to integrate rational functions, namely

(1) polynomials — we shall see below that these are needed when the degree[59] of $N(x)$ is equal to or strictly bigger than the degree of $D(x)$, and

(2) rational functions of the particularly simple form $\frac{A}{(ax+b)^n}$ and

(3) rational functions of the form $\frac{Ax+B}{(ax^2+bx+c)^m}$.

We already know how to integrate the first two forms, and we'll see how to integrate the third form in the near future.

To begin to explore this method of decomposition, let us go back to the example we just saw

$$x+\frac{1}{x+1}+\frac{1}{x-1}=\frac{x(x+1)(x-1)+(x-1)+(x+1)}{(x+1)(x-1)}=\frac{x^3+x}{x^2-1}$$

The technique that we will use is based on two observations:

(1) The denominators on the left-hand side are the factors of the denominator $x^2-1=(x-1)(x+1)$ on the right-hand side.

(2) Use $P(x)$ to denote the polynomial on the left hand side, and then use $N(x)$ and $D(x)$ to denote the numerator and denominator of the right hand side. That is

$$P(x)=x \qquad N(x)=x^3+x \qquad D(x)=x^2-1.$$

Then the degree of $N(x)$ is the sum of the degrees of $P(x)$ and $D(x)$. This is because the highest degree term in $N(x)$ is x^3, which comes from multiplying $P(x)$ by $D(x)$, as we see in

$$x+\frac{1}{x+1}+\frac{1}{x-1}=\frac{\overbrace{x}^{P(x)}\overbrace{(x+1)(x-1)}^{D(x)}+(x-1)+(x+1)}{(x+1)(x-1)}=\frac{x^3+x}{x^2-1}$$

59 The degree of a polynomial is the largest power of x. For example, the degree of $2x^3+4x^2+6x+8$ is three.

More generally, the presence of a polynomial on the left hand side is signalled on the right hand side by the fact that the degree of the numerator is at least as large as the degree of the denominator.

1.10.1 ▸▸ Partial Fraction Decomposition Examples

Rather than writing up the technique — known as the partial fraction decomposition — in full generality, we will instead illustrate it through a sequence of examples.

Example 1.10.1 $\left(\int \frac{x-3}{x^2-3x+2}\mathrm{d}x\right)$

In this example, we integrate $\frac{N(x)}{D(x)} = \frac{x-3}{x^2-3x+2}$.

Solution.

- *Step 1.* We first check to see if a polynomial $P(x)$ is needed. To do so, we check to see if the degree of the numerator, $N(x)$, is strictly smaller than the degree of the denominator $D(x)$. In this example, the numerator, $x-3$, has degree one and that is indeed strictly smaller than the degree of the denominator, x^2-3x+2, which is two. In this case[60] we do not need to extract a polynomial $P(x)$ and we move on to step 2.

- *Step 2.* The second step is to factor the denominator

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

  In this example it is quite easy, but in future examples (and quite possibly in your homework, quizzes and exam) you will have to work harder to factor the denominator. In Appendix A.16 we have written up some simple tricks for factoring polynomials. We will illustrate them in Example 1.10.3 below.

- *Step 3.* The third step is to write $\frac{x-3}{x^2-3x+2}$ in the form

$$\frac{x-3}{x^2-3x+2} = \frac{A}{x-1} + \frac{B}{x-2}$$

  for some constants A and B. More generally, if the denominator consists of n different linear factors, then we decompose the ratio as

$$\text{rational function} = \frac{A_1}{\text{linear factor 1}} + \frac{A_2}{\text{linear factor 2}} + \cdots + \frac{A_n}{\text{linear factor n}}$$

  To proceed we need to determine the values of the constants A, B and there are several different methods to do so. Here are two methods

60 We will soon get to an example (Example 1.10.2 in fact) in which the numerator degree is at least as large as the denominator degree — in that situation we have to extract a polynomial $P(x)$ before we can move on to step 2.

- *Step 3 – Algebra Method.* This approach has the benefit of being conceptually clearer and easier, but the downside is that it is more tedious.

  To determine the values of the constants A, B, we put[61] the right-hand side back over the common denominator $(x-1)(x-2)$.

$$\frac{x-3}{x^2-3x+2} = \frac{A}{x-1} + \frac{B}{x-2} = \frac{A(x-2)+B(x-1)}{(x-1)(x-2)}$$

  The fraction on the far left is the same as the fraction on the far right if and only if their numerators are the same.

$$x-3 = A(x-2)+B(x-1)$$

  Write the right hand side as a polynomial in standard form (i.e. collect up all x terms and all constant terms)

$$x-3 = (A+B)x + (-2A-B)$$

  For these two polynomials to be the same, the coefficient of x on the left hand side and the coefficient of x on the right hand side must be the same. Similarly the coefficients of x^0 (i.e. the constant terms) must match. This gives us a system of two equations.

$$A+B=1 \qquad\qquad -2A-B=-3$$

  in the two unknowns A, B. We can solve this system by

  - using the first equation, namely $A+B=1$, to determine A in terms of B:

$$A = 1-B$$

  - Substituting this into the remaining equation eliminates the A from second equation, leaving one equation in the one unknown B, which can then be solved for B:

$$\begin{aligned} -2A-B &= -3 && \text{substitute } A=1-B \\ -2(1-B)-B &= -3 && \text{clean up} \\ -2+B &= -3 && \text{so } B=-1 \end{aligned}$$

  - Once we know B, we can substitute it back into $A=1-B$ to get A.

$$A = 1-B = 1-(-1) = 2$$

  Hence

$$\frac{x-3}{x^2-3x+2} = \frac{2}{x-1} - \frac{1}{x-2}$$

61 That is, we take the decomposed form and sum it back together.

- *Step 3 – Sneaky Method.* This takes a little more work to understand, but it is more efficient than the algebra method.

  We wish to find A and B for which

  $$\frac{x-3}{(x-1)(x-2)} = \frac{A}{x-1} + \frac{B}{x-2}$$

  Note that the denominator on the left hand side has been written in factored form.

  – To determine A, we multiply both sides of the equation by A's denominator, which is $x-1$,

  $$\frac{x-3}{x-2} = A + \frac{(x-1)B}{x-2}$$

  and then we completely eliminate B from the equation by evaluating at $x = 1$. This value of x is chosen to make $x - 1 = 0$.

  $$\left.\frac{x-3}{x-2}\right|_{x=1} = A + \left.\frac{(x-1)B}{x-2}\right|_{x=1} \implies A = \frac{1-3}{1-2} = 2$$

  – To determine B, we multiply both sides of the equation by B's denominator, which is $x-2$,

  $$\frac{x-3}{x-1} = \frac{(x-2)A}{x-1} + B$$

  and then we completely eliminate A from the equation by evaluating at $x = 2$. This value of x is chosen to make $x - 2 = 0$.

  $$\left.\frac{x-3}{x-1}\right|_{x=2} = \left.\frac{(x-2)A}{x-1}\right|_{x=2} + B \implies B = \frac{2-3}{2-1} = -1$$

  Hence we have (the thankfully consistent answer)

  $$\frac{x-3}{x^2-3x+2} = \frac{2}{x-1} - \frac{1}{x-2}$$

  Notice that no matter which method we use to find the constants we can easily check our answer by summing the terms back together:

  $$\frac{2}{x-1} - \frac{1}{x-2} = \frac{2(x-2)-(x-1)}{(x-2)(x-1)} = \frac{2x-4-x+1}{x^2-3x+2} = \frac{x-3}{x^2-3x+2}\checkmark$$

Step 4. The final step is to integrate.

$$\int \frac{x-3}{x^2-3x+2}\mathrm{d}x = \int \frac{2}{x-1}\mathrm{d}x + \int \frac{-1}{x-2}\mathrm{d}x = 2\log|x-1| - \log|x-2| + C$$

Example 1.10.1

Perhaps the first thing that you notice is that this process takes quite a few steps[62]. However no single step is all that complicated; it only takes practice. With that said, let's do another, slightly more complicated, one.

Example 1.10.2 $\left(\int \frac{3x^3-8x^2+4x-1}{x^2-3x+2}\mathrm{d}x\right)$

In this example, we integrate $\frac{N(x)}{D(x)} = \frac{3x^3-8x^2+4x-1}{x^2-3x+2}$.

Solution.

- *Step 1.* We first check to see if the degree of the numerator $N(x)$ is strictly smaller than the degree of the denominator $D(x)$. In this example, the numerator, $3x^3 - 8x^2 + 4x - 1$, has degree three and the denominator, $x^2 - 3x + 2$, has degree two. As $3 \geqslant 2$, we have to implement the first step.

  The goal of the first step is to write $\frac{N(x)}{D(x)}$ in the form

  $$\frac{N(x)}{D(x)} = P(x) + \frac{R(x)}{D(x)}$$

  with $P(x)$ being a polynomial and $R(x)$ being a polynomial of degree strictly smaller than the degree of $D(x)$. The right hand side is $\frac{P(x)D(x)+R(x)}{D(x)}$, so we have to express the numerator in the form $N(x) = P(x)D(x) + R(x)$, with $P(x)$ and $R(x)$ being polynomials and with the degree of R being strictly smaller than the degree of D. $P(x)D(x)$ is a sum of expressions of the form $ax^nD(x)$. We want to pull as many expressions of this form as possible out of the numerator $N(x)$, leaving only a low degree remainder $R(x)$.

  We do this using long division — the same long division you learned in school, but with the base 10 replaced by x.

  - We start by observing that to get from the highest degree term in the denominator (x^2) to the highest degree term in the numerator ($3x^3$), we have to multiply it by $3x$. So we write,

    $$\begin{array}{r} 3x \\ x^2-3x+2 \,\overline{)\, 3x^3-8x^2+4x-1} \end{array}$$

    In the above expression, the denominator is on the left, the numerator is on the right and $3x$ is written above the highest order term of the numerator. Always put lower powers of x to the right of higher powers of x — this mirrors how you do long division with numbers; lower powers of ten sit to the right of higher powers of ten.
  - Now we subtract $3x$ times the denominator, $x^2 - 3x + 2$, which is $3x^3 - 9x^2 + 6x$, from the numerator.

62 Though, in fairness, we did step 3 twice — and that is the most tedious bit... Actually — sometimes factoring the denominator can be quite challenging. We'll consider this issue in more detail shortly.

$$\begin{array}{r} 3x \\ x^2-3x+2 \,\overline{\big|\, 3x^3-8x^2+4x-1} \\ \underline{3x^3-9x^2+6x} \quad \longleftarrow 3x(x^2-3x+2) \\ x^2-2x-1 \end{array}$$

– This has left a remainder of x^2-2x-1. To get from the highest degree term in the denominator (x^2) to the highest degree term in the remainder (x^2), we have to multiply by 1. So we write,

$$\begin{array}{r} 3x+1 \\ x^2-3x+2 \,\overline{\big|\, 3x^3-8x^2+4x-1} \\ \underline{3x^3-9x^2+6x} \\ x^2-2x-1 \end{array}$$

– Now we subtract 1 times the denominator, x^2-3x+2, which is x^2-3x+2, from the remainder.

$$\begin{array}{r} 3x+1 \\ x^2-3x+2 \,\overline{\big|\, 3x^3-8x^2+4x-1} \\ \underline{3x^3-9x^2+6x} \quad \longleftarrow 3x(x^2-3x+2) \\ x^2-2x-1 \\ \underline{x^2-3x+2} \quad \longleftarrow 1\,(x^2-3x+2) \\ x-3 \end{array}$$

– This leaves a remainder of $x-3$. Because the remainder has degree 1, which is smaller than the degree of the denominator (being degree 2), we stop.

– In this example, when we subtracted $3x(x^2-3x+2)$ and $1(x^2-3x+2)$ from $3x^3-8x^2+4x-1$ we ended up with $x-3$. That is,

$$3x^3-8x^2+4x-1 \;-\; 3x(x^2-3x+2) \;-\; 1(x^2-3x+2) \;=\; x-3$$

or, collecting the two terms proportional to (x^2-3x+2)

$$3x^3-8x^2+4x-1 \;-\; (3x+1)(x^2-3x+2) \;=\; x-3$$

Moving the $(3x+1)(x^2-3x+2)$ to the right hand side and dividing the whole equation by x^2-3x+2 gives

$$\frac{3x^3-8x^2+4x-1}{x^2-3x+2} \;=\; 3x+1 \;+\; \frac{x-3}{x^2-3x+2}$$

And we can easily check this expression just by summing the two terms on the right-hand side.

We have written the integrand in the form $\frac{N(x)}{D(x)} = P(x) + \frac{R(x)}{D(x)}$, with the degree of $R(x)$ strictly smaller than the degree of $D(x)$, which is what we wanted. Observe that $R(x)$ is the final remainder of the long division procedure and $P(x)$ is at the top of the long division computation.

$$\begin{array}{rl} & 3x+1 \quad \longleftarrow P(x) \\ D(x)\to x^2-3x+2 \,\big|\, & 3x^3-8x^2+4x-1 \quad \longleftarrow N(x) \\ & \underline{3x^3-9x^2+6x} \quad \longleftarrow 3x\cdot D(x) \\ & x^2-2x-1 \quad \longleftarrow N(x)-3x\cdot D(x) \\ & \underline{x^2-3x+2} \quad \longleftarrow 1\cdot D(x) \\ & x-3 \quad \longleftarrow R(x)=N(x)-(3x+1)D(x) \end{array}$$

This is the end of Step 1. Oof! You should definitely practice this step.

- *Step 2.* The second step is to factor the denominator

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

We already did this in Example 1.10.1.

- *Step 3.* The third step is to write $\frac{x-3}{x^2-3x+2}$ in the form

$$\frac{x-3}{x^2-3x+2}=\frac{A}{x-1}+\frac{B}{x-2}$$

for some constants A and B. We already did this in Example 1.10.1. We found $A=2$ and $B=-1$.

- *Step 4.* The final step is to integrate.

$$\begin{aligned}\int\frac{3x^3-8x^2+4x-1}{x^2-3x+2}\mathrm{d}x &= \int\big[3x+1\big]\mathrm{d}x+\int\frac{2}{x-1}\mathrm{d}x+\int\frac{-1}{x-2}\mathrm{d}x \\ &= \frac{3}{2}x^2+x+2\log|x-1|-\log|x-2|+C\end{aligned}$$

You can see that the integration step is quite quick — almost all the work is in preparing the integrand.

Example 1.10.2

Here is a very solid example. It is quite long and the steps are involved. However please persist. No single step is too difficult.

Example 1.10.3 $\left(\int\frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5}\mathrm{d}x\right)$

In this example, we integrate $\frac{N(x)}{D(x)}=\frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5}$.

Solution.

- *Step 1.* Again, we start by comparing the degrees of the numerator and denominator. In this example, the numerator, $x^4+5x^3+16x^2+26x+22$, has degree four and the

denominator, x^3+3x^2+7x+5, has degree three. As $4 \geqslant 3$, we must execute the first step, which is to write $\frac{N(x)}{D(x)}$ in the form

$$\frac{N(x)}{D(x)} = P(x) + \frac{R(x)}{D(x)}$$

with $P(x)$ being a polynomial and $R(x)$ being a polynomial of degree strictly smaller than the degree of $D(x)$. This step is accomplished by long division, just as we did in Example 1.10.2. We'll go through the whole process in detail again.

Actually — before you read on ahead, please have a go at the long division. It is good practice.

– We start by observing that to get from the highest degree term in the denominator (x^3) to the highest degree term in the numerator (x^4), we have to multiply by x. So we write,

$$\begin{array}{r} x \\ \hline x^3+3x^2+7x+5 \,\big|\, x^4+5x^3+16x^2+26x+22 \end{array}$$

– Now we subtract x times the denominator x^3+3x^2+7x+5, which is $x^4+3x^3+7x^2+5x$, from the numerator.

$$\begin{array}{rl} & x \\ \hline x^3+3x^2+7x+5 \,\big|\, & x^4+5x^3+16x^2+26x+22 \\ & \underline{x^4+3x^3+\ 7x^2+\ 5x} \quad \longleftarrow x(x^3+3x^2+7x+5) \\ & \quad 2x^3+\ 9x^2+21x+22 \end{array}$$

– The remainder was $2x^3+9x^2+21x+22$. To get from the highest degree term in the denominator (x^3) to the highest degree term in the remainder ($2x^3$), we have to multiply by 2. So we write,

$$\begin{array}{rl} & x+\ 2 \\ \hline x^3+3x^2+7x+5 \,\big|\, & x^4+5x^3+16x^2+26x+22 \\ & \underline{x^4+3x^3+\ 7x^2+\ 5x} \\ & \quad 2x^3+\ 9x^2+21x+22 \end{array}$$

– Now we subtract 2 times the denominator x^3+3x^2+7x+5, which is $2x^3+6x^2+14x+10$, from the remainder.

$$\begin{array}{rl} & x+\ 2 \\ \hline x^3+3x^2+7x+5 \,\big|\, & x^4+5x^3+16x^2+26x+22 \\ & \underline{x^4+3x^3+\ 7x^2+\ 5x} \quad \longleftarrow x(x^3+3x^2+7x+5) \\ & \quad 2x^3+\ 9x^2+21x+22 \\ & \quad \underline{2x^3+\ 6x^2+14x+10} \quad \longleftarrow 2(x^3+3x^2+7x+5) \\ & \qquad\quad 3x^2+\ 7x+12 \end{array}$$

– This leaves a remainder of $3x^2+7x+12$. Because the remainder has degree 2, which is smaller than the degree of the denominator, which is 3, we stop.

– In this example, when we subtracted $x(x^3+3x^2+7x+5)$ and $2(x^3+3x^2+7x+5)$ from $x^4+5x^3+16x^2+26x+22$ we ended up with $3x^2+7x+12$. That is,

$$x^4+5x^3+16x^2+26x+22 \;-\; x(x^3+3x^2+7x+5) \;-\; 2(x^3+3x^2+7x+5) = 3x^2+7x+12$$

or, collecting the two terms proportional to (x^3+3x^2+7x+5)

$$x^4+5x^3+16x^2+26x+22 \;-\; (x+2)(x^3+3x^2+7x+5) \;=\; 3x^2+7x+12$$

Moving the $(x+2)(x^3+3x^2+7x+5)$ to the right hand side and dividing the whole equation by x^3+3x^2+7x+5 gives

$$\frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5} = x+2+\frac{3x^2+7x+12}{x^3+3x^2+7x+5}$$

This is of the form $\frac{N(x)}{D(x)} = P(x)+\frac{R(x)}{D(x)}$, with the degree of $R(x)$ strictly smaller than the degree of $D(x)$, which is what we wanted. Observe, once again, that $R(x)$ is the final remainder of the long division procedure and $P(x)$ is at the top of the long division computation.

$$\begin{array}{rl} & x+\;2 \;\longleftarrow\; P(x) \\ x^3+3x^2+7x+5 \;\big|\; & \overline{x^4+5x^3+16x^2+26x+22} \\ & \underline{x^4+3x^3+\;7x^2+\;5x} \\ & \quad 2x^3+\;9x^2+21x+22 \\ & \quad \underline{2x^3+\;6x^2+14x+10} \\ & \quad\quad 3x^2+\;7x+12 \;\longleftarrow\; R(x) \end{array}$$

- *Step 2.* The second step is to factor the denominator $D(x) = x^3+3x^2+7x+5$. In the "real world" factorisation of polynomials is often very hard. Fortunately[63], this is not the "real world" and there is a trick available to help us find this factorisation. The reader should take some time to look at Appendix A.16 before proceeding.

  – The trick exploits the fact that most polynomials that appear in homework assignments and on tests have integer coefficients and some integer roots. Any integer root of a polynomial that has integer coefficients, like $D(x) = x^3+3x^2+7x+5$, must divide the constant term of the polynomial exactly. Why this is true is explained[64] in Appendix A.16.

  – So any integer root of x^3+3x^2+7x+5 must divide 5 exactly. Thus the only integers which can be roots of $D(x)$ are ±1 and ±5. Of course, not all of these give roots of the polynomial — in fact there is no guarantee that any of them will be. We have to test each one.

63 One does not typically think of mathematics assignments or exams as nice kind places…The polynomials that appear in the "real world" are not so forgiving. Nature, red in tooth and claw — to quote Tennyson inappropriately (especially when this author doesn't know any other words from the poem).

64 Appendix A.16 contains several simple tricks for factoring polynomials. We recommend that you have a look at them.

– To test if $+1$ is a root, we sub $x = 1$ into $D(x)$:

$$D(1) = 1^3 + 3(1)^2 + 7(1) + 5 = 16$$

As $D(1) \neq 0$, 1 is not a root of $D(x)$.

– To test if -1 is a root, we sub it into $D(x)$:

$$D(-1) = (-1)^3 + 3(-1)^2 + 7(-1) + 5 = -1 + 3 - 7 + 5 = 0$$

As $D(-1) = 0$, -1 is a root of $D(x)$. As -1 is a root of $D(x)$, $\big(x - (-1)\big) = (x + 1)$ must factor $D(x)$ exactly. We can factor the $(x + 1)$ out of $D(x) = x^3 + 3x^2 + 7x + 5$ by long division once again.

– Dividing $D(x)$ by $(x + 1)$ gives:

$$\begin{array}{rl} & x^2+\ 2x+\ \ 5 \\ x+1\ \big| & x^3+3x^2+7x+5 \\ & \underline{x^3+\ \ x^2} \qquad\qquad \longleftarrow x^2(x+1) \\ & \quad 2x^2+7x+5 \\ & \quad \underline{2x^2+2x} \qquad\ \longleftarrow 2x(x+1) \\ & \qquad\ 5x+5 \\ & \qquad\ \underline{5x+5} \ \longleftarrow\ 5(x+1) \\ & \qquad\ \ 0 \end{array}$$

This time, when we subtracted $x^2(x + 1)$ and $2x(x + 1)$ and $5(x + 1)$ from $x^3 + 3x^2 + 7x + 5$ we ended up with 0 — as we knew would happen, because we knew that $x + 1$ divides $x^3 + 3x^2 + 7x + 5$ exactly. Hence

$$x^3 + 3x^2 + 7x + 5 \ - \ x^2(x+1) \ - \ 2x(x+1) \ - \ 5(x+1) \ = \ 0$$

or

$$x^3 + 3x^2 + 7x + 5 \ = \ x^2(x+1) \ + \ 2x(x+1) \ + \ 5(x+1)$$

or

$$x^3 + 3x^2 + 7x + 5 = (x^2 + 2x + 5)(x + 1)$$

– It isn't quite time to stop yet; we should attempt to factor the quadratic factor, $x^2 + 2x + 5$. We can use the quadratic formula[65] to find the roots of $x^2 + 2x + 5$:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-2 \pm \sqrt{4 - 20}}{2} = \frac{-2 \pm \sqrt{-16}}{2}$$

Since this expression contains the square root of a negative number the equation $x^2 + 2x + 5 \ = \ 0$ has no real solutions; without the use of complex numbers, $x^2 + 2x + 5$ cannot be factored.

65 To be precise, the quadratic equation $ax^2 + bx + c = 0$ has solutions

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

The term $b^2 - 4ac$ is called the discriminant and it tells us about the number of solutions. If the discriminant is positive then there are two real solutions. When it is zero, there is a single solution. And if it is negative, there is no real solutions (you need complex numbers to say more than this).

We have reached the end of step 2. At this point we have

$$\frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5} = x+2+\frac{3x^2+7x+12}{(x+1)(x^2+2x+5)}$$

- *Step 3.* The third step is to write $\frac{3x^2+7x+12}{(x+1)(x^2+2x+5)}$ in the form

$$\frac{3x^2+7x+12}{(x+1)(x^2+2x+5)} = \frac{A}{x+1}+\frac{Bx+C}{x^2+2x+5}$$

for some constants A, B and C.

Note that the numerator, $Bx+C$ of the second term on the right hand side is not just a constant. It is of degree one, which is exactly one smaller than the degree of the denominator, x^2+2x+5. More generally, if the denominator consists of n different linear factors and m different quadratic factors, then we decompose the ratio as

$$\begin{aligned}\text{rational function} = &\frac{A_1}{\text{linear factor 1}}+\frac{A_2}{\text{linear factor 2}}+\cdots+\frac{A_n}{\text{linear factor n}}\\ &+\frac{B_1x+C_1}{\text{quadratic factor 1}}+\frac{B_2x+C_2}{\text{quadratic factor 2}}+\cdots+\frac{B_mx+C_m}{\text{quadratic factor m}}\end{aligned}$$

To determine the values of the constants A, B, C, we put the right hand side back over the common denominator $(x+1)(x^2+2x+5)$.

$$\frac{3x^2+7x+12}{(x+1)(x^2+2x+5)} = \frac{A}{x+1}+\frac{Bx+C}{x^2+2x+5} = \frac{A(x^2+2x+5)+(Bx+C)(x+1)}{(x+1)(x^2+2x+5)}$$

The fraction on the far left is the same as the fraction on the far right if and only if their numerators are the same.

$$3x^2+7x+12 = A(x^2+2x+5)+(Bx+C)(x+1)$$

Again, as in Example 1.10.1, there are a couple of different ways to determine the values of A, B and C from this equation.

- *Step 3 – Algebra Method.* The conceptually clearest procedure is to write the right hand side as a polynomial in standard form (i.e. collect up all x^2 terms, all x terms and all constant terms)

$$3x^2+7x+12 = (A+B)x^2+(2A+B+C)x+(5A+C)$$

For these two polynomials to be the same, the coefficient of x^2 on the left hand side and the coefficient of x^2 on the right hand side must be the same. Similarly the coefficients of x^1 must match and the coefficients of x^0 must match.

This gives us a system of three equations

$$A+B=3 \qquad 2A+B+C=7 \qquad 5A+C=12$$

in the three unknowns A, B, C. We can solve this system by

- using the first equation, namely $A + B = 3$, to determine A in terms of B: $A = 3 - B$.
- Substituting this into the remaining two equations eliminates the A's from these two equations, leaving two equations in the two unknowns B and C.

$$\begin{array}{rrr} A = 3 - B & 2A + B + C = 7 & 5A + C = 12 \\ \Rightarrow & 2(3 - B) + B + C = 7 & 5(3 - B) + C = 12 \\ \Rightarrow & -B + C = 1 & -5B + C = -3 \end{array}$$

- Now we can use the equation $-B + C = 1$, to determine B in terms of C: $B = C - 1$.
- Substituting this into the remaining equation eliminates the B's leaving an equation in the one unknown C, which is easy to solve.

$$\begin{array}{rrr} B = C - 1 & -5B + C = -3 \\ \Rightarrow & -5(C - 1) + C = -3 \\ \Rightarrow & -4C = -8 \end{array}$$

- So $C = 2$, and then $B = C - 1 = 1$, and then $A = 3 - B = 2$. Hence

$$\frac{3x^2 + 7x + 12}{(x+1)(x^2 + 2x + 5)} = \frac{2}{x+1} + \frac{x+2}{x^2 + 2x + 5}$$

- *Step 3 – Sneaky Method.* While the above method is transparent, it is rather tedious. It is arguably better to use the second, sneakier and more efficient, procedure. In order for

$$3x^2 + 7x + 12 = A(x^2 + 2x + 5) + (Bx + C)(x + 1)$$

the equation must hold for all values of x.

- In particular, it must be true for $x = -1$. When $x = -1$, the factor $(x + 1)$ multiplying $Bx + C$ is exactly zero. So B and C disappear from the equation, leaving us with an easy equation to solve for A:

$$3x^2 + 7x + 12\Big|_{x=-1} = \Big[A(x^2 + 2x + 5) + (Bx + C)(x + 1)\Big]_{x=-1} \Longrightarrow 8 = 4A \Longrightarrow A = 2$$

- Sub this value of A back in and simplify.

$$\begin{aligned} 3x^2 + 7x + 12 &= 2(x^2 + 2x + 5) + (Bx + C)(x + 1) \\ x^2 + 3x + 2 &= (Bx + C)(x + 1) \end{aligned}$$

Since $(x + 1)$ is a factor on the right hand side, it must also be a factor on the left hand side.

$$(x+2)(x+1) = (Bx + C)(x + 1) \quad \Rightarrow \quad (x + 2) = (Bx + C) \quad \Rightarrow \quad B = 1,\ C = 2$$

So again we find that

$$\frac{3x^2+7x+12}{(x+1)(x^2+2x+5)} = \frac{2}{x+1} + \frac{x+2}{x^2+2x+5} \checkmark$$

Thus our integrand can be written as

$$\frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5} = x+2+\frac{2}{x+1}+\frac{x+2}{x^2+2x+5}.$$

- *Step 4.* Now we can finally integrate! The first two pieces are easy.

$$\int (x+2)\mathrm{d}x = \tfrac{1}{2}x^2+2x \qquad \int \frac{2}{x+1}\mathrm{d}x = 2\log|x+1|$$

(We're leaving the arbitrary constant to the end of the computation.)

The final piece is a little harder. The idea is to complete the square[66] in the denominator

$$\frac{x+2}{x^2+2x+5} = \frac{x+2}{(x+1)^2+4}$$

and then make a change of variables to make the fraction look like $\frac{ay+b}{y^2+1}$. In this case

$$\frac{x+2}{(x+1)^2+4} = \frac{1}{4}\frac{x+2}{(\frac{x+1}{2})^2+1}$$

so we make the change of variables $y = \frac{x+1}{2}, \mathrm{d}y = \frac{\mathrm{d}x}{2},\ x = 2y-1, \mathrm{d}x = 2\,\mathrm{d}y$

$$\begin{aligned}\int \frac{x+2}{(x+1)^2+4}\mathrm{d}x &= \frac{1}{4}\int \frac{x+2}{(\frac{x+1}{2})^2+1}\mathrm{d}x \\ &= \frac{1}{4}\int \frac{(2y-1)+2}{y^2+1}\,2\,\mathrm{d}y \;=\; \frac{1}{2}\int \frac{2y+1}{y^2+1}\,\mathrm{d}y \\ &= \int \frac{y}{y^2+1}\,\mathrm{d}y + \frac{1}{2}\int \frac{1}{y^2+1}\,\mathrm{d}y\end{aligned}$$

66 This same idea arose in Section 1.9. Given a quadratic written as

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

rewrite it as

$$Q(x) = a(x+d)^2+e.$$

We can determine d and e by expanding and comparing coefficients of x:

$$ax^2+bx+c = a(x^2+2dx+d^2)+e = ax^2+2dax+(e+ad^2)$$

Hence $d = b/2a$ and $e = c-ad^2$.

Both integrals are easily evaluated, using the substitution $u = y^2 + 1$, $\mathrm{d}u = 2y\,\mathrm{d}y$ for the first.

$$\int \frac{y}{y^2+1}\,\mathrm{d}y \;=\; \int \frac{1}{u}\frac{\mathrm{d}u}{2} \;=\; \frac{1}{2}\log|u| = \frac{1}{2}\log(y^2+1) \;=\; \frac{1}{2}\log\left[\left(\frac{x+1}{2}\right)^2+1\right]$$

$$\frac{1}{2}\int \frac{1}{y^2+1}\,\mathrm{d}y \;=\; \frac{1}{2}\arctan y \;=\; \frac{1}{2}\arctan\left(\frac{x+1}{2}\right)$$

That's finally it. Putting all of the pieces together

$$\int \frac{x^4+5x^3+16x^2+26x+22}{x^3+3x^2+7x+5}\mathrm{d}x = \frac{1}{2}x^2+2x+2\log|x+1| + \frac{1}{2}\log\left[\left(\frac{x+1}{2}\right)^2+1\right] + \frac{1}{2}\arctan\left(\frac{x+1}{2}\right) + C$$

Example 1.10.3

The best thing after working through a few a nice long examples is to do another nice long example — it is excellent practice[67]. We recommend that the reader attempt the problem before reading through our solution.

Example 1.10.4 $\left(\int \frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4}\mathrm{d}x\right)$

In this example, we integrate $\frac{N(x)}{D(x)} = \frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4}$.

- *Step 1.* The degree of the numerator $N(x)$ is equal to the degree of the denominator $D(x)$, so the first step to write $\frac{N(x)}{D(x)}$ in the form

$$\frac{N(x)}{D(x)} = P(x) + \frac{R(x)}{D(x)}$$

with $P(x)$ being a polynomial (which should be of degree 0, i.e. just a constant) and $R(x)$ being a polynomial of degree strictly smaller than the degree of $D(x)$. By long division

$$\begin{array}{r} 4 \\ x^3+5x^2+8x+4 \,\overline{)\, 4x^3+23x^2+45x+27} \\ \underline{4x^3+20x^2+32x+16} \\ 3x^2+13x+11 \end{array}$$

67 At the risk of quoting Nietzsche, "That which does not kill us makes us stronger." Though this author always preferred the logically equivalent contrapositive — "That which does not make us stronger will kill us." However no one is likely to be injured by practicing partial fractions or looking up quotes on Wikipedia. Its also a good excuse to remind yourself of what a contrapositive is — though we will likely look at them again when we get to sequences and series.

so

$$\frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4} = 4 + \frac{3x^2+13x+11}{x^3+5x^2+8x+4}$$

- *Step 2.* The second step is to factorise $D(x) = x^3+5x^2+8x+4$.
  - To start, we'll try and guess an integer root. Any integer root of $D(x)$ must divide the constant term, 4, exactly. Only ±1, ±2, ±4 can be integer roots of x^3+5x^2+8x+4.
  - We test to see if ±1 are roots.

$$\begin{aligned} D(1) = (1)^3+5(1)^2+8(1)+4 \neq 0 \quad &\Rightarrow \quad x=1 \text{ is not a root} \\ D(-1) = (-1)^3+5(-1)^2+8(-1)+4 = 0 \quad &\Rightarrow \quad x=-1 \text{ is a root} \end{aligned}$$

    So $(x+1)$ must divide x^3+5x^2+8x+4 exactly.
  - By long division

$$\begin{array}{r} x^2+4x+4 \\ x+1 \overline{)\, x^3+5x^2+8x+4} \\ \underline{x^3+x^2} \\ 4x^2+8x+4 \\ \underline{4x^2+4x} \\ 4x+4 \\ \underline{4x+4} \\ 0 \end{array}$$

    so

$$x^3+5x^2+8x+4 = (x+1)(x^2+4x+4) = (x+1)(x+2)(x+2)$$

  - Notice that we could have instead checked whether or not ±2 are roots

$$\begin{aligned} D(2) = (2)^3+5(2)^2+8(2)+4 \neq 0 \quad &\Rightarrow \quad x=2 \text{ is not a root} \\ D(-2) = (-2)^3+5(-2)^2+8(-2)+4 = 0 \quad &\Rightarrow \quad x=-2 \text{ is a root} \end{aligned}$$

    We now know that both -1 and -2 are roots of x^3+5x^2+8x+4 and hence both $(x+1)$ and $(x+2)$ are factors of x^3+5x^2+8x+4. Because x^3+5x^2+8x+4 is of degree three and the coefficient of x^3 is 1, we must have $x^3+5x^2+8x+4 = (x+1)(x+2)(x+a)$ for some constant a. Multiplying out the right hand side shows that the constant term is $2a$. So $2a=4$ and $a=2$.

This is the end of step 2. We now know that

$$\frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4} \;=\; 4 + \frac{3x^2+13x+11}{(x+1)(x+2)^2}$$

- *Step 3.* The third step is to write $\frac{3x^2+13x+11}{(x+1)(x+2)^2}$ in the form

$$\frac{3x^2+13x+11}{(x+1)(x+2)^2} = \frac{A}{x+1} + \frac{B}{x+2} + \frac{C}{(x+2)^2}$$

for some constants A, B and C.

Note that there are two terms on the right hand arising from the factor $(x+2)^2$. One has denominator $(x+2)$ and one has denominator $(x+2)^2$. More generally, for each factor $(x+a)^n$ in the denominator of the rational function on the left hand side, we include

$$\frac{A_1}{x+a} + \frac{A_2}{(x+a)^2} + \cdots + \frac{A_n}{(x+a)^n}$$

in the partial fraction decomposition on the right hand side[68].

To determine the values of the constants A, B, C, we put the right hand side back over the common denominator $(x+1)(x+2)^2$.

$$\begin{aligned}\frac{3x^2+13x+11}{(x+1)(x+2)^2} &= \frac{A}{x+1} + \frac{B}{x+2} + \frac{C}{(x+2)^2} \\ &= \frac{A(x+2)^2 + B(x+1)(x+2) + C(x+1)}{(x+1)(x+2)^2}\end{aligned}$$

The fraction on the far left is the same as the fraction on the far right if and only if their numerators are the same.

$$3x^2+13x+11 = A(x+2)^2 + B(x+1)(x+2) + C(x+1)$$

As in the previous examples, there are a couple of different ways to determine the values of A, B and C from this equation.

- *Step 3 – Algebra Method.* The conceptually clearest procedure is to write the right hand side as a polynomial in standard form (i.e. collect up all x^2 terms, all x terms and all constant terms)

$$3x^2+13x+11 = (A+B)x^2 + (4A+3B+C)x + (4A+2B+C)$$

For these two polynomials to be the same, the coefficient of x^2 on the left hand side and the coefficient of x^2 on the right hand side must be the same. Similarly the coefficients of x^1 and the coefficients of x^0 (i.e. the constant terms) must match. This gives us a system of three equations,

$$A+B=3 \qquad 4A+3B+C=13 \qquad 4A+2B+C=11$$

in the three unknowns A, B, C. We can solve this system by

  - using the first equation, namely $A+B=3$, to determine A in terms of B: $A=3-B$.

68 This is justified in the (optional) subsection "Justification of the Partial Fraction Decompositions" below.

– Substituting this into the remaining equations eliminates the A, leaving two equations in the two unknown B, C.

$$4(3-B)+3B+C=13 \qquad 4(3-B)+2B+C=11$$

or

$$-B+C=1 \qquad -2B+C=-1$$

– We can now solve the first of these equations, namely $-B+C=1$, for B in terms of C, giving $B=C-1$.

– Substituting this into the last equation, namely $-2B+C=-1$, gives $-2(C-1)+C=-1$ which is easily solved to give

– $C=3$, and then $B=C-1=2$ and then $A=3-B=1$.

Hence

$$\frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4}=4+\frac{3x^2+13x+11}{(x+1)(x+2)^2}=4+\frac{1}{x+1}+\frac{2}{x+2}+\frac{3}{(x+2)^2}$$

- *Step 3 – Sneaky Method.* The second, sneakier, method for finding A, B and C exploits the fact that $3x^2+13x+11=A(x+2)^2+B(x+1)(x+2)+C(x+1)$ must be true for all values of x. In particular, it must be true for $x=-1$. When $x=-1$, the factor $(x+1)$ multiplying B and C is exactly zero. So B and C disappear from the equation, leaving us with an easy equation to solve for A:

$$\begin{aligned} 3x^2+13x+11\Big|_{x=-1} &= \Big[A(x+2)^2+B(x+1)(x+2)+C(x+1)\Big]_{x=-1} \\ \Longrightarrow 1&=A \end{aligned}$$

Sub this value of A back in and simplify.

$$\begin{aligned} 3x^2+13x+11&=(1)(x+2)^2+B(x+1)(x+2)+C(x+1) \\ 2x^2+9x+7&=B(x+1)(x+2)+C(x+1)=(xB+2B+C)(x+1) \end{aligned}$$

Since $(x+1)$ is a factor on the right hand side, it must also be a factor on the left hand side.

$$(2x+7)(x+1)=(xB+2B+C)(x+1) \quad \Rightarrow \quad (2x+7)=(xB+2B+C)$$

For the coefficients of x to match, B must be 2. For the constant terms to match, $2B+C$ must be 7, so C must be 3. Hence we again have

$$\frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4}=4+\frac{3x^2+13x+11}{(x+1)(x+2)^2}=4+\frac{1}{x+1}+\frac{2}{x+2}+\frac{3}{(x+2)^2}$$

- *Step 4.* The final step is to integrate

$$\begin{aligned} \int\frac{4x^3+23x^2+45x+27}{x^3+5x^2+8x+4}\mathrm{d}x &= \int 4\mathrm{d}x+\int\frac{1}{x+1}\mathrm{d}x+\int\frac{2}{x+2}\mathrm{d}x+\int\frac{3}{(x+2)^2}\mathrm{d}x \\ &= 4x+\log|x+1|+2\log|x+2|-\frac{3}{x+2}+C \end{aligned}$$

Example 1.10.4

The method of partial fractions is not just confined to the problem of integrating rational functions. There are other integrals — such as $\int \sec x \mathrm{d}x$ and $\int \sec^3 x \mathrm{d}x$ — that can be transformed (via substitutions) into integrals of rational functions. We encountered both of these integrals in Sections 1.8 and 1.9 on trigonometric integrals and substitutions.

Example 1.10.5 ($\int \sec x \mathrm{d}x$)

Solution. In this example, we integrate $\sec x$. It is not yet clear what this integral has to do with partial fractions. To get to a partial fractions computation, we first make one of our old substitutions.

$$\begin{aligned}
\int \sec x \mathrm{d}x &= \int \frac{1}{\cos x} \mathrm{d}x && \text{massage the expression a little} \\
&= \int \frac{\cos x}{\cos^2 x} \mathrm{d}x && \text{substitute } u = \sin x, \mathrm{d}u = \cos x \mathrm{d}x \\
&= -\int \frac{\mathrm{d}u}{u^2 - 1} && \text{and use } \cos^2 x = 1 - \sin^2 x = 1 - u^2
\end{aligned}$$

So we now have to integrate $\frac{1}{u^2-1}$, which is a rational function of u, and so is perfect for partial fractions.

- *Step 1.* The degree of the numerator, 1, is zero, which is strictly smaller than the degree of the denominator, $u^2 - 1$, which is two. So the first step is skipped.
- *Step 2.* The second step is to factor the denominator:

$$u^2 - 1 = (u - 1)(u + 1)$$

- *Step 3.* The third step is to write $\frac{1}{u^2-1}$ in the form

$$\frac{1}{u^2 - 1} = \frac{1}{(u - 1)(u + 1)} = \frac{A}{u - 1} + \frac{B}{u + 1}$$

  for some constants A and B.

- *Step 3 – Sneaky Method.*
  - Multiply through by the denominator to get

$$1 = A(u + 1) + B(u - 1)$$

    This equation must be true for all u.
  - If we now set $u = 1$ then we eliminate B from the equation leaving us with

$$1 = 2A \qquad \text{so } A = 1/2.$$

– Similarly, if we set $u = -1$ then we eliminate A, leaving

$$1 = -2B \qquad \text{which implies } B = -1/2.$$

We have now found that $A = 1/2, B = -1/2$, so

$$\frac{1}{u^2-1} = \frac{1}{2}\Big[\frac{1}{u-1} - \frac{1}{u+1}\Big].$$

- It is always a good idea to check our work.

$$\frac{1/2}{u-1} + \frac{-1/2}{u+1} = \frac{1/2(u+1) - 1/2(u-1)}{(u-1)(u+1)} = \frac{1}{(u-1)(u+1)} \checkmark$$

- *Step 4.* The final step is to integrate.

$$\begin{aligned}
\int \sec x \mathrm{d}x &= -\int \frac{\mathrm{d}u}{u^2-1} && \text{after substitution} \\
&= -\frac{1}{2}\int \frac{\mathrm{d}u}{u-1} + \frac{1}{2}\int \frac{\mathrm{d}u}{u+1} && \text{partial fractions} \\
&= -\frac{1}{2}\log|u-1| + \frac{1}{2}\log|u+1| + C \\
&= -\frac{1}{2}\log|\sin(x)-1| + \frac{1}{2}\log|\sin(x)+1| + C && \text{rearrange a little} \\
&= \frac{1}{2}\log\left|\frac{1+\sin x}{1-\sin x}\right| + C
\end{aligned}$$

Notice that since $-1 \leqslant \sin x \leqslant 1$, we are free to drop the absolute values in the last line if we wish.

Example 1.10.5

Another example in the same spirit, though a touch harder. Again, we saw this problem in Section 1.8 and 1.9.

Example 1.10.6 ($\int \sec^3 x \mathrm{d}x$)

Solution.

- We'll start by converting it into the integral of a rational function using the substitution $u = \sin x$, $\mathrm{d}u = \cos x \mathrm{d}x$.

$$\begin{aligned}
\int \sec^3 x \mathrm{d}x &= \int \frac{1}{\cos^3 x}\mathrm{d}x && \text{massage this a little} \\
&= \int \frac{\cos x}{\cos^4 x}\mathrm{d}x && \text{replace } \cos^2 x = 1 - \sin^2 x = 1 - u^2 \\
&= \int \frac{\cos x \mathrm{d}x}{\left[1-\sin^2 x\right]^2} \\
&= \int \frac{\mathrm{d}u}{[1-u^2]^2}
\end{aligned}$$

- We could now find the partial fraction decomposition of the integrand $\frac{1}{[1-u^2]^2}$ by executing the usual four steps. But it is easier to use

$$\frac{1}{u^2-1} = \frac{1}{2}\Big[\frac{1}{u-1} - \frac{1}{u+1}\Big]$$

which we worked out in Example 1.10.5 above.

- Squaring this gives

$$\begin{aligned}\frac{1}{[1-u^2]^2} &= \frac{1}{4}\Big[\frac{1}{u-1} - \frac{1}{u+1}\Big]^2 \\ &= \frac{1}{4}\Big[\frac{1}{(u-1)^2} - \frac{2}{(u-1)(u+1)} + \frac{1}{(u+1)^2}\Big] \\ &= \frac{1}{4}\Big[\frac{1}{(u-1)^2} - \frac{1}{u-1} + \frac{1}{u+1} + \frac{1}{(u+1)^2}\Big]\end{aligned}$$

where we have again used $\frac{1}{u^2-1} = \frac{1}{2}\left[\frac{1}{u-1} - \frac{1}{u+1}\right]$ in the last step.

- It only remains to do the integrals and simplify.

$$\begin{aligned}\int \sec^3 x\,dx &= \frac{1}{4}\int \Big[\frac{1}{(u-1)^2} - \frac{1}{u-1} + \frac{1}{u+1} + \frac{1}{(u+1)^2}\Big]du \\ &= \frac{1}{4}\Big[-\frac{1}{u-1} - \log|u-1| + \log|u+1| - \frac{1}{u+1}\Big] + C && \text{group carefully} \\ &= \frac{-1}{4}\Big[\frac{1}{u-1} + \frac{1}{u+1}\Big] + \frac{1}{4}\Big[\log|u+1| - \log|u-1|\Big] + C && \text{sum carefully} \\ &= -\frac{1}{4}\frac{2u}{u^2-1} + \frac{1}{4}\log\Big|\frac{u+1}{u-1}\Big| + C && \text{clean up} \\ &= \frac{1}{2}\frac{u}{1-u^2} + \frac{1}{4}\log\Big|\frac{u+1}{u-1}\Big| + C && \text{put } u = \sin x \\ &= \frac{1}{2}\frac{\sin x}{\cos^2 x} + \frac{1}{4}\log\Big|\frac{\sin x+1}{\sin x-1}\Big| + C\end{aligned}$$

Example 1.10.6

1.10.2 ▸▸ The Form of Partial Fraction Decompositions

In the examples above we used the partial fractions method to decompose rational functions into easily integrated pieces. Each of those examples was quite involved and we had to spend quite a bit of time factoring and doing long division. The key step in each of the computations was Step 3 — in that step we decomposed the rational function $\frac{N(x)}{D(x)}$ (or $\frac{R(x)}{D(x)}$), for which the degree of the numerator is strictly smaller than the degree of the denominator, into a sum of particularly simple rational functions, like $\frac{A}{x-a}$. We did not, however, give a systematic description of those decompositions.

In this subsection we fill that gap by describing the general[69] form of partial fraction decompositions. The justification of these forms is not part of the course, but the interested reader is invited to read the next (optional) subsection where such justification is given. In the following it is assumed that

- $N(x)$ and $D(x)$ are polynomials with the degree of $N(x)$ strictly smaller than the degree of $D(x)$.
- K is a constant.
- $a_1, a_2, \cdots, a_j$ are all different numbers.
- $m_1, m_2, \cdots, m_j$, and $n_1, n_2, \cdots, n_k$ are all strictly positive integers.
- $x^2 + b_1 x + c_1$, $x^2 + b_2 x + c_2$, $\cdots$, $x^2 + b_k x + c_k$ are all different.

▸▸▸ Simple Linear Factor Case

If the denominator $D(x) = K(x - a_1)(x - a_2)\cdots(x - a_j)$ is a product of j different linear factors, then

Equation 1.10.7.

$$\frac{N(x)}{D(x)} = \frac{A_1}{x - a_1} + \frac{A_2}{x - a_2} + \cdots + \frac{A_j}{x - a_j}$$

We can then integrate each term

$$\int \frac{A}{x - a} \mathrm{d}x = A \log|x - a| + C.$$

▸▸▸ General Linear Factor Case

If the denominator $D(x) = K(x - a_1)^{m_1}(x - a_2)^{m_2}\cdots(x - a_j)^{m_j}$ then

Equation 1.10.8.

$$\begin{aligned}\frac{N(x)}{D(x)} = &\frac{A_{1,1}}{x - a_1} + \frac{A_{1,2}}{(x - a_1)^2} + \cdots + \frac{A_{1,m_1}}{(x - a_1)^{m_1}} \\ &+ \frac{A_{2,1}}{x - a_2} + \frac{A_{2,2}}{(x - a_2)^2} + \cdots + \frac{A_{2,m_2}}{(x - a_2)^{m_2}} + \cdots \\ &+ \frac{A_{j,1}}{x - a_j} + \frac{A_{j,2}}{(x - a_j)^2} + \cdots + \frac{A_{j,m_j}}{(x - a_j)^{m_j}}\end{aligned}$$

69 Well — not the completely general form, in the sense that we are not allowing the use of complex numbers. As a result we have to use both linear and quadratic factors in the denominator. If we could use complex numbers we would be able to restrict ourselves to linear factors.

Notice that we could rewrite each line as

$$\frac{A_1}{x-a}+\frac{A_2}{(x-a)^2}+\cdots+\frac{A_m}{(x-a)^m}=\frac{A_1(x-a)^{m-1}+A_2(x-a)^{m-2}+\cdots+A_m}{(x-a)^m}$$
$$=\frac{B_1x^{m-1}+B_2x^{m-2}+\cdots+B_m}{(x-a)^m}$$

which is a polynomial whose degree, $m-1$, is strictly smaller than that of the denominator $(x-a)^m$. But the form of Equation (1.10.8) is preferable because it is easier to integrate.

$$\int\frac{A}{x-a}\mathrm{d}x=A\log|x-a|+C$$
$$\int\frac{A}{(x-a)^k}\mathrm{d}x=-\frac{1}{k-1}\cdot\frac{A}{(x-a)^{k-1}}\qquad\text{provided } k>1.$$

▸▸▸ Simple Linear and Quadratic Factor Case

If $D(x)=K(x-a_1)\cdots(x-a_j)(x^2+b_1x+c_1)\cdots(x^2+b_kx+c_k)$ then

Equation 1.10.9.

$$\frac{N(x)}{D(x)}=\frac{A_1}{x-a_1}+\cdots+\frac{A_j}{x-a_j}+\frac{B_1x+C_1}{x^2+b_1x+c_1}+\cdots+\frac{B_kx+C_k}{x^2+b_kx+c_k}$$

Note that the numerator of each term on the right hand side has degree one smaller than the degree of the denominator.

The quadratic terms $\frac{Bx+C}{x^2+bx+c}$ are integrated in a two-step process that is best illustrated with a simple example (see also Example 1.10.3 above).

Example 1.10.10 $\left(\int\frac{2x+7}{x^2+4x+13}\mathrm{d}x\right)$

Solution.

- Start by completing the square in the denominator:

$$x^2+4x+13=(x+2)^2+9\qquad\text{and thus}$$
$$\frac{2x+7}{x^2+4x+13}=\frac{2x+7}{(x+2)^2+3^2}$$

- Now set $y=(x+2)/3, \mathrm{d}y=\frac{1}{3}\mathrm{d}x$, or equivalently $x=3y-2, \mathrm{d}x=3\mathrm{d}y$:

$$\int\frac{2x+7}{x^2+4x+13}\mathrm{d}x=\int\frac{2x+7}{(x+2)^2+3^2}\mathrm{d}x$$
$$=\int\frac{6y-4+7}{3^2y^2+3^2}\cdot3\mathrm{d}y$$
$$=\int\frac{6y+3}{3(y^2+1)}\mathrm{d}y$$
$$=\int\frac{2y+1}{y^2+1}\mathrm{d}y$$

Notice that we chose 3 in $y = (x+2)/3$ precisely to transform the denominator into the form y^2+1.

- Now almost always the numerator will be a linear polynomial of y and we decompose as follows

$$\begin{aligned}\int \frac{2x+7}{x^2+4x+13}\mathrm{d}x &= \int \frac{2y+1}{y^2+1}\mathrm{d}y \\ &= \int \frac{2y}{y^2+1}\mathrm{d}y + \int \frac{1}{y^2+1}\mathrm{d}y \\ &= \log|y^2+1| + \arctan y + C \\ &= \log\left|\left(\frac{x+2}{3}\right)^2+1\right| + \arctan\left(\frac{x+2}{3}\right) + C\end{aligned}$$

Example 1.10.10

▸▸▸ Optional — General Linear and Quadratic Factor Case

If $D(x) = K(x-a_1)^{m_1}\cdots(x-a_j)^{m_j}(x^2+b_1x+c_1)^{n_1}\cdots(x^2+b_kx+c_k)^{n_k}$ then

Equation 1.10.11.

$$\begin{aligned}\frac{N(x)}{D(x)} = &\frac{A_{1,1}}{x-a_1} + \frac{A_{1,2}}{(x-a_1)^2} + \cdots + \frac{A_{1,m_1}}{(x-a_1)^{m_1}} + \cdots \\ &+ \frac{A_{j,1}}{x-a_j} + \frac{A_{j,2}}{(x-a_j)^2} + \cdots + \frac{A_{j,m_j}}{(x-a_j)^{m_j}} \\ &+ \frac{B_{1,1}x+C_{1,1}}{x^2+b_1x+c_1} + \frac{B_{1,2}x+C_{1,2}}{(x^2+b_1x+c_1)^2} + \cdots + \frac{B_{1,n_1}x+C_{1,n_1}}{(x^2+b_1x+c_1)^{n_1}} + \cdots \\ &+ \frac{B_{k,1}x+C_{k,1}}{x^2+b_kx+c_k} + \frac{B_{k,2}x+C_{k,2}}{(x^2+b_kx+c_k)^2} + \cdots + \frac{B_{k,n_k}x+C_{k,n_k}}{(x^2+b_kx+c_k)^{n_k}}\end{aligned}$$

We have already seen how to integrate the simple and general linear terms, and the simple quadratic terms. Integrating general quadratic terms is not so straightforward.

Example 1.10.12 $\left(\int \frac{\mathrm{d}x}{(x^2+1)^n}\right)$

This example is not so easy, so it should definitely be considered optional.

Solution. In what follows write

$$I_n = \int \frac{\mathrm{d}x}{(x^2+1)^n}.$$

- When $n = 1$ we know that

$$\int \frac{\mathrm{d}x}{x^2+1} = \arctan x + C$$

- Now assume that $n > 1$, then

$$\begin{aligned}\int \frac{1}{(x^2+1)^n}\mathrm{d}x &= \int \frac{(x^2+1-x^2)}{(x^2+1)^n}\mathrm{d}x && \text{sneaky}\\ &= \int \frac{1}{(x^2+1)^{n-1}}\mathrm{d}x - \int \frac{x^2}{(x^2+1)^n}\mathrm{d}x\\ &= I_{n-1} - \int \frac{x^2}{(x^2+1)^n}\mathrm{d}x\end{aligned}$$

  So we can write I_n in terms of I_{n-1} and this second integral.

- We can use integration by parts to compute the second integral:

$$\int \frac{x^2}{(x^2+1)^n}\mathrm{d}x = \int \frac{x}{2}\cdot\frac{2x}{(x^2+1)^n}\mathrm{d}x \qquad \text{sneaky}$$

  We set $u = x/2$ and $\mathrm{d}v = \frac{2x}{(x^2+1)^n}\mathrm{d}x$, which gives $\mathrm{d}u = \frac{1}{2}\mathrm{d}x$ and $v = -\frac{1}{n-1}\cdot\frac{1}{(x^2+1)^{n-1}}$. You can check v by differentiating. Integration by parts gives

$$\begin{aligned}\int \frac{x}{2}\cdot\frac{2x}{(x^2+1)^n}\mathrm{d}x &= -\frac{x}{2(n-1)(x^2+1)^{n-1}} + \int \frac{\mathrm{d}x}{2(n-1)(x^2+1)^{n-1}}\\ &= -\frac{x}{2(n-1)(x^2+1)^{n-1}} + \frac{1}{2(n-1)}\cdot I_{n-1}\end{aligned}$$

- Now put everything together:

$$\begin{aligned}I_n &= \int \frac{1}{(x^2+1)^n}\mathrm{d}x\\ &= I_{n-1} + \frac{x}{2(n-1)(x^2+1)^{n-1}} - \frac{1}{2(n-1)}\cdot I_{n-1}\\ &= \frac{2n-3}{2(n-1)}I_{n-1} + \frac{x}{2(n-1)(x^2+1)^{n-1}}\end{aligned}$$

- We can then use this recurrence to write down I_n for the first few n:

$$\begin{aligned}I_2 &= \frac{1}{2}I_1 + \frac{x}{2(x^2+1)} + C\\ &= \frac{1}{2}\arctan x + \frac{x}{2(x^2+1)}\\ I_3 &= \frac{3}{4}I_2 + \frac{x}{4(x^2+1)^2}\\ &= \frac{3}{8}\arctan x + \frac{3x}{8(x^2+1)} + \frac{x}{4(x^2+1)^2} + C\\ I_4 &= \frac{5}{6}I_3 + \frac{x}{6(x^2+1)^3}\\ &= \frac{5}{16}\arctan x + \frac{5x}{16(x^2+1)} + \frac{5x}{24(x^2+1)^2} + \frac{x}{6(x^2+1)^3} + C\end{aligned}$$

  and so forth. You can see why partial fraction questions involving denominators with repeated quadratic factors do not often appear on exams.

Example 1.10.12

1.10.3 ▸▸ Optional — Justification of the Partial Fraction Decompositions

We will now see the justification for the form of the partial fraction decompositions. We start by considering the case in which the denominator has only linear factors. Then we'll consider the case in which quadratic factors are allowed too[70].

▸▸▸ The Simple Linear Factor Case

In the most common partial fraction decomposition, we split up

$$\frac{N(x)}{(x-a_1)\times\cdots\times(x-a_d)}$$

into a sum of the form

$$\frac{A_1}{x-a_1}+\cdots+\frac{A_d}{x-a_d}$$

We now show that this decomposition can always be achieved, under the assumptions that the a_i's are all different and $N(x)$ is a polynomial of degree at most $d-1$. To do so, we shall repeatedly apply the following Lemma.

Lemma 1.10.13.

Let $N(x)$ and $D(x)$ be polynomials of degree n and d respectively, with $n \leqslant d$. Suppose that a is NOT a zero of $D(x)$. Then there is a polynomial $P(x)$ of degree $p < d$ and a number A such that

$$\frac{N(x)}{D(x)\,(x-a)}=\frac{P(x)}{D(x)}+\frac{A}{x-a}$$

Proof.

- To save writing, let $z = x-a$. We then write $\tilde{N}(z) = N(z+a)$ and $\tilde{D}(z) = D(z+a)$, which are again polynomials of degree n and d respectively. We also know that $\tilde{D}(0) = D(a) \neq 0$.
- In order to complete the proof we need to find a polynomial $\tilde{P}(z)$ of degree $p < d$ and a number A such that

$$\frac{\tilde{N}(z)}{\tilde{D}(z)\,z}=\frac{\tilde{P}(z)}{\tilde{D}(z)}+\frac{A}{z}=\frac{\tilde{P}(z)z+A\tilde{D}(z)}{\tilde{D}(z)\,z}$$

  or equivalently, such that

$$\tilde{P}(z)z+A\tilde{D}(z)=\tilde{N}(z).$$

70 In fact, quadratic factors are completely avoidable because, if we use complex numbers, then every polynomial can be written as a product of linear factors. This is the fundamental theorem of algebra.

- Now look at the polynomial on the left hand side. Every term in $\tilde{P}(z)z$, has at least one power of z. So the constant term on the left hand side is exactly the constant term in $A\tilde{D}(z)$, which is equal to $A\tilde{D}(0)$. The constant term on the right hand side is equal to $\tilde{N}(0)$. So the constant terms on the left and right hand sides are the same if we choose $A = \frac{\tilde{N}(0)}{\tilde{D}(0)}$. Recall that $\tilde{D}(0)$ cannot be zero, so A is well defined.

- Now move $A\tilde{D}(z)$ to the right hand side.

$$\tilde{P}(z)z = \tilde{N}(z) - A\tilde{D}(z)$$

  The constant terms in $\tilde{N}(z)$ and $A\tilde{D}(z)$ are the same, so the right hand side contains no constant term and the right hand side is of the form $\tilde{N}_1(z)z$ for some polynomial $\tilde{N}_1(z)$.

- Since $\tilde{N}(z)$ is of degree at most d and $A\tilde{D}(z)$ is of degree exactly d, $\tilde{N}_1$ is a polynomial of degree $d-1$. It now suffices to choose $\tilde{P}(z) = \tilde{N}_1(z)$.

□

Now back to

$$\frac{N(x)}{(x-a_1)\times\cdots\times(x-a_d)}$$

Apply Lemma 1.10.13, with $D(x) = (x-a_2)\times\cdots\times(x-a_d)$ and $a = a_1$. It says

$$\frac{N(x)}{(x-a_1)\times\cdots\times(x-a_d)} = \frac{A_1}{x-a_1} + \frac{P(x)}{(x-a_2)\times\cdots\times(x-a_d)}$$

for some polynomial P of degree at most $d-2$ and some number A_1.

Apply Lemma 1.10.13 a second time, with $D(x) = (x-a_3)\times\cdots\times(x-a_d)$, $N(x) = P(x)$ and $a = a_2$. It says

$$\frac{P(x)}{(x-a_2)\times\cdots\times(x-a_d)} = \frac{A_2}{x-a_2} + \frac{Q(x)}{(x-a_3)\times\cdots\times(x-a_d)}$$

for some polynomial Q of degree at most $d-3$ and some number A_2.

At this stage, we know that

$$\frac{N(x)}{(x-a_1)\times\cdots\times(x-a_d)} = \frac{A_1}{x-a_1} + \frac{A_2}{x-a_2} + \frac{Q(x)}{(x-a_3)\times\cdots\times(x-a_d)}$$

If we just keep going, repeatedly applying Lemma 1, we eventually end up with

$$\frac{N(x)}{(x-a_1)\times\cdots\times(x-a_d)} = \frac{A_1}{x-a_1} + \cdots + \frac{A_d}{x-a_d}$$

as required.

▸▸▸ The General Linear Factor Case

Now consider splitting

$$\frac{N(x)}{(x-a_1)^{n_1}\times\cdots\times(x-a_d)^{n_d}}$$

into a sum of the form[71]

$$\left[\frac{A_{1,1}}{x-a_1}+\cdots+\frac{A_{1,n_1}}{(x-a_1)^{n_1}}\right]+\cdots+\left[\frac{A_{d,1}}{x-a_d}+\cdots+\frac{A_{d,n_d}}{(x-a_d)^{n_d}}\right]$$

We now show that this decomposition can always be achieved, under the assumptions that the a_i's are all different and $N(x)$ is a polynomial of degree at most $n_1+\cdots+n_d-1$. To do so, we shall repeatedly apply the following Lemma.

Lemma 1.10.14.

Let $N(x)$ and $D(x)$ be polynomials of degree n and d respectively, with $n<d+m$. Suppose that a is NOT a zero of $D(x)$. Then there is a polynomial $P(x)$ of degree $p<d$ and numbers $A_1,\cdots,A_m$ such that

$$\frac{N(x)}{D(x)\,(x-a)^m}=\frac{P(x)}{D(x)}+\frac{A_1}{x-a}+\frac{A_2}{(x-a)^2}+\cdots+\frac{A_m}{(x-a)^m}$$

Proof.

- As we did in the proof of the previous lemma, we write $z=x-a$. Then $\tilde N(z)=N(z+a)$ and $\tilde D(z)=D(z+a)$ are polynomials of degree n and d respectively, $\tilde D(0)=D(a)\neq 0$.
- In order to complete the proof we have to find a polynomial $\tilde P(z)$ of degree $p<d$ and numbers $A_1,\cdots,A_m$ such that

$$\begin{aligned}\frac{\tilde N(z)}{\tilde D(z)\,z^m}&=\frac{\tilde P(z)}{\tilde D(z)}+\frac{A_1}{z}+\frac{A_2}{z^2}+\cdots+\frac{A_m}{z^m}\\&=\frac{\tilde P(z)z^m+A_1z^{m-1}\tilde D(z)+A_2z^{m-2}\tilde D(z)+\cdots+A_m\tilde D(z)}{\tilde D(z)\,z^m}\end{aligned}$$

  or equivalently, such that

$$\tilde P(z)z^m+A_1z^{m-1}\tilde D(z)+A_2z^{m-2}\tilde D(z)+\cdots+A_{m-1}z\tilde D(z)+A_m\tilde D(z)=\tilde N(z)$$

- Now look at the polynomial on the left hand side. Every single term on the left hand side, except for the very last one, $A_m\tilde D(z)$, has at least one power of z. So the constant term on the left hand side is exactly the constant term in $A_m\tilde D(z)$, which is equal to $A_m\tilde D(0)$. The constant term on the right hand side is equal to $\tilde N(0)$. So the constant terms on the left and right hand sides are the same if we choose $A_m=\frac{\tilde N(0)}{\tilde D(0)}$. Recall that $\tilde D(0)\neq 0$ so A_m is well defined.

71 If we allow ourselves to use complex numbers as roots, this is the general case. We don't need to consider quadratic (or higher) factors since all polynomials can be written as products of linear factors with complex coefficients.

- Now move $A_m\tilde{D}(z)$ to the right hand side.

$$\tilde{P}(z)z^m + A_1 z^{m-1}\tilde{D}(z) + A_2 z^{m-2}\tilde{D}(z) + \cdots + A_{m-1}z\tilde{D}(z) = \tilde{N}(z) - A_m\tilde{D}(z)$$

  The constant terms in $\tilde{N}(z)$ and $A_m\tilde{D}(z)$ are the same, so the right hand side contains no constant term and the right hand side is of the form $\tilde{N}_1(z)z$ with $\tilde{N}_1$ a polynomial of degree at most $d+m-2$. (Recall that $\tilde{N}$ is of degree at most $d+m-1$ and $\tilde{D}$ is of degree at most d.) Divide the whole equation by z to get

$$\tilde{P}(z)z^{m-1} + A_1 z^{m-2}\tilde{D}(z) + A_2 z^{m-3}\tilde{D}(z) + \cdots + A_{m-1}\tilde{D}(z) = \tilde{N}_1(z).$$

- Now, we can repeat the previous argument. The constant term on the left hand side, which is exactly equal to $A_{m-1}\tilde{D}(0)$ matches the constant term on the right hand side, which is equal to $\tilde{N}_1(0)$ if we choose $A_{m-1} = \frac{\tilde{N}_1(0)}{\tilde{D}(0)}$. With this choice of A_{m-1}

$$\begin{aligned}\tilde{P}(z)z^{m-1} + A_1 z^{m-2}\tilde{D}(z) + A_2 z^{m-3}\tilde{D}(z) + \cdots + A_{m-2}z\tilde{D}(z)& \\ = \tilde{N}_1(z) - A_{m-1}\tilde{D}(z) = \tilde{N}_2(z)z&\end{aligned}$$

  with $\tilde{N}_2$ a polynomial of degree at most $d+m-3$. Divide by z and continue.

- After m steps like this, we end up with

$$\tilde{P}(z)z = \tilde{N}_{m-1}(z) - A_1\tilde{D}(z)$$

  after having chosen $A_1 = \frac{\tilde{N}_{m-1}(0)}{\tilde{D}(0)}$.

- There is no constant term on the right side so that $\tilde{N}_{m-1}(z) - A_1\tilde{D}(z)$ is of the form $\tilde{N}_m(z)z$ with $\tilde{N}_m$ a polynomial of degree $d-1$. Choosing $\tilde{P}(z) = \tilde{N}_m(z)$ completes the proof.

□

Now back to

$$\frac{N(x)}{(x-a_1)^{n_1}\times\cdots\times(x-a_d)^{n_d}}$$

Apply Lemma 1.10.14, with $D(x) = (x-a_2)^{n_2}\times\cdots\times(x-a_d)^{n_d}$, $m = n_1$ and $a = a_1$. It says

$$\begin{aligned}&\frac{N(x)}{(x-a_1)^{n_1}\times\cdots\times(x-a_d)^{n_d}}\\ &\qquad = \frac{A_{1,1}}{x-a_1} + \frac{A_{1,2}}{(x-a_1)^2} + \cdots + \frac{A_{1,n_1}}{(x-a)^{n_1}} + \frac{P(x)}{(x-a_2)^{n_2}\times\cdots\times(x-a_d)^{n_d}}\end{aligned}$$

Apply Lemma 1.10.14 a second time, with $D(x) = (x-a_3)^{n_3}\times\cdots\times(x-a_d)^{n_d}$, $N(x) = P(x)$, $m = n_2$ and $a = a_2$. And so on. Eventually, we end up with

$$\left[\frac{A_{1,1}}{x-a_1} + \cdots + \frac{A_{1,n_1}}{(x-a_1)^{n_1}}\right] + \cdots + \left[\frac{A_{d,1}}{x-a_d} + \cdots + \frac{A_{d,n_d}}{(x-a_d)^{n_d}}\right]$$

which is exactly what we were trying to show.

▸▸▸ Really Optional — The Fully General Case

We are now going to see that, in general, if $N(x)$ and $D(x)$ are polynomials with the degree of N being strictly smaller than the degree of D (which we'll denote $\deg(N) < \deg(D)$) and if

$$D(x) = K(x-a_1)^{m_1}\cdots(x-a_j)^{m_j}(x^2+b_1x+c_1)^{n_1}\cdots(x^2+b_kx+c_k)^{n_k} \tag{E1}$$

(with $b_\ell^2 - 4c_\ell < 0$ for all $1 \leqslant \ell \leqslant k$ so that no quadratic factor can be written as a product of linear factors with real coefficients) then there are real numbers $A_{i,j}$, $B_{i,j}$, $C_{i,j}$ such that

$$\begin{aligned}\frac{N(x)}{D(x)} = &\frac{A_{1,1}}{x-a_1} + \frac{A_{1,2}}{(x-a_1)^2} + \cdots + \frac{A_{1,m_1}}{(x-a_1)^{m_1}} + \cdots\\
&+ \frac{A_{j,1}}{x-a_j} + \frac{A_{j,2}}{(x-a_j)^2} + \cdots + \frac{A_{j,m_j}}{(x-a_j)^{m_j}}\\
&+ \frac{B_{1,1}x+C_{1,1}}{x^2+b_1x+c_1} + \frac{B_{1,2}x+C_{1,2}}{(x^2+b_1x+c_1)^2} + \cdots + \frac{B_{1,n_1}x+C_{1,n_1}}{(x^2+b_1x+c_1)^{n_1}} + \cdots\\
&+ \frac{B_{k,1}x+C_{k,1}}{x^2+b_kx+c_k} + \frac{B_{k,2}x+C_{k,2}}{(x^2+b_kx+c_k)^2} + \cdots + \frac{B_{k,n_k}x+C_{1,n_k}}{(x^2+b_kx+c_k)^{n_k}}\end{aligned}$$

This was (1.10.11).

We start with two simpler results, that we'll use repeatedly to get (1.10.11). In the first simpler result, we consider the fraction $\frac{P(x)}{Q_1(x)\,Q_2(x)}$ with $P(x)$, $Q_1(x)$ and $Q_2(x)$ being polynomials with real coefficients and we are going to assume that when $P(x)$, $Q_1(x)$ and $Q_2(x)$ are factored as in (E1), no two of them have a common linear or quadratic factor. As an example, no two of

$$\begin{aligned}P(x) &= 2(x-3)(x-4)(x^2+3x+3)\\ Q_1(x) &= 2(x-1)(x^2+2x+2)\\ Q_2(x) &= 2(x-2)(x^2+2x+3)\end{aligned}$$

have such a common factor. But, for

$$\begin{aligned}P(x) &= 2(x-3)(x-4)(x^2+x+1)\\ Q_1(x) &= 2(x-1)(x^2+2x+2)\\ Q_2(x) &= 2(x-2)(x^2+x+1)\end{aligned}$$

$P(x)$ and $Q_2(x)$ have the common factor x^2+x+1.

Lemma 1.10.15.

Let $P(x)$, $Q_1(x)$ and $Q_2(x)$ be polynomials with real coefficients and with $\deg(P) < \deg(Q_1Q_2)$. Assume that no two of $P(x)$, $Q_1(x)$ and $Q_2(x)$ have a common linear or quadratic factor. Then there are polynomials P_1, P_2 with $\deg(P_1) < \deg(Q_1)$, $\deg(P_2) < \deg(Q_2)$, and

$$\frac{P(x)}{Q_1(x)\,Q_2(x)} = \frac{P_1(x)}{Q_1(x)} + \frac{P_2(x)}{Q_2(x)}$$

Proof. We are to find polynomials P_1 and P_2 that obey

$$P(x) = P_1(x)\,Q_2(x) + P_2(x)\,Q_1(x)$$

Actually, we are going to find polynomials p_1 and p_2 that obey

$$p_1(x)\,Q_1(x) + p_2(x)\,Q_2(x) = C \tag{E2}$$

for some nonzero constant C, and then just multiply (E2) by $\frac{P(x)}{C}$. To find p_1, p_2 and C we are going to use something called the Euclidean algorithm. It is an algorithm[72] that is used to efficiently find the greatest common divisors of two numbers. Because $Q_1(x)$ and $Q_2(x)$ have no common factors of degree 1 or 2, their "greatest common divisor" has degree 0, i.e. is a constant.

- The first step is to apply long division to $\frac{Q_1(x)}{Q_2(x)}$ to find polynomials $n_0(x)$ and $r_0(x)$ such that
$$\frac{Q_1(x)}{Q_2(x)} = n_0(x) + \frac{r_0(x)}{Q_2(x)} \qquad \text{with } \deg(r_0) < \deg(Q_2)$$
or, equivalently,
$$Q_1(x) = n_0(x)\,Q_2(x) + r_0(x) \qquad \text{with } \deg(r_0) < \deg(Q_2)$$
- The second step is to apply long division to $\frac{Q_2(x)}{r_0(x)}$ to find polynomials $n_1(x)$ and $r_1(x)$ such that
$$Q_2(x) = n_1(x)\,r_0(x) + r_1(x) \qquad \text{with } \deg(r_1) < \deg(r_0) \text{ or } r_1(x) = 0$$
- The third step (assuming that $r_1(x)$ was not zero) is to apply long division to $\frac{r_0(x)}{r_1(x)}$ to find polynomials $n_2(x)$ and $r_2(x)$ such that
$$r_0(x) = n_2(x)\,r_1(x) + r_2(x) \qquad \text{with } \deg(r_2) < \deg(r_1) \text{ or } r_2(x) = 0$$
- And so on.

As the degree of the remainder $r_i(x)$ decreases by at least one each time i is increased by one, the above iteration has to terminate with some $r_{\ell+1}(x) = 0$. That is, we choose ℓ to be index of the last nonzero remainder. Here is a summary of all of the long division steps.

$$\begin{aligned}
Q_1(x) &= n_0(x)\,Q_2(x) + r_0(x) && \text{with } \deg(r_0) < \deg(Q_2)\\
Q_2(x) &= n_1(x)\,r_0(x) + r_1(x) && \text{with } \deg(r_1) < \deg(r_0)\\
r_0(x) &= n_2(x)\,r_1(x) + r_2(x) && \text{with } \deg(r_2) < \deg(r_1)\\
r_1(x) &= n_3(x)\,r_2(x) + r_3(x) && \text{with } \deg(r_3) < \deg(r_2)\\
&\vdots\\
r_{\ell-2}(x) &= n_\ell(x)\,r_{\ell-1}(x) + r_\ell(x) && \text{with } \deg(r_\ell) < \deg(r_{\ell-1})\\
r_{\ell-1}(x) &= n_{\ell+1}(x)\,r_\ell(x) + r_{\ell+1}(x) && \text{with } r_{\ell+1} = 0
\end{aligned}$$

Now we are going to take a closer look at all of the different remainders that we have generated.

72 It appears in Euclid's Elements, which was written about 300 BC, and it was probably known even before that.

- From first long division step, namely $Q_1(x) = n_0(x)\,Q_2(x) + r_0(x)$ we have that the remainder
$$r_0(x) = Q_1(x) - n_0(x)\,Q_2(x)$$
- From the second long division step, namely $Q_2(x) = n_1(x)\,r_0(x) + r_1(x)$ we have that the remainder
$$\begin{aligned} r_1(x) &= Q_2(x) - n_1(x)\,r_0(x) = Q_2(x) - n_1(x)\big[Q_1(x) - n_0(x)\,Q_2(x)\big] \\ &= A_1(x)\,Q_1(x) + B_1(x)\,Q_2(x) \end{aligned}$$
with $A_1(x) = -n_1(x)$ and $B_1(x) = 1 + n_0(x)\,n_1(x)$.
- From the third long division step (assuming that $r_1(x)$ was not zero), namely $r_0(x) = n_2(x)\,r_1(x) + r_2(x)$, we have that the remainder
$$\begin{aligned} r_2(x) &= r_0(x) - n_2(x)\,r_1(x) \\ &= \big[Q_1(x) - n_0(x)\,Q_2(x)\big] - n_2(x)\big[A_1(x)\,Q_1(x) + B_1(x)\,Q_2(x)\big] \\ &= A_2(x)\,Q_1(x) + B_2(x)\,Q_2(x) \end{aligned}$$
with $A_2(x) = 1 - n_2(x)\,A_1(x)$ and $B_2(x) = -n_0(x) - n_2(x)\,B_1(x)$.
- And so on. Continuing in this way, we conclude that the final nonzero remainder $r_\ell(x) = A_\ell(x)\,Q_1(x) + B_\ell(x)\,Q_2(x)$ for some polynomials A_ℓ and B_ℓ.

Now the last nonzero remainder $r_\ell(x)$ has to be a nonzero constant C because

- it is nonzero by the definition of $r_\ell(x)$ and
- if $r_\ell(x)$ were a polynomial of degree at least one, then
  - $r_\ell(x)$ would be a factor of $r_{\ell-1}(x)$ because $r_{\ell-1}(x) = n_{\ell+1}(x)\,r_\ell(x)$ and
  - $r_\ell(x)$ would be a factor of $r_{\ell-2}(x)$ because $r_{\ell-2}(x) = n_\ell(x)\,r_{\ell-1}(x) + r_\ell(x)$ and
  - $r_\ell(x)$ would be a factor of $r_{\ell-3}(x)$ because $r_{\ell-3}(x) = n_{\ell-1}(x)\,r_{\ell-2}(x) + r_{\ell-1}(x)$ and
  - $\cdots$ and $\cdots$
  - $r_\ell(x)$ would be a factor of $r_1(x)$ because $r_1(x) = n_3(x)\,r_2(x) + r_3(x)$ and
  - $r_\ell(x)$ would be a factor of $r_0(x)$ because $r_0(x) = n_2(x)\,r_1(x) + r_2(x)$ and
  - $r_\ell(x)$ would be a factor of $Q_2(x)$ because $Q_2(x) = n_1(x)\,r_0(x) + r_1(x)$ and
  - $r_\ell(x)$ would be a factor of $Q_1(x)$ because $Q_1(x) = n_0(x)\,Q_2(x) + r_0(x)$
- so that $r_\ell(x)$ would be a common factor for $Q_1(x)$ and $Q_2(x)$, in contradiction to the hypothesis that no two of $P(x)$, $Q_1(x)$ and $Q_2(x)$ have a common linear or quadratic factor.

We now have that $A_\ell(x)\,Q_1(x) + B_\ell(x)\,Q_2(x) = r_\ell(x) = C$. Multiplying by $\frac{P(x)}{C}$ gives

$$\tilde{P}_2(x)\,Q_1(x) + \tilde{P}_1(x)\,Q_2(x) = P(x) \quad \text{or} \quad \frac{\tilde{P}_1(x)}{Q_1(x)} + \frac{\tilde{P}_2(x)}{Q_2(x)} = \frac{P(x)}{Q_1(x)\,Q_2(x)}$$

with $\tilde{P}_2(x) = \frac{P(x)\,A_\ell(x)}{C}$ and $\tilde{P}_1(x) = \frac{P(x)\,B_\ell(x)}{C}$. We're not quite done, because there is still the danger that $\deg(\tilde{P}_1) \geqslant \deg(Q_1)$ or $\deg(\tilde{P}_2) \geqslant \deg(Q_2)$. To deal with that possibility, we long divide $\frac{\tilde{P}_1(x)}{Q_1(x)}$ and call the remainder $P_1(x)$.

$$\frac{\tilde{P}_1(x)}{Q_1(x)} = N(x) + \frac{P_1(x)}{Q_1(x)} \qquad \text{with } \deg(P_1) < \deg(Q_1)$$

Therefore we have that

$$\begin{aligned}\frac{P(x)}{Q_1(x)\,Q_2(x)} &= \frac{P_1(x)}{Q_1(x)} + N(x) + \frac{\tilde{P}_2(x)}{Q_2(x)} \\ &= \frac{P_1(x)}{Q_1(x)} + \frac{\tilde{P}_2(x) + N(x)Q_2(x)}{Q_2(x)}\end{aligned}$$

Denoting $P_2(x) = \tilde{P}_2(x) + N(x)Q_2(x)$ gives $\frac{P}{Q_1Q_2} = \frac{P_1}{Q_1} + \frac{P_2}{Q_2}$ and since $\deg(P_1) < \deg(Q_1)$, the only thing left to prove is that $\deg(P_2) < \deg(Q_2)$.

We assume that $\deg(P_2) \geqslant \deg(Q_2)$ and look for a contradiction. We have

$$\begin{aligned}&\deg(P_2Q_1) \geqslant \deg(Q_1Q_2) > \deg(P_1Q_2) \\ &\implies \deg(P) = \deg(P_1Q_2 + P_2Q_1) = \deg(P_2Q_1) \geqslant \deg(Q_1Q_2)\end{aligned}$$

which contradicts the hypothesis that $\deg(P) < \deg(Q_1Q_2)$ and the proof is complete. □

For the second of the two simpler results, that we'll shortly use repeatedly to get (1.10.11), we consider $\frac{P(x)}{(x-a)^m}$ and $\frac{P(x)}{(x^2+bx+c)^m}$.

Lemma 1.10.16.

Let $m \geqslant 2$ be an integer, and let $Q(x)$ be either $x - a$ or $x^2 + bx + c$, with a, b and c being real numbers. Let $P(x)$ be a polynomial with real coefficients, which does not contain $Q(x)$ as a factor, and with $\deg(P) < \deg(Q^m) = m\deg(Q)$. Then, for each $1 \leqslant i \leqslant m$, there is a polynomial P_i with $\deg(P_i) < \deg(Q)$ or $P_i = 0$, such that

$$\frac{P(x)}{Q(x)^m} = \frac{P_1(x)}{Q(x)} + \frac{P_2(x)}{Q(x)^2} + \frac{P_3(x)}{Q(x)^3} + \cdots + \frac{P_{m-1}(x)}{Q(x)^{m-1}} + \frac{P_m(x)}{Q(x)^m}.$$

In particular, if $Q(x) = x - a$, then each $P_i(x)$ is just a constant A_i, and if $Q(x) = x^2 + bx + c$, then each $P_i(x)$ is a polynomial $B_ix + C_i$ of degree at most one.

Proof. We simply repeatedly use long divison to get

$$\begin{aligned}\frac{P(x)}{Q(x)^m} &= \frac{P(x)}{Q(x)}\frac{1}{Q(x)^{m-1}} = \left\{n_1(x) + \frac{r_1(x)}{Q(x)}\right\}\frac{1}{Q(x)^{m-1}} \\ &= \frac{r_1(x)}{Q(x)^m} + \frac{n_1(x)}{Q(x)}\frac{1}{Q(x)^{m-2}} = \frac{r_1(x)}{Q(x)^m} + \left\{n_2(x) + \frac{r_2(x)}{Q(x)}\right\}\frac{1}{Q(x)^{m-2}} \\ &= \frac{r_1(x)}{Q(x)^m} + \frac{r_2(x)}{Q(x)^{m-1}} + \frac{n_2(x)}{Q(x)}\frac{1}{Q(x)^{m-3}} \\ &\vdots \\ &= \frac{r_1(x)}{Q(x)^m} + \frac{r_2(x)}{Q(x)^{m-1}} + \cdots + \frac{r_{m-2}(x)}{Q(x)^3} + \frac{n_{m-2}(x)}{Q(x)}\frac{1}{Q(x)} \\ &= \frac{r_1(x)}{Q(x)^m} + \frac{r_2(x)}{Q(x)^{m-1}} + \cdots + \frac{r_{m-2}(x)}{Q(x)^3} + \left\{n_{m-1}(x) + \frac{r_{m-1}(x)}{Q(x)}\right\}\frac{1}{Q(x)} \\ &= \frac{r_1(x)}{Q(x)^m} + \frac{r_2(x)}{Q(x)^{m-1}} + \cdots + \frac{r_{m-2}(x)}{Q(x)^3} + \frac{r_{m-1}(x)}{Q(x)^2} + \frac{n_{m-1}(x)}{Q(x)}\end{aligned}$$

By the rules of long division every $\deg(r_i) < \deg(Q)$. It is also true that the final numerator, n_{m-1}, has $\deg(n_{m-1}) < \deg(Q)$ — that is, we kept dividing by Q until the degree of the quotient was less than the degree of Q. To see this, note that $\deg(P) < m\deg(Q)$ and

$$\begin{aligned}
\deg(n_1) &= \deg(P) - \deg(Q) \\
\deg(n_2) &= \deg(n_1) - \deg(Q) = \deg(P) - 2\deg(Q) \\
&\vdots \\
\deg(n_{m-1}) &= \deg(n_{m-2}) - \deg(Q) = \deg(P) - (m-1)\deg(Q) \\
&< m\deg(Q) - (m-1)\deg(Q) \\
&= \deg(Q)
\end{aligned}$$

So, if $\deg(Q) = 1$, then $r_1, r_2, \ldots, r_{m-1}, n_{m-1}$ are all real numbers, and if $\deg(Q) = 2$, then $r_1, r_2, \ldots, r_{m-1}, n_{m-1}$ all have degree at most one. □

We are now in a position to get (1.10.11). We use (E1) to factor[73] $D(x) = (x-a_1)^{m_1}Q_2(x)$ and use Lemma 1.10.15 to get

$$\frac{N(x)}{D(x)} = \frac{N(x)}{(x-a_1)^{m_1}Q_2(x)} = \frac{P_1(x)}{(x-a_1)^{m_1}} + \frac{P_2(x)}{Q_2(x)}$$

where $\deg(P_1) < m_1$, and $\deg(P_2) < \deg(Q_2)$. Then we use Lemma 1.10.16 to get

$$\frac{N(x)}{D(x)} = \frac{P_1(x)}{(x-a_1)^{m_1}} + \frac{P_2(x)}{Q_2(x)} = \frac{A_{1,1}}{x-a_1} + \frac{A_{1,2}}{(x-a_1)^2} + \cdots + \frac{A_{1,m_1}}{(x-a_1)^{m_1}} + \frac{P_2(x)}{Q_2(x)}$$

We continue working on $\frac{P_2(x)}{Q_2(x)}$ in this way, pulling off of the denominator one $(x-a_i)^{m_i}$ or one $(x^2+b_ix+c_i)^{n_i}$ at a time, until we exhaust all of the factors in the denominator $D(x)$.

1.11▲ 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[74]. Such integrals are not merely mathematical curiosities, but arise very naturally in many contexts. For example, the error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

is extremely important in many areas of mathematics, and also in many practical applications of statistics.

73 This is assuming that there is at least one linear factor. If not, we factor $D(x) = (x^2+b_1x+c_1)^{n_1}Q_2(x)$ instead.

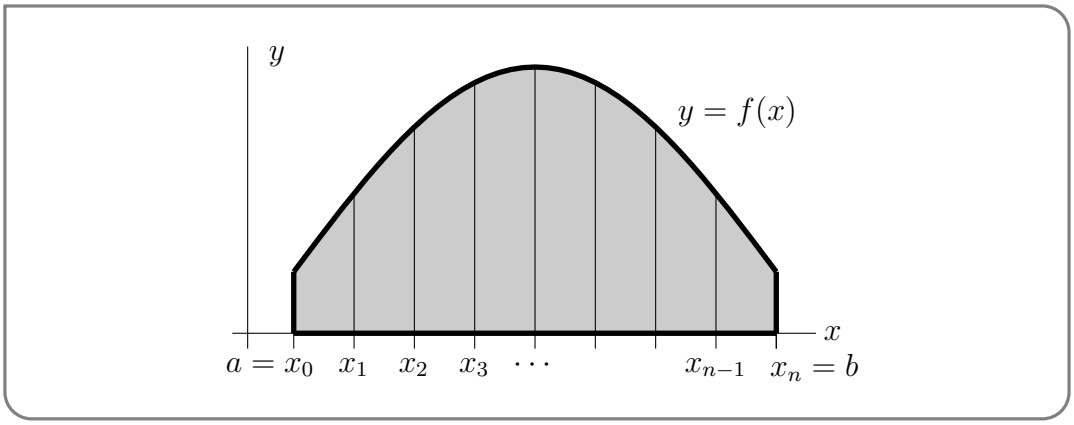
74 We apologise for being a little sloppy here — but we just want to say that it can be very hard or even impossible to write some integrals as some finite sized expression involving polynomials, exponentials, logarithms and trigonometric functions. We don't want to get into a discussion of computability, though that is a very interesting topic.

In 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. To develop these methods we return to Riemann sums and our geometric interpretation of the definite integral as the signed area.

We start by describing (and applying) three simple algorithms for generating, numerically, approximate values for the definite integral $\int_a^b f(x)\,\mathrm{d}x$. In each algorithm, we begin in much the same way as we approached Riemann sums.

- We first select an integer $n > 0$, called the "number of steps".

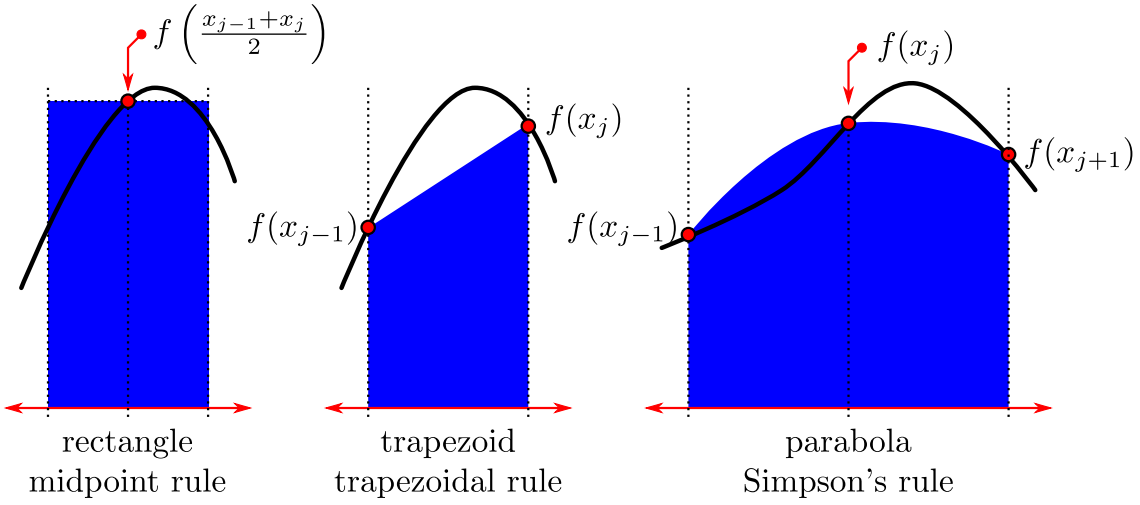
- We then divide the interval of integration, $a \leqslant x \leqslant b$, into n equal subintervals, each of length $\Delta x = \frac{b-a}{n}$. The first subinterval runs from $x_0 = a$ to $x_1 = a + \Delta x$. The second runs from x_1 to $x_2 = a + 2\Delta x$, and so on. The last runs from $x_{n-1} = b - \Delta x$ to $x_n = b$.



This splits the original integral into n pieces:

$$\int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x$$

Each subintegral $\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x$ is approximated by the area of a simple geometric figure. The three algorithms we consider approximate the area by rectangles, trapezoids and parabolas (respectively).



We will explain these rules in detail below, but we give a brief overview here:

(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

$$\int_{x_{j-1}}^{x_j} f(x)\mathrm{d}x \approx f\left(\frac{x_{j-1}+x_j}{2}\right)\Delta x$$

This is illustrated in the leftmost figure above.

(2) The trapezoidal rule approximates each subintegral by the area of a trapezoid with vertices at $(x_{j-1},0),(x_{j-1},f(x_{j-1})),(x_j,f(x_j)),(x_j,0)$:

$$\int_{x_{j-1}}^{x_j} f(x)\mathrm{d}x \approx \frac{1}{2}\left[f(x_{j-1})+f(x_j)\right]\Delta x$$

The trapezoid is illustrated in the middle figure above. We shall derive the formula for the area shortly.

(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 $(x_{j+1},f(x_{j+1}))$:

$$\int_{x_{j-1}}^{x_{j+1}} f(x)\mathrm{d}x \approx \frac{1}{3}\left[f(x_{j-1})+4f(x_j)+f(x_{j+1})\right]\Delta x$$

The parabola is illustrated in the right hand figure above. We shall derive the formula for the area shortly.

Notation 1.11.1 (Midpoints).

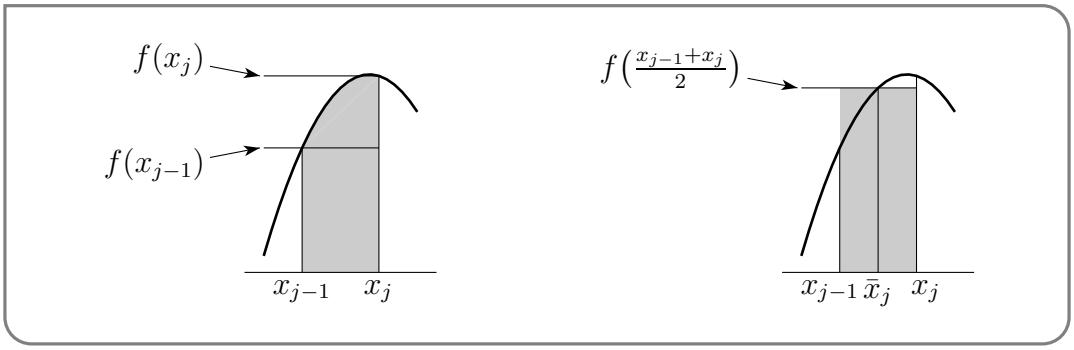
In what follows we need to refer to the midpoint between x_{j-1} and x_j very frequently. To save on writing (and typing) we introduce the notation

$$\bar{x}_j = \frac{1}{2}\left(x_{j-1}+x_j\right).$$

1.11.1 ▸▸ The Midpoint Rule

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 Δx. Our other approximations will have similar forms.

In summary:

Equation 1.11.2 (The midpoint rule).

The midpoint rule approximation is

$$\int_a^b f(x)\,dx \approx \Big[f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)\Big]\Delta x$$

where $\Delta x = \frac{b-a}{n}$ and

$$\begin{array}{lllll} x_0 = a & x_1 = a + \Delta x & x_2 = a + 2\Delta x & \cdots & x_{n-1} = b - \Delta x \quad x_n = b \\ & \bar{x}_1 = \frac{x_0+x_1}{2} & \bar{x}_2 = \frac{x_1+x_2}{2} & \cdots & \bar{x}_{n-1} = \frac{x_{n-2}+x_{n-1}}{2} \quad \bar{x}_n = \frac{x_{n-1}+x_n}{2} \end{array}$$

Example 1.11.3 $\left(\int_0^1 \frac{4}{1+x^2}\,\mathrm{d}x\right)$

We approximate the above integral using the midpoint rule with $n = 8$ step.

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

$$x_0 = 0 \qquad x_1 = \tfrac{1}{8} \qquad x_2 = \tfrac{2}{8} \qquad \cdots \qquad x_7 = \tfrac{7}{8} \qquad x_8 = \tfrac{8}{8} = 1$$

  Consequently

$$\bar{x}_1 = \tfrac{1}{16} \qquad \bar{x}_2 = \tfrac{3}{16} \qquad \bar{x}_3 = \tfrac{5}{16} \qquad \cdots \qquad \bar{x}_8 = \tfrac{15}{16}$$

- We now apply Equation (1.11.2) to the integrand $f(x) = \frac{4}{1+x^2}$:

$$\begin{aligned}
\int_0^1 \frac{4}{1+x^2}\,\mathrm{d}x &\approx \left[\overbrace{\frac{4}{1+\bar{x}_1^2}}^{f(\bar{x}_1)} + \overbrace{\frac{4}{1+\bar{x}_2^2}}^{f(\bar{x}_2)} + \cdots + \overbrace{\frac{4}{1+\bar{x}_7^2}}^{f(\bar{x}_{n-1})} + \overbrace{\frac{4}{1+\bar{x}_8^2}}^{f(\bar{x}_n)}\right]\Delta x \\
&= \left[\frac{4}{1+\frac{1}{16^2}} + \frac{4}{1+\frac{3^2}{16^2}} + \frac{4}{1+\frac{5^2}{16^2}} + \frac{4}{1+\frac{7^2}{16^2}} + \frac{4}{1+\frac{9^2}{16^2}} + \frac{4}{1+\frac{11^2}{16^2}} + \frac{4}{1+\frac{13^2}{16^2}} + \frac{4}{1+\frac{15^2}{16^2}}\right]\frac{1}{8} \\
&= \left[3.98444 + 3.86415 + 3.64413 + 3.35738 + 3.03858 + 2.71618 + 2.40941 + 2.12890\right]\frac{1}{8} \\
&= 3.1429
\end{aligned}$$

  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):

$$\int_0^1 \frac{4}{1+x^2}\mathrm{d}x = 4\arctan x\Big|_0^1 = \pi$$

- 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 1.11.3

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[75]. Of course, it would be very helpful to quantify what we mean by "good" in this context and that requires us to discuss errors.

Definition 1.11.4.

Suppose that α 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 1.11.4 below.

Example 1.11.5 $\left(\int_0^\pi \sin x\,\mathrm{d}x\right)$

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

$$x_0 = 0 \qquad x_1 = \tfrac{\pi}{8} \qquad x_2 = \tfrac{2\pi}{8} \qquad \cdots \qquad x_7 = \tfrac{7\pi}{8} \qquad x_8 = \tfrac{8\pi}{8} = \pi$$

  Consequently,

$$\bar{x}_1 = \tfrac{\pi}{16} \qquad \bar{x}_2 = \tfrac{3\pi}{16} \qquad \cdots \qquad \bar{x}_7 = \tfrac{13\pi}{16} \qquad \bar{x}_8 = \tfrac{15\pi}{16}$$

- Now apply Equation (1.11.2) to the integrand $f(x) = \sin x$:

$$\begin{aligned}
&\int_0^\pi \sin x\,\mathrm{d}x \approx \Big[\sin(\bar{x}_1) + \sin(\bar{x}_2) + \cdots + \sin(\bar{x}_8)\Big]\Delta x \\
&= \Big[\sin(\tfrac{\pi}{16}) + \sin(\tfrac{3\pi}{16}) + \sin(\tfrac{5\pi}{16}) + \sin(\tfrac{7\pi}{16}) + \sin(\tfrac{9\pi}{16}) + \sin(\tfrac{11\pi}{16}) + \sin(\tfrac{13\pi}{16}) + \sin(\tfrac{15\pi}{16})\Big]\tfrac{\pi}{8} \\
&= \Big[0.1951 + 0.5556 + 0.8315 + 0.9808 + 0.9808 + 0.8315 + 0.5556 + 0.1951\Big] \times 0.3927 \\
&= 5.1260 \times 0.3927 = 2.013
\end{aligned}$$

- Again, we have chosen this example so that we can compare it against the exact value:

$$\int_0^\pi \sin x\mathrm{d}x = \big[-\cos x\big]_0^\pi = -\cos\pi + \cos 0 = 2.$$

75 Thankfully it is very easy to write a program to apply the midpoint rule.

- So with eight steps of the midpoint rule we achieved

$$\text{absolute error} = |2.013 - 2| = 0.013$$
$$\text{relative error} = \frac{|2.013 - 2|}{2} = 0.0065$$
$$\text{percentage error} = 100 \times \frac{|2.013 - 2|}{2} = 0.65\%$$

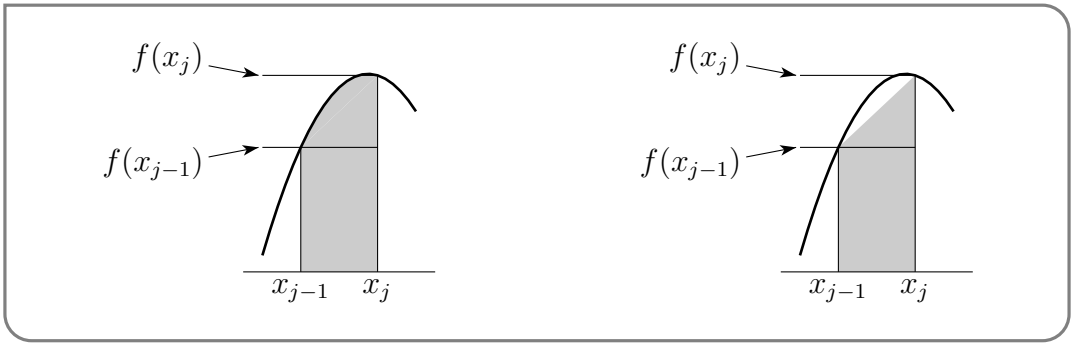
With little work we have managed to estimate the integral to within 1% of its true value.

Example 1.11.5

1.11.2 ▸▸ The Trapezoidal Rule

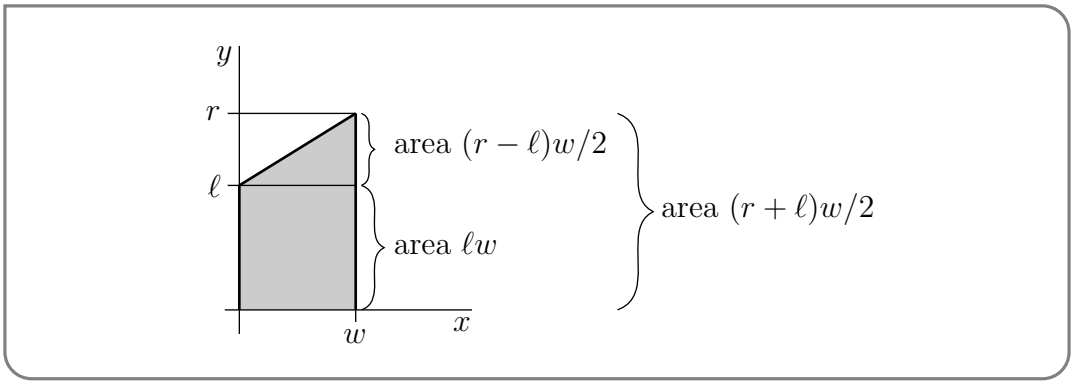
Consider again the area represented by the integral $\int_{x_{j-1}}^{x_j} f(x)\,dx$. The trapezoidal rule[76] (unsurprisingly) approximates this area by a trapezoid[77] whose vertices lie at

$$(x_{j-1}, 0), (x_{j-1}, f(x_{j-1})), (x_j, f(x_j)) \text{ and } (x_j, 0).$$



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.



76 This method is also called the "trapezoid rule" and "trapezium rule".

77 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

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_{x_0}^{x_1} f(x)\,dx \;+\; \int_{x_1}^{x_2} f(x)\,dx \;+\cdots+\; \int_{x_{n-1}}^{x_n} f(x)\,dx\\ &\approx \tfrac{f(x_0)+f(x_1)}{2}\Delta x + \tfrac{f(x_1)+f(x_2)}{2}\Delta x \;+\cdots+ \tfrac{f(x_{n-1})+f(x_n)}{2}\Delta x\\ &= \left[\tfrac{1}{2}f(x_0)+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\tfrac{1}{2}f(x_n)\right]\Delta x\end{aligned}$$

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 1.11.6 (The trapezoidal rule).

The trapezoidal rule approximation is

$$\int_a^b f(x)\,dx \approx \left[\tfrac{1}{2}f(x_0)+f(x_1)+f(x_2)+\cdots+f(x_{n-1})+\tfrac{1}{2}f(x_n)\right]\Delta x$$

where

$$\Delta x = \tfrac{b-a}{n}, \quad x_0 = a, \quad x_1 = a+\Delta x, \quad x_2 = a+2\Delta x, \quad \cdots, \quad x_{n-1} = b-\Delta x, \quad x_n = b$$

To compare and contrast we apply the trapezoidal rule to the examples we did above with the midpoint rule.

Example 1.11.7 ($\int_0^1 \frac{4}{1+x^2}\,dx$ — using the trapezoidal rule)

Solution. We proceed very similarly to Example 1.11.3 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

$$x_0 = 0 \qquad x_1 = \tfrac{1}{8} \qquad x_2 = \tfrac{2}{8} \qquad \cdots \qquad x_7 = \tfrac{7}{8} \qquad x_8 = \tfrac{8}{8} = 1$$

- Applying the trapezoidal rule, Equation (1.11.6), gives

$$\begin{aligned}
\int_0^1 \frac{4}{1+x^2}\,dx &\approx \Big[\frac{1}{2}\overbrace{\frac{4}{1+x_0^2}}^{f(x_0)} + \overbrace{\frac{4}{1+x_1^2}}^{f(x_1)} + \cdots + \overbrace{\frac{4}{1+x_7^2}}^{f(x_{n-1})} + \frac{1}{2}\overbrace{\frac{4}{1+x_8^2}}^{f(x_n)}\Big]\Delta x \\
&= \Big[\frac{1}{2}\frac{4}{1+0^2} + \frac{4}{1+\frac{1}{8^2}} + \frac{4}{1+\frac{2^2}{8^2}} + \frac{4}{1+\frac{3^2}{8^2}} \\
&\quad + \frac{4}{1+\frac{4^2}{8^2}} + \frac{4}{1+\frac{5^2}{8^2}} + \frac{4}{1+\frac{6^2}{8^2}} + \frac{4}{1+\frac{7^2}{8^2}} + \frac{1}{2}\frac{4}{1+\frac{8^2}{8^2}}\Big]\frac{1}{8} \\
&= \Big[\frac{1}{2}\times 4 + 3.939 + 3.765 + 3.507 \\
&\quad + 3.2 + 2.876 + 2.56 + 2.266 + \frac{1}{2}\times 2\Big]\frac{1}{8} \\
&= 3.139
\end{aligned}$$

  to three decimal places.

- The exact value of the integral is still π. 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 1.11.3.

Example 1.11.7

Let us also redo Example 1.11.5 using the trapezoidal rule.

Example 1.11.8 ($\int_0^\pi \sin x\,dx$ — using the trapezoidal rule)

Solution. We proceed very similarly to Example 1.11.5 and again use $n = 8$ steps.

- We again have $a = 0$, $b = \pi$, $\Delta x = \frac{\pi}{8}$ and

$$x_0 = 0 \qquad x_1 = \tfrac{\pi}{8} \qquad x_2 = \tfrac{2\pi}{8} \qquad \cdots \qquad x_7 = \tfrac{7\pi}{8} \qquad x_8 = \tfrac{8\pi}{8} = \pi$$

- Applying the trapezoidal rule, Equation (1.11.6), gives

$$\begin{aligned}
&\int_0^\pi \sin x\,dx \approx \Big[\tfrac{1}{2}\sin(x_0) + \sin(x_1) + \cdots + \sin(x_7) + \tfrac{1}{2}\sin(x_8)\Big]\Delta x \\
&= \Big[\tfrac{1}{2}\sin 0 + \sin\tfrac{\pi}{8} + \sin\tfrac{2\pi}{8} + \sin\tfrac{3\pi}{8} + \sin\tfrac{4\pi}{8} + \sin\tfrac{5\pi}{8} + \sin\tfrac{6\pi}{8} + \sin\tfrac{7\pi}{8} + \tfrac{1}{2}\sin\tfrac{8\pi}{8}\Big]\tfrac{\pi}{8} \\
&= \Big[\tfrac{1}{2}\times 0 + 0.3827 + 0.7071 + 0.9239 + 1.0000 + 0.9239 + 0.7071 + 0.3827 + \tfrac{1}{2}\times 0\Big] \times 0.3927 \\
&= 5.0274 \times 0.3927 = 1.974
\end{aligned}$$

- 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 1.11.5 using the midpoint rule.

Example 1.11.8

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 Section 1.11.4.

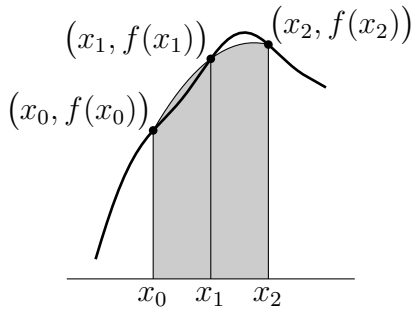
1.11.3 ▸▸ Simpson's Rule

When we use the trapezoidal rule we approximate the area $\int_{x_{j-1}}^{x_j} f(x)\mathrm{d}x$ by the area between the x-axis and a straight line that runs from $(x_{j-1}, f(x_{j-1}))$ to $(x_j, f(x_j))$ — that is, we approximate the function $f(x)$ on this interval by a linear function that agrees with the function at each endpoint. An obvious way to extend this — just as we did when extending linear approximations to quadratic approximations in our differential calculus course — is to approximate the function with a quadratic. This is precisely what Simpson's[78] rule does.

Simpson's 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

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x = \int_{x_0}^{x_2} f(x)\,\mathrm{d}x$$

by the area bounded by the parabola that passes through the three points $\big(x_0, f(x_0)\big)$, $\big(x_1, f(x_1)\big)$ and $\big(x_2, f(x_2)\big)$, 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 $\int_{x_2}^{x_4} f(x)\,\mathrm{d}x$ by the area between the x–axis and the part of a parabola with $x_2 \leqslant x \leqslant x_4$. This parabola passes through the three points $\big(x_2, f(x_2)\big)$, $\big(x_3, f(x_3)\big)$ and $\big(x_4, f(x_4)\big)$. 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 $\big(x_0, f(x_0)\big)$, $\big(x_1, f(x_1)\big)$ and $\big(x_2, f(x_2)\big)$. Then we find the area

78 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.

between the x–axis and the part of that parabola with $x_0 \leqslant x \leqslant x_2$. To simplify this computation consider a parabola passing through the points $(-h, y_{-1}), (0, y_0)$ and (h, y_1).

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

$$\begin{aligned}\int_{-h}^{h} \big[Ax^2 + Bx + C\big]\,\mathrm{d}x &= \left[\frac{A}{3}x^3 + \frac{B}{2}x^2 + Cx\right]_{-h}^{h} \\ &= \frac{2A}{3}h^3 + 2Ch \qquad \text{it is helpful to write it as} \\ &= \frac{h}{3}\left(2Ah^2 + 6C\right)\end{aligned}$$

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

$$\begin{aligned}\text{area} &= \int_{-h}^{h} \big[Ax^2 + Bx + C\big]\,\mathrm{d}x = \frac{h}{3}\left(2Ah^2 + 6C\right) \\ &= \frac{h}{3}\left(y_{-1} + 4y_0 + y_1\right)\end{aligned}$$

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

$$\int_{x_0}^{x_2} f(x)\,\mathrm{d}x \approx \tfrac{1}{3}\Delta x\big[f(x_0)+4f(x_1)+f(x_2)\big]$$

and

$$\int_{x_2}^{x_4} f(x)\,\mathrm{d}x \approx \tfrac{1}{3}\Delta x\big[f(x_2)+4f(x_3)+f(x_4)\big]$$

and so on. Summing these all together gives:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \int_{x_0}^{x_2} f(x)\,\mathrm{d}x + \int_{x_2}^{x_4} f(x)\,\mathrm{d}x + \int_{x_4}^{x_6} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-2}}^{x_n} f(x)\,\mathrm{d}x\\ &\approx \tfrac{\Delta x}{3}\big[f(x_0)+4f(x_1)+f(x_2)\big] + \tfrac{\Delta x}{3}\big[f(x_2)+4f(x_3)+f(x_4)\big]\\ &\quad + \tfrac{\Delta x}{3}\big[f(x_4)+4f(x_5)+f(x_6)\big] + \cdots + \tfrac{\Delta x}{3}\big[f(x_{n-2})+4f(x_{n-1})+f(x_n)\big]\\ &= \Big[f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+2f(x_4)+\cdots+2f(x_{n-2})+4f(x_{n-1})+f(x_n)\Big]\tfrac{\Delta x}{3}\end{aligned}$$

In summary

Equation 1.11.9 (Simpson's rule).

The Simpson's rule approximation is

$$\int_a^b f(x)\,\mathrm{d}x \approx \Big[f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+2f(x_4)+\cdots \cdots+2f(x_{n-2})+4f(x_{n-1})+f(x_n)\Big]\tfrac{\Delta x}{3}$$

where n is even and

$$\Delta x = \tfrac{b-a}{n}, \quad x_0 = a, \quad x_1 = a+\Delta x, \quad x_2 = a+2\Delta x, \quad \cdots, \quad x_{n-1} = b-\Delta x, \quad x_n = b$$

Notice that Simpson's rule requires essentially no more work than the trapezoidal rule. In both rules we must evaluate $f(x)$ at $x = x_0, x_1, \ldots, x_n$, but we add those terms multiplied by different constants[79].

Let's put it to work on our two running examples.

Example 1.11.10 ($\int_0^1 \frac{4}{1+x^2}\,\mathrm{d}x$ — using Simpson's rule)

Solution. We proceed almost identically to Example 1.11.7 and again use $n = 8$ steps.

79 There is an easy generalisation of Simpson's rule that uses cubics instead of parabolas. It leads to the formula

$$\int_a^b f(x)\mathrm{d}x = [f(x_0)+3f(x_1)+3f(x_2)+2f(x_3)+2f(x_4)+3f(x_5)+3f(x_6)+2f(x_7)+\cdots+f(x_n)]\,\frac{3\Delta x}{8}$$

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.

- We have the same $\Delta, a, b, x_0, \ldots, x_n$ as Example 1.11.7.
- Applying Equation 1.11.9 gives

$$\begin{aligned}\int_0^1 \frac{4}{1+x^2}\,dx \approx &\left[\frac{4}{1+0^2}+4\frac{4}{1+\frac{1}{8^2}}+2\frac{4}{1+\frac{2^2}{8^2}}+4\frac{4}{1+\frac{3^2}{8^2}}\right.\\ &\left.+2\frac{4}{1+\frac{4^2}{8^2}}+4\frac{4}{1+\frac{5^2}{8^2}}+2\frac{4}{1+\frac{6^2}{8^2}}+4\frac{4}{1+\frac{7^2}{8^2}}+\frac{4}{1+\frac{8^2}{8^2}}\right]\frac{1}{8\times 3}\\ =&\left[4+4\times 3.938461538+2\times 3.764705882+4\times 3.506849315\right.\\ &\left.+2\times 3.2+4\times 2.876404494+2\times 2.56+4\times 2.265486726+2\right]\frac{1}{8\times 3}\\ =&\ 3.14159250\end{aligned}$$

  to eight decimal places.
- This agrees with π (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 $100\frac{|3.14159250-\pi|}{\pi}\% = 5\times 10^{-6}\%$ of the exact answer.

Example 1.11.10

It is striking that the absolute error approximating with Simpson's rule is so much smaller than the error from the midpoint and trapezoidal rules.

$$\begin{aligned}\text{midpoint error} &= 0.0013\\ \text{trapezoid error} &= 0.0026\\ \text{Simpson error} &= 0.00000015\end{aligned}$$

Buoyed by this success, we will also redo Example 1.11.8 using Simpson's rule.

Example 1.11.11 ($\int_0^\pi \sin x\,dx$ — Simpson's rule)

Solution. We proceed almost identically to Example 1.11.8 and again use $n = 8$ steps.

- We have the same $\Delta, a, b, x_0, \ldots, x_n$ as Example 1.11.7.
- Applying Equation 1.11.9 gives

$$\begin{aligned}\int_0^\pi \sin x\,dx &\approx \left[\sin(x_0)+4\sin(x_1)+2\sin(x_2)+\cdots+4\sin(x_7)+\sin(x_8)\right]\tfrac{\Delta x}{3}\\ &= \left[\sin(0)+4\sin(\tfrac{\pi}{8})+2\sin(\tfrac{2\pi}{8})+4\sin(\tfrac{3\pi}{8})+2\sin(\tfrac{4\pi}{8})\right.\\ &\qquad\left.+4\sin(\tfrac{5\pi}{8})+2\sin(\tfrac{6\pi}{8})+4\sin(\tfrac{7\pi}{8})+\sin(\tfrac{8\pi}{8})\right]\tfrac{\pi}{8\times 3}\\ &= \left[0+4\times 0.382683+2\times 0.707107+4\times 0.923880+2\times 1.0\right.\\ &\qquad\left.+4\times 0.923880+2\times 0.707107+4\times 0.382683+0\right]\tfrac{\pi}{8\times 3}\\ &= 15.280932\times 0.130900\\ &= 2.00027\end{aligned}$$

- With only eight steps of Simpson's rule we achieved $100\frac{2.00027-2}{2} = 0.014\%$ accuracy.

Example 1.11.11

Again we contrast the error we achieved with the other two rules:

$$\begin{aligned}\text{midpoint error} &= 0.013\\ \text{trapezoid error} &= 0.026\\ \text{Simpson error} &= 0.00027\end{aligned}$$

This completes our derivation of the midpoint, trapezoidal and Simpson's rules for approximating the values of definite integrals. So far we have not attempted to see how efficient and how accurate the algorithms are in general. That's our next task.

1.11.4 ▸▸ Three Simple Numerical Integrators – Error Behaviour

Now we are armed with our three (relatively simple) methods for numerical integration we should give thought to how practical they might be in the real world[80]. Two obvious considerations when deciding whether or not a given algorithm is of any practical value are

(a) the amount of computational effort required to execute the algorithm and

(b) the accuracy that this computational effort yields.

For algorithms like our simple integrators, 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 the midpoint rule is n, while the number required for n steps of the trapezoidal and Simpson's rules is $n+1$. So all three of our rules require essentially the same amount of effort – one evaluation of $f(x)$ per step.

To get a first impression of the error behaviour of these methods, we apply them to a problem whose answer we know exactly:

$$\int_0^\pi \sin x\,\mathrm{d}x = -\cos x\Big|_0^\pi = 2.$$

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.

	Midpoint		**Trapezoidal**		**Simpson's**	
n	error	# evals	error	# evals	error	# evals
10	8.2×10^{-3}	10	1.6×10^{-2}	11	1.1×10^{-4}	11
100	8.2×10^{-5}	100	1.6×10^{-4}	101	1.1×10^{-8}	101
1000	8.2×10^{-7}	1000	1.6×10^{-6}	1001	1.1×10^{-12}	1001

80 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!

Observe that

- Using 101 evaluations of f worth of Simpson's rule gives an error 75 times smaller than 1000 evaluations of f worth of the midpoint rule.

- The trapezoidal rule error with n steps is about twice the midpoint rule error with n steps.

- 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 $10^2 = n^2$.

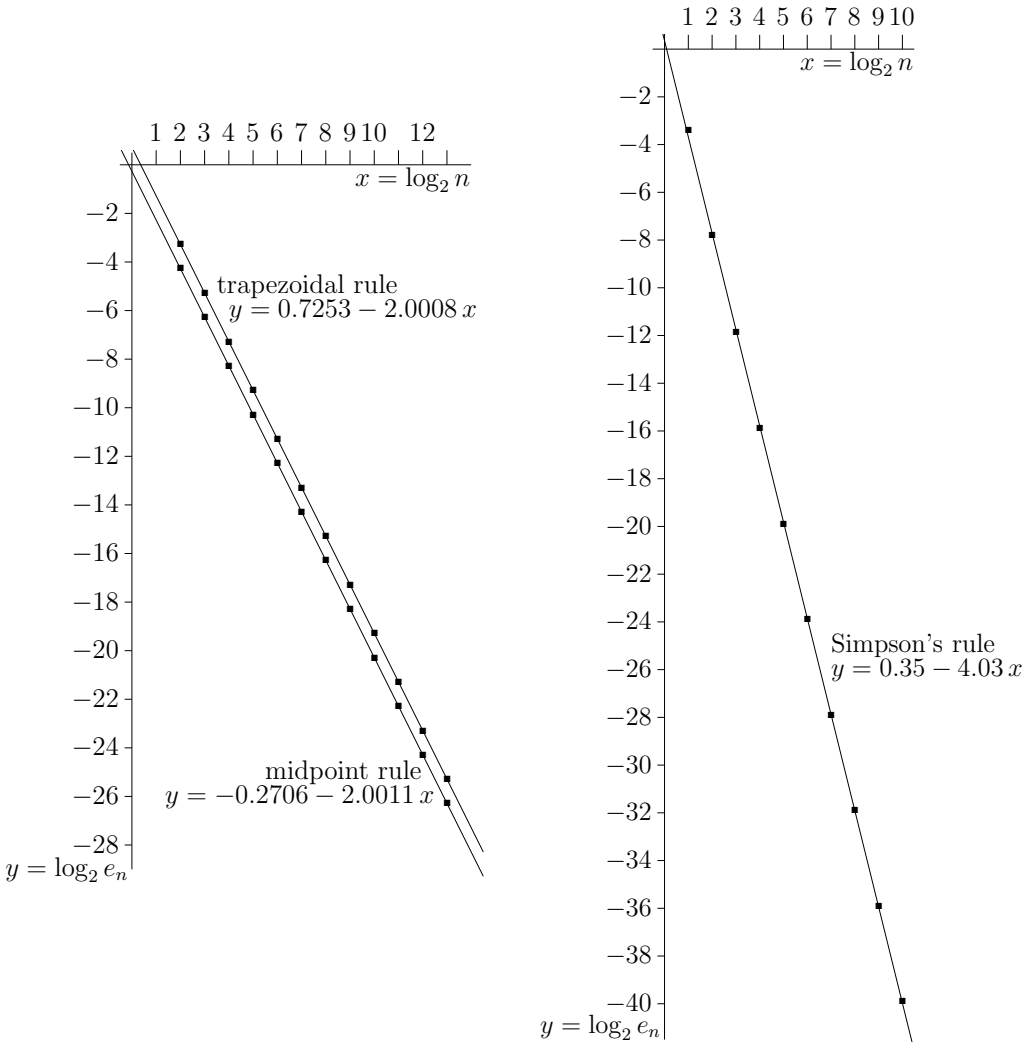
- With Simpson's rule, increasing the number of steps by a factor of 10 appears to reduce the error by about a factor of $10^4 = n^4$.

So it looks like

$$\begin{aligned}
\text{approx value of } \int_a^b f(x)\,\mathrm{d}x \text{ given by } n \text{ midpoint steps} &\approx \int_a^b f(x)\,\mathrm{d}x + K_M \cdot \frac{1}{n^2} \\
\text{approx value of } \int_a^b f(x)\,\mathrm{d}x \text{ given by } n \text{ trapezoidal steps} &\approx \int_a^b f(x)\,\mathrm{d}x + K_T \cdot \frac{1}{n^2} \\
\text{approx value of } \int_a^b f(x)\,\mathrm{d}x \text{ given by } n \text{ Simpson's steps} &\approx \int_a^b f(x)\,\mathrm{d}x + K_S \cdot \frac{1}{n^4}
\end{aligned}$$

with some constants K_M, K_T and K_S. It also seems that $K_T \approx 2K_M$.

Figure 1.11.1.



A log-log plot of the error in the n step approximation to $\int_0^\pi \sin x \, dx$.

To test these conjectures for the behaviour of the errors we apply our three rules with about ten different choices of n of the form $n = 2^m$ with m integer. Figure 1.11.1 contains two graphs of the results. The left-hand plot shows the results for the midpoint and trapezoidal rules and the right-hand plot shows the results for Simpson's rule.

For each rule we are expecting (based on our conjectures above) that the error

$$e_n = |\text{exact value} - \text{approximate value}|$$

with n steps is (roughly) of the form

$$e_n = K\frac{1}{n^k}$$

for some constants K and k. We would like to test if this is really the case, by graphing $Y = e_n$ against $X = n$ and seeing if the graph "looks right". But it is not easy to tell

whether or not a given curve really is $Y = \frac{K}{X^k}$, for some specific k, by just looking at it. However, your eye is pretty good at determining whether or not a graph is a straight line. Fortunately, there is a little trick that turns the curve $Y = \frac{K}{X^k}$ into a straight line – no matter what k is.

Instead of plotting Y against X, we plot $\log Y$ against $\log X$. This transformation[81] works because when $Y = \frac{K}{X^k}$

$$\log Y = \log K - k \log X$$

So plotting $y = \log Y$ against $x = \log X$ gives the straight line $y = \log K - kx$, which has slope $-k$ and y–intercept $\log K$.

The three graphs in Figure 1.11.1 plot $y = \log_2 e_n$ against $x = \log_2 n$ for our three rules. Note that we have chosen to use logarithms[82] with this "unusual base" because it makes it very clear how much the error is improved if we *double* the number of steps used. To be more precise — one unit step along the x-axis represents changing $n \mapsto 2n$. For example, applying Simpson's rule with $n = 2^4$ steps results in an error of 0.0000166, so the point $(x = \log_2 2^4 = 4, y = \log_2 0.0000166 = \frac{\log 0.0000166}{\log 2} = -15.9)$ has been included on the graph. Doubling the effort used — that is, doubling the number of steps to $n = 2^5$— results in an error of 0.00000103. So, the data point $(x = \log_2 2^5 = 5\,,\ y = \log_2 0.00000103 = \frac{\ln 0.00000103}{\ln 2} = -19.9)$ lies on the graph. Note that the x-coordinates of these points differ by 1 unit.

For each of the three sets of data points, a straight line has also been plotted "through" the data points. A procedure called linear regression[83] has been used to decide precisely which straight line to plot. It provides a formula for the slope and y–intercept of the straight line which "best fits" any given set of data points. From the three lines, it sure looks like $k = 2$ for the midpoint and trapezoidal rules and $k = 4$ for Simpson's rule. It also looks like the ratio between the value of K for the trapezoidal rule, namely $K = 2^{0.7253}$, and the value of K for the midpoint rule, namely $K = 2^{-0.2706}$, is pretty close to 2: $2^{0.7253}/2^{-0.2706} = 2^{0.9959}$.

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

81 There is a variant of this trick that works even when you don't know the answer to the integral ahead of time. Suppose that you suspect that the approximation satisfies

$$M_n = A + K\tfrac{1}{n^k}$$

where A is the exact value of the integral and suppose that you don't know the values of A, K and k. Then

$$M_n - M_{2n} = K\tfrac{1}{n^k} - K\tfrac{1}{(2n)^k} = K\big(1 - \tfrac{1}{2^k}\big)\tfrac{1}{n^k}$$

so plotting $y = \log(M_n - M_{2n})$ against $x = \log n$ gives the straight line $y = \log\big[K\big(1 - \frac{1}{2^k}\big)\big] - kx$.

82 Now is a good time for a quick revision of logarithms — see "Whirlwind review of logarithms" in Section 2.7 of the CLP-1 text.

83 Linear regression is not part of this course as its derivation requires some multivariable calculus. It is a very standard technique in statistics.

Theorem 1.11.12 (Numerical integration errors).

Assume that $|f''(x)| \leqslant M$ for all $a \leqslant x \leqslant b$. Then

the total error introduced by the midpoint rule is bounded by $\dfrac{M}{24}\dfrac{(b-a)^3}{n^2}$

and

the total error introduced by the trapezoidal rule is bounded by $\dfrac{M}{12}\dfrac{(b-a)^3}{n^2}$

when approximating $\displaystyle\int_a^b f(x)\mathrm{d}x$. Further, if $|f^{(4)}(x)| \leqslant L$ for all $a \leqslant x \leqslant b$, then

the total error introduced by Simpson's rule is bounded by $\dfrac{L}{180}\dfrac{(b-a)^5}{n^4}$.

The first of these error bounds in proven in the following (optional) section. Here are some examples which illustrate how they are used. First let us check that the above result is consistent with our data in Figure 1.11.1

Example 1.11.13 (Midpoint rule error approximating $\int_0^\pi \sin x\,\mathrm{d}x$)

- The integral $\int_0^\pi \sin x\,\mathrm{d}x$ has $b-a=\pi$.
- The second derivative of the integrand satisfies

$$\left|\frac{\mathrm{d}^2}{\mathrm{d}x^2}\sin x\right| = |-\sin x| \leqslant 1$$

  So we take $M=1$.

- So the error, e_n, introduced when n steps are used is bounded by

$$\begin{aligned}|e_n| &\leqslant \frac{M}{24}\frac{(b-a)^3}{n^2}\\ &= \frac{\pi^3}{24}\frac{1}{n^2}\\ &\approx 1.29\frac{1}{n^2}\end{aligned}$$

- The data in the graph in Figure 1.11.1 gives

$$|e_n| \approx 2^{-.2706}\frac{1}{n^2} = 0.83\frac{1}{n^2}$$

  which is consistent with the bound $|e_n| \leqslant \frac{\pi^3}{24}\frac{1}{n^2}$.

Example 1.11.13

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 $\frac{1}{10}^{\text{th}}$ of a millimeter, there is no point in making design specifications more accurate than $\frac{1}{10}^{\text{th}}$ of a millimeter.

Example 1.11.14

Suppose, for example, that we wish to use the midpoint rule to evaluate[84]

$$\int_0^1 e^{-x^2}\mathrm{d}x$$

to within an accuracy of 10^{-6}.

Solution.

- The integral has $a = 0$ and $b = 1$.
- The first two derivatives of the integrand are

$$\frac{\mathrm{d}}{\mathrm{d}x}e^{-x^2} = -2xe^{-x^2} \qquad \text{and}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}e^{-x^2} = \frac{\mathrm{d}}{\mathrm{d}x}\big(-2xe^{-x^2}\big) = -2e^{-x^2} + 4x^2e^{-x^2} = 2(2x^2-1)e^{-x^2}$$

- As x runs from 0 to 1, $2x^2 - 1$ increases from -1 to 1, so that

$$0 \leqslant x \leqslant 1 \implies |2x^2-1| \leqslant 1,\ e^{-x^2} \leqslant 1 \implies \left|2(2x^2-1)e^{-x^2}\right| \leqslant 2$$

  So we take $M = 2$.

- The error introduced by the n step midpoint rule is at most

$$\begin{aligned} e_n &\leqslant \frac{M}{24}\frac{(b-a)^3}{n^2} \\ &\leqslant \frac{2}{24}\frac{(1-0)^3}{n^2} = \frac{1}{12n^2} \end{aligned}$$

- We need this error to be smaller than 10^{-6} so

$$\begin{aligned} e_n \leqslant \frac{1}{12n^2} &\leqslant 10^{-6} && \text{and so} \\ 12n^2 &\geqslant 10^6 && \text{clean up} \\ n^2 &\geqslant \frac{10^6}{12} = 83333.3\ldots && \text{square root both sides} \\ n &\geqslant 288.7 \end{aligned}$$

  So 289 steps of the midpoint rule will do the job.

84 This is our favourite running example of an integral that cannot be evaluated algebraically — we need to use numerical methods.

- In fact $n = 289$ results in an error of about 3.7×10^{-7}.

Example 1.11.14

That seems like far too much work, and the trapezoidal rule will have twice the error. So we should look at Simpson's rule.

Example 1.11.15

Suppose now that we wish evaluate $\int_0^1 e^{-x^2}\,dx$ to within an accuracy of 10^{-6} — but now using Simpson's rule. How many steps should we use?

Solution.

- Again we have $a = 0, b = 1$.
- We then need to bound $\frac{d^4}{dx^4}e^{-x^2}$ on the domain of integration, $0 \leqslant x \leqslant 1$.

$$\begin{aligned}\frac{d^3}{dx^3}e^{-x^2} &= \frac{d}{dx}\{2(2x^2-1)e^{-x^2}\} = 8xe^{-x^2} - 4x(2x^2-1)e^{-x^2}\\ &= 4(-2x^3+3x)e^{-x^2}\\ \frac{d^4}{dx^4}e^{-x^2} &= \frac{d}{dx}\{4(-2x^3+3x)e^{-x^2}\} = 4(-6x^2+3)e^{-x^2} - 8x(-2x^3+3x)e^{-x^2}\\ &= 4(4x^4-12x^2+3)e^{-x^2}\end{aligned}$$

- Now, for any x, $e^{-x^2} \leqslant 1$. Also, for $0 \leqslant x \leqslant 1$,

$$\begin{aligned} 0 \leqslant x^2, x^4 &\leqslant 1 && \text{so}\\ 3 \leqslant 4x^4 + 3 &\leqslant 7 && \text{and}\\ -12 \leqslant -12x^2 &\leqslant 0 && \text{adding these together gives}\\ -9 \leqslant 4x^4 - 12x^2 + 3 &\leqslant 7 \end{aligned}$$

  Consequently, $|4x^4 - 12x^2 + 3|$ is bounded by 9 and so

$$\left|\frac{d^4}{dx^4}e^{-x^2}\right| \leqslant 4 \times 9 = 36$$

  So take $L = 36$.

- The error introduced by the n step Simpson's rule is at most

$$\begin{aligned} e_n &\leqslant \frac{L}{180}\frac{(b-a)^5}{n^4}\\ &\leqslant \frac{36}{180}\frac{(1-0)^5}{n^4} = \frac{1}{5n^4}\end{aligned}$$

- In order for this error to be no more than 10^{-6} we require n to satisfy

$$\begin{aligned} e_n &\leqslant \frac{1}{5n^4} \leqslant 10^{-6} && \text{and so} \\ 5n^4 &\geqslant 10^6 \\ n^4 &\geqslant 200000 && \text{take fourth root} \\ n &\geqslant 21.15 \end{aligned}$$

  So 22 steps of Simpson's rule will do the job.

- $n = 22$ steps actually results in an error of 3.5×10^{-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 $\left|\frac{\mathrm{d}^4}{\mathrm{d}x^4} f(x)\right| \leqslant 36$. If we are more careful then we will get a slightly smaller n. It actually turns out[85] that you only need $n = 10$ to approximate within 10^{-6}.

Example 1.11.15

1.11.5 ▸▸ Optional — An Error Bound for the Midpoint Rule

We now try develop some understanding as to why we got the above experimental results. We start with the error generated by a single step of the midpoint rule. That is, the error introduced by the approximation

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x \approx f(\bar{x}_1)\Delta x \qquad \text{where } \Delta x = x_1 - x_0,\ \bar{x}_1 = \tfrac{x_0+x_1}{2}$$

To do this we are going to need to apply integration by parts in a sneaky way. Let us start by considering[86] a subinterval $\alpha \leqslant x \leqslant \beta$ 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

$$\int_\alpha^\beta f(x)\mathrm{d}x \approx 2q \cdot f(\alpha + q) = qf(\alpha + q) + qf(\beta - q)$$

since the interval has width $2q$ and the midpoint is $\alpha + q = \beta - q$.

The sneaky trick we will employ is to write

$$\int_\alpha^\beta f(x)\mathrm{d}x = \int_\alpha^{\alpha+q} f(x)\mathrm{d}x + \int_{\beta-q}^\beta f(x)\mathrm{d}x$$

85 The authors tested this empirically.
86 We chose this interval so that we didn't have lots of subscripts floating around in the algebra.

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

$$\int_{\alpha}^{\alpha+q} f(x)\mathrm{d}x \approx qf(\alpha+q) + \text{small error term}$$
$$\int_{\beta-q}^{\beta} f(x)\mathrm{d}x \approx qf(\beta-q) + \text{small error term}$$

Let us apply integration by parts to $\int_{\alpha}^{\alpha+q} f(x)\mathrm{d}x$ — with $u = f(x), \mathrm{d}v = \mathrm{d}x$ so $\mathrm{d}u = f'(x)\mathrm{d}x$ and we will make the slightly non-standard choice of $v = x - \alpha$:

$$\begin{aligned}\int_{\alpha}^{\alpha+q} f(x)\mathrm{d}x &= \left[(x-\alpha)f(x)\right]_{\alpha}^{\alpha+q} - \int_{\alpha}^{\alpha+q}(x-\alpha)f'(x)\mathrm{d}x \\ &= qf(\alpha+q) - \int_{\alpha}^{\alpha+q}(x-\alpha)f'(x)\mathrm{d}x\end{aligned}$$

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 $u = f'(x), \mathrm{d}v = (x-\alpha)\mathrm{d}x$, so $\mathrm{d}u = f''(x), v = \frac{(x-\alpha)^2}{2}$:

$$\begin{aligned}\int_{\alpha}^{\alpha+q} f(x)\mathrm{d}x &= qf(\alpha+q) - \int_{\alpha}^{\alpha+q}(x-\alpha)f'(x)\mathrm{d}x \\ &= qf(\alpha+q) - \left[\frac{(x-\alpha)^2}{2}f'(x)\right]_{\alpha}^{\alpha+q} + \int_{\alpha}^{\alpha+q}\frac{(x-\alpha)^2}{2}f''(x)\mathrm{d}x \\ &= qf(\alpha+q) - \frac{q^2}{2}f'(\alpha+q) + \int_{\alpha}^{\alpha+q}\frac{(x-\alpha)^2}{2}f''(x)\mathrm{d}x\end{aligned}$$

To obtain a similar expression for the other integral, we repeat the above steps and obtain:

$$\int_{\beta-q}^{\beta} f(x)\mathrm{d}x = qf(\beta-q) + \frac{q^2}{2}f'(\beta-q) + \int_{\beta-q}^{\beta}\frac{(x-\beta)^2}{2}f''(x)\mathrm{d}x$$

Now add together these two expressions

$$\begin{aligned}\int_{\alpha}^{\alpha+q} f(x)\mathrm{d}x + \int_{\beta-q}^{\beta} f(x)\mathrm{d}x &= qf(\alpha+q) + qf(\beta-q) + \frac{q^2}{2}\left(f'(\beta-q) - f'(\alpha+q)\right) \\ &\quad + \int_{\alpha}^{\alpha+q}\frac{(x-\alpha)^2}{2}f''(x)\mathrm{d}x + \int_{\beta-q}^{\beta}\frac{(x-\beta)^2}{2}f''(x)\mathrm{d}x\end{aligned}$$

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:

$$\int_{\alpha}^{\beta} f(x)\mathrm{d}x = 2qf(\alpha+q) + \int_{\alpha}^{\alpha+q}\frac{(x-\alpha)^2}{2}f''(x)\mathrm{d}x + \int_{\beta-q}^{\beta}\frac{(x-\beta)^2}{2}f''(x)\mathrm{d}x$$

Rearrange this expression a little and take absolute values

$$\left|\int_\alpha^\beta f(x)\mathrm{d}x - 2qf(\alpha+q)\right| \leqslant \left|\int_\alpha^{\alpha+q} \frac{(x-\alpha)^2}{2} f''(x)\mathrm{d}x\right| + \left|\int_{\beta-q}^{\beta} \frac{(x-\beta)^2}{2} f''(x)\mathrm{d}x\right|$$

where we have also made use of the triangle inequality[87]. By assumption $|f''(x)| \leqslant M$ on the interval $\alpha \leqslant x \leqslant \beta$, so

$$\begin{aligned}\left|\int_\alpha^\beta f(x)\mathrm{d}x - 2qf(\alpha+q)\right| &\leqslant M\int_\alpha^{\alpha+q} \frac{(x-\alpha)^2}{2}\mathrm{d}x + M\int_{\beta-q}^{\beta} \frac{(x-\beta)^2}{2}\mathrm{d}x \\ &= \frac{Mq^3}{3} = \frac{M(\beta-\alpha)^3}{24}\end{aligned}$$

where we have used $q = \frac{\beta-\alpha}{2}$ in the last step.

Thus on any interval $x_i \leqslant x \leqslant x_{i+1} = x_i + \Delta x$

$$\left|\int_{x_i}^{x_{i+1}} f(x)\mathrm{d}x - \Delta x f\left(\frac{x_i + x_{i+1}}{2}\right)\right| \leqslant \frac{M}{24}(\Delta x)^3$$

Putting everything together we see that the error using the midpoint rule is bounded by

$$\begin{aligned}&\left|\int_a^b f(x)\mathrm{d}x - \left[f(\bar{x}_1) + f(\bar{x}_2) + \cdots + f(\bar{x}_n)\right]\Delta x\right| \\ &\qquad \leqslant \left|\int_{x_0}^{x_1} f(x)\mathrm{d}x - \Delta x f(\bar{x}_1)\right| + \cdots + \left|\int_{x_{n-1}}^{x_n} f(x)\mathrm{d}x - \Delta x f(\bar{x}_n)\right| \\ &\qquad\qquad \leqslant n \times \frac{M}{24}(\Delta x)^3 = n \times \frac{M}{24}\left(\frac{b-a}{n}\right)^3 = \frac{M(b-a)^3}{24n^2}\end{aligned}$$

as required.

A very similar analysis shows that, as was stated in Theorem 1.11.12 above,

- the total error introduced by the trapezoidal rule is bounded by $\frac{M}{12}\frac{(b-a)^3}{n^2}$,
- the total error introduced by Simpson's rule is bounded by $\frac{M}{180}\frac{(b-a)^5}{n^4}$

1.12▲ Improper Integrals

1.12.1 ▸▸ Definitions

To this point we have only considered nicely behaved integrals $\int_a^b f(x)\mathrm{d}x$. Though the algebra involved in some of our examples was quite difficult, all the integrals had

87 The triangle inequality says that for any real numbers x, y

$$|x+y| \leqslant |x| + |y|.$$

- 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.

Definition 1.12.1.

An integral having either an infinite limit of integration or an unbounded integrand is called an improper integral.

Two examples are

$$\int_0^\infty \frac{\mathrm{d}x}{1+x^2} \qquad \text{and} \qquad \int_0^1 \frac{\mathrm{d}x}{x}$$

The first has an infinite domain of integration and the integrand of the second tends to ∞ 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 1.12.2 $\left(\int_{-1}^1 \frac{1}{x^2}\mathrm{d}x\right)$

Consider the integral

$$\int_{-1}^1 \frac{1}{x^2}\mathrm{d}x$$

If we "do" this integral completely naively then we get

$$\begin{aligned}\int_{-1}^1 \frac{1}{x^2}\,\mathrm{d}x &= \frac{x^{-1}}{-1}\Big|_{-1}^1 \\ &= \frac{1}{-1} - \frac{-1}{-1} \\ &= -2\end{aligned}$$

which is *wrong*[88]. 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

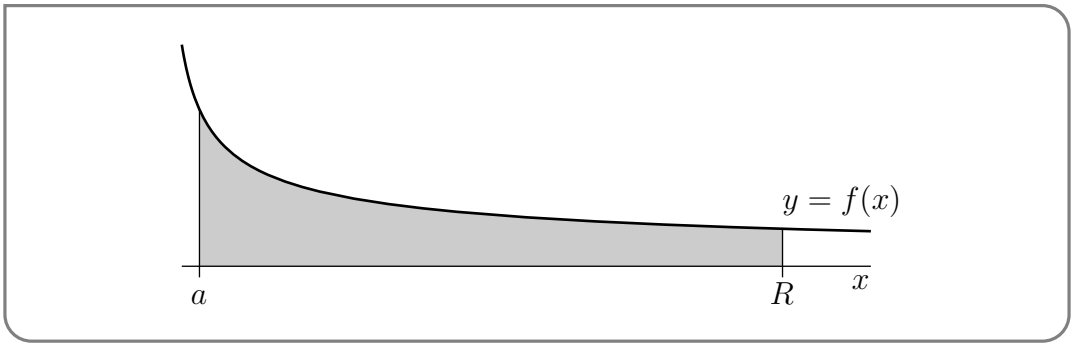
$$\text{if } F'(x) = f(x) \text{ then } \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$$

88 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 searchengineing and look at the concept of falisfyability.

is applicable only when $F'(x)$ exists and equals $f(x)$ for *all* $a \leqslant x \leqslant b$. 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 1.12.2

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 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 $R \to \infty$. Let us put this into practice:



Example 1.12.3 $\left(\int_a^\infty \frac{dx}{1+x^2}\right)$

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 $R \to +\infty$:

$$\begin{aligned}\int_a^\infty \frac{dx}{1+x^2} &= \lim_{R\to\infty} \int_a^R \frac{dx}{1+x^2} \\ &= \lim_{R\to\infty} \big[\arctan R - \arctan a\big] \\ &= \frac{\pi}{2} - \arctan a.\end{aligned}$$

Example 1.12.3

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 1.12.4 (Improper integral with infinite domain of integration).

(a) If the integral $\int_a^R f(x)\,\mathrm{d}x$ exists for all $R > a$, then

$$\int_a^\infty f(x)\,\mathrm{d}x = \lim_{R\to\infty}\int_a^R f(x)\,\mathrm{d}x$$

when the limit exists (and is finite).

(b) If the integral $\int_r^b f(x)\,\mathrm{d}x$ exists for all $r < b$, then

$$\int_{-\infty}^b f(x)\,\mathrm{d}x = \lim_{r\to-\infty}\int_r^b f(x)\,\mathrm{d}x$$

when the limit exists (and is finite).

(c) If the integral $\int_r^R f(x)\,\mathrm{d}x$ exists for all $r < R$, then

$$\int_{-\infty}^\infty f(x)\,\mathrm{d}x = \lim_{r\to-\infty}\int_r^c f(x)\,\mathrm{d}x + \lim_{R\to\infty}\int_c^R f(x)\,\mathrm{d}x$$

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 deal with an integral like $\int_0^1 \frac{\mathrm{d}x}{x}$ that has a finite domain of integration, but whose integrand is unbounded near one limit of integration[89]. Our approach is similar — we sneak up on the problem. We compute the integral on a smaller domain, such as $\int_t^1 \frac{\mathrm{d}x}{x}$, with $t > 0$, and then take the limit $t \to 0+$.

Example 1.12.5 $\left(\int_0^1 \frac{1}{x}\mathrm{d}x\right)$

Solution.

- Since the integrand is unbounded near $x = 0$, we integrate on the smaller domain $t \leqslant x \leqslant 1$ with $t > 0$:

$$\int_t^1 \frac{1}{x}\mathrm{d}x = \log|x|\Big|_t^1 = -\log|t|$$

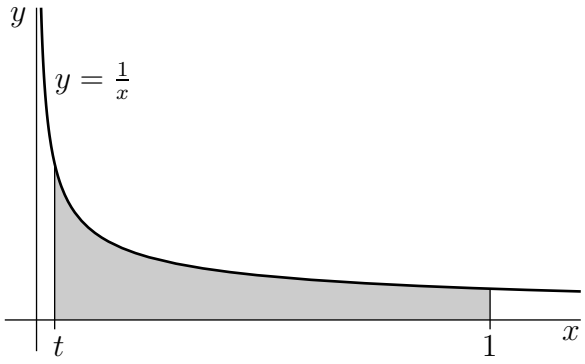
- We then take the limit as $t \to 0^+$ to obtain

$$\int_0^1 \frac{1}{x}\mathrm{d}x = \lim_{t\to0^+}\int_t^1 \frac{1}{x}\mathrm{d}x = \lim_{t\to0^+} -\log|t| = +\infty$$

89 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 1.12.5



Indeed, we *define* integrals with unbounded integrands via this process:

Definition 1.12.6 (Improper integral with unbounded integrand).

(a) If the integral $\int_t^b f(x)\,\mathrm{d}x$ exists for all $a < t < b$, then

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{t\to a+} \int_t^b f(x)\,\mathrm{d}x$$

when the limit exists (and is finite).

(b) If the integral $\int_a^T f(x)\,\mathrm{d}x$ exists for all $a < T < b$, then

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{T\to b-} \int_a^T f(x)\,\mathrm{d}x$$

when the limit exists (and is finite).

(c) Let $a < c < b$. If the integrals $\int_a^T f(x)\,\mathrm{d}x$ and $\int_t^b f(x)\,\mathrm{d}x$ exist for all $a < T < c$ and $c < t < b$, then

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{T\to c-} \int_a^T f(x)\,\mathrm{d}x + \lim_{t\to c+} \int_t^b f(x)\,\mathrm{d}x$$

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

$$\int_{-1}^{1} \frac{1}{x^2} \mathrm{d}x$$

A quick computation shows that this integral diverges to $+\infty$

$$\begin{aligned}\int_{-1}^{1} \frac{1}{x^2} \mathrm{d}x &= \lim_{a\to 0^-} \int_{-1}^{a} \frac{1}{x^2} \mathrm{d}x + \lim_{b\to 0^+} \int_{b}^{1} \frac{1}{x^2} \mathrm{d}x \\ &= \lim_{a\to 0^-} \left[1 - \frac{1}{a}\right] + \lim_{b\to 0^+} \left[\frac{1}{b} - 1\right] \\ &= +\infty\end{aligned}$$

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 1.12.7 $\left(\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{(x-2)x^2}\right)$

Consider the integral

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{(x-2)x^2}$$

- 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

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{(x-2)x^2} = \int_{-\infty}^{a} \frac{\mathrm{d}x}{(x-2)x^2} &+ \int_{a}^{0} \frac{\mathrm{d}x}{(x-2)x^2} + \int_{0}^{b} \frac{\mathrm{d}x}{(x-2)x^2} \\ &+ \int_{b}^{2} \frac{\mathrm{d}x}{(x-2)x^2} + \int_{2}^{c} \frac{\mathrm{d}x}{(x-2)x^2} + \int_{c}^{\infty} \frac{\mathrm{d}x}{(x-2)x^2}\end{aligned}$$

  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 $x \to 0-$,
  - the third integral has an integrand that becomes unbounded as $x \to 0+$,
  - the fourth integral has an integrand that becomes unbounded as $x \to 2-$,
  - the fifth integral has an integrand that becomes unbounded as $x \to 2+$, 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 1.12.7

1.12.2 ▸▸ Examples

With the more formal definitions out of the way, we are now ready for some (important) examples.

Example 1.12.8 $\left(\int_1^\infty \frac{\mathrm{d}x}{x^p} \text{ with } p > 0\right)$

Solution.

- Fix any $p > 0$.
- The domain of the integral $\int_1^\infty \frac{\mathrm{d}x}{x^p}$ 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
$$\int_1^\infty \frac{\mathrm{d}x}{x^p} = \lim_{R\to\infty} \int_1^R \frac{\mathrm{d}x}{x^p}$$
- 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 $p > 1$,
$$\begin{aligned}\int_1^R \frac{\mathrm{d}x}{x^p} &= \frac{1}{1-p} x^{1-p}\Big|_1^R \\ &= \frac{R^{1-p}-1}{1-p}\end{aligned}$$
Taking the limit as $R \to \infty$ gives
$$\begin{aligned}\int_1^\infty \frac{\mathrm{d}x}{x^p} &= \lim_{R\to\infty} \int_1^R \frac{\mathrm{d}x}{x^p} \\ &= \lim_{R\to\infty} \frac{R^{1-p}-1}{1-p} \\ &= \frac{-1}{1-p} = \frac{1}{p-1}\end{aligned}$$

since $1-p<0$.

- Similarly when $p<1$ we have

$$\int_1^\infty \frac{\mathrm{d}x}{x^p} = \lim_{R\to\infty}\int_1^R \frac{\mathrm{d}x}{x^p} \qquad = \lim_{R\to\infty}\frac{R^{1-p}-1}{1-p}$$
$$= +\infty$$

because $1-p>0$ and the term R^{1-p} diverges to $+\infty$.

- Finally when $p=1$

$$\int_1^R \frac{\mathrm{d}x}{x} = \log|R| - \log 1 = \log R$$

Then taking the limit as $R\to\infty$ gives us

$$\int_1^\infty \frac{\mathrm{d}x}{x^p} = \lim_{R\to\infty}\log|R| = +\infty.$$

- So summarising, we have

$$\int_1^\infty \frac{\mathrm{d}x}{x^p} = \begin{cases}\text{divergent} & \text{if } p\leqslant 1\\ \frac{1}{p-1} & \text{if } p>1\end{cases}$$

Example 1.12.8

Example 1.12.9 $\left(\int_0^1 \frac{\mathrm{d}x}{x^p} \text{ with } p>0\right)$

Solution.

- Again fix any $p>0$.
- The domain of integration of the integral $\int_0^1 \frac{\mathrm{d}x}{x^p}$ 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

$$\int_0^1 \frac{\mathrm{d}x}{x^p} = \lim_{t\to 0+}\int_t^1 \frac{\mathrm{d}x}{x^p}$$

- Again, the antiderivative changes at $p=1$, so we split the problem into three cases.
- When $p>1$ we have

$$\int_t^1 \frac{\mathrm{d}x}{x^p} = \frac{1}{1-p}x^{1-p}\Big|_t^1$$
$$= \frac{1-t^{1-p}}{1-p}$$

Since $1-p<0$ when we take the limit as $t\to 0$ the term t^{1-p} diverges to $+\infty$ and we obtain

$$\int_0^1 \frac{\mathrm{d}x}{x^p} = \lim_{t\to 0^+} \frac{1-t^{1-p}}{1-p} = +\infty$$

- When $p=1$ we similarly obtain

$$\begin{aligned}\int_0^1 \frac{\mathrm{d}x}{x} &= \lim_{t\to 0+} \int_t^1 \frac{\mathrm{d}x}{x} \\ &= \lim_{t\to 0+} \big(-\log|t|\big) \\ &= +\infty\end{aligned}$$

- Finally, when $p<1$ we have

$$\begin{aligned}\int_0^1 \frac{\mathrm{d}x}{x^p} &= \lim_{t\to 0^+} \int_t^1 \frac{\mathrm{d}x}{x^p} \\ &= \lim_{t\to 0^+} \frac{1-t^{1-p}}{1-p} = \frac{1}{1-p}\end{aligned}$$

since $1-p>0$.

- In summary

$$\int_0^1 \frac{\mathrm{d}x}{x^p} = \begin{cases} \frac{1}{1-p} & \text{if } p<1 \\ \text{divergent} & \text{if } p\geqslant 1 \end{cases}$$

Example 1.12.9

Example 1.12.10 $\left(\int_0^\infty \frac{\mathrm{d}x}{x^p} \text{ with } p>0\right)$

Solution.

- Yet again fix $p>0$.
- This time the domain of integration of the integral $\int_0^\infty \frac{\mathrm{d}x}{x^p}$ 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 $x=1$:

$$\int_0^\infty \frac{\mathrm{d}x}{x^p} = \int_0^1 \frac{\mathrm{d}x}{x^p} + \int_1^\infty \frac{\mathrm{d}x}{x^p}$$

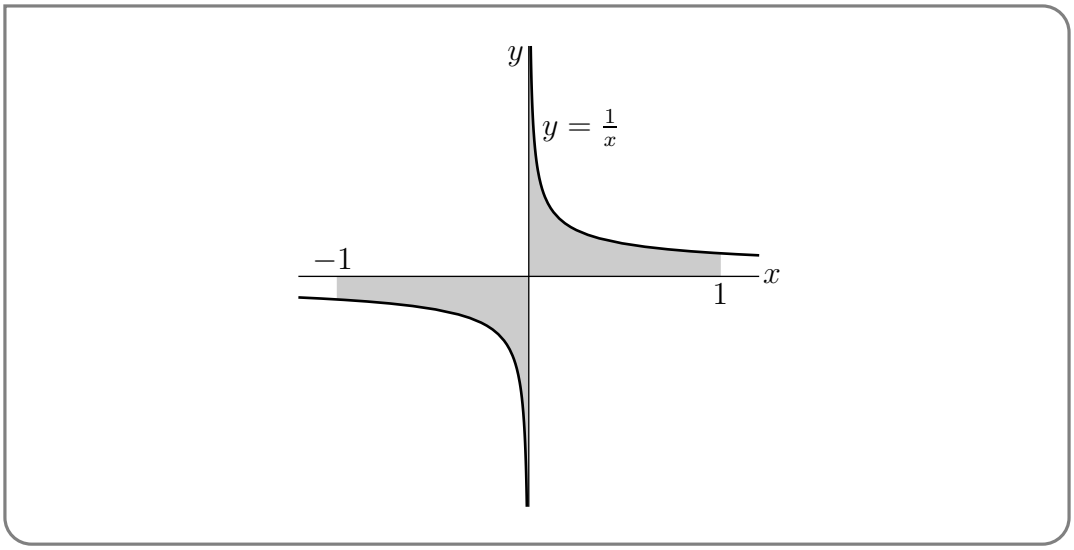
- We saw, in Example 1.12.9, that the first integral diverged whenever $p\geqslant 1$, and we also saw, in Example 1.12.8, that the second integral diverged whenever $p\leqslant 1$.

- So the integral $\int_0^\infty \frac{dx}{x^p}$ diverges for all values of p.

Example 1.12.10

Example 1.12.11 $\left(\int_{-1}^{1} \frac{dx}{x}\right)$

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

$$\int_0^1 \frac{dx}{x} = \lim_{t\to 0+} \int_t^1 \frac{dx}{x} = \lim_{t\to 0+} \Big[\log x\Big]_t^1 = \lim_{t\to 0+} \log \frac{1}{t} = +\infty$$
$$\int_{-1}^0 \frac{dx}{x} = \lim_{T\to 0-} \int_{-1}^T \frac{dx}{x} = \lim_{T\to 0-} \Big[\log|x|\Big]_{-1}^T = \lim_{T\to 0-} \log|T| = -\infty$$

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

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

and that the area to the left of the y–axis is (substitute $-7t$ for T above)

$$\lim_{t\to 0+} \int_{-1}^{-7t} \frac{dx}{x} = -\infty$$

If $\infty - \infty = 0$, the following limit should be 0.

$$\begin{aligned}\lim_{t\to 0+}\left[\int_t^1 \frac{\mathrm{d}x}{x} + \int_{-1}^{-7t}\frac{\mathrm{d}x}{x}\right] &= \lim_{t\to 0+}\left[\log\frac{1}{t} + \log|-7t|\right]\\ &= \lim_{t\to 0+}\left[\log\frac{1}{t} + \log(7t)\right]\\ &= \lim_{t\to 0+}\left[-\log t + \log 7 + \log t\right] = \lim_{t\to 0+}\log 7\\ &= \log 7\end{aligned}$$

This appears to give $\infty - \infty = \log 7$. Of course the number 7 was picked at random. You can make $\infty - \infty$ be any number at all, by making a suitable replacement for 7.

Example 1.12.11

Example 1.12.12 (Example 1.12.2 revisited)

The careful computation of the integral of Example 1.12.2 is

$$\begin{aligned}\int_{-1}^1 \frac{1}{x^2}\,\mathrm{d}x &= \lim_{T\to 0-}\int_{-1}^T \frac{1}{x^2}\,\mathrm{d}x + \lim_{t\to 0+}\int_t^1 \frac{1}{x^2}\,\mathrm{d}x\\ &= \lim_{T\to 0-}\left[-\frac{1}{x}\right]_{-1}^T + \lim_{t\to 0+}\left[-\frac{1}{x}\right]_t^1\\ &= \infty + \infty\end{aligned}$$

Hence the integral diverges to $+\infty$.

Example 1.12.12

Example 1.12.13 $\left(\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{1+x^2}\right)$

Since

$$\begin{aligned}\lim_{R\to\infty}\int_0^R \frac{\mathrm{d}x}{1+x^2} &= \lim_{R\to\infty}\Big[\arctan x\Big]_0^R = \lim_{R\to\infty}\arctan R = \frac{\pi}{2}\\ \lim_{r\to-\infty}\int_r^0 \frac{\mathrm{d}x}{1+x^2} &= \lim_{r\to-\infty}\Big[\arctan x\Big]_r^0 = \lim_{r\to-\infty} -\arctan r = \frac{\pi}{2}\end{aligned}$$

The integral $\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{1+x^2}$ converges and takes the value π.

Example 1.12.13

Example 1.12.14

For what values of p does $\int_e^{\infty}\frac{\mathrm{d}x}{x(\log x)^p}$ converge?

Solution.

- For $x \geqslant e$, the denominator $x(\log x)^p$ 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

$$\begin{aligned}
\int_e^\infty \frac{\mathrm{d}x}{x(\log x)^p} &= \lim_{R\to\infty} \int_e^R \frac{\mathrm{d}x}{x(\log x)^p} && \text{use substitution} \\
&= \lim_{R\to\infty} \int_1^{\log R} \frac{\mathrm{d}u}{u^p} && \text{with } u = \log x, \mathrm{d}u = \frac{\mathrm{d}x}{x} \\
&= \lim_{R\to\infty} \begin{cases} \frac{1}{1-p}\Big[(\log R)^{1-p} - 1\Big] & \text{if } p \neq 1 \\ \log(\log R) & \text{if } p = 1 \end{cases} \\
&= \begin{cases} \text{divergent} & \text{if } p \leqslant 1 \\ \frac{1}{p-1} & \text{if } p > 1 \end{cases}
\end{aligned}$$

  In this last step we have used similar logic that that used in Example 1.12.8, but with R replaced by $\log R$.

Example 1.12.14

Example 1.12.15 (the gamma function)

The gamma function $\Gamma(x)$ is defined by the improper integral

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, \mathrm{d}x$$

We shall now compute $\Gamma(n)$ for all natural numbers n.

- To get started, we'll compute

$$\Gamma(1) = \int_0^\infty e^{-x}\, \mathrm{d}x = \lim_{R\to\infty} \int_0^R e^{-x}\, \mathrm{d}x = \lim_{R\to\infty} \Big[-e^{-x} \Big]_0^R = 1$$

- Then compute

$$\begin{aligned}
\Gamma(2) &= \int_0^\infty xe^{-x}\,\mathrm{d}x \\
&= \lim_{R\to\infty}\int_0^R xe^{-x}\,\mathrm{d}x && \text{use integration by parts with} \\
& && u = x, \mathrm{d}v = e^{-x}\mathrm{d}x, \\
& && v = -e^{-x}, \mathrm{d}u = \mathrm{d}x \\
&= \lim_{R\to\infty}\left[-xe^{-x}\Big|_0^R + \int_0^R e^{-x}\,\mathrm{d}x\right] \\
&= \lim_{R\to\infty}\left[-xe^{-x} - e^{-x}\right]_0^R \\
&= 1
\end{aligned}$$

  For the last equality, we used that $\lim_{x\to\infty} xe^{-x} = 0$.

- 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,

$$\begin{aligned}
\Gamma(n+1) &= \int_0^\infty x^n e^{-x}\,\mathrm{d}x \\
&= \lim_{R\to\infty}\int_0^R x^n e^{-x}\,\mathrm{d}x && \text{again integrate by parts with} \\
& && u = x^n, \mathrm{d}v = e^{-x}\mathrm{d}x, \\
& && v = -e^{-x}, \mathrm{d}u = nx^{n-1}\mathrm{d}x \\
&= \lim_{R\to\infty}\left[-x^n e^{-x}\Big|_0^R + \int_0^R nx^{n-1}e^{-x}\,\mathrm{d}x\right] \\
&= \lim_{R\to\infty} n\int_0^R x^{n-1}e^{-x}\,\mathrm{d}x \\
&= n\Gamma(n)
\end{aligned}$$

  To get to the third row, we used that $\lim_{x\to\infty} x^n e^{-x} = 0$.

- 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[90] the Gamma function shifted by one.

Example 1.12.15

1.12.3 ▸▸ 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 1.11). You want to be sure that at least the integral converges before feeding it into a computer[91]. Fortunately it is usually possible to determine whether or not an improper integral converges even when you cannot evaluate it explicitly.

Remark 1.12.16. For pedagogical purposes, we are going to concentrate on the problem of determining whether or not an integral $\int_a^\infty f(x)\,\mathrm{d}x$ converges, when $f(x)$ has no singularities for $x \geqslant a$. 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. And, at the end of this section, we'll give the analogous tools that are appropriate for use with an integrand which is singular at one end of a finite domain of integration.

Now let's start. Imagine that we have an improper integral $\int_a^\infty f(x)\,\mathrm{d}x$, that $f(x)$ has no singularities for $x \geqslant a$ and that $f(x)$ is complicated enough that we cannot evaluate the integral explicitly[92]. The idea is find another improper integral $\int_a^\infty g(x)\,\mathrm{d}x$

- with $g(x)$ simple enough that we can evaluate the integral $\int_a^\infty g(x)\,\mathrm{d}x$ explicitly, or at least determine easily whether or not $\int_a^\infty g(x)\,\mathrm{d}x$ converges, and
- with $g(x)$ behaving enough like $f(x)$ for large x that the integral $\int_a^\infty f(x)\,\mathrm{d}x$ converges if and only if $\int_a^\infty g(x)\,\mathrm{d}x$ converges.

So far, this is a pretty vague strategy. Here is a theorem which starts to make it more precise.

90 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 $0, -1, -2, -3, \ldots$. For example, one can show that

$$\Gamma(1-z)\Gamma(z) = \frac{\pi}{\sin \pi z}.$$

91 Applying numerical integration methods to a divergent integral may result in perfectly reasonably looking but very wrong answers.

92 You could, for example, think of something like our running example $\int_a^\infty e^{-t^2}\mathrm{d}t$.

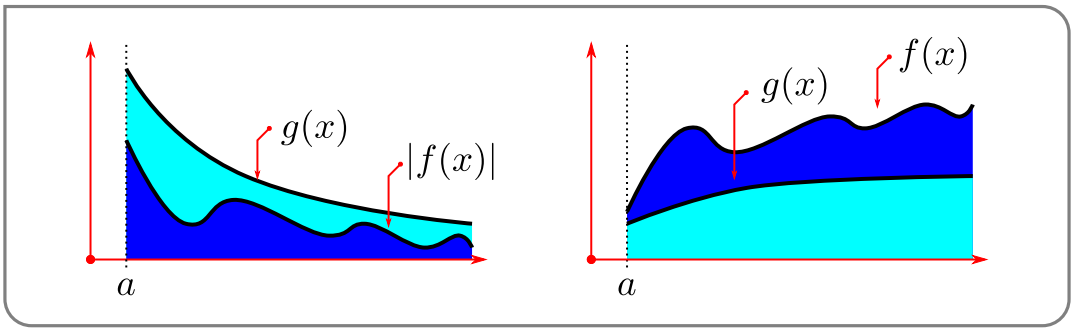
Theorem 1.12.17 (Comparison).

Let a be a real number. Let f and g be functions that are defined and continuous for all $x \geqslant a$ and assume that $g(x) \geqslant 0$ for all $x \geqslant a$.

(a) If $|f(x)| \leqslant g(x)$ for all $x \geqslant a$ and if $\int_a^\infty g(x)\,dx$ converges then $\int_a^\infty f(x)\,dx$ also converges.

(b) If $f(x) \geqslant g(x)$ for all $x \geqslant a$ and if $\int_a^\infty g(x)\,dx$ diverges then $\int_a^\infty f(x)\,dx$ also 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
$$\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant g(x)\,\} \text{ is finite.}$$
When $|f(x)| \leqslant g(x)$, the region
$\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant |f(x)|\,\}$ is contained inside $\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant g(x)\}$
and so must also have finite area. Consequently the areas of both the regions
$$\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant f(x)\,\} \text{ and } \{\,(x,y) \mid x \geqslant a,\ f(x) \leqslant y \leqslant 0\}$$
are finite too[93].

- If $\int_a^\infty g(x)\,dx$ diverges, then the area of
$$\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant g(x)\,\} \text{ is infinite.}$$
When $f(x) \geqslant g(x)$, the region
$\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant f(x)\,\}$ contains the region $\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant g(x)\,\}$
and so also has infinite area.

Example 1.12.18 $\left(\int_1^\infty e^{-x^2}\,dx\right)$

We cannot evaluate the integral $\int_1^\infty e^{-x^2}\,dx$ explicitly[94], however we would still like to un-

93 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 $\{\,(x,y) \mid x \geqslant a,\ 0 \leqslant y \leqslant f(x)\,\}$ and $\{\,(x,y) \mid x \geqslant a,\ f(x) \leqslant y \leqslant 0\,\}$.

94 It has been the subject of many remarks and footnotes.

derstand if it is finite or not — does it converge or diverge?

Solution. We will use Theorem 1.12.17 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 $x \geqslant 1$, we have $x^2 \geqslant x$ and hence $e^{-x^2} \leqslant e^{-x}$. Thus we can use Theorem 1.12.17 to compare

$$\int_1^\infty e^{-x^2}dx \text{ with } \int_1^\infty e^{-x}dx$$

- The integral

$$\begin{aligned}\int_1^\infty e^{-x}\,dx &= \lim_{R\to\infty}\int_1^R e^{-x}\,dx\\ &= \lim_{R\to\infty}\Big[-e^{-x}\Big]_1^R\\ &= \lim_{R\to\infty}\left[e^{-1}-e^{-R}\right] = e^{-1}\end{aligned}$$

  converges.
- So, by Theorem 1.12.17, 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 1.12.18

Example 1.12.19 $\left(\int_{1/2}^\infty e^{-x^2}\,dx\right)$

Solution.

- 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 1.12.18. But we cannot just repeat the argument of Example 1.12.18 because it is not true that $e^{-x^2} \leqslant e^{-x}$ 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 1.12.18 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 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 1.12.19

Theorem 1.12.20.

Let a and c be real numbers with $a < c$ and let the function $f(x)$ be continuous for all $x \geqslant a$. 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

$$\begin{aligned}\lim_{R\to\infty}\int_a^R f(x)\,dx &= \lim_{R\to\infty}\left[\int_a^c f(x)\,dx + \int_c^R f(x)\,dx\right]\\ &= \int_a^c f(x)\,dx + \lim_{R\to\infty}\int_c^R f(x)\,dx\end{aligned}$$

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 $\lim_{R\to\infty}\int_c^R f(x)\,dx$ exists and is finite, which in turn is the case if and only if the integral $\int_c^{\infty} f(x)\,dx$ converges. □

Example 1.12.21

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 $x \geqslant 1$ 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[95]. 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.

95 This takes practice, practice and more practice. At the risk of alliteration — please perform plenty of practice problems.

- 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 1.12.8, 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 1.12.17 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 $x \geqslant 1$ (i.e. on the domain of integration)

$$\frac{\sqrt{x}}{x^2+x} \leqslant \frac{A}{x^{3/2}}$$

  for some constant A. Let's try this.

- Since $x \geqslant 1$ 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 1.12.17(a) and Example 1.12.8, with $p = 3/2$ do indeed show that the integral $\int_1^\infty \frac{\sqrt{x}}{x^2+x}\,dx$ converges.

Example 1.12.21

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 1.12.17. 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 1.12.17 that is often easier to apply and that also fits well with the sort of intuition that we developed to solve Example 1.12.21.

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

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} \text{ exists and is a finite nonzero number.}$$

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

$$\frac{1}{2}L \leqslant \frac{f(x)}{g(x)} \leqslant 2L \qquad \text{for all } x > B$$

  Equivalently, $f(x)$ lies between $\frac{L}{2}g(x)$ and $2Lg(x)$, for all $x \geqslant B$.
- Consequently, the integral of $f(x)$ converges if and only if the integral of $g(x)$ converges, by Theorems 1.12.17 and 1.12.20.

These considerations lead to the following variant of Theorem 1.12.17.

Theorem 1.12.22 (Limiting comparison)**.**

Let $-\infty < a < \infty$. Let f and g be functions that are defined and continuous for all $x \geqslant a$ and assume that $g(x) \geqslant 0$ for all $x \geqslant a$.

(a) If $\int_a^\infty g(x)\,dx$ converges and the limit

$$\lim_{x\to\infty} \frac{f(x)}{g(x)}$$

exists, then $\int_a^\infty f(x)\,dx$ converges.

(b) If $\int_a^\infty g(x)\,dx$ diverges and the limit

$$\lim_{x\to\infty} \frac{f(x)}{g(x)}$$

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 1.12.22 is used.

Example 1.12.23 $\left(\int_1^\infty \frac{x+\sin x}{e^{-x}+x^2}\,dx\right)$

Does the integral $\displaystyle\int_1^\infty \frac{x+\sin x}{e^{-x}+x^2}\,dx$ 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
  - $|\sin x| \leqslant 1 \ll x$, 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

$$f(x) = \frac{x+\sin x}{e^{-x}+x^2} \qquad\qquad g(x) = \frac{1}{x}$$

  and compute

$$\begin{aligned}\lim_{x\to\infty} \frac{f(x)}{g(x)} &= \lim_{x\to\infty} \frac{x+\sin x}{e^{-x}+x^2} \div \frac{1}{x} \\ &= \lim_{x\to\infty} \frac{(1+\sin x/x)x}{(e^{-x}/x^2+1)x^2} \times x \\ &= \lim_{x\to\infty} \frac{1+\sin x/x}{e^{-x}/x^2+1} \\ &= 1\end{aligned}$$

- Since $\int_1^\infty g(x)\,dx = \int_1^\infty \frac{dx}{x}$ diverges, by Example 1.12.8 with $p = 1$, Theorem 1.12.22(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 1.12.23

We finish this section by giving variants of Theorems 1.12.17 and 1.12.22 that are useful in dealing with an integrand that is singular at one end of a (finite) domain of integration. For the rest of this section, let $a < b$ and let $[a,b] = \{x \,|\, a \leqslant x \leqslant b\}$ be the domain of integration. We allow the integrand to be singular at one end, that we'll call x_s, of the domain of integration. So

- if the integrand is singular at $x = a$, we set $x_s = a$ and write

$$I = (a, b] = \{x \,|\, a < x \leqslant b\}$$

and

- if the integrand is singular at $x = b$ we set $x_s = b$ and write

$$I = [a, b) = \{x \,|\, a \leqslant x < b\}$$

Recall that "$x \in I$" means "x is an element of the set I". Here is the analogue of Theorem 1.12.17

Theorem 1.12.24 (Comparison).

Let f and g be functions that are defined and continuous for all $x \in I$ and assume that $g(x) \geqslant 0$ for all $x \in I$.

(a) If $|f(x)| \leqslant g(x)$ for all $x \in I$ and if $\int_a^b g(x)\,dx$ converges then $\int_a^b f(x)\,dx$ also converges.

(b) If $f(x) \geqslant g(x)$ for all $x \in I$ and if $\int_a^b g(x)\,dx$ diverges then $\int_a^b f(x)\,dx$ also diverges.

and here is the analogue of Theorem 1.12.22

Theorem 1.12.25 (Limiting comparison).

Let f and g be functions that are defined and continuous for all $x \in I$ and assume that $g(x) \geqslant 0$ for all $x \in I$.

(a) If $\int_a^b g(x)\,dx$ converges and the limit

$$\lim_{\substack{x \to x_s \\ x \in I}} \frac{f(x)}{g(x)}$$

exists, then $\int_a^b f(x)\,dx$ converges.

(b) If $\int_a^b g(x)\,dx$ diverges and the limit

$$\lim_{\substack{x \to x_s \\ x \in I}} \frac{f(x)}{g(x)}$$

exists and is nonzero, then $\int_a^b f(x)\,dx$ 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.)

Chapter 2

APPLICATIONS OF INTEGRATION

In the previous chapter we defined the definite integral, based on its interpretation as the area of a region in the xy–plane. We also developed a bunch of theory to help us work with integrals. This abstract definition, and the associated theory, turns out to be extremely useful simply because "areas of regions in the xy–plane" appear in a huge number of different settings, many of which seem superficially not to involve "areas of regions in the xy–plane". Here are some examples.

- The work involved in moving a particle or in pumping a fluid out of a reservoir. See section 2.1.
- The average value of a function. See section 2.2.
- The center of mass of an object. See section 2.3.
- The time dependence of temperature. See section 2.4.
- Radiocarbon dating. See section 2.4.

Let us start with the first of these examples.

2.1▲ Work

While computing areas and volumes are nice mathematical applications of integration we can also use integration to compute quantities of importance in physics and statistics. One such quantity is work. Work is a way of quantifying the amount of energy that is required to act against a force[1]. In SI[2] metric units the force F has units newtons (which

1 For example — if your expensive closed-source textbook has fallen on the floor, work quantifies the amount of energy required to lift the object from the floor acting against the force of gravity.

2 SI is short for "le système international d'unités" which is French for "the international system of units". It is the most recent internationally sanctioned version of the metric system, published in 1960. It aims to establish sensible units of measurement (no cubic furlongs per hogshead-Fahrenheit). It defines seven base units — metre (length), kilogram (mass), second (time), kelvin (temperature), ampere (electric current), mole (quantity of substance) and candela (luminous intensity). From these one can then establish derived units — such as metres per second for velocity and speed.

are kilogram–metres per second squared), x has units metres and the work W has units joules (which are newton–metres or kilogram–metres squared per second squared).

Definition 2.1.1.

The work done by a force $F(x)$ in moving an object from $x = a$ to $x = b$ is

$$W = \int_a^b F(x)\,\mathrm{d}x$$

In particular, if the force is a constant, F, independent of x, the work is $F \cdot (b - a)$.

Here is some motivation for this definition. Consider a particle of mass m moving along the x–axis. Let the position of the particle at time t be $x(t)$. The particle starts at position a at time α, moves to the right, finishing at position $b > a$ at time β. While the particle moves, it is subject to a position-dependent force $F(x)$. Then Newton's law of motion[3] says[4] that force is mass times acceleration

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}(t) = F\big(x(t)\big)$$

Now consider our definition of work above. It tells us that the work done in moving the particle from $x = a$ to $x = b$ is

$$W = \int_a^b F(x)\mathrm{d}x$$

However, we know the position as a function of time, so we can substitute $x = x(t)$, $\mathrm{d}x = \frac{\mathrm{d}x}{\mathrm{d}t}\mathrm{d}t$ (using Theorem 1.4.6) and rewrite the above integral:

$$W = \int_a^b F(x)\mathrm{d}x = \int_{t=\alpha}^{t=\beta} F(x(t))\frac{\mathrm{d}x}{\mathrm{d}t}\mathrm{d}t$$

Using Newton's second law we can rewrite our integrand:

$$\begin{aligned}
&= m\int_\alpha^\beta \frac{\mathrm{d}^2x}{\mathrm{d}t^2}\frac{\mathrm{d}x}{\mathrm{d}t}\mathrm{d}t \\
&= m\int_\alpha^\beta \frac{\mathrm{d}v}{\mathrm{d}t}v(t)\mathrm{d}t && \text{since } v(t) = \frac{\mathrm{d}x}{\mathrm{d}t} \\
&= m\int_\alpha^\beta \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}v(t)^2\right)\mathrm{d}t
\end{aligned}$$

3 Specifically, the second of Newton's three law of motion. These were first published in 1687 in his "Philosophiæ Naturalis Principia Mathematica".

4 It actually says something more graceful in Latin - Mutationem motus proportionalem esse vi motrici impressae, et fieri secundum lineam rectam qua vis illa imprimitur. Or — The alteration of motion is ever proportional to the motive force impressed; and is made in the line in which that force is impressed. It is amazing what you can find on the internet.

What happened here? By the chain rule, for any function $f(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}f(t)^2\right) = f(t)f'(t).$$

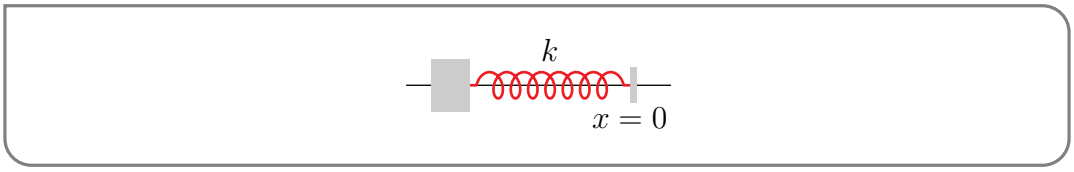
In the above computation we have used this fact with $f(t) = v(t)$. Now using the fundamental theorem of calculus (Theorem 1.3.1 part 2), we have

$$\begin{aligned} W &= m\int_\alpha^\beta \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}v(t)^2\right)\mathrm{d}t \\ &= \frac{1}{2}mv(\beta)^2 - \frac{1}{2}mv(\alpha)^2. \end{aligned}$$

By definition, the function $\frac{1}{2}mv(t)^2$ is the kinetic energy[5] of the particle at time t. So the work W of Definition 2.1.1 is the change in kinetic energy from the time the particle was at $x = a$ to the time it was at $x = b$.

Example 2.1.2 (Hooke's Law)

Imagine that a spring lies along the x-axis. The left hand end is fixed to a wall, but the right hand end lies freely at $x = 0$. So the spring is at its "natural length".



- Now suppose that we wish to stretch out the spring so that its right hand end is at $x = L$.
- Hooke's Law[6] says that when a (linear) spring is stretched (or compressed) by x units beyond its natural length, it exerts a force of magnitude kx, where the constant k is the spring constant of that spring.
- In our case, once we have stretched the spring by x units to the right, the spring will be trying to pull back the right hand end by applying a force of magnitude kx directed to the left.
- For us to continue stretching the spring we will have to apply a compensating force of magnitude kx directed to the right. That is, we have to apply the force $F(x) = +kx$.

5 This is not a physics text so we will not be too precise. Roughly speaking, kinetic energy is the energy an object possesses due to it being in motion, as opposed to potential energy, which is the energy of the object due to its position in a force field. Leibniz and Bernoulli determined that kinetic energy is proportional to the square of the velocity, while the modern term "kinetic energy" was first used by Lord Kelvin (back while he was still William Thompson).

6 Robert Hooke (1635–1703) was an English contemporary of Isaac Newton (1643–1727). It was in a 1676 letter to Hooke that Newton wrote "If I have seen further it is by standing on the shoulders of Giants." There is some thought that this was sarcasm and Newton was actually making fun of Hooke, who had a spinal deformity. However at that time Hooke and Newton were still friends. Several years later they did have a somewhat public falling-out over some of Newton's work on optics.

- So to stretch a spring by L units from its natural length we have to supply the work

$$W = \int_0^L kx\mathrm{d}x = \frac{1}{2}kL^2$$

Example 2.1.2

Example 2.1.3 (Spring)

A spring has a natural length of 0.1m. If a 12N force is needed to keep it stretched to a length of 0.12m, how much work is required to stretch it from 0.12m to 0.15m?

Solution. In order to answer this question we will need to determine the spring constant and then integrate the appropriate function.

- Our first task is to determine the spring constant k. We are told that when the spring is stretched to a length of 0.12m, i.e. to a length of $0.12 - 0.1 = 0.02$m beyond its natural length, then the spring generates a force of magnitude 12N.
- Hooke's law states that the force exerted by the spring, when it is stretched by x units, has magnitude kx, so

$$12 = k \cdot 0.02 = k \cdot \frac{2}{100} \qquad \text{thus}$$
$$k = 600.$$

- So to stretch the spring
  - from a length of 0.12m, i.e. a length of $x = 0.12 - 0.1 = 0.02$m beyond its natural length,
  - to a length of 0.15m, i.e. a length of $x = 0.15 - 0.1 = 0.05$m beyond its natural length,

  takes work

$$\begin{aligned} W &= \int_{0.02}^{0.05} kx\mathrm{d}x = \left[\frac{1}{2}kx^2\right]_{0.02}^{0.05} \\ &= 300\left(0.05^2 - 0.02^2\right) \\ &= 0.63\text{J} \end{aligned}$$

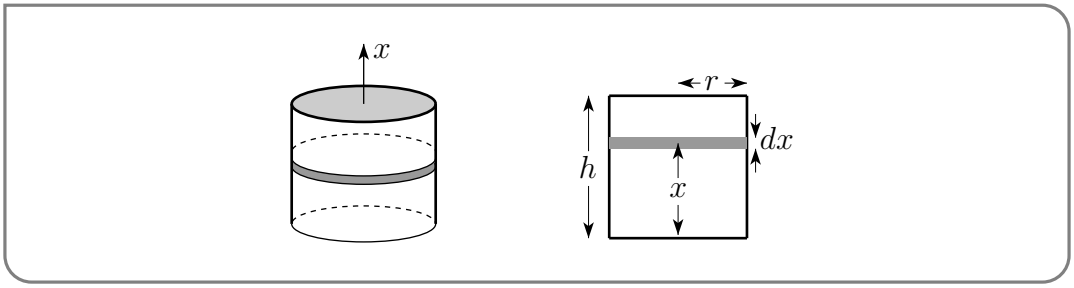
Example 2.1.3

Example 2.1.4 (Pumping Out a Reservoir)

A cylindrical reservoir[7] of height h and radius r is filled with a fluid of density ρ. We

7 We could assign units to these measurements — such as metres for the lengths h and r, and kilograms per cubic metre for the density ρ.

would like to know how much work is required to pump all of the fluid out the top of the reservoir.



Solution. We are going to tackle this problem by applying the standard integral calculus "slice into small pieces" strategy. This is how we computed areas and volumes — slice the problem into small pieces, work out how much each piece contributes, and then add up the contributions using an integral.

- Start by slicing the reservoir (or rather the fluid inside it) into thin, horizontal, cylindrical pancakes, as in the figure above. We proceed by determining how much work is required to pump out this pancake volume of fluid[8].
- Each pancake is a squat cylinder with thickness $\mathrm{d}x$ and circular cross section of radius r and area πr^2. Hence it has volume $\pi r^2 \mathrm{d}x$ and mass $\rho \times \pi r^2 \mathrm{d}x$.
- Near the surface of the Earth gravity exerts a downward force of mg on a body of mass m. The constant $g = 9.8\mathrm{m/sec}^2$ is called the *standard acceleration due to gravity*[9]. For us to raise the pancake we have to apply a compensating upward force of mg, which, for our pancake, is

$$F = g\rho \times \pi r^2 \mathrm{d}x$$

- To remove the pancake at height x from the reservoir we need to raise it to height h. So we have to lift it a distance $h - x$ using the force $F = \pi\rho g r^2 \mathrm{d}x$, which takes work $\pi\rho g r^2 \,(h - x)\, \mathrm{d}x$.
- The total work to empty the whole reservoir is

$$\begin{aligned} W &= \int_0^h \pi\, \rho g\, r^2 (h - x) \mathrm{d}x = \pi\, \rho g\, r^2 \int_0^h (h - x) \mathrm{d}x \\ &= \pi\, \rho g\, r^2 \Big[hx - \frac{x^2}{2} \Big]_0^h \\ &= \frac{\pi}{2}\, \rho g\, r^2 h^2 \end{aligned}$$

8 Potential for a bad "work out how much work out" pun here.

9 This quantity is not actually constant — it varies slightly across the surface of earth depending on local density, height above sea-level and centrifugal force from the earth's rotation. It is, for example, slightly higher in Oslo and slightly lower in Singapore. It is actually *defined* to be $9.80665\ \mathrm{m/sec}^2$ by the International Organisation for Standardization.

- If we measure lengths in metres and mass in kilograms, then this quantity has units of Joules. If we instead used feet and pounds[10] then this would have units of "foot–pounds". One foot-pound is equal to 1.355817… Joules.

Example 2.1.4

Example 2.1.5 (Escape Velocity)

Suppose that you shoot a probe straight up from the surface of the Earth — at what initial speed must the probe move in order to escape Earth's gravity?

Solution. We determine this by computing how much work must be done in order to escape Earth's gravity. If we assume that all of this work comes from the probe's initial kinetic energy, then we can establish the minimum initial velocity required.

- The work done by gravity when a mass moves from the surface of the Earth to a height h above the surface is

$$W = \int_0^h F(x)\mathrm{d}x$$

  where $F(x)$ is the gravitational force acting on the mass at height x above the Earth's surface.

- The gravitational force[11] of the Earth acting on a particle of mass m at a height x above the surface of the Earth is

$$F = -\frac{GMm}{(R+x)^2},$$

  where G is the gravitational constant, M is the mass of the Earth and R is the radius of the Earth. Note that $R + x$ is the distance from the object to the centre of the Earth. Additionally, note that this force is negative because gravity acts downward.

10 It is extremely mysterious to the authors why a person would do science or engineering in imperial units. One of the authors still has nightmares about having had to do so as a student.

11 Newton published his inverse square law of universal gravitation in his Principia in 1687. His law states that the gravitational force between two masses m_1 and m_2 is

$$F = -G\frac{m_1 m_2}{r^2}$$

where r is the distance separating the (centres of the) masses and $G = 6.674 \times 10^{-11}\mathrm{Nm}^2/\mathrm{kg}^2$ is the gravitational constant. Notice that r measures the separation between the centres of the masses not the distance between the surfaces of the objects.
Also, do not confuse G with g — standard acceleration due to gravity. The first measurement of G was performed by Henry Cavendish in 1798 — the interested reader should look up the "Cavendish experiment" for details of this very impressive work.

- So the work done by gravity on the probe, as it travels from the surface of the Earth to a height h, is

$$\begin{aligned} W &= -\int_0^h \frac{GMm}{(R+x)^2}\mathrm{d}x \\ &= -GMm\int_0^h \frac{1}{(R+x)^2}\mathrm{d}x \end{aligned}$$

  A quick application of the substitution rule with $u = R + x$ gives

$$\begin{aligned} &= -GMm\int_{u(0)}^{u(h)} \frac{1}{u^2}\mathrm{d}u \\ &= -GMm\left[-\frac{1}{u}\right]_{u=R}^{u=R+h} \\ &= \frac{GMm}{R+h} - \frac{GMm}{R} \end{aligned}$$

- So if the probe completely escapes the Earth and travels all the way to $h = \infty$, gravity does work

$$\lim_{h\to\infty}\left[\frac{GMm}{R+h} - \frac{GMm}{R}\right] = -\frac{GMm}{R}$$

  The minus sign means that gravity has removed energy $\frac{GMm}{R}$ from the probe.

- To finish the problem we need one more assumption. Let us assume that all of this energy comes from the probe's initial kinetic energy and that the probe is not fitted with any sort of rocket engine. Hence the initial kinetic energy $\frac{1}{2}mv^2$ (coming from an initial velocity v) must be at least as large as the work computed above. That is we need

$$\frac{1}{2}mv^2 \geqslant \frac{GMm}{R} \qquad \text{which rearranges to give}$$
$$v \geqslant \sqrt{\frac{2GM}{R}}$$

- The right hand side of this inequality, $\sqrt{\frac{2GM}{R}}$, is called the escape velocity.

Example 2.1.5

Example 2.1.6 (Lifting a Cable)

A 10–metre–long cable of mass 5kg is used to lift a bucket of water, with mass 8kg, out of a well. Find the work done.

Solution. Denote by y the height of the bucket above the top of the water in the well. So the bucket is raised from $y = 0$ to $y = 10$. The cable has mass density 0.5kg/m. So when the bucket is at height y,

- the cable that remains to be lifted has mass $0.5(10-y)$ kg and
- the remaining cable and water is subject to a downward gravitational force of magnitude $\big[0.5(10-y)+8\big]g = \big[13-\frac{y}{2}\big]g$, where $g = 9.8\ \mathrm{m/sec^2}$.

So to raise the bucket from height y to height $y+\mathrm{d}y$ we need to apply a compensating upward force of $\big[13-\frac{y}{2}\big]g$ through distance $\mathrm{d}y$. This takes work $\big[13-\frac{y}{2}\big]g\,\mathrm{d}y$. So the total work required is

$$\int_0^{10}\Big[13-\frac{y}{2}\Big]g\,\mathrm{d}y = g\left[13y-\frac{y^2}{4}\right]_0^{10} = \big[130-25\big]g = 105g = 1029\ \mathrm{J}$$

Example 2.1.6

2.2▲ Averages

Another frequent[12] application of integration is computing averages and other statistical quantities. We will not spend too much time on this topic — that is best left to a proper course in statistics — however, we will demonstrate the application of integration to the problem of computing averages.

Let us start with the definition[13] of the average of a finite set of numbers.

12 Awful pun. The two main approaches to statistics are frequentism and Bayesianism; the latter named after Bayes' Theorem which is, in turn, named for Reverend Thomas Bayes. While this (both the approaches to statistics and their history and naming) is a very interesting and quite philosophical topic, it is beyond the scope of this course. The interested reader has plenty of interesting reading here to interest them.

13 We are being a little loose here with the distinction between mean and average. To be much more pedantic — the average is the arithmetic mean. Other interesting "means" are the geometric and harmonic means:

$$\begin{aligned}\text{arithmetic mean} &= \frac{1}{n}\left(y_1+y_2+\cdots+y_n\right)\\ \text{geometric mean} &= \left(y_1\cdot y_2\cdots y_n\right)^{1/n}\\ \text{harmonic mean} &= \left[\frac{1}{n}\left(\frac{1}{y_1}+\frac{1}{y_2}+\cdots\frac{1}{y_n}\right)\right]^{-1}\end{aligned}$$

All of these quantities, along with the median and mode, are ways to measure the typical value of a set of numbers. They all have advantages and disadvantages — another interesting topic beyond the scope of this course, but plenty of fodder for the interested reader and their favourite search engine. But let us put pedantry (and beyond-the-scope-of-the-course-reading) aside and just use the terms average and mean interchangeably for our purposes here.

Definition 2.2.1.

The average (mean) of a set of n numbers $y_1, y_2, \cdots, y_n$ is

$$y_{\text{ave}} = \bar{y} = \langle y \rangle = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

The notations y_{ave}, $\bar{y}$ and $\langle y \rangle$ are all commonly used to represent the average.

Now suppose that we want to take the average of a function $f(x)$ with x running continuously from a to b. How do we even define what that means? A natural approach is to

- select, for each natural number n, a sample of n, more or less uniformly distributed, values of x between a and b,
- take the average of the values of f at the selected points,
- and then take the limit as n tends to infinity.

Unsurprisingly, this process looks very much like how we computed areas and volumes previously. So let's get to it.

- First fix any natural number n.
- Subdivide the interval $a \leqslant x \leqslant b$ into n equal subintervals, each of width $\Delta x = \frac{b-a}{n}$.
- The subinterval number i runs from x_{i-1} to x_i with $x_i = a + i\frac{b-a}{n}$.
- Select, for each $1 \leqslant i \leqslant n$, one value of x from subinterval number i and call it x_i^*. So $x_{i-1} \leqslant x_i^* \leqslant x_i$.
- The average value of f at the selected points is

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i^*) = \frac{1}{b-a}\sum_{i=1}^{n} f(x_i^*)\Delta x \qquad \text{since } \Delta x = \frac{b-a}{n}$$

  giving us a Riemann sum.

Now when we take the limit $n \to \infty$ we get exactly $\frac{1}{b-a}\int_a^b f(x)\mathrm{d}x$. That's why we define

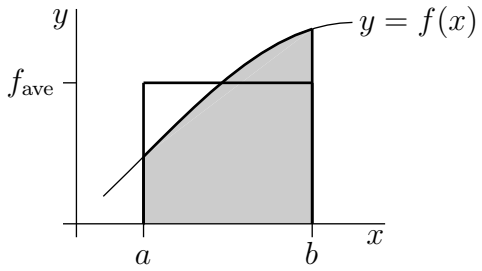
Definition 2.2.2.

Let $f(x)$ be an integrable function defined on the interval $a \leqslant x \leqslant b$. The average value of f on that interval is

$$f_{\text{ave}} = \bar{f} = \langle f \rangle = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x$$

Consider the case when $f(x)$ is positive. Then rewriting Definition 2.2.2 as

$$f_{\text{ave}}\,(b-a) = \int_a^b f(x)\mathrm{d}x$$



gives us a link between the average value and the area under the curve. The right-hand side is the area of the region

$$\left\{(x,y) \;\middle|\; a \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\ \right\}$$

while the left-hand side can be seen as the area of a rectangle of width $b-a$ and height f_{ave}. Since these areas must be the same, we interpret f_{ave} as the height of the rectangle which has the same width and the same area as $\left\{(x,y) \;\middle|\; a \leqslant x \leqslant b,\ 0 \leqslant y \leqslant f(x)\ \right\}$.

Let us start with a couple of simple examples and then work our way up to harder ones.

Example 2.2.3

Let $f(x) = x$ and $g(x) = x^2$ and compute their average values over $1 \leqslant x \leqslant 5$.

Solution. We can just plug things into the definition.

$$\begin{aligned} f_{\text{ave}} &= \frac{1}{5-1}\int_1^5 x\mathrm{d}x \\ &= \frac{1}{4}\left[\frac{x^2}{2}\right]_1^5 \\ &= \frac{1}{8}(25-1) = \frac{24}{8} \\ &= 3 \end{aligned}$$

as we might expect. And then

$$\begin{aligned} g_{\text{ave}} &= \frac{1}{5-1}\int_1^5 x^2\mathrm{d}x \\ &= \frac{1}{4}\left[\frac{x^3}{3}\right]_1^5 \\ &= \frac{1}{12}(125-1) = \frac{124}{12} \\ &= \frac{31}{3} \end{aligned}$$

Example 2.2.3

Something a little more trigonometric

Example 2.2.4

Find the average value of $\sin(x)$ over $0 \leqslant x \leqslant \pi/2$.

Solution. Again, we just need the definition.

$$\begin{aligned}\text{average} &= \frac{1}{\pi/2 - 0}\int_0^{\pi/2} \sin(x)\mathrm{d}x \\ &= \frac{2}{\pi}\cdot\Big[-\cos(x)\Big]_0^{\pi/2} \\ &= \frac{2}{\pi}(-\cos(\pi/2) + \cos(0)) \\ &= \frac{2}{\pi}.\end{aligned}$$

Example 2.2.4

We could keep going... But better to do some more substantial examples.

Example 2.2.5 (Average velocity)

Let $x(t)$ be the position at time t of a car moving along the x–axis. The velocity of the car at time t is the derivative $v(t) = x'(t)$. The average velocity of the car over the time interval $a \leqslant t \leqslant b$ is

$$\begin{aligned}v_{\text{ave}} &= \frac{1}{b-a}\int_a^b v(t)\mathrm{d}t \\ &= \frac{1}{b-a}\int_a^b x'(t)\mathrm{d}t \\ &= \frac{x(b)-x(a)}{b-a} \qquad \text{by the fundamental theorem of calculus.}\end{aligned}$$

The numerator in this formula is just the displacement (net distance travelled — if $x'(t) \geqslant 0$, it's the distance travelled) between time a and time b and the denominator is just the time it took.

Notice that this is exactly the formula we used way back at the start of your *differential* calculus class to help introduce the idea of the derivative. Of course this is a very circuitous way to get to this formula — but it is reassuring that we get the same answer.

Example 2.2.5

A very physics example.

Example 2.2.6 (Peak vs RMS voltage)

When you plug a light bulb into a socket[14] and turn it on, it is subjected to a voltage

$$V(t) = V_0 \sin(\omega t - \delta)$$

where

- $V_0 = 170$ volts,
- $\omega = 2\pi \times 60$ (which corresponds to 60 cycles per second[15]) and
- the constant δ is an (unimportant) phase. It just shifts the time at which the voltage is zero

The voltage V_0 is the "peak voltage" — the maximum value the voltage takes over time. More typically we quote the "root mean square" voltage[16] (or RMS-voltage). In this example we explain the difference, but to simplify the calculations, let us simplify the voltage function and just use

$$V(t) = V_0 \sin(t)$$

Since the voltage is a sine-function, it takes both positive and negative values. If we take its simple average over 1 period then we get

$$\begin{aligned} V_{\text{ave}} &= \frac{1}{2\pi - 0} \int_0^{2\pi} V_0 \sin(t) \mathrm{d}t \\ &= \frac{V_0}{2\pi} \Big[-\cos(t) \Big]_0^{2\pi} \\ &= \frac{V_0}{2\pi} \left(-\cos(2\pi) + \cos 0 \right) = \frac{V_0}{2\pi} (-1 + 1) \\ &= 0 \end{aligned}$$

This is clearly not a good indication of the typical voltage.

What we actually want here is a measure of how far the voltage is from zero. Now we could do this by taking the average of $|V(t)|$, but this is a little harder to work with. Instead we take the average of the square[17] of the voltage (so it is always positive) and

14 A normal household socket delivers alternating current, rather than the direct current USB supplies. At the risk of yet another "the interested reader" suggestion — the how and why household plugs supply AC current is another worthwhile and interesting digression from studying integration. The interested reader should look up the "War of Currents". The diligent and interested reader should bookmark this, finish the section and come back to it later.

15 Some countries supply power at 50 cycles per second. Japan actually supplies both — 50 cycles in the east of the country and 60 in the west.

16 This example was written in North America where the standard voltage supplied to homes is 120 volts. Most of the rest of the world supplies homes with 240 volts. The main reason for this difference is the development of the light bulb. The USA electrified earlier when the best voltage for bulb technology was 110 volts. As time went on, bulb technology improved and countries that electrified later took advantage of this (and the cheaper transmission costs that come with higher voltage) and standardised at 240 volts. So many digressions in this section!

17 For a finite set of numbers one can compute the "quadratic mean" which is another way to generalise the notion of the average:

$$\text{quadratic mean} = \sqrt{\frac{1}{n} \left(y_1^2 + y_2^2 + \cdots + y_n^2 \right)}$$

then take the square root at the end. That is

$$\begin{aligned} V_{\text{rms}} &= \sqrt{\frac{1}{2\pi - 0}\int_0^{2\pi} V(t)^2 \mathrm{d}t} \\ &= \sqrt{\frac{1}{2\pi}\int_0^{2\pi} V_0^2 \sin^2(t)\mathrm{d}t} \\ &= \sqrt{\frac{V_0^2}{2\pi}\int_0^{2\pi} \sin^2(t)\mathrm{d}t} \end{aligned}$$

This is called the "root mean square" voltage.

Though we do know how to integrate sine and cosine, we don't (yet) know how to integrate their squares. A quick look at double-angle formulas[18] gives us a way to eliminate the square:

$$\cos(2\theta) = 1 - 2\sin^2\theta \implies \sin^2\theta = \frac{1-\cos(2\theta)}{2}$$

Using this we manipulate our integrand a little more:

$$\begin{aligned} V_{\text{rms}} &= \sqrt{\frac{V_0^2}{2\pi}\int_0^{2\pi} \frac{1}{2}(1-\cos(2t))\mathrm{d}t} \\ &= \sqrt{\frac{V_0^2}{4\pi}\left[t - \frac{1}{2}\sin(2t)\right]_0^{2\pi}} \\ &= \sqrt{\frac{V_0^2}{4\pi}\left(2\pi - \frac{1}{2}\sin(4\pi) - 0 + \frac{1}{2}\sin(0)\right)} \\ &= \sqrt{\frac{V_0^2}{4\pi}\cdot 2\pi} \\ &= \frac{V_0}{\sqrt{2}} \end{aligned}$$

So if the peak voltage is 170 volts then the RMS voltage is $\frac{170}{\sqrt{2}} \approx 120.2$.

Example 2.2.6

Continuing this very physics example:

Example 2.2.7

Let us take our same light bulb with voltage (after it is plugged in) given by

$$V(t) = V_0 \sin(\omega t - \delta)$$

where

18 A quick glance at Appendix A.14 will refresh your memory.

- V_0 is the peak voltage,
- $\omega = 2\pi \times 60$, and
- the constant δ is an (unimportant) phase.

If the light bulb is "100 watts", then what is its resistance?

To answer this question we need the following facts from physics.

- If the light bulb has resistance R ohms, this causes, by Ohm's law, a current of

$$I(t) = \frac{1}{R}V(t)$$

  (amps) to flow through the light bulb.
- The current I is the number of units of charge moving through the bulb per unit time.
- The voltage is the energy required to move one unit of charge through the bulb.
- The power is the energy used by the bulb per unit time and is measured in watts.

So the power is the product of the current times the voltage and, so

$$P(t) = I(t)V(t) = \frac{V(t)^2}{R} = \frac{V_0^2}{R}\sin^2(\omega t - \delta)$$

The average power used over the time interval $a \leqslant t \leqslant b$ is

$$P_{\text{ave}} = \frac{1}{b-a}\int_a^b P(t)\,\mathrm{d}t = \frac{V_0^2}{R(b-a)}\int_a^b \sin^2(\omega t - \delta)\,\mathrm{d}t$$

Notice that this is almost exactly the form we had in the previous example when computing the root mean square voltage.

Again we simplify the integrand using the identity

$$\cos(2\theta) = 1 - 2\sin^2\theta \implies \sin^2\theta = \frac{1-\cos(2\theta)}{2}$$

So

$$\begin{aligned}
P_{\text{ave}} &= \frac{1}{b-a}\int_a^b P(t)\,\mathrm{d}t = \frac{V_0^2}{2R(b-a)}\int_a^b \big[1-\cos(2\omega t - 2\delta)\big]\,\mathrm{d}t \\
&= \frac{V_0^2}{2R(b-a)}\left[t - \frac{\sin(2\omega t - 2\delta)}{2\omega}\right]_a^b \\
&= \frac{V_0^2}{2R(b-a)}\left[b - a - \frac{\sin(2\omega b - 2\delta)}{2\omega} + \frac{\sin(2\omega a - 2\delta)}{2\omega}\right] \\
&= \frac{V_0^2}{2R} - \frac{V_0^2}{4\omega R(b-a)}\big[\sin(2\omega b - 2\delta) - \sin(2\omega a - 2\delta)\big]
\end{aligned}$$

In the limit as the length of the time interval $b-a$ tends to infinity, this converges to $\frac{V_0^2}{2R}$. The resistance R of a "100 watt bulb" obeys

$$\frac{V_0^2}{2R} = 100 \qquad \text{so that} \qquad R = \frac{V_0^2}{200}.$$

We finish this example off with two side remarks.

- If we translate the peak voltage to the root mean square voltage using
$$V_0 = V_{\text{rms}} \cdot \sqrt{2}$$
then we have
$$P = \frac{V_{\text{rms}}^2}{R}$$
- If we were using direct voltage rather than alternating current then the computation is much simpler. The voltage and current are constants, so
$$\begin{aligned} P &= V \cdot I && \text{but } I = V/R \text{ by Ohm's law} \\ &= \frac{V^2}{R} \end{aligned}$$
So if we have a direct current giving voltage equal to the root mean square voltage, then we would expend the same power.

Example 2.2.7

▸▸ Optional — Return to the Mean Value Theorem

Here is another application of the Definition 2.2.2 of the average value of a function on an interval. The following theorem can be thought of as an analogue of the mean value theorem (which was covered in your differential calculus class) but for integrals. The theorem says that a continuous function $f(x)$ must be exactly equal to its average value for some x. For example, if you went for a drive along the x–axis and you were at $x(a)$ at time a and at $x(b)$ at time b, then your velocity $x'(t)$ had to be exactly your average velocity $\frac{x(b)-x(a)}{b-a}$ at some time t between a and b. In particular, if your average velocity was greater than the speed limit, you were definitely speeding at some point during the trip. This is, of course, no great surprise[19].

Theorem 2.2.8 (Mean Value Theorem for Integrals)**.**

Let $f(x)$ be a continuous function on the interval $a \leqslant x \leqslant b$. Then there is some c obeying $a < c < b$ such that

$$\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x = f(c) \qquad \text{or} \qquad \int_a^b f(x)\mathrm{d}x = f(c)\,(b-a)$$

19 There are many unsurprising things that are true, but there are also many unsurprising things that surprisingly turn out to be false. Mathematicians like to prove things - surprising or not.

Proof. We will apply the mean value theorem (Theorem 2.13.4 in the CLP-1 text) to the function

$$F(x) = \int_a^x f(t)\,\mathrm{d}t$$

By the part 1 of the fundamental theorem of calculus (Theorem 1.3.1), $F'(x) = f(x)$, so the mean value theorem says that there is a $a < c < b$ with

$$\begin{aligned} f(c) = F'(c) = \frac{F(b) - F(a)}{b-a} &= \frac{1}{b-a}\left\{\int_a^b f(t)\,\mathrm{d}t - \int_a^a f(t)\,\mathrm{d}t\right\} \\ &= \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x \end{aligned}$$

□

In the next section, we will encounter an application in which we want to take the average value of a function $f(x)$, but in doing so we want some values of x to count more than other values of x. That is, we want to weight some x's more than other x's. To do so, we choose a "weight function" $w(x) \geqslant 0$ with $w(x)$ larger for more important x's. Then we define the weighted average of f as follows.

Definition 2.2.9.

Let $f(x)$ and $w(x)$ be integrable functions defined on the interval $a \leqslant x \leqslant b$ with $w(x) \geqslant 0$ for all $a \leqslant x \leqslant b$ and with $\int_a^b w(x)\,\mathrm{d}x > 0$. The average value of f on that interval, weighted by w, is

$$\frac{\int_a^b f(x)\,w(x)\,\mathrm{d}x}{\int_a^b w(x)\,\mathrm{d}x}$$

We typically refer to this simply as the weighted average of f.

Here are a few remarks concerning this definition.

- The definition has been rigged so that, if $f(x) = 1$ for all x, then the weighted average of f is 1, no matter what weight function $w(x)$ is used.
- If the weight function $w(x) = C$ for some constant $C > 0$ then the weighted average

  $$\frac{\int_a^b f(x)\,w(x)\,\mathrm{d}x}{\int_a^b w(x)\,\mathrm{d}x} = \frac{\int_a^b f(x)\,C\,\mathrm{d}x}{\int_a^b C\,\mathrm{d}x} = \frac{\int_a^b f(x)\,\mathrm{d}x}{b-a}$$

  is just the usual average.
- For any function $w(x) \geqslant 0$ and any $a < b$, we have $\int_a^b w(x)\,\mathrm{d}x \geqslant 0$. But for the definition of weighted average to make sense, we need to be able to divide by $\int_a^b w(x)\,\mathrm{d}x$. So we need $\int_a^b w(x)\,\mathrm{d}x \neq 0$.

The next theorem says that a continuous function $f(x)$ must be equal to its weighted average at some point x.

Theorem 2.2.10 (Mean Value Theorem for Weighted Integrals)**.**

Let $f(x)$ and $w(x)$ be continuous functions on the interval $a \leqslant x \leqslant b$. Assume that $w(x) > 0$ for all $a < x < b$. Then there is some c obeying $a < c < b$ such that

$$\frac{\int_a^b f(x)\,w(x)\,\mathrm{d}x}{\int_a^b w(x)\,\mathrm{d}x} = f(c) \qquad \text{or} \qquad \int_a^b f(x)\,w(x)\,\mathrm{d}x = f(c)\int_a^b w(x)\,\mathrm{d}x$$

Proof. We will apply the generalised mean value theorem (Theorem 3.4.38 in the CLP-1 text) to

$$F(x) = \int_a^x f(t)\,w(t)\mathrm{d}t \qquad G(x) = \int_a^x w(t)\mathrm{d}t$$

By the part 1 of the fundamental theorem of calculus (Theorem 1.3.1), $F'(x) = f(x)w(x)$ and $G'(x) = w(x)$, so the generalised mean value theorem says that there is a $a < c < b$ with

$$\begin{aligned} f(c) &= \frac{F'(c)}{G'(c)} = \frac{F(b)-F(a)}{G(b)-G(a)} = \frac{\int_a^b f(t)\,w(t)\mathrm{d}t - \int_a^a f(t)w(t)\,\mathrm{d}t}{\int_a^b w(t)\mathrm{d}t - \int_a^a w(t)\mathrm{d}t} \\ &= \frac{\int_a^b f(t)\,w(t)\,\mathrm{d}t}{\int_a^b w(t)\,\mathrm{d}t} \end{aligned}$$

□

Example 2.2.11

In this example, we will take a number of weighted averages of the simple function $f(x) = x$ over the simple interval $a = 1 \leqslant x \leqslant 2 = b$. As x increases from 1 to 2, the function $f(x)$ increases linearly from 1 to 2. So it is no shock that the ordinary average of f is exactly its middle value:

$$\frac{1}{b-a}\int_a^b f(t)\,\mathrm{d}t = \frac{1}{2-1}\int_1^2 t\,\mathrm{d}t = \frac{3}{2}$$

Pick any natural number $N \geqslant 1$ and consider the weight function $w_N(x) = x^N$. Note that $w_N(x)$ increases as x increases. So $w_N(x)$ weights bigger x's more than it weights smaller x's. In particular w_N weights the point $x = 2$ by a factor of 2^N (which is greater than 1 and grows to infinity as N grows to infinity) more than it weights the point $x = 1$. The

weighted average of f is

$$\frac{\int_a^b f(t)\,w_N(t)\,\mathrm{d}t}{\int_a^b w_N(t)\,\mathrm{d}t} = \frac{\int_1^2 t^{N+1}\,\mathrm{d}t}{\int_1^2 t^N\,\mathrm{d}t} = \frac{\frac{2^{N+2}-1}{N+2}}{\frac{2^{N+1}-1}{N+1}} = \frac{N+1}{N+2}\,\frac{2^{N+2}-1}{2^{N+1}-1}$$
$$= \begin{cases} \frac{2\times 7}{3\times 3} = 1.555 & \text{if } N=1 \\ \frac{3\times 15}{4\times 7} = 1.607 & \text{if } N=2 \\ \frac{4\times 31}{5\times 15} = 1.653 & \text{if } N=3 \\ \frac{5\times 63}{6\times 31} = 1.694 & \text{if } N=4 \\ 1.889 & \text{if } N=16 \\ 1.992 & \text{if } N=256 \end{cases}$$

As we would expect, the w_N-weighted average is between 1.5 (which is the ordinary, unweighted, average) and 2 (which is the biggest value of f in the interval) and grows as N grows. The limit as $N\to\infty$ of the w_N-weighted average is

$$\lim_{N\to\infty} \frac{N+1}{N+2}\,\frac{2^{N+2}-1}{2^{N+1}-1} = \lim_{N\to\infty} \frac{N+2-1}{N+2}\,\frac{2^{N+2}-2+1}{2^{N+1}-1}$$
$$= \lim_{N\to\infty} \left[1-\frac{1}{N+2}\right]\left[2+\frac{1}{2^{N+1}-1}\right]$$
$$= 2$$

Example 2.2.11

Example 2.2.12

Here is an example which shows what can go wrong with Theorem 2.2.10 if we allow the weight function $w(x)$ to change sign. Let $a=-0.99$ and $b=1$. Let

$$w(x) = \begin{cases} 1 & \text{if } x \geqslant 0 \\ -1 & \text{if } x < 0 \end{cases}$$
$$f(x) = \begin{cases} x & \text{if } x \geqslant 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Then

$$\int_a^b f(x)\,w(x)\,\mathrm{d}x = \int_0^1 x\,\mathrm{d}x = \frac{1}{2}$$
$$\int_a^b w(x)\,\mathrm{d}x = \int_0^1 \mathrm{d}x - \int_{-0.99}^0 \mathrm{d}x = 1-0.99 = 0.01$$

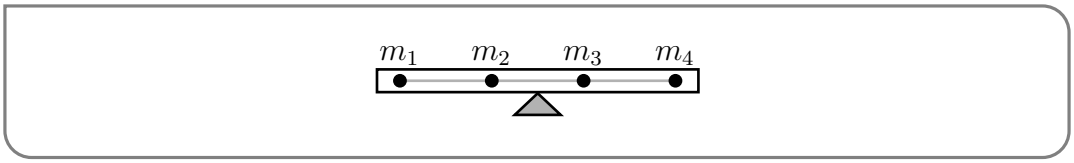
As c runs from a to b, $f(c)\int_a^b w(x)\,\mathrm{d}x = 0.01 f(c)$ runs from 0 to 0.01 and, in particular, never takes a value anywhere near $\int_a^b f(x)\,w(x)\,\mathrm{d}x = \frac{1}{2}$. There is no c value which works.

Example 2.2.12

2.3▲ Centre of Mass and Torque

2.3.1 ▸▸ Centre of Mass

If you support a body at its center of mass (in a uniform gravitational field) it balances perfectly. That's the definition of the center of mass of the body. If the body consists of a



finite number of masses $m_1, \cdots, m_n$ attached to an infinitely strong, weightless (idealized) rod with mass number i attached at position x_i, then the center of mass is at the (weighted) average value of x:

$$\bar{x} = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i} \tag{2.3.1}$$

The denominator $m = \sum_{i=1}^{n} m_i$ is the total mass of the body. This formula for the center of mass is derived in the following (optional) section. See (2.3.8).

For many (but certainly not all) purposes an (extended rigid) body acts like a point particle located at its center of mass. For example it is very common to treat the Earth as a point particle. Here is a more detailed example in which we think of a body as being made up of a number of component parts and compute the center of mass of the body as a whole by using the center of masses of the component parts. Suppose that we have a dumbbell which consists of

- a left end made up of particles of masses $m_{l,1}, \cdots, m_{l,3}$ located at $x_{l,1}, \cdots, x_{l,3}$ and
- a right end made up of particles of masses $m_{r,1}, \cdots, m_{r,4}$ located at $x_{r,1}, \cdots, x_{r,4}$ and
- an infinitely strong, weightless (idealized) rod joining all of the particles.

Then the mass and center of mass of the left end are

$$M_l = m_{l,1} + \cdots + m_{l,3} \qquad \bar{X}_l = \frac{m_{l,1}x_{l,1} + \cdots + m_{l,3}x_{l,3}}{M_l}$$

and the mass and center of mass of the right end are

$$M_r = m_{r,1} + \cdots + m_{r,4} \qquad \bar{X}_r = \frac{m_{r,1}x_{r,1} + \cdots + m_{r,4}x_{r,4}}{M_r}$$

The mass and center of mass of the entire dumbbell are

$$\begin{aligned} M &= m_{l,1} + \cdots + m_{l,3} \ + \ m_{r,1} + \cdots + m_{r,4} \\ &= M_l + M_r \\ \bar{x} &= \frac{m_{l,1}x_{l,1} + \cdots + m_{l,3}x_{l,3} \ + \ m_{r,1}x_{r,1} + \cdots + m_{r,4}x_{r,4}}{M} \\ &= \frac{M_l\bar{X}_l + M_r\bar{X}_r}{M_r + M_l} \end{aligned}$$

So we can compute the center of mass of the entire dumbbell by treating it as being made up of two point particles, one of mass M_l located at the centre of mass of the left end, and one of mass M_r located at the center of mass of the right end.

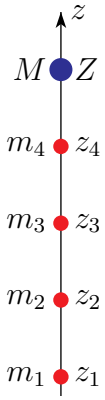
Example 2.3.1 (Work and Centre of Mass)

Here is another example in which an extended body acts like a point particle located at its centre of mass. Imagine that there are a finite number of masses $m_1, \cdots, m_n$ arrayed along a (vertical) z–axis with mass number i attached at height z_i. Note that the total mass of the array is $M = \sum_{i=1}^n m_i$ and that the centre of mass of the array is at height

$$\bar{z} = \frac{\sum_{i=1}^n m_i z_i}{\sum_{i=1}^n m_i} = \frac{1}{M}\sum_{i=1}^n m_i z_i$$

Now suppose that we lift all of the masses, against gravity, to height Z. So after the lift there is a total mass M located at height Z. The i^{th} mass is subject to a downward gravitational force of $m_i g$. So to lift the i^{th} mass we need to apply a compensating upward force of $m_i g$ through a distance of $Z - z_i$. This takes work $m_i g(Z - z_i)$. So the total work required to lift all n masses is

$$\begin{aligned}\text{Work} &= \sum_{i=1}^n m_i g(Z - z_i)\\ &= gZ\sum_{i=1}^n m_i - g\sum_{i=1}^n m_i z_i\\ &= gZM - gM\bar{z}\\ &= Mg(Z - \bar{z})\end{aligned}$$



So the work required to lift the array of n particles is identical to the work required to lift a single particle, whose mass, M, is the total mass of the array, from height $\bar{z}$, the centre of mass of the array, to height Z.

Example 2.3.1

Example 2.3.2 (Example 2.3.1, continued)

Imagine, as in Example 2.3.1, that there are a finite number of masses $m_1, \cdots, m_n$ arrayed along a (vertical) z–axis with mass number i attached at height z_i. Again, the total mass and centre of mass of the array are

$$M = \sum_{i=1}^n m_i \qquad \bar{z} = \frac{\sum_{i=1}^n m_i z_i}{\sum_{i=1}^n m_i} = \frac{1}{M}\sum_{i=1}^n m_i z_i$$

Now suppose that we lift, for each $1 \leqslant i \leqslant n$, mass number i, against gravity, from its initial height z_i to a final height Z_i. So after the lift we have a new array of masses with total mass and centre of mass

$$M = \sum_{i=1}^n m_i \qquad \bar{Z} = \frac{\sum_{i=1}^n m_i Z_i}{\sum_{i=1}^n m_i} = \frac{1}{M}\sum_{i=1}^n m_i Z_i$$

To lift the i^{th} mass took work $m_i g(Z_i - z_i)$. So the total work required to lift all n masses was

$$\begin{aligned}\text{Work} &= \sum_{i=1}^{n} m_i g(Z_i - z_i) \\ &= g\sum_{i=1}^{n} m_i Z_i - g\sum_{i=1}^{n} m_i z_i \\ &= gM\bar{Z} - gM\bar{z} = Mg(\bar{Z} - \bar{z})\end{aligned}$$

So the work required to lift the array of n particles is identical to the work required to lift a single particle, whose mass, M, is the total mass of the array, from height $\bar{z}$, the initial centre of mass of the array, to height $\bar{Z}$, the final centre of mass of the array.

Example 2.3.2

Now we'll extend the above ideas to cover more general classes of bodies. If the body consists of mass distributed continuously along a straight line, say with mass density $\rho(x)$kg/m and with x running from a to b, rather than consisting of a finite number of point masses, the formula for the center of mass becomes

$$\bar{x} = \frac{\int_a^b x\,\rho(x)\,\mathrm{d}x}{\int_a^b \rho(x)\,\mathrm{d}x} \tag{2.3.2}$$

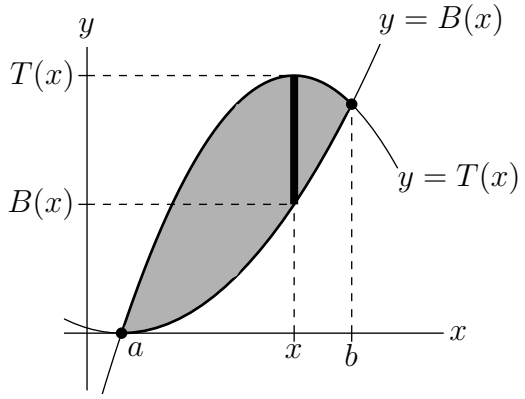
Think of $\rho(x)\,\mathrm{d}x$ as the mass of the "almost point particle" between x and $x + \mathrm{d}x$.

If the body is a two dimensional object, like a metal plate, lying in the xy–plane, its center of mass is a point $(\bar{x}, \bar{y})$ with $\bar{x}$ being the (weighted) average value of the x–coordinate over the body and $\bar{y}$ being the (weighted) average value of the y–coordinate over the body. To be concrete, suppose the body fills the region

$$\left\{\ (x,y)\ \middle|\ a \leqslant x \leqslant b,\ B(x) \leqslant y \leqslant T(x)\ \right\}$$

in the xy–plane. For simplicity, we will assume that the density of the body is a constant, say ρ. When the density is constant, the center of mass is also called the *centroid* and is thought of as the geometric center of the body.

To find the centroid of the body, we use our standard "slicing" strategy. We slice the body into thin vertical strips, as illustrated in the figure below. Here is a detailed descrip-



tion of a generic strip.

- The strip has width $\mathrm{d}x$.
- Each point of the strip has essentially the same x–coordinate. Call it x.
- The top of the strip is at $y = T(x)$ and the bottom of the strip is at $y = B(x)$.
- So the strip has
  - height $T(x) - B(x)$
  - area $[T(x) - B(x)]\,\mathrm{d}x$
  - mass $\rho[T(x) - B(x)]\,\mathrm{d}x$
  - centroid, i.e. middle point, $\big(x\,,\ \frac{B(x)+T(x)}{2}\big)$.

In computing the centroid of the entire body, we may treat each strip as a single particle of mass $\rho[T(x) - B(x)]\,\mathrm{d}x$ located at $\big(x\,,\ \frac{B(x)+T(x)}{2}\big)$. So the mass of the entire body is

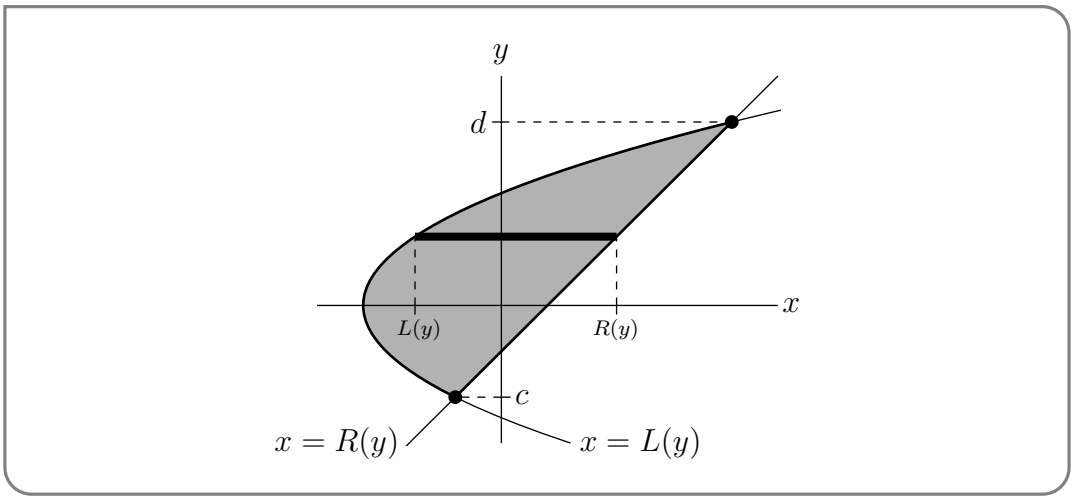
$$M = \rho \int_a^b [T(x) - B(x)]\,\mathrm{d}x = \rho A \tag{2.3.3a}$$

where $A = \int_a^b [T(x) - B(x)]\,\mathrm{d}x$ is the area of the region. The coordinates of the centroid are

$$\bar{x} = \frac{\int_a^b x\ \overbrace{\rho[T(x) - B(x)]\,\mathrm{d}x}^{\text{mass of slice}}}{M} = \frac{\int_a^b x[T(x) - B(x)]\,\mathrm{d}x}{A} \tag{2.3.3b}$$

$$\bar{y} = \frac{\int_a^b \overbrace{\frac{B(x)+T(x)}{2}}^{\text{average } y \text{ on slice}}\ \overbrace{\rho[T(x) - B(x)]\,\mathrm{d}x}^{\text{mass of slice}}}{M} = \frac{\int_a^b [T(x)^2 - B(x)^2]\,\mathrm{d}x}{2A} \tag{2.3.3c}$$

We can of course also slice up the body using horizontal slices. If the body has constant



density ρ and fills the region

$$\big\{\ (x,y)\ \big|\ L(y) \leqslant x \leqslant R(y),\ c \leqslant y \leqslant d\ \big\}$$

then the same computation as above gives the mass of the body to be

$$M = \rho \int_c^d [R(y) - L(y)]\,\mathrm{d}y = \rho A \tag{2.3.4a}$$

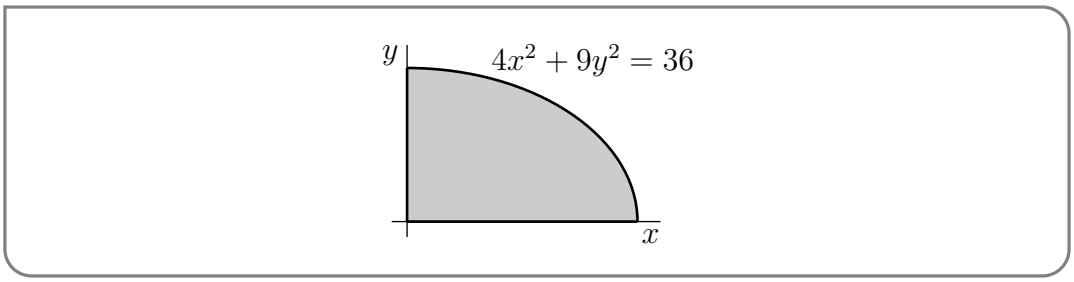
where $A = \int_c^d [R(y) - L(y)]\,\mathrm{d}y$ is the area of the region, and gives the coordinates of the centroid to be

$$\bar{x} = \frac{\int_c^d \overbrace{\frac{R(y)+L(y)}{2}}^{\text{average } x \text{ on slice}} \overbrace{\rho[R(y) - L(y)]\,\mathrm{d}y}^{\text{mass of slice}}}{M} = \frac{\int_c^d [R(y)^2 - L(y)^2]\,\mathrm{d}y}{2A} \tag{2.3.4b}$$

$$\bar{y} = \frac{\int_c^d y\ \overbrace{\rho[R(y) - L(y)]\,\mathrm{d}y}^{\text{mass of slice}}}{M} = \frac{\int_c^d y[R(y) - L(y)]\,\mathrm{d}y}{A} \tag{2.3.4c}$$

Example 2.3.3

Find the x–coordinate of the centroid (centre of gravity) of the plane region R that lies in the first quadrant $x \geqslant 0,\ y \geqslant 0$ and inside the ellipse $4x^2 + 9y^2 = 36$. (The area bounded by the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ is πab square units.)



Solution. In standard form $4x^2 + 9y^2 = 36$ is $\frac{x^2}{9} + \frac{y^2}{4} = 1$. So, on R, x runs from 0 to 3 and R has area $A = \frac{1}{4}\pi \times 3 \times 2 = \frac{3}{2}\pi$. For each fixed x, between 0 and 3, y runs from 0 to $2\sqrt{1 - \frac{x^2}{9}}$. So, applying (2.3.3.b) with $a = 0$, $b = 3$, $T(x) = 2\sqrt{1 - \frac{x^2}{9}}$ and $B(x) = 0$,

$$\bar{x} = \frac{1}{A}\int_0^3 x\,T(x)\,\mathrm{d}x = \frac{1}{A}\int_0^3 x\,2\sqrt{1 - \frac{x^2}{9}}\,\mathrm{d}x = \frac{4}{3\pi}\int_0^3 x\sqrt{1 - \frac{x^2}{9}}\,\mathrm{d}x$$

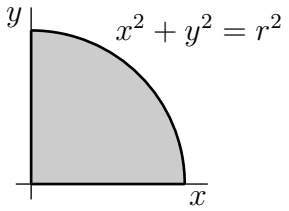
Sub in $u = 1 - \frac{x^2}{9}$, $\mathrm{d}u = -\frac{2}{9}x\,\mathrm{d}x$.

$$\bar{x} = -\frac{9}{2}\frac{4}{3\pi}\int_1^0 \sqrt{u}\,\mathrm{d}u = -\frac{9}{2}\frac{4}{3\pi}\Big[\frac{u^{3/2}}{3/2}\Big]_1^0 = -\frac{9}{2}\frac{4}{3\pi}\Big[-\frac{2}{3}\Big] = \frac{4}{\pi}$$

Example 2.3.3

Example 2.3.4

Find the centroid of the quarter circular disk $x \geqslant 0$, $y \geqslant 0$, $x^2 + y^2 \leqslant r^2$.



Solution. By symmetry, $\bar{x} = \bar{y}$. The area of the quarter disk is $A = \frac{1}{4}\pi r^2$. By (2.3.3.b) with $a = 0$, $b = r$, $T(x) = \sqrt{r^2 - x^2}$ and $B(x) = 0$,

$$\bar{x} = \frac{1}{A}\int_0^r x\sqrt{r^2 - x^2}\,\mathrm{d}x$$

To evaluate the integral, sub in $u = r^2 - x^2$, $\mathrm{d}u = -2x\,\mathrm{d}x$.

$$\int_0^r x\sqrt{r^2 - x^2}\,\mathrm{d}x = \int_{r^2}^0 \sqrt{u}\,\frac{\mathrm{d}u}{-2} = -\frac{1}{2}\left[\frac{u^{3/2}}{3/2}\right]_{r^2}^0 = \frac{r^3}{3} \tag{2.3.5}$$

So

$$\bar{x} = \frac{4}{\pi r^2}\left[\frac{r^3}{3}\right] = \frac{4r}{3\pi}$$

As we observed above, we should have $\bar{x} = \bar{y}$. But, just for practice, let's compute $\bar{y}$ by the integral formula (2.3.3.c), again with $a = 0$, $b = r$, $T(x) = \sqrt{r^2 - x^2}$ and $B(x) = 0$,

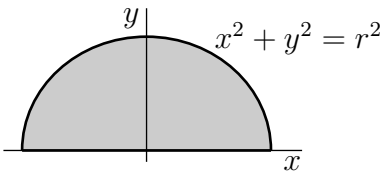
$$\begin{aligned}
\bar{y} = \frac{1}{2A}\int_0^r \left(\sqrt{r^2 - x^2}\right)^2 \mathrm{d}x &= \frac{2}{\pi r^2}\int_0^r \left(r^2 - x^2\right)\mathrm{d}x \\
&= \frac{2}{\pi r^2}\left[r^2 x - \frac{x^3}{3}\right]_0^r \qquad = \frac{2}{\pi r^2}\frac{2r^3}{3} \\
&= \frac{4r}{3\pi}
\end{aligned}$$

as expected.

Example 2.3.4

Example 2.3.5

Find the centroid of the half circular disk $y \geqslant 0$, $x^2 + y^2 \leqslant r^2$.



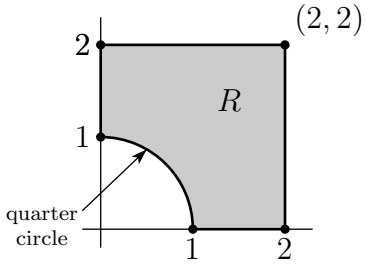
Solution. Once again, we have a symmetry —- namely the half disk is symmetric about the y–axis. So the centroid lies on the y–axis and $\bar{x} = 0$. The area of the half disk is $A = \frac{1}{2}\pi r^2$. By (2.3.3.c), with $a = -r$, $b = r$, $T(x) = \sqrt{r^2 - x^2}$ and $B(x) = 0$,

$$\begin{aligned}
\bar{y} &= \frac{1}{2A}\int_{-r}^{r} \left(\sqrt{r^2 - x^2}\right)^2 \mathrm{d}x \;= \frac{1}{\pi r^2}\int_{-r}^{r} \left(r^2 - x^2\right) \mathrm{d}x \\
&= \frac{2}{\pi r^2}\int_{0}^{r} \left(r^2 - x^2\right) \mathrm{d}x \qquad \text{since the integrand is even} \\
&= \frac{2}{\pi r^2}\left[r^2 x - \frac{x^3}{3}\right]_0^r \\
&= \frac{4r}{3\pi}
\end{aligned}$$

Example 2.3.5

Example 2.3.6

Find the centroid of the region R in the diagram.



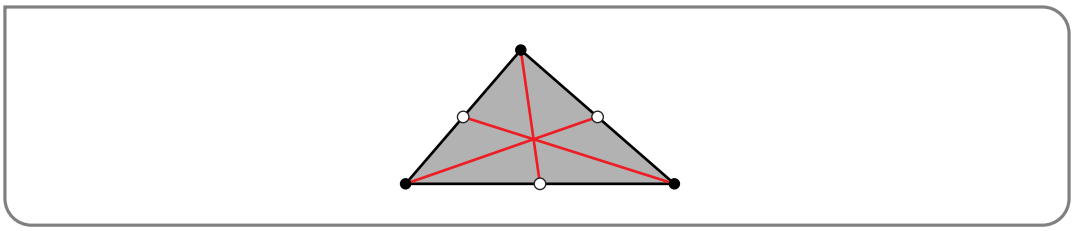
Solution. By symmetry, $\bar{x} = \bar{y}$. The region R is a 2×2 square with one quarter of a circle of radius 1 removed and so has area $2 \times 2 - \frac{1}{4}\pi = \frac{16-\pi}{4}$. The top of R is $y = T(x) = 2$. The bottom is $y = B(x)$ with $B(x) = \sqrt{1 - x^2}$ when $0 \leqslant x \leqslant 1$ and $B(x) = 0$ when $1 \leqslant x \leqslant 2$. So

$$\begin{aligned}
\bar{y} = \bar{x} &= \frac{1}{A}\left[\int_0^1 x[2 - \sqrt{1 - x^2}]\,\mathrm{d}x + \int_1^2 x[2 - 0]\,\mathrm{d}x\right] \\
&= \frac{4}{16 - \pi}\left[x^2\big|_0^1 + x^2\big|_1^2 - \int_0^1 x\sqrt{1 - x^2}\,\mathrm{d}x\right] \\
&= \frac{4}{16 - \pi}\left[4 - \frac{1}{3}\right] \qquad \text{by (2.3.5) with } r = 1 \\
&= \frac{44}{48 - 3\pi}
\end{aligned}$$

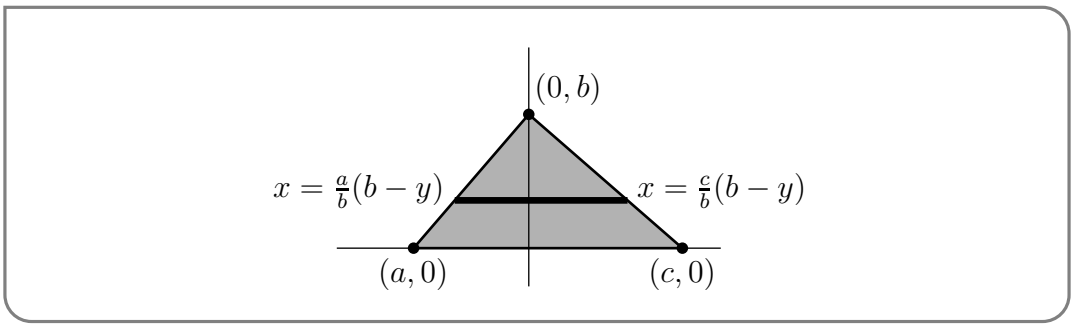
Example 2.3.6

Example 2.3.7

Prove that the centroid of any triangle is located at the point of intersection of the medians. A median of a triangle is a line segment joining a vertex to the midpoint of the opposite side.



Solution. Choose a coordinate system so that the vertices of the triangle are located at $(a,0)$, $(0,b)$ and $(c,0)$. (In the figure below, a is negative.) The line joining $(a,0)$ and $(0,b)$



has equation $bx + ay = ab$. (Check that $(a,0)$ and $(0,b)$ both really are on this line.) The line joining $(c,0)$ and $(0,b)$ has equation $bx + cy = bc$. (Check that $(c,0)$ and $(0,b)$ both really are on this line.) Hence for each fixed y between 0 and b, x runs from $a - \frac{a}{b}y$ to $c - \frac{c}{b}y$.

We'll use horizontal strips to compute $\bar{x}$ and $\bar{y}$. We could just apply (2.3.4) with $c = 0$, $d = b$, $R(y) = \frac{c}{b}(b-y)$ (which is gotten by solving $bx + cy = bc$ for x) and $L(y) = \frac{a}{b}(b-y)$ (which is gotten by solving $bx + ay = ab$ for x).

But rather than memorizing or looking up those formulae, we'll derive them for this example. So consider a thin strip at height y as illustrated in the figure above.

- The strip has length
$$\ell(y) = \Big[\frac{c}{b}(b-y) - \frac{a}{b}(b-y)\Big] = \frac{c-a}{b}(b-y)$$
- The strip has width $\mathrm{d}y$.
- On this strip, y has average value y.
- On this strip, x has average value $\frac{1}{2}\big[\frac{a}{b}(b-y) + \frac{c}{b}(b-y)\big] = \frac{a+c}{2b}(b-y)$.

As the area of the triangle is $A = \frac{1}{2}(c-a)b$,

$$\bar{y} = \frac{1}{A}\int_0^b y\,\ell(y)\,\mathrm{d}y = \frac{2}{(c-a)b}\int_0^b y\frac{c-a}{b}(b-y)\,\mathrm{d}y = \frac{2}{b^2}\int_0^b (by - y^2)\,\mathrm{d}y = \frac{2}{b^2}\Big(b\frac{b^2}{2} - \frac{b^3}{3}\Big)$$
$$= \frac{2}{b^2}\frac{b^3}{6} = \frac{b}{3}$$

$$\bar{x} = \frac{1}{A}\int_0^b \frac{a+c}{2b}(b-y)\,\ell(y)\,dy = \frac{2}{(c-a)b}\int_0^b \frac{a+c}{2b}(b-y)\frac{c-a}{b}(b-y)\,dy = \frac{a+c}{b^3}\int_0^b (y-b)^2\,dy$$
$$= \frac{a+c}{b^3}\Big[\frac{1}{3}(y-b)^3\Big]_0^b = \frac{a+c}{b^3}\,\frac{b^3}{3} = \frac{a+c}{3}$$

We have found that the centroid of the triangle is at $(\bar{x},\bar{y}) = \big(\frac{a+c}{3},\frac{b}{3}\big)$. We shall now show that this point lies on all three medians.

- One vertex is at $(a,0)$. The opposite side runs from $(0,b)$ and $(c,0)$ and so has midpoint $\frac{1}{2}(c,b)$. The line from $(a,0)$ to $\frac{1}{2}(c,b)$ has slope $\frac{b/2}{c/2-a} = \frac{b}{c-2a}$ and so has equation $y = \frac{b}{c-2a}(x-a)$. As $\frac{b}{c-2a}(\bar{x}-a) = \frac{b}{c-2a}\big(\frac{a+c}{3}-a\big) = \frac{1}{3}\frac{b}{c-2a}(c+a-3a) = \frac{b}{3} = \bar{y}$, the centroid does indeed lie on this median. In this computation we have implicitly assumed that $c \neq 2a$ so that the denominator $c-2a \neq 0$. In the event that $c = 2a$, the median runs from $(a,0)$ to $\big(a,\frac{b}{2}\big)$ and so has equation $x = a$. When $c = 2a$ we also have $\bar{x} = \frac{a+c}{3} = a$, so that the centroid still lies on the median.
- Another vertex is at $(c,0)$. The opposite side runs from $(a,0)$ and $(0,b)$ and so has midpoint $\frac{1}{2}(a,b)$. The line from $(c,0)$ to $\frac{1}{2}(a,b)$ has slope $\frac{b/2}{a/2-c} = \frac{b}{a-2c}$ and so has equation $y = \frac{b}{a-2c}(x-c)$. As $\frac{b}{a-2c}(\bar{x}-c) = \frac{b}{a-2c}\big(\frac{a+c}{3}-c\big) = \frac{1}{3}\frac{b}{a-2c}(a+c-3c) = \frac{b}{3} = \bar{y}$, the centroid does indeed lie on this median. In this computation we have implicitly assumed that $a \neq 2c$ so that the denominator $a-2c \neq 0$. In the event that $a = 2c$, the median runs from $(c,0)$ to $\big(c,\frac{b}{2}\big)$ and so has equation $x = c$. When $a = 2c$ we also have $\bar{x} = \frac{a+c}{3} = c$, so that the centroid still lies on the median.
- The third vertex is at $(0,b)$. The opposite side runs from $(a,0)$ and $(c,0)$ and so has midpoint $\big(\frac{a+c}{2},0\big)$. The line from $(0,b)$ to $\big(\frac{a+c}{2},0\big)$ has slope $\frac{-b}{(a+c)/2} = -\frac{2b}{a+c}$ and so has equation $y = b - \frac{2b}{a+c}x$. As $b - \frac{2b}{a+c}\bar{x} = b - \frac{2b}{a+c}\frac{a+c}{3} = \frac{b}{3} = \bar{y}$, the centroid does indeed lie on this median. This time, we have implicitly assumed that $a+c \neq 0$. In the event that $a+c = 0$, the median runs from $(0,b)$ to $(0,0)$ and so has equation $x = 0$. When $a+c = 0$ we also have $\bar{x} = \frac{a+c}{3} = 0$, so that the centroid still lies on the median.

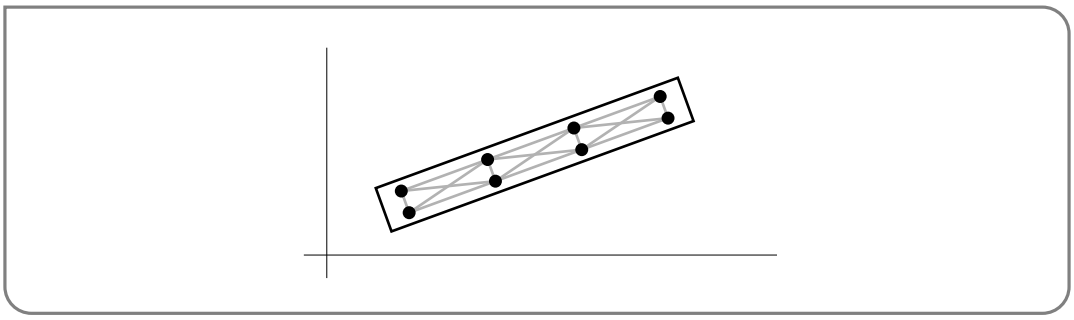
Example 2.3.7

2.3.2 ▸▸ Optional — Torque

Newton's law of motion says that the position $x(t)$ of a single particle moving under the influence of a force F obeys $mx''(t) = F$. Similarly, the positions $x_i(t)$, $1 \leqslant i \leqslant n$, of a set of particles moving under the influence of forces F_i obey $mx_i''(t) = F_i$, $1 \leqslant i \leqslant n$. Often systems of interest consist of some small number of rigid bodies. Suppose that we are interested in the motion of a single rigid body, say a piece of wood. The piece of wood is made up of a huge number of atoms. So the system of equations determining the motion of all of the individual atoms in the piece of wood is huge. On the other hand, because the piece of wood is rigid, its configuration is completely determined by the position of, for example, its centre of mass and its orientation. (Rather than get into what is precisely meant by "orientation", let's just say that it is certainly determined by, for example, the

positions of a few of the corners of the piece of wood). It is possible to extract from the huge system of equations that determine the motion of all of the individual atoms, a small system of equations that determine the motion of the centre of mass and the orientation. We can avoid some vector analysis, that is beyond the scope of this course, by assuming that our rigid body is moving in two rather than three dimensions.

So, imagine a piece of wood moving in the xy–plane. Furthermore, imagine that the



piece of wood consists of a huge number of particles joined by a huge number of weightless but very strong steel rods. The steel rod joining particle number one to particle number two just represents a force acting between particles number one and two. Suppose that

- there are n particles, with particle number i having mass m_i
- at time t, particle number i has x–coordinate $x_i(t)$ and y–coordinate $y_i(t)$
- at time t, the external force (gravity and the like) acting on particle number i has x–coordinate $H_i(t)$ and y–coordinate $V_i(t)$. Here H stands for horizontal and V stands for vertical.
- at time t, the force acting on particle number i, due to the steel rod joining particle number i to particle number j has x–coordinate $H_{i,j}(t)$ and y–coordinate $V_{i,j}(t)$. If there is no steel rod joining particles number i and j, just set $H_{i,j}(t) = V_{i,j}(t) = 0$. In particular, $H_{i,i}(t) = V_{i,i}(t) = 0$.

The only assumptions that we shall make about the steel rod forces are

(A1) for each $i \neq j$, $H_{i,j}(t) = -H_{j,i}(t)$ and $V_{i,j}(t) = -V_{j,i}(t)$. In words, the steel rod joining particles i and j applies equal and opposite forces to particles i and j.

(A2) for each $i \neq j$, there is a function $M_{i,j}(t)$ such that $H_{i,j}(t) = M_{i,j}(t)\big[x_i(t) - x_j(t)\big]$ and $V_{i,j}(t) = M_{i,j}(t)\big[y_i(t) - y_j(t)\big]$. In words, the force due to the rod joining particles i and j acts parallel to the line joining particles i and j. For (A1) to be true, we need $M_{i,j}(t) = M_{j,i}(t)$.

Newton's law of motion, applied to particle number i, now tells us that

$$m_i x_i''(t) = H_i(t) + \sum_{j=1}^{n} H_{i,j}(t) \qquad (X_i)$$

$$m_i y_i''(t) = V_i(t) + \sum_{j=1}^{n} V_{i,j}(t) \qquad (Y_i)$$

Adding up all of the equations (X_i), for $i = 1, 2, 3, \cdots, n$ and adding up all of the equations (Y_i), for $i = 1, 2, 3, \cdots, n$ gives

$$\sum_{i=1}^{n} m_i x_i''(t) = \sum_{i=1}^{n} H_i(t) + \sum_{1 \leqslant i,j \leqslant n} H_{i,j}(t) \qquad (\Sigma_i X_i)$$

$$\sum_{i=1}^{n} m_i y_i''(t) = \sum_{i=1}^{n} V_i(t) + \sum_{1 \leqslant i,j \leqslant n} V_{i,j}(t) \qquad (\Sigma_i Y_i)$$

The sum $\sum_{1 \leqslant i,j \leqslant n} H_{i,j}(t)$ contains $H_{1,2}(t)$ exactly once and it also contains $H_{2,1}(t)$ exactly once and these two terms cancel exactly, by assumption (A1). In this way, all terms in $\sum_{1 \leqslant i,j \leqslant n} H_{i,j}(t)$ with $i \neq j$ exactly cancel. All terms with $i = j$ are assumed to be zero. So $\sum_{1 \leqslant i,j \leqslant n} H_{i,j}(t) = 0$. Similarly, $\sum_{1 \leqslant i,j \leqslant n} V_{i,j}(t) = 0$, so the equations $(\Sigma_i X_i)$ and $(\Sigma_i Y_i)$ simplify to

$$\sum_{i=1}^{n} m_i x_i''(t) = \sum_{i=1}^{n} H_i(t) \qquad (\Sigma_i X_i)$$

$$\sum_{i=1}^{n} m_i y_i''(t) = \sum_{i=1}^{n} V_i(t) \qquad (\Sigma_i Y_i)$$

Denote by

$$M = \sum_{i=1}^{n} m_i$$

the total mass of the system, by

$$X(t) = \frac{1}{M} \sum_{i=1}^{n} m_i x_i(t) \qquad \text{and} \qquad Y(t) = \frac{1}{M} \sum_{i=1}^{n} m_i y_i(t)$$

the x– and y–coordinates of the centre of mass of the system at time t and by

$$H(t) = \sum_{i=1}^{n} H_i(t) \qquad \text{and} \qquad V(t) = \sum_{i=1}^{n} V_i(t)$$

the x– and y–coordinates of the total external force acting on the system at time t. In this notation, the equations $(\Sigma_i X_i)$ and $(\Sigma_i Y_i)$ are

$$MX''(t) = H(t) \qquad MY''(t) = V(t) \tag{2.3.6}$$

So the centre of mass of the system moves just like a single particle of mass M subject to the total external force.

Now multiply equation (Y_i) by $x_i(t)$, subtract from it equation (X_i) multiplied by $y_i(t)$, and sum over i. This gives the equation $\sum_i \big[x_i(t)\,(Y_i) - y_i(t)\,(X_i)\big]$:

$$\sum_{i=1}^{n} m_i \big[x_i(t) y_i''(t) - y_i(t) x_i''(t)\big] = \sum_{i=1}^{n} \big[x_i(t) V_i(t) - y_i(t) H_i(t)\big] + \sum_{1 \leqslant i,j \leqslant n} \big[x_i(t) V_{i,j}(t) - y_i(t) H_{i,j}(t)\big]$$

By the assumption (A2)

$$\begin{aligned} x_1(t)V_{1,2}(t) - y_1(t)H_{1,2}(t) &= x_1(t)M_{1,2}(t)\big[y_1(t) - y_2(t)\big] - y_1(t)M_{1,2}(t)\big[x_1(t) - x_2(t)\big] \\ &= M_{1,2}(t)\big[y_1(t)x_2(t) - x_1(t)y_2(t)\big] \\ x_2(t)V_{2,1}(t) - y_2(t)H_{2,1}(t) &= x_2(t)M_{2,1}(t)\big[y_2(t) - y_1(t)\big] - y_2(t)M_{2,1}(t)\big[x_2(t) - x_1(t)\big] \\ &= M_{2,1}(t)\big[-y_1(t)x_2(t) + x_1(t)y_2(t)\big] \\ &= M_{1,2}(t)\big[-y_1(t)x_2(t) + x_1(t)y_2(t)\big] \end{aligned}$$

So the $i = 1$, $j = 2$ term in $\sum_{1\leqslant i,j\leqslant n}\big[x_i(t)V_{i,j}(t) - y_i(t)H_{i,j}(t)\big]$ exactly cancels the $i = 2$, $j = 1$ term. In this way all of the terms in $\sum_{1\leqslant i,j\leqslant n}\big[x_i(t)V_{i,j}(t) - y_i(t)H_{i,j}(t)\big]$ with $i \neq j$ cancel. Each term with $i = j$ is exactly zero. So $\sum_{1\leqslant i,j\leqslant n}\big[x_i(t)V_{i,j}(t) - y_i(t)H_{i,j}(t)\big] = 0$ and

$$\sum_{i=1}^{n} m_i\big[x_i(t)y_i''(t) - y_i(t)x_i''(t)\big] = \sum_{i=1}^{n}\big[x_i(t)V_i(t) - y_i(t)H_i(t)\big]$$

Define

$$\begin{aligned} L(t) &= \sum_{i=1}^{n} m_i\big[x_i(t)y_i'(t) - y_i(t)x_i'(t)\big] \\ T(t) &= \sum_{i=1}^{n}\big[x_i(t)V_i(t) - y_i(t)H_i(t)\big] \end{aligned}$$

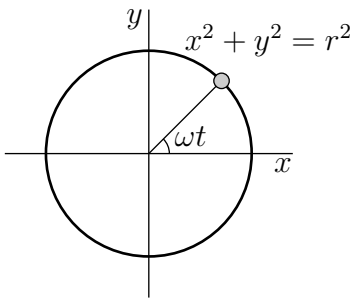
In this notation

$$\frac{d}{dt}L(t) = T(t) \tag{2.3.7}$$

- Equation (2.3.7) plays the role of Newton's law of motion for rotational motion.
- $T(t)$ is called the torque and plays the role of "rotational force".
- $L(t)$ is called the angular momentum (about the origin) and is a measure of the rate at which the piece of wood is rotating.
  - For example, if a particle of mass m is traveling in a circle of radius r, centred on the origin, at ω radians per unit time, then $x(t) = r\cos(\omega t)$, $y(t) = r\sin(\omega t)$ and

    $$\begin{aligned} m\big[x(t)y'(t) - y(t)x'(t)\big] &= m\big[r\cos(\omega t)\; r\omega\cos(\omega t) - r\sin(\omega t)\;\big(-r\omega\sin(\omega t)\big)\big] \\ &= mr^2\,\omega \end{aligned}$$

    is proportional to ω, which is the rate of rotation about the origin.



In any event, in order for the piece of wood to remain stationary, that is to have $x_i(t)$ and $y_i(t)$ be constant for all $1 \leqslant i \leqslant n$, we need to have

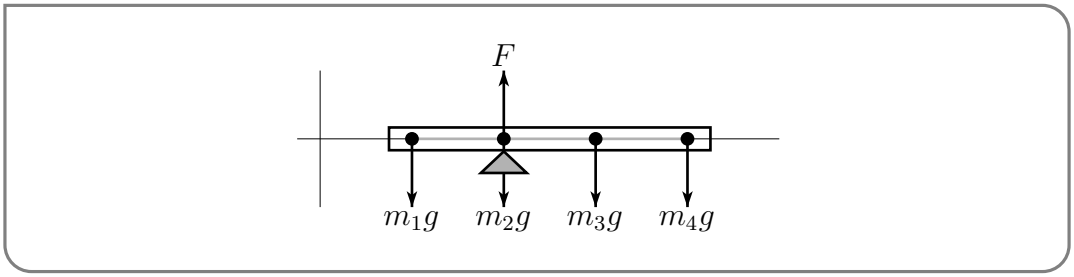
$$X''(y) = Y''(t) = L(t) = 0$$

and then equations (2.3.6) and (2.3.7) force

$$H(t) = V(t) = T(t) = 0$$

Now suppose that the piece of wood is a seesaw that is long and thin and is lying on the x–axis, supported on a fulcrum at $x = p$. Then every $y_i = 0$ and the torque simplifies to $T(t) = \sum_{i=1}^{n} x_i(t)V_i(t)$. The forces consist of

- gravity, $m_i g$, acting downwards on particle number i, for each $1 \leqslant i \leqslant n$ and the
- force F imposed by the fulcrum that is pushing straight up on the particle at $x = p$.



So

- The net vertical force is $V(t) = F - \sum_{i=1}^{n} m_i g = F - Mg$. If the seesaw is to remain stationary, this must be zero so that $F = Mg$.
- The total torque (about the origin) is

  $$T = Fp - \sum_{i=1}^{n} m_i g x_i = Mgp - \sum_{i=1}^{n} m_i g x_i$$

  If the seesaw is to remain stationary, this must also be zero and the fulcrum must be placed at

  $$p = \frac{1}{M} \sum_{i=1}^{n} m_i x_i \tag{2.3.8}$$

  which is the centre of mass of the piece of wood.

2.4▲ Separable 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 β–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 will just look at one special, but important, type of equation.

2.4.1 ▸▸ Separate and Integrate

Definition 2.4.1.

A *separable differential equation* is an equation for a function $y(x)$ of the form

$$\frac{dy}{dx}(x) = f(x)\, g\big(y(x)\big)$$

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[20]

$$\frac{dy}{dx} = f(x)\, g(y)$$

and solves such an equation by cross multiplying/dividing to get all of the y's, including the dy on one side of the equation and all of the x's, including the dx, on the other side of the equation.

$$\frac{dy}{g(y)} = f(x)\, dx$$

20 Look at the right hand side of the equation. The x–dependence is separated from the y–dependence. That's the reason for the name "separable".

(We are of course assuming that $g(y)$ is nonzero.) Then you integrate both sides

$$\int \frac{dy}{g(y)} = \int f(x)\,dx \tag{2.4.1}$$

This looks illegal, and indeed is illegal — $\frac{dy}{dx}$ is not a fraction. But we'll now see that the answer is still correct. This procedure is simply a mnemonic device to help you remember the answer (2.4.1).

- Our goal is to find all functions $y(x)$ that obey $\frac{dy}{dx}(x) = f(x)\,g\big(y(x)\big)$.
- Assuming that g is nonzero,

$$\begin{aligned} y'(x) = f(x)\,g(y(x)) &\iff \frac{y'(x)}{g(y(x))} = f(x) \iff \int \frac{y'(x)}{g(y(x))}\,dx = \int f(x)\,dx \\ &\iff \int \frac{dy}{g(y)}\bigg|_{y=y(x)} = \int f(x)\,dx \\ &\qquad \text{with the substitution } y = y(x),\ dy = y'(x)\,dx \end{aligned}$$

- That's our answer (2.4.1) again.

Let $G(y)$ be an antiderivative of $\frac{1}{g(y)}$ (i.e. $G'(y) = \frac{1}{g(y)}$) and $F(x)$ be an antiderivative of $f(x)$ (i.e. $F'(x) = f(x)$). If we reinstate the argument of y, (2.4.1) is

$$G\big(y(x)\big) = F(x) + C \tag{2.4.2}$$

Observe that the solution (2.4.2) contains an arbitrary constant, C. The value of this arbitrary constant *can not* be determined by the differential equation. You need additional data to determine it. Often this data consists of the value of the unknown function for one value of x. That is, often the problem you have to solve is of the form

$$\frac{dy}{dx}(x) = f(x)\,g\big(y(x)\big) \qquad y(x_0) = y_0$$

where $f(x)$ and $g(y)$ are given functions and x_0 and y_0 are given numbers. This type of problem is called an "initial value problem". It is solved by first using the method above to find the general solution to the differential equation, including the arbitrary constant C, and then using the "initial condition" $y(x_0) = y_0$ to determine the value of C. We'll see examples of this shortly.

Example 2.4.2

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 to the right hand side.

$$e^{y}\,dy = x\,dx$$

Then we integrate both sides

$$\int e^y \mathrm{d}y = \int x \mathrm{d}x \iff e^y = \frac{x^2}{2} + C$$

The C on the right hand side contains both the arbitrary constant for the indefinite integral $\int e^y \mathrm{d}y$ and the arbitrary constant for the indefinite integral $\int x \mathrm{d}x$. Finally, we solve for y, which is really a function of x.

$$y(x) = \log\left(\frac{x^2}{2} + C\right)$$

Recall that we are using log to refer to the natural (base e) logarithm.

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. For example, suppose that we wish to find a function $y(x)$ that obeys both

$$\frac{\mathrm{d}y}{\mathrm{d}x} = xe^{-y} \qquad \text{and} \qquad y(0) = 1$$

We know that, to have $\frac{\mathrm{d}y}{\mathrm{d}x} = xe^{-y}$ satisfied, we must have $y(x) = \log\left(\frac{x^2}{2} + C\right)$, for some constant C. To also have $y(0) = 1$, we must have

$$1 = y(0) = \log\left(\frac{x^2}{2} + C\right)\Big|_{x=0} = \log C \iff \log C = 1 \iff C = e$$

So our final solution is $y(x) = \log\left(\frac{x^2}{2} + e\right)$.

Example 2.4.2

Example 2.4.3

Let a and b be any two constants. We'll now solve the family of differential equations

$$\frac{\mathrm{d}y}{\mathrm{d}x} = a(y - b)$$

using our mnemonic device.

$$\begin{aligned}\frac{\mathrm{d}y}{y-b} = a\,\mathrm{d}x \implies \int \frac{\mathrm{d}y}{y-b} = \int a\,\mathrm{d}x \implies \log|y-b| = ax + c \implies |y-b| = e^{ax+c} = e^c e^{ax} \\ \implies y - b = Ce^{ax}\end{aligned}$$

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 $\frac{\mathrm{d}y}{\mathrm{d}x}$ and $a(y - b)$ are zero. So the general solution to

$\frac{dy}{dx} = a(y-b)$ is $y(x) = Ce^{ax} + b$, where the constant C can be any real number. Note that when $y(x) = Ce^{ax} + b$ 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 2.4.3

This is worth stating as a theorem.

Theorem 2.4.4.

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

$$y(x) = \{y(0) - b\}\, e^{ax} + b$$

Example 2.4.5

Solve $\frac{dy}{dx} = y^2$

Solution. When $y \neq 0$,

$$\frac{dy}{dx} = y^2 \implies \frac{dy}{y^2} = dx \implies \frac{y^{-1}}{-1} = x + C \implies y = -\frac{1}{x+C}$$

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:

$$y(x) = 0 \implies y'(x) = 0,\ y(x)^2 = 0 \implies y'(x) = y(x)^2$$

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 2.4.5

Example 2.4.6

When a raindrop falls it increases in size so that its mass $m(t)$, is a function of time t. The rate of growth of mass, i.e. $\frac{dm}{dt}$, is $km(t)$ for some positive constant k. According to Newton's law of motion, $\frac{d}{dt}(mv) = gm$, where v is the velocity of the raindrop (with v being positive for downward motion) and g is the acceleration due to gravity. Find the terminal velocity, $\lim_{t\to\infty} v(t)$, of a raindrop.

Solution. In this problem we have two unknown functions, $m(t)$ and $v(t)$, and two differential equations, $\frac{dm}{dt} = km$ and $\frac{\mathrm{d}}{\mathrm{d}t}(mv) = gm$. The first differential equation, $\frac{dm}{dt} = km$, involves only $m(t)$, not $v(t)$, so we use it to determine $m(t)$. By Theorem 2.4.4, with $b = 0$, $a = k$, y replaced by m and x replaced by t,

$$\frac{\mathrm{d}m}{\mathrm{d}t} = km \implies m(t) = m(0)e^{kt}$$

Now that we know $m(t)$ (except for the value of the constant $m(0)$), we can substitute it into the second differential equation, which we can then use to determine the remaining unknown function $v(t)$. Observe that the second equation, $\frac{\mathrm{d}}{\mathrm{d}t}(mv) = gm(t) = gm(0)e^{kt}$ tells that the derivative of the function $y(t) = m(t)v(t)$ is $gm(0)e^{kt}$. So $y(t)$ is just an antiderivative of $gm(0)e^{kt}$.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = gm(t) = gm(0)e^{kt} \implies y(t) = \int gm(0)e^{kt}\,\mathrm{d}t = gm(0)\frac{e^{kt}}{k} + C$$

Now that we know $y(t) = m(t)v(t) = m(0)e^{kt}v(t)$, we can get $v(t)$ just by dividing out the $m(0)e^{kt}$.

$$y(t) = gm(0)\frac{e^{kt}}{k} + C \implies m(0)e^{kt}v(t) = gm(0)\frac{e^{kt}}{k} + C \implies v(t) = \frac{g}{k} + \frac{C}{m(0)e^{kt}}$$

Our solution, $v(t)$, contains two arbitrary constants, namely C and $m(0)$. They will be determined by, for example, the mass and velocity at time $t = 0$. But since we are only interested in the terminal velocity $\lim_{t\to\infty} v(t)$, we don't need to know C and $m(0)$. Since $k > 0$, $\lim_{t\to\infty} \frac{C}{e^{kt}} = 0$ and the terminal velocity $\lim_{t\to\infty} v(t) = \frac{g}{k}$.

Example 2.4.6

Example 2.4.7

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

$$\frac{\mathrm{d}C}{\mathrm{d}t} = r - kC$$

where k is a positive constant of proportionality.

(a) Express $C(t)$ in terms of k and $C(0)$.

(b) Find $\lim_{t\to\infty} C(t)$.

Solution. (a) Since $r - kC = -k\big(C - \frac{r}{k}\big)$ the given equation is

$$\frac{\mathrm{d}C}{\mathrm{d}t} = -k\big(C - \frac{r}{k}\big)$$

which is of the form solved in Theorem 2.4.4 with $a = -k$ and $b = \frac{r}{k}$. So the solution is

$$C(t) = \frac{r}{k} + \Big(C(0) - \frac{r}{k}\Big)e^{-kt}$$

(b) For any $k > 0$, $\lim_{t\to\infty} e^{-kt} = 0$. Consequently, for any $C(0)$ and any $k > 0$, $\lim_{t\to\infty} C(t) = \frac{r}{k}$. 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 $t \to \infty$ gives

$$0 = r - kC_e \implies C_e = \frac{r}{k}$$

Example 2.4.7

2.4.2 ▸▸ Optional — Carbon Dating

Scientists can determine the age of objects containing organic material by a method called *carbon dating* or *radiocarbon dating*[21]. The bombardment of the upper atmosphere by cosmic rays converts nitrogen to a radioactive isotope of carbon, ^{14}C, with a half–life of about 5730 years. Vegetation absorbs carbon dioxide from the atmosphere through photosynthesis and animals acquire ^{14}C by eating plants. When a plant or animal dies, it stops replacing its carbon and the amount of ^{14}C begins to decrease through radioactive decay. Therefore the level of radioactivity also decreases. More precisely, let $Q(t)$ denote the amount of ^{14}C in the plant or animal t years after it dies. The number of radioactive decays per unit time, at time t, is proportional to the amount of ^{14}C present at time t, which is $Q(t)$. Thus

$$\frac{\mathrm{d}Q}{\mathrm{d}t}(t) = -kQ(t) \tag{2.4.3}$$

Here k is a constant of proportionality that is determined by the half–life. We shall explain what half–life is, and also determine the value of k, in Example 2.4.8, below.

Before we do so, let's think about the sign in (2.4.3).

- Recall that $Q(t)$ denotes a quantity, namely the amount of ^{14}C present at time t. There cannot be a negative amount of ^{14}C. Nor can this quantity be zero. (We would not use carbon dating when there is no ^{14}C present.) Consequently, $Q(t) > 0$.
- As the time t increases, $Q(t)$ decreases, because ^{14}C is being continuously converted into ^{14}N by radioactive decay[22]. Thus $\frac{\mathrm{d}Q}{\mathrm{d}t}(t) < 0$.
- The signs $Q(t) > 0$ and $\frac{\mathrm{d}Q}{\mathrm{d}t}(t) < 0$ are consistent with (2.4.3) provided the constant of proportionality $k > 0$.

21 Willard Libby, of Chicago University was awarded the Nobel Prize in Chemistry in 1960, for developing radiocarbon dating.

22 The precise transition is $^{14}C \to {}^{14}N + e^- + \bar{\nu}_e$ where e^- is an electron and $\bar{\nu}_e$ is an electron neutrino.

- In (2.4.3), we chose to call the constant of proportionality "$-k$". We did so in order to make $k > 0$. We could just as well have chosen to call the constant of proportionality "K". That is, we could have replaced (2.4.3) by $\frac{dQ}{dt}(t) = KQ(t)$. The constant of proportionality K would have to be negative, (and K and k would be related by $K = -k$).

Example 2.4.8

In this example, we determine the value of the constant of proportionality k in (2.4.3) that corresponds to the half–life of ^{14}C, which is 5730 years.

- Imagine that some plant or animal contains a quantity Q_0 of ^{14}C at its time of death. Let's choose the zero point of time $t = 0$ to be the instant that the plant or animal died.
- Denote by $Q(t)$ the amount of ^{14}C in the plant or animal t years after it died. Then $Q(t)$ must obey both (2.4.3) and $Q(0) = Q_0$.
- Theorem 2.4.4, with $b = 0$ and $a = -k$, then tells us that $Q(t) = Q_0 e^{-kt}$ for all $t \geqslant 0$.
- By definition, the half–life of ^{14}C is the length of time that it takes for half of the ^{14}C to decay. That is, the half–life $t_{1/2}$ is determined by

$$\begin{aligned} Q(t_{1/2}) = \tfrac{1}{2}Q(0) = \tfrac{1}{2}Q_0 \qquad & \text{but we know that } Q(t) = Q_0 e^{-kt} \\ Q_0 e^{-kt_{1/2}} = \tfrac{1}{2}Q_0 \qquad & \text{now cancel } Q_0 \\ e^{-kt_{1/2}} = \tfrac{1}{2} & \end{aligned}$$

  Taking the logarithm of both sides gives

$$-kt_{1/2} = \log\frac{1}{2} = -\log 2 \implies k = \frac{\log 2}{t_{1/2}}$$

  Recall that, in this text, we use $\log x$ to indicate the natural logarithm. That is,

$$\log x = \log_e x = \log x$$

We are told that, for ^{14}C, the half–life $t_{1/2} = 5730$, so

$$k = \frac{\log 2}{5730} = 0.000121 \qquad \text{to 6 decimal places}$$

Example 2.4.8

From the work in the above example we have accumulated enough new facts to make a corollary to Theorem 2.4.4.

Corollary 2.4.9.

The function $Q(t)$ satisfies the equation

$$\frac{dQ}{dt} = -kQ(t)$$

if and only if

$$Q(t) = Q(0)\, e^{-kt}$$

The half–life is defined to be the time $t_{1/2}$ which obeys

$$Q(t_{1/2}) = \frac{1}{2}\, Q(0)$$

The half–life is related to the constant k by

$$t_{1/2} = \frac{\log 2}{k}$$

Now here is a typical problem that is solved using Corollary 2.4.9.

Example 2.4.10

A particular piece of parchment contains about 64% as much ^{14}C as plants do today. Estimate the age of the parchment.

Solution. Let $Q(t)$ denote the amount of ^{14}C in the parchment t years after it was first created. By (2.4.3) and Example 2.4.8

$$\frac{dQ}{dt}(t) = -kQ(t) \qquad \text{with } k = \frac{\log 2}{5730} = 0.000121$$

By Corollary 2.4.9

$$Q(t) = Q(0)\, e^{-kt}$$

The time at which $Q(t)$ reaches $0.64\, Q(0)$ is determined by

$$\begin{aligned} Q(t) &= 0.64\, Q(0) && \text{but } Q(t) = Q(0)\, e^{-kt} \\ Q(0)\, e^{-kt} &= 0.64\, Q(0) && \text{cancel } Q(0) \\ e^{-kt} &= 0.64 && \text{take logarithms} \\ -kt &= \log 0.64 \\ t &= \frac{\log 0.64}{-k} = \frac{\log 0.64}{-0.000121} = 3700 && \text{to 2 significant digits} \end{aligned}$$

That is, the parchment[23] is about 37 centuries old.

23 The British Museum has an Egyptian mathematical text from the seventeenth century B.C.

Example 2.4.10

We have stated that the half-life of ^{14}C is 5730 years. How can this be determined? We can explain this using the following example.

Example 2.4.11

A scientist in a B-grade science fiction film is studying a sample of the rare and fictitious element, implausium. With great effort he has produced a sample of pure implausium. The next day — 17 hours later — he comes back to his lab and discovers that his sample is now only 37% pure. What is the half-life of the element?

Solution. We can again set up our problem using Corollary 2.4.9. Let $Q(t)$ denote the quantity of implausium at time t, measured in hours. Then we know

$$Q(t) = Q(0) \cdot e^{-kt}$$

We also know that

$$Q(17) = 0.37Q(0).$$

That enables us to determine k via

$$\begin{aligned} Q(17) = 0.37Q(0) &= Q(0)e^{-17k} && \text{divide both sides by } Q(0) \\ 0.37 &= e^{-17k} \end{aligned}$$

and so

$$k = -\frac{\log 0.37}{17} = 0.05849$$

We can then convert this to the half life using Corollary 2.4.9:

$$t_{1/2} = \frac{\log 2}{k} \approx 11.85 \text{ hours}$$

While this example is entirely fictitious, one really can use this approach to measure the half-life of materials.

Example 2.4.11

2.4.3 ▸▸ Optional — Newton's Law of Cooling

Newton's law of cooling says:

> The rate of change of temperature of an object is proportional to the difference in temperature between the object and its surroundings. The temperature of the surroundings is sometimes called the ambient temperature.

If we denote by $T(t)$ the temperature of the object at time t and by A the temperature of its surroundings, Newton's law of cooling says that there is some constant of proportionality, K, such that

$$\frac{\mathrm{d}T}{\mathrm{d}t}(t) = K\big[T(t) - A\big] \tag{2.4.4}$$

This mathematical model of temperature change works well when studying a small object in a large, fixed temperature, environment. For example, a hot cup of coffee in a large room[24]. Let's start by thinking a little about the sign of the constant of proportionality. At any time t, there are three possibilities.

- If $T(t) > A$, that is, if the body is warmer than its surroundings, we would expect heat to flow from the body into its surroundings and so we would expect the body to cool off so that $\frac{\mathrm{d}T}{\mathrm{d}t}(t) < 0$. For this expectation to be consistent with (2.4.4), we need $K < 0$.
- If $T(t) < A$, that is the body is cooler than its surroundings, we would expect heat to flow from the surroundings into the body and so we would expect the body to warm up so that $\frac{\mathrm{d}T}{\mathrm{d}t}(t) > 0$. For this expectation to be consistent with (2.4.4), we again need $K < 0$.
- Finally if $T(t) = A$, that is the body and its environment have the same temperature, we would not expect any heat to flow between the two and so we would expect that $\frac{\mathrm{d}T}{\mathrm{d}t}(t) = 0$. This does not impose any condition on K.

In conclusion, we would expect $K < 0$. Of course, we could have chosen to call the constant of proportionality $-k$, rather than K. Then the differential equation would be $\frac{\mathrm{d}T}{\mathrm{d}t} = -k\big(T - A\big)$ and we would expect $k > 0$.

Example 2.4.12

The temperature of a glass of iced tea is initially 5°. After 5 minutes, the tea has heated to 10° in a room where the air temperature is 30°.

(a) Determine the temperature as a function of time.

(b) What is the temperature after 10 minutes?

(c) Determine when the tea will reach a temperature of 20°.

Solution. (a)

- Denote by $T(t)$ the temperature of the tea t minutes after it was removed from the fridge, and let $A = 30$ be the ambient temperature.
- By Newton's law of cooling,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = K(T - A) = K(T - 30)$$

for some, as yet unknown, constant of proportionality K.

24 It does not work so well when the object is of a similar size to its surroundings since the temperature of the surroundings will rise as the object cools. It also fails when there are phase transitions involved — for example, an ice-cube melting in a warm room does not obey Newton's law of cooling.

- By Theorem 2.4.4 with $a = K$ and $b = 30$,

$$T(t) = [T(0) - 30]\, e^{Kt} + 30 = 30 - 25e^{Kt}$$

since the initial temperature $T(0) = 5$.

- This solution is not complete because it still contains an unknown constant, namely K. We have not yet used the given data that $T(5) = 10$. We can use it to determine K. At $t = 5$,

$$T(5) = 30 - 25e^{5K} = 10 \implies e^{5K} = \frac{20}{25} \implies 5K = \log\frac{20}{25}$$
$$\implies K = \frac{1}{5}\log\frac{4}{5} = -0.044629$$

to six decimal places.

(b) To find the temperature at 10 minutes we can just use the solution we have determined above.

$$\begin{aligned} T(10) &= 30 - 25e^{10K} \\ &= 30 - 25e^{10\times\frac{1}{5}\log\frac{4}{5}} \\ &= 30 - 25e^{2\log\frac{4}{5}} = 30 - 25e^{\log\frac{16}{25}} \\ &= 30 - 16 = 14^\circ \end{aligned}$$

(c) The temperature is 20° when

$$30 - 25e^{Kt} = 20 \implies e^{Kt} = \frac{10}{25} \implies Kt = \log\frac{10}{25}$$
$$\implies t = \frac{1}{K}\log\frac{2}{5} = 20.5 \text{ min}$$

to one decimal place.

Example 2.4.12

Example 2.4.13

A dead body is discovered at 3:45pm in a room where the temperature is 20°C. At that time the temperature of the body is 27°C. Two hours later, at 5:45pm, the temperature of the body is 25.3°C. What was the time of death? Note that the normal (adult human) body temperature is 37°C.

Solution. We will assume that the body's temperature obeys Newton's law of cooling.

- Denote by $T(t)$ the temperature of the body at time t, with $t = 0$ corresponding to 3:45pm. We wish to find the time of death — call it t_d.
- There is a lot of data in the statement of the problem. We are told

(1) the ambient temperature: $A = 20$

(2) the temperature of the body when discovered: $T(0) = 27$

(3) the temperature of the body 2 hours later: $T(2) = 25.3$

(4) assuming the person was a healthy adult right up until he died, the temperature at the time of death: $T(t_d) = 37$.

- Theorem 2.4.4 with $a = K$ and $b = A = 20$

$$T(t) = [T(0) - A]\, e^{Kt} + A = 20 + 7e^{Kt}$$

Two unknowns remain, K and t_d.

- We can find the first, K, by using the condition (3), which says $T(2) = 25.3$.

$$\begin{aligned} 25.3 = T(2) = 20 + 7e^{2K} &\implies 7e^{2K} = 5.3 \implies 2K = \log\left(\tfrac{5.3}{7}\right) \\ &\implies K = \tfrac{1}{2}\log\left(\tfrac{5.3}{7}\right) = -0.139 \end{aligned}$$

- Finally, t_d is determined by the condition (4).

$$\begin{aligned} 37 = T(t_d) = 20 + 7e^{-0.139t_d} &\implies e^{-0.139t_d} = \tfrac{17}{7} \implies -0.139t_d = \log\left(\tfrac{17}{7}\right) \\ &\implies t_d = -\tfrac{1}{0.139}\log\left(\tfrac{17}{7}\right) = -6.38 \end{aligned}$$

to two decimal places. Now 6.38 hours is 6 hours and $0.38 \times 60 = 23$ minutes. So the time of death was 6 hours and 23 minutes before 3:45pm, which is 9:22am.

Example 2.4.13

A slightly tricky example — we need to determine the ambient temperature from three measurements at different times.

Example 2.4.14

A glass of room-temperature water is carried out onto a balcony from an apartment where the temperature is 22°C. After one minute the water has temperature 26°C and after two minutes it has temperature 28°C. What is the outdoor temperature?

Solution. We will assume that the temperature of the water obeys Newton's law of cooling.

- Let A be the outdoor temperature and $T(t)$ be the temperature of the water t minutes after it is taken outside.

- By Newton's law of cooling,

$$T(t) = A + \big(T(0) - A\big)e^{Kt}$$

Theorem 2.4.4 with $a = K$ and $b = A$. Notice there are 3 unknowns here — A, $T(0)$ and K — so we need three pieces of information to find them all.

- We are told $T(0) = 22$, so

$$T(t) = A + \big(22 - A\big)e^{Kt}.$$

- We are also told $T(1) = 26$, which gives

$$\begin{aligned} 26 &= A + (22 - A)e^K && \text{rearrange things} \\ e^K &= \frac{26 - A}{22 - A} \end{aligned}$$

- Finally, $T(2) = 28$, so

$$\begin{aligned} 28 &= A + (22 - A)e^{2K} && \text{rearrange} \\ e^{2K} &= \frac{28 - A}{22 - A} && \text{but } e^K = \frac{26 - A}{22 - A}, \text{ so} \\ \left(\frac{26 - A}{22 - A}\right)^2 &= \frac{28 - A}{22 - A} && \text{multiply through by } (22 - A)^2 \\ (26 - A)^2 &= (28 - A)(22 - A) \end{aligned}$$

We can expand out both sides and collect up terms to get

$$\begin{aligned} \underbrace{26^2}_{=676} - 52A + A^2 &= \underbrace{28 \times 22}_{=616} - 50A + A^2 \\ 60 &= 2A \\ 30 &= A \end{aligned}$$

So the temperature outside is 30°.

Example 2.4.14

2.4.4 ▸▸ Optional — Population 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 β offspring (for some constant β) 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

$$\frac{P(t + t_g) - P(t)}{t_g} \approx \frac{\mathrm{d}P}{\mathrm{d}t}(t)$$

we get the differential equation

$$\frac{dP}{dt} = bP(t) \tag{2.4.5}$$

By Corollary 2.4.9, with $-k$ replaced by b,

$$P(t) = P(0) \cdot e^{bt} \tag{2.4.6}$$

This is called the Malthusian[25] 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* T to the time, the population size is *multiplied* by e^{bT}. In particular, the population size doubles every $\frac{\log 2}{b}$ units of time. The Malthusian growth model can be a reasonably good model only when the population size is very small compared to its environment[26]. A more sophisticated model of population growth, that takes into account the "carrying capacity of the environment" is considered below.

Example 2.4.15

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. We follow our usual pattern for dealing with such problems.

- 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 (2.4.5). So, by (2.4.6)

  $$P(t) = P(0) \cdot e^{bt}$$

  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

$$\begin{aligned} 4 &= 2 \cdot e^{47b} && \text{clean up} \\ e^{47b} &= 2 && \text{take the log and clean up} \\ b &= \frac{\log 2}{47} = 0.0147 && \text{to 3 decimal places} \end{aligned}$$

25 This is named after Rev. Thomas Robert Malthus. He described this model in a 1798 paper called "An essay on the principle of population".

26 That is, the population has plenty of food and space to grow.

- We now know $P(t)$ completely, so we can easily determine the predicted population[27] in 2100, i.e. at $t = 2100 - 1927 = 173$.

$$P(173) = 2e^{173b} = 2e^{173\times 0.0147} = 25.4 \text{ billion}$$

- Finally, our crude model predicts that the population is 6 billion at the time t that obeys

$$\begin{aligned} P(t) = 2e^{bt} &= 6 && \text{clean up} \\ e^{bt} &= 3 && \text{take the log and clean up} \\ t &= \frac{\log 3}{b} = 47\frac{\log 3}{\log 2} = 74.5 \end{aligned}$$

which corresponds[28] to the middle of 2001.

Example 2.4.15

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

$$P'(t) = b_0\left(1 - \frac{P(t)}{K}\right)P(t)$$

This is a separable differential equation and we can solve it explicitly. We shall do so shortly. See Example 2.4.16, below. But, before doing that, we'll see what we can learn 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

$$\frac{\mathrm{d}P}{\mathrm{d}t}(t) = \big(\,6000 - 3P(t)\,\big)\,P(t)$$

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 $(6000 - 3P)\,P = 3(2000 - P)\,P$

27 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. Our prediction of a population of 25.4 billion at time $t = 173$ looks pretty poor.

28 The world population really reached 6 billion in about 1999. So our prediction that the population would reach 6 billion at about time $t = 74.5$ is not so bad.

- is zero for $P = 0,\ 2000$,
- is strictly positive for $0 < P < 2000$ and
- is strictly negative for $P > 2000$.

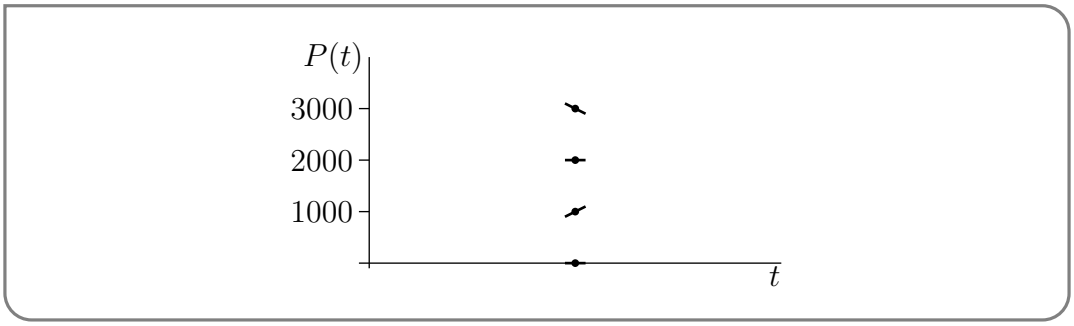
Consequently

$$\frac{\mathrm{d}P}{\mathrm{d}t}(t) \begin{cases} = 0 & \text{if } P(t) = 0 \\ > 0 & \text{if } 0 < P(t) < 2000 \\ = 0 & \text{if } P(t) = 2000 \\ < 0 & \text{if } P(t) > 2000 \end{cases}$$

Thus if $P(t)$ is some function that obeys $\frac{\mathrm{d}P}{\mathrm{d}t}(t) = \big(6000 - 3P(t)\big)P(t)$, then as the graph of $P(t)$ passes through the point $\big(t, P(t)\big)$

$$\text{the graph has } \begin{cases} \text{slope zero,} & \text{i.e. is horizontal,} & \text{if } P(t) = 0 \\ \text{positive slope,} & \text{i.e. is increasing,} & \text{if } 0 < P(t) < 2000 \\ \text{slope zero,} & \text{i.e. is horizontal,} & \text{if } P(t) = 2000 \\ \text{negative slope,} & \text{i.e. is decreasing,} & \text{if } 0 < P(t) < 2000 \end{cases}$$

as illustrated in the figure

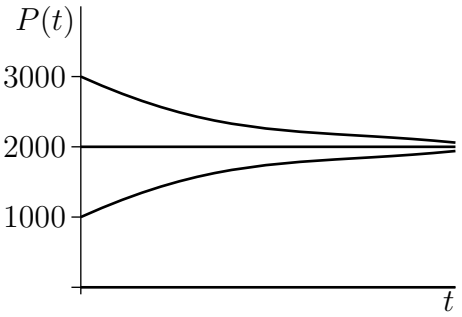


As a result,

- 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 $\frac{\mathrm{d}P}{\mathrm{d}t}(t) = \big(6000 - 3P(t)\big)P(t)$ 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 $\frac{\mathrm{d}P}{\mathrm{d}t}(t) = \big(6000 - 3P(t)\big)P(t)$ 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 $t \to \infty$.



Now we'll do an example in which we explicitly solve the logistic growth equation.

Example 2.4.16

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

$$\begin{aligned}
\frac{dy}{dt} = ay(K-y) &\implies \frac{dy}{y(K-y)} = a\,dt \implies \int \frac{1}{K}\Big[\frac{1}{y} - \frac{1}{y-K}\Big]dy = \int a\,dt \\
&\implies \frac{1}{K}[\log|y| - \log|y-K|] = at + C \\
&\implies \log\frac{|y|}{|y-K|} = aKt + CK \implies \Big|\frac{y}{y-K}\Big| = De^{aKt}
\end{aligned}$$

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.

$$\begin{aligned} y = (100-y)\frac{5}{95}e^{2t/95} &\implies 95y = (500-5y)e^{2t/95} \\ &\implies \left(95+5e^{2t/95}\right)y = 500e^{2t/95} \\ &\implies y = \frac{500e^{2t/95}}{95+5e^{2t/95}} = \frac{100e^{2t/95}}{19+e^{2t/95}} = \frac{100}{1+19e^{-2t/95}} \end{aligned}$$

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 2.4.16

2.4.5 ▸▸ Optional — Mixing Problems

Example 2.4.17

At time $t = 0$, where t is measured in minutes, a tank with a 5–litre capacity contains 3 litres of water in which 1 kg of salt is dissolved. Fresh water enters the tank at a rate of 2 litres per minute and the fully mixed solution leaks out of the tank at the *varying* rate of $2t$ litres per minute.

(a) Determine the volume of solution $V(t)$ in the tank at time t.

(b) Determine the amount of salt $Q(t)$ in solution when the amount of water in the tank is at maximum.

Solution. (a) The rate of change of the volume in the tank, at time t, is $2-2t$, because water is entering at a rate 2 and solution is leaking out at a rate $2t$. Thus

$$\frac{\mathrm{d}V}{\mathrm{d}t} = 2-2t \implies \mathrm{d}V = (2-2t)\,\mathrm{d}t \implies V = \int (2-2t)\,\mathrm{d}t = 2t - t^2 + C$$

at least until $V(t)$ reaches either the capacity of the tank or zero. When $t=0$, $V=3$ so $C=3$ and $V(t) = 3+2t-t^2$. Observe that $V(t)$ is at a maximum when $\frac{\mathrm{d}V}{\mathrm{d}t} = 2-2t = 0$, or $t=1$.

(b) In the very short time interval from time t to time $t+\mathrm{d}t$, $2t\,\mathrm{d}t$ litres of brine leaves the tank. That is, the fraction $\frac{2t\,\mathrm{d}t}{V(t)}$ of the total salt in the tank, namely $Q(t)\frac{2t\,\mathrm{d}t}{V(t)}$ kilograms, leaves. Thus salt is leaving the tank at the rate

$$\frac{Q(t)\frac{2t\,\mathrm{d}t}{V(t)}}{\mathrm{d}t} = \frac{2tQ(t)}{V(t)} = \frac{2tQ(t)}{3+2t-t^2} \text{ kilograms per minute}$$

so

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = -\frac{2tQ(t)}{3+2t-t^2} \implies \frac{dQ}{Q} = -\frac{2t}{3+2t-t^2}\mathrm{d}t = -\frac{2t}{(3-t)(1+t)}\mathrm{d}t = \Big[\frac{3/2}{t-3} + \frac{1/2}{t+1}\Big]\mathrm{d}t$$
$$\implies \log Q = \frac{3}{2}\log|t-3| + \frac{1}{2}\log|t+1| + C$$

We are interested in the time interval $0 \leqslant t \leqslant 1$. In this time interval $|t-3| = 3-t$ and $|t+1| = t+1$ so

$$\log Q = \frac{3}{2}\log(3-t) + \frac{1}{2}\log(t+1) + C$$

At $t=0$, Q is 1 so

$$\log 1 = \frac{3}{2}\log(3-0) + \frac{1}{2}\log(0+1) + C \implies C = \log 1 - \frac{3}{2}\log 3 - \frac{1}{2}\log 1 = -\frac{3}{2}\log 3$$

At $t=1$

$$\log Q = \frac{3}{2}\log(3-1) + \frac{1}{2}\log(1+1) - \frac{3}{2}\log 3 = 2\log 2 - \frac{3}{2}\log 3 = \log 4 - \log 3^{3/2}$$

so $Q = \frac{4}{3^{3/2}}$.

Example 2.4.17

Example 2.4.18

A tank contains 1500 liters of brine with a concentration of 0.3 kg of salt per liter. Another brine solution, this with a concentration of 0.1 kg of salt per liter is poured into the tank at a rate of 20 li/min. At the same time, 20 li/min of the solution in the tank, which is stirred continuously, is drained from the tank.

(a) How many kilograms of salt will remain in the tank after half an hour?

(b) How long will it take to reduce the concentration to 0.2 kg/li?

Solution. Denote by $Q(t)$ the amount of salt in the tank at time t. In a very short time interval $\mathrm{d}t$, the incoming solution adds $20\,\mathrm{d}t$ liters of a solution carrying 0.1 kg/li. So the incoming solution adds $0.1 \times 20\,\mathrm{d}t = 2\,\mathrm{d}t$ kg of salt. In the same time interval $20\,\mathrm{d}t$ liters is drained from the tank. The concentration of the drained brine is $\frac{Q(t)}{1500}$. So $\frac{Q(t)}{1500}20\,\mathrm{d}t$ kg were removed. All together, the change in the salt content of the tank during the short time interval is

$$dQ = 2\,\mathrm{d}t - \frac{Q(t)}{1500}20\,\mathrm{d}t = \Big(2 - \frac{Q(t)}{75}\Big)\mathrm{d}t$$

The rate of change of salt content per unit time is

$$\frac{dQ}{dt} = 2 - \frac{Q(t)}{75} = -\frac{1}{75}\big(Q(t) - 150\big)$$

The solution of this equation is

$$Q(t) = \big\{Q(0) - 150\big\}e^{-t/75} + 150$$

by Theorem 2.4.4, with $a = -\frac{1}{75}$ and $b = 150$. At time 0, $Q(0) = 1500 \times 0.3 = 450$. So

$$Q(t) = 150 + 300e^{-t/75}$$

(a) At $t = 30$

$$Q(30) = 150 + 300e^{-30/75} = 351.1 \text{ kg}$$

(b) $Q(t) = 0.2 \times 1500 = 300$ kg is achieved when

$$150 + 300e^{-t/75} = 300 \implies 300e^{-t/75} = 150 \implies e^{-t/75} = 0.5$$
$$\implies -\frac{t}{75} = \log(0.5) \implies t = -75\log(0.5) = 51.99 \text{ min}$$

Example 2.4.18

2.4.6 ▸▸ Optional — Interest on Investments

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 $\big(\frac{1}{n}\big)^{\text{th}}$ 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 $P + \frac{1}{n} \times \frac{r}{100} \times P = \big(1 + \frac{r}{100n}\big)P$.
- 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 $\big(1 + \frac{r}{100n}\big)P$ and interest is being paid for $\big(\frac{1}{n}\big)^{\text{th}}$ of a year, the second interest payment is $\frac{1}{n} \times \frac{r}{100} \times \big(1 + \frac{r}{100n}\big)P$. After the second interest payment, the balance in the account is $\big(1 + \frac{r}{100n}\big)P + \frac{1}{n} \times \frac{r}{100} \times \big(1 + \frac{r}{100n}\big)P = \big(1 + \frac{r}{100n}\big)^2 P$.

- And so on.

In general, at time $t = \frac{m}{n}$ (just after the m^{th} interest payment), the balance in the account is

$$B(t) = \left(1 + \frac{r}{100n}\right)^m P = \left(1 + \frac{r}{100n}\right)^{nt} P \tag{2.4.7}$$

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 $n \to \infty$ is called continuous compounding[29]. Under continuous compounding, the balance at time t is

$$B(t) = \lim_{n\to\infty} \left(1 + \frac{r}{100n}\right)^{nt} P$$

You may have already seen the limit

$$\lim_{x\to 0}(1+x)^{a/x} = e^a \tag{2.4.8}$$

If so, you can evaluate $B(t)$ by applying (2.4.8) with $x = \frac{r}{100n}$ and $a = \frac{rt}{100}$ (so that $\frac{a}{x} = nt$). As $n \to \infty$, $x \to 0$ so that

$$B(t) = \lim_{n\to\infty} \left(1 + \frac{r}{100n}\right)^{nt} P = \lim_{x\to 0}(1+x)^{a/x} P = e^a P = e^{rt/100} P \tag{2.4.9}$$

If you haven't seen (2.4.8) before, that's OK. In the following example, we rederive (2.4.9) using a differential equation instead of (2.4.8).

Example 2.4.19

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 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 $\frac{1}{n} = h$,

$$B(t+h) = B(t) + \frac{r}{100} B(t)h \qquad \text{or} \qquad \frac{B(t+h) - B(t)}{h} = \frac{r}{100} B(t)$$

To get continuous compounding we take the limit $n \to \infty$ or, equivalently, $h \to 0$. This gives

$$\lim_{h\to 0} \frac{B(t+h) - B(t)}{h} = \frac{r}{100} B(t) \qquad \text{or} \qquad \frac{\mathrm{d}B}{\mathrm{d}t}(t) = \frac{r}{100} B(t)$$

By Theorem 2.4.4, with $a = \frac{r}{100}$ and $b = 0$, (or Corollary 2.4.9 with $k = -\frac{r}{100}$),

$$B(t) = e^{rt/100} B(0) = e^{rt/100} P$$

once again.

Example 2.4.19

29 There are banks that advertise continuous compounding. You can find some by googling "interest is compounded continuously and paid"

Example 2.4.20

(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 (2.4.9). After 10 years you have $Pe^{r/10}$. This is supposed to be $2P$, so

$$Pe^{r/10} = 2P \implies e^{r/10} = 2 \implies \frac{r}{10} = \log 2 \implies r = 10\log 2 = 6.93\%$$

(b) Let the interest rate be r% per year. If you start with \$$P$, then after t years, you have $P\big(1+\frac{r}{100\times 12}\big)^{12t}$, under monthly compounding. This was (2.4.7). After 10 years you have $P\big(1+\frac{r}{100\times 12}\big)^{120}$. This is supposed to be $2P$, so

$$P\big(1+\frac{r}{100\times 12}\big)^{120} = 2P \implies \big(1+\frac{r}{1200}\big)^{120} = 2 \implies 1+\frac{r}{1200} = 2^{1/120}$$
$$\implies \frac{r}{1200} = 2^{1/120}-1 \implies r = 1200\big(2^{1/120}-1\big) = 6.95\%$$

Example 2.4.20

Example 2.4.21

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 $\Delta t \to 0$, the approximation becomes exact and we get

$$\frac{\mathrm{d}A}{\mathrm{d}t} = 0.05A + 2000$$

(b) The amount of money at time t obeys

$$\frac{\mathrm{d}A}{\mathrm{d}t} = 0.05A(t) + 2{,}000 = 0.05\big(A(t) + 40{,}000\big)$$

So by Theorem 2.4.4 (with $a = 0.05$ and $b = -40{,}000$),

$$A(t) = \big(A(0) + 40{,}000\big)e^{0.05t} - 40{,}000$$

At time 0 (when the graduate is 25), $A(0) = 50{,}000$, so the amount of money at time t is

$$A(t) = 90{,}000\, e^{0.05t} - 40,000$$

In particular, when the graduate is 65 years old, $t = 40$ and

$$A(40) = 90{,}000\, e^{0.05\times 40} - 40,000 = \$625{,}015.05$$

(c) When the graduate stops depositing money and instead starts withdrawing money at a rate W, the equation for A becomes

$$\frac{\mathrm{d}A}{\mathrm{d}t} = 0.05A - W = 0.05(A - 20W)$$

assuming that the interest rate remains 5%. This time, Theorem 2.4.4 (with $a = 0.05$ and $b = 20W$) gives

$$A(t) = \big(A(0) - 20W\big)e^{0.05t} + 20W$$

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 2.4.21

Chapter 3

SEQUENCES AND SERIES

You have probably learned about Taylor polynomials[1] 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 to 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 $\lim_{n\to\infty} E_n(x) = 0$ 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

1 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	$1 + 1 + \frac{1}{2} + \frac{1}{6} = 2.666666\ldots$
5 terms	$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} = 2.708333\ldots$
6 terms	$1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} = 2.716666\ldots$

By looking at the infinite sum in this way, we naturally obtain a sequence of numbers

$$\{\, 1\,,\, 2\,,\, 2.5\,,\, 2.666666\,,\ldots,\, 2.708333\,,\ldots,\, 2.716666\,,\ldots,\, \cdots\,\}.$$

The key to understanding the original infinite sum is to understand the behaviour of this sequence of numbers — in particularly, what do the numbers do as we go further and further? Does it settle down [2] to a given limit?

3.1▲ Sequences

In the discussion above we used the term "sequence" without giving it a precise mathematical meaning. Let us rectify this now.

Definition 3.1.1.

A sequence is a list of infinitely[3] many numbers with a specified order. It is denoted

$$\{a_1,\, a_2,\, a_3,\, \cdots,\, a_n,\, \cdots\} \quad \text{or} \quad \{a_n\} \quad \text{or} \quad \{a_n\}_{n=1}^{\infty}$$

We will often specify a sequence by writing it more explicitly, like

$$\Big\{a_n = f(n)\Big\}_{n=1}^{\infty}$$

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 countably infinite list of numbers. The interested (pedantic or otherwise) reader should look up countable and uncountable sets.

Example 3.1.2

Here are three sequences.

$$\left\{1,\ \frac{1}{2},\ \frac{1}{3},\ \cdots,\ \frac{1}{n},\ \cdots\right\} \qquad \text{or} \qquad \left\{a_n = \frac{1}{n}\right\}_{n=1}^{\infty}$$

$$\left\{1,\ 2,\ 3,\ \cdots,\ n,\ \cdots\right\} \qquad \text{or} \qquad \left\{a_n = n\right\}_{n=1}^{\infty}$$

$$\left\{1,\ -1,\ 1,\ -1,\ \cdots,\ (-1)^{n-1},\ \cdots\right\} \qquad \text{or} \qquad \left\{a_n = (-1)^{n-1}\right\}_{n=1}^{\infty}$$

It is not necessary that there be a simple explicit formula for the n^{th} term of a sequence. For example the decimal digits of π is a perfectly good sequence

$$\{3,\ 1,\ 4,\ 1,\ 5,\ 9,\ 2,\ 6,\ 5,\ 3,\ 5,\ 8,\ 9,\ 7,\ 9,\ 3,\ 2,\ 3,\ 8,\ 4,\ 6,\ 2,\ 6,\ 4,\ 3,\ 3,\ 8,\ \cdots\ \}$$

but there is no simple formula[4] for the n^{th} digit.

Example 3.1.2

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 3.1.3.

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

$$\lim_{n\to\infty} a_n = A \qquad \text{or} \qquad a_n \to A \text{ as } n \to \infty$$

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

$$\lim_{x\to\infty} f(x) = L \qquad \text{if} \qquad f(x) \to L \text{ as } x \to \infty$$

Example 3.1.4

Three of the four sequences in Example 3.1.2 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^{th} binary digit of π (i.e. writing π 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 π also diverges, though the proof that this is the case is a bit beyond us right now[5].

The other sequence in Example 3.1.2 has $a_n = \frac{1}{n}$. As n tends to infinity, $\frac{1}{n}$ tends to zero. So

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

Example 3.1.4

Example 3.1.5 $\left(\lim_{n\to\infty} \frac{n}{2n+1}\right)$

Here is a little less trivial example. To study the behaviour of $\frac{n}{2n+1}$ as $n \to \infty$, it is a good idea to write it as

$$\frac{n}{2n+1} = \frac{1}{2+\frac{1}{n}}$$

As $n \to \infty$, 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

$$\lim_{n\to\infty} \frac{n}{2n+1} = \lim_{n\to\infty} \frac{1}{2+\frac{1}{n}} = \frac{1}{2}$$

Example 3.1.5

Notice that in this last example, we are really using techniques that we used before to study infinite limits like $\lim_{x\to\infty} f(x)$. This experience can be easily transferred to dealing with $\lim_{n\to\infty} a_n$ limits by using the following result.

Theorem 3.1.6.

If

$$\lim_{x\to\infty} f(x) = L$$

and if $a_n = f(n)$ for all positive integers n, then

$$\lim_{n\to\infty} a_n = L$$

5 If the digits of π were to converge, then π would have to be a rational number. The irrationality of π (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 3.1.7 $\left(\lim_{n\to\infty} e^{-n}\right)$

Set $f(x) = e^{-x}$. Then $e^{-n} = f(n)$ and

$$\text{since } \lim_{x\to\infty} e^{-x} = 0 \qquad \text{we know that} \qquad \lim_{n\to\infty} e^{-n} = 0$$

Example 3.1.7

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 $\lim_{x\to\infty} f(x)$ also apply to limits like $\lim_{n\to\infty} a_n$.

Theorem 3.1.8 (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

$$\lim_{n\to\infty} a_n = A \qquad \lim_{n\to\infty} b_n = B$$

Then the following limits hold.

(a) $\lim_{n\to\infty} [a_n + b_n] = A + B$
(The limit of the sum is the sum of the limits.)

(b) $\lim_{n\to\infty} [a_n - b_n] = A - B$
(The limit of the difference is the difference of the limits.)

(c) $\lim_{n\to\infty} C a_n = CA.$

(d) $\lim_{n\to\infty} a_n\, b_n = A\, B$
(The limit of the product is the product of the limits.)

(e) If $B \neq 0$ then $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{A}{B}$
(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 3.1.9

Combining Examples 3.1.5 and 3.1.7,

$$\begin{aligned}
\lim_{n\to\infty}\left[\frac{n}{2n+1}+7e^{-n}\right] &= \lim_{n\to\infty}\frac{n}{2n+1}+\lim_{n\to\infty}7e^{-n} && \text{by Theorem 3.1.8.a}\\
&= \lim_{n\to\infty}\frac{n}{2n+1}+7\lim_{n\to\infty}e^{-n} && \text{by Theorem 3.1.8.c}\\
&= \frac{1}{2}+7\cdot 0 && \text{by Examples 3.1.5 and 3.1.7}\\
&= \frac{1}{2}
\end{aligned}$$

Example 3.1.9

There is also a squeeze theorem for sequences.

Theorem 3.1.10 (Squeeze theorem).

If $a_n \leqslant c_n \leqslant b_n$ for all natural numbers n, and if

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = L$$

then

$$\lim_{n\to\infty} c_n = L$$

Example 3.1.11

In this example we use the squeeze theorem to evaluate

$$\lim_{n\to\infty}\left[1+\frac{\pi_n}{n}\right]$$

where π_n is the n^{th} decimal digit of π. That is,

$$\pi_1 = 3 \quad \pi_2 = 1 \quad \pi_3 = 4 \quad \pi_4 = 1 \quad \pi_5 = 5 \quad \pi_6 = 9 \quad \cdots$$

We do not have a simple formula for π_n. But we do know that

$$0 \leqslant \pi_n \leqslant 9 \implies 0 \leqslant \frac{\pi_n}{n} \leqslant \frac{9}{n} \implies 1 \leqslant 1+\frac{\pi_n}{n} \leqslant 1+\frac{9}{n}$$

and we also know that

$$\lim_{n\to\infty} 1 = 1 \qquad \lim_{n\to\infty}\left[1+\frac{9}{n}\right] = 1$$

So the squeeze theorem with $a_n = 1$, $c_n = 1+\frac{\pi_n}{n}$, and $b_n = 1+\frac{9}{n}$ gives

$$\lim_{n\to\infty}\left[1+\frac{\pi_n}{n}\right] = 1$$

Example 3.1.11

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 3.1.12 (Continuous functions of limits).

If $\lim_{n\to\infty} a_n = L$ and if the function $g(x)$ is continuous at L, then

$$\lim_{n\to\infty} g(a_n) = g(L)$$

Example 3.1.13 $\left(\lim_{n\to\infty} \sin\frac{\pi n}{2n+1}\right)$

Write $\sin\frac{\pi n}{2n+1} = g\left(\frac{n}{2n+1}\right)$ with $g(x) = \sin(\pi x)$. We saw, in Example 3.1.5 that

$$\lim_{n\to\infty} \frac{n}{2n+1} = \frac{1}{2}$$

Since $g(x) = \sin(\pi x)$ is continuous at $x = \frac{1}{2}$, which is the limit of $\frac{n}{2n+1}$, we have

$$\lim_{n\to\infty} \sin\frac{\pi n}{2n+1} = \lim_{n\to\infty} g\left(\frac{n}{2n+1}\right) = g\left(\frac{1}{2}\right) = \sin\frac{\pi}{2} = 1$$

Example 3.1.13

With this introduction to sequences and some tools to determine their limits, we can now return to the problem of understanding infinite sums.

3.2▲ Series

A series is a sum

$$a_1 + a_2 + a_3 + \cdots + a_n + \cdots$$

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 $\frac{1}{3}$:

$$\frac{1}{3} = 0.3333\cdots$$

Recall that the expansion written in this way actually means

$$0.333333\cdots = \frac{3}{10} + \frac{3}{100} + \frac{3}{1000} + \frac{3}{10000} + \cdots = \sum_{n=1}^{\infty} \frac{3}{10^n}$$

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:

$$\sum_{n=1}^{\infty} \frac{3}{10^n} = \overbrace{\frac{3}{10}}^{n=1} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \cdots$$

$$\sum_{j=2}^{\infty} \frac{3}{10^{j-1}} = \overbrace{\frac{3}{10}}^{j=2} + \overbrace{\frac{3}{100}}^{j=3} + \overbrace{\frac{3}{1000}}^{j=4} + \cdots$$

$$\sum_{\ell=0}^{\infty} \frac{3}{10^{\ell+1}} = \overbrace{\frac{3}{10}}^{\ell=0} + \overbrace{\frac{3}{100}}^{\ell=1} + \overbrace{\frac{3}{1000}}^{\ell=3} + \cdots$$

$$\frac{3}{10} + \sum_{n=2}^{\infty} \frac{3}{10^n} = \frac{3}{10} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \cdots$$

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 3.2.1 $\left(\sum_{n=1}^{\infty} \frac{3}{10^n}\right)$

As we have just seen above the series $\sum_{n=1}^{\infty} \frac{3}{10^n}$ is

$$\sum_{n=1}^{\infty} \frac{3}{10^n} = \overbrace{\frac{3}{10}}^{n=1} + \overbrace{\frac{3}{100}}^{n=2} + \overbrace{\frac{3}{1000}}^{n=3} + \cdots$$

Notice that the n^{th} term in that sum is

$$3 \times 10^{-n} = 0.\overbrace{00\cdots0}^{n-1 \text{ zeroes}}3$$

So the sum of the first 5, 10, 15 and 20 terms in that series are

$$\sum_{n=1}^{5} \frac{3}{10^n} = 0.33333 \qquad \sum_{n=1}^{10} \frac{3}{10^n} = 0.3333333333$$
$$\sum_{n=1}^{15} \frac{3}{10^n} = 0.333333333333333 \qquad \sum_{n=1}^{20} \frac{3}{10^n} = 0.33333333333333333333$$

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[6] to define $\sum_{n=1}^{\infty} \frac{3}{10^n}$ to be $\frac{1}{3}$.

Example 3.2.1

Example 3.2.2 $\left(\sum_{n=1}^{\infty} 1 \text{ and } \sum_{n=1}^{\infty}(-1)^n\right)$

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

$$\sum_{n=1}^{\infty}(-1)^n = \overbrace{(-1)}^{n=1} + \overbrace{1}^{n=2} + \overbrace{(-1)}^{n=3} + \overbrace{1}^{n=4} + \overbrace{(-1)}^{n=5} + \cdots$$

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 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 3.2.2

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.

6 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.

Definition 3.2.3.

The N^{th} partial sum of the series $\sum_{n=1}^{\infty} a_n$ is the sum of its first N terms

$$S_N = \sum_{n=1}^{N} a_n.$$

Conversely, for each $N \geqslant 2$, $a_N = S_N - S_{N-1}$.
The partial sums form a sequence $\{S_N\}_{N=1}^{\infty}$. If this sequence of partial sums converges $S_N \to S$ as $N \to \infty$ 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.

In the last definition, we introduced the notation S_N for the N^{th} partial sum of a series $\sum_{n=1}^{\infty} a_n$ that has the smallest value of the index n being 1. It is sometimes convenient, as we shall see in the next example, to express a series with the index having a smallest value other than one, like for example $\sum_{n=0}^{\infty} a_n$. Then there are a couple of different commonly used definitions for the partial sum notation S_N. One is $S_N = \sum_{n=0}^{N} a_n$, which is the sum of all terms having indices less than or equal to N. With this definition, we still have $a_N = S_N - S_{N-1}$. A second commonly used definition is the sum of the first N terms in the series, i.e. $S_N = \sum_{n=0}^{N-1} a_n$. Under this definition $a_N = S_{N+1} - S_N$. You may of course use whatever definition you like. But you should state clearly what definition you are using.

Example 3.2.4 (Geometric Series)

Let a and r be any two fixed real numbers with $a \neq 0$. The series

$$a + ar + ar^2 + \cdots + ar^n + \cdots = \sum_{n=0}^{\infty} ar^n$$

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[7] term is the $n = 0$ term, which is $ar^0 = a$. The second term is the $n = 1$ term,

7 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:

$$\mathbb{N} = \{0, 1, 2, \dots\}$$

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.

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 $ar^{n-1}\big|_{n=1} = ar^{1-1} = a$, the second term is $ar^{n-1}\big|_{n=2} = ar^{2-1} = ar$, and so on[8]. 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

$$S_N = \sum_{n=0}^{N} ar^n = a + ar + ar^2 + \cdots + ar^N.$$

The secret to evaluating this sum is to see what happens when we multiply it by r:

$$\begin{aligned} rS_N &= r\left(a + ar + ar^2 + \cdots + ar^N\right) \\ &= ar + ar^2 + ar^3 + \cdots + ar^{N+1} \end{aligned}$$

Notice that this is almost the same[9] 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

$$\begin{aligned} S_N &= a + ar + ar^2 + \cdots + ar^N \\ rS_N &= \quad ar + ar^2 + \cdots + ar^N + ar^{N+1} \end{aligned}$$

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

$$S_N = \underbrace{a + a + \cdots + a}_{N+1 \text{ terms}} = a(N+1)$$

So in summary:

$$S_N = \begin{cases} a\frac{1-r^{N+1}}{1-r} & \text{if } r \neq 1 \\ a(N+1) & \text{if } r = 1 \end{cases} \tag{3.2.1}$$

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 $N \to \infty$, so that S_N converges to $\frac{a}{1-r}$ as $N \to \infty$ and

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r} \text{ provided } |r| < 1. \tag{3.2.2}$$

8 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.

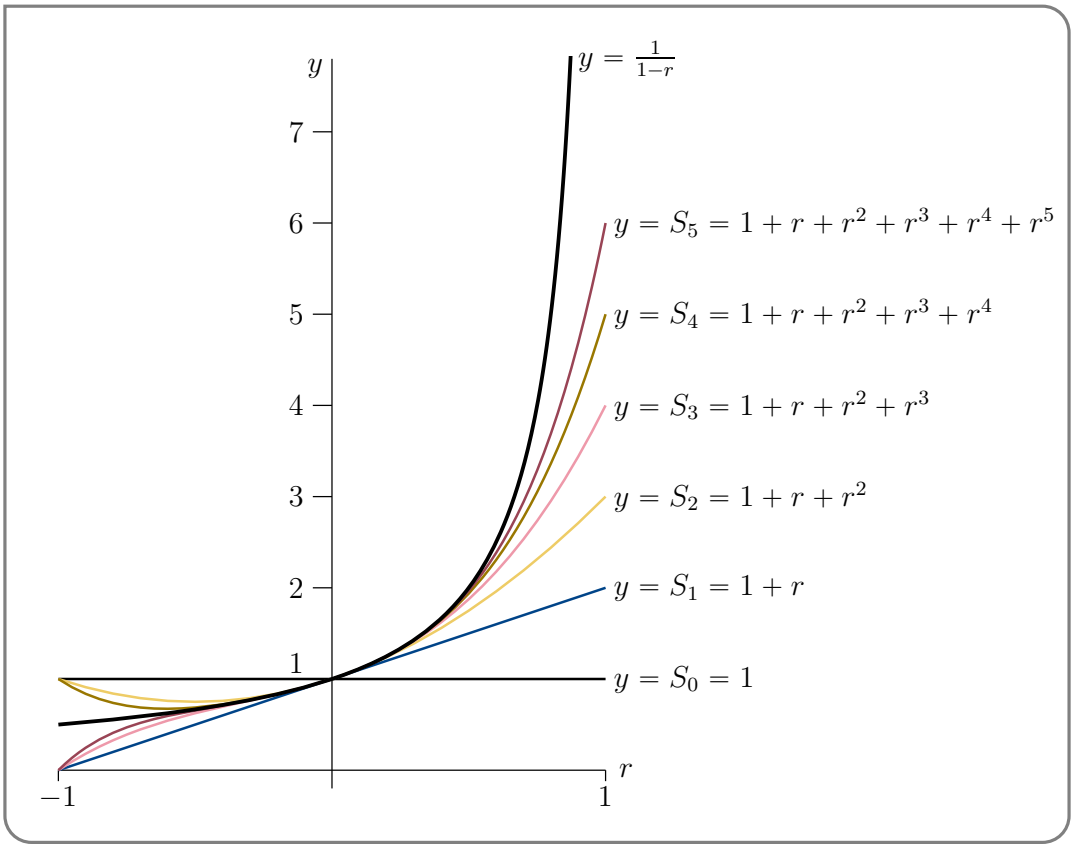
9 One can find similar properties of other special series, that allow us, with some work, to cancel many terms in the partial sums. We will shortly see a good example of this. The interested reader should look up "creative telescoping" to see how this idea might be used more generally, though it is somewhat beyond this course.

On the other hand if $|r| \geqslant 1$, S_N diverges. To understand this divergence, consider the following 4 cases:

- If $r > 1$, then r^N grows to ∞ as $N \to \infty$.
- If $r < -1$, then the magnitude of r^N grows to ∞, and the sign of r^N oscillates between $+$ and $-$, as $N \to \infty$.
- If $r = +1$, then $N + 1$ grows to ∞ as $N \to \infty$.
- If $r = -1$, then r^N just oscillates between $+1$ and -1 as $N \to \infty$.

In each case the sequence of partial sums does not converge and so the series does not converge.

Here are some sketches of the graphs of $\frac{1}{1-r}$ and S_N, $0 \leqslant N \leqslant 5$, for $a = 1$ and $-1 \leqslant r < 1$.



In these sketches we see that

- when $0 < r < 1$, and also when $-1 < r < 0$ with N odd, we have $S_N = \frac{1-r^{N+1}}{1-r} < \frac{1}{1-r}$. On the other hand, when $-1 < r < 0$ with N even, we have $S_N = \frac{1-r^{N+1}}{1-r} > \frac{1}{1-r}$.
- When $0 < |r| < 1$, $S_N = \frac{1-r^{N+1}}{1-r}$ gets closer and closer to $\frac{1}{1-r}$ as N increases.
- When $r = -1$, S_N just alternates between 0, when N is odd, and 1, when N is even.

Example 3.2.4

Now that we know how to handle geometric series let's return to Example 3.2.1.

Example 3.2.5 (Decimal Expansions)

The decimal expansion

$$0.3333\cdots = \frac{3}{10} + \frac{3}{100} + \frac{3}{1000} + \frac{3}{10000} + \cdots = \sum_{n=1}^{\infty} \frac{3}{10^n}$$

is a geometric series with the first term $a = \frac{3}{10}$ and the ratio $r = \frac{1}{10}$. So, by Example 3.2.4,

$$0.3333\cdots = \sum_{n=1}^{\infty} \frac{3}{10^n} = \frac{3/10}{1 - 1/10} = \frac{3/10}{9/10} = \frac{1}{3}$$

just as we would have expected.

We can push this idea further. Consider the repeating decimal expansion:

$$0.16161616\cdots = \frac{16}{100} + \frac{16}{10000} + \frac{16}{1000000} + \cdots$$

This is another geometric series with the first term $a = \frac{16}{100}$ and the ratio $r = \frac{1}{100}$. So, by Example 3.2.4,

$$0.16161616\cdots = \sum_{n=1}^{\infty} \frac{16}{100^n} = \frac{16/100}{1 - 1/100} = \frac{16/100}{99/100} = \frac{16}{99}$$

again, as expected. In this way any periodic decimal expansion converges to a ratio of two integers — that is, to a rational number[10].

Here is another more complicated example.

$$\begin{aligned}
0.1234343434\cdots &= \frac{12}{100} + \frac{34}{10000} + \frac{34}{1000000} + \cdots \\
&= \frac{12}{100} + \sum_{n=2}^{\infty} \frac{34}{100^n} \\
&= \frac{12}{100} + \frac{34}{10000}\frac{1}{1 - 1/100} \quad \text{by Example 3.2.4 with } a = \frac{34}{100^2} \text{ and } r = \frac{1}{100} \\
&= \frac{12}{100} + \frac{34}{10000}\frac{100}{99} \\
&= \frac{1222}{9900}
\end{aligned}$$

Example 3.2.5

10 We have included a (more) formal proof of this fact in the optional §3.7 at the end of this chapter. Proving that a repeating decimal expansion gives a rational number isn't too hard. Proving the converse — that every rational number has a repeating decimal expansion is a little trickier, but we also do that in the same optional section.

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 3.2.6 (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[11] that has been rigged to illustrate a phenomenon called "telescoping". Notice that the n^{th} term can be rewritten as

$$\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$$

and so we have

$$a_n = b_n - b_{n+1} \qquad \text{where } b_n = \frac{1}{n}.$$

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.

$$\begin{aligned} S_N &= \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \cdots + \frac{1}{N\cdot(N+1)} \\ &= \left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{N} - \frac{1}{N+1}\right) \end{aligned}$$

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

$$\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_{N\to\infty} S_N = \lim_{N\to\infty} \left(1 - \frac{1}{N+1}\right) = 1$$

Example 3.2.6

More generally, if we can write

$$a_n = b_n - b_{n+1}$$

11 Well... this sort of series does show up when you start to look at the Maclaurin polynomial of functions like $(1-x)\log(1-x)$. 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.

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

$$\sum_{n=1}^{\infty} a_n = b_1 - \lim_{N\to\infty} b_{N+1}$$

provided this limit exists. Often $\lim_{N\to\infty} b_{N+1} = 0$ and then $\sum_{n=1}^{\infty} a_n = b_1$. But this does not always happen. Here is an example.

Example 3.2.7 (A Divergent Telescoping Series)

In this example, we are going to study the series $\sum_{n=1}^{\infty} \log\left(1+\frac{1}{n}\right)$. Let's start by just writing out the first few terms.

$$\begin{aligned}\sum_{n=1}^{\infty} \log\left(1+\frac{1}{n}\right) &= \overbrace{\log\left(1+\frac{1}{1}\right)}^{n=1} + \overbrace{\log\left(1+\frac{1}{2}\right)}^{n=2} + \overbrace{\log\left(1+\frac{1}{3}\right)}^{n=3} + \overbrace{\log\left(1+\frac{1}{4}\right)}^{n=4} + \cdots \\ &= \log(2) \quad + \log\left(\frac{3}{2}\right) \quad + \log\left(\frac{4}{3}\right) \quad + \log\left(\frac{5}{4}\right) + \cdots\end{aligned}$$

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:

$$\begin{aligned}S_N &= \sum_{n=1}^{N} \log\left(1+\frac{1}{n}\right) \\ &= \overbrace{\log\left(1+\frac{1}{1}\right)}^{n=1} + \overbrace{\log\left(1+\frac{1}{2}\right)}^{n=2} + \overbrace{\log\left(1+\frac{1}{3}\right)}^{n=3} + \cdots + \overbrace{\log\left(1+\frac{1}{N-1}\right)}^{n=N-1} + \overbrace{\log\left(1+\frac{1}{N}\right)}^{n=N} \\ &= \log(2) \quad + \log\left(\frac{3}{2}\right) \quad + \log\left(\frac{4}{3}\right) \quad + \cdots + \log\left(\frac{N}{N-1}\right) \quad + \log\left(\frac{N+1}{N}\right) \\ &= \log\left(2 \times \frac{3}{2} \times \frac{4}{3} \times \cdots \times \frac{N}{N-1} \times \frac{N+1}{N}\right) \\ &= \log(N+1)\end{aligned}$$

Uh oh!

$$\lim_{N\to\infty} S_N = \lim_{N\to\infty} \log(N+1) = +\infty$$

This telescoping series diverges! There is an important lesson here. Telescoping series *can diverge*. They do not always converge to b_1.

Example 3.2.7

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 3.2.8 (Arithmetic of series)**.**

Let C, S and T 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

$$\sum_{n=1}^{\infty} a_n = S \qquad\qquad \sum_{n=1}^{\infty} b_n = T$$

Then the following hold.

(a) $\displaystyle\sum_{n=1}^{\infty} \big[a_n + b_n\big] = S + T$ and $\displaystyle\sum_{n=1}^{\infty} \big[a_n - b_n\big] = S - T$

(b) $\displaystyle\sum_{n=1}^{\infty} Ca_n = CS.$

Example 3.2.9

As a simple example of how we use the arithmetic of series Theorem 3.2.8, 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 3.2.4) 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 3.2.6. We can now use Theorem 3.2.8 to build the specified "complicated" series out of these two "simple" pieces.

$$\begin{aligned}\sum_{n=1}^{\infty}\Big[\frac{1}{7^n}+\frac{2}{n(n+1)}\Big] &= \sum_{n=1}^{\infty}\frac{1}{7^n}+\sum_{n=1}^{\infty}\frac{2}{n(n+1)} && \text{by Theorem 3.2.8.a}\\ &= \sum_{n=1}^{\infty}\frac{1}{7^n}+2\sum_{n=1}^{\infty}\frac{1}{n(n+1)} && \text{by Theorem 3.2.8.b}\\ &= \frac{1}{6}+2\cdot 1=\frac{13}{6}\end{aligned}$$

Example 3.2.9

Example 3.2.10 (Arithmetic Progressions)

An arithmetic progression is a sequence for which the difference between successive terms is a constant. That is, there are constants a_1 and d such that the n^{th} term in the sequence is $a_n = a_1 + (n-1)d$. For example, the sequence $1, 2, 3, 4, \cdots$ has $a_1 = 1$, $d = 1$ and $a_n = n$. We will now compute the partial sum $S_N = \sum_{n=1}^{N} n$ of the corresponding series $\sum_{n=1}^{\infty} n$. To do so, we will use a trick that is often attributed to Carl Friedrich Gauss[12]. The trick consists of writing S_N twice, once in forward order and once in reverse order

$$\begin{aligned}S_N &= 1 + \quad 2 \quad + \quad 3 \quad + \cdots + N\\ S_N &= N + (N-1) + (N-2) + \cdots + 1\end{aligned}$$

and then adding the two representations of S_N together.

$$\begin{aligned}2S_N &= \overbrace{(N+1)+(N+1)+(N+1)+\cdots(N+1)}^{N \text{ terms}}\\ &= N(N+1)\\ \implies S_N &= \frac{N(N+1)}{2}\end{aligned}$$

Using this and a little arithmetic, we can compute the partial sums for the series that corresponds to the general arithmetic progression $a_n = a_1 + (n-1)d$. (Of course, you can get the same result by applying the above trick to $S_N = \sum_{n=1}^{N}\{a_1 + (n-1)d\}$.)

$$\begin{aligned}S_N &= \sum_{n=1}^{N}\{a_1+(n-1)d\} = \sum_{n=1}^{N}\{(a_1-d)+nd\}\\ &= (a_1-d)\sum_{n=1}^{N}1 + d\sum_{n=1}^{N}n = (a_1-d)N + d\tfrac{N(N+1)}{2}\end{aligned}$$

12 Carl Friedrich Gauss (1777–1855) was a German mathematician and physicist who ranks among the world's most influential mathematicians. He is said to have reinvented this trick when he was in primary school. In fact, the trick was known many centuries before him. You can use your favourite search engine to learn more.

$$= \tfrac{N}{2}\Big\{a_1 + [a_1 + d(N-1)]\Big\}$$
$$= \tfrac{N}{2}\{\text{first term} + \text{last term}\}$$

Example 3.2.10

Example 3.2.11 (Power Sequence)

We are now going to consider partial sums for the series $\sum_{n=1}^{\infty} n^k$, for several values of the parameter k. In the last example, we found the partial sum

$$\sum_{n=1}^{N} n = \tfrac{1}{2}N(N+1)$$

for the case $k = 1$. In Theorem 1.1.6 we stated the partial sums

$$\sum_{n=1}^{N} n^2 = \tfrac{1}{6}N(N+1)(2N+1)$$
$$\sum_{n=1}^{N} n^3 = \Big[\tfrac{1}{2}N(N+1)\Big]^2$$

for the cases $k = 2$ and $k = 3$. In that theorem, while we gave a strategy that could be used to prove those formulae, we did not actually give the proofs. We are now going to give the proofs, using a different strategy — a strategy that can be used to find the partial sums for any natural number k. As a warmup, we'll start with $k = 1$, even though we have already handled that case. The strategy is based on the observation that the telescoping sum

$$\sum_{n=1}^{N} \{(n+1)^2 - n^2\} = (N+1)^2 - 1 = N^2 + 2N$$

On the other hand, the same sum

$$\sum_{n=1}^{N} \{(n+1)^2 - n^2\} = \sum_{n=1}^{N} \{2n+1\} = 2\left[\sum_{n=1}^{N} n\right] + \left[\sum_{n=1}^{N} 1\right]$$
$$= 2\left[\sum_{n=1}^{N} n\right] + N$$

Equating the two formulae for $\sum_{n=1}^{N} \{(n+1)^2 - n^2\}$ gives

$$N^2 + 2N = 2\left[\sum_{n=1}^{N} n\right] + N$$
$$\implies \sum_{n=1}^{N} n = \tfrac{1}{2}\big[N^2 + N\big] = \tfrac{1}{2}N(N+1)$$

as we found in Example 3.2.10.

Now we move on to $k = 2$. This time we use the telescoping sum

$$\sum_{n=1}^{N} \left\{(n+1)^3 - n^3\right\} = (N+1)^3 - 1 = N^3 + 3N^2 + 3N$$

On the other hand, the same sum

$$\begin{aligned}\sum_{n=1}^{N} \left\{(n+1)^3 - n^3\right\} = \sum_{n=1}^{N} \left\{3n^2 + 3n + 1\right\} &= 3\left[\sum_{n=1}^{N} n^2\right] + 3\left[\sum_{n=1}^{N} n\right] + \left[\sum_{n=1}^{N} 1\right] \\ &= 3\left[\sum_{n=1}^{N} n^2\right] + \tfrac{3}{2}N(N+1) + N\end{aligned}$$

Equating the two formulae for $\sum_{n=1}^{N} \left\{(n+1)^3 - n^3\right\}$ gives

$$\begin{aligned}N^3 + 3N^2 + 3N &= 3\left[\sum_{n=1}^{N} n^2\right] + \tfrac{3}{2}N^2 + \tfrac{5}{2}N \\ \implies \sum_{n=1}^{N} n^2 &= \tfrac{1}{3}\left[N^3 + \tfrac{3}{2}N^2 + \tfrac{1}{2}N\right] = \tfrac{1}{6}N\left[2N^2 + 3N + 1\right] \\ &= \tfrac{1}{6}N(N+1)(2N+1)\end{aligned}$$

as stated in Theorem 1.1.6.

Finally, we get to $k = 3$. This time we use the telescoping sum

$$\sum_{n=1}^{N} \left\{(n+1)^4 - n^4\right\} = (N+1)^4 - 1$$

On the other hand, the same sum

$$\begin{aligned}\sum_{n=1}^{N} \left\{(n+1)^4 - n^4\right\} &= \sum_{n=1}^{N} \left\{4n^3 + 6n^2 + 4n + 1\right\} \\ &= 4\left[\sum_{n=1}^{N} n^3\right] + 6\left[\sum_{n=1}^{N} n^2\right] + 4\left[\sum_{n=1}^{N} n\right] + \left[\sum_{n=1}^{N} 1\right] \\ &= 4\left[\sum_{n=1}^{N} n^3\right] + N(N+1)(2N+1) + 2N(N+1) + N\end{aligned}$$

Equating the two formulae for $\sum_{n=1}^{N} \left\{(n+1)^4 - n^4\right\}$ gives

$$(N+1)^4 - 1 = 4\left[\sum_{n=1}^{N} n^3\right] + (N+1)(2N^2 + 3N) + N$$

$$\begin{aligned}\implies \sum_{n=1}^{N} n^3 &= \tfrac{1}{4}\Big[(N+1)^4 - 1 - (N+1)(2N^3+3N) - N\Big] \\ &= \tfrac{1}{4}(N+1)\Big[(N+1)^3 - (2N^2+3N) - 1\Big] \\ &= \tfrac{1}{4}(N+1)\Big[N^3+N^2\Big] \\ &= \tfrac{1}{4}N^2(N+1)^2\end{aligned}$$

as stated in Theorem 1.1.6.

Example 3.2.11

3.3▲ Convergence 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[13][14]. 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.

3.3.1 ▸▸ The Divergence Test

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 $N \to \infty$, we have $S_N \to S$ and, because $N-1 \to \infty$ too, we also have $S_{N-1} \to S$.

13 The authors should be a little more careful making such a blanket statement. While it is true that it is not wise to approximate a divergent series by taking N terms with N large, there are cases when one can get a very good approximation by taking N terms with N small! For example, the Taylor remainder theorem shows us that when the n^{th} derivative of a function $f(x)$ grows very quickly with n, Taylor polynomials of degree N, with N large, can give bad approximations of $f(x)$, while the Taylor polynomials of degree one or two can still provide very good approximations of $f(x)$ when x is very small. As an example of this, one of the triumphs of quantum electrodynamics, namely the computation of the anomalous magnetic moment of the electron, depends on precisely this. A number of important quantities were predicted using the first few terms of divergent power series. When those quantities were measured experimentally, the predictions turned out to be incredibly accurate.

14 The field of asymptotic analysis often makes use of the first few terms of divergent series to generate approximate solutions to problems; this, along with numerical computations, is one of the most important techniques in applied mathematics. Indeed, there is a whole wonderful book (which, unfortunately, is too advanced for most Calculus 2 students) devoted to playing with divergent series called, unsurprisingly, "Divergent Series" by G.H. Hardy. This is not to be confused with the "Divergent" series by V. Roth set in a post-apocalyptic dystopian Chicago. That latter series diverges quite dramatically from mathematical topics, while the former does not have a film adaptation (yet).

- So $a_N = S_N - S_{N-1} \to S - S = 0$.

This tells us that, if we already know that a given series $\sum a_n$ is convergent, then the n^{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[15] of the statement is a useful test for *divergence*.

Theorem 3.3.1 (Divergence Test).

If the sequence $\{a_n\}_{n=1}^{\infty}$ fails to converge to zero as $n \to \infty$, then the series $\sum_{n=1}^{\infty} a_n$ diverges.

Example 3.3.2

Let $a_n = \frac{n}{n+1}$. Then

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \frac{n}{n+1} = \lim_{n\to\infty} \frac{1}{1+1/n} = 1 \neq 0$$

So the series $\sum_{n=1}^{\infty} \frac{n}{n+1}$ diverges.

Example 3.3.2

Warning 3.3.3.

The divergence test is a "one way test". It tells us that if $\lim_{n\to\infty} a_n$ is nonzero, or fails to exist, then the series $\sum_{n=1}^{\infty} a_n$ diverges. But it tells us *absolutely nothing* when $\lim_{n\to\infty} a_n = 0$. In particular, it is perfectly possible for a series $\sum_{n=1}^{\infty} a_n$ to *diverge* even though $\lim_{n\to\infty} a_n = 0$. An example is $\sum_{n=1}^{\infty} \frac{1}{n}$. We'll show in Example 3.3.6, 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 — see the optional §3.3.9 —, 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.

15 We have discussed the contrapositive a few times in the CLP notes, but it doesn't hurt to discuss it again here (or for the reader to quickly look up the relevant footnote in Section 1.3 of the CLP-1 text). At any rate, 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 ($\sum a_n$ converges) then (a_n converges to 0).

The contrapositive of this statement is then

If (a_n does not converge to 0) then ($\sum a_n$ does not converge).

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

$$\sum_{n=1}^{\infty} \frac{1}{n^3}$$

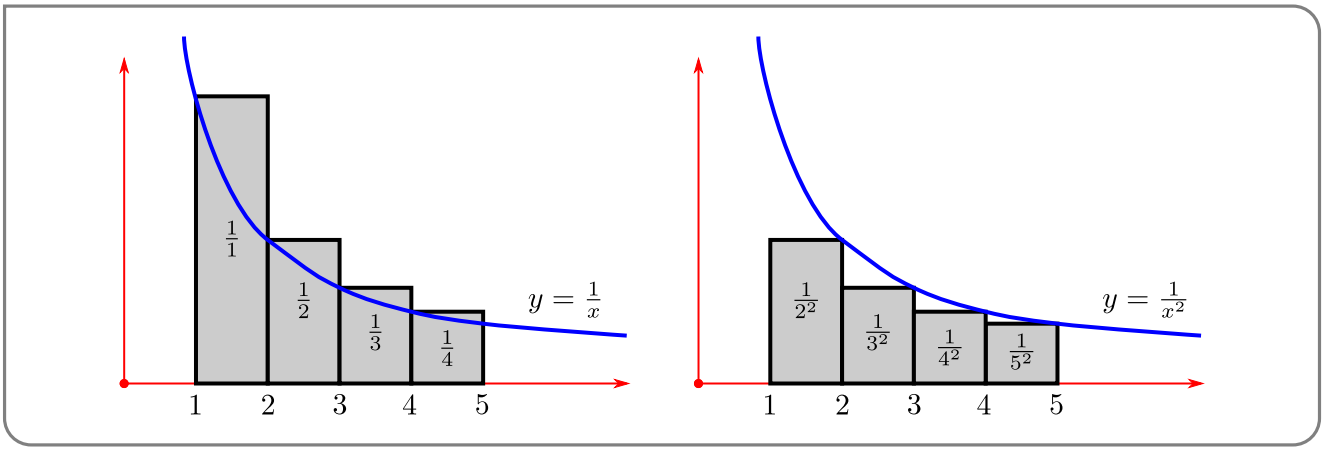
converges. However it does not tell us its value[16].

3.3.2 ▸▸ The Integral Test

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 1.12.17, 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 3.3.4

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.



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 $1 \leqslant x \leqslant 5$. That is

$$\sum_{n=1}^{4} \frac{1}{n} \geqslant \int_1^5 \frac{1}{x} \mathrm{d}x$$

16 This series converges to Apéry's constant 1.2020569031 .... 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.

If we were to continue drawing the columns all the way out to infinity, then we would have

$$\sum_{n=1}^{\infty} \frac{1}{n} \geqslant \int_1^{\infty} \frac{1}{x} \mathrm{d}x$$

We are able to compute this improper integral exactly:

$$\int_1^{\infty} \frac{1}{x} \mathrm{d}x = \lim_{R \to \infty} \Big[\log |x| \Big]_1^R = +\infty$$

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*[17] for the series

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

Indeed we see from the sketch on the right above that

$$\sum_{n=2}^{N} \frac{1}{n^2} \leqslant \int_1^{N} \frac{1}{x^2} \mathrm{d}x$$

and hence

$$\sum_{n=2}^{\infty} \frac{1}{n^2} \leqslant \int_1^{\infty} \frac{1}{x^2} \mathrm{d}x$$

This last improper integral is easy to evaluate:

$$\begin{aligned} \int_1^{\infty} \frac{1}{x^2} \mathrm{d}x &= \lim_{R \to \infty} \left[-\frac{1}{x} \right]_1^R \\ &= \lim_{R \to \infty} \left(\frac{1}{1} - \frac{1}{R} \right) = 1 \end{aligned}$$

Thus we know that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \sum_{n=2}^{\infty} \frac{1}{n^2} \leqslant 2.$$

and so the series must converge.

Example 3.3.4

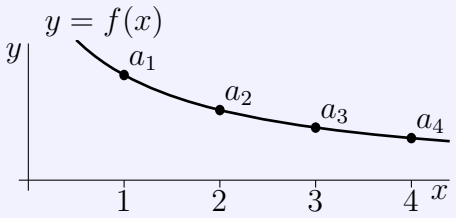
The above arguments are formalised in the following theorem.

17 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.*

Theorem 3.3.5 (The Integral Test).

Let N_0 be any natural number. If $f(x)$ is a function which is defined and continuous for all $x \geqslant N_0$ and which obeys

(i) $f(x) \geqslant 0$ for all $x \geqslant N_0$ and
(ii) $f(x)$ decreases as x increases and
(iii) $f(n) = a_n$ for all $n \geqslant N_0$.



Then

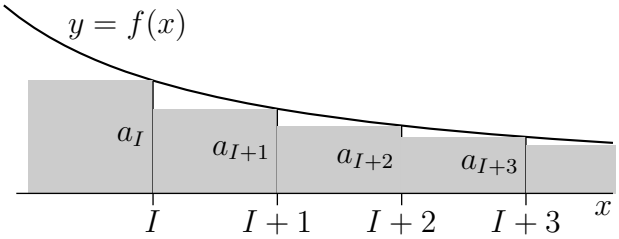
$$\sum_{n=1}^{\infty} a_n \text{ converges} \iff \int_{N_0}^{\infty} f(x)\,dx \text{ converges}$$

Furthermore, when the series converges, the truncation error

$$0 \leqslant \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{N} a_n \leqslant \int_{N}^{\infty} f(x)\,dx \qquad \text{for all } N \geqslant N_0$$

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 $a_n \geqslant 0$ for all $n \geqslant I > N_0$, the sequence of partial sums $s_\ell = \sum_{n=I}^{\ell} a_n$ obeys $s_{\ell+1} = s_\ell + a_{n+1} \geqslant s_\ell$. That is, s_ℓ increases as ℓ 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

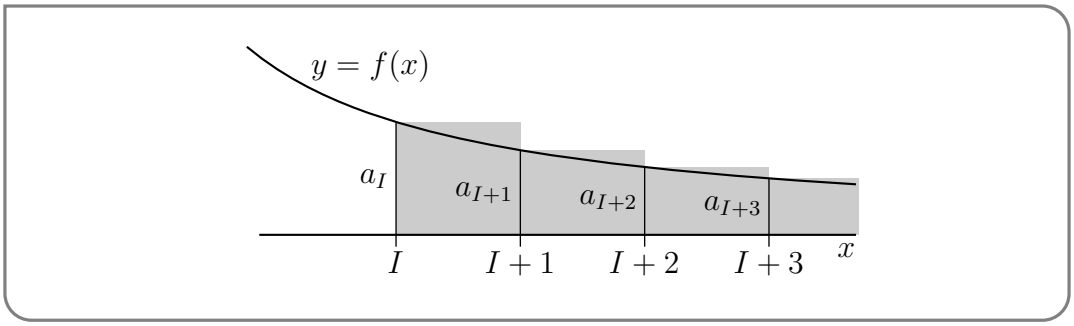
- the second shaded rectangle has height a_{I+1} and width 1, and hence area a_{I+1}, and so on

This shaded area is smaller than the area under the curve $y = f(x)$ for $I - 1 \leqslant x < \infty$. So

$$0 \leqslant \sum_{n=I}^{\infty} a_n \leqslant \int_{I-1}^{\infty} f(x)\,\mathrm{d}x$$

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 $I = N + 1$:

$$0 \leqslant \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{N} a_n = \sum_{n=N+1}^{\infty} a_n \leqslant \int_{N}^{\infty} f(x)\,\mathrm{d}x$$



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 $I \leqslant x < \infty$. So

$$\sum_{n=I}^{\infty} a_n \geqslant \int_{I}^{\infty} f(x)\,\mathrm{d}x$$

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[18]

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges.

Example 3.3.6 (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 $n \geqslant 1$. 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 3.3.5:
  (i) $f(x) > 0$ for all $x \geqslant N_0 = 1$ and
  (ii) $f(x)$ decreases as x increases because $f'(x) = -p\frac{1}{x^{p+1}} < 0$ for all $x \geqslant N_0 = 1$.
- 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 1.12.8, 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

18 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

$$\zeta(p) = \sum_{n=1}^{\infty} \frac{1}{n^p} = \prod_{\text{P prime}} \left(1 - \text{P}^{-p}\right)^{-1}$$

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.

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

$$\int_N^\infty \frac{\mathrm{d}x}{x^{1.000001}} = \lim_{R\to\infty} \int_N^R \frac{\mathrm{d}x}{x^{1.000001}} = \lim_{R\to\infty} -\frac{1}{0.000001}\left[\frac{1}{R^{0.000001}} - \frac{1}{N^{0.000001}}\right] = \frac{10^6}{N^{0.000001}}$$

This does tend to zero as $N \to \infty$, but really slowly.

Example 3.3.6

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^{th} term needs to shrink in order for the series to converge. We know that for large x, the function $\log x$ 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

$$\sum_{n=2}^{\infty} \frac{1}{n \log n}$$

converges. Notice that we sum from $n = 2$ because when $n = 1, n \log n = 0$. And we don't need to stop there[19]. We can analyse the convergence of this sum with any power of $\log n$.

Example 3.3.7 $\left(\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}\right)$

Let $p > 0$. We'll now use the integral test to determine whether or not the series $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}$ converges.

- As in the last example, we start by choosing a function that obeys $f(n) = a_n = \frac{1}{n(\log n)^p}$ for all n bigger than some N_0. Certainly $f(x) = \frac{1}{x(\log x)^p}$ obeys $f(n) = \frac{1}{n(\log n)^p}$ for all $n \geqslant 2$. So let's use that f and try $N_0 = 2$.
- Now let's check the other two conditions of Theorem 3.3.5:
  (i) Both x and $\log x$ are positive for all $x > 1$, so $f(x) > 0$ for all $x \geqslant N_0 = 2$.
  (ii) As x increases both x and $\log x$ increase and so $x(\log x)^p$ increases and $f(x)$ decreases.
- So the integral test tells us that the series $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}$ converges if and only if the integral $\int_2^\infty \frac{\mathrm{d}x}{x(\log x)^p}$ converges.

19 We could go even further and see what happens if we include powers of $\log(\log(n))$ and other more exotic slow growing functions.

- To test the convergence of the integral, we make the substitution $u = \log x$, $\mathrm{d}u = \frac{\mathrm{d}x}{x}$.

$$\int_2^R \frac{\mathrm{d}x}{x(\log x)^p} = \int_{\log 2}^{\log R} \frac{\mathrm{d}u}{u^p}$$

  We already know that the integral $\int_1^\infty \frac{\mathrm{d}u}{u^p}$, and hence the integral $\int_2^R \frac{\mathrm{d}x}{x(\log x)^p}$, converges if and only if $p > 1$.

So we conclude that $\sum_{n=2}^{\infty} \frac{1}{n(\log n)^p}$ converges if and only if $p > 1$.

Example 3.3.7

3.3.3 ▸▸ The Comparison Test

Our next convergence test is the comparison test. It is much like the comparison test for improper integrals (see Theorem 1.12.17) 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 3.3.8 (The Comparison Test).

Let N_0 be a natural number and let $K > 0$.

(a) If $|a_n| \leqslant Kc_n$ for all $n \geqslant N_0$ and $\sum_{n=0}^{\infty} c_n$ converges, then $\sum_{n=0}^{\infty} a_n$ converges.

(b) If $a_n \geqslant Kd_n \geqslant 0$ for all $n \geqslant N_0$ 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 optional section 3.3.10.

(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 ∞) 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 ∞, 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 3.3.6), for some $p > 0$, turns out to be just what is needed.

Example 3.3.9 $\left(\sum_{n=1}^{\infty} \frac{1}{n^2+2n+3}\right)$

We could determine whether or not the series $\sum_{n=1}^{\infty} \frac{1}{n^2+2n+3}$ converges by applying the integral test. But it is not worth the effort[20]. Whether or not any series converges is determined by the behaviour of the summand[21] 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[22] $n^2 \gg 2n \gg 3$ so that $\frac{1}{n^2+2n+3} \approx \frac{1}{n^2}$. We already know, from Example 3.3.6, 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 $n \geqslant 1$, $n^2 + 2n + 3 > n^2$, so that $\frac{1}{n^2+2n+3} \leqslant \frac{1}{n^2}$. So the comparison test, Theorem 3.3.8, 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.

Example 3.3.9

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 3.3.9, 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

20 Go back and quickly scan Theorem 3.3.5; to apply it we need to show that $\frac{1}{n^2+2n+3}$ is positive and decreasing (it is), and then we need to integrate $\int \frac{1}{x^2+2x+3} dx$. To do that we reread the notes on partial fractions, then rewrite $x^2 + 2x + 3 = (x+1)^2 + 2$ and so

$$\int_1^{\infty} \frac{1}{x^2+2x+3} dx = \int_1^{\infty} \frac{1}{(x+1)^2+2} dx \cdots$$

and then arctangent appears, etc etc. Urgh. Okay — let's go back to the text now and see how to avoid this.

21 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.

22 The symbol "$\gg$" means "much larger than". Similarly, the symbol "$\ll$" means "much less than". Good shorthand symbols can be quite expressive.

pretty rare that $a_n \leqslant b_n$ for all n. It is much more common that $a_n \leqslant Kb_n$ for some constant K. This is enough to allow application of the comparison test. Here is an example.

Example 3.3.10 $\left(\sum_{n=1}^{\infty} \frac{n+\cos n}{n^3-1/3}\right)$

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 3.3.6, 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[23] 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 $n \geqslant 1$.
  - First consider the numerator $1 + (\cos n)\frac{1}{n}$. For all $n \geqslant 1$
    - $\frac{1}{n} \leqslant 1$ and
    - $|\cos n| \leqslant 1$

    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 $n \geqslant 1$, $\frac{1}{3n^3}$ lies between $\frac{1}{3}$ and 0 so that
    - $1 - \frac{1}{3n^3}$ is between $\frac{2}{3}$ and 1 and consequently
    - $\frac{1}{1-1/3n^3}$ is between $\frac{3}{2}$ and 1.
  - 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
    $$\frac{1+\frac{\cos n}{n}}{1-\frac{1}{3n^3}} \leqslant 2\left(\frac{3}{2}\right) = 3$$

23 This is very similar to how we computed limits at infinity way way back near the beginning of CLP-1.

We now know that

$$|a_n| = \frac{1}{n^2}\,\frac{1+2/n}{1-1/3n^3} \leqslant \frac{3}{n^2}$$

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 3.3.10

The last example was actually a relatively simple application of the comparison theorem — finding a suitable constant K can be *really* tedious[24]. 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 3.3.11 (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

$$\lim_{n\to\infty} \frac{a_n}{b_n} = L$$

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 $\lim_{n\to\infty} \frac{a_n}{b_n} = L$, 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 $\left|\frac{a_n}{b_n}\right| \leqslant |L| + 1$, for all $n \geqslant N_0$, and hence
- $|a_n| \leqslant K b_n$ with $K = |L| + 1$, for all $n \geqslant N_0$.
- The comparison Theorem 3.3.8 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 $\lim_{n\to\infty} \frac{a_n}{b_n} = L$, we know that,

24 Really, really tedious. And you thought some of those partial fractions computations were bad ...

- when n is large, $\frac{a_n}{b_n}$ is very close to L.
- In particular, there is some natural number N so that $\frac{a_n}{b_n} \geqslant \frac{L}{2}$, and hence
- $a_n \geqslant Kb_n$ with $K = \frac{L}{2} > 0$, for all $n \geqslant N$.
- The comparison Theorem 3.3.8 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 3.3.12 $\left(\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{n^2-2n+3}\right)$

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 3.3.6 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

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{\frac{\sqrt{n+1}}{n^2-2n+3}}{\frac{1}{n^{3/2}}} = \lim_{n\to\infty} \frac{n^{3/2}\sqrt{n+1}}{n^2-2n+3}$$

Again it is a good idea to factor the dominant term out of the numerator and the dominant term out of the denominator.

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{n^2\sqrt{1+1/n}}{n^2\left(1-2/n+3/n^2\right)} = \lim_{n\to\infty} \frac{\sqrt{1+1/n}}{1-2/n+3/n^2} = 1$$

We already know that the series $\sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} \frac{1}{n^{3/2}}$ converges by Example 3.3.6 with $p = \frac{3}{2}$. So our series converges by the limit comparison test, Theorem 3.3.11.

Example 3.3.12

Example 3.3.13 $\left(\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{n^2-2n+3}$, again$\right)$

We can also try to deal with the series of Example 3.3.12, using the comparison test directly. But that requires us to find K so that

$$\frac{\sqrt{n+1}}{n^2-2n+3} \leqslant \frac{K}{n^{3/2}}$$

We might do this by examining the numerator and denominator separately:

- The numerator isn't too bad since for all $n \geqslant 1$:

$$\begin{aligned} n+1 &\leqslant 2n \qquad \text{and so} \\ \sqrt{n+1} &\leqslant \sqrt{2n} \end{aligned}$$

- The denominator is quite a bit more tricky, since we need a *lower* bound, rather than an upper bound, and we cannot just write $|n^2-2n+3| \geqslant n^2$, 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 $n^2-2n+3 \geqslant K'n^2$, i.e. $\frac{1}{n^2-2n+3} \leqslant \frac{1}{K'n^2}$ for all $n \geqslant N_0$. For $n \geqslant 4$, we have $2n = \frac{1}{2}4n \leqslant \frac{1}{2}n \cdot n = \frac{1}{2}n^2$. So for $n \geqslant 4$,

$$n^2-2n+3 \geqslant n^2 - \frac{1}{2}n^2 + 3 \geqslant \frac{1}{2}n^2$$

Putting the numerator and denominator back together we have

$$\frac{\sqrt{n+1}}{n^2-2n+3} \leqslant \frac{\sqrt{2n}}{n^2/2} = 2\sqrt{2}\frac{1}{n^{3/2}} \qquad \text{for all } n \geqslant 4$$

and the comparison test then tells us that our series converges. It is pretty clear that the approach of Example 3.3.12 was much more straightforward.

Example 3.3.13

3.3.4 ▸▸ The Alternating Series Test

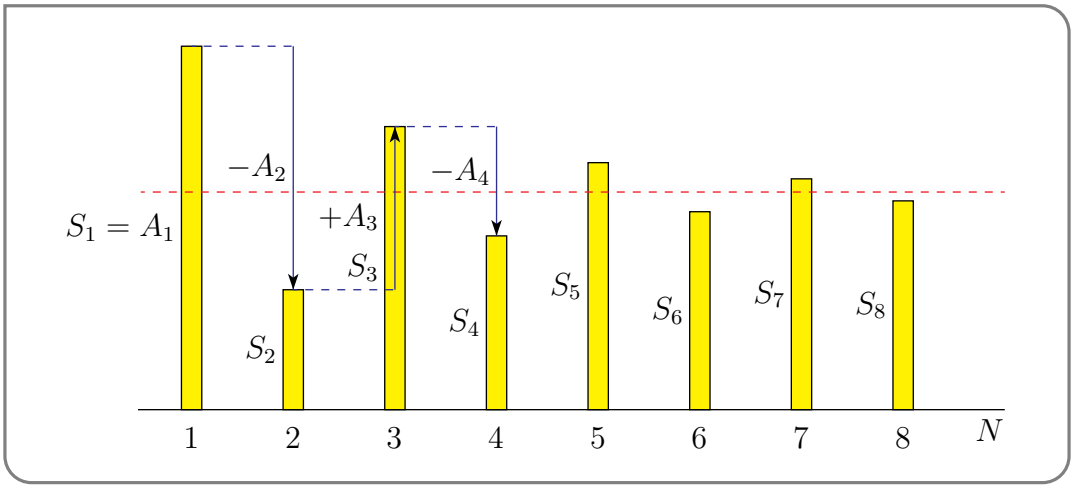
When the signs of successive terms in a series alternate between $+$ and $-$, like for example in $1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\cdots$, the series is called an *alternating series*. More generally, the series

$$A_1 - A_2 + A_3 - A_4 + \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} A_n$$

is alternating if every $A_n \geqslant 0$. Often (but not always) the terms in alternating series get successively smaller. That is, then $A_1 \geqslant A_2 \geqslant A_3 \geqslant \cdots$. 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 $A_3 \leqslant A_2$, 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 $A_4 \leqslant A_3$, 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 ∞, the amplitude of oscillation tends to zero and the sequence $S_1, S_2, S_3, \cdots$ converges to some limit S. This is illustrated in the figure



Here is a convergence test for alternating series that exploits this structure, and that is really easy to apply.

Theorem 3.3.14 (Alternating Series Test)**.**

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of real numbers that obeys

(i) $A_n \geqslant 0$ for all $n \geqslant 1$ and
(ii) $A_{n+1} \leqslant A_n$ for all $n \geqslant 1$ (i.e. the sequence is monotone decreasing) and
(iii) $\lim_{n\to\infty} A_n = 0$.

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^{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 optional section 3.3.10. 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)

$$E_N = S - S_N = \sum_{n=N+1}^{\infty} (-1)^{n-1} A_n = (-1)^N \Big[A_{N+1} - A_{N+2} + A_{N+3} - A_{N+4} + \cdots \Big]$$

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,

$$(-1)^N S_{N,\ell'} = \overbrace{(A_{N+1} - A_{N+2})}^{\geqslant 0} + \overbrace{(A_{N+3} - A_{N+4})}^{\geqslant 0} + \cdots + \overbrace{(A_{\ell'-1} - A_{\ell'})}^{\geqslant 0} \geqslant 0 \quad \text{and}$$

$$(-1)^N S_{N,\ell'+1} = \overbrace{(-1)^N S_{N,\ell'}}^{\geqslant 0} + \overbrace{A_{\ell'+1}}^{\geqslant 0} \geqslant 0$$

  This tells us that $(-1)^N S_{N,\ell} \geqslant 0$ for all $\ell > N+1$, both even and odd.

- Similarly, if $\ell' > N+1$, with $\ell' - N$ odd,

$$(-1)^N S_{N,\ell'} = A_{N+1} - (\overbrace{A_{N+2} - A_{N+3}}^{\geqslant 0}) - (\overbrace{A_{N+4} - A_{N+5}}^{\geqslant 0}) - \cdots - \overbrace{(A_{\ell'-1} - A_{\ell'})}^{\geqslant 0} \leqslant A_{N+1}$$

$$(-1)^N S_{N,\ell'+1} = \overbrace{(-1)^N S_{N,\ell'}}^{\leqslant A_{N+1}} - \overbrace{A_{\ell'+1}}^{\geqslant 0} \leqslant A_{N+1}$$

  This tells us that $(-1)^N S_{N,\ell} \leqslant A_{N+1}$ 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 3.3.10), this implies that, since $A_{N+1} \to 0$ as $N \to \infty$, the series converges and that

$$S - S_N = \lim_{\ell \to \infty} S_{N,\ell}$$

lies between $(-1)^N A_{N+1}$ and 0.

□

Example 3.3.15

We have already seen, in Example 3.3.6, 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) $A_n = \frac{1}{n} \geqslant 0$ for all $n \geqslant 1$, 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) $\lim_{n\to\infty} A_n = \lim_{n\to\infty} \frac{1}{n} = 0$.

so that all of the hypotheses of the alternating series test, i.e. of Theorem 3.3.14, are satisfied. We shall see, in Example 3.5.20, that

$$\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} = \log 2.$$

Example 3.3.15

Example 3.3.16 (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 3.6.3, below. In particular

$$\frac{1}{e} = e^{-1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \cdots$$

is an alternating series and satisfies all of the conditions of the alternating series test, Theorem 3.3.14a:

(i) The terms in the series alternate in sign.
(ii) The magnitude of the n^{th} term in the series decreases monotonically as n increases.
(iii) The n^{th} term in the series converges to zero as $n \to \infty$.

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

$$\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} \leqslant \frac{1}{e} \leqslant \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} + \frac{1}{10!}$$

so that

$$\frac{1}{\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!} + \frac{1}{10!}} \leqslant e \leqslant \frac{1}{\frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \frac{1}{6!} - \frac{1}{7!} + \frac{1}{8!} - \frac{1}{9!}}$$

which, to seven decimal places says

$$2.7182816 \leqslant e \leqslant 2.7182837$$

(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[25]. For example $20! \approx 2.4 \times 10^{27}$.

Example 3.3.16

Example 3.3.17

We will shortly see, in Example 3.5.20, that if $-1 < x \leqslant 1$, then

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{x^n}{n}$$

25 The interested reader may wish to check out "Stirling's approximation", which says that $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$.

Suppose that we have to compute $\log \frac{11}{10}$ to within an accuracy of 10^{-12}. Since $\frac{11}{10} = 1 + \frac{1}{10}$, we can get $\log \frac{11}{10}$ by evaluating $\log(1+x)$ at $x = \frac{1}{10}$, so that

$$\log \frac{11}{10} = \log\left(1 + \frac{1}{10}\right) = \frac{1}{10} - \frac{1}{2\times 10^2} + \frac{1}{3\times 10^3} - \frac{1}{4\times 10^4} + \cdots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{1}{n\times 10^n}$$

By the alternating series test, this series converges. Also by the alternating series test, approximating $\log \frac{11}{10}$ by throwing away all but the first N terms

$$\log \frac{11}{10} \approx \frac{1}{10} - \frac{1}{2\times 10^2} + \frac{1}{3\times 10^3} - \frac{1}{4\times 10^4} + \cdots + (-1)^{N-1} \frac{1}{N\times 10^N} = \sum_{n=1}^{N} (-1)^{n-1} \frac{1}{n\times 10^n}$$

introduces an error whose magnitude is no more than the magnitude of the first term that we threw away.

$$\text{error} \leqslant \frac{1}{(N+1)\times 10^{N+1}}$$

To achieve an error that is no more than 10^{-12}, we have to choose N so that

$$\frac{1}{(N+1)\times 10^{N+1}} \leqslant 10^{-12}$$

The best way to do so is simply to guess — we are not going to be able to manipulate the inequality $\frac{1}{(N+1)\times 10^{N+1}} \leqslant \frac{1}{10^{12}}$ into the form $N \leqslant \cdots$, 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 $N+1 \geqslant 12$, i.e. whenever $N \geqslant 11$. So we will achieve an error of less than 10^{-12} if we choose $N = 11$.

$$\left.\frac{1}{(N+1)\times 10^{N+1}}\right|_{N=11} = \frac{1}{12\times 10^{12}} < \frac{1}{10^{12}}$$

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 $\frac{1}{(N+1)\times 10^{N+1}} \leqslant \frac{1}{10^{12}}$, you can next just try $N = 10$, then $N = 9$, and so on.

$$\left.\frac{1}{(N+1)\times 10^{N+1}}\right|_{N=11} = \frac{1}{12\times 10^{12}} < \frac{1}{10^{12}}$$
$$\left.\frac{1}{(N+1)\times 10^{N+1}}\right|_{N=10} = \frac{1}{11\times 10^{11}} < \frac{1}{10\times 10^{11}} = \frac{1}{10^{12}}$$
$$\left.\frac{1}{(N+1)\times 10^{N+1}}\right|_{N=9} = \frac{1}{10\times 10^{10}} = \frac{1}{10^{11}} > \frac{1}{10^{12}}$$

So in this problem, the smallest acceptable $N = 10$.

Example 3.3.17

3.3.5 ▸▸ 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 $n \to \infty$ does. This is the basis of the ratio test.

Theorem 3.3.18 (Ratio Test).

Let N be any positive integer and assume that $a_n \neq 0$ for all $n \geqslant N$.

(a) If $\lim\limits_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = L < 1$, then $\sum\limits_{n=1}^{\infty} a_n$ converges.

(b) If $\lim\limits_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = L > 1$, or $\lim\limits_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = +\infty$, then $\sum\limits_{n=1}^{\infty} a_n$ diverges.

Warning 3.3.19.

Beware that the ratio test provides absolutely no conclusion about the convergence or divergence of the series $\sum\limits_{n=1}^{\infty} a_n$ if $\lim\limits_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = 1$. See Example 3.3.22, below.

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 $n \to \infty$. In particular there must be some natural number M so that $\left|\frac{a_{n+1}}{a_n}\right| \leqslant R$ for all $n \geqslant M$. So $|a_{n+1}| \leqslant R|a_n|$ for all $n \geqslant M$. In particular

$$\begin{aligned}
|a_{M+1}| &\leqslant R\,|a_M| \\
|a_{M+2}| &\leqslant R\,|a_{M+1}| \leqslant R^2\,|a_M| \\
|a_{M+3}| &\leqslant R\,|a_{M+2}| \leqslant R^3\,|a_M| \\
&\vdots \\
|a_{M+\ell}| &\leqslant R^{\ell}\,|a_M|
\end{aligned}$$

for all $\ell \geqslant 0$. 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 $\sum\limits_{\ell=1}^{\infty} a_{M+\ell} = \sum\limits_{n=M+1}^{\infty} a_n$ converges. So the series $\sum\limits_{n=1}^{\infty} a_n$ converges too.

(b) We are assuming that $\left|\frac{a_{n+1}}{a_n}\right|$ approaches $L > 1$ as $n \to \infty$. In particular there must be some natural number $M > N$ so that $\left|\frac{a_{n+1}}{a_n}\right| \geqslant 1$ for all $n \geqslant M$. So $|a_{n+1}| \geqslant |a_n|$ for all $n \geqslant M$. That is, $|a_n|$ increases as n increases as long as $n \geqslant M$. So $|a_n| \geqslant |a_M|$ for all $n \geqslant M$ and a_n cannot converge to zero as $n \to \infty$. So the series diverges by the divergence test. □

Example 3.3.20 $\left(\sum_{n=0}^{\infty} anx^{n-1}\right)$

Fix any two nonzero real numbers a and x. We have already seen in Example 3.2.4 — we have just renamed r to x — that the geometric series $\sum_{n=0}^{\infty} ax^n$ converges when $|x| < 1$ and diverges when $|x| \geqslant 1$. We are now going to consider a new series, constructed by differentiating[26] each term in the geometric series $\sum_{n=0}^{\infty} ax^n$. This new series is

$$\sum_{n=0}^{\infty} a_n \qquad \text{with} \quad a_n = a\, n\, x^{n-1}$$

Let's apply the ratio test.

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{a\,(n+1)\,x^n}{a\,n\,x^{n-1}}\right| = \frac{n+1}{n}|x| = \left(1+\frac{1}{n}\right)|x| \to L = |x| \quad \text{as } n \to \infty$$

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 $n \to \infty$ and the series diverges by the divergence test.

Example 3.3.20

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 $|x| > R$.

Of course, we will still have to to determine what happens when $x = +R, -R$. To determine convergence or divergence in those cases we will need to use one of the other tests we have seen.

26 We shall see later, in Theorem 3.5.13, that the function $\sum_{n=0}^{\infty} anx^{n-1}$ is indeed the derivative of the function $\sum_{n=0}^{\infty} ax^n$. 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.

Example 3.3.21 $\left(\sum_{n=0}^{\infty} \frac{a}{n+1} X^{n+1}\right)$

Once again, fix any two nonzero real numbers a and X. We again start with the geometric series $\sum_{n=0}^{\infty} ax^n$ but this time we construct a new series by integrating[27] each term, ax^n, from $x = 0$ to $x = X$ giving $\frac{a}{n+1}X^{n+1}$. The resulting new series is

$$\sum_{n=0}^{\infty} a_n \qquad \text{with } a_n = \frac{a}{n+1}X^{n+1}$$

To apply the ratio test we need to compute

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{\frac{a}{n+2}X^{n+2}}{\frac{a}{n+1}X^{n+1}}\right| = \frac{n+1}{n+2}|X| = \frac{1+\frac{1}{n}}{1+\frac{2}{n}}|X| \to L = |X| \quad \text{as } n \to \infty$$

The ratio test now tells us that the series $\sum_{n=0}^{\infty} \frac{a}{n+1}X^{n+1}$ converges if $|X| < 1$ and diverges if $|X| > 1$. It says nothing about the cases $X = \pm 1$.

If $X = 1$, the series reduces to

$$\sum_{n=0}^{\infty} \frac{a}{n+1}X^{n+1}\bigg|_{X=1} = \sum_{n=0}^{\infty} \frac{a}{n+1} = a\sum_{m=1}^{\infty} \frac{1}{m} \qquad \text{with } m = n+1$$

which is just a times the harmonic series, which we know diverges, by Example 3.3.6.

If $X = -1$, the series reduces to

$$\sum_{n=0}^{\infty} \frac{a}{n+1}X^{n+1}\bigg|_{X=-1} = \sum_{n=0}^{\infty} (-1)^{n+1}\frac{a}{n+1}$$

which converges by the alternating series test. See Example 3.3.15.

In conclusion, the series $\sum_{n=0}^{\infty} \frac{a}{n+1}X^{n+1}$ converges if and only if $-1 \leqslant X < 1$.

Example 3.3.21

The ratio test is often quite easy to apply, but one must always be careful when the limit of the ratio is 1. The next example illustrates this.

Example 3.3.22 $(L = 1)$

In this example, we are going to see three different series that all have $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = 1$. One is going to diverge and the other two are going to converge.

- The first series is the harmonic series

$$\sum_{n=1}^{\infty} a_n \qquad \text{with } a_n = \frac{1}{n}$$

We have already seen, in Example 3.3.6, that this series diverges. It has

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{\frac{1}{n+1}}{\frac{1}{n}}\right| = \frac{n}{n+1} = \frac{1}{1+\frac{1}{n}} \to L = 1 \quad \text{as } n \to \infty$$

27 We shall also see later, in Theorem 3.5.13, that the function $\sum_{n=0}^{\infty} \frac{a}{n+1}x^{n+1}$ is indeed an antiderivative of the function $\sum_{n=0}^{\infty} ax^n$.

- The second series is the alternating harmonic series

$$\sum_{n=1}^{\infty} a_n \qquad \text{with } a_n = (-1)^{n-1}\frac{1}{n}$$

We have already seen, in Example 3.3.15, that this series converges. But it also has

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(-1)^n \frac{1}{n+1}}{(-1)^{n-1}\frac{1}{n}}\right| = \frac{n}{n+1} = \frac{1}{1+\frac{1}{n}} \to L = 1 \quad \text{as } n \to \infty$$

- The third series is

$$\sum_{n=1}^{\infty} a_n \qquad \text{with } a_n = \frac{1}{n^2}$$

We have already seen, in Example 3.3.6 with $p = 2$, that this series converges. But it also has

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{\frac{1}{(n+1)^2}}{\frac{1}{n^2}}\right| = \frac{n^2}{(n+1)^2} = \frac{1}{(1+\frac{1}{n})^2} \to L = 1 \quad \text{as } n \to \infty$$

Example 3.3.22

Let's do a somewhat artificial example that forces us to combine a few of the techniques we have seen.

Example 3.3.23 $\left(\sum_{n=1}^{\infty} \frac{(-3)^n\sqrt{n+1}}{2n+3}x^n\right)$

Again, the convergence of this series will depend on x.

- Let us start with the ratio test — so we compute

$$\begin{aligned}\left|\frac{a_{n+1}}{a_n}\right| &= \left|\frac{(-3)^{n+1}\sqrt{n+2}(2n+3)x^{n+1}}{(-3)^n\sqrt{n+1}(2n+5)x^n}\right| \\ &= |-3| \cdot \frac{\sqrt{n+2}}{\sqrt{n+1}} \cdot \frac{2n+3}{2n+5} \cdot |x|\end{aligned}$$

So in the limit as $n \to \infty$ we are left with

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = 3|x|$$

- The ratio test then tells us that if $3|x| > 1$ the series diverges, while when $3|x| < 1$ the series converges.

- This leaves us with the cases $x = +\frac{1}{3}$ and $-\frac{1}{3}$.

- Setting $x = \frac{1}{3}$ gives the series

$$\sum_{n=1}^{\infty} \frac{(-1)^n \sqrt{n+1}}{2n+3}$$

The fact that the terms alternate here suggests that we use the alternating series test. That will show that this series converges provided $\frac{\sqrt{n+1}}{2n+3}$ decreases as n increases. So we define the function

$$f(t) = \frac{\sqrt{t+1}}{2t+3}$$

(which is constructed by replacing the n in $\frac{\sqrt{n+1}}{2n+3}$ with t) and verify that $f(t)$ is a decreasing function of t. To prove that, it suffices to show its derivative is negative when $t \geqslant 1$:

$$\begin{aligned} f'(t) &= \frac{(2t+3) \cdot \frac{1}{2} \cdot (t+1)^{-1/2} - 2\sqrt{t+1}}{(2t+3)^2} \\ &= \frac{(2t+3) - 4(t+1)}{2\sqrt{t+1}(2t+3)^2} \\ &= \frac{-2t-1}{2\sqrt{t+1}(2t+3)^2} \end{aligned}$$

When $t \geqslant 1$ this is negative and so $f(t)$ is a decreasing function. Thus we can apply the alternating series test to show that the series converges when $x = \frac{1}{3}$.

- When $x = -\frac{1}{3}$ the series becomes

$$\sum_{n=1}^{\infty} \frac{\sqrt{n+1}}{2n+3}.$$

Notice that when n is large, the summand is approximately $\frac{\sqrt{n}}{2n}$ which suggests that the series will diverge by comparison with $\sum n^{-1/2}$. To formalise this, we can use the limit comparison theorem:

$$\begin{aligned} \lim_{n\to\infty} \frac{\sqrt{n+1}}{2n+3} \frac{1}{n^{-1/2}} &= \lim_{n\to\infty} \frac{\sqrt{n} \cdot \sqrt{1+1/n}}{n(2+3/n)} \cdot n^{1/2} \\ &= \lim_{n\to\infty} \frac{n \cdot \sqrt{1+1/n}}{n(2+3/n)} \\ &= \frac{1}{2} \end{aligned}$$

So since this ratio has a finite limit and the series $\sum n^{-1/2}$ diverges, we know that our series also diverges.

So in summary the series converges when $-\frac{1}{3} < x \leqslant \frac{1}{3}$ and diverges otherwise.

Example 3.3.23

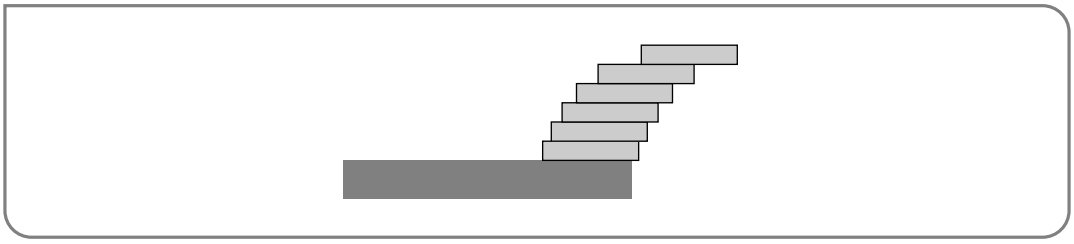
3.3.6 ▸▸ Convergence Test List

We now have half a dozen convergence tests:

- *Divergence Test*
  - works well when the n^{th} term in the series *fails* to converge to zero as n tends to infinity
- *Alternating Series Test*
  - works well when successive terms in the series alternate in sign
  - don't forget to check that successive terms decrease in magnitude and tend to zero as n tends to infinity
- *Integral Test*
  - works well when, if you substitute x for n in the n^{th} term you get a function, $f(x)$, that you can integrate
  - don't forget to check that $f(x) \geqslant 0$ 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 $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = L$
  - 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^{th} term a_n is approximately the same as a simpler term b_n (see Example 3.3.10) and it is easy to determine whether or not $\sum_{n=1}^{\infty} b_n$ converges
  - don't forget to check that $b_n \geqslant 0$
  - usually the Limit Comparison Test is easier to apply than the Comparison Test

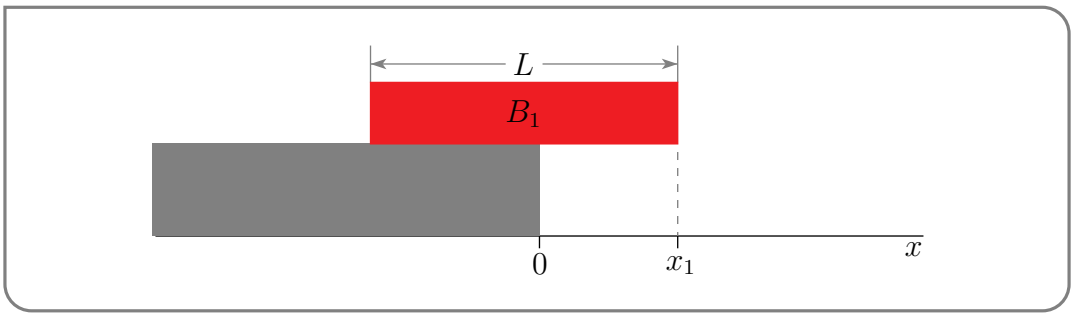
3.3.7 ▸▸ Optional — The Leaning Tower of Books

Imagine that you are about to stack a bunch of identical books on a table. But you don't want to just stack them exactly vertically. You want to built a "leaning tower of books" that overhangs the edge of the table as much as possible.



How big an overhang can you get? The answer to that question, which we'll now derive, uses a series!

- Let's start by just putting book #1 on the table. It's the red book labelled "B_1" in the figure below.



Use a horizontal x–axis with $x = 0$ corresponding to the right hand edge of the table. Imagine that we have placed book #1 so that its right hand edge overhangs the end of the table by a distance x_1.

  - In order for the book to not topple off of the table, we need its centre of mass to lie above the table. That is, we need the x–coordinate of the centre mass of B_1, which we shall denote $\bar{X}(B_1)$, to obey

$$\bar{X}(B_1) \leqslant 0$$

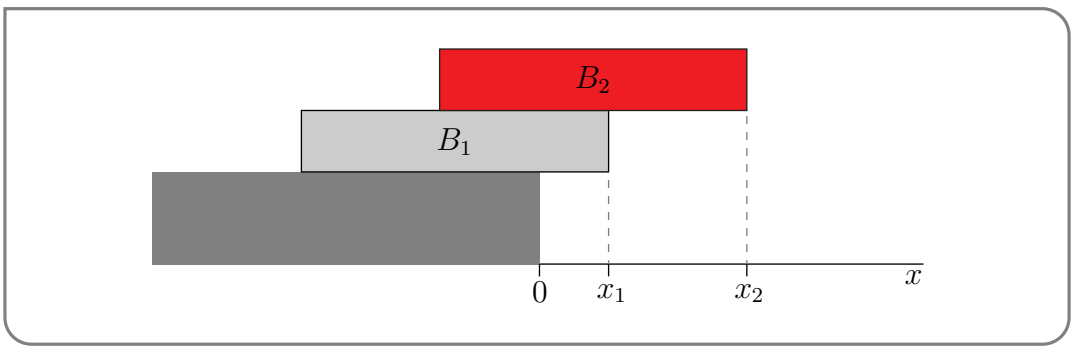
Assuming that our books have uniform density and are of length L, $\bar{X}(B_1)$ will be exactly half way between the right hand end of the book, which is at $x = x_1$, and the left hand end of the book, which is at $x = x_1 - L$. So

$$\bar{X}(B_1) = \frac{1}{2}x_1 + \frac{1}{2}(x_1 - L) = x_1 - \frac{L}{2}$$

Thus book #1 does not topple off of the table provided

$$x_1 \leqslant \frac{L}{2}$$

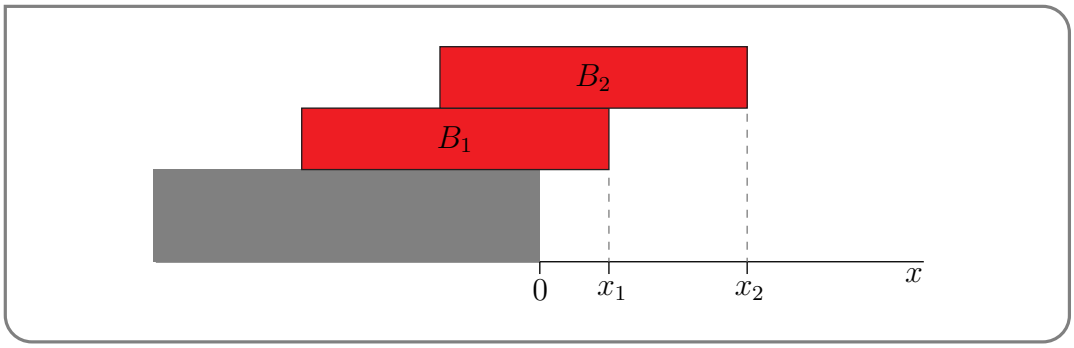
- Now let's put books #1 and #2 on the table, with the right hand edge of book #1 at $x = x_1$ and the right hand edge of book #2 at $x = x_2$, as in the figure below.



- In order for book #2 to not topple off of book #1, we need the centre of mass of book #2 to lie above book #1. That is, we need the x–coordinate of the centre mass of B_2, which is $\bar{X}(B_2) = x_2 - \frac{L}{2}$, to obey

$$\bar{X}(B_2) \leqslant x_1 \iff x_2 - \frac{L}{2} \leqslant x_1 \iff x_2 \leqslant x_1 + \frac{L}{2}$$

- Assuming that book #2 does not topple off of book #1, we still need to arrange that the pair of books does not topple off of the table. Think of the pair of books as the combined red object in the figure



In order for the combined red object to not topple off of the table, we need the centre of mass of the combined red object to lie above the table. That is, we need the x–coordinate of the centre mass of the combined red object, which we shall denote $\bar{X}(B_1 \cup B_2)$, to obey

$$\bar{X}(B_1 \cup B_2) \leqslant 0$$

The centre of mass of the combined red object is the weighted average[28] of the centres of mass of B_1 and B_2. As B_1 and B_2 have the same weight,

$$\begin{aligned}\bar{X}(B_1 \cup B_2) &= \frac{1}{2}\bar{X}(B_1) + \frac{1}{2}\bar{X}(B_2) = \frac{1}{2}\Big(x_1 - \frac{L}{2}\Big) + \frac{1}{2}\Big(x_2 - \frac{L}{2}\Big) \\ &= \frac{1}{2}(x_1 + x_2) - \frac{L}{2}\end{aligned}$$

and the combined red object does not topple off of the table if

$$\bar{X}(B_1 \cup B_2) = \frac{1}{2}(x_1 + x_2) - \frac{L}{2} \leqslant 0 \iff x_1 + x_2 \leqslant L$$

In conclusion, our two–book tower survives if

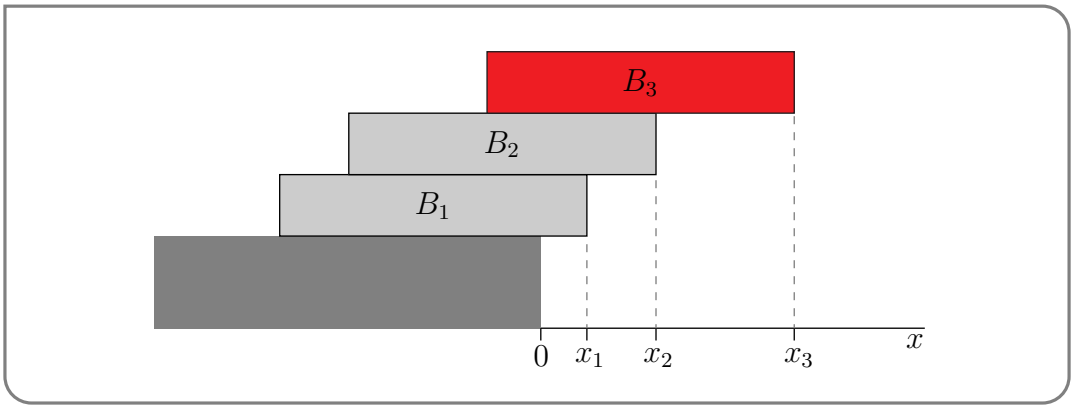
$$x_2 \leqslant x_1 + \frac{L}{2} \qquad \text{and} \qquad x_1 + x_2 \leqslant L$$

In particular we may choose x_1 and x_2 to satisfy $x_2 = x_1 + \frac{L}{2}$ and $x_1 + x_2 = L$. Then, substituting $x_2 = x_1 + \frac{L}{2}$ into $x_1 + x_2 = L$ gives

$$x_1 + \Big(x_1 + \frac{L}{2}\Big) = L \iff 2x_1 = \frac{L}{2} \iff x_1 = \frac{L}{2}\Big(\frac{1}{2}\Big), \quad x_2 = \frac{L}{2}\Big(1 + \frac{1}{2}\Big)$$

28 It might be a good idea to review the beginning of §2.3 at this point.

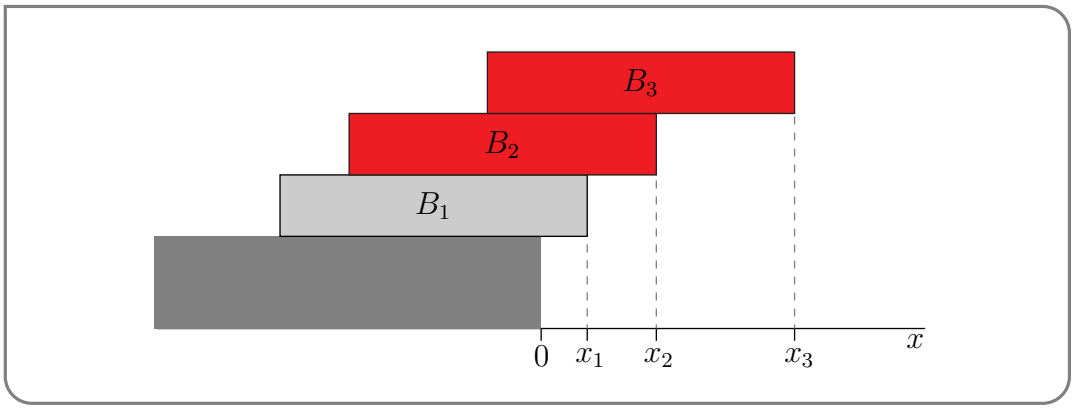
- Before considering the general "n–book tower", let's now put books #1, #2 and #3 on the table, with the right hand edge of book #1 at $x = x_1$, the right hand edge of book #2 at $x = x_2$, and the right hand edge of book #3 at $x = x_3$, as in the figure below.



  - In order for book #3 to not topple off of book #2, we need the centre of mass of book #3 to lie above book #2. That is, we need the x–coordinate of the centre mass of B_3, which is $\bar{X}(B_3) = x_3 - \frac{L}{2}$, to obey

$$\bar{X}(B_3) \leqslant x_2 \iff x_3 - \frac{L}{2} \leqslant x_2 \iff x_3 \leqslant x_2 + \frac{L}{2}$$

  - Assuming that book #3 does not topple off of book #2, we still need to arrange that the pair of books, book #2 plus book #3 (the red object in the figure below), does not topple off of book #1.



    In order for this combined red object to not topple off of book #1, we need the x–coordinate of its centre mass, which we denote $\bar{X}(B_2 \cup B_3)$, to obey

$$\bar{X}(B_2 \cup B_3) \leqslant x_1$$

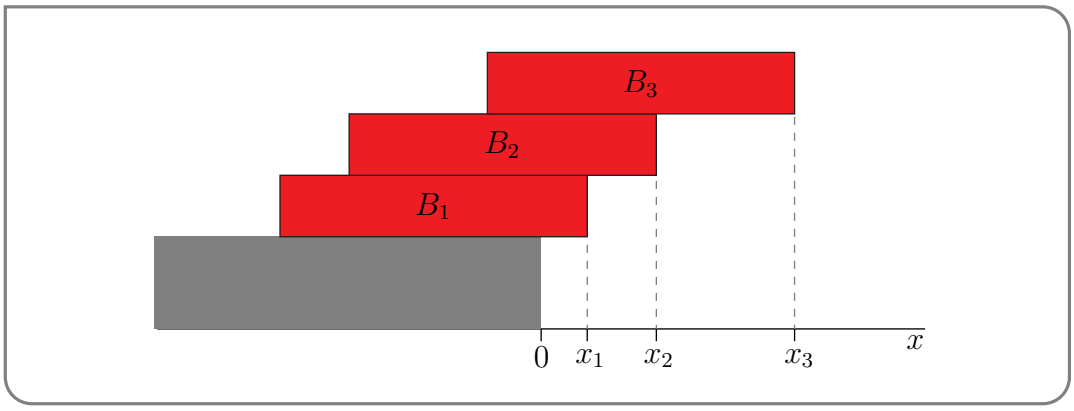
    The centre of mass of the combined red object is the weighted average of the centre of masses of B_2 and B_3. As B_2 and B_3 have the same weight,

$$\begin{aligned}\bar{X}(B_2 \cup B_3) &= \frac{1}{2}\bar{X}(B_2) + \frac{1}{2}\bar{X}(B_3) = \frac{1}{2}\Big(x_2 - \frac{L}{2}\Big) + \frac{1}{2}\Big(x_3 - \frac{L}{2}\Big) \\ &= \frac{1}{2}(x_2 + x_3) - \frac{L}{2}\end{aligned}$$

and the combined red object does not topple off of book #1 if

$$\frac{1}{2}(x_2+x_3)-\frac{L}{2}\leqslant x_1 \iff x_2+x_3\leqslant 2x_1+L$$

- Assuming that book #3 does not topple off of book #2, and also that the combined book #2 plus book #3 does not topple off of book #1, we still need to arrange that the whole tower of books, book #1 plus book #2 plus book #3 (the red object in the figure below), does not topple off of the table.



In order for this combined red object to not topple off of the table, we need the x–coordinate of its centre mass, which we denote $\bar{X}(B_1\cup B_2\cup B_3)$, to obey

$$\bar{X}(B_1\cup B_2\cup B_3)\leqslant 0$$

The centre of mass of the combined red object is the weighted average of the centre of masses of B_1 and B_2 and B_3. As they all have the same weight,

$$\begin{aligned}\bar{X}(B_1\cup B_2\cup B_3)&=\frac{1}{3}\bar{X}(B_1)+\frac{1}{3}\bar{X}(B_2)+\frac{1}{3}\bar{X}(B_3)\\&=\frac{1}{3}\Big(x_1-\frac{L}{2}\Big)+\frac{1}{3}\Big(x_2-\frac{L}{2}\Big)+\frac{1}{3}\Big(x_3-\frac{L}{2}\Big)\\&=\frac{1}{3}(x_1+x_2+x_3)-\frac{L}{2}\end{aligned}$$

and the combined red object does not topple off of the table if

$$\frac{1}{3}(x_1+x_2+x_3)-\frac{L}{2}\leqslant 0 \iff x_1+x_2+x_3\leqslant\frac{3L}{2}$$

In conclusion, our three–book tower survives if

$$x_3\leqslant x_2+\frac{L}{2}\qquad\text{and}\qquad x_2+x_3\leqslant 2x_1+L\qquad\text{and}\qquad x_1+x_2+x_3\leqslant\frac{3L}{2}$$

In particular, we may choose x_1, x_2 and x_3 to satisfy

$$\begin{aligned}x_1+x_2+x_3&=\frac{3L}{2}\qquad\text{and}\\x_2+x_3&=2x_1+L\qquad\text{and}\\x_3&=\frac{L}{2}+x_2\end{aligned}$$

Substituting the second equation into the first gives

$$3x_1 + L = \frac{3L}{2} \implies x_1 = \frac{L}{2}\Big(\frac{1}{3}\Big)$$

Next substituting the third equation into the second, and then using the formula above for x_1, gives

$$2x_2 + \frac{L}{2} = 2x_1 + L = \frac{L}{3} + L \implies x_2 = \frac{L}{2}\Big(\frac{1}{2} + \frac{1}{3}\Big)$$

and finally

$$x_3 = \frac{L}{2} + x_2 = \frac{L}{2}\Big(1 + \frac{1}{2} + \frac{1}{3}\Big)$$

- We are finally ready for the general "n–book tower". Stack n books on the table, with book B_1 on the bottom and book B_n at the top, and with the right hand edge of book #j at $x = x_j$. The same centre of mass considerations as above show that the tower survives if

$$\begin{aligned}
\bar{X}(B_n) &\leqslant x_{n-1} & x_n - \frac{L}{2} &\leqslant x_{n-1}\\
\bar{X}(B_{n-1} \cup B_n) &\leqslant x_{n-2} & \frac{1}{2}(x_{n-1} + x_n) - \frac{L}{2} &\leqslant x_{n-2}\\
&\vdots & &\vdots\\
\bar{X}(B_3 \cup \cdots \cup B_n) &\leqslant x_2 & \frac{1}{n-2}(x_3 + \cdots + x_n) - \frac{L}{2} &\leqslant x_2\\
\bar{X}(B_2 \cup B_3 \cup \cdots \cup B_n) &\leqslant x_1 & \frac{1}{n-1}(x_2 + x_3 + \cdots + x_n) - \frac{L}{2} &\leqslant x_1\\
\bar{X}(B_1 \cup B_2 \cup B_3 \cup \cdots \cup B_n) &\leqslant 0 & \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) - \frac{L}{2} &\leqslant 0
\end{aligned}$$

In particular, we may choose the x_j's to obey

$$\begin{aligned}
\frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) &= \frac{L}{2}\\
\frac{1}{n-1}(x_2 + x_3 + \cdots + x_n) &= \frac{L}{2} + x_1\\
\frac{1}{n-2}(x_3 + \cdots + x_n) &= \frac{L}{2} + x_2\\
&\vdots & &\vdots\\
\frac{1}{2}(x_{n-1} + x_n) &= \frac{L}{2} + x_{n-2}\\
x_n &= \frac{L}{2} + x_{n-1}
\end{aligned}$$

Substituting $x_2 + x_3 + \cdots + x_n = (n-1)x_1 + \frac{L}{2}(n-1)$ from the second equation into the first equation gives

$$\frac{1}{n}\Big\{\overbrace{x_1 + (n-1)x_1}^{nx_1} + \frac{L}{2}(n-1)\Big\} = \frac{L}{2} \implies x_1 + \frac{L}{2}\Big(1 - \frac{1}{n}\Big) = \frac{L}{2}\Big(\frac{1}{2}\Big)$$
$$\implies x_1 = \frac{L}{2}\Big(\frac{1}{n}\Big)$$

Substituting $x_3 + \cdots + x_n = (n-2)x_2 + \frac{L}{2}(n-2)$ from the third equation into the second equation gives

$$\frac{1}{n-1}\Big\{\overbrace{x_2 + (n-2)x_2}^{(n-1)x_2} + \frac{L}{2}(\overbrace{n-2}^{(n-1)-1})\Big\} = \frac{L}{2} + x_1 = \frac{L}{2}\Big(1 + \frac{1}{n}\Big)$$
$$\implies x_2 + \frac{L}{2}\Big(1 - \frac{1}{n-1}\Big) = \frac{L}{2}\Big(1 + \frac{1}{n}\Big)$$
$$\implies x_2 = \frac{L}{2}\Big(\frac{1}{n-1} + \frac{1}{n}\Big)$$

Just keep going. We end up with

$$\begin{aligned}
x_1 &= \frac{L}{2}\Big(\frac{1}{n}\Big)\\
x_2 &= \frac{L}{2}\Big(\frac{1}{n-1} + \frac{1}{n}\Big)\\
x_3 &= \frac{L}{2}\Big(\frac{1}{n-2} + \frac{1}{n-1} + \frac{1}{n}\Big)\\
&\vdots\\
x_{n-2} &= \frac{L}{2}\Big(\frac{1}{3} + \cdots + \frac{1}{n}\Big)\\
x_{n-1} &= \frac{L}{2}\Big(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\Big)\\
x_n &= \frac{L}{2}\Big(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\Big)
\end{aligned}$$

Our overhang is $x_n = \frac{L}{2}\big(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\big)$. This is $\frac{L}{2}$ times the n^{th} partial sum of the harmonic series $\sum_{m=1}^{\infty} \frac{1}{m}$. As we saw in Example 3.3.6 (the p test), the harmonic series diverges. So, as n goes to infinity $1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ also goes to infinity. We may make the overhang as large[29] as we like!

3.3.8 ▸▸ Optional — The Root Test

There is another test that is very similar in spirit to the ratio test. It also comes from a reexamination of the geometric series

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} ar^n$$

29 At least if our table is strong enough.

The ratio test was based on the observation that r, which largely determines whether or not the series converges, could be found by computing the ratio $r = a_{n+1}/a_n$. The root test is based on the observation that $|r|$ can also be determined by looking that the n^{th} root of the n^{th} term with n very large:

$$\lim_{n\to\infty} \sqrt[n]{|ar^n|} = |r| \lim_{n\to\infty} \sqrt[n]{|a|} = |r| \qquad \text{if } a \neq 0$$

Of course, in general, the n^{th} term is not exactly ar^n. However, if for very large n, the n^{th} term is approximately proportional to r^n, with $|r|$ given by the above limit, we would expect the series to converge when $|r| < 1$ and diverge when $|r| > 1$. That is indeed the case.

Theorem 3.3.24 (Root Test).

Assume that

$$L = \lim_{n\to\infty} \sqrt[n]{|a_n|}$$

exists or is $+\infty$.

(a) If $L < 1$, then $\sum_{n=1}^{\infty} a_n$ converges.

(b) If $L > 1$, or $L = +\infty$, then $\sum_{n=1}^{\infty} a_n$ diverges.

Warning 3.3.25.

Beware that the root test provides absolutely no conclusion about the convergence or divergence of the series $\sum_{n=1}^{\infty} a_n$ if $\lim_{n\to\infty} \sqrt[n]{|a_n|} = 1$.

Proof. (a) Pick any number R obeying $L < R < 1$. We are assuming that $\sqrt[n]{|a_n|}$ approaches L as $n \to \infty$. In particular there must be some natural number M so that $\sqrt[n]{|a_n|} \leqslant R$ for all $n \geqslant M$. So $|a_n| \leqslant R^n$ for all $n \geqslant M$ and the series $\sum_{n=1}^{\infty} a_n$ converges by comparison to the geometric series $\sum_{n=1}^{\infty} R^n$

(b) We are assuming that $\sqrt[n]{|a_n|}$ approaches $L > 1$ (or grows unboundedly) as $n \to \infty$. In particular there must be some natural number M so that $\sqrt[n]{|a_n|} \geqslant 1$ for all $n \geqslant M$. So $|a_n| \geqslant 1$ for all $n \geqslant M$ and the series diverges by the divergence test.

□

Example 3.3.26 $\left(\sum_{n=1}^{\infty} \frac{(-3)^n\sqrt{n+1}}{2n+3} x^n\right)$

We have already used the ratio test, in Example 3.3.23, to show that this series converges

when $|x| < \frac{1}{3}$ and diverges when $|x| > \frac{1}{3}$. We'll now use the root test to draw the same conclusions.

- Write $a_n = \frac{(-3)^n\sqrt{n+1}}{2n+3}x^n$.
- We compute

$$\begin{aligned}\sqrt[n]{|a_n|} &= \sqrt[n]{\left|\frac{(-3)^n\sqrt{n+1}}{2n+3}x^n\right|} \\ &= 3|x|\big(n+1\big)^{1/2n}\big(2n+3\big)^{-1/n}\end{aligned}$$

- We'll now show that the limit of $\big(n+1\big)^{1/2n}$ as $n \to \infty$ is exactly 1. To do, so we first compute the limit of the logarithm.

$$\begin{aligned}\lim_{n\to\infty} \log\big(n+1\big)^{1/2n} &= \lim_{n\to\infty} \frac{\log\big(n+1\big)}{2n} && \text{now apply Theorem 3.1.6} \\ &= \lim_{t\to\infty} \frac{\log\big(t+1\big)}{2t} \\ &= \lim_{t\to\infty} \frac{\frac{1}{t+1}}{2} && \text{by l'Hôpital} \\ &= 0\end{aligned}$$

  So

$$\lim_{n\to\infty} \big(n+1\big)^{1/2n} = \lim_{n\to\infty} \exp\big\{\log\big(n+1\big)^{1/2n}\big\} = e^0 = 1$$

  An essentially identical computation also gives that $\lim_{n\to\infty}\big(2n+3\big)^{-1/n} = e^0 = 1$.

- So

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = 3|x|$$

and the root test also tells us that if $3|x| > 1$ the series diverges, while when $3|x| < 1$ the series converges.

Example 3.3.26

We have done the last example once, in Example 3.3.23, using the ratio test and once, in Example 3.3.26, using the root test. It was clearly much easier to use the ratio test. Here is an example that is most easily handled by the root test

Example 3.3.27 $\left(\sum_{n=1}^{\infty}\left(\frac{n}{n+1}\right)^{n^2}\right)$

Write $a_n = \left(\frac{n}{n+1}\right)^{n^2}$. Then

$$\sqrt[n]{|a_n|} = \sqrt[n]{\left(\frac{n}{n+1}\right)^{n^2}} = \left(\frac{n}{n+1}\right)^{n} = \left(1+\frac{1}{n}\right)^{-n}$$

Now we take the limit,

$$\begin{aligned}\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{-n} &= \lim_{X\to\infty}\left(1+\frac{1}{X}\right)^{-X} && \text{by Theorem 3.1.6}\\ &= \lim_{x\to 0}\left(1+x\right)^{-1/x} && \text{where } x=\frac{1}{X}\\ &= e^{-1}\end{aligned}$$

by Example 3.7.20 in the CLP-1 text with $a=-1$. As the limit is strictly smaller than 1, the series $\sum_{n=1}^{\infty}\left(\frac{n}{n+1}\right)^{n^2}$ converges.

To draw the same conclusion using the ratio test, one would have to show that the limit of

$$\frac{a_{n+1}}{a_n}=\left(\frac{n+1}{n+2}\right)^{(n+1)^2}\left(\frac{n+1}{n}\right)^{n^2}$$

as $n\to\infty$ is strictly smaller than 1. It's clearly better to stick with the root test.

Example 3.3.27

3.3.9 ▸▸ Optional — Harmonic and Basel Series

▸▸▸ The Harmonic Series

The series

$$\sum_{n=1}^{\infty}\frac{1}{n}$$

that appeared in Warning 3.3.3, is called the Harmonic series[30], and its partial sums

$$H_N=\sum_{n=1}^{N}\frac{1}{n}$$

are called the Harmonic numbers. Though these numbers have been studied at least as far back as Pythagoras, the divergence of the series was first proved in around 1350 by Nicholas Oresme (1320-5 – 1382), though the proof was lost for many years and rediscovered by Mengoli (1626–1686) and the Bernoulli brothers (Johann 1667–1748 and Jacob 1655–1705).

Oresme's proof is beautiful and all the more remarkable that it was produced more than 300 years before calculus was developed by Newton and Leibnitz. It starts by group-

30 The interested reader should use their favourite search engine to read more on the link between this series and musical harmonics. You can also find interesting links between the Harmonic series and the so-called "jeep problem" and also the problem of stacking a tower of dominoes to create an overhang that does not topple over.

ing the terms of the harmonic series carefully:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1}{n} &= 1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\cdots\\ &= 1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)+\left(\frac{1}{9}+\frac{1}{10}+\cdots+\frac{1}{15}+\frac{1}{16}\right)+\cdots\\ &> 1+\frac{1}{2}+\left(\frac{1}{4}+\frac{1}{4}\right)+\left(\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\right)+\left(\frac{1}{16}+\frac{1}{16}+\cdots+\frac{1}{16}+\frac{1}{16}\right)+\cdots\\ &= 1+\frac{1}{2}+\left(\frac{2}{4}\right)+\left(\frac{4}{8}\right)+\left(\frac{8}{16}\right)+\cdots\end{aligned}$$

So one can see that this is $1+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\cdots$ and so must diverge[31].

There are many variations on Oresme's proof — for example, using groups of two or three. A rather different proof relies on the inequality

$$e^x > 1+x \qquad \text{for} \quad x>0$$

which follows immediately from the Taylor series for e^x given in Theorem 3.6.5. From this we can bound the exponential of the Harmonic numbers:

$$\begin{aligned}e^{H_n} &= e^{1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}}\\ &= e^1\cdot e^{1/2}\cdot e^{1/3}\cdot e^{1/4}\cdots e^{1/n}\\ &> (1+1)\cdot(1+1/2)\cdot(1+1/3)\cdot(1+1/4)\cdots(1+1/n)\\ &= \frac{2}{1}\cdot\frac{3}{2}\cdot\frac{4}{3}\cdot\frac{5}{4}\cdots\frac{n+1}{n}\\ &= n+1\end{aligned}$$

Since e^{H_n} grows unboundedly with n, the harmonic series diverges.

▸▸▸ The Basel Problem

The problem of determining the exact value of the sum of the series

$$\sum_{n=1}^{\infty}\frac{1}{n^2}$$

is called the Basel problem. The problem is named after the home town of Leonhard Euler, who solved it. One can use telescoping series to show that this series must converge. Notice that

$$\frac{1}{n^2} < \frac{1}{n(n-1)} = \frac{1}{n-1}-\frac{1}{n}$$

31 The grouping argument can be generalised further and the interested reader should look up Cauchy's condensation test.

Hence we can bound the partial sum:

$$\begin{aligned} S_k = \sum_{n=1}^{k} \frac{1}{n^2} &< 1 + \sum_{n=2}^{k} \frac{1}{n(n-1)} && \text{avoid dividing by 0} \\ &= 1 + \sum_{n=2}^{k} \left(\frac{1}{n-1} - \frac{1}{n} \right) && \text{which telescopes to} \\ &= 1 + 1 - \frac{1}{k} \end{aligned}$$

Thus, as k increases, the partial sum S_k increases (the series is a sum of positive terms), but is always smaller than 2. So the sequence of partial sums converges.

Mengoli posed the problem of evaluating the series exactly in 1644 and it was solved — not entirely rigorously — by Euler in 1734. A rigorous proof had to wait another 7 years. Euler used some extremely cunning observations and manipulations of the sine function to show that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

He used the Maclaurin series

$$\sin x = 1 - \frac{x^3}{6} + \frac{x^5}{24} - \cdots$$

and a product formula for sine

$$\begin{aligned} \sin x &= x \cdot \left(1 - \frac{x}{\pi}\right) \cdot \left(1 + \frac{x}{\pi}\right) \cdot \left(1 - \frac{x}{2\pi}\right) \cdot \left(1 + \frac{x}{2\pi}\right) \cdot \left(1 - \frac{x}{3\pi}\right) \cdot \left(1 + \frac{x}{3\pi}\right) \cdots \\ &= x \cdot \left(1 - \frac{x^2}{\pi}\right) \cdot \left(1 - \frac{x^2}{4\pi}\right) \cdot \left(1 - \frac{x^2}{9\pi}\right) \cdots \end{aligned} \tag{3.3.1}$$

Extracting the coefficient of x^3 from both expansions gives the desired result. The proof of the product formula is well beyond the scope of this course. But notice that at least the values of x which make the left hand side of (3.3.1) zero, namely $x = n\pi$ with n integer, are exactly the same as the values of x which make the right hand side of (3.3.1) zero[32].

This approach can also be used to compute $\sum_{n=1}^{\infty} n^{-2p}$ for $p = 1, 2, 3, \ldots$ and show that they are rational multiples[33] of π^{2p}. The corresponding series of odd powers are significantly nastier and getting closed form expressions for them remains a famous open problem.

3.3.10 ▸▸ Optional — Some Proofs

In this optional section we provide proofs of two convergence tests. We shall repeatedly use the fact that any sequence a_1, a_2, a_3, $\cdots$, of real numbers which is increasing (i.e.

32 Knowing that the left and right hand sides of (3.3.1) are zero for the same values of x is far from the end of the story. Two functions $f(x)$ and $g(x)$ having the same zeros, need not be equal. It is certainly possible that $f(x) = g(x) * A(x)$ where $A(x)$ is a function that is nowhere zero. The interested reader should look up the Weierstrass factorisation theorem

33 Search–engine your way to "Riemann zeta function".

$a_{n+1} \geqslant a_n$ for all n) and bounded (i.e. there is a constant M such that $a_n \leqslant M$ for all n) converges. We shall not prove this fact[34].

We start with the comparison test, and then move on to the alternating series test.

Theorem 3.3.28 (The Comparison Test)**.**

Let N_0 be a natural number and let $K > 0$.

(a) If $|a_n| \leqslant Kc_n$ for all $n \geqslant N_0$ and $\sum_{n=0}^{\infty} c_n$ converges, then $\sum_{n=0}^{\infty} a_n$ converges.

(b) If $a_n \geqslant Kd_n \geqslant 0$ for all $n \geqslant N_0$ 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 3.2.8,

$$\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} Kc_n - \sum_{n=0}^{\infty} [Kc_n - a_n]$$

will converge too. But for all $n \geqslant N_0$, $Kc_n - a_n \geqslant 0$ so that, for all $N \geqslant N_0$, 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 $N \to \infty$.

(b) For all $N > N_0$, the partial sum

$$S_N = \sum_{n=0}^{N} a_n \geqslant \sum_{n=0}^{N_0} a_n + K \sum_{n=N_0+1}^{N} d_n$$

By hypothesis, $\sum_{n=N_0+1}^{N} d_n$, and hence S_N, grows without bound as $N \to \infty$. So $S_N \to \infty$ as $N \to \infty$. □

34 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".

Theorem 3.3.28 (Alternating Series Test).

Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of real numbers that obeys

(i) $a_n \geqslant 0$ for all $n \geqslant 1$ and
(ii) $a_{n+1} \leqslant a_n$ for all $n \geqslant 1$ (i.e. the sequence is monotone decreasing) and
(iii) $\lim_{n\to\infty} a_n = 0$.

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^{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

$$\begin{aligned} S_{2n} &= \overbrace{(a_1 - a_2)}^{\geqslant 0} + \overbrace{(a_3 - a_4)}^{\geqslant 0} + \cdots + \overbrace{(a_{2n-1} - a_{2n})}^{\geqslant 0} \\ &\leqslant \overbrace{(a_1 - a_2)}^{\geqslant 0} + \overbrace{(a_3 - a_4)}^{\geqslant 0} + \cdots + \overbrace{(a_{2n-1} - a_{2n})}^{\geqslant 0} + \overbrace{(a_{2n+1} - a_{2n+2})}^{\geqslant 0} = S_{2(n+1)} \end{aligned}$$

and

$$\begin{aligned} S_{2n} &= a_1 - \overbrace{(a_2 - a_3)}^{\geqslant 0} - \overbrace{(a_4 - a_5)}^{\geqslant 0} - \cdots - \overbrace{(a_{2n-2} - a_{2n-1})}^{\geqslant 0} - \overbrace{a_{2n}}^{\geqslant 0} \\ &\leqslant a_1 \end{aligned}$$

So the sequence $S_2, S_4, S_6, \cdots$ 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 $n \to \infty$, 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 §3.3.4. □

3.4▲ 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 examples 3.3.20 and 3.3.21 showed us, that we can multiply or divide the n^{th} term by n and it will still converge. We can even multiply the n^{th} term by $(-1)^n$ (making it an alternating series), 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 the following two series converge:

$$\sum_{n=1}^{\infty} \frac{1}{n^{1.00000001}} \qquad\qquad \sum_{n=1}^{\infty} (-1)^n \frac{1}{n}.$$

This suggests that the divergence of the Harmonic series is much more delicate. In this section, we discuss one way to characterise this sort of delicate convergence — especially in the presence of changes of sign.

3.4.1 ▸▸ Definitions

Definition 3.4.1 (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 3.3.8.a, with $c_n = |a_n|$.

Theorem 3.4.2 (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 3.4.2 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 3.4.3 $\left(\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n}\right)$

The alternating harmonic series $\sum\limits_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n}$ of Example 3.3.15 converges (by the alternating series test). But the harmonic series $\sum\limits_{n=1}^{\infty}\frac{1}{n}$ of Example 3.3.6 diverges (by the integral test). So the alternating harmonic series $\sum\limits_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n}$ converges conditionally.

Example 3.4.3

Example 3.4.4 $\left(\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n^2}\right)$

Because the series $\sum_{n=1}^{\infty}\left|(-1)^{n-1}\frac{1}{n^2}\right| = \sum\limits_{n=1}^{\infty}\frac{1}{n^2}$ of Example 3.3.6 converges (by the integral test), the series $\sum\limits_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n^2}$ converges absolutely, and hence converges.

Example 3.4.4

Example 3.4.5 (random signs)

Imagine flipping a coin infinitely many times. Set $\sigma_n = +1$ if the n^{th} flip comes up heads and $\sigma_n = 0$ if the n^{th} flip comes up tails. The series $\sum_{n=1}^{\infty}(-1)^{\sigma_n}\frac{1}{n^2}$ is not in general an alternating series. But we know that the series $\sum_{n=1}^{\infty}\left|(-1)^{\sigma_n}\frac{1}{n^2}\right| = \sum\limits_{n=1}^{\infty}\frac{1}{n^2}$ converges. So $\sum_{n=1}^{\infty}(-1)^{\sigma_n}\frac{1}{n^2}$ converges absolutely, and hence converges.

Example 3.4.5

3.4.2 ▸▸ Optional — The Delicacy of Conditionally Convergent Series

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 3.4.6

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 3.5.20, we'll show that this series is equal to $\log 2$. However, by rearranging the terms we can make it sum to $\frac{1}{2}\log 2$. The usual order is

$$\frac{1}{1}-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\cdots$$

For the moment think of the terms being paired as follows:

$$\left(\frac{1}{1}-\frac{1}{2}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\left(\frac{1}{5}-\frac{1}{6}\right)+\cdots$$

so the denominators go odd-even odd-even. Now rearrange the terms so the denominators are odd-even-even odd-even-even:

$$\left(1-\frac{1}{2}-\frac{1}{4}\right)+\left(\frac{1}{3}-\frac{1}{6}-\frac{1}{8}\right)+\left(\frac{1}{5}-\frac{1}{10}-\frac{1}{12}\right)+\cdots$$

Now notice that the first term in each triple is exactly twice the second term. If we now combine those terms we get

$$\begin{aligned}&\left(\underbrace{1-\frac{1}{2}}_{=1/2}-\frac{1}{4}\right)+\left(\underbrace{\frac{1}{3}-\frac{1}{6}}_{=1/6}-\frac{1}{8}\right)+\left(\underbrace{\frac{1}{5}-\frac{1}{10}}_{=1/10}-\frac{1}{12}\right)+\cdots\\ &=\left(\frac{1}{2}-\frac{1}{4}\right)+\left(\frac{1}{6}-\frac{1}{8}\right)+\left(\frac{1}{10}-\frac{1}{12}\right)+\cdots\end{aligned}$$

We can now extract a factor of $\frac{1}{2}$ from each term, so

$$\begin{aligned}&=\frac{1}{2}\left(\frac{1}{1}-\frac{1}{2}\right)+\frac{1}{2}\left(\frac{1}{3}-\frac{1}{4}\right)+\frac{1}{2}\left(\frac{1}{5}-\frac{1}{6}\right)+\cdots\\ &=\frac{1}{2}\left[\left(\frac{1}{1}-\frac{1}{2}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\left(\frac{1}{5}-\frac{1}{6}\right)+\cdots\right]\end{aligned}$$

So by rearranging the terms, the sum of the series is now exactly half the original sum!

Example 3.4.6

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[35]. 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 3.4.7

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.

$$1,\ \frac{1}{3},\ \frac{1}{5},\ \frac{1}{7},\ \cdots \qquad \text{and} \qquad -\frac{1}{2},\ -\frac{1}{4},\ -\frac{1}{6},\ \cdots$$

- Notice that that if we add together the numbers in the second list,we get

$$-\frac{1}{2}\left[1+\frac{1}{2}+\frac{1}{3}+\cdots\right]$$

  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

$$1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}\cdots \qquad \text{which is greater than} \qquad \frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\cdots$$

  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 $+\infty$.

- Next move just enough numbers from the beginning of the second list into the reordered 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 $-\infty$.

35 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 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 3.4.7

3.5▲ 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 3.2.4), for $|x|<1$ this series converges to a limit, that varies with x, while for $|x|\geqslant 1$ the series diverges. Consequently we can consider this series to be a function of x

$$f(x)=\sum_{n=0}^{\infty} x^n \qquad \text{on the domain } |x|<1.$$

Furthermore (also from Example 3.2.4) 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 3.5.13). This means, for example, that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{1}{1-x}\right\} &= \frac{\mathrm{d}}{\mathrm{d}x}\sum_{n=0}^{\infty} x^n && \text{provided } |x| < 1 \\
&= \sum_{n=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x} x^n && \text{just differentiate term by term} \\
&= \sum_{n=0}^{\infty} n x^{n-1}
\end{aligned}$$

and in a very similar way

$$\begin{aligned}
\int \frac{1}{1-x}\mathrm{d}x &= \int \sum_{n=0}^{\infty} x^n \mathrm{d}x && \text{provided } |x| < 1 \\
&= \sum_{n=0}^{\infty} \int x^n \mathrm{d}x && \text{just integrate term by term} \\
&= C + \sum_{n=0}^{\infty} \frac{1}{n+1} x^{n+1}
\end{aligned}$$

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 section, the development of machinery to define and understand power series. This will allow us to answer questions[36] like

$$\text{Is } e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \,?$$

Our starting point (now that we have equipped ourselves with basic ideas about series), is the definition of power series.

36 Recall that $n! = 1 \times 2 \times 3 \times \cdots \times n$ is called "n factorial". By convention $0! = 1$.

3.5.1 ▸▸ Radius and Interval of Convergence

Definition 3.5.1.

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

$$A_0 + A_1 x + A_2 x^2 + A_3 x^3 + \cdots = \sum_{n=0}^{\infty} A_n x^n$$

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.

One possible value of x is $x = c$ and then the series reduces[37] to

$$\begin{aligned}\sum_{n=0}^{\infty} A_n(x-c)^n\Big|_{x=c} &= \sum_{n=0}^{\infty} A_n(c-c)^n \\ &= \underbrace{A_0}_{n=0} + \underbrace{0}_{n=1} + \underbrace{0}_{n=2} + \underbrace{0}_{n=3} + \cdots\end{aligned}$$

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^{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

$$\begin{aligned}\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| &= \lim_{n\to\infty}\left|\frac{A_{n+1}(x-c)^{n+1}}{A_n(x-c)^n}\right| \\ &= \lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right|\cdot|x-c| \\ &= |x-c|\cdot\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right|.\end{aligned}$$

When we do so there are several possible outcomes.

37 *By convention*, when the term $(x-c)^0$ appears in a power series, it has value 1 for all values of x, even $x = c$.

- If the limit of ratios exists and is non-zero

$$\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right| = A \neq 0,$$

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 3.5.2.

$$R = \frac{1}{A} = \left[\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right|\right]^{-1}$$

is called the *radius of convergence* of the series[38].

- If the limit of ratios exists and is zero

$$\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right| = 0$$

then $\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right||x-c| = 0$ 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 $+\infty$

$$\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right| = +\infty$$

then $\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right||x-c| = +\infty$ 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 $A_0+0+0+0+0+\cdots$, 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 $n\to\infty$, 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.

38 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.

Definition 3.5.3.

(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 3.5.4 (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 $|x| \geqslant 1$. So, in the terminology of Definition 3.5.3, the geometric series has radius of convergence $R = 1$. As a consistency check, we can also compute R using (3.5.2). The series $\sum_{n=0}^{\infty} ax^n$ has $A_n = a$. So

$$R = \left[\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right|\right]^{-1} = \left[\lim_{n\to\infty} 1\right]^{-1} = 1$$

as expected.

Example 3.5.4

Example 3.5.5 (Radius of convergence = $+\infty$)

The series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ has $A_n = \frac{1}{n!}$. So

$$\begin{aligned}\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right| &= \lim_{n\to\infty}\frac{1/(n+1)!}{1/n!} = \lim_{n\to\infty}\frac{n!}{(n+1)!} = \lim_{n\to\infty}\frac{1\times 2\times 3\times\cdots\times n}{1\times 2\times 3\times\cdots\times n\times(n+1)}\\ &= \lim_{n\to\infty}\frac{1}{n+1}\\ &= 0\end{aligned}$$

and $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ has radius of convergence ∞. It converges for every x.

Example 3.5.5

Example 3.5.6 (Radius of convergence = 0)

The series $\sum_{n=0}^{\infty} n!x^n$ has $A_n = n!$. So

$$\begin{aligned}\lim_{n\to\infty}\left|\frac{A_{n+1}}{A_n}\right| &= \lim_{n\to\infty}\frac{(n+1)!}{n!} = \lim_{n\to\infty}\frac{1\times 2\times 3\times 4\times\cdots\times n\times(n+1)}{1\times 2\times 3\times 4\times\cdots\times n}\\ &= \lim_{n\to\infty}(n+1)\\ &= +\infty\end{aligned}$$

and $\sum_{n=0}^{\infty} n!x^n$ has radius of convergence zero[39]. It converges only for $x = 0$, where it takes the value $0! = 1$.

Example 3.5.6

Example 3.5.7

Comparing the series

$$1 + 2x + x^2 + 2x^3 + x^4 + 2x^5 + \cdots$$

to

$$\sum_{n=1}^{\infty} A_n x^n = A_0 + A_1 x + A_2 x^2 + A_3 x^3 + A_4 x^4 + A_5 x^5 + \cdots$$

we see that

$$A_0 = 1 \quad A_1 = 2 \qquad A_2 = 1 \qquad A_3 = 2 \qquad A_4 = 1 \qquad A_5 = 2 \qquad \cdots$$

so that

$$\frac{A_1}{A_0} = 2 \qquad \frac{A_2}{A_1} = \frac{1}{2} \qquad \frac{A_3}{A_2} = 2 \qquad \frac{A_4}{A_3} = \frac{1}{2} \qquad \frac{A_5}{A_4} = 2 \qquad \cdots$$

and $\frac{A_{n+1}}{A_n}$ does not converge as $n \to \infty$. 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^{th} term in our series obeys

  $$\left|A_n x^n\right| \leqslant 2|x|^n$$

  and so is smaller than the n^{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.

39 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.

- Since every A_n is at least one, the n^{th} term in our series obeys

$$\left|A_n x^n\right| \geqslant |x|^n$$

  If $|x| \geqslant 1$, this $a_n = A_n x^n$ cannot converge to zero as $n \to \infty$, 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 3.5.7

Example 3.5.8

Lets construct a series from the digits of π. Now to avoid dividing by zero, let us set

$$A_n = 1 + \text{the } n^{\text{th}} \text{ digit of } \pi$$

Since $\pi = 3.141591\ldots$

$$A_0 = 4 \qquad A_1 = 2 \qquad A_2 = 5 \qquad A_3 = 2 \qquad A_4 = 6 \qquad A_5 = 10 \qquad A_6 = 2 \qquad \cdots$$

Consequently every A_n is an integer between 1 and 10 and gives us the series

$$\sum_{n=0}^{\infty} A_n x^n = 4 + 2x + 5x^2 + 2x^3 + 6x^4 + 10x^5 + \cdots$$

The number π is irrational[40] and consequently the ratio $\frac{A_{n+1}}{A_n}$ cannot have a limit as $n \to \infty$. If you do not understand why this is the case then don't worry too much about it[41]. As in 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^{th} term in our series obeys

$$\left|A_n x^n\right| \leqslant 10|x|^n$$

  and so is smaller than the n^{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.

40 We give a proof of this in the optional §3.7 at the end of this chapter.

41 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 π would be eventually periodic — something that only rational numbers do (a proof is given in the optional §3.7 at the end of this chapter).

- Since every A_n is at least one, the n^{th} term in our series obeys

$$\left|A_n x^n\right| \geqslant |x|^n$$

If $|x| \geqslant 1$, this $a_n = A_n x^n$ cannot converge to zero as $n \to \infty$, 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 3.5.8

Though we won't prove it, it is true that every power series has a radius of convergence, whether or not the limit $\lim_{n\to\infty} \left|\frac{A_{n+1}}{A_n}\right|$ exists.

Theorem 3.5.9.

Let $\sum_{n=0}^{\infty} A_n(x-c)^n$ be a power series. Then one of the following alternatives must hold.

(a) The power series converges for every number x. In this case we say that the radius of convergence is ∞.

(b) 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.

(c) 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 3.5.10.

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 3.5.9, 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
- $c-R \leqslant x < c+R$, which is also denoted $[c-R\,,\,c+R)$, or
- $c-R < x \leqslant c+R$, which is also denoted $(c-R\,,\,c+R]$, or
- $c-R \leqslant x \leqslant c+R$, 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 $x = c \pm R$. The following example shows that all four possibilities can occur.

Example 3.5.11

Let p be any real number[42] and consider the series $\sum_{n=1}^{\infty} \frac{x^n}{n^p}$. This series has $A_n = \frac{1}{n^p}$. Since

$$\lim_{n\to\infty} \left|\frac{A_{n+1}}{A_n}\right| = \lim_{n\to\infty} \frac{n^p}{(n+1)^p} = \lim_{n\to\infty} \frac{1}{(1+1/n)^p} = 1$$

the series has radius of convergence 1. So it certainly converges for $|x| < 1$ and diverges for $|x| > 1$. That just leaves $x = \pm 1$.

- When $x = 1$, the series reduces to $\sum_{n=1}^{\infty} \frac{1}{n^p}$. We know, from Example 3.3.6, that this series converges if and only if $p > 1$.
- When $x = -1$, the series reduces to $\sum_{n=1}^{\infty} \frac{(-1)^n}{n^p}$. By the alternating series test, Theorem 3.3.14, this series converges whenever $p > 0$ (so that $\frac{1}{n^p}$ tends to zero as n tends to infinity). When $p \leqslant 0$ (so that $\frac{1}{n^p}$ does *not* tend to zero as n tends to infinity), it diverges by the divergence test, Theorem 3.3.1.

So

- The power series $\sum_{n=1}^{\infty} x^n$ (i.e. $p = 0$) has interval of convergence $-1 < x < 1$.
- The power series $\sum_{n=1}^{\infty} \frac{x^n}{n}$ (i.e. $p = 1$) has interval of convergence $-1 \leqslant x < 1$.
- The power series $\sum_{n=1}^{\infty} \frac{(-1)^n}{n} x^n$ (i.e. $p = 1$) has interval of convergence $-1 < x \leqslant 1$.
- The power series $\sum_{n=1}^{\infty} \frac{x^n}{n^2}$ (i.e. $p = 2$) has interval of convergence $-1 \leqslant x \leqslant 1$.

42 We avoid problems with 0^p by starting the series from $n = 1$.

Example 3.5.11

Example 3.5.12

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

$$\sum_{n\geqslant 0} A_n (x-3)^n.$$

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 bull's eye marks the centre.

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 $|x-3| > R$.

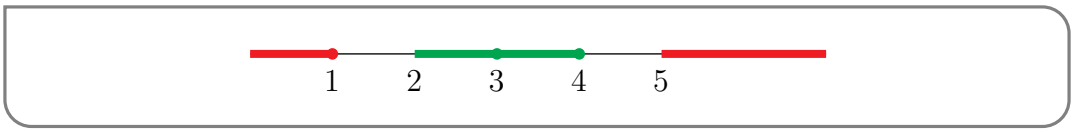
We have been told that

- the series converges when $x = 4$, which tells us that
  - $x = 4$ cannot obey $|x-3| > R$ so
  - $x = 4$ must obey $|x-3| \leqslant R$, i.e. $|4-3| \leqslant R$, i.e. $R \geqslant 1$
- the series diverges when $x = 1$ so we also know that
  - $x = 1$ cannot obey $|x-3| < R$ so
  - $x = 1$ must obey $|x-3| \geqslant R$, i.e. $|1-3| \geqslant R$, i.e. $R \leqslant 2$

We still don't know R exactly. But we do know that $1 \leqslant R \leqslant 2$. 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 3.5.12

3.5.2 ▸▸ 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 3.5.13 (Operations on Power Series).

Assume that the functions $f(x)$ and $g(x)$ are given by the power series

$$f(x) = \sum_{n=0}^{\infty} A_n(x-c)^n \qquad g(x) = \sum_{n=0}^{\infty} B_n(x-c)^n$$

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

$$\begin{aligned}
f(x)+g(x) &= \sum_{n=0}^{\infty} [A_n + B_n]\,(x-c)^n \\
Kf(x) &= \sum_{n=0}^{\infty} K\,A_n\,(x-c)^n \\
(x-c)^N f(x) &= \sum_{n=0}^{\infty} A_n\,(x-c)^{n+N} \quad \text{for any integer } N \geqslant 1 \\
&= \sum_{k=N}^{\infty} A_{k-N}\,(x-c)^k \quad \text{where } k = n+N \\
f'(x) &= \sum_{n=0}^{\infty} A_n\,n\,(x-c)^{n-1} = \sum_{n=1}^{\infty} A_n\,n\,(x-c)^{n-1} \\
\int_c^x f(t)\,\mathrm{d}t &= \sum_{n=0}^{\infty} A_n \frac{(x-c)^{n+1}}{n+1} \\
\int f(x)\,\mathrm{d}x &= \left[\sum_{n=0}^{\infty} A_n \frac{(x-c)^{n+1}}{n+1}\right] + C \quad \text{with } C \text{ an arbitrary constant}
\end{aligned}$$

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 $\sum_{n=0}^{\infty} KA_n\,(x-c)^n$ when $K = 0$.

Example 3.5.14

The last statement of Theorem 3.5.13 might seem a little odd, but consider the following two power series centred at 0:

$$\sum_{n=0}^{\infty} 2^n x^n \text{ and } \sum_{n=0}^{\infty} (1-2^n)x^n.$$

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 3.5.14

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

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \qquad \text{for all } |x| < 1$$

Example 3.5.15 $\left(\frac{1}{1-x^2}\right)$

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.

$$\begin{aligned}
\frac{1}{1-x^2} = \frac{1}{1-y}\bigg|_{y=x^2} &= \left[\sum_{n=0}^{\infty} y^n\right]_{y=x^2} \quad \text{if } |y| < 1, \text{ i.e. } |x| < 1 \\
= \sum_{n=0}^{\infty} \left(x^2\right)^n &= \sum_{n=0}^{\infty} x^{2n} \\
= 1 + x^2 + x^4 + x^6 + \cdots &
\end{aligned}$$

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

$$|y| < 1 \iff \left|x^2\right| < 1 \iff |x| < 1$$

So our power series has radius of convergence 1 and interval of convergence $-1 < x < 1$.

Example 3.5.15

Example 3.5.16 $\left(\frac{x}{2+x^2}\right)$

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.

$$\begin{aligned}
\frac{x}{2+x^2} &= \frac{x}{2}\,\frac{1}{1+x^2/2} = \frac{x}{2}\,\frac{1}{1-\left(-x^2/2\right)} && \text{set } -\frac{x^2}{2} = y \\
&= \frac{x}{2}\,\frac{1}{1-y}\bigg|_{y=-\frac{x^2}{2}} = \frac{x}{2}\left[\sum_{n=0}^{\infty} y^n\right]_{y=-\frac{x^2}{2}} && \text{if } |y| < 1 \\
&= \frac{x}{2}\sum_{n=0}^{\infty}\left(-\frac{x^2}{2}\right)^n = \frac{x}{2}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n}x^{2n} = \sum_{n=0}^{\infty}\frac{(-1)^n}{2^{n+1}}x^{2n+1} && \text{by Theorem 3.5.13, twice} \\
&= \frac{x}{2} - \frac{x^3}{4} + \frac{x^5}{8} - \frac{x^7}{16} + \cdots
\end{aligned}$$

The geometric series that we used in the second line converges when

$$|y| < 1 \iff \left|-x^2/2\right| < 1 \iff |x|^2 < 2 \iff |x| < \sqrt{2}$$

So the given power series has radius of convergence $\sqrt{2}$ and interval of convergence $-\sqrt{2} < x < \sqrt{2}$.

Example 3.5.16

Example 3.5.17 (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.

$$\begin{aligned}
\frac{1}{5-x} = \frac{1}{2-(x-3)} &= \frac{1}{2}\,\frac{1}{1-\frac{x-3}{2}} \qquad \text{set } \frac{x-3}{2} = y \\
&= \frac{1}{2}\,\frac{1}{1-y}\bigg|_{y=\frac{x-3}{2}} = \frac{1}{2}\left[\sum_{n=0}^{\infty} y^n\right]_{y=\frac{x-3}{2}} \qquad \text{if } |y|<1 \\
&= \frac{1}{2}\sum_{n=0}^{\infty}\left(\frac{x-3}{2}\right)^n = \sum_{n=0}^{\infty}\frac{(x-3)^n}{2^{n+1}} \\
&= \frac{1}{2} + \frac{x-3}{4} + \frac{(x-3)^2}{8} + \frac{(x-3)^3}{16} + \cdots
\end{aligned}$$

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 3.5.17

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 3.5.18 (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

$$f\left(Kx^k\right) = \sum_{n=0}^{\infty} A_n K^n\, x^{kn}$$

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 3.5.19 $\left(\frac{1}{(1-x)^2}\right)$

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

$$\begin{aligned}\frac{1}{(1-x)^2} &= \frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{1}{1-x}\right\} \\ &= \frac{\mathrm{d}}{\mathrm{d}x}\left\{\sum_{n=0}^{\infty} x^n\right\} \\ &= \sum_{n=1}^{\infty} n x^{n-1} \qquad \text{by Theorem 3.5.13}\end{aligned}$$

Note that the $n = 0$ term has disappeared because, for $n = 0$,

$$\frac{\mathrm{d}}{\mathrm{d}x}x^n = \frac{\mathrm{d}}{\mathrm{d}x}x^0 = \frac{\mathrm{d}}{\mathrm{d}x}1 = 0$$

Also note that the radius of convergence of this series is one. We can see this via Theorem 3.5.13. 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 $n \to \infty$ 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 3.5.19

Notice that, in this last example, we differentiated a known series to get to our answer. As per Theorem 3.5.13, 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[43]. 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 3.5.20 ($\log(1+x)$)

Find a power series representation for $\log(1+x)$.

43 Consider the power series $\sum_{n=1}^{\infty} \frac{x^n}{n}$. We know that its interval of convergence is $-1 \leqslant x < 1$. (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 $\frac{\mathrm{d}}{\mathrm{d}x}\log(1+x) = \frac{1}{1+x}$ so that $\log(1+t)$ is an antiderivative of $\frac{1}{1+t}$ and

$$\begin{aligned}\log(1+x) &= \int_0^x \frac{\mathrm{d}t}{1+t} = \int_0^x \Big[\sum_{n=0}^{\infty}(-t)^n\Big]\,\mathrm{d}t \\ &= \sum_{n=0}^{\infty}\int_0^x (-t)^n\,\mathrm{d}t \qquad \text{by Theorem 3.5.13} \\ &= \sum_{n=0}^{\infty}(-1)^n\frac{x^{n+1}}{n+1} \\ &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots\end{aligned}$$

Theorem 3.5.13 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

$$\log(1+x) = \sum_{n=0}^{\infty}(-1)^n\frac{x^{n+1}}{n+1} \qquad \text{for all} \quad -1 < x < 1$$

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 — $\log(1+(-1)) = \log 0 = -\infty$. When $x = 1$, the series converges by the alternating series test. It is possible to prove, by continuity, though we won't do so here, that the sum is $\log 2$. So the interval of convergence is $-1 < x \leqslant 1$.

Example 3.5.20

Example 3.5.21 $(\arctan x)$

Find a power series representation for $\arctan x$.

Solution. Recall that $\frac{\mathrm{d}}{\mathrm{d}x}\arctan x = \frac{1}{1+x^2}$ so that $\arctan t$ is an antiderivative of $\frac{1}{1+t^2}$ and

$$\begin{aligned}\arctan x &= \int_0^x \frac{\mathrm{d}t}{1+t^2} = \int_0^x \Big[\sum_{n=0}^{\infty}(-t^2)^n\Big]\,\mathrm{d}t = \sum_{n=0}^{\infty}\int_0^x (-1)^n t^{2n}\,\mathrm{d}t \\ &= \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{2n+1} \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots\end{aligned}$$

Theorem 3.5.13 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

$$\arctan x = \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{2n+1} \qquad \text{for all } -1 < x < 1$$

When $x = \pm 1$, the series converges by the alternating series test. So the interval of convergence is $-1 \leqslant x \leqslant 1$. It is possible to prove, though once again we won't do so here,

that when $x = \pm 1$, the series $\sum_{n=0}^{\infty}(-1)^n \frac{x^{2n+1}}{2n+1}$ converges to the value of the left hand side, $\arctan x$, at $x = \pm 1$. That is, to $\arctan(\pm 1) = \pm\frac{\pi}{4}$.

Example 3.5.21

The operations on power series dealt with in Theorem 3.5.13 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[44]. Unfortunately Theorem 3.5.13 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[45] to Taylor series.

3.6▲ Taylor Series

3.6.1 ▸▸ Extending Taylor Polynomials

Recall[46] 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 $0 \leqslant k \leqslant n$. 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

44 As always, a quick visit to your favourite search engine will direct the interested reader to more information.

45 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.

46 Please review your notes from last term if this material is feeling a little unfamiliar.

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, we saw in (3.4.33) of the CLP-1 text that, for any natural number n,

Equation 3.6.1.

$$f(x) = T_n(x) + E_n(x)$$

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 the approximation $f(x) \approx T_n(x)$. The Taylor polynomial[47] is given by the formula

Equation 3.6.1-a

$$T_n(x) = f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{n!} f^{(n)}(a)\,(x-a)^n$$

while the error satisfies[48]

Equation 3.6.1-b

$$E_n(x) = \tfrac{1}{(n+1)!} f^{(n+1)}(c)\,(x-a)^{n+1}$$

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[49].

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 $n \to \infty$, and in that case we can write $f(x)$ as the limit

$$f(x) = \lim_{n\to\infty} T_n(x) = \lim_{n\to\infty} \sum_{k=0}^{n} \tfrac{1}{k!} f^{(k)}(a)\,(x-a)^k$$

This is really a limit of partial sums, and so we can write

$$f(x) = \sum_{k=0}^{\infty} \tfrac{1}{k!} f^{(k)}(a)\,(x-a)^k$$

which is a power series representation of the function. Let us formalise this in a definition.

47 Did you take a quick look at your notes?

48 This is probably the most commonly used formula for the error. But there is another fairly commonly used formula. It, and some less commonly used formulae, are given in the next (optional) subsection "More about the Taylor Remainder".

49 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 3.6.2 (Taylor series).

The Taylor series for the function $f(x)$ expanded around a is the power series

$$\sum_{n=0}^{\infty} \tfrac{1}{n!} f^{(n)}(a)\,(x-a)^n$$

When $a = 0$ it is also called the Maclaurin series of $f(x)$. If $\lim_{n\to\infty} E_n(x) = 0$, then

$$f(x) = \sum_{n=0}^{\infty} \tfrac{1}{n!} f^{(n)}(a)\,(x-a)^n$$

Demonstrating that, for a given function, $\lim_{n\to\infty} E_n(x) = 0$ 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 3.6.3 (Exponential Series)

Find the Maclaurin series 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 $k \geqslant 0$.

We know that $\frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x$ and so

$$\begin{aligned} e^x = f(x) = f'(x) = f''(x) = \cdots = f^{(k)}(x) = \cdots \qquad & \text{which gives} \\ 1 = f(0) = f'(0) = f''(0) = \cdots = f^{(k)}(0) = \cdots . & \end{aligned}$$

Equations (3.6.1) and (3.6.1-a) then give us

$$e^x = f(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + E_n(x)$$

We shall see, in the optional Example 3.6.6 below, that, for any fixed x, $\lim\limits_{n\to\infty} E_n(x) = 0$. Consequently, for all x,

$$e^x = \lim_{n\to\infty} \left[1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{n!}x^n\right] = \sum_{n=0}^{\infty} \frac{1}{n!}x^n$$

Example 3.6.3

We have now seen power series representations for the functions

$$\frac{1}{1-x} \qquad \frac{1}{(1-x)^2} \qquad \log(1+x) \qquad \arctan(x) \qquad e^x.$$

We do not think that you, the reader, will be terribly surprised to see that we develop series for sine and cosine next.

Example 3.6.4 (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.

$$\begin{array}{llllll} f(x) = \sin x & f'(x) = \cos x & f''(x) = -\sin x & f^{(3)}(x) = -\cos x & f^{(4)}(x) = \sin x & \cdots \\ g(x) = \cos x & g'(x) = -\sin x & g''(x) = -\cos x & g^{(3)}(x) = \sin x & g^{(4)}(x) = \cos x & \cdots \end{array}$$

Now set $x = a = 0$.

$$\begin{array}{llllll} f(x) = \sin x & f(0) = 0 & f'(0) = 1 & f''(0) = 0 & f^{(3)}(0) = -1 & f^{(4)}(0) = 0 \quad \cdots \\ g(x) = \cos x & g(0) = 1 & g'(0) = 0 & g''(0) = -1 & g^{(3)}(0) = 0 & g^{(4)}(0) = 1 \quad \cdots \end{array}$$

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

$$\begin{aligned} \sin x &\approx x - \tfrac{1}{3!}x^3 + \tfrac{1}{5!}x^5 - \cdots \\ \cos x &\approx 1 - \tfrac{1}{2!}x^2 + \tfrac{1}{4!}x^4 - \cdots \end{aligned}$$

We shall see, in the optional Example 3.6.8 below, that, for both $\sin x$ and $\cos x$, we have $\lim_{n\to\infty} E_n(x) = 0$ so that

$$\begin{aligned} f(x) &= \lim_{n\to\infty} \left[f(0) + f'(0)\,x + \cdots + \tfrac{1}{n!} f^{(n)}(0)\,x^n \right] \\ g(x) &= \lim_{n\to\infty} \left[g(0) + g'(0)\,x + \cdots + \tfrac{1}{n!} g^{(n)}(0)\,x^n \right] \end{aligned}$$

Reviewing the patterns we found in the derivatives, we conclude that, for all x,

$$\begin{aligned} \sin x &= x - \tfrac{1}{3!}x^3 + \tfrac{1}{5!}x^5 - \cdots = \sum_{n=0}^{\infty} (-1)^n \tfrac{1}{(2n+1)!} x^{2n+1} \\ \cos x &= 1 - \tfrac{1}{2!}x^2 + \tfrac{1}{4!}x^4 - \cdots = \sum_{n=0}^{\infty} (-1)^n \tfrac{1}{(2n)!} x^{2n} \end{aligned}$$

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 $n \to \infty$ these ratios go to zero and consequently both series are convergent for all x. (This is very similar to what was observed in Example 3.5.5.)

Example 3.6.4

We have developed power series representations for a number of important functions[50] . Here is a theorem that summarizes them.

Theorem 3.6.5.

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots \quad \text{for all } -\infty < x < \infty$$

$$\sin(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)!}x^{2n+1} = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots \quad \text{for all } -\infty < x < \infty$$

$$\cos(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n)!}x^{2n} = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \cdots \quad \text{for all } -\infty < x < \infty$$

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + \cdots \quad \text{for all } -1 < x < 1$$

$$\log(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \quad \text{for all } -1 < x \leqslant 1$$

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots \quad \text{for all } -1 \leqslant x \leqslant 1$$

Notice that the series for sine and cosine sum to something that looks very similar to

50 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:

$$\begin{aligned}\sin(x)+\cos(x) &= \left(x-\frac{1}{3!}x^3+\frac{1}{5!}x^5-\cdots\right)+\left(1-\frac{1}{2!}x^2+\frac{1}{4!}x^4-\cdots\right)\\ &= 1+x-\frac{1}{2!}x^2-\frac{1}{3!}x^3+\frac{1}{4!}x^4+\frac{1}{5!}x^5-\cdots\\ e^x &= 1+x+\frac{1}{2!}x^2+\frac{1}{3!}x^3+\frac{1}{4!}x^4+\frac{1}{5!}x^5+\cdots\end{aligned}$$

So both series have coefficients with the same absolute value (namely $\frac{1}{n!}$), but there are differences in sign[51]. This is not a coincidence and we direct the interested reader to the optional Section 3.6.3 where will show how these series are linked through $\sqrt{-1}$.

Example 3.6.6 (Optional — Why $\sum_{n=0}^{\infty}\frac{1}{n!}x^n$ is e^x.)

We have already seen, in Example 3.6.3, that

$$e^x = 1+x+\frac{x^2}{2!}+\cdots+\frac{x^n}{n!}+E_n(x)$$

By (3.6.1-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 $n\to\infty$.

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 $x\leqslant c\leqslant 0$ and hence $e^x\leqslant e^c\leqslant e^0=1$.
- On the other hand, if $x\geqslant 0$ then $0\leqslant c\leqslant x$ and so $1=e^0\leqslant e^c\leqslant e^x$.

In either case we have that $0\leqslant e^c\leqslant 1+e^x$. Because of this the error term

$$|E_n(x)| = \left|\frac{e^c}{(n+1)!}x^{n+1}\right| \leqslant [e^x+1]\frac{|x|^{n+1}}{(n+1)!}$$

We claim that this upper bound, and hence the error $E_n(x)$, quickly shrinks to zero as $n\to\infty$.

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 $n\to\infty$, let's write it as follows.

$$e_n(x) = \frac{|x|^{n+1}}{(n+1)!} = \overbrace{\frac{|x|}{1}\cdot\frac{|x|}{2}\cdot\frac{|x|}{3}\cdots\frac{|x|}{n}\cdot\frac{|x|}{|n+1|}}^{n+1\text{ factors}}$$

51 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

$$\begin{aligned} e_n(x) &= \overbrace{\left(\frac{|x|}{1}\cdot\frac{|x|}{2}\cdot\frac{|x|}{3}\cdots\frac{|x|}{k}\right)}^{k \text{ factors}}\cdot\left(\frac{|x|}{k+1}\cdots\frac{|x|}{|n+1|}\right) \\ &\leqslant \underbrace{\left(\frac{|x|}{1}\cdot\frac{|x|}{2}\cdot\frac{|x|}{3}\cdots\frac{|x|}{k}\right)}_{=Q(x)}\cdot\left(\frac{|x|}{k+1}\right)^{n+1-k} \\ &= Q(x)\cdot\left(\frac{|x|}{k+1}\right)^{n+1-k} \end{aligned}$$

Since k does not depend not 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 $n \to \infty$ 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 $n \to \infty$.

Consequently, for all x, $\lim_{n\to\infty} E_n(x) = 0$, as claimed, and we really have

$$e^x = \lim_{n\to\infty}\left[1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{n!}x^n\right] = \sum_{n=0}^{\infty}\frac{1}{n!}x^n$$

Example 3.6.6

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 $n \to \infty$, we can show that the series satisfies the same simple differential equation[52] and the same initial condition as the function.

Example 3.6.7 (Optional — Another approach to showing that $\sum_{n=0}^{\infty}\frac{1}{n!}x^n$ is e^x.)

We already know from Example 3.5.5, 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

52 Recall, you studied that differential equation in the section on separable differential equations (Theorem 2.4.4 in Section 2.4) as well as wayyyy back in the section on exponential growth and decay in differential calculus.

initial conditions[53]. We know that $y = e^x$ satisfies

$$\frac{dy}{dx} = y \qquad \text{and} \qquad y(0) = 1$$

and by Theorem 2.4.4 (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

$$\frac{df}{dx} = f(x) \qquad \text{and} \qquad f(0) = 1.$$

- By Theorem 3.5.13,

$$\begin{aligned}\frac{df}{dx} &= \frac{d}{dx}\left\{\sum_{n=0}^{\infty}\frac{1}{n!}x^n\right\} = \sum_{n=1}^{\infty}\frac{n}{n!}x^{n-1} = \sum_{n=1}^{\infty}\frac{1}{(n-1)!}x^{n-1}\\ &= \overbrace{1}^{n=1} + \overbrace{x}^{n=2} + \overbrace{\frac{x^2}{2!}}^{n=3} + \overbrace{\frac{x^3}{3!}}^{n=4} + \cdots\\ &= f(x)\end{aligned}$$

- When we substitute $x = 0$ into the series we get (see the discussion after Definition 3.5.1)

$$f(0) = 1 + \frac{0}{1!} + \frac{0}{2!} + \cdots = 1.$$

Hence $f(x)$ solves the same initial value problem and we must have $f(x) = e^x$.

Example 3.6.7

We can show that the error terms in Maclaurin polynomials for sine and cosine go to zero as $n \to \infty$ using very much the same approach as in Example 3.6.6.

Example 3.6.8 (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 (3.6.1-b) we always have $\left|f^{(n+1)}(c)\right| \leqslant 1$ and hence

$$|E_n(x)| \leqslant \frac{|x|^{n+1}}{(n+1)!}.$$

53 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 2.4.

In Example 3.6.3, we showed that $\frac{|x|^{n+1}}{(n+1)!} \to 0$ as $n \to \infty$ — 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

$$f(x) = \lim_{n\to\infty} \left[f(0) + f'(0)\, x + \cdots + \tfrac{1}{n!} f^{(n)}(0)\, x^n \right]$$

as required.

Example 3.6.8

▸▸ Optional — More about the Taylor Remainder

In this section, we fix a real number a and a natural number n, suppose that all derivatives of the function $f(x)$ exist, and we study the error

$$\begin{aligned} E_n(a,x) &= f(x) - T_n(a,x) \\ \text{where } T_n(a,x) &= f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{n!} f^{(n)}(a)\,(x-a)^n \end{aligned}$$

made when we approximate $f(x)$ by the Taylor polynomial $T_n(a,x)$ of degree n for the function $f(x)$, expanded about a. We have already seen, in (3.6.1-b), one formula, probably the most commonly used formula, for $E_n(a,x)$. In the next theorem, we repeat that formula and give a second, commonly used, formula. After an example, we give a second theorem that contains some less commonly used formulae.

Theorem 3.6.9 (Commonly used formulae for the Taylor remainder)**.**

The Taylor remainder $E_n(a,x)$ is given by

(a) (integral form)

$$E_n(a,x) = \int_a^x \frac{1}{n!} f^{(n+1)}(t)\,(x-t)^n \,\mathrm{d}t$$

(b) (Lagrange form)

$$E_n(a,x) = \frac{1}{(n+1)!}\, f^{(n+1)}(c)\,(x-a)^{n+1}$$

for some c strictly between a and x.

Notice that the integral form of the error is explicit - we could, in principle, compute it exactly. (Of course if we could do that, we probably wouldn't need to use a Taylor expansion to approximate f.) This contrasts with the Lagrange form which is an 'existential' statement - it tells us that 'c' exists, but not how to compute it.

Proof. (a) We will give two proofs. The first is shorter and simpler, but uses some trickery. The second is longer, but is more straightforward. It uses a technique called mathematical induction.

Proof 1: We are going to use a little trickery to get a simple proof. We simply view x as being fixed and study the dependence of $E_n(a,x)$ on a. To emphasise that that is what we are doing, we define

$$S(t) = f(x) - f(t) - f'(t)\,(x-t) - \tfrac{1}{2}f''(t)\,(x-t)^2 - \cdots - \tfrac{1}{n!}f^{(n)}(t)\,(x-t)^n \qquad (*)$$

and observe that $E_n(a,x) = S(a)$.

By the fundamental theorem of calculus (Theorem 1.3.1), the function $S(t)$ is determined by its derivative, $S'(t)$, and its value at a single point. Finding a value of $S(t)$ for one value of t is easy. Substitute $t = x$ into $(*)$ to yield $S(x) = 0$. To find $S'(t)$, apply $\frac{\mathrm{d}}{\mathrm{d}t}$ to both sides of $(*)$. Recalling that x is just a constant parameter,

$$\begin{aligned} S'(t) &= 0 - f'(t) - \big[f''(t)(x-t) - f'(t)\big] - \big[\tfrac{1}{2}f^{(3)}(t)(x-t)^2 - f''(t)(x-t)\big] \\ &\qquad - \cdots - \big[\tfrac{1}{n!}f^{(n+1)}(t)\,(x-t)^n - \tfrac{1}{(n-1)!}f^{(n)}(t)\,(x-t)^{n-1}\big] \\ &= -\tfrac{1}{n!}f^{(n+1)}(t)\,(x-t)^n \end{aligned}$$

So, by the fundamental theorem of calculus, $S(x) = S(a) + \int_a^x S'(t)\,\mathrm{d}t$ and

$$\begin{aligned} E_n(a,x) &= -\big[S(x) - S(a)\big] = -\int_a^x S'(t)\,\mathrm{d}t \\ &= \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t \end{aligned}$$

Proof 2: The proof that we have just given was short, but also very tricky — almost noone could create that proof without big hints. Here is another much less tricky, but also commonly used, proof.

- First consider the case $n = 0$. When $n = 0$,

  $$E_0(a,x) = f(x) - T_0(a,x) = f(x) - f(a)$$

  The fundamental theorem of calculus gives

  $$f(x) - f(a) = \int_a^x f'(t)\,\mathrm{d}t$$

  so that

  $$E_0(a,x) = \int_a^x f'(t)\,\mathrm{d}t$$

  That is exactly the $n = 0$ case of part (a).
- Next fix any integer $n \geqslant 0$ and suppose that we already know that

  $$E_n(a,x) = \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t$$

  Apply integration by parts (Theorem 1.7.2) to this integral with

  $$u(t) = f^{(n+1)}(t) \qquad \mathrm{d}v = \frac{1}{n!}(x-t)^n\,\mathrm{d}t, \quad v(t) = -\frac{1}{(n+1)!}(x-t)^{n+1}$$

Since $v(x) = 0$, integration by parts gives

$$\begin{aligned} E_n(a,x) &= u(x)v(x) - u(a)v(a) - \int_a^x v(t)u'(t)\,\mathrm{d}t \\ &= \frac{1}{(n+1)!}f^{(n+1)}(a)\,(x-a)^{n+1} + \int_a^x \frac{1}{(n+1)!}f^{(n+2)}(t)\,(x-t)^{n+1}\,\mathrm{d}t \end{aligned} \tag{$**$}$$

Now, we defined

$$E_n(a,x) = f(x) - f(a) - f'(a)\,(x-a) - \tfrac{1}{2}f''(a)\,(x-a)^2 - \cdots - \tfrac{1}{n!}f^{(n)}(a)\,(x-a)^n$$

so

$$E_{n+1}(a,x) = E_n(a,x) - \tfrac{1}{(n+1)!}f^{(n+1)}(a)\,(x-a)^{n+1}$$

This formula expresses $E_{n+1}(a,x)$ in terms of $E_n(a,x)$. That's called a reduction formula. Combining the reduction formula with ($**$) gives

$$E_{n+1}(a,x) = \int_a^x \frac{1}{(n+1)!}f^{(n+2)}(t)\,(x-t)^{n+1}\,\mathrm{d}t$$

- Let's pause to summarise what we have learned in the last two bullets. Use the notation $P(n)$ to stand for the statement "$E_n(a,x) = \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t$". To prove part (a) of the theorem, we need to prove that the statement $P(n)$ is true for all integers $n \geqslant 0$. In the first bullet, we showed that the statement $P(0)$ is true. In the second bullet, we showed that if, for some integer $n \geqslant 0$, the statement $P(n)$ is true, then the statement $P(n+1)$ is also true. Consequently,
  - $P(0)$ is true by the first bullet and then
  - $P(1)$ is true by the second bullet with $n = 0$ and then
  - $P(2)$ is true by the second bullet with $n = 1$ and then
  - $P(3)$ is true by the second bullet with $n = 2$
  - and so on, for ever and ever.

  That tells us that $P(n)$ is true for all integers $n \geqslant 0$, which is exactly part (a) of the theorem. This proof technique is called mathematical induction[54].

(b) We have already seen one proof in the optional Section 3.4.9 of the CLP-1 text. We will see two more proofs here.

Proof 1: We apply the generalised mean value theorem, which is Theorem 3.4.38 in the CLP-1 text. It says that

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(c)}{G'(c)} \tag{GMVT}$$

54 While the use of the ideas of induction goes back over 2000 years, the first recorded rigorous use of induction appeared in the work of Levi ben Gershon (1288–1344, better known as Gersonides). The first explicit formulation of mathematical induction was given by the French mathematician Blaise Pascal in 1665.

for some c strictly[55] between a and b. We apply (GVMT) with $b = x$, $F(t) = S(t)$ and $G(t) = (x-t)^{n+1}$. This gives

$$\begin{aligned} E_n(a,x) &= -\big[S(x) - S(a)\big] = -\frac{S'(c)}{G'(c)}\big[G(x) - G(a)\big] \\ &= -\frac{-\frac{1}{n!}f^{(n+1)}(c)\,(x-c)^n}{-(n+1)(x-c)^n}\,\big[0 - (x-a)^{n+1}\big] \\ &= \frac{1}{(n+1)!}f^{(n+1)}(c)(x-a)^{n+1} \end{aligned}$$

Don't forget, when computing $G'(c)$, that G is a function of t with x just a fixed parameter.

Proof 2: We apply Theorem 2.2.10 (the mean value theorem for weighted integrals). If $a < x$, we use the weight function $w(t) = \frac{1}{n!}(x-t)^n$, which is strictly positive for all $a < t < x$. By part (a) this gives

$$\begin{aligned} E_n(a,x) &= \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t \\ &= f^{(n+1)}(c)\int_a^x \frac{1}{n!}(x-t)^n\,\mathrm{d}t \qquad \text{for some } a < c < x \\ &= f^{(n+1)}(c)\left[-\frac{1}{n!}\frac{(x-t)^{n+1}}{n+1}\right]_a^x \\ &= \frac{1}{(n+1)!}f^{(n+1)}(c)\,(x-a)^{n+1} \end{aligned}$$

If $x < a$, we instead use the weight function $w(t) = \frac{1}{n!}(t-x)^n$, which is strictly positive for all $x < t < a$. This gives

$$\begin{aligned} E_n(a,x) &= \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t = -(-1)^n\int_x^a \frac{1}{n!}f^{(n+1)}(t)\,(t-x)^n\,\mathrm{d}t \\ &= (-1)^{n+1}f^{(n+1)}(c)\int_x^a \frac{1}{n!}(t-x)^n\,\mathrm{d}t \qquad \text{for some } x < c < a \\ &= (-1)^{n+1}f^{(n+1)}(c)\left[\frac{1}{n!}\frac{(t-x)^{n+1}}{n+1}\right]_x^a \\ &= \frac{1}{(n+1)!}f^{(n+1)}(c)\,(-1)^{n+1}(a-x)^{n+1} \\ &= \frac{1}{(n+1)!}f^{(n+1)}(c)\,(x-a)^{n+1} \end{aligned}$$

□

Theorem 3.6.9 has provided us with two formulae for the Taylor remainder $E_n(a,x)$. The formula of part (b), $E_n(a,x) = \frac{1}{(n+1)!}f^{(n+1)}(c)\,(x-a)^{n+1}$, is probably the easiest to

55 In Theorem 3.4.38 in the CLP-1 text, we assumed, for simplicity, that $a < b$. To get (GVMT) when $b < a$ simply exchange a and b in Theorem 3.4.38.

use, and the most commonly used, formula for $E_n(a,x)$. The formula of part (a), $E_n(a,x) = \int_a^x \frac{1}{n!} f^{(n+1)}(t)\,(x-t)^n\,dt$, while a bit harder to apply, gives a bit better bound than that of part (b) (in the proof of Theorem 3.6.9 we showed that part (b) follows from part (a)). Here is an example in which we use both parts.

Example 3.6.10

In Theorem 3.6.5 we stated that

$$\log(1+x) = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \qquad \text{for all } -1 < x \leqslant 1 \qquad \text{(S1)}$$

But, so far, we have not justified this statement. We do so now, using (both parts of) Theorem 3.6.9. We start by setting $f(x) = \log(1+x)$ and finding the Taylor polynomials $T_n(0,x)$, and the corresponding errors $E_n(0,x)$, for $f(x)$.

$$\begin{aligned}
f(x) &= \log(1+x) & f(0) &= \log 1 = 0 \\
f'(x) &= \frac{1}{1+x} & f'(0) &= 1 \\
f''(x) &= \frac{-1}{(1+x)^2} & f''(0) &= -1 \\
f'''(x) &= \frac{2}{(1+x)^3} & f'''(0) &= 2 \\
f^{(4)}(x) &= \frac{-2\times 3}{(1+x)^4} & f^{(4)}(0) &= -3! \\
f^{(5)}(x) &= \frac{2\times 3\times 4}{(1+x)^5} & f^{(5)}(0) &= 4! \\
&\vdots & &\vdots \\
f^{(n)}(x) &= \frac{(-1)^{n+1}(n-1)!}{(1+x)^n} & f^{(n)}(0) &= (-1)^{n+1}(n-1)!
\end{aligned}$$

So the Taylor polynomial of degree n for the function $f(x) = \log(1+x)$, expanded about $a = 0$, is

$$\begin{aligned}
T_n(0,x) &= f(0) + f'(0)\,x + \cdots + \tfrac{1}{n!} f^{(n)}(0)\,x^n \\
&= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \frac{1}{5}x^5 + \cdots + \frac{(-1)^{n+1}}{n}x^n
\end{aligned}$$

Theorem 3.6.9 gives us two formulae for the error $E_n(0,x) = f(x) - T_n(0,x)$ made when we approximate $f(x)$ by $T_n(0,x)$. Part (a) of the theorem gives

$$E_n(0,x) = \int_0^x \frac{1}{n!} f^{(n+1)}(t)\,(x-t)^n\,dt = (-1)^n \int_0^x \frac{(x-t)^n}{(1+t)^{n+1}}\,dt \qquad \text{(Ea)}$$

and part (b) gives

$$E_n(0,x) = \frac{1}{(n+1)!}\,f^{(n+1)}(c)\,x^{n+1} = (-1)^n\,\frac{1}{n+1}\,\frac{x^{n+1}}{(1+c)^{n+1}} \qquad \text{(Eb)}$$

for some (unknown) c between 0 and x. The statement (S1), that we wish to prove, is equivalent to the statement

$$\lim_{n\to\infty} E_n(0,x) = 0 \qquad \text{for all } -1 < x \leqslant 1 \tag{S2}$$

and we will now show that (S2) is true.

The case $x = 0$: This case is trivial, since, when $x = 0$, $E_n(0,x) = 0$ for all n.

The case $0 < x \leqslant 1$: This case is relatively easy to deal with using (Eb). When $0 < x \leqslant 1$, the c of (Eb) must be positive, so that

$$|E_n(0,x)| = \frac{1}{n+1}\frac{x^{n+1}}{(1+c)^{n+1}} \leqslant \frac{1}{n+1}\frac{1^{n+1}}{(1+0)^{n+1}} = \frac{1}{n+1}$$

converges to zero as $n \to \infty$.

The case $-1 < x < 0$: When $-1 < x < 0$ is close to -1, (Eb) is not sufficient to show that (S2) is true. To see this, let's consider the example $x = -0.8$. All we know about the c of (Eb) is that it has to be between 0 and -0.8. For example, (Eb) certainly allows c to be -0.6 and then

$$\left|(-1)^n \frac{1}{n+1}\frac{x^{n+1}}{(1+c)^{n+1}}\right|_{\substack{x=-0.8\\c=-0.6}} = \frac{1}{n+1}\frac{0.8^{n+1}}{(1-0.6)^{n+1}} = \frac{1}{n+1}2^{n+1}$$

goes to $+\infty$ as $n \to \infty$.

Note that, while this does tell us that (Eb) is not sufficient to prove (S2), when x is close to -1, it does not also tell us that $\lim_{n\to\infty} |E_n(0,-0.8)| = +\infty$ (which would imply that (S2) is false) — c could equally well be -0.2 and then

$$\left|(-1)^n \frac{1}{n+1}\frac{x^{n+1}}{(1+c)^{n+1}}\right|_{\substack{x=-0.8\\c=-0.2}} = \frac{1}{n+1}\frac{0.8^{n+1}}{(1-0.2)^{n+1}} = \frac{1}{n+1}$$

goes to 0 as $n \to \infty$.

We'll now use (Ea) (which has the advantage of not containing any unknown free parameter c) to verify (S2) when $-1 < x < 0$. Rewrite the right hand side of (Ea)

$$\begin{aligned}(-1)^n \int_0^x \frac{(x-t)^n}{(1+t)^{n+1}}\,\mathrm{d}t &= -\int_x^0 \frac{(t-x)^n}{(1+t)^{n+1}}\,\mathrm{d}t \\ &= -\int_0^{-x} \frac{s^n}{(1+x+s)^{n+1}}\,\mathrm{d}s \quad s = t-x,\ \mathrm{d}s = \mathrm{d}t\end{aligned}$$

The exact evaluation of this integral is very messy and not very illuminating. Instead, we bound it. Note that, for $1 + x > 0$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{s}{1+x+s}\right) = \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{1+x+s-(1+x)}{1+x+s}\right) &= \frac{\mathrm{d}}{\mathrm{d}s}\left(1-\frac{1+x}{1+x+s}\right) \\ &= \frac{1+x}{(1+x+s)^2} > 0\end{aligned}$$

so that $\frac{s}{1+x+s}$ increases as s increases. Consequently, the biggest value that $\frac{s}{1+x+s}$ takes on the domain of integration $0 \leqslant s \leqslant -x = |x|$ is

$$\left.\frac{s}{1+x+s}\right|_{s=-x} = -x = |x|$$

and the integrand

$$0 \leqslant \frac{s^n}{[1+x+s]^{n+1}} = \left(\frac{s}{1+x+s}\right)^n \frac{1}{1+x+s} \leqslant \frac{|x|^n}{1+x+s}$$

Consequently,

$$\begin{aligned} |E_n(0,x)| &= \left|(-1)^n \int_0^x \frac{(x-t)^n}{(1+t)^{n+1}}\,dt\right| = \int_0^{-x} \frac{s^n}{[1+x+s]^{n+1}}\,ds \\ &\leqslant |x|^n \int_0^{-x} \frac{1}{1+x+s}\,ds = |x|^n \Big[\log(1+x+s)\Big]_{s=0}^{s=-x} \\ &= |x|^n[-\log(1+x)] \end{aligned}$$

converges to zero as $n \to \infty$ fore each fixed $-1 < x < 0$.

So we have verified (S2), as desired.

Example 3.6.10

As we said above, Theorem 3.6.9 gave the two most commonly used formulae for the Taylor remainder. Here are some less commonly used, but occasionally useful, formulae.

Theorem 3.6.11 (More formulae for the Taylor remainder).

(a) If $G(t)$ is differentiable[56] and $G'(c)$ is nonzero for all c strictly between a and x, then the Taylor remainder

$$E_n(a,x) = \frac{1}{n!} f^{(n+1)}(c)\, \frac{G(x)-G(a)}{G'(c)}\, (x-c)^n$$

for some c strictly between a and x.

(b) (Cauchy form)

$$E_n(a,x) = \frac{1}{n!} f^{(n+1)}(c)\, (x-c)^n (x-a)$$

for some c strictly between a and x.

56 Note that the function G need not be related to f. It just has to be differentiable with a nonzero derivative.

Proof. As in the proof of Theorem 3.6.9, we define

$$S(t) = f(x) - f(t) - f'(t)\,(x-t) - \tfrac{1}{2}f''(t)\,(x-t)^2 - \cdots - \tfrac{1}{n!}f^{(n)}(t)\,(x-t)^n$$

and observe that $E_n(a,x) = S(a)$ and $S(x) = 0$ and $S'(t) = -\frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n$.

(a) Recall that the generalised mean-value theorem, which is Theorem 3.4.38 in the CLP-1 text, says that

$$\frac{F(b)-F(a)}{G(b)-G(a)} = \frac{F'(c)}{G'(c)} \tag{GMVT}$$

for some c strictly between a and b. We apply this theorem with $b = x$ and $F(t) = S(t)$. This gives

$$\begin{aligned} E_n(a,x) &= -\big[S(x) - S(a)\big] = -\frac{S'(c)}{G'(c)}\big[G(x) - G(a)\big] \\ &= -\frac{-\frac{1}{n!}f^{(n+1)}(c)\,(x-c)^n}{G'(c)}\,\big[G(x) - G(a)\big] \\ &= \frac{1}{n!}f^{(n+1)}(c)\,\frac{G(x)-G(a)}{G'(c)}\,(x-c)^n \end{aligned}$$

(b) Apply part (a) with $G(x) = x$. This gives

$$\begin{aligned} E_n(a,x) &= \frac{1}{n!}f^{(n+1)}(c)\,\frac{x-a}{1}\,(x-c)^n \\ &= \frac{1}{n!}f^{(n+1)}(c)\,(x-c)^n(x-a) \end{aligned}$$

for some c strictly between a and b.

□

Example 3.6.12 (Example 3.6.10, continued)

In Example 3.6.10 we verified that

$$\log(1+x) = \sum_{n=0}^{\infty}(-1)^n\frac{x^{n+1}}{n+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \tag{S1}$$

for all $-1 < x \leqslant 1$. There we used the Lagrange form,

$$E_n(a,x) = \frac{1}{(n+1)!}\,f^{(n+1)}(c)\,(x-a)^{n+1}$$

for the Taylor remainder to verify (S1) when $0 \leqslant x \leqslant 1$, but we also saw that it is not possible to use the Lagrange form to verify (S1) when x is close to -1. We instead used the integral form

$$E_n(a,x) = \int_a^x \frac{1}{n!}f^{(n+1)}(t)\,(x-t)^n\,\mathrm{d}t$$

We will now use the Cauchy form (part (b) of Theorem 3.6.11)

$$E_n(a,x) = \frac{1}{n!} f^{(n+1)}(c)\,(x-c)^n(x-a)$$

to verify

$$\lim_{n\to\infty} E_n(0,x) = 0 \tag{S2}$$

when $-1 < x < 0$. We have already noted that (S2) is equivalent to (S1).

Write $f(x) = \log(1+x)$. We saw in Example 3.6.10 that

$$f^{(n+1)}(x) = \frac{(-1)^n n!}{(1+x)^{n+1}}$$

So, in this example, the Cauchy form is

$$E_n(0,x) = (-1)^n \frac{(x-c)^n x}{(1+c)^{n+1}}$$

for some $x < c < 0$. When $-1 < x < c < 0$,

- c and x are negative and $1+x$, $1+c$ and $c-x$ are (strictly) positive so that

$$\begin{aligned} c(1+x) < 0 &\implies c < -cx \implies c - x < -x - xc = |x|(1+c) \\ &\implies \left|\frac{x-c}{1+c}\right| = \frac{c-x}{1+c} < |x| \end{aligned}$$

  so that $\left|\frac{x-c}{1+c}\right|^n < |x|^n$ and

- the distance from -1 to c, namely $c-(-1) = 1+c$ is greater than the distance from -1 to x, namely $x-(-1) = 1+x$, so that $\frac{1}{1+c} < \frac{1}{1+x}$.

So, for $-1 < x < c < 0$,

$$|E_n(0,x)| = \left|\frac{x-c}{1+c}\right|^n \frac{|x|}{1+c} < \frac{|x|^{n+1}}{1+c} < \frac{|x|^{n+1}}{1+x}$$

goes to zero as $n \to \infty$.

Example 3.6.12

3.6.2 ▸▸ 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 π — a problem that dates back to the Babylonians.

Example 3.6.13 (Computing the number π)

There are numerous methods for computing π to any desired degree of accuracy[57]. Many of them use the Maclaurin expansion

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$$

of Theorem 3.6.5. Since $\arctan(1) = \frac{\pi}{4}$, the series gives us a very pretty formula for π:

$$\frac{\pi}{4} = \arctan 1 = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$$
$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right)$$

Unfortunately, this series is not very useful for computing π because it converges so slowly. If we approximate the series by its N^{th} partial sum, then the alternating series test (Theorem 3.3.14) tells us that the error is bounded by the first term we drop. To guarantee that we have 2 decimal digits of π correct, we need to sum about the first 200 terms!

A much better way to compute π using this series is to take advantage of the fact that $\tan \frac{\pi}{6} = \frac{1}{\sqrt{3}}$:

$$\begin{aligned}
\pi &= 6 \arctan\left(\frac{1}{\sqrt{3}}\right) = 6 \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \frac{1}{\left(\sqrt{3}\right)^{2n+1}} \\
&= 2\sqrt{3} \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \frac{1}{3^n} \\
&= 2\sqrt{3}\left(1 - \frac{1}{3\times 3} + \frac{1}{5\times 9} - \frac{1}{7\times 27} + \frac{1}{9\times 81} - \frac{1}{11\times 243} + \cdots\right)
\end{aligned}$$

Again, this is an alternating series and so (via Theorem 3.3.14) 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 π — by hand!

This is just one of very many ways to compute π. Another one, which still uses the Maclaurin expansion of $\arctan x$, but is much more efficient, is

$$\pi = 16 \arctan \frac{1}{5} - 4 \arctan \frac{1}{239}$$

57 The computation of π 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 π" by David Blatner.

This formula was used by John Machin in 1706 to compute π to 100 decimal digits — again, by hand.

Example 3.6.13

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 3.6.14 (Error function)

The *error function*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t$$

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 exponential. Start with the Maclaurin series 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 3.5.13 to integrate term-by-term:

$$\begin{aligned} \operatorname{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x \left[\sum_{n=0}^{\infty} \frac{(-t^2)^n}{n!} \right] \mathrm{d}t \\ &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)n!} \end{aligned}$$

For example, for the bell curve, the probability of being within one standard deviation of the mean[58], is

$$\begin{aligned} \operatorname{erf}\left(1/\sqrt{2}\right) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{\left(1/\sqrt{2}\right)^{2n+1}}{(2n+1)n!} = \frac{2}{\sqrt{2\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{1}{(2n+1)2^n n!} \\ &= \sqrt{\frac{2}{\pi}} \left(1 - \frac{1}{3 \times 2} + \frac{1}{5 \times 2^2 \times 2} - \frac{1}{7 \times 2^3 \times 3!} + \frac{1}{9 \times 2^4 \times 4!} - \cdots \right) \end{aligned}$$

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 3.3.14 again, an error between zero and

58 If you don't know what this means (forgive the pun) don't worry, because it is not part of the course. Standard deviation is a way of quantifying variation within a population.

the first dropped term, which is minus

$$\sqrt{\frac{2}{\pi}}\,\frac{1}{11\times 2^5\times 5!} < 2\times 10^{-5}$$

Example 3.6.14

Example 3.6.15

Evaluate

$$\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n3^n} \qquad \text{and} \qquad \sum_{n=1}^{\infty}\frac{1}{n3^n}$$

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 one of the series in Theorem 3.6.5, just with a specific value of x. The left hand given series is

$$\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\,\frac{1}{3^n} = \frac{1}{3}-\frac{1}{2}\,\frac{1}{3^2}+\frac{1}{3}\,\frac{1}{3^3}-\frac{1}{4}\,\frac{1}{3^4}+\cdots$$

The series in Theorem 3.6.5 that this most closely resembles is

$$\log(1+x) = x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}-\cdots$$

Indeed

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{(-1)^{n-1}}{n}\,\frac{1}{3^n} &= \frac{1}{3}-\frac{1}{2}\,\frac{1}{3^2}+\frac{1}{3}\,\frac{1}{3^3}-\frac{1}{4}\,\frac{1}{3^4}+\cdots\\ &= \left[x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}-\cdots\right]_{x=\frac{1}{3}}\\ &= \Big[\log(1+x)\Big]_{x=\frac{1}{3}}\\ &= \log\frac{4}{3}\end{aligned}$$

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.

$$\begin{aligned}\Big[\log(1+x)\Big]_{x=-\frac{1}{3}} &= \left[x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}-\cdots\right]_{x=-\frac{1}{3}}\\ &= -\frac{1}{3}-\frac{1}{2}\,\frac{1}{3^2}-\frac{1}{3}\,\frac{1}{3^3}-\frac{1}{4}\,\frac{1}{3^4}+\cdots\end{aligned}$$

which is exactly minus the desired right hand series. So

$$\sum_{n=1}^{\infty}\frac{1}{n3^n} = -\Big[\log(1+x)\Big]_{x=-\frac{1}{3}} = -\log\frac{2}{3} = \log\frac{3}{2}$$

Example 3.6.15

Example 3.6.16

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[59]. There is a much more efficient approach that exploits two pieces of knowledge that we have.

- From equation (3.6.1-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,

$$\sin y = y - \frac{1}{3!}y^3 + \frac{1}{5!}y^5 - \cdots$$

- Just substituting $y = 2x^3$, we have

$$\begin{aligned}\sin(2x^3) &= 2x^3 - \frac{1}{3!}(2x^3)^3 + \frac{1}{5!}(2x^3)^5 - \cdots \\ &= 2x^3 - \frac{8}{3!}x^9 + \frac{2^5}{5!}x^{15} - \cdots\end{aligned}$$

- 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 3.6.16

Example 3.6.17 (Optional — Computing the number e)

Back in Example 3.6.6, we saw that

$$e^x = 1 + x + \tfrac{x^2}{2!} + \cdots + \tfrac{x^n}{n!} + \tfrac{1}{(n+1)!}e^c x^{n+1}$$

59 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.

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

$$e = 1 + 1 + \tfrac{1}{2!} + \cdots + \tfrac{1}{n!} + \tfrac{1}{(n+1)!}e^c$$

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 in an increasing function[60] of c, and so $1 = e^0 \leqslant e^c \leqslant e^1 = e$. Thus we know that

$$\frac{1}{(n+1)!} \leqslant e - \left(1 + 1 + \tfrac{1}{2!} + \cdots + \tfrac{1}{n!}\right) \leqslant \frac{e}{(n+1)!}$$

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.

But all is not lost. Let's look a little more closely at the right-hand inequality when $n = 1$:

$$\begin{aligned} e - (1+1) &\leqslant \frac{e}{2} && \text{move the } e\text{'s to one side} \\ \frac{e}{2} &\leqslant 2 && \text{and clean it up} \\ e &\leqslant 4. \end{aligned}$$

Now this is a pretty crude bound[61] but it isn't hard to improve. Try this again with $n = 2$:

$$\begin{aligned} e - (1+1+\frac{1}{2}) &\leqslant \frac{e}{6} && \text{move } e\text{'s to one side} \\ \frac{5e}{6} &\leqslant \frac{5}{2} \\ e &\leqslant 3. \end{aligned}$$

Better. Now we can rewrite our bound:

$$\frac{1}{(n+1)!} \leqslant e - \left(1 + 1 + \tfrac{1}{2!} + \cdots + \tfrac{1}{n!}\right) \leqslant \frac{e}{(n+1)!} \leqslant \frac{3}{(n+1)!}$$

If we set $n = 4$ in this we get

$$\frac{1}{120} = \frac{1}{5!} \leqslant e - \left(1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24}\right) \leqslant \frac{3}{120}$$

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

$$\frac{1}{10!} \leqslant e - \left(1 + 1 + \frac{1}{2} + \cdots + \frac{1}{9!}\right) \leqslant \frac{3}{10!}$$

60 Check the derivative!

61 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.

Since $10! = 3628800$, the upper bound on the error is $\frac{3}{3628800} < \frac{3}{3000000} = 10^{-6}$, and we can approximate e by

$$\begin{aligned}&1+1+\tfrac{1}{2!}+\tfrac{1}{3!}+\tfrac{1}{4!}+\tfrac{1}{5!}+\tfrac{1}{6!}+\tfrac{1}{7!}+\tfrac{1}{8!}+\tfrac{1}{9!}\\ =&1+1+0.5+0.1\dot{6}+0.041\dot{6}+0.008\dot{3}+0.00138\dot{8}+0.0001984+0.0000248+0.0000028\\ =&2.718282\end{aligned}$$

and it is correct to six decimal places.

Example 3.6.17

3.6.3 ▸▸ Optional — Linking e^x with Trigonometric Functions

Let us return to the observation that we made earlier about the Maclaurin series for sine, cosine and the exponential functions:

$$\begin{aligned}\cos x + \sin x &= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \cdots\\ e^x &= 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \cdots\end{aligned}$$

We see that these series are identical except for the differences in the signs of the coefficients. Let us try to make them look even more alike by introducing extra constants A, B and q into the equations. Consider

$$\begin{aligned}A\cos x + B\sin x &= A + Bx - \frac{A}{2!}x^2 - \frac{B}{3!}x^3 + \frac{A}{4!}x^4 + \frac{B}{5!}x^5 - \cdots\\ e^{qx} &= 1 + qx + \frac{q^2}{2!}x^2 + \frac{q^3}{3!}x^3 + \frac{q^4}{4!}x^4 + \frac{q^5}{5!}x^5 + \cdots\end{aligned}$$

Let's try to choose A, B and q so that these to expressions are equal. To do so we must make sure that the coefficients of the various powers of x agree. Looking just at the coefficients of x^0 and x^1, we see that we need

$$A = 1 \qquad \text{and} \qquad B = q$$

Substituting this into our expansions gives

$$\begin{aligned}\cos x + q\sin x &= 1 + qx - \frac{1}{2!}x^2 - \frac{q}{3!}x^3 + \frac{1}{4!}x^4 + \frac{q}{5!}x^5 - \cdots\\ e^{qx} &= 1 + qx + \frac{q^2}{2!}x^2 + \frac{q^3}{3!}x^3 + \frac{q^4}{4!}x^4 + \frac{q^5}{5!}x^5 + \cdots\end{aligned}$$

Now the coefficients of x^0 and x^1 agree, but the coefficient of x^2 tells us that we need q to be a number so that $q^2 = -1$, or

$$q = \sqrt{-1}$$

We know that no such *real* number q exists. But for the moment let us see what happens if we just assume[62] that we can find q so that $q^2 = -1$. Then we will have that

$$q^3 = -q \qquad q^4 = 1 \qquad q^5 = q \qquad \cdots$$

so that the series for $\cos x + q \sin x$ and e^{qx} are identical. That is

$$e^{qx} = \cos x + q \sin x$$

If we now write this with the more usual notation $q = \sqrt{-1} = i$ we arrive at what is now known as Euler's formula

Equation 3.6.18.

$$e^{ix} = \cos x + i \sin x$$

Euler's proof of this formula (in 1740) was based on Maclaurin expansions (much like our explanation above). Euler's formula[63] is widely regarded as one of the most important and beautiful in all of mathematics.

Of course having established Euler's formula one can find slicker demonstrations. For example, let

$$f(x) = e^{-ix}(\cos x + i \sin x)$$

Differentiating (with product and chain rules and the fact that $i^2 = -1$) gives us

$$\begin{aligned} f'(x) &= -ie^{-ix}(\cos x + i \sin x) + e^{-ix}(-\sin x + i \cos x) \\ &= 0 \end{aligned}$$

Since the derivative is zero, the function $f(x)$ must be a constant. Setting $x = 0$ tells us that

$$f(0) = e^0(\cos 0 + i \sin 0) = 1.$$

Hence $f(x) = 1$ for all x. Rearranging then arrives at

$$e^{ix} = \cos x + i \sin x$$

62 We do not wish to give a primer on imaginary and complex numbers here. The interested reader can start by looking at Appendix B.

63 It is worth mentioning here that history of this topic is perhaps a little rough on Roger Cotes (1682–1716) who was one of the strongest mathematicians of his time and a collaborator of Newton. Cotes published a paper on logarithms in 1714 in which he states

$$ix = \log(\cos x + i \sin x).$$

(after translating his results into more modern notation). He proved this result by computing in two different ways the surface area of an ellipse rotated about one axis and equating the results. Unfortunately Cotes died only 2 years later at the age of 33. Upon hearing of his death Newton is supposed to have said "If he had lived, we might have known something." The reader might think this a rather weak statement, however coming from Newton it was high praise.

as required.

Substituting $x = \pi$ into Euler's formula we get Euler's identity

$$e^{i\pi} = -1$$

which is more often stated

Equation 3.6.19.

$$e^{i\pi} + 1 = 0$$

which links the 5 most important constants in mathematics, $1, 0, \pi, e$ and $\sqrt{-1}$.

3.6.4 ▸▸ 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 3.6.20

In this example, we'll start with a relatively simple limit, namely

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

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 3.6.4,

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots$$

Consequently[64]

$$\frac{\sin x}{x} = 1 - \frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \cdots$$

64 We are hiding some mathematics behind this "consequently". What we are really using is 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 "$\cdots$". 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.

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,

$$\lim_{x\to 0}\Big[-\frac{1}{3!}x^2+\frac{1}{5!}x^4-\cdots\Big]=0$$

We won't justify that statement here, but it will be justified in the following (optional) subsection. So

$$\begin{aligned}\lim_{x\to 0}\frac{\sin x}{x}&=\lim_{x\to 0}\Big[1-\frac{1}{3!}x^2+\frac{1}{5!}x^4-\cdots\Big]\\&=1+\lim_{x\to 0}\Big[-\frac{1}{3!}x^2+\frac{1}{5!}x^4-\cdots\Big]\\&=1\end{aligned}$$

Example 3.6.20

The limit in the previous example can also be evaluated relatively easily using l'Hôpital's rule[65]. While the following limit can also, in principal, be evaluated using l'Hôpital's rule, it is much more efficient to use Taylor series[66].

Example 3.6.21

In this example we evaluate

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

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 3.5.21,

$$\arctan x=x-\frac{x^3}{3}+\frac{x^5}{5}-\cdots$$

so the numerator

$$\arctan x-x=-\frac{x^3}{3}+\frac{x^5}{5}-\cdots$$

By Example 3.6.4,

$$\sin x=x-\frac{1}{3!}x^3+\frac{1}{5!}x^5-\cdots$$

65 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.

66 It takes 3 applications of l'Hôpital's rule and some careful cleaning up of the intermediate expressions. Oof!

so the denominator

$$\sin x - x = -\frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots$$

and the ratio

$$\frac{\arctan x - x}{\sin x - x} = \frac{-\frac{x^3}{3} + \frac{x^5}{5} - \cdots}{-\frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots}$$

Notice that every term in both the numerator and the denominator contains a common factor of x^3, which we can cancel out.

$$\frac{\arctan x - x}{\sin x - x} = \frac{-\frac{1}{3} + \frac{x^2}{5} - \cdots}{-\frac{1}{3!} + \frac{1}{5!}x^2 - \cdots}$$

As x tends to zero,

- 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.

$$\begin{aligned}
\lim_{x\to 0} \frac{\arctan x - x}{\sin x - x} &= \lim_{x\to 0} \frac{-\frac{1}{3} + \frac{x^2}{5} - \cdots}{-\frac{1}{3!} + \frac{1}{5!}x^2 - \cdots} \\
&= \frac{\lim_{x\to 0}\left[-\frac{1}{3} + \frac{x^2}{5} - \cdots\right]}{\lim_{x\to 0}\left[-\frac{1}{3!} + \frac{1}{5!}x^2 - \cdots\right]} \\
&= \frac{-1/3}{-1/3!} \\
&= 2
\end{aligned}$$

Example 3.6.21

3.6.5 ▸▸ Optional — The Big O Notation

In Example 3.6.20 we used, without justification[67], that, as x tends to zero, not only does every term in

$$\frac{\sin x}{x} - 1 = -\frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \cdots = \sum_{n=1}^{\infty} (-1)^n \frac{1}{(2n+1)!}x^{2n}$$

converge to zero, but in fact the sum of all infinitely many terms also converges to zero. We did something similar twice in Example 3.6.21; once in computing the limit of the numerator and once in computing the limit of the denominator.

67 Though there were a few comments in a footnote.

We'll now develop some machinery that provides the justification. We start by recalling, from equation (3.6.1), that if, for some natural number n, the function $f(x)$ has $n+1$ derivatives near the point a, then

$$f(x) = T_n(x) + E_n(x)$$

where

$$T_n(x) = f(a) + f'(a)\,(x-a) + \cdots + \tfrac{1}{n!} f^{(n)}(a)\,(x-a)^n$$

is the Taylor polynomial of degree n for the function $f(x)$ and expansion point a and

$$E_n(x) = f(x) - T_n(x) = \tfrac{1}{(n+1)!} f^{(n+1)}(c)\,(x-a)^{n+1}$$

is the error introduced when we approximate $f(x)$ by the polynomial $T_n(x)$. Here c is some unknown number between a and x. As c is not known, we do not know exactly what the error $E_n(x)$ is. But that is usually not a problem.

In the present context[68] we are interested in taking the limit as $x \to a$. So we are only interested in x-values that are very close to a, and because c lies between x and a, c is also very close to a. Now, as long as $f^{(n+1)}(x)$ is continuous at a, as $x \to a$, $f^{(n+1)}(c)$ must approach $f^{(n+1)}(a)$ which is some finite value. This, in turn, means that there must be constants $M, D > 0$ such that $\left|f^{(n+1)}(c)\right| \leqslant M$ for all c's within a distance D of a. If so, there is another constant C (namely $\frac{M}{(n+1)!}$) such that

$$\left|E_n(x)\right| \leqslant C|x-a|^{n+1} \qquad \text{whenever } |x-a| \leqslant D$$

There is some notation for this behaviour.

Definition 3.6.22 (Big O).

Let a and m be real numbers. We say that the function "$g(x)$ is of order $|x-a|^m$ near a" and we write $g(x) = O\big(|x-a|^m\big)$ if there exist constants[69] $C, D > 0$ such that

Equation 3.6.23.

$$\left|g(x)\right| \leqslant C|x-a|^m \qquad \text{whenever } |x-a| \leqslant D$$

Whenever $O\big(|x-a|^m\big)$ appears in an algebraic expression, it just stands for some (unknown) function $g(x)$ that obeys (3.6.23). This is called "big O" notation.

How should we parse the big O notation when we see it? Consider the following

$$g(x) = O(|x-3|^2)$$

68 It is worth pointing out that our Taylor series must be expanded about the point to which we are limiting — i.e. a. To work out a limit as $x \to a$ we need Taylor series expanded about a and not some other point.

69 To be precise, C and D do not depend on x, though they may, and usually do, depend on m.

First of all, we know from the definition that the notation only tells us something about $g(x)$ for x near the point a. The equation above contains "$O(|x-3|^2)$" which tells us something about what the function looks like when x is close to 3. Further, because it is "$|x-3|$" squared, it says that the graph of the function lies below a parabola $y = C(x-3)^2$ and above a parabola $y = -C(x-3)^2$ near $x = 3$. The notation doesn't tell us anything more than this — we don't know, for example, that the graph of $g(x)$ is concave up or concave down. It also tells us that Taylor expansion of $g(x)$ around $x = 3$ does not contain any constant or linear term — the first non-zero term in the expansion is of degree at least two. For example, all of the following functions are $O(|x-3|^2)$.

$$5(x-3)^2+6(x-3)^3, \qquad -7(x-3)^2-8(x-3)^4, \qquad (x-3)^3, \qquad (x-3)^{5/2}$$

In the next few examples we will rewrite a few of the Taylor polynomials that we know using this big O notation.

Example 3.6.24

Let $f(x) = \sin x$ and $a = 0$. Then

$$\begin{array}{lllll} f(x)=\sin x & f'(x)=\cos x & f''(x)=-\sin x & f^{(3)}(x)=-\cos x & f^{(4)}(x)=\sin x \quad \cdots \\ f(0)=0 & f'(0)=1 & f''(0)=0 & f^{(3)}(0)=-1 & f^{(4)}(0)=0 \quad \cdots \end{array}$$

and the pattern repeats. So every derivative is plus or minus either sine or cosine and, as we saw in previous examples, this makes analysing the error term for the sine and cosine series quite straightforward. In particular, $\left|f^{(n+1)}(c)\right| \leqslant 1$ for all real numbers c and all natural numbers n. So the Taylor polynomial of, for example, degree 3 and its error term are

$$\begin{aligned} \sin x &= x - \tfrac{1}{3!}x^3 + \tfrac{\cos c}{5!}x^5 \\ &= x - \tfrac{1}{3!}x^3 + O(|x|^5) \end{aligned}$$

under Definition 3.6.22, with $C = \frac{1}{5!}$ and any $D > 0$. Similarly, for any natural number n,

Equation 3.6.25.

$$\begin{aligned} \sin x &= x - \tfrac{1}{3!}x^3 + \tfrac{1}{5!}x^5 - \cdots + (-1)^n \tfrac{1}{(2n+1)!}x^{2n+1} + O\big(|x|^{2n+3}\big) \\ \cos x &= 1 - \tfrac{1}{2!}x^2 + \tfrac{1}{4!}x^4 - \cdots + (-1)^n \tfrac{1}{(2n)!}x^{2n} + O\big(|x|^{2n+2}\big) \end{aligned}$$

Example 3.6.24

When we studied the error in the expansion of the exponential function (way back in optional Example 3.6.6), we had to go to some length to understand the behaviour of the error term well enough to prove convergence for all numbers x. However, in the big O notation, we are free to assume that x is close to 0. Furthermore we do not need to derive

an explicit bound on the size of the coefficient C. This makes it quite a bit easier to verify that the big O notation is correct.

Example 3.6.26

Let n be any natural number. Since $\frac{\mathrm{d}}{\mathrm{d}x}e^x = e^x$, we know that $\frac{\mathrm{d}^k}{\mathrm{d}x^k}\{e^x\} = e^x$ for every integer $k \geqslant 0$. Thus

$$e^x = 1 + x + \tfrac{x^2}{2!} + \tfrac{x^3}{3!} + \cdots + \tfrac{x^n}{n!} + \tfrac{e^c}{(n+1)!}x^{n+1}$$

for some c between 0 and x. If, for example, $|x| \leqslant 1$, then $|e^c| \leqslant e$, so that the error term

$$\left|\tfrac{e^c}{(n+1)!}x^{n+1}\right| \leqslant C|x|^{n+1} \qquad \text{with } C = \tfrac{e}{(n+1)!} \qquad \text{whenever } |x| \leqslant 1$$

So, under Definition 3.6.22, with $C = \frac{e}{(n+1)!}$ and $D = 1$,

Equation 3.6.27.

$$e^x = 1 + x + \tfrac{x^2}{2!} + \tfrac{x^3}{3!} + \cdots + \tfrac{x^n}{n!} + O\big(|x|^{n+1}\big)$$

You can see that, because we only have to consider x's that are close to the expansion point (in this example, 0) it is relatively easy to derive the bounds that are required to justify the use of the big O notation.

Example 3.6.26

Example 3.6.28

Let $f(x) = \log(1+x)$ and $a = 0$. Then

$$f'(x) = \tfrac{1}{1+x} \quad f''(x) = -\tfrac{1}{(1+x)^2} \quad f^{(3)}(x) = \tfrac{2}{(1+x)^3} \quad f^{(4)}(x) = -\tfrac{2\times3}{(1+x)^4} \quad f^{(5)}(x) = \tfrac{2\times3\times4}{(1+x)^5}$$
$$f'(0) = 1 \quad f''(0) = -1 \quad f^{(3)}(0) = 2 \quad f^{(4)}(0) = -3! \quad f^{(5)}(0) = 4!$$

We can see a pattern for $f^{(n)}(x)$ forming here — $f^{(n)}(x)$ is a sign times a ratio with

- the sign being $+$ when n is odd and being $-$ when n is even. So the sign is $(-1)^{n-1}$.
- The denominator is $(1+x)^n$.
- The numerator[70] is the product $2 \times 3 \times 4 \times \cdots \times (n-1) = (n-1)!$.

70 Remember that $n! = 1 \times 2 \times 3 \times \cdots \times n$, and that we use the convention $0! = 1$.

Thus[71], for any natural number n,

$$f^{(n)}(x) = (-1)^{n-1}\tfrac{(n-1)!}{(1+x)^n} \qquad \text{which means that}$$
$$\tfrac{1}{n!}f^{(n)}(0)\,x^n = (-1)^{n-1}\tfrac{(n-1)!}{n!}x^n = (-1)^{n-1}\tfrac{x^n}{n}$$

so

$$\log(1+x) = x - \tfrac{x^2}{2} + \tfrac{x^3}{3} - \cdots + (-1)^{n-1}\tfrac{x^n}{n} + E_n(x)$$

with

$$E_n(x) = \tfrac{1}{(n+1)!}f^{(n+1)}(c)\,(x-a)^{n+1} = \tfrac{1}{n+1}\cdot\tfrac{(-1)^n}{(1+c)^{n+1}}\cdot x^{n+1}$$

If we choose, for example $D = \frac{1}{2}$, then[72] for any x obeying $|x| \leqslant D = \frac{1}{2}$, we have $|c| \leqslant \frac{1}{2}$ and $|1+c| \geqslant \frac{1}{2}$ so that

$$|E_n(x)| \leqslant \tfrac{1}{(n+1)(1/2)^{n+1}}|x|^{n+1} = O\big(|x|^{n+1}\big)$$

under Definition 3.6.22, with $C = \frac{2^{n+1}}{n+1}$ and $D = \frac{1}{2}$. Thus we may write

Equation 3.6.29.

$$\log(1+x) = x - \tfrac{x^2}{2} + \tfrac{x^3}{3} - \cdots + (-1)^{n-1}\tfrac{x^n}{n} + O\big(|x|^{n+1}\big)$$

Example 3.6.28

Remark 3.6.30. The big O notation has a few properties that are useful in computations and taking limits. All follow immediately from Definition 3.6.22.

(a) If $p > 0$, then

$$\lim_{x\to 0} O(|x|^p) = 0$$

(b) For any real numbers p and q,

$$O(|x|^p)\,O(|x|^q) = O(|x|^{p+q})$$

71 It is not too hard to make this rigorous using the principle of mathematical induction. The interested reader should do a little search-engine-ing. Induction is a very standard technique for proving statements of the form "For every natural number n,...". For example

$$\text{For every natural number } n,\ \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \qquad \text{or}$$
$$\text{For every natural number } n,\ \frac{\mathrm{d}^n}{\mathrm{d}x^n}\{\log(1+x)\} = (-1)^{n-1}\frac{(n-1)!}{(1+x)^n}$$

It was also used by Polya (1887–1985) to give a very convincing (but subtly (and deliberately) flawed) proof that all horses have the same colour.

72 Since $|c| \leqslant \frac{1}{2}$, $-\frac{1}{2} \leqslant c \leqslant \frac{1}{2}$. If we now add 1 to every term we get $\frac{1}{2} \leqslant 1+c \leqslant \frac{3}{2}$ and so $|1+c| \geqslant \frac{1}{2}$. You can also do this with the triangle inequality which tells us that for any x, y we know that $|x+y| \leqslant |x| + |y|$. Actually, you want the reverse triangle inequality (which is a simple corollary of the triangle inequality) which says that for any x, y we have $|x+y| \geqslant \big||x| - |y|\big|$.

(This is just because $C|x|^p \times C'|x|^q = (CC')|x|^{p+q}$.) In particular,

$$ax^m\, O(|x|^p) = O(|x|^{p+m})$$

for any constant a and any integer m.

(c) For any real numbers p and q,

$$O(|x|^p) + O(|x|^q) = O(|x|^{\min\{p,q\}})$$

(For example, if $p = 2$ and $q = 5$, then $C|x|^2 + C'|x|^5 = \big(C + C'|x|^3\big)|x|^2 \leqslant (C + C')|x|^2$ whenever $|x| \leqslant 1$.)

(d) For any real numbers p and q with $p > q$, any function which is $O(|x|^p)$ is also $O(|x|^q)$ because $C|x|^p = C|x|^{p-q}|x|^q \leqslant C|x|^q$ whenever $|x| \leqslant 1$.

(e) All of the above observations also hold for more general expressions with $|x|$ replaced by $|x - a|$, i.e. for $O(|x - a|^p)$. The only difference being in (a) where we must take the limit as $x \to a$ instead of $x \to 0$.

3.6.6 ▸▸ Optional — Evaluating Limits Using Taylor Expansions — More Examples

Example 3.6.31 (Example 3.6.20 revisited)

In this example, we'll return to the limit

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

of Example 3.6.20 and treat it more carefully. By Example 3.6.24,

$$\sin x = x - \frac{1}{3!}x^3 + O(|x|^5)$$

That is, for small x, $\sin x$ is the same as $x - \frac{1}{3!}x^3$, up to an error that is bounded by some constant times $|x|^5$. So, dividing by x, $\frac{\sin x}{x}$ is the same as $1 - \frac{1}{3!}x^2$, up to an error that is bounded by some constant times x^4 — see Remark 3.6.30(b). That is

$$\frac{\sin x}{x} = 1 - \frac{1}{3!}x^2 + O(x^4)$$

But any function that is bounded by some constant times x^4 (for all x smaller than some constant $D > 0$) necessarily tends to 0 as $x \to 0$ — see Remark 3.6.30(a). . Thus

$$\lim_{x\to 0} \frac{\sin x}{x} = \lim_{x\to 0} \Big[1 - \frac{1}{3!}x^2 + O(x^4)\Big] = \lim_{x\to 0} \Big[1 - \frac{1}{3!}x^2\Big] = 1$$

Reviewing the above computation, we see that we did a little more work than we had to. It wasn't necessary to keep track of the $-\frac{1}{3!}x^3$ contribution to $\sin x$ so carefully. We could have just said that

$$\sin x = x + O(|x|^3)$$

so that

$$\lim_{x\to 0}\frac{\sin x}{x}=\lim_{x\to 0}\frac{x+O(|x|^3)}{x}=\lim_{x\to 0}\left[1+O(x^2)\right]=1$$

We'll spend a little time in the later, more complicated, examples learning how to choose the number of terms we keep in our Taylor expansions so as to make our computations as efficient as possible.

Example 3.6.31

Example 3.6.32

In this example, we'll use the Taylor polynomial of Example 3.6.28 to evaluate $\lim_{x\to 0}\frac{\log(1+x)}{x}$ and $\lim_{x\to 0}(1+x)^{a/x}$. The Taylor expansion of equation (3.6.29) with $n=1$ tells us that

$$\log(1+x)=x+O(|x|^2)$$

That is, for small x, $\log(1+x)$ is the same as x, up to an error that is bounded by some constant times x^2. So, dividing by x, $\frac{1}{x}\log(1+x)$ is the same as 1, up to an error that is bounded by some constant times $|x|$. That is

$$\frac{1}{x}\log(1+x)=1+O(|x|)$$

But any function that is bounded by some constant times $|x|$, for all x smaller than some constant $D>0$, necessarily tends to 0 as $x\to 0$. Thus

$$\lim_{x\to 0}\frac{\log(1+x)}{x}=\lim_{x\to 0}\frac{x+O(|x|^2)}{x}=\lim_{x\to 0}\left[1+O(|x|)\right]=1$$

We can now use this limit to evaluate

$$\lim_{x\to 0}(1+x)^{a/x}.$$

Now, we could either evaluate the limit of the logarithm of this expression, or we can carefully rewrite the expression as $e^{(\text{something})}$. Let us do the latter.

$$\begin{aligned}\lim_{x\to 0}(1+x)^{a/x}&=\lim_{x\to 0}e^{a/x\log(1+x)}\\&=\lim_{x\to 0}e^{\frac{a}{x}[x+O(|x|^2)]}\\&=\lim_{x\to 0}e^{a+O(|x|)}=e^a\end{aligned}$$

Here we have used that if $F(x)=O(|x|^2)$ then $\frac{a}{x}F(x)=O(x)$ — see Remark 3.6.30(b). We have also used that the exponential is continuous — as x tends to zero, the exponent of $e^{a+O(|x|)}$ tends to a so that $e^{a+O(|x|)}$ tends to e^a — see Remark 3.6.30(a).

Example 3.6.32

Example 3.6.33

In this example, we'll evaluate[73] the harder limit

$$\lim_{x\to 0} \frac{\cos x - 1 + \frac{1}{2}x\sin x}{[\log(1+x)]^4}$$

The first thing to notice about this limit is that, as x tends to zero, the numerator

$$\cos x - 1 + \tfrac{1}{2}x\sin x \to \cos 0 - 1 + \tfrac{1}{2}\cdot 0 \cdot \sin 0 = 0$$

and the denominator

$$[\log(1+x)]^4 \to [\log(1+0)]^4 = 0$$

too. So both the numerator and denominator tend to zero and we may not simply evaluate the limit of the ratio by taking the limits of the numerator and denominator and dividing.

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 develop a strategy for evaluating this limit, let's do a "little scratch work", starting by taking a closer look at the denominator. By Example 3.6.28,

$$\log(1+x) = x + O(x^2)$$

This tells us that $\log(1+x)$ looks a lot like x for very small x. So the denominator $[x + O(x^2)]^4$ looks a lot like x^4 for very small x. Now, what about the numerator?

- If the numerator looks like some constant times x^p with $p > 4$, for very small x, then the ratio will look like the constant times $\frac{x^p}{x^4} = x^{p-4}$ and, as $p-4 > 0$, will tend to 0 as x tends to zero.
- If the numerator looks like some constant times x^p with $p < 4$, for very small x, then the ratio will look like the constant times $\frac{x^p}{x^4} = x^{p-4}$ and will, as $p-4 < 0$, tend to infinity, and in particular diverge, as x tends to zero.
- If the numerator looks like Cx^4, for very small x, then the ratio will look like $\frac{Cx^4}{x^4} = C$ and will tend to C as x tends to zero.

The moral of the above "scratch work" is that we need to know the behaviour of the numerator, for small x, up to order x^4. Any contributions of order x^p with $p > 4$ may be put into error terms $O(|x|^p)$.

Now we are ready to evaluate the limit. Because the expressions are a little involved, we will simplify the numerator and denominator separately and then put things together.

73 Use of l'Hôpital's rule here could be characterised as a "courageous decision". The interested reader should search-engine their way to Sir Humphrey Appleby and 'Yes Minister" to better understand this reference (and the workings of government in the Westminster system). Discretion being the better part of valour, we'll stop and think a little before limiting (ha) our choices.

Using the expansions we developed in Example 3.6.24, the numerator,

$$\begin{aligned}
\cos x - 1 + \frac{1}{2}x\sin x &= \left(1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 + O(|x|^6)\right) \\
&\quad -1 + \frac{x}{2}\left(x - \frac{1}{3!}x^3 + O(|x|^5)\right) && \text{expand} \\
&= \left(\frac{1}{24} - \frac{1}{12}\right)x^4 + O(|x|^6) + \frac{x}{2}O(|x|^5) \\
&= -\frac{1}{24}x^4 + O(|x|^6) + O(|x|^6) && \text{by Remark 3.6.30(b)} \\
&= -\frac{1}{24}x^4 + O(|x|^6) && \text{by Remark 3.6.30(c)}
\end{aligned}$$

Similarly, using the expansion that we developed in Example 3.6.28,

$$\begin{aligned}
[\log(1+x)]^4 &= \left[x + O(|x|^2)\right]^4 \\
&= \left[x + xO(|x|)\right]^4 && \text{by Remark 3.6.30(b)} \\
&= x^4[1 + O(|x|)]^4
\end{aligned}$$

Now put these together and take the limit as $x \to 0$:

$$\begin{aligned}
\lim_{x\to 0}\frac{\cos x - 1 + \frac{1}{2}x\sin x}{[\log(1+x)]^4} &= \lim_{x\to 0}\frac{-\frac{1}{24}x^4 + O(|x|^6)}{x^4[1+O(|x|)]^4} \\
&= \lim_{x\to 0}\frac{-\frac{1}{24}x^4 + x^4O(|x|^2)}{x^4[1+O(|x|)]^4} && \text{by Remark 3.6.30(b)} \\
&= \lim_{x\to 0}\frac{-\frac{1}{24} + O(|x|^2)}{[1+O(|x|)]^4} \\
&= -\frac{1}{24} && \text{by Remark 3.6.30(a).}
\end{aligned}$$

Example 3.6.33

The next two limits have much the same flavour as those above — expand the numerator and denominator to high enough order, do some cancellations and then take the limit. We have increased the difficulty a little by introducing "expansions of expansions".

Example 3.6.34

In this example we'll evaluate another harder limit, namely

$$\lim_{x\to 0}\frac{\log\left(\frac{\sin x}{x}\right)}{x^2}$$

The first thing to notice about this limit is that, as x tends to zero, the denominator x^2 tends to 0. So, yet again, to find the limit, we are going to have to show that the numerator also

tends to 0 and we are going to have to exhibit a cancellation between the numerator and the denominator.

Because the denominator is x^2 any terms in the numerator, $\log\left(\frac{\sin x}{x}\right)$ that are of order x^3 or higher will contribute terms in the ratio $\frac{\log(\frac{\sin x}{x})}{x^2}$ that are of order x or higher. Those terms in the ratio will converge to zero as $x \to 0$. The moral of this discussion is that we need to compute $\log \frac{\sin x}{x}$ to order x^2 with errors of order x^3. Now we saw, in Example 3.6.31, that

$$\frac{\sin x}{x} = 1 - \frac{1}{3!}x^2 + O(x^4)$$

We also saw, in equation (3.6.29) with $n = 1$, that

$$\log(1+X) = X + O(X^2)$$

Substituting[74] $X = -\frac{1}{3!}x^2 + O(x^4)$, and using that $X^2 = O(x^4)$ (by Remark 3.6.30(b,c)), we have that the numerator

$$\log\left(\frac{\sin x}{x}\right) = \log(1+X) = X + O(X^2) = -\frac{1}{3!}x^2 + O(x^4)$$

and the limit

$$\lim_{x\to 0} \frac{\log\left(\frac{\sin x}{x}\right)}{x^2} = \lim_{x\to 0} \frac{-\frac{1}{3!}x^2 + O(x^4)}{x^2} = \lim_{x\to 0}\Big[-\frac{1}{3!} + O(x^2)\Big] = -\frac{1}{3!} = -\frac{1}{6}$$

Example 3.6.34

Example 3.6.35

Evaluate

$$\lim_{x\to 0} \frac{e^{x^2} - \cos x}{\log(1+x) - \sin x}$$

Solution.

Step 1: Find the limit of the denominator.

$$\lim_{x\to 0}\big[\log(1+x) - \sin x\big] = \log(1+0) - \sin 0 = 0$$

This tells us that we can't evaluate the limit just by finding the limits of the numerator and denominator separately and then dividing.

Step 2: Determine the leading order behaviour of the denominator near $x = 0$. By equations (3.6.29) and (3.6.25),

$$\log(1+x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \cdots$$
$$\sin x = x - \tfrac{1}{3!}x^3 + \tfrac{1}{5!}x^5 - \cdots$$

74 In our derivation of $\log(1+X) = X + O(X^2)$ in Example 3.6.28, we required only that $|X| \leqslant \frac{1}{2}$. So we are free to substitute $X = -\frac{1}{3!}x^2 + O(x^4)$ for any x that is small enough that $\big|-\frac{1}{3!}x^2 + O(x^4)\big| < \frac{1}{2}$.

Taking the difference of these expansions gives

$$\log(1+x) - \sin x = -\tfrac{1}{2}x^2 + \left(\tfrac{1}{3} + \tfrac{1}{3!}\right)x^3 + \cdots$$

This tells us that, for x near zero, the denominator is $-\frac{x^2}{2}$ (that's the leading order term) plus contributions that are of order x^3 and smaller. That is

$$\log(1+x) - \sin x = -\tfrac{x^2}{2} + O(|x|^3)$$

Step 3: Determine the behaviour of the numerator near $x = 0$ to order x^2 with errors of order x^3 and smaller (just like the denominator). By equation (3.6.27)

$$e^X = 1 + X + O\big(X^2\big)$$

Substituting $X = x^2$

$$e^{x^2} = 1 + x^2 + O\big(x^4\big)$$
$$\cos x = 1 - \tfrac{1}{2}x^2 + O\big(x^4\big)$$

by equation (3.6.25). Subtracting, the numerator

$$e^{x^2} - \cos x = \tfrac{3}{2}x^2 + O\big(x^4\big)$$

Step 4: Evaluate the limit.

$$\lim_{x\to 0} \frac{e^{x^2} - \cos x}{\log(1+x) - \sin x} = \lim_{x\to 0} \frac{\frac{3}{2}x^2 + O(x^4)}{-\frac{x^2}{2} + O(|x|^3)} = \lim_{x\to 0} \frac{3/2 + O(x^2)}{-1/2 + O(|x|)} = \frac{3/2}{-1/2} = -3$$

Example 3.6.35

3.7▲ Optional — Rational and Irrational Numbers

In this optional section we shall use series techniques to look a little at rationality and irrationality of real numbers. We shall see the following results.

- A real number is rational (i.e. a ratio of two integers) if and only if its decimal expansion is eventually periodic. "Eventually periodic" means that, if we denote the n^{th} decimal place by d_n, then there are two positive integers k and p such that $d_{n+p} = d_n$ whenever $n > k$. So the part of the decimal expansion after the decimal point looks like

  $$.\underbrace{a_1a_2a_3\cdots a_k}\,\underbrace{b_1b_2\cdots b_p}\,\underbrace{b_1b_2\cdots b_p}\,\underbrace{b_1b_2\cdots b_p}\cdots$$

  It is possible that a finite number of decimal places right after the decimal point do not participate in the periodicity. It is also possible that $p = 1$ and $b_1 = 0$, so that the decimal expansion ends with an infinite string of zeros.
- e is irrational.
- π is irrational.

▸▸ Decimal Expansions of Rational Numbers

We start by showing that a real number is rational if and only if its decimal expansion is eventually periodic. We need only consider the expansions of numbers $0 < x < 1$. If a number is negative then we can just multiply it by -1 and not change the expansion. Similarly if the number is larger than 1 then we can just subtract off the integer part of the number and leave the expansion unchanged.

▸▸▸ Eventually Periodic Implies Rational

Let us assume that a number $0 < x < 1$ has a decimal expansion that is eventually periodic. Hence we can write

$$x = 0.\underbrace{a_1a_2a_3\cdots a_k}\,\underbrace{b_1b_2\cdots b_p}\,\underbrace{b_1b_2\cdots b_p}\,\underbrace{b_1b_2\cdots b_p}\cdots$$

Let $\alpha = a_1a_2a_3\cdots a_k$ and $\beta = b_1b_2\cdots b_p$. In particular, α has at most k digits and β has at most p digits. Then we can (carefully) write

$$\begin{aligned} x &= \frac{\alpha}{10^k} + \frac{\beta}{10^{k+p}} + \frac{\beta}{10^{k+2p}} + \frac{\beta}{10^{k+3p}} + \cdots \\ &= \frac{\alpha}{10^k} + \frac{\beta}{10^{k+p}} \sum_{j=0}^{\infty} 10^{-p} \end{aligned}$$

This sum is just a geometric series (see Example 3.2.4) and we can evaluate it:

$$\begin{aligned} &= \frac{\alpha}{10^k} + \frac{\beta}{10^{k+p}} \cdot \frac{1}{1-10^{-p}} = \frac{\alpha}{10^k} + \frac{\beta}{10^k} \cdot \frac{1}{10^p - 1} \\ &= \frac{1}{10^k}\left(\alpha + \frac{\beta}{10^p - 1}\right) = \frac{\alpha(10^p-1)+\beta}{10^k(10^p-1)} \end{aligned}$$

This is a ratio of integers, so x is a rational number.

▸▸▸ Rational Implies Eventually Periodic

Let $0 < x < 1$ be rational with $x = \frac{a}{b}$, where a and b are positive integers. We wish to show that x's decimal expansion is eventually periodic. Start by looking at the last formula we derived in the "eventually periodic implies rational" subsection. If we can express the denominator b in the form $\frac{10^k(10^p-1)}{q}$ with k, p and q integers, we will be in business because $\frac{a}{b} = \frac{aq}{10^k(10^p-1)}$. From this we can generate the desired decimal expansion by running the argument of the last subsection backwards. So we want to find integers k, p, q such that $10^{k+p} - 10^k = b \cdot q$. To do so consider the powers of 10 up to 10^b:

$$1, 10^1, 10^2, 10^3, \ldots, 10^b$$

For each $j = 0, 1, 2, \ldots, b$, find integers c_j and $0 \leqslant r_j < b$ so that

$$10^j = b \cdot c_j + r_j$$

To do so, start with 10^j and repeatedly subtract b from it until the remainder drops strictly below b. The r_j's can take at most b different values, namely $0, 1, 2, \cdots, b-1$, and we now have $b+1$ r_j's, namely $r_0, r_1, \cdots, r_b$. So we must be able to find two powers of 10 which give the same remainder[75]. That is there must be $0 \leqslant k < l \leqslant b$ so that $r_k = r_l$. Hence

$$\begin{aligned} 10^l - 10^k &= (bc_l + r_l) - (bc_k + r_k) \\ &= b(c_l - c_k) \qquad \text{since } r_k = r_l. \end{aligned}$$

and we have

$$b = \frac{10^k(10^p - 1)}{q}$$

where $p = l - k$ and $q = c_l - c_k$ are both strictly positive integers, since $l > k$ so that $10^l - 10^k > 0$. Thus we can write

$$\frac{a}{b} = \frac{aq}{10^k(10^p - 1)}$$

Next divide the numerator aq by $10^p - 1$ and compute the remainder. That is, write $aq = \alpha(10^p - 1) + \beta$ with $0 \leqslant \beta < 10^p - 1$. Notice that $0 \leqslant \alpha < 10^k$, as otherwise $x = \frac{a}{b} \geqslant 1$. That is, α has at most k digits and β has at most p digits. This, finally, gives us

$$\begin{aligned} x = \frac{a}{b} &= \frac{\alpha(10^p - 1) + \beta}{10^k(10^p - 1)} \\ &= \frac{\alpha}{10^k} + \frac{\beta}{10^k(10^p - 1)} \\ &= \frac{\alpha}{10^k} + \frac{\beta}{10^{k+p}(1 - 10^{-p})} \\ &= \frac{\alpha}{10^k} + \frac{\beta}{10^{k+p}} \sum_{j=0}^{\infty} 10^{-pj} \end{aligned}$$

which gives the required eventually periodic expansion.

▸▸ Irrationality of e

We will give 2 proofs that the number e is irrational, the first due to Fourier (1768–1830) and the second due to Pennisi (1918–2010). Both are proofs by contradiction[76] — we first assume that e is rational and then show that this implies a contradiction. In both cases we reach the contradiction by showing that a given quantity (related to the series expression for e) must be both a positive integer and also strictly less than 1.

75 This is an application of the pigeon hole principle — the very simple but surprisingly useful idea that if you have n items which you have to put in m boxes, and if $n > m$, then at least one box must contain more than one item.

76 Proof by contradiction is a standard and very powerful method of proof in mathematics. It relies on the law of the excluded middle which states that any given mathematical statement P is either true or false. Because of this, if we can show that the statement P being false implies something contradictory — like $1 = 0$ or $a > a$ — then we can conclude that P must be true. The interested reader can certainly find many examples (and a far more detailed explanation) using their favourite search engine.

▸▸▸ Proof 1

This proof is due to Fourier. Let us assume that the number e is rational so we can write it as

$$e = \frac{a}{b}$$

where a, b are positive integers. Using the Maclaurin series for e^x we have

$$\frac{a}{b} = e^1 = \sum_{n=0}^{\infty} \frac{1}{n!}$$

Now multiply both sides by $b!$ to get

$$a\frac{b!}{b} = \sum_{n=0}^{\infty} \frac{b!}{n!}$$

The left-hand side of this expression is an integer. We complete the proof by showing that the right-hand side cannot be an integer (and hence that we have a contradiction).

First split the series on the right-hand side into two piece as follows

$$\sum_{n=0}^{\infty} \frac{b!}{n!} = \underbrace{\sum_{n=0}^{b} \frac{b!}{n!}}_{=A} + \underbrace{\sum_{n=b+1}^{\infty} \frac{b!}{n!}}_{=B}$$

The first sum, A, is finite sum of integers:

$$A = \sum_{n=0}^{b} \frac{b!}{n!} = \sum_{n=0}^{b} (n+1)(n+2)\cdots(b-1)b.$$

Consequently A must be an integer. Notice that we simplified the ratio of factorials using the fact that when $b \geqslant n$ we have

$$\frac{b!}{n!} = \frac{1 \cdot 2 \cdots n(n+1)(n+2)\cdots(b-1)b}{1 \cdot 2 \cdots n} = (n+1)(n+2)\cdots(b-1)b.$$

Now we turn to the second sum. Since it is a sum of strictly positive terms we must have

$$B > 0$$

We complete the proof by showing that $B < 1$. To do this we bound each term from above:

$$\begin{aligned} \frac{b!}{n!} &= \frac{1}{\underbrace{(b+1)(b+2)\cdots(n-1)n}_{n-b \text{ factors}}} \\ &\leqslant \frac{1}{\underbrace{(b+1)(b+1)\cdots(b+1)(b+1)}_{n-b \text{ factors}}} = \frac{1}{(b+1)^{n-b}} \end{aligned}$$

Indeed the inequality is strict except when $n = b + 1$. Hence we have that

$$\begin{aligned} B &< \sum_{n=b+1}^{\infty} \frac{1}{(b+1)^{n-b}} \\ &= \frac{1}{(b+1)} + \frac{1}{(b+1)^2} + \frac{1}{(b+1)^3} + \cdots \end{aligned}$$

This is just a geometric series (see Example 3.2.4) and equals

$$\begin{aligned} &= \frac{1}{(b+1)} \frac{1}{1 - \frac{1}{b+1}} \\ &= \frac{1}{b+1-1} = \frac{1}{b} \end{aligned}$$

And since b is a positive integer, we have shown that

$$0 < B < 1$$

and thus B cannot be an integer.

Thus we have that

$$\underbrace{a\frac{b!}{b}}_{\text{integer}} = \underbrace{A}_{\text{integer}} + \underbrace{B}_{\text{not integer}}$$

which gives a contradiction. Thus e cannot be rational.

▸▸▸ Proof 2

This proof is due to Pennisi (1953). Let us (again) assume that the number e is rational. Hence it can be written as

$$e = \frac{a}{b},$$

where a, b are positive integers. This means that we can write

$$e^{-1} = \frac{b}{a}.$$

Using the Maclaurin series for e^x we have

$$\frac{b}{a} = e^{-1} = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!}$$

Before we do anything else, we multiply both sides by $(-1)^{a+1}a!$ — this might seem a little strange at this point, but the reason will become clear as we proceed through the proof. The expression is now

$$(-1)^{a+1} b \frac{a!}{a} = \sum_{n=0}^{\infty} \frac{(-1)^{n+a+1} a!}{n!}$$

The left-hand side of the expression is an integer. We again complete the proof by showing that the right–hand side cannot be an integer.

We split the series on the right-hand side into two pieces:

$$\sum_{n=0}^{\infty} \frac{(-1)^{n+a+1}a!}{n!} = \underbrace{\sum_{n=0}^{a} \frac{(-1)^{n+a+1}a!}{n!}}_{=A} + \underbrace{\sum_{n=a+1}^{\infty} \frac{(-1)^{n+a+1}a!}{n!}}_{=B}$$

We will show that A is an integer while $0 < B < 1$; this gives the required contradiction.

Every term in the sum A is an integer. To see this we simplify the ratio of factorials as we did in the previous proof:

$$A = \sum_{n=0}^{a} \frac{(-1)^{n+a+1}a!}{n!} = \sum_{n=0}^{a} (-1)^{n+a+1}(n+1)(n+2)\cdots(a-1)a$$

Let us now examine the series B. Again clean up the ratio of factorials:

$$\begin{aligned} B &= \sum_{n=a+1}^{\infty} \frac{(-1)^{n+a+1}a!}{n!} = \sum_{n=a+1}^{\infty} \frac{(-1)^{n+a+1}}{(a+1)\cdot(a+2)\cdots(n-1)\cdot n} \\ &= \frac{(-1)^{2a+2}}{a+1} + \frac{(-1)^{2a+3}}{(a+1)(a+2)} + \frac{(-1)^{2a+4}}{(a+1)(a+2)(a+3)} + \cdots \\ &= \frac{1}{a+1} - \frac{1}{(a+1)(a+2)} + \frac{1}{(a+1)(a+2)(a+3)} - \cdots \end{aligned}$$

Hence B is an alternating series of decreasing terms and by the alternating series test (Theorem 3.3.14) it converges. Further, it must converge to a number between its first and second partial sums (see the discussion before Theorem 3.3.14). Hence the right-hand side lies between

$$\frac{1}{a+1} \qquad \text{and} \qquad \frac{1}{a+1} - \frac{1}{(a+1)(a+2)} = \frac{1}{a+2}$$

Since a is a positive integer the above tells us that B converges to a real number strictly greater than 0 and strictly less than 1. Hence it cannot be an integer.

This gives us a contradiction and hence e cannot be rational.

▸▸ Irrationality of π

This proof is due to Niven (1946) and doesn't require any mathematics beyond the level of this course. Much like the proofs above we will start by assuming that π is rational and then reach a contradiction. Again this contradiction will be that a given quantity must be an integer but at the same time must lie strictly between 0 and 1.

Assume that π is a rational number and so can be written as $\pi = \frac{a}{b}$ with a, b positive integers. Now let n be a positive integer and define the polynomial

$$f(x) = \frac{x^n(a-bx)^n}{n!}.$$

It is certainly not immediately obvious why and how Niven chose this polynomial, but you will see that it has been very carefully crafted to make the proof work. In particular we will show — under our assumption that π is rational — that, if n is really big, then

$$I_n = \int_0^\pi f(x)\sin(x)\mathrm{d}x$$

is an integer and it also lies strictly between 0 and 1, giving the required contradiction.

▸▸▸ Bounding the Integral

Consider again the polynomial

$$f(x) = \frac{x^n(a-bx)^n}{n!}.$$

Notice that

$$\begin{aligned} f(0) &= 0 \\ f(\pi) &= f(a/b) = 0. \end{aligned}$$

Furthermore, for $0 \leqslant x \leqslant \pi = a/b$, we have $x \leqslant \frac{a}{b}$ and $a - bx \leqslant a$ so that

$$0 \leqslant x(a-bx) \leqslant a^2/b.$$

We could work out a more precise[77] upper bound, but this one is sufficient for the analysis that follows. Hence

$$0 \leqslant f(x) \leqslant \left(\frac{a^2}{b}\right)^n \frac{1}{n!}$$

We also know that for $0 \leqslant x \leqslant \pi = a/b$, $0 \leqslant \sin(x) \leqslant 1$. Thus

$$0 \leqslant f(x)\sin(x) \leqslant \left(\frac{a^2}{b}\right)^n \frac{1}{n!}$$

for all $0 \leqslant x \leqslant 1$. Using this inequality we bound

$$0 < I_n = \int_0^\pi f(x)\sin(x)\mathrm{d}x < \left(\frac{a^2}{b}\right)^n \frac{1}{n!}.$$

We will later show that, if n is really big, then $\left(\frac{a^2}{b}\right)^n \frac{1}{n!} < 1$. We'll first show, starting now, that I_n is an integer.

77 You got lots of practice finding the maximum and minimum values of continuous functions on closed intervals when you took calculus last term.

▸▸▸ Integration by Parts

In order to show that the value of this integral is an integer we will use integration by parts. You have already practiced using integration by parts to integrate quantities like

$$\int x^2 \sin(x)\,\mathrm{d}x$$

and this integral isn't much different. For the moment let us just use the fact that $f(x)$ is a polynomial of degree $2n$. Using integration by parts with $u = f(x)$, $\mathrm{d}v = \sin(x)$ and $v = -\cos(x)$ gives us

$$\int f(x)\sin(x)\,\mathrm{d}x = -f(x)\cos(x) + \int f'(x)\cos(x)\,\mathrm{d}x$$

Use integration by parts again with $u = f'(x)$, $\mathrm{d}v = \cos(x)$ and $v = \sin(x)$.

$$= -f(x)\cos(x) + f'(x)\sin(x) - \int f''(x)\sin(x)\,\mathrm{d}x$$

Use integration by parts yet again, with $u = f''(x)$, $\mathrm{d}v = \sin(x)$ and $v = -\cos(x)$.

$$= -f(x)\cos(x) + f'(x)\sin(x) + f''(x)\cos(x) - \int f'''(x)\cos(x)\,\mathrm{d}x$$

And now we can see the pattern; we get alternating signs, and then derivatives multiplied by sines and cosines:

$$\begin{aligned}\int f(x)\sin(x)\mathrm{d}x = \cos(x)&\left(-f(x) + f''(x) - f^{(4)}(x) + f^{(6)}(x) - \cdots\right)\\ &+ \sin(x)\left(f'(x) - f'''(x) + f^{(5)}(x) - f^{(7)}(x) + \cdots\right)\end{aligned}$$

This terminates at the $2n^{\text{th}}$ derivative since $f(x)$ is a polynomial of degree $2n$. We can check this computation by differentiating the terms on the right-hand side:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}&\left(\cos(x)\left(-f(x) + f''(x) - f^{(4)}(x) + f^{(6)}(x) - \cdots\right)\right)\\ &= -\sin(x)\left(-f(x) + f''(x) - f^{(4)}(x) + f^{(6)}(x) - \cdots\right)\\ &\quad + \cos(x)\left(-f'(x) + f'''(x) - f^{(5)}(x) + f^{(7)}(x) - \cdots\right)\end{aligned}$$

and similarly

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}&\left(\sin(x)\left(f'(x) - f'''(x) + f^{(5)}(x) - f^{(7)}(x) + \cdots\right)\right)\\ &= \cos(x)\left(f'(x) - f'''(x) + f^{(5)}(x) - f^{(7)}(x) + \cdots\right)\\ &\quad + \sin(x)\left(f''(x) - f^{(4)}(x) + f^{(6)}(x) - \cdots\right)\end{aligned}$$

When we add these two expressions together all the terms cancel except $f(x)\sin(x)$, as required.

Now when we take the definite integral from 0 to π , all the sine terms give 0 because $\sin(0) = \sin(\pi) = 0$. Since $\cos(\pi) = -1$ and $\cos(0) = +1$, we are just left with:

$$\int_0^\pi f(x)\sin(x)\mathrm{d}x = \Big(f(0) - f''(0) + f^{(4)}(0) - f^{(6)}(0) + \cdots + (-1)^n f^{(2n)}(0)\Big) + \Big(f(\pi) - f''(\pi) + f^{(4)}(\pi) - f^{(6)}(\pi) + \cdots + (-1)^n f^{(2n)}(\pi)\Big)$$

So to show that I_n is an integer, it now suffices to show that $f^{(j)}(0)$ and $f^{(j)}(\pi)$ are integers.

▸▸ The Derivatives are Integers

Recall that

$$f(x) = \frac{x^n(a-bx)^n}{n!}$$

and expand it:

$$f(x) = \frac{c_0}{n!}x^0 + \frac{c_1}{n!}x^1 + \cdots + \frac{c_n}{n!}x^n + \cdots + \frac{c_{2n}}{n!}x^{2n}$$

All the c_j are integers, and clearly $c_j = 0$ for all $j = 0, 1, \ldots, n-1$, because of the factor x^n in $f(x)$.

Now take the k^{th} derivative and set $x = 0$. Note that, if $j < k$, then $\frac{\mathrm{d}^k}{\mathrm{d}x^k}x^j = 0$ for all x and, if $j > k$, then $\frac{\mathrm{d}^k}{\mathrm{d}x^k}x^j$ is some number times x^{j-k} which evaluates to zero when we set $x = 0$. So

$$f^{(k)}(0) = \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(\frac{c_k}{k!}x^k\right) = \frac{k!c_k}{n!}$$

If $k < n$, then this is zero since $c_k = 0$. If $k > n$, this is an integer because c_k is an integer and $k!/n! = (n+1)(n+2)\cdots(k-1)k$ is an integer. If $k = n$, then $f^{(k)}(0) = c_n$ is again an integer. Thus all the derivatives of $f(x)$ evaluated at $x = 0$ are integers.

But what about the derivatives at $\pi = a/b$? To see this, we can make use of a handy symmetry. Notice that

$$f(x) = f(\pi - x) = f(a/b - x)$$

You can confirm this by just grinding through the algebra:

$$\begin{aligned}
f(x) &= \frac{x^n(a-bx)^n}{n!} && \text{now replace } x \text{ with } a/b - x\\
f(a/b-x) &= \frac{(a/b-x)^n(a-b(a/b-x))^n}{n!} && \text{start cleaning this up:}\\
&= \frac{\left(\frac{a-bx}{b}\right)^n(a-a+bx)^n}{n!}\\
&= \frac{\left(\frac{a-bx}{b}\right)^n(bx)^n}{n!}\\
&= \frac{(a-bx)^n x^n}{n!} = f(x)
\end{aligned}$$

Using this symmetry (and the chain rule) we see that

$$f'(x) = -f'(\pi - x)$$

and if we keep differentiating

$$f^{(k)}(x) = (-1)^k f^{(k)}(\pi - x)$$

Setting $x = 0$ in this tells us that

$$f^{(k)}(0) = (-1)^k f^{(k)}(\pi)$$

So because all the derivatives at $x = 0$ are integers, we know that all the derivatives at $x = \pi$ are also integers.

Hence the integral we are interested in

$$\int_0^\pi f(x)\sin(x)\mathrm{d}x$$

must be an integer.

▸▸▸ Putting It Together

Based on our assumption that $\pi = a/b$ is rational, we have shown that the integral

$$I_n = \int_0^\pi \frac{x^n(a - bx)}{n!}\sin(x)\mathrm{d}x$$

satisfies

$$0 < I_n < \left(\frac{a^2}{b}\right)^n \frac{1}{n!}$$

and also that I_n is an integer.

We are, however, free to choose n to be any positive integer we want. If we take n to be very large — in particular much much larger than a — then $n!$ will be much much larger than a^{2n} (we showed this in Example 3.6.6), and consequently

$$0 < I_n < \left(\frac{a^2}{b}\right)^n \frac{1}{n!} < 1$$

Which means that the integral cannot be an integer. This gives the required contradiction, showing that π is irrational.

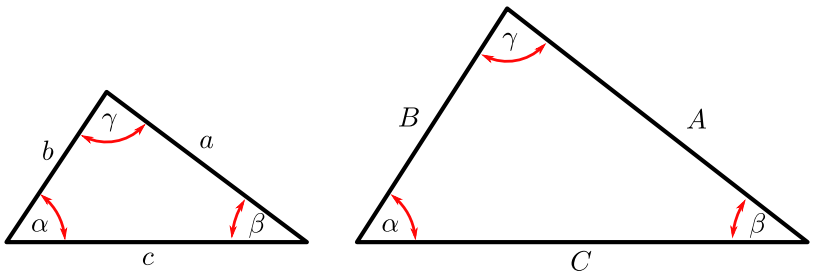
Appendix A

HIGH SCHOOL MATERIAL

This chapter is really split into three parts.

- Sections A.1 to A.11 contains results that we expect you to understand and know.
- Then Section A.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.

A.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

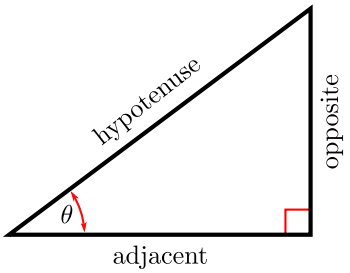
$$\frac{A}{a} = \frac{B}{b} = \frac{C}{c}$$

- (SAS — side angle side) Two sides have lengths in the same ratio and the angle between them is the same. For example

$$\frac{A}{a} = \frac{C}{c} \text{ and angle } \beta \text{ is same}$$

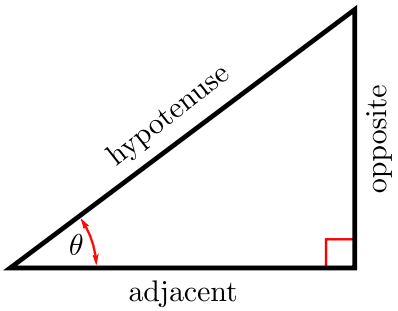
A.2▲ Pythagoras

For a right-angled triangle the length of the hypotenuse is related to the lengths of the other two sides by



$$(\text{adjacent})^2 + (\text{opposite})^2 = (\text{hypotenuse})^2$$

A.3▲ Trigonometry — Definitions

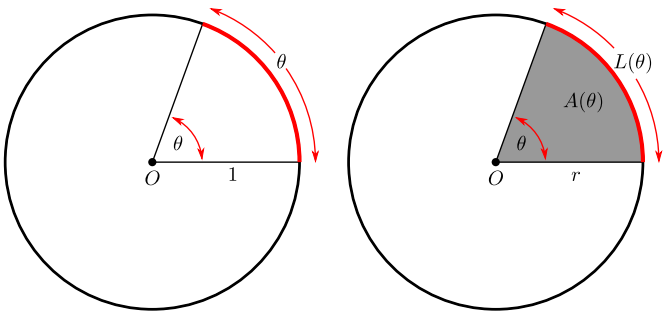


$$\sin\theta = \frac{\text{opposite}}{\text{hypotenuse}} \qquad \csc\theta = \frac{1}{\sin\theta}$$

$$\cos\theta = \frac{\text{adjacent}}{\text{hypotenuse}} \qquad \sec\theta = \frac{1}{\cos\theta}$$

$$\tan\theta = \frac{\text{opposite}}{\text{adjacent}} \qquad \cot\theta = \frac{1}{\tan\theta}$$

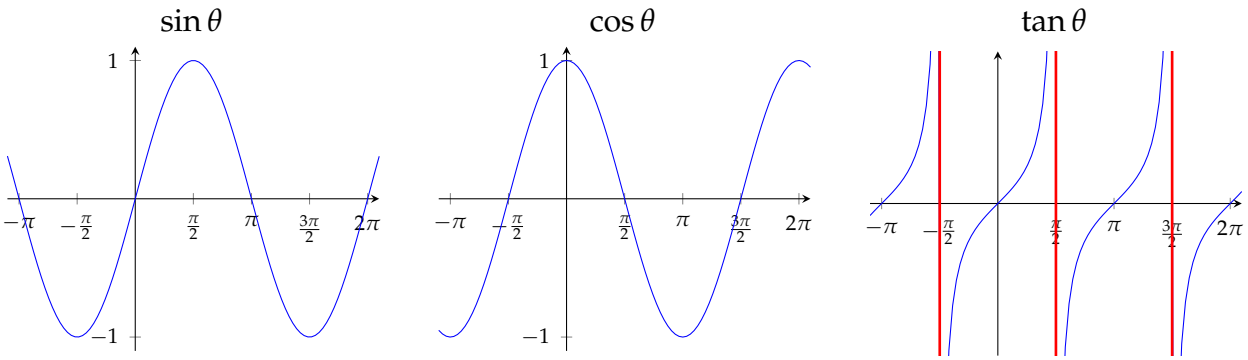
A.4▲ Radians, Arcs and Sectors



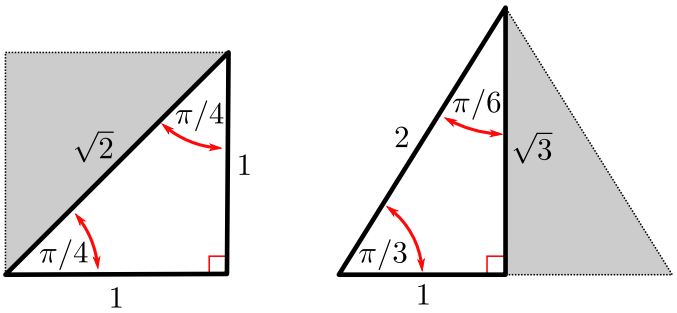
For a circle of radius r and angle of θ radians:

- Arc length $L(\theta) = r\theta$.
- Area of sector $A(\theta) = \frac{\theta}{2}r^2$.

A.5▲ Trigonometry — Graphs



A.6▲ Trigonometry — Special Triangles



From the above pair of special triangles we have

$$\sin\frac{\pi}{4}=\frac{1}{\sqrt{2}} \qquad \sin\frac{\pi}{6}=\frac{1}{2} \qquad \sin\frac{\pi}{3}=\frac{\sqrt{3}}{2}$$
$$\cos\frac{\pi}{4}=\frac{1}{\sqrt{2}} \qquad \cos\frac{\pi}{6}=\frac{\sqrt{3}}{2} \qquad \cos\frac{\pi}{3}=\frac{1}{2}$$
$$\tan\frac{\pi}{4}=1 \qquad \tan\frac{\pi}{6}=\frac{1}{\sqrt{3}} \qquad \tan\frac{\pi}{3}=\sqrt{3}$$

A.7▲ Trigonometry — Simple Identities

- Periodicity

$$\sin(\theta+2\pi)=\sin(\theta) \qquad \cos(\theta+2\pi)=\cos(\theta)$$

- Reflection

$$\sin(-\theta)=-\sin(\theta) \qquad \cos(-\theta)=\cos(\theta)$$

- Reflection around $\pi/4$

$$\sin\left(\tfrac{\pi}{2}-\theta\right)=\cos\theta \qquad \cos\left(\tfrac{\pi}{2}-\theta\right)=\sin\theta$$

- Reflection around $\pi/2$

$$\sin(\pi-\theta)=\sin\theta \qquad \cos(\pi-\theta)=-\cos\theta$$

- Rotation by π

$$\sin(\theta+\pi)=-\sin\theta \qquad \cos(\theta+\pi)=-\cos\theta$$

- Pythagoras

$$\sin^2\theta+\cos^2\theta=1$$

A.8▲ Trigonometry — Add and Subtract Angles

- Sine

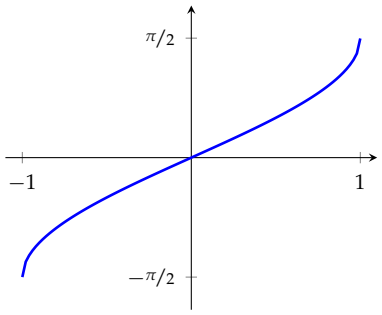
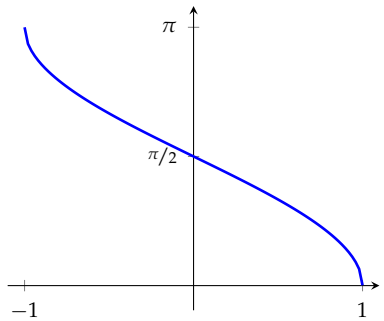
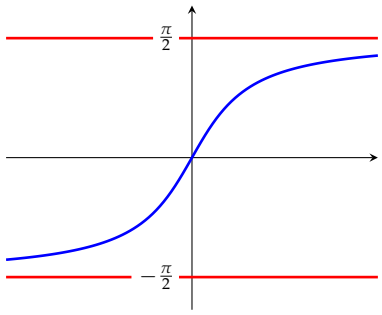
$$\sin(\alpha\pm\beta)=\sin(\alpha)\cos(\beta)\pm\cos(\alpha)\sin(\beta)$$

- Cosine

$$\cos(\alpha\pm\beta)=\cos(\alpha)\cos(\beta)\mp\sin(\alpha)\sin(\beta)$$

A.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.

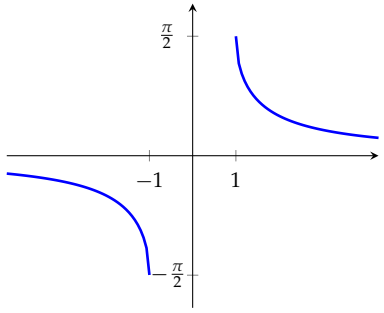
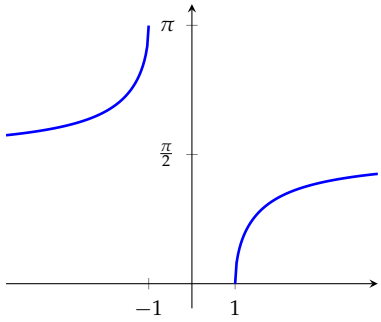
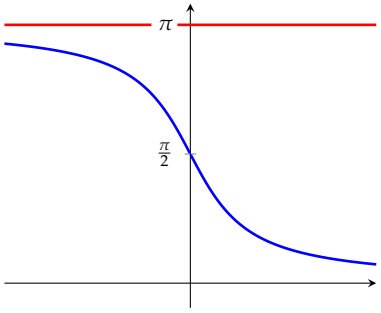
$\arcsin x$	$\arccos x$	$\arctan x$
Domain: $-1\leqslant x\leqslant 1$ Range: $-\frac{\pi}{2}\leqslant \arcsin x\leqslant\frac{\pi}{2}$	Domain: $-1\leqslant x\leqslant 1$ Range: $0\leqslant \arccos x\leqslant\pi$	Domain: all real numbers Range: $-\frac{\pi}{2}< \arctan x<\frac{\pi}{2}$
		

Since these functions are inverses of each other we have

$$\arcsin(\sin\theta)=\theta \qquad -\frac{\pi}{2}\leqslant\theta\leqslant\frac{\pi}{2}$$
$$\arccos(\cos\theta)=\theta \qquad 0\leqslant\theta\leqslant\pi$$
$$\arctan(\tan\theta)=\theta \qquad -\frac{\pi}{2}<\theta<\frac{\pi}{2}$$

and also

$$\begin{aligned} \sin(\arcsin x) &= x & -1 \leqslant x \leqslant 1 \\ \cos(\arccos x) &= x & -1 \leqslant x \leqslant 1 \\ \tan(\arctan x) &= x & \text{any real } x \end{aligned}$$

$\operatorname{arccsc} x$	$\operatorname{arcsec} x$	$\operatorname{arccot} x$
Domain: $\lvert x\rvert \geqslant 1$ Range: $-\frac{\pi}{2} \leqslant \operatorname{arccsc} x \leqslant \frac{\pi}{2}$ $\operatorname{arccsc} x \neq 0$	Domain: $\lvert x\rvert \geqslant 1$ Range: $0 \leqslant \operatorname{arcsec} x \leqslant \pi$ $\operatorname{arcsec} x \neq \frac{\pi}{2}$	Domain: all real numbers Range: $0 < \operatorname{arccot} x < \pi$
		

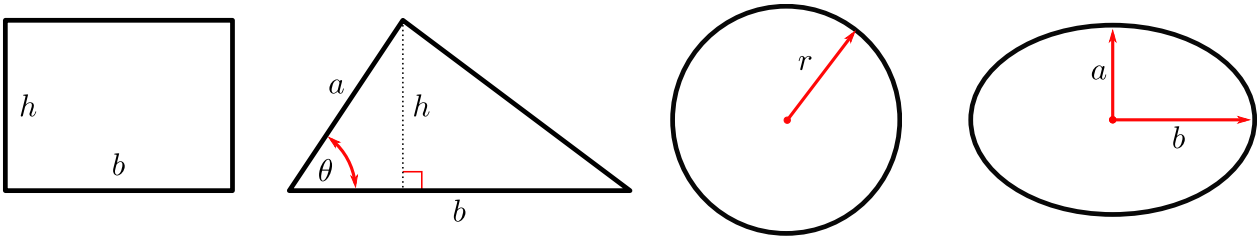
Again

$$\begin{aligned} \operatorname{arccsc}(\csc\theta) &= \theta & -\frac{\pi}{2} \leqslant \theta \leqslant \frac{\pi}{2},\ \theta \neq 0 \\ \operatorname{arcsec}(\sec\theta) &= \theta & 0 \leqslant \theta \leqslant \pi,\ \theta \neq \frac{\pi}{2} \\ \operatorname{arccot}(\cot\theta) &= \theta & 0 < \theta < \pi \end{aligned}$$

and

$$\begin{aligned} \csc(\operatorname{arccsc} x) &= x & \lvert x\rvert \geqslant 1 \\ \sec(\operatorname{arcsec} x) &= x & \lvert x\rvert \geqslant 1 \\ \cot(\operatorname{arccot} x) &= x & \text{any real } x \end{aligned}$$

A.10 Areas



- Area of a rectangle

$$A = bh$$

- Area of a triangle

$$A = \frac{1}{2}bh = \frac{1}{2}ab\sin\theta$$

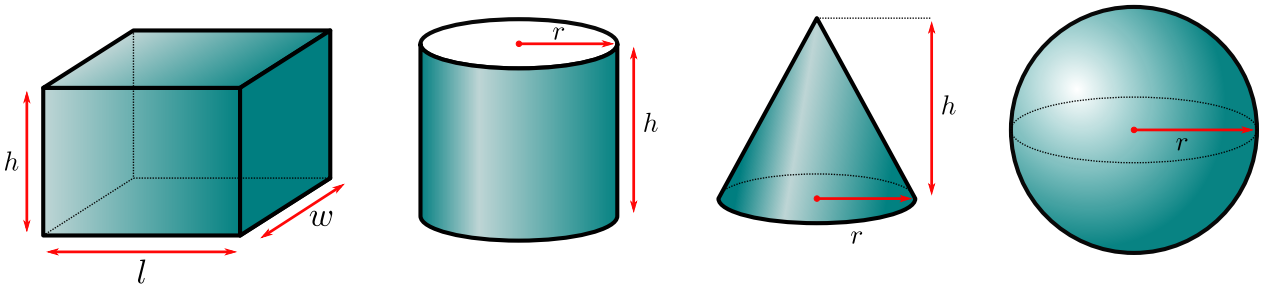
- Area of a circle

$$A = \pi r^2$$

- Area of an ellipse

$$A = \pi ab$$

A.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

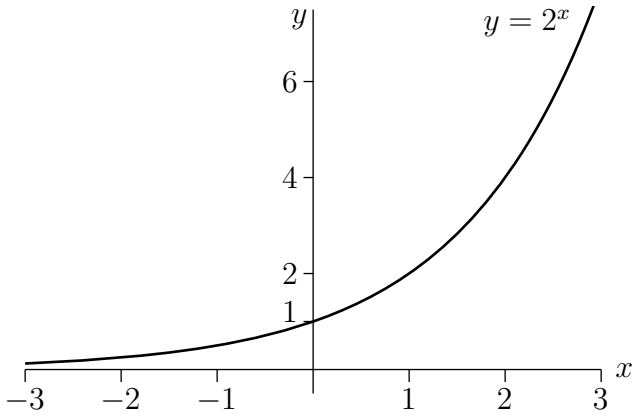
$$V = \frac{4}{3}\pi r^3$$

A.12 Powers

In the following, x and y are arbitrary real numbers, and q is an arbitrary constant that is strictly bigger than zero.

- $q^0 = 1$

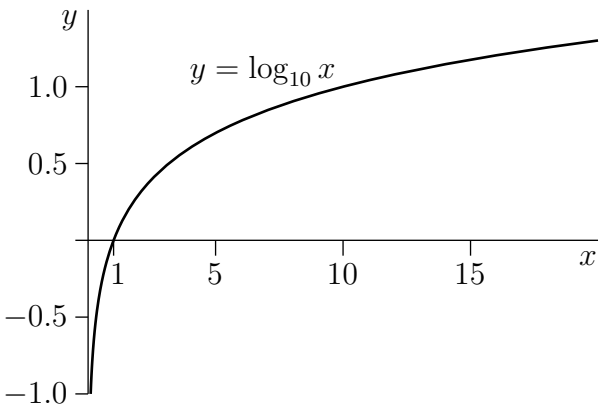
- $q^{x+y} = q^x q^y$, $q^{x-y} = \frac{q^x}{q^y}$
- $q^{-x} = \frac{1}{q^x}$
- $\left(q^x\right)^y = q^{xy}$
- $\lim_{x\to\infty} q^x = \infty$, $\lim_{x\to-\infty} q^x = 0$ if $q > 1$
- $\lim_{x\to\infty} q^x = 0$, $\lim_{x\to-\infty} q^x = \infty$ if $0 < q < 1$
- The graph of 2^x is given below. The graph of q^x, for any $q > 1$, is similar.



A.13▲ Logarithms

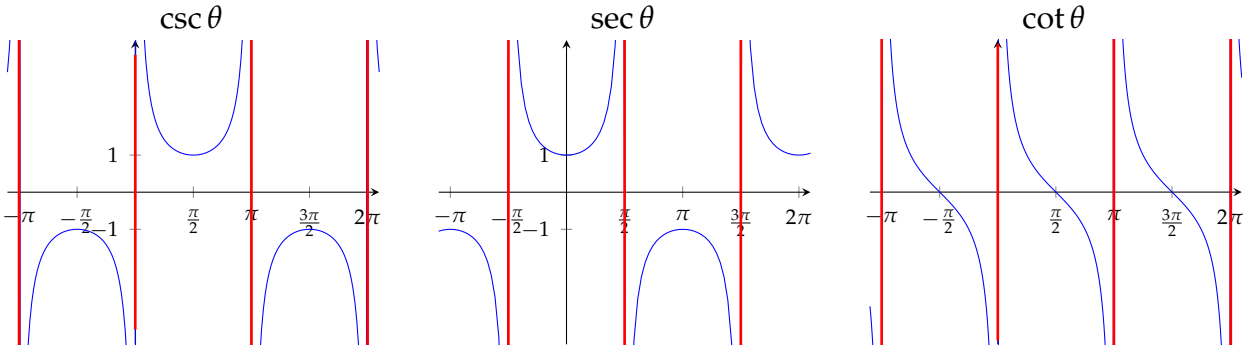
In the following, x and y are arbitrary real numbers that are strictly bigger than 0, and p and q are arbitrary constants that are strictly bigger than one.

- $q^{\log_q x} = x$, $\log_q\left(q^x\right) = x$
- $\log_q x = \frac{\log_p x}{\log_p q}$
- $\log_q 1 = 0$, $\log_q q = 1$
- $\log_q(xy) = \log_q x + \log_q y$
- $\log_q\left(\frac{x}{y}\right) = \log_q x - \log_q y$
- $\log_q\left(\frac{1}{y}\right) = -\log_q y$,
- $\log_q(x^y) = y \log_q x$
- $\lim_{x\to\infty} \log_q x = \infty$, $\lim_{x\to0+} \log_q x = -\infty$
- The graph of $\log_{10} x$ is given below. The graph of $\log_q x$, for any $q > 1$, is similar.



A.14▲ Highschool Material You Should be Able to Derive

- Graphs of $\csc\theta$, $\sec\theta$ and $\cot\theta$:



- More Pythagoras

$$\sin^2\theta+\cos^2\theta=1 \xrightarrow{\text{divide by } \cos^2\theta} \tan^2\theta+1=\sec^2\theta$$

$$\sin^2\theta+\cos^2\theta=1 \xrightarrow{\text{divide by } \sin^2\theta} 1+\cot^2\theta=\csc^2\theta$$

- 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)

$$\begin{aligned}\cos(2\alpha)&=\cos^2(\alpha)-\sin^2(\alpha)\\&=2\cos^2(\alpha)-1 && (\text{use } \sin^2(\alpha)=1-\cos^2(\alpha))\\&=1-2\sin^2(\alpha) && (\text{use } \cos^2(\alpha)=1-\sin^2(\alpha))\end{aligned}$$

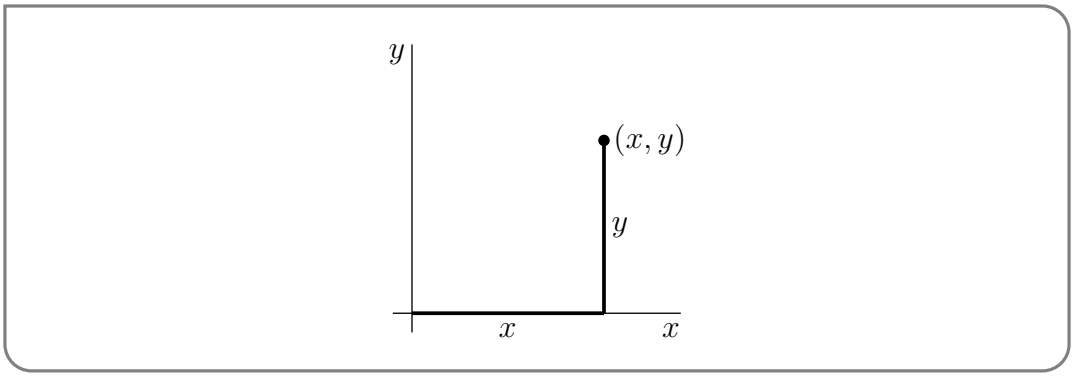
- Composition of trigonometric and inverse trigonometric functions:

$$\cos(\arcsin x)=\sqrt{1-x^2} \qquad \sec(\arctan x)=\sqrt{1+x^2}$$

and similar expressions.

A.15▲ Cartesian Coordinates

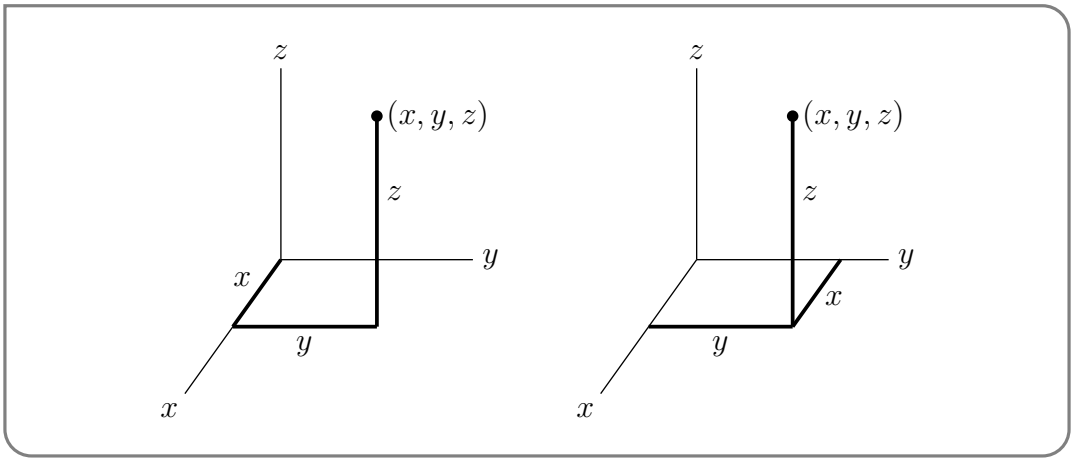
Each point in two dimensions may be labeled by two coordinates (x, y) which specify the position of the point in some units with respect to some axes as in the figure below.



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.



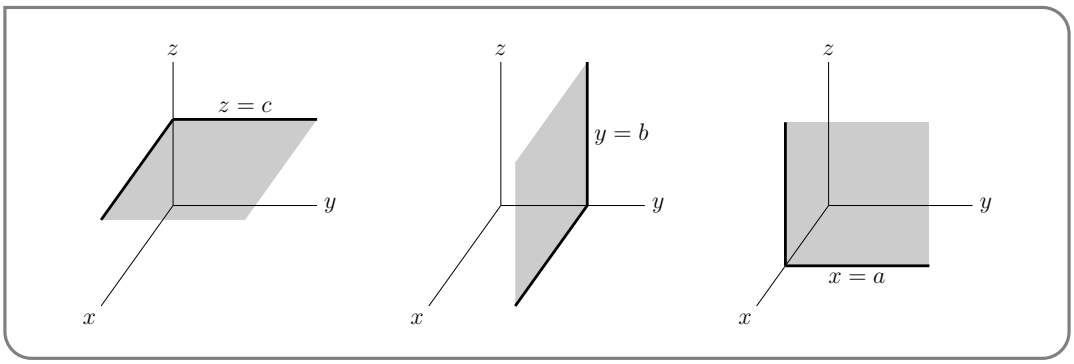
The set of all points in three dimensions is denoted $\mathbb{R}^3$. The plane that contains, for example, the x– and y–axes is called the xy–plane.

- 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

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

A.16▲ 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[1] are extremely cumbersome, and no such formula exists for polynomials of degree 5 and higher[2].

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[3] 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[4] 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[5] 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 A.16.1 (A very useful trick).

If r or $-r$ is an integer root of a polynomial $P(x) = a_n x^n + \cdots + a_1 x + a_0$ 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

$$a_n \cdot r^n + \cdots + a_1 \cdot r + a_0 = 0$$

If we isolate a_0 in this expression we get

$$a_0 = -\left[a_n r^n + \cdots + a_1 r\right]$$

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. □

Let us put this observation to work.

Example A.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

$$\begin{aligned} P(1) &= 2 & P(-1) &= 0 \\ P(2) &= 6 & P(-2) &= -10 \end{aligned}$$

- Thus the only integer root is -1.

Example A.16.1

Example A.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):

$$\begin{aligned} P(1) &= 0 & P(-1) &= 4 \\ P(2) &= 40 & P(-2) &= 12 \\ P(3) &= 132 & P(-3) &= 0 \\ P(6) &= 900 & P(-6) &= -336 \end{aligned}$$

- Thus the only integer roots are 1 and -3.

Example A.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 A.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

$$a_n b^n + \cdots + a_1 b d^{n-1} + a_0 d^n = 0$$

Move terms around to isolate $a_0 d^n$:

$$a_0 d^n = -\left[a_n b^n + \cdots + a_1 b d^{n-1}\right]$$

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 $a_n b^n$:

$$a_n b^n = -\left[a_{n-1} b^{n-1} d + \cdots + a_1 b d^{n-1} + a_0 d^n\right]$$

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 A.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[6]

$$\begin{aligned} P(1) &= -2 & P(-1) &= 0 \\ P(3) &= 12 & P(-3) &= 18 \\ P\left(\tfrac{1}{2}\right) &= -3 & P\left(-\tfrac{1}{2}\right) &= -2 \\ P\left(\tfrac{3}{2}\right) &= 0 & P\left(-\tfrac{3}{2}\right) &= 3 \end{aligned}$$

so the roots are -1 and $\frac{3}{2}$.

6 Again, this is a little tedious, but not difficult. Its 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 A.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 divison.

Trick A.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[7] in school, but with 10 replaced by x.

Example A.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.

$$\begin{array}{r|l} & x^2 \\ x + 1 & x^3 - x^2 + 2 \\ & x^3 + x^2 \quad \longleftarrow x^2(x+1) \end{array}$$

7 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.

$$\begin{array}{rl} & x^2 \\ x+1 \,\big|\, & x^3 - x^2 + \quad 2 \\ & \underline{x^3 + x^2 \qquad\quad} \quad \longleftarrow x^2(x+1) \\ & -2x^2 \end{array}$$

- 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.

$$\begin{array}{rll} & x^2 - 2x & \\ x+1 \,\big|\, & x^3 - x^2 + \quad 2 & \\ & \underline{x^3 + x^2 \qquad\quad} & \longleftarrow x^2(x+1) \\ & -2x^2 & \\ & -2x^2 - 2x & \longleftarrow -2x(x+1) \end{array}$$

And so on.

$$\begin{array}{rll} & x^2 - 2x + 2 & \\ x+1 \,\big|\, & x^3 - x^2 + \quad 2 & \\ & \underline{x^3 + x^2 \qquad\quad} & \longleftarrow x^2(x+1) \\ & -2x^2 & \\ & \underline{-2x^2 - 2x} & \longleftarrow -2x(x+1) \\ & 2x + 2 & \\ & \underline{2x + 2} & \longleftarrow 2(x+1) \\ & 0 & \end{array}$$

- 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[8].

Example A.16.4

8 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 A.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

$$\begin{aligned} \text{coefficient of } x^3\text{:} &\qquad a = 1 \\ \text{coefficient of } x^2\text{:} &\qquad a + b = -1 \\ \text{coefficient of } x^1\text{:} &\qquad b + c = 0 \\ \text{coefficient of } x^0\text{:} &\qquad c = 2 \end{aligned}$$

- This gives us a system of equations that we can solve quite directly. Indeed it tells us immediately that 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 A.16.5

Appendix B

Complex Numbers and Exponentials

B.1▲ Definition and Basic Operations

We'll start with the definition of a complex number and its addition and multiplication rules. You may find the multiplication rule quite mysterious. Don't worry. We'll soon gets lots of practice using it and seeing how useful it is.

Definition B.1.1.

(a) The complex plane is simply the xy-plane equipped with an addition operation and a multiplication operation. A complex number is nothing more than a point in that xy-plane. It is conventional to use the notation $x + iy$[1] to stand for the complex number (x, y). In other words, it is conventional to write x in place of $(x, 0)$ and i in place of $(0, 1)$.

(b) The first component, x, of the complex number $x + iy$ is called its real part and the second component, y, is called its imaginary part, even though there is nothing imaginary[2] about it.

(c) The sum of the complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ is defined by

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

That is, you just separately add the real parts and the imaginary parts.

1 In electrical engineering it is conventional to use $x + jy$ instead of $x + iy$.

2 Don't attempt to attribute any special significance to the word "complex" in "complex number", or to the word "real" in "real number" and "real part", or to the word "imaginary" in "imaginary part". All are *just names*. The name "imaginary" was introduced by René Descartes in 1637. 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. Originally, "imaginary" was a derogatory term and imaginary numbers were thought to be useless. But they turned out to be incredibly useful!

Definition B.1.1 (continued).

(d) The product of the complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ is defined by

$$z_1 z_2 = x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1)$$

Do not memorise this multiplication rule. We'll see a simple, effective memory aid for it shortly. The heart of that memory aid is the observation that the complex number i has the special property that

$$i^2 = (0 + 1i)(0 + 1i) = (0 \times 0 - 1 \times 1) + i(0 \times 1 + 1 \times 0) = -1$$

Addition and multiplication of complex numbers obey the familiar addition rules

$$\begin{aligned} z_1 + z_2 &= z_2 + z_1 \\ z_1 + (z_2 + z_3) &= (z_1 + z_2) + z_3 \\ 0 + z_1 &= z_1 \end{aligned}$$

and multiplication rules

$$\begin{aligned} z_1 z_2 &= z_2 z_1 \\ z_1(z_2 z_3) &= (z_1 z_2) z_3 \\ 1 z_1 &= z_1 \end{aligned}$$

and distributive laws

$$\begin{aligned} z_1(z_2 + z_3) &= z_1 z_2 + z_1 z_3 \\ (z_1 + z_2) z_3 &= z_1 z_3 + z_2 z_3 \end{aligned}$$

To remember how to multiply complex numbers, you just have to supplement the familiar rules of the real number system with $i^2 = -1$. The previous sentence is the memory aid referred to in Definition B.1.1(d).

Example B.1.2

If $z = 1 + 2i$ and $w = 3 + 4i$, then

$$\begin{aligned} z + w &= (1 + 2i) + (3 + 4i) = 4 + 6i \\ zw &= (1 + 2i)(3 + 4i) \quad = 3 + 4i + 6i + 8i^2 = 3 + 4i + 6i - 8 = -5 + 10i \end{aligned}$$

Example B.1.2

Definition B.1.3.

(a) The negative of any complex number $z = x + iy$ is defined by

$$-z = -x + (-y)i$$

and obviously obeys $z + (-z) = 0$.

(b) The reciprocal[3], z^{-1} or $\frac{1}{z}$, of any complex number $z = x + iy$, other than 0, is defined by

$$\frac{1}{z} z = 1$$

We shall see below that it is given by the formula

$$z^{-1} = \frac{1}{z} = \frac{x}{x^2+y^2} + \frac{-y}{x^2+y^2} i$$

Example B.1.4

It is possible to derive the formula for $\frac{1}{z}$ by observing that

$$(a + ib)(x + iy) = [ax - by] + i[ay + bx]$$

equals $1 = 1 + i0$ if and only if

$$\begin{aligned} ax - by &= 1 \\ ay + bx &= 0 \end{aligned}$$

and solving these equations for a and b. We will see a much shorter derivation in Remark B.1.6 below. For now, we'll content ourselves with just verifying that $\frac{x}{x^2+y^2} + \frac{-y}{x^2+y^2} i$ is the inverse of $x + iy$ by multiplying out

$$\begin{aligned} \left(\frac{x}{x^2+y^2} - \frac{y}{x^2+y^2} i \right) (x + iy) &= \frac{x^2}{x^2+y^2} - \frac{xy}{x^2+y^2} i + \frac{xy}{x^2+y^2} i - \frac{y^2}{x^2+y^2} i^2 \\ &= \frac{x^2 - i^2 y^2}{x^2+y^2} = \frac{x^2+y^2}{x^2+y^2} = 1 \end{aligned}$$

Example B.1.4

3 The reciprocal z^{-1} is also called the multiplicative inverse of z.

Definition B.1.5.

(a) The complex conjugate of $z = x + iy$ is denoted $\bar{z}$ and is defined to be

$$\bar{z} = \overline{x + iy} = x - iy$$

That is, to take the complex conjugate, one replaces every i by $-i$ and vice versa.

(b) The distance from $z = x + iy$ (recall that this is the point (x, y) in the xy-plane) to 0 is denoted $|z|$ and is called the absolute value, or modulus, of z. It is given by

$$|z| = \sqrt{x^2 + y^2}$$

Note that

$$z\bar{z} = (x + iy)(x - iy) = x^2 - ixy + ixy + y^2 = x^2 + y^2$$

is always a nonnegative real number and that

$$|z| = \sqrt{z\,\bar{z}}$$

Remark B.1.6. Let $z = x + iy$ with x and y real. Since $|z|^2 = z\,\bar{z}$, we have

$$\frac{1}{z} = \frac{1}{z}\frac{\bar{z}}{\bar{z}} = \frac{\bar{z}}{|z|^2} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} + \frac{-y}{x^2 + y^2}i$$

which is the formula for $\frac{1}{z}$ given in Definition B.1.3(b).

Example B.1.7

It is easy to divide a complex number by a real number. For example

$$\frac{11 + 2i}{25} = \frac{11}{25} + \frac{2}{25}i$$

In general, the complex conjugate provides us with a trick for rewriting any ratio of complex numbers as a ratio with a real denominator. For example, suppose that we want to find $\frac{1+2i}{3+4i}$. The trick is to multiply by $1 = \frac{3-4i}{3-4i}$. The number $3 - 4i$ is the complex conjugate of the denominator $3 + 4i$. Since $(3 + 4i)(3 - 4i) = 9 - 12i + 12i - 16i^2 = 9 + 16 = 25$

$$\frac{1 + 2i}{3 + 4i} = \frac{1 + 2i}{3 + 4i}\,\frac{3 - 4i}{3 - 4i} = \frac{(1 + 2i)(3 - 4i)}{25} = \frac{11 + 2i}{25} = \frac{11}{25} + \frac{2}{25}i$$

Example B.1.7

Definition B.1.8.

The notations[4] $\Re z$ and $\Im z$ stand for the real and imaginary parts of the complex number z, respectively. If $z = x + iy$ (with x and y real) they are defined by

$$\Re z = x \qquad \Im z = y$$

Note that both $\Re z$ and $\Im z$ are real numbers. Just subbing in $\bar{z} = x - iy$, you can verify that

$$\Re z = \frac{1}{2}(z + \bar{z}) \qquad \Im z = \frac{1}{2i}(z - \bar{z})$$

Lemma B.1.9.

If $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, then

$$|z_1 z_2| = |z_1|\,|z_2|$$

Proof. Since $z_1 z_2 = (x_1 + iy_1)(x_2 + iy_2) = (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1)$,

$$\begin{aligned} |z_1 z_2| &= \sqrt{(x_1 x_2 - y_1 y_2)^2 + (x_1 y_2 + x_2 y_1)^2} \\ &= \sqrt{x_1^2 x_2^2 - 2x_1 x_2 y_1 y_2 + y_1^2 y_2^2 + x_1^2 y_2^2 + 2x_1 y_2 x_2 y_1 + x_2^2 y_1^2} \\ &= \sqrt{x_1^2 x_2^2 + y_1^2 y_2^2 + x_1^2 y_2^2 + x_2^2 y_1^2} = \sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2)} \\ &= |z_1|\,|z_2| \end{aligned}$$

□

B.2▲ The Complex Exponential

B.2.1 ▸▸ Definition and Basic Properties.

There are two equivalent standard definitions of the exponential, e^z, of the complex number $z = x + iy$. For the more intuitive definition, one simply replaces the real number x in the Taylor series expansion[5] $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ with the complex number z, giving

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{EZ}$$

We instead highlight the more computationally useful definition.

4 The symbols $\Re$ and $\Im$ are the letters R and I in the Fraktur font, which was created in the early 1500's and became common in the German-speaking world. A standard alternative notation is $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$.

5 See Theorem 3.6.5.

Definition B.2.1.

For any complex number $z = x + iy$, with x and y real, the exponential e^z, is defined by

$$e^{x+iy} = e^x \cos y + ie^x \sin y$$

In particular[6], $e^{iy} = \cos y + i \sin y$.

We will not fully prove that the intuitive definition (EZ) and the computational Definition B.2.1 are equivalent. But we will do so in the special case that $z = iy$, with y real. Under (EZ),

$$e^{iy} = 1 + iy + \frac{(iy)^2}{2!} + \frac{(iy)^3}{3!} + \frac{(iy)^4}{4!} + \frac{(iy)^5}{5!} + \frac{(iy)^6}{6!} + \cdots$$

The even terms in this expansion are

$$1 + \frac{(iy)^2}{2!} + \frac{(iy)^4}{4!} + \frac{(iy)^6}{6!} + \cdots = 1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \frac{y^6}{6!} + \cdots = \cos y$$

and the odd terms in this expansion are

$$iy + \frac{(iy)^3}{3!} + \frac{(iy)^5}{5!} + \cdots = i\Big(y - \frac{y^3}{3!} + \frac{y^5}{5!} + \cdots\Big) = i \sin y$$

Adding the even and odd terms together gives us that, under (EZ), e^{iy} is indeed equal to $\cos y + i \sin y$.[7] As a consequence, we have

$$e^{i\pi} = -1$$

which gives an amazing linking between calculus (e), geometry (π), algebra (i) and the basic number -1.

In the next lemma we verify that the complex exponential obeys a couple of familiar computational properties.

Lemma B.2.2.

(a) For any complex numbers z_1 and z_2,

$$e^{z_1+z_2} = e^{z_1} e^{z_2}$$

(b) For any complex number c,

$$\frac{\mathrm{d}}{\mathrm{d}t} e^{ct} = ce^{ct}$$

6 The equation $e^{iy} = \cos y + i \sin y$ is known as Euler's formula. Leonhard Euler (1707–1783) was a Swiss mathematician and physicist who spent most of his adult life in Saint Petersberg and Berlin. He gave the name π to the ratio of a circle's circumference to its diameter. He also developed the constant e. His collected works fill 92 volumes.

7 It is obvious that, in the special case that $z = x$ with x real, the definitions (EZ) and B.2.1 are equivalent. So to complete the proof of equivalence in the general case $z = x + iy$, it suffices to prove that $e^{x+iy} = e^x e^{iy}$ under both (EZ) and Definition B.2.1. For Definition B.2.1, this follows from Lemma B.2.2, below.

Proof. (a) For any two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$, with x_1, y_1, x_2, y_2 real,

$$\begin{aligned} e^{z_1}e^{z_2} &= e^{x_1}(\cos y_1 + i\sin y_1)e^{x_2}(\cos y_2 + i\sin y_2) \\ &= e^{x_1+x_2}(\cos y_1 + i\sin y_1)(\cos y_2 + i\sin y_2) \\ &= e^{x_1+x_2}\{(\cos y_1\cos y_2 - \sin y_1\sin y_2) + i(\cos y_1\sin y_2 + \cos y_2\sin y_1)\} \\ &= e^{x_1+x_2}\{\cos(y_1+y_2) + i\sin(y_1+y_2)\} \\ &\qquad\qquad \text{by the trig identities of Appendix A.8} \\ &= e^{(x_1+x_2)+i(y_1+y_2)} \\ &= e^{z_1+z_2} \end{aligned}$$

so that the familiar multiplication formula also applies to complex exponentials.

(b) For any real number t and any complex number $c = \alpha + i\beta$, with α, β real,

$$e^{ct} = e^{\alpha t + i\beta t} = e^{\alpha t}[\cos(\beta t) + i\sin(\beta t)]$$

so that the derivative with respect to t

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}e^{ct} &= \alpha e^{\alpha t}[\cos(\beta t) + i\sin(\beta t)] + e^{\alpha t}[-\beta\sin(\beta t) + i\beta\cos(\beta t)] \\ &= (\alpha + i\beta)e^{\alpha t}[\cos(\beta t) + i\sin(\beta t)] \\ &= ce^{ct} \end{aligned}$$

is also the familiar one.

□

B.2.2 ▸▸ Relationship with sin and cos.

Equation B.2.3.

When θ is a real number

$$\begin{aligned} e^{i\theta} &= \cos\theta + i\sin\theta \\ e^{-i\theta} &= \cos\theta - i\sin\theta = \overline{e^{i\theta}} \end{aligned}$$

are complex numbers of modulus one.

Solving for $\cos\theta$ and $\sin\theta$ (by adding and subtracting the two equations) gives

Equation B.2.4.

$$\begin{aligned} \cos\theta &= \frac{1}{2}(e^{i\theta} + e^{-i\theta}) &&= \Re e^{i\theta} \\ \sin\theta &= \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) &&= \Im e^{i\theta} \end{aligned}$$

Example B.2.5

These formulae make it easy derive trig identities. For example,

$$\begin{aligned}\cos\theta\cos\phi &= \frac{1}{4}\big(e^{i\theta}+e^{-i\theta}\big)\big(e^{i\phi}+e^{-i\phi}\big)\\ &= \frac{1}{4}\big(e^{i(\theta+\phi)}+e^{i(\theta-\phi)}+e^{i(-\theta+\phi)}+e^{-i(\theta+\phi)}\big)\\ &= \frac{1}{4}\big(e^{i(\theta+\phi)}+e^{-i(\theta+\phi)}+e^{i(\theta-\phi)}+e^{i(-\theta+\phi)}\big)\\ &= \frac{1}{2}\big(\cos(\theta+\phi)+\cos(\theta-\phi)\big)\end{aligned}$$

and, using $(a+b)^3 = a^3+3a^2b+3ab^2+b^3$,

$$\begin{aligned}\sin^3\theta &= -\frac{1}{8i}\big(e^{i\theta}-e^{-i\theta}\big)^3\\ &= -\frac{1}{8i}\big(e^{i3\theta}-3e^{i\theta}+3e^{-i\theta}-e^{-i3\theta}\big)\\ &= \frac{3}{4}\frac{1}{2i}\big(e^{i\theta}-e^{-i\theta}\big)-\frac{1}{4}\frac{1}{2i}\big(e^{i3\theta}-e^{-i3\theta}\big)\\ &= \frac{3}{4}\sin\theta-\frac{1}{4}\sin(3\theta)\end{aligned}$$

and

$$\begin{aligned}\cos(2\theta) &= \Re\big(e^{2\theta i}\big) = \Re\big(e^{i\theta}\big)^2\\ &= \Re\big(\cos\theta+i\sin\theta\big)^2\\ &= \Re\big(\cos^2\theta+2i\sin\theta\cos\theta-\sin^2\theta\big)\\ &= \cos^2\theta-\sin^2\theta\end{aligned}$$

Example B.2.5

B.2.3 ▸▸ Polar Coordinates.

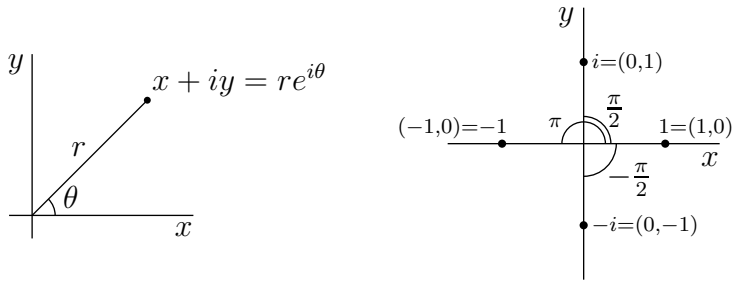
Let $z = x+iy$ be any complex number. Writing x and y in polar coordinates in the usual way, i.e. $x = r\cos(\theta)$, $y = r\sin(\theta)$, gives

$$x+iy = r\cos\theta+ir\sin\theta = re^{i\theta}$$

See the figure on the left below. In particular

$$\begin{aligned}1 &= e^{i0} &&= e^{2\pi i} &&= e^{2k\pi i} && \text{for } k=0,\pm1,\pm2,\cdots\\ -1 &= e^{i\pi} &&= e^{3\pi i} &&= e^{(1+2k)\pi i} && \text{for } k=0,\pm1,\pm2,\cdots\\ i &= e^{i\pi/2} &&= e^{\frac{5}{2}\pi i} &&= e^{(\frac{1}{2}+2k)\pi i} && \text{for } k=0,\pm1,\pm2,\cdots\\ -i &= e^{-i\pi/2} &&= e^{\frac{3}{2}\pi i} &&= e^{(-\frac{1}{2}+2k)\pi i} && \text{for } k=0,\pm1,\pm2,\cdots\end{aligned}$$

See the figure on the right below.



The polar coordinate $\theta = \arctan \frac{y}{x}$ associated with the complex number $z = x + iy$, i.e. the point (x, y) in the xy-plane, is also called the argument of z.

The polar coordinate representation makes it easy to find square roots, third roots and so on. Fix any positive integer n. The n^{th} roots of unity are, by definition, all solutions z of

$$z^n \;=\; 1$$

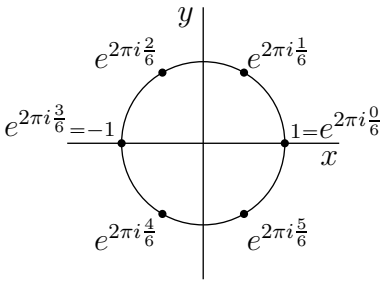
Writing $z = re^{i\theta}$

$$r^n e^{n\theta i} \;=\; 1e^{0i}$$

The polar coordinates (r, θ) and (r', θ') represent the same point in the xy-plane if and only if $r = r'$ and $\theta = \theta' + 2k\pi$ for some integer k. So $z^n = 1$ if and only if $r^n = 1$, i.e. $r = 1$, and $n\theta = 2k\pi$ for some integer k. The n^{th} roots of unity are all the complex numbers $e^{2\pi i \frac{k}{n}}$ with k integer. There are precisely n distinct n^{th} roots of unity because $e^{2\pi i \frac{k}{n}} = e^{2\pi i \frac{k'}{n}}$ if and only if $2\pi\frac{k}{n} - 2\pi\frac{k'}{n} = 2\pi\frac{k-k'}{n}$ is an integer multiple of 2π. That is, if and only if $k - k'$ is an integer multiple of n. The n distinct n^{th} roots of unity are

$$1 \;,\; e^{2\pi i \frac{1}{n}} \;,\; e^{2\pi i \frac{2}{n}} \;,\; e^{2\pi i \frac{3}{n}} \;,\; \cdots \;,\; e^{2\pi i \frac{n-1}{n}}$$

For example, the 6^{th} roots of unity are depicted below.



B.2.4 ▸▸ Exploiting Complex Exponentials in Calculus Computations

You have learned how to evaluate integrals involving trigonometric functions by using integration by parts, various trigonometric identities and various substitutions. It is often much easier to just use (B.2.3) and (B.2.4). Part of the utility of complex numbers comes from how well they interact with calculus through the exponential function. Here are two examples

Example B.2.6

$$\begin{aligned}\int e^x \cos x \, dx &= \frac{1}{2}\int e^x \left[e^{ix} + e^{-ix}\right] dx = \frac{1}{2}\int \left[e^{(1+i)x} + e^{(1-i)x}\right] dx \\ &= \frac{1}{2}\left[\frac{1}{1+i}e^{(1+i)x} + \frac{1}{1-i}e^{(1-i)x}\right] + C\end{aligned}$$

This form of the indefinite integral looks a little weird because of the i's. While it looks complex because of the i's, it is actually purely real (and correct), because $\frac{1}{1-i}e^{(1-i)x}$ is the complex conjugate of $\frac{1}{1+i}e^{(1+i)x}$. We can convert the indefinite integral into a more familiar form just by subbing back in $e^{\pm ix} = \cos x \pm i \sin x$, $\frac{1}{1+i} = \frac{1-i}{(1+i)(1-i)} = \frac{1-i}{2}$ and $\frac{1}{1-i} = \overline{\frac{1}{1+i}} = \frac{1+i}{2}$.

$$\begin{aligned}\int e^x \cos x \, dx &= \frac{1}{2}e^x\left[\frac{1}{1+i}e^{ix} + \frac{1}{1-i}e^{-ix}\right] + C \\ &= \frac{1}{2}e^x\left[\frac{1-i}{2}(\cos x + i\sin x) + \frac{1+i}{2}(\cos x - i \sin x)\right] + C \\ &= \frac{1}{2}e^x \cos x + \frac{1}{2}e^x \sin x + C\end{aligned}$$

You can quickly verify this by differentiating (or by comparing with Example 1.7.11).

Example B.2.6

Example B.2.7

Evaluating the integral $\int \cos^n x \, dx$ using the methods of Section 1.8 can be a real pain. It is much easier if we convert to complex exponentials. Using $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$,

$$\begin{aligned}\int \cos^4 x \, dx &= \frac{1}{2^4}\int \left[e^{ix} + e^{-ix}\right]^4 dx = \frac{1}{2^4}\int \left[e^{4ix} + 4e^{2ix} + 6 + 4e^{-2ix} + e^{-4ix}\right] dx \\ &= \frac{1}{2^4}\left[\frac{1}{4i}e^{4ix} + \frac{4}{2i}e^{2ix} + 6x + \frac{4}{-2i}e^{-2ix} + \frac{1}{-4i}e^{-4ix}\right] + C \\ &= \frac{1}{2^4}\left[\frac{1}{2}\frac{1}{2i}(e^{4ix} - e^{-4ix}) + \frac{4}{2i}(e^{2ix} - e^{-2ix}) + 6x\right] + C \\ &= \frac{1}{2^4}\left[\frac{1}{2}\sin 4x + 4\sin 2x + 6x\right] + C \\ &= \frac{1}{32}\sin 4x + \frac{1}{4}\sin 2x + \frac{3}{8}x + C\end{aligned}$$

Example B.2.7

B.2.5 ▸▸ Exploiting Complex Exponentials in Differential Equation Computations

Complex exponentials are also widely used to simplify the process of guessing solutions to ordinary differential equations. We'll start with (possibly a review of) some basic definitions and facts about differential equations.

Definition B.2.8.

(a) A *differential equation* is an equation for an unknown function that contains the derivatives of that unknown function. For example $y''(t) + y(t) = 0$ is a differential equation for the unknown function $y(t)$.

(b) In the differential calculus text CLP-1, we treated only derivatives of functions of one variable. Such derivatives are called ordinary derivatives. A differential equation is called an *ordinary differential equation* (often shortened to "ODE") if only ordinary derivatives appear. That is, if the unknown function has only a single independent variable.

In CLP-3 we will treat derivatives of functions of more than one variable. For example, let $f(x, y)$ be a function of two variables. If you treat y as a constant and take the derivative of the resulting function of the single variable x, the result is called the partial derivative of f with respect to x. A differential equation is called a *partial differential equation* (often shortened to "PDE") if partial derivatives appear. That is, if the unknown function has more than one independent variable. For example $y''(t) + y(t) = 0$ is an ODE while $\frac{\partial^2 u}{\partial t^2}(x,t) = c^2 \frac{\partial^2 u}{\partial x^2}(x,t)$ is a PDE.

(c) The *order* of a differential equation is the order of the highest derivative that appears. For example $y''(t) + y(t) = 0$ is a second order ODE.

(d) An ordinary differential equation that is of the form

$$a_0(t)y^{(n)}(t) + a_1(t)y^{(n-1)}(t) + \cdots + a_{n-1}(t)y'(t) + a_n(t)y(t) = F(t) \quad \text{(B.2.1)}$$

with given coefficient functions $a_0(t), \cdots, a_n(t)$ and $F(t)$ is said to be *linear*. Otherwise, the ODE is said to be *nonlinear*. For example, $y'(t)^2 + y(t) = 0$, $y'(t)y''(t) + y(t) = 0$ and $y'(t) = e^{y(t)}$ are all nonlinear.

(e) The ODE (B.2.1) is said to have *constant coefficients* if the coefficients $a_0(t)$, $a_1(t)$, $\cdots$, $a_n(t)$ are all constants. Otherwise, it is said to have *variable coefficients*. For example, the ODE $y''(t) + 7y(t) = \sin t$ is constant coefficient, while $y''(t) + ty(t) = \sin t$ is variable coefficient.

(f) The ODE (B.2.1) is said to be *homogeneous* if $F(t)$ is identically zero. Otherwise, it is said to be *inhomogeneous* or *nonhomogeneous*. For example, the ODE $y''(t) + 7y(t) = 0$ is homogeneous, while $y''(t) + 7y(t) = \sin t$ is inhomogeneous. A homogeneous ODE always has the trivial solution $y(t) = 0$.

Definition B.2.8 (continued).

(g) An *initial value problem* is a problem in which one is to find an unknown function $y(t)$ that satisfies both a given ODE and given initial conditions, like $y(t_0) = 1, y'(t_0) = 0$. Note that all of the conditions involve the function $y(t)$ (or its derivatives) evaluated at a single time $t = t_0$.

(h) A *boundary value problem* is a problem in which one is to find an unknown function $y(t)$ that satisfies both a given ODE and given boundary conditions, like $y(t_0) = 0, y(t_1) = 0$. Note that the conditions involve the function $y(t)$ (or its derivatives) evaluated at two different times.

The following theorem gives the form of solutions to the linear[8] ODE (B.2.1).

Theorem B.2.9.

Assume that the coefficients $a_0(t), a_1(t), \cdots, a_{n-1}(t), a_n(t)$ and $F(t)$ are continuous functions and that $a_0(t)$ is not zero.

(a) The general solution to the linear ODE (B.2.1) is of the form

$$y(t) = y_p(t) + C_1 y_1(t) + C_2 y_2(t) + \cdots + C_n y_n(t) \tag{B.2.2}$$

where

- n is the order of (B.2.1)
- $y_p(t)$ is **any** solution to (B.2.1)
- $C_1, C_2, \cdots, C_n$ are arbitrary constants
- $y_1, y_2, \cdots, y_n$ are n independent solutions to the homogenous equation

$$a_0(t)y^{(n)}(t) + a_1(t)y^{(n-1)}(t) + \cdots + a_{n-1}(t)y'(t) + a_n(t)y(t) = 0$$

associated to (B.2.1). "Independent" just means that no y_i can be written as a linear combination of the other y_j's. For example, $y_1(t)$ cannot be expressed in the form $b_2y_2(t) + \cdots + b_ny_n(t)$.

In (B.2.2), y_p is called the "particular solution" and $C_1y_1(t) + C_2y_2(t) + \cdots + C_ny_n(t)$ is called the "complementary solution".

(b) Given any constants $b_0, \cdots, b_{n-1}$ there is exactly one function $y(t)$ that obeys the ODE (B.2.1) and the initial conditions

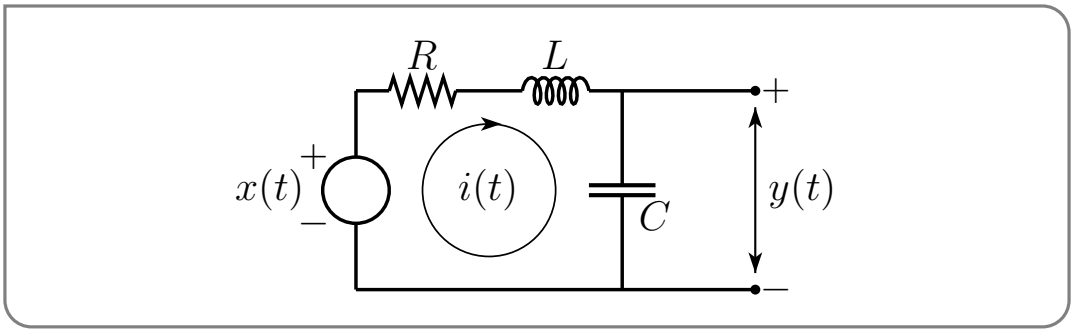
$$y(0) = b_0 \qquad y'(0) = b_1 \qquad \cdots \qquad y^{(n-1)}(0) = b_{n-1}$$

8 There are a some special classes of nonlinear ODE's, like the separable differential equations of §2.4, that are relatively easy to solve. But generally, nonlinear ODE's are much harder to solve than linear ODE's.

In the following example we'll derive one widely used linear constant coefficient ODE.

Example B.2.10 (RLC circuit)

As an example of how ODE's arise, we consider the RLC circuit, which is the electrical circuit consisting of a resistor of resistance R, a coil (or solenoid) of inductance L, a capacitor of capacitance C and a voltage source arranged in series, as shown below. Here R, L and C are all nonnegative constants.



We're going to think of the voltage $x(t)$ as an input signal, and the voltage $y(t)$ as an output signal. The goal is to determine the output signal produced by a given input signal. If $i(t)$ is the current flowing at time t in the loop as shown and $q(t)$ is the charge on the capacitor, then the voltages across R, L and C, respectively, at time t are $Ri(t)$, $L\frac{\mathrm{d}i}{\mathrm{d}t}(t)$ and $y(t) = \frac{q(t)}{C}$. By the Kirchhoff's law[9] that says that the voltage between any two points has to be independent of the path used to travel between the two points, these three voltages must add up to $x(t)$ so that

$$Ri(t) + L\frac{\mathrm{d}i}{\mathrm{d}t}(t) + \frac{q(t)}{C} = x(t) \tag{B.2.3}$$

Assuming that R, L, C and $x(t)$ are known, this is still one differential equation in two unknowns, the current $i(t)$ and the charge $q(t)$. Fortunately, there is a relationship between the two. Because the current entering the capacitor is the rate of change of the charge on the capacitor

$$i(t) = \frac{\mathrm{d}q}{\mathrm{d}t}(t) = Cy'(t) \tag{B.2.4}$$

This just says that the capacitor cannot create or destroy charge on its own; all charging of the capacitor must come from the current. Substituting (B.2.4) into (B.2.3) gives

$$LCy''(t) + RCy'(t) + y(t) = x(t)$$

which is a second order linear constant coefficient ODE. As a concrete example, we'll take an ac voltage source and choose the origin of time so that $x(0) = 0$, $x(t) = E_0 \sin(\omega t)$. Then the differential equation becomes

$$LCy''(t) + RCy'(t) + y(t) = E_0 \sin(\omega t) \tag{B.2.5}$$

9 Gustav Robert Kirchhoff (1824–1887) was a German physicist. There are several sets of Kirchhoff's laws that are named after him — Kirchhoff's circuit laws, that we are using in this example, Krichhoff's spectroscopy laws and Kirchhoff's law of thermochemistry. Kirchhoff and his collaborator Robert Bunsen, of Bunsen burner fame, invented the spectroscope.

Example B.2.10

Finally, here are two examples in which we use complex exponentials to solve an ODE.

Example B.2.11

By Theorem B.2.9(a), the general solution to the ordinary differential equation

$$y''(t) + 4y'(t) + 5y(t) = 0 \qquad \text{(ODE)}$$

is of the form $C_1u_1(t) + C_2u_2(t)$ with $u_1(t)$ and $u_2(t)$ being two (independent) solutions to (ODE) and with C_1 and C_2 being arbitrary constants. The easiest way to find $u_1(t)$ and $u_2(t)$ is to guess them. And the easiest way to guess them is to try[10] $y(t) = e^{rt}$, with r being a constant to be determined. Substituting $y(t) = e^{rt}$ into (ODE) gives

$$r^2e^{rt} + 4re^{rt} + 5e^{rt} = 0 \iff \big(r^2 + 4r + 5\big)e^{rt} = 0 \iff r^2 + 4r + 5 = 0$$

This quadratic equation for r can be solved either by using the high school formula or by completing the square.

$$\begin{aligned} r^2 + 4r + 5 = 0 &\iff (r+2)^2 + 1 = 0 \iff (r+2)^2 = -1 \iff r + 2 = \pm i \\ &\iff r = -2 \pm i \end{aligned}$$

So the general solution to (ODE) is

$$y(t) = C_1e^{(-2+i)t} + C_2e^{(-2-i)t}$$

This is one way to write the general solution, but there are many others. In particular there are quite a few people in the world who are (foolishly) afraid[11] of complex exponentials. We can hide them by using (B.2.3) and (B.2.4).

$$\begin{aligned} y(t) &= C_1e^{(-2+i)t} + C_2e^{(-2-i)t} = C_1e^{-2t}e^{it} + C_2e^{-2t}e^{-it} \\ &= C_1e^{-2t}\big(\cos t + i\sin t\big) + C_2e^{-2t}\big(\cos t - i\sin t\big) \\ &= (C_1 + C_2)e^{-2t}\cos t + (iC_1 - iC_2)e^{-2t}\sin t \\ &= D_1e^{-2t}\cos t + D_2e^{-2t}\sin t \end{aligned}$$

with $D_1 = C_1 + C_2$ and $D_2 = iC_1 - iC_2$ being two other arbitrary constants. Don't make the mistake of thinking that D_2 must be complex because i appears in the formula $D_2 = iC_1 - iC_2$ relating D_2 and C_1, C_2. No one said that C_1 and C_2 are real numbers. In fact, in typical applications, the arbitrary constants are determined by initial conditions and often D_1 and D_2 turn out to be real and C_1 and C_2 turn out to be complex. For example, the initial conditions $y(0) = 0$, $y'(0) = 2$ force

$$\begin{aligned} 0 &= y(0) = C_1 + C_2 \\ 2 &= y'(0) = (-2+i)C_1 + (-2-i)C_2 \end{aligned}$$

10 The reason that $y(t) = e^{rt}$ is a good guess is that, with this guess, all of $y(t)$, $y'(t)$ and $y''(t)$ are constants times e^{rt}. So the left hand side of the differential equation is also a constant, that depends on r, times e^{rt}. So we just have to choose r so that the constant is zero.

11 Embracing the complexity leads to simplicity.

The first equation gives $C_2 = -C_1$ and then the second equation gives

$$(-2+i)C_1 - (-2-i)C_1 = 2 \iff 2iC_1 = 2 \iff iC_1 = 1 \iff C_1 = -i,\ C_2 = i$$

and

$$D_1 = C_1 + C_2 = 0 \qquad D_2 = iC_1 - iC_2 = 2$$

Example B.2.11

Example B.2.12

We shall now guess one solution (i.e. a particular solution) to the differential equation

$$y''(t) + 2y'(t) + 3y(t) = \cos t \tag{ODE1}$$

Equations like this arise, for example, in the study of the RLC circuit. We shall simplify the computation by exploiting that $\cos t = \Re e^{it}$. First, we shall guess a function $Y(t)$ obeying

$$Y'' + 2Y' + 3Y = e^{it} \tag{ODE2}$$

Then, taking complex conjugates,

$$\bar{Y}'' + 2\bar{Y}' + 3\bar{Y} = e^{-it} \tag{$\overline{\text{ODE2}}$}$$

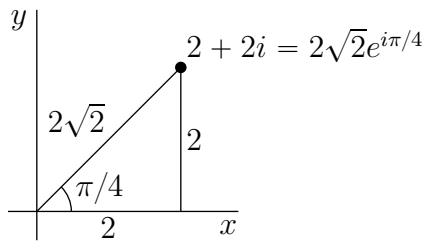
and, adding $\frac{1}{2}$(ODE2) and $\frac{1}{2}(\overline{\text{ODE2}})$ together will give

$$(\Re Y)'' + 2(\Re Y)' + 3(\Re Y) = \Re e^{it} = \cos t$$

which shows that $\Re Y(t)$ is a solution to (ODE1). Let's try $Y(t) = Ae^{it}$, with A a constant to be determined. This is a solution of (ODE2) if and only if

$$\begin{aligned} & \frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(Ae^{it}\right) + 2\frac{\mathrm{d}}{\mathrm{d}t}\left(Ae^{it}\right) + 3Ae^{it} = e^{it} \\ \iff \quad & (i^2 + 2i + 3)Ae^{it} = e^{it} \\ \iff \quad & A = \frac{1}{2+2i} \end{aligned}$$

So $\frac{e^{it}}{2+2i}$ is a solution to (ODE2) and $\Re\frac{e^{it}}{2+2i}$ is a solution to (ODE1). To simplify this, write $2+2i$ in polar coordinates. From the sketch



we have $2+2i = 2\sqrt{2}e^{i\frac{\pi}{4}}$. So

$$\frac{e^{it}}{2+2i} = \frac{e^{it}}{2\sqrt{2}e^{i\frac{\pi}{4}}} = \frac{1}{2\sqrt{2}}e^{i(t-\frac{\pi}{4})}$$
$$\implies \Re\frac{e^{it}}{2+2i} = \frac{1}{2\sqrt{2}}\Re e^{i(t-\frac{\pi}{4})} = \frac{1}{2\sqrt{2}}\cos\left(t-\frac{\pi}{4}\right)$$

Example B.2.12

Appendix C

More About Numerical Integration

C.1▲ Richardson Extrapolation

There are many approximation procedures in which one first picks a step size h and then generates an approximation $A(h)$ to some desired quantity $\mathcal{A}$. For example, $\mathcal{A}$ might be the value of some integral $\int_a^b f(x)\,dx$. For the trapezoidal rule with n steps, $\Delta x = \frac{b-a}{n}$ plays the role of the step size. Often the order of the error generated by the procedure is known. This means

$$\mathcal{A} = A(h) + Kh^k + K_1h^{k+1} + K_2h^{k+2} + \cdots \tag{E1}$$

with k being some known constant, called the order of the error, and $K,\ K_1,\ K_2,\ \cdots$ being some other (usually unknown) constants. If $A(h)$ is the approximation to $\mathcal{A} = \int_a^b f(x)\,dx$ produced by the trapezoidal rule with $\Delta x = h$, then $k = 2$. If Simpson's rule is used, $k = 4$.

Let's first suppose that h is small enough that the terms $K_1h^{k+1} + K_2h^{k+2} + \cdots$ in (E1) are small enough[1] that dropping them has essentially no impact. This would give

$$\mathcal{A} = A(h) + Kh^k \tag{E2}$$

Imagine that we know k, but that we do not know $\mathcal{A}$ or K, and think of (E2) as an equation that the unknowns $\mathcal{A}$ and K have to solve. It may look like we have one equation in the two unknowns K, $\mathcal{A}$, but that is *not* the case. The reason is that (E2) is (essentially) true for all (sufficiently small) choices of h. If we pick some h, say h_1, and use the algorithm to determine $A(h_1)$ then (E2), with h replaced by h_1, gives one equation in the two unknowns $\mathcal{A}$ and K, and if we then pick some different h, say h_2, and use the algorithm a second time to determine $A(h_2)$ then (E2), with h replaced by h_2, gives a second equation in the two unknowns $\mathcal{A}$ and K. The two equations will then determine both $\mathcal{A}$ and K.

1 Typically, we don't have access to, and don't care about, the exact error. We only care about its order of magnitude. So if h is small enough that $K_1h^{k+1} + K_2h^{k+2} + \cdots$ is a factor of at least, for example, one hundred smaller than Kh^k, then dropping $K_1h^{k+1} + K_2h^{k+2} + \cdots$ would not bother us at all.

To be more concrete, suppose that we have picked some specific value of h, and have chosen $h_1 = h$ and $h_2 = \frac{h}{2}$, and that we have evaluated $A(h)$ and $A(h/2)$. Then the two equations are

$$\mathcal{A} = A(h) + Kh^k \tag{E3a}$$

$$\mathcal{A} = A(h/2) + K\left(\tfrac{h}{2}\right)^k \tag{E3b}$$

While these equations are nonlinear in h, they are linear in the constants K and $\mathcal{A}$, so it is easy to solve for both K and $\mathcal{A}$. To get K, just subtract (E3b) from (E3a).

$$(\text{E3a}) - (\text{E3b}): \quad 0 = A(h) - A(h/2) + \left(1 - \tfrac{1}{2^k}\right)Kh^k \quad \Longrightarrow \quad K = \frac{A(h/2) - A(h)}{[1 - 2^{-k}]h^k} \tag{E4a}$$

To get $\mathcal{A}$, multiply (E3b) by 2^k and then subtract (E3a).

$$2^k(\text{E3b}) - (\text{E3a}): \quad [2^k - 1]\mathcal{A} = 2^k A(h/2) - A(h) \quad \Longrightarrow \quad \mathcal{A} = \frac{2^k A(h/2) - A(h)}{2^k - 1} \tag{E4b}$$

The generation of a "new improved" approximation for $\mathcal{A}$ from two $A(h)$'s with different values of h is called Richardson[2] Extrapolation. Here is a summary

Equation C.1.1 (Richardson extrapolation).

Let $A(h)$ be a step size h approximation to $\mathcal{A}$. If

$$\mathcal{A} = A(h) + Kh^k$$

then

$$K = \frac{A(h/2) - A(h)}{[1 - 2^{-k}]h^k} \qquad \mathcal{A} = \frac{2^k A(h/2) - A(h)}{2^k - 1}$$

This works very well since, by computing $A(h)$ for two different h's, we can remove the biggest error term in (E1), and so get a much more precise approximation to $\mathcal{A}$ for little additional work.

Example C.1.2

Applying the trapezoidal rule (1.11.6) to the integral $\mathcal{A} = \int_0^1 \frac{4}{1+x^2}\mathrm{d}x$ with step sizes $\frac{1}{8}$ and $\frac{1}{16}$ (i.e. with $n = 8$ and $n = 16$) gives, with $h = \frac{1}{8}$,

$$A(h) = 3.1389884945 \qquad A(h/2) = 3.1409416120$$

So (E4b), with $k = 2$, gives us the "new improved" approximation

$$\frac{2^2 \times 3.1409416120 - 3.1389884945}{2^2 - 1} = 3.1415926512$$

We saw in Example 1.11.3 that $\int_0^1 \frac{4}{1+x^2}\mathrm{d}x = \pi$, so this new approximation really is "improved":

2 Richardson extrapolation was introduced by the Englishman Lewis Fry Richardson (1881–1953) in 1911.

- $A(1/8)$ agrees with π to two decimal places,
- $A(1/16)$ agrees with π to three decimal places and
- the new approximation agrees with π to eight decimal places.

Beware that (E3b), namely $\mathcal{A} = A(h/2) + K\left(\frac{h}{2}\right)^k$, is saying that $K\left(\frac{h}{2}\right)^k$ is (approximately) the error in $A(h/2)$, *not* the error in $\mathcal{A}$. You cannot get an "even more improved" approximation by using (E4a) to compute K and then adding $K\left(\frac{h}{2}\right)^k$ to the "new improved" $\mathcal{A}$ of (E4b) — doing so just gives $\mathcal{A} + K\left(\frac{h}{2}\right)^k$, not a more accurate $\mathcal{A}$.

Example C.1.2

Example C.1.3 (Example 1.11.15 revisited)

Suppose again that we wish to use Simpson's rule (1.11.9) to evaluate $\int_0^1 e^{-x^2}\,dx$ to within an accuracy of 10^{-6}, but that we do not need the degree of certainty provided by Example 1.11.15. Observe that we need (approximately) that $|K|h^4 < 10^{-6}$, so if we can estimate K (using our Richardson trick) then we can estimate the required h. A commonly used strategy, based on this observation, is to

- first apply Simpson's rule twice with some relatively small number of steps and
- then use (E4a), with $k = 4$, to estimate K and
- then use the condition $|K|h^k \leqslant 10^{-6}$ to determine, approximately, the number of steps required
- and finally apply Simpson's rule with the number of steps just determined.

Let's implement this strategy. First we estimate K by applying Simpson's rule with step sizes $\frac{1}{4}$ and $\frac{1}{8}$. Writing $\frac{1}{4} = h'$, we get

$$A(h') = 0.74685538 \qquad A(h'/2) = 0.74682612$$

so that (E4a), with $k = 4$ and h replaced by h', yields

$$K = \frac{0.74682612 - 0.74685538}{[1 - 2^{-4}](1/4)^4} = -7.990 \times 10^{-3}$$

We want to use a step size h obeying

$$|K|h^4 \leqslant 10^{-6} \iff 7.990 \times 10^{-3} h^4 \leqslant 10^{-6} \iff h \leqslant \sqrt[4]{\frac{1}{7990}} = \frac{1}{9.45}$$

like, for example, $h = \frac{1}{10}$. Applying Simpson's rule with $h = \frac{1}{10}$ gives

$$A(1/10) = 0.74682495$$

The exact answer, to eight decimal places, is 0.74682413 so the error in $A(1/10)$ is indeed just under 10^{-6}.

Suppose now that we change our minds. We want an accuracy of 10^{-12}, rather than 10^{-6}. We have already estimated K. So now we want to use a step size h obeying

$$|K|h^4 \leqslant 10^{-12} \iff 7.99 \times 10^{-3} h^4 \leqslant 10^{-12} \iff h \leqslant \sqrt[4]{\frac{1}{7.99 \times 10^9}} = \frac{1}{299.0}$$

like, for example, $h = \frac{1}{300}$. Applying Simpson's rule with $h = \frac{1}{300}$ gives, to fourteen decimal places,

$$A(1/300) = 0.74682413281344$$

The exact answer, to fourteen decimal places, is 0.74682413281243 so the error in $A(1/300)$ is indeed just over 10^{-12}.

Example C.1.3

C.2▲ Romberg Integration

The formulae (E4a,b) for K and $\mathcal{A}$ are, of course, only[3] approximate since they are based on (E2), which is an approximation to (E1). Let's repeat the derivation that leads to (E4), but using the full (E1),

$$\mathcal{A} = A(h) + Kh^k + K_1 h^{k+1} + K_2 h^{k+2} + \cdots$$

Once again, suppose that we have chosen some h and that we have evaluated $A(h)$ and $A(h/2)$. They obey

$$\mathcal{A} = A(h) + Kh^k + K_1 h^{k+1} + K_2 h^{k+2} + \cdots \tag{E5a}$$

$$\mathcal{A} = A(h/2) + K\left(\tfrac{h}{2}\right)^k + K_1\left(\tfrac{h}{2}\right)^{k+1} + K_2\left(\tfrac{h}{2}\right)^{k+2} + \cdots \tag{E5b}$$

Now, as we did in the derivation of (E4b), multiply (E5b) by 2^k and then subtract (E5a). This gives

$$\left(2^k - 1\right)\mathcal{A} = 2^k A(h/2) - A(h) - \tfrac{1}{2}K_1 h^{k+1} - \tfrac{3}{4}K_2 h^{k+2} + \cdots$$

and then, dividing across by $\left(2^k - 1\right)$,

$$\mathcal{A} = \frac{2^k A(h/2) - A(h)}{2^k - 1} - \frac{1/2}{2^k - 1}K_1 h^{k+1} - \frac{3/4}{2^k - 1}K_2 h^{k+2} + \cdots$$

Hence if we define our "new improved approximation"

$$B(h) = \frac{2^k A(h/2) - A(h)}{2^k - 1} \quad \text{and} \quad \tilde{K} = -\frac{1/2}{2^k - 1}K_1 \quad \text{and} \quad \tilde{K}_1 = -\frac{3/4}{2^k - 1}K_2 \tag{E6}$$

we have

$$\mathcal{A} = B(h) + \tilde{K}h^{k+1} + \tilde{K}_1 h^{k+2} + \cdots$$

3 "Only" is a bit strong. Don't underestimate the power of a good approximation (pun intended).

which says that $B(h)$ is an approximation to $\mathcal{A}$ whose error is of order $k+1$, one better[4] than $A(h)$'s.

If $A(h)$ has been computed for three values of h, we can generate $B(h)$ for two values of h and repeat the above procedure with a new value of k. And so on. One widely used numerical integration algorithm, called Romberg integration[5], applies this procedure repeatedly to the trapezoidal rule. It is known that the trapezoidal rule approximation $T(h)$ to an integral I has error behaviour (assuming that the integrand $f(x)$ is smooth)

$$I = T(h) + K_1 h^2 + K_2 h^4 + K_3 h^6 + \cdots$$

Only even powers of h appear. Hence

$T(h)$	has error of order 2, so that, using (E6) with $k=2$,
$T_1(h) = \frac{4T(h/2)-T(h)}{3}$	has error of order 4, so that, using (E6) with $k=4$,
$T_2(h) = \frac{16T_1(h/2)-T_1(h)}{15}$	has error of order 6, so that, using (E6) with $k=6$,
$T_3(h) = \frac{64T_2(h/2)-T_2(h)}{63}$	has error of order 8 and so on

We know another method which produces an error of order 4 — Simpson's rule. In fact, $T_1(h)$ is exactly Simpson's rule (for step size $\frac{h}{2}$).

Equation C.2.1 (Romberg integration).

Let $T(h)$ be the trapezoidal rule approximation, with step size h, to an integral I. The Romberg integration algorithm is

$$\begin{aligned} T_1(h) &= \frac{4T(h/2)-T(h)}{3} \\ T_2(h) &= \frac{16T_1(h/2)-T_1(h)}{15} \\ T_3(h) &= \frac{64T_2(h/2)-T_2(h)}{63} \\ &\vdots \\ T_k(h) &= \frac{2^{2k}T_{k-1}(h/2)-T_{k-1}(h)}{2^{2k}-1} \\ &\vdots \end{aligned}$$

Example C.2.2

The following table[6] illustrates Romberg integration by applying it to the area A of the integral $A = \int_0^1 \frac{4}{1+x^2}\mathrm{d}x$. The exact value of this integral is π which is 3.14159265358979, to fourteen decimal places.

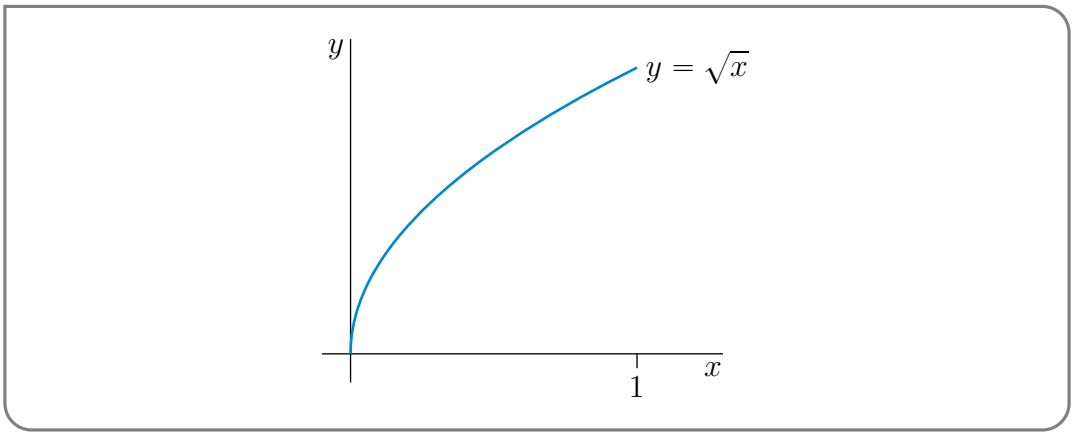
4 That is, the error decays as h^{k+1} as opposed to h^k — so, as h decreases, it gets smaller faster.
5 Romberg Integration was introduced by the German Werner Romberg (1909–2003) in 1955.
6 The second column, for example, of the table only reports 5 decimal places for $T(h)$. But many more decimal places of $T(h)$ were used in the computations of $T_1(h)$ etc.

h	$T(h)$	$T_1(h)$	$T_2(h)$	$T_3(h)$
1/4	3.13118	3.14159250246	3.14159266114	3.14159265359003
1/8	3.13899	3.141592651225	3.141592653708	
1/16	3.14094	3.141592653553		
1/32	3.14143			

This computation required the evaluation of $f(x) = \frac{4}{1+x^2}$ only for $x = n/32$ with $0 \leqslant n \leqslant 32$ — that is, a total of 33 evaluations of f. Those 33 evaluations gave us 12 correct decimal places. By way of comparison, $T\big(1/32\big)$ used the same 33 evaluations of f, but only gave us 3 correct decimal places.

Example C.2.2

As we have seen, Richardson extrapolation can be used to choose the step size so as to achieve some desired degree of accuracy. We are next going to consider a family of algorithms that extend this idea to use small step sizes in the part of the domain of integration where it is hard to get good accuracy and large step sizes in the part of the domain of integration where it is easy to get good accuracy. We will illustrate the ideas by applying them to the integral $\int_0^1 \sqrt{x}\,dx$. The integrand $\sqrt{x}$ changes very quickly when x is small and changes slowly when x is large. So we will make the step size small near $x = 0$ and make the step size large near $x = 1$.



C.3▲ Adaptive Quadrature

"Adaptive quadrature" refers to a family of algorithms that use small step sizes in the part of the domain of integration where it is hard to get good accuracy and large step sizes in the part of the domain of integration where it is easy to get good accuracy.

We'll illustrate the idea using Simpson's rule applied to the integral $\int_a^b f(x)\,dx$, and assuming that we want the error to be no more than (approximately) some fixed constant ε. For example, ε could be 10^{-6}. Denote by $S(a', b'; h')$, the answer given when Simpson's rule is applied to the integral $\int_{a'}^{b'} f(x)\,dx$ with step size h'.

- *Step 1.* We start by applying Simpson's rule, combined with Richardson extrapolation so as to get an error estimate, with the largest possible step size h. Namely, set

$h = \frac{b-a}{2}$ and compute

$$f(a) \quad f\big(a+\tfrac{h}{2}\big) \quad f(a+h) = f\big(\tfrac{a+b}{2}\big) \quad f\big(a+\tfrac{3h}{2}\big) \quad f(a+2h) = f(b)$$

Then

$$S\big(a,b\,;h\big) = \tfrac{h}{3}\big\{f(a) + 4f\big(a+h\big) + f(b)\big\}$$

and

$$\begin{aligned} S\big(a,b\,;\tfrac{h}{2}\big) &= \tfrac{h}{6}\big\{f(a) + 4f\big(a+\tfrac{h}{2}\big) + 2f\big(a+h\big) + 4f\big(a+\tfrac{3h}{2}\big) + f(b)\big\} \\ &= S\big(a,\tfrac{a+b}{2}\,;\tfrac{h}{2}\big) + S\big(\tfrac{a+b}{2},b\,;\tfrac{h}{2}\big) \end{aligned}$$

with

$$\begin{aligned} S\big(a,\tfrac{a+b}{2}\,;\,\tfrac{h}{2}\big) &= \tfrac{h}{6}\big\{f(a) + 4f\big(a+\tfrac{h}{2}\big) + f\big(\tfrac{a+b}{2}\big)\big\} \\ S\big(\tfrac{a+b}{2},b\,;\,\tfrac{h}{2}\big) &= \tfrac{h}{6}\big\{f\big(\tfrac{a+b}{2}\big) + 4f\big(a+\tfrac{3h}{2}\big) + f(b)\big\} \end{aligned}$$

Using the Richardson extrapolation formula (E4a) with $k = 4$ gives that the error in $S\big(a,b\,;\,\frac{h}{2}\big)$ is (approximately)

$$\big|K\big(\tfrac{h}{2}\big)^4\big| = \tfrac{1}{15}\Big|S\big(a,b\,;\,\tfrac{h}{2}\big) - S\big(a,b\,;\,h\big)\Big| \tag{E7}$$

If this is smaller than ε, we have (approximately) the desired accuracy and stop[7].

- *Step 2.* If (E7) is larger than ε, we divide the original integral $I = \int_a^b f(x)\,\mathrm{d}x$ into two "half-sized" integrals, $I_1 = \int_a^{\frac{a+b}{2}} f(x)\,\mathrm{d}x$ and $I_2 = \int_{\frac{a+b}{2}}^b f(x)\,\mathrm{d}x$ and repeat the procedure of Step 1 on each of them, but with h replaced by $\frac{h}{2}$ and ε replaced by $\frac{\varepsilon}{2}$ — if we can find an approximation, $\tilde{I}_1$, to I_1 with an error less than $\frac{\varepsilon}{2}$ and an approximation, $\tilde{I}_2$, to I_2 with an error less than $\frac{\varepsilon}{2}$, then $\tilde{I}_1 + \tilde{I}_2$ approximates I with an error less than ε. Here is more detail.
  - If the error in the approximation $\tilde{I}_1$ to I_1 and the error in the approximation $\tilde{I}_2$ to I_2 are both acceptable, then we use $\tilde{I}_1$ as our final approximation to I_1 and we use $\tilde{I}_2$ as our final approximation to I_2.
  - If the error in the approximation $\tilde{I}_1$ to I_1 is acceptable but the error in the approximation $\tilde{I}_2$ to I_2 is not acceptable, then we use $\tilde{I}_1$ as our final approximation to I_1 but we subdivide the integral I_2.
  - If the error in the approximation $\tilde{I}_1$ to I_1 is not acceptable but the error in the approximation $\tilde{I}_2$ to I_2 is acceptable, then we use $\tilde{I}_2$ as our final approximation to I_2 but we subdivide the integral I_1.
  - If the error in the approximation $\tilde{I}_1$ to I_1 and the error in the approximation $\tilde{I}_2$ to I_2 are both not acceptable, then we subdivide both of the integrals I_1 and I_2.

  So we *adapt* the step size as we go.

7 It is very common to build in a bit of a safety margin and require that, for example, $\big|K\big(\frac{h}{2}\big)^4\big|$ be smaller than $\frac{\varepsilon}{2}$ rather than ε.

- *Steps 3, 4, 5, $\cdots$* Repeat as required.

Example C.3.1

Let's apply adaptive quadrature using Simpson's rule as above with the goal of computing $\int_0^1 \sqrt{x}\,\mathrm{d}x$ with an error of at most $\varepsilon = 0.0005 = 5 \times 10^{-4}$. Observe that $\frac{\mathrm{d}}{\mathrm{d}x}\sqrt{x} = \frac{1}{2\sqrt{x}}$ blows up as x tends to zero. The integrand changes very quickly when x is small. So we will probably need to make the step size small near the limit of integration $x = 0$.

- *Step 1 — the interval* $[0,1]$. (The notation $[0,1]$ stands for the interval $0 \leqslant x \leqslant 1$.)

$$S(0,1\,;\tfrac{1}{2}) = 0.63807119$$
$$S(0,\tfrac{1}{2}\,;\tfrac{1}{4}) = 0.22559223$$
$$S(\tfrac{1}{2},1\,;\tfrac{1}{4}) = 0.43093403$$
$$\text{error} = \tfrac{1}{15}\left|S(0,\tfrac{1}{2}\,;\tfrac{1}{4}) + S(\tfrac{1}{2},1\,;\tfrac{1}{4}) - S(0,1\,;\tfrac{1}{2})\right| = 0.0012 > \varepsilon = 0.0005$$

  This is unacceptably large, so we subdivide the interval $[0,1]$ into the two halves $\left[0,\frac{1}{2}\right]$ and $\left[\frac{1}{2},1\right]$ and apply the procedure separately to each half.

- *Step 2a — the interval* $[0,\frac{1}{2}]$.

$$S(0,\tfrac{1}{2}\,;\tfrac{1}{4}) = 0.22559223$$
$$S(0,\tfrac{1}{4}\,;\tfrac{1}{8}) = 0.07975890$$
$$S(\tfrac{1}{4},\tfrac{1}{2}\,;\tfrac{1}{8}) = 0.15235819$$
$$\text{error} = \tfrac{1}{15}\left|S(0,\tfrac{1}{4}\,;\tfrac{1}{8}) + S(\tfrac{1}{4},\tfrac{1}{2}\,;\tfrac{1}{8}) - S(0,\tfrac{1}{2}\,;\tfrac{1}{4})\right| = 0.00043 > \tfrac{\varepsilon}{2} = 0.00025$$

  This error is unacceptably large.

- *Step 2b — the interval* $[\frac{1}{2},1]$.

$$S(\tfrac{1}{2},1\,;\tfrac{1}{4}) = 0.43093403$$
$$S(\tfrac{1}{2},\tfrac{3}{4}\,;\tfrac{1}{8}) = 0.19730874$$
$$S(\tfrac{3}{4},1\,;\tfrac{1}{8}) = 0.23365345$$
$$\text{error} = \tfrac{1}{15}\left|S(\tfrac{1}{2},\tfrac{3}{4}\,;\tfrac{1}{8}) + S(\tfrac{3}{4},1\,;\tfrac{1}{8}) - S(\tfrac{1}{2},1\,;\tfrac{1}{4})\right| = 0.0000019 < \tfrac{\varepsilon}{2} = 0.00025$$

  This error is acceptable.

- *Step 2 resumé.* The error for the interval $[\frac{1}{2},1]$ is small enough, so we accept

$$S(\tfrac{1}{2},1\,;\tfrac{1}{8}) = S(\tfrac{1}{2},\tfrac{3}{4}\,;\tfrac{1}{8}) + S(\tfrac{3}{4},1\,;\tfrac{1}{8}) = 0.43096219$$

  as the approximate value of $\int_{1/2}^1 \sqrt{x}\,\mathrm{d}x$.

  The error for the interval $[0,\frac{1}{2}]$ is unacceptably large, so we subdivide the interval $[0,\frac{1}{2}]$ into the two halves $[0,\frac{1}{4}]$ and $[\frac{1}{4},\frac{1}{2}]$ and apply the procedure separately to each half.

- *Step 3a — the interval* $[0, \frac{1}{4}]$.

$$S(0, \tfrac{1}{4}\,;\tfrac{1}{8}) = 0.07975890$$
$$S(0, \tfrac{1}{8}\,;\tfrac{1}{16}) = 0.02819903$$
$$S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{16}) = 0.05386675$$
$$\text{error} = \tfrac{1}{15}\left|S(0, \tfrac{1}{8}\,;\tfrac{1}{16}) + S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{16}) - S(0, \tfrac{1}{4}\,;\tfrac{1}{8})\right| = 0.000153792 > \tfrac{\varepsilon}{4} = 0.000125$$

  This error is unacceptably large.

- *Step 3b — the interval* $[\frac{1}{4}, \frac{1}{2}]$.

$$S(\tfrac{1}{4}, \tfrac{1}{2}\,;\tfrac{1}{8}) = 0.15235819$$
$$S(\tfrac{1}{4}, \tfrac{3}{8}\,;\tfrac{1}{16}) = 0.06975918$$
$$S(\tfrac{3}{8}, \tfrac{1}{2}\,;\tfrac{1}{16}) = 0.08260897$$
$$\text{error} = \tfrac{1}{15}\left|S(\tfrac{1}{4}, \tfrac{3}{8}\,;\tfrac{1}{16}) + S(\tfrac{3}{8}, \tfrac{1}{2}\,;\tfrac{1}{16}) - S(\tfrac{1}{4}, \tfrac{1}{2}\,;\tfrac{1}{8})\right| = 0.00000066 < \tfrac{\varepsilon}{4} = 0.000125$$

  This error is acceptable.

- *Step 3 resumé.* The error for the interval $[\frac{1}{4}, \frac{1}{2}]$ is small enough, so we accept

$$S(\tfrac{1}{4}, \tfrac{1}{2}\,;\tfrac{1}{16}) = S(\tfrac{1}{4}, \tfrac{3}{8}\,;\tfrac{1}{16}) + S(\tfrac{3}{8}, \tfrac{1}{2}\,;\tfrac{1}{16}) = 0.15236814$$

  as the approximate value of $\int_{1/4}^{1/2} \sqrt{x}\,\mathrm{d}x$.

  The error for the interval $[0, \frac{1}{4}]$ is unacceptably large, so we subdivide the interval $[0, \frac{1}{4}]$ into the two halves $[0, \frac{1}{8}]$ and $[\frac{1}{8}, \frac{1}{4}]$ and apply the procedure separately to each half.

- *Step 4a — the interval* $[0, \frac{1}{8}]$.

$$S(0, \tfrac{1}{8}\,;\tfrac{1}{16}) = 0.02819903$$
$$S(0, \tfrac{1}{16}\,;\tfrac{1}{32}) = 0.00996986$$
$$S(\tfrac{1}{16}, \tfrac{1}{8}\,;\tfrac{1}{32}) = 0.01904477$$
$$\text{error} = \tfrac{1}{15}\left|S(0, \tfrac{1}{16}\,;\tfrac{1}{32}) + S(\tfrac{1}{16}, \tfrac{1}{8}\,;\tfrac{1}{32}) - S(0, \tfrac{1}{8}\,;\tfrac{1}{16})\right| = 0.000054 < \tfrac{\varepsilon}{8} = 0.0000625$$

  This error is acceptable.

- *Step 4b — the interval* $[\frac{1}{8}, \frac{1}{4}]$.

$$S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{16}) = 0.05386675$$
$$S(\tfrac{1}{8}, \tfrac{3}{16}\,;\tfrac{1}{32}) = 0.02466359$$
$$S(\tfrac{3}{16}, \tfrac{1}{4}\,;\tfrac{1}{32}) = 0.02920668$$
$$\text{error} = \tfrac{1}{15}\left|S(\tfrac{1}{8}, \tfrac{3}{16}\,;\tfrac{1}{32}) + S(\tfrac{3}{6}, \tfrac{1}{4}\,;\tfrac{1}{32}) - S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{16})\right| = 0.00000024 < \tfrac{\varepsilon}{8} = 0.0000625$$

  This error is acceptable.

- *Step 4 resumé.* The error for the interval $[0, \frac{1}{8}]$ is small enough, so we accept

  $$S(0, \tfrac{1}{8}\,;\tfrac{1}{32}) = S(0, \tfrac{1}{16}\,;\tfrac{1}{32}) + S(\tfrac{1}{16}, \tfrac{1}{8}\,;\tfrac{1}{32}) = 0.02901464$$

  as the approximate value of $\int_0^{1/8} \sqrt{x}\,\mathrm{d}x$.

  The error for the interval $[\frac{1}{8}, \frac{1}{4}]$ is small enough, so we accept

  $$S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{32}) = S(\tfrac{1}{8}, \tfrac{3}{16}\,;\tfrac{1}{32}) + S(\tfrac{3}{16}, \tfrac{1}{4}\,;\tfrac{1}{32}) = 0.05387027$$

  as the approximate value of $\int_{1/8}^{1/4} \sqrt{x}\,\mathrm{d}x$.

- *Conclusion* The approximate value for $\int_0^1 \sqrt{x}\,\mathrm{d}x$ is

  $$S(0, \tfrac{1}{8}\,;\tfrac{1}{32}) + S(\tfrac{1}{8}, \tfrac{1}{4}\,;\tfrac{1}{32}) + S(\tfrac{1}{4}, \tfrac{1}{2}\,;\tfrac{1}{16}) + S(\tfrac{1}{2}, 1\,;\tfrac{1}{8}) = 0.66621525 \qquad \text{(E8)}$$

Of course the exact value of $\int_0^1 \sqrt{x}\,\mathrm{d}x = \frac{2}{3}$, so the actual error in our approximation is

$$\tfrac{2}{3} - 0.66621525 = 0.00045 < \varepsilon = 0.0005$$

Here is what Simpson's rule gives us when applied with some fixed step sizes.

$$\begin{aligned} S(0, 1\,;\tfrac{1}{8}) &= 0.66307928 \\ S(0, 1\,;\tfrac{1}{16}) &= 0.66539819 \\ S(0, 1\,;\tfrac{1}{32}) &= 0.66621818 \\ S(0, 1\,;\tfrac{1}{64}) &= 0.66650810 \end{aligned}$$

So to get an error comparable to that in (E8) from Simpson's rule with a fixed step size, we need to use $h = \frac{1}{32}$. In (E8) the step size $h = \frac{1}{32}$ was just used on the subinterval $\left[0, \frac{1}{4}\right]$.

Example C.3.1

Appendix D

NUMERICAL SOLUTION OF ODE'S

In Section 2.4 we solved a number of inital value problems of the form

$$y'(t) = f\big(t, y(t)\big)$$
$$y(t_0) = y_0$$

Here $f(t,y)$ is a given function, t_0 is a given initial time and y_0 is a given initial value for y. The unknown in the problem is the function $y(t)$. There are a number of other techniques for analytically solving some problems of this type. However it is often simply not possible to find an explicit solution. This appendix introduces some simple algorithms for generating approximate numerical solutions to such problems.

D.1▲ Simple ODE Solvers — Derivation

The first order of business is to derive three simple algorithms for generating approximate numerical solutions to the initial value problem

$$y'(t) = f\big(t, y(t)\big)$$
$$y(t_0) = y_0$$

The first is called Euler's method because it was developed by (surprise!) Euler[1].

D.1.1 ▸▸ Euler's Method

Our goal is to approximate (numerically) the unknown function

$$\begin{aligned} y(t) &= y(t_0) + \int_{t_0}^{t} y'(\tau)\,\mathrm{d}\tau \\ &= y(t_0) + \int_{t_0}^{t} f\big(\tau, y(\tau)\big)\,\mathrm{d}\tau \end{aligned}$$

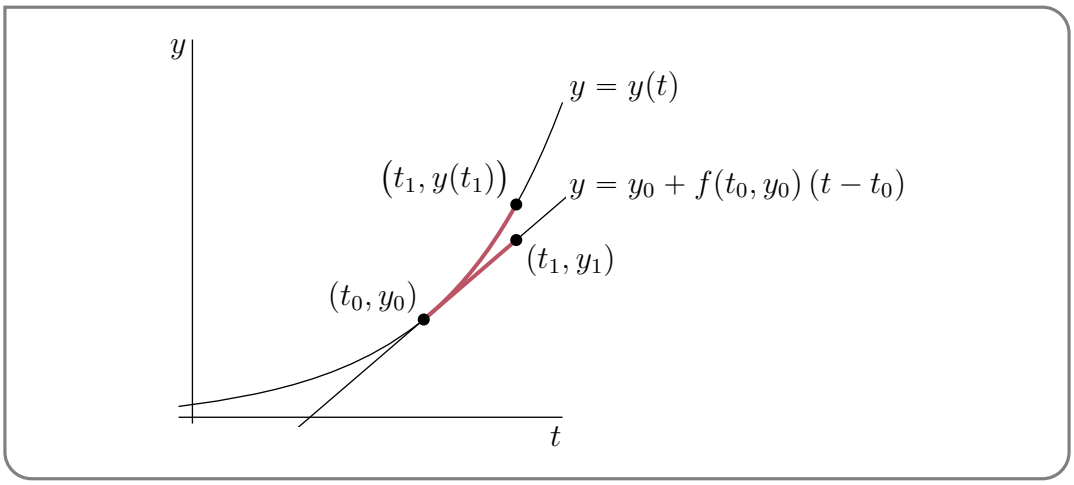
1 Leonhard Euler (1707–1783) was a Swiss mathematician and physicist who spent most of his adult life in Saint Petersberg and Berlin. He gave the name π to the ratio of a circle's circumference to its diameter. He also developed the constant e.

for $t \geqslant t_0$. We are told explicitly the value of $y(t_0)$, namely y_0. So we know $f\big(\tau, y(\tau)\big)\big|_{\tau=t_0} = f\big(t_0, y_0\big)$. But we do not know the integrand $f\big(\tau, y(\tau)\big)$ for $\tau > t_0$. On the other hand, if τ is close t_0, then $f\big(\tau, y(\tau)\big)$ will remain close[2] to $f\big(t_0, y_0\big)$. So pick a small number h and define

$$\begin{aligned} t_1 &= t_0 + h \\ y_1 &= y(t_0) + \int_{t_0}^{t_1} f(t_0, y_0)\,\mathrm{d}\tau = y_0 + f\big(t_0, y_0\big)(t_1 - t_0) \\ &= y_0 + f\big(t_0, y_0\big)h \end{aligned}$$

By the above argument

$$y(t_1) \approx y_1$$



Now we start over from the new point (t_1, y_1). We now know an approximate value for y at time t_1. If $y(t_1)$ were exactly y_1, then the instantaneous rate of change of y at time t_1, namely $y'(t_1) = f\big(t_1, y(t_1)\big)$, would be exactly $f(t_1, y_1)$ and $f\big(t, y(t)\big)$ would remain close to $f(t_1, y_1)$ for t close to t_1. Defining

$$\begin{aligned} t_2 &= t_1 + h = t_0 + 2h \\ y_2 &= y_1 + \int_{t_1}^{t_2} f(t_1, y_1)\,\mathrm{d}t = y_1 + f\big(t_1, y_1\big)(t_2 - t_1) \\ &= y_1 + f\big(t_1, y_1\big)h \end{aligned}$$

we have

$$y(t_2) \approx y_2$$

We just repeat this argument ad infinitum. Define, for $n = 0, 1, 2, 3, \cdots$

$$t_n = t_0 + nh$$

Suppose that, for some value of n, we have already computed an approximate value y_n for $y(t_n)$. Then the rate of change of $y(t)$ for t close to t_n is $f\big(t, y(t)\big) \approx f\big(t_n, y(t_n)\big) \approx f\big(t_n, y_n\big)$ and

2 This will be the case as long as $f(t, y)$ is continuous.

Equation D.1.1 (Euler's Method).

$$y(t_{n+1}) \approx y_{n+1} = y_n + f(t_n, y_n)h$$

This algorithm is called *Euler's Method*. The parameter h is called the *step size*.

Here is a table applying a few steps of Euler's method to the initial value problem

$$y' = -2t + y$$
$$y(0) = 3$$

with step size $h = 0.1$. For this initial value problem

$$f(t, y) = -2t + y$$
$$t_0 = 0$$
$$y_0 = 3$$

Of course this initial value problem has been chosen for illustrative purposes only. The exact solution is[3] $y(t) = 2 + 2t + e^t$.

n	t_n	y_n	$f(t_n, y_n) = -2t_n + y_n$	$y_{n+1} = y_n + f(t_n, y_n) * h$
0	0.0	3.000	-2*0.0+3.000=3.000	3.000+3.000*0.1=3.300
1	0.1	3.300	-2*0.1+3.300=3.100	3.300+3.100*0.1=3.610
2	0.2	3.610	-2* 0.2+3.610=3.210	3.610+3.210*0.1=3.931
3	0.3	3.931	-2* 0.3+3.931=3.331	3.931+3.331*0.1=4.264
4	0.4	4.264	-2* 0.4+4.264=3.464	4.264+3.464*0.1=4.611
5	0.5	4.611		

The exact solution at $t = 0.5$ is 4.6487, to four decimal places. We expect that Euler's method will become more accurate as the step size becomes smaller. But, of course, the amount of effort goes up as well. If we recompute using $h = 0.01$, we get (after much more work) 4.6446.

D.1.2 ▸▸ The Improved Euler's Method

Euler's method is one algorithm which generates approximate solutions to the initial value problem

$$y'(t) = f(t, y(t))$$
$$y(t_0) = y_0$$

In applications, $f(t, y)$ is a given function and t_0 and y_0 are given numbers. The function $y(t)$ is unknown. Denote by $\varphi(t)$ the exact solution[4] for this initial value problem. In other

3 Even if you haven't learned how to solve initial value problems like this one, you can check that $y(t) = 2 + 2t + e^t$ obeys both $y'(t) = -2t + y(t)$ and $y(0) = 3$.

4 Under reasonable hypotheses on f, there is exactly one such solution. The interested reader should search engine their way to the Picard-Lindelöf theorem.

words $\varphi(t)$ is the function that obeys

$$\begin{aligned}\varphi'(t) &= f\big(t, \varphi(t)\big)\\ \varphi(t_0) &= y_0\end{aligned}$$

exactly.

Fix a step size h and define $t_n = t_0 + nh$. By turning the problem into one of approximating integrals, we now derive another algorithm that generates approximate values for φ at the sequence of equally spaced time values $t_0,\ t_1,\ t_2,\ \cdots$. We shall denote the approximate values y_n with

$$y_n \approx \varphi(t_n)$$

By the fundamental theorem of calculus and the differential equation, the exact solution obeys

$$\begin{aligned}\varphi(t_{n+1}) &= \varphi(t_n) + \int_{t_n}^{t_{n+1}} \varphi'(t)\ \mathrm{d}t\\ &= \varphi(t_n) + \int_{t_n}^{t_{n+1}} f\big(t, \varphi(t)\big)\ \mathrm{d}t\end{aligned}$$

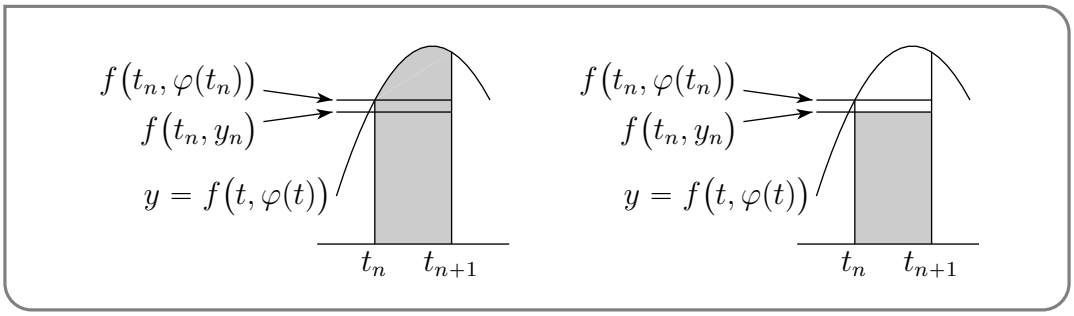
Fix any n and suppose that we have already found $y_0,\ y_1,\ \cdots,\ y_n$. Our algorithm for computing y_{n+1} will be of the form

$$y_{n+1} = y_n + \text{ approximate value of } \int_{t_n}^{t_{n+1}} f\big(t, \varphi(t)\big)\ \mathrm{d}t$$

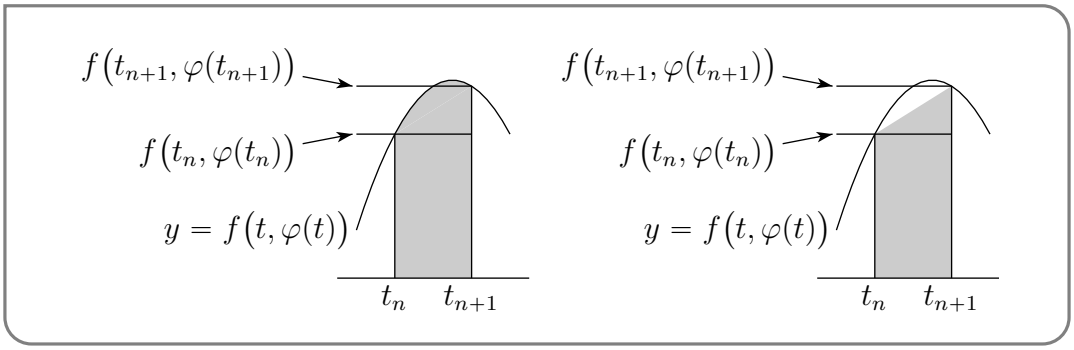
In Euler's method, we approximated $f\big(t, \varphi(t)\big)$ for $t_n \leqslant t \leqslant t_{n+1}$ by the constant $f\big(t_n, y_n\big)$. Thus

$$\text{Euler's approximate value for } \int_{t_n}^{t_{n+1}} f\big(t, \varphi(t)\big)\ \mathrm{d}t \text{ is } \int_{t_n}^{t_{n+1}} f\big(t_n, y_n\big)\ \mathrm{d}t = f\big(t_n, y_n\big)h$$

So Euler's method approximates the area of the complicated region $0 \leqslant y \leqslant f\big(t, \varphi(t)\big)$, $t_n \leqslant t \leqslant t_{n+1}$ (represented by the shaded region under the parabola in the left half of the figure below) by the area of the rectangle $0 \leqslant y \leqslant f\big(t_n, y_n\big)$, $t_n \leqslant t \leqslant t_{n+1}$ (the shaded rectangle in the right half of the figure below).



Our second algorithm, the improved Euler's method, gets a better approximation by using the trapezoidal rule. That is, we approximate the integral by the area of the trapezoid on the right below, rather than the rectangle on the right above. The exact area of



this trapezoid is the length h of the base multiplied by the average of the heights of the two sides, which is $\frac{1}{2}\big[f\big(t_n,\varphi(t_n)\big)+f\big(t_{n+1},\varphi(t_{n+1})\big)\big]$. Of course we do not know $\varphi(t_n)$ or $\varphi(t_{n+1})$ exactly.

Recall that we have already found $y_0,\cdots,y_n$ and are in the process of finding y_{n+1}. So we already have an approximation for $\varphi(t_n)$, namely y_n. But we still need to approximate $\varphi(t_{n+1})$. We can do so by using one step of the original Euler method! That is

$$\varphi(t_{n+1})\approx\varphi(t_n)+\varphi'(t_n)h\approx y_n+f(t_n,y_n)h$$

So our approximation of $\frac{1}{2}\big[f\big(t_n,\varphi(t_n)\big)+f\big(t_{n+1},\varphi(t_{n+1})\big)\big]$ is

$$\frac{1}{2}\Big[f\big(t_n,y_n\big)+f\Big(t_{n+1},y_n+f(t_n,y_n)h\Big)\Big]$$

and

$$\text{Improved Euler's approximate value for } \int_{t_n}^{t_{n+1}} f\big(t,\varphi(t)\big)\,\mathrm{d}t \text{ is}$$
$$\frac{1}{2}\Big[f\big(t_n,y_n\big)+f\Big(t_{n+1},y_n+f(t_n,y_n)h\Big)\Big]h$$

Putting everything together[5], the improved Euler's method algorithm is

Equation D.1.2 (Improved Euler).

$$y(t_{n+1})\approx y_{n+1}=y_n+\frac{1}{2}\Big[f\big(t_n,y_n\big)+f\Big(t_{n+1},y_n+f(t_n,y_n)h\Big)\Big]h$$

Here are the first two steps of the improved Euler's method applied to

$$y'=-2t+y\qquad y(0)=3$$

5 Notice that we have made a first approximation for $\varphi(t_{n+1})$ by using Euler's method. Then improved Euler uses the first approximation to build a better approximation for $\varphi(t_{n+1})$. Building an approximation on top of another approximation does not always work, but it works very well here.

with $h = 0.1$. In each step we compute $f(t_n, y_n)$, followed by $y_n + f(t_n, y_n)h$, which we denote $\tilde{y}_{n+1}$, followed by $f(t_{n+1}, \tilde{y}_{n+1})$, followed by $y_{n+1} = y_n + \frac{1}{2}\big[f\big(t_n, y_n\big) + f\big(t_{n+1}, \tilde{y}_{n+1}\big)\big]h$.

$$
\begin{aligned}
t_0 = 0 \quad y_0 = 3 \quad &\Longrightarrow f(t_0, y_0) = -2 * 0 + 3 = 3 \\
&\Longrightarrow \quad \tilde{y}_1 = 3 + 3 * 0.1 = 3.3 \\
&\Longrightarrow f(t_1, \tilde{y}_1) = -2 * 0.1 + 3.3 = 3.1 \\
&\Longrightarrow \quad y_1 = 3 + \frac{1}{2}[3 + 3.1] * 0.1 = 3.305 \\
t_1 = 0.1 \quad y_1 = 3.305 \ &\Longrightarrow f(t_1, y_1) = -2 * 0.1 + 3.305 = 3.105 \\
&\Longrightarrow \quad \tilde{y}_2 = 3.305 + 3.105 * 0.1 = 3.6155 \\
&\Longrightarrow f(t_2, \tilde{y}_2) = -2 * 0.2 + 3.6155 = 3.2155 \\
&\Longrightarrow \quad y_2 = 3.305 + \frac{1}{2}[3.105 + 3.2155] * 0.1 = 3.621025
\end{aligned}
$$

Here is a table which gives the first five steps.

n	t_n	y_n	$f(t_n, y_n)$	$\tilde{y}_{n+1}$	$f(t_{n+1}, \tilde{y}_{n+1})$	y_{n+1}
0	0.0	3.000	3.000	3.300	3.100	3.305
1	0.1	3.305	3.105	3.616	3.216	3.621
2	0.2	3.621	3.221	3.943	3.343	3.949
3	0.3	3.949	3.349	4.284	3.484	4.291
4	0.4	4.291	3.491	4.640	3.640	4.647
5	0.5	4.647				

As we saw at the end of Section D.1.1, the exact $y(0.5)$ is 4.6487, to four decimal places, and Euler's method gave 4.611.

D.1.3 ▸▸ The Runge-Kutta Method

The Runge-Kutta[6] algorithm is similar to the Euler and improved Euler methods in that it also uses, in the notation of the last subsection,

$$y_{n+1} = y_n + \text{ approximate value for } \int_{t_n}^{t_{n+1}} f\big(t, \varphi(t)\big)\, \mathrm{d}t$$

But rather than approximating $\int_{t_n}^{t_{n+1}} f\big(t, \varphi(t)\big)\, \mathrm{d}t$ by the area of a rectangle, as does Euler, or by the area of a trapezoid, as does improved Euler, it approximates by the area under a parabola. That is, it uses Simpson's rule. According to Simpson's rule (which is derived in §1.11.3)

$$\int_{t_n}^{t_n+h} f\big(t, \varphi(t)\big)\, \mathrm{d}t \approx \tfrac{h}{6}\Big[f\big(t_n, \varphi(t_n)\big) + 4f\big(t_n + \tfrac{h}{2}, \varphi(t_n + \tfrac{h}{2})\big) + f\big(t_n + h, \varphi(t_n + h)\big)\Big]$$

6 Carl David Tolmé Runge (1856–1927) and Martin Wilhelm Kutta (1867–1944) were German mathematicians.

Analogously to what happened in our development of the improved Euler method, we don't know $\varphi(t_n)$, $\varphi(t_n + \frac{h}{2})$ or $\varphi(t_n + h)$. So we have to approximate them as well. The Runge-Kutta algorithm, incorporating all these approximations, is[7]

Equation D.1.3 (Runge-Kutta).

$$\begin{aligned} k_{1,n} &= f(t_n, y_n) \\ k_{2,n} &= f(t_n + \tfrac{1}{2}h, y_n + \tfrac{h}{2}k_{1,n}) \\ k_{3,n} &= f(t_n + \tfrac{1}{2}h, y_n + \tfrac{h}{2}k_{2,n}) \\ k_{4,n} &= f(t_n + h, y_n + hk_{3,n}) \\ y_{n+1} &= y_n + \tfrac{h}{6}\left[k_{1,n} + 2k_{2,n} + 2k_{3,n} + k_{4,n}\right] \end{aligned}$$

That is, Runge-Kutta uses

- $k_{1,n}$ to approximate $f\big(t_n, \varphi(t_n)\big) = \varphi'(t_n)$,
- both $k_{2,n}$ and $k_{3,n}$ to approximate $f\big(t_n + \frac{h}{2}, \varphi(t_n + \frac{h}{2})\big) = \varphi'(t_n + \frac{h}{2})$, and
- $k_{4,n}$ to approximate $f\big(t_n + h, \varphi(t_n + h)\big)$.

Here are the first two steps of the Runge-Kutta algorithm applied to

$$y' = -2t + y \qquad y(0) = 3$$

with $h = 0.1$.

$$\begin{aligned} t_0 = 0 \quad & y_0 = 3 \\ \implies & k_{1,0} = f(0,3) = -2 * 0 + 3 = 3 \\ \implies & \quad y_0 + \tfrac{h}{2}k_{1,0} = 3 + 0.05 * 3 = 3.15 \\ \implies & k_{2,0} = f(0.05, 3.15) = -2 * 0.05 + 3.15 = 3.05 \\ \implies & \quad y_0 + \tfrac{h}{2}k_{2,0} = 3 + 0.05 * 3.05 = 3.1525 \\ \implies & k_{3,0} = f(0.05, 3.1525) = -2 * 0.05 + 3.1525 = 3.0525 \\ \implies & \quad y_0 + hk_{3,0} = 3 + 0.1 * 3.0525 = 3.30525 \\ \implies & k_{4,0} = f(0.1, 3.30525) = -2 * 0.1 + 3.30525 = 3.10525 \\ \implies & \quad y_1 = 3 + \tfrac{0.1}{6}[3 + 2 * 3.05 + 2 * 3.0525 + 3.10525] = 3.3051708 \end{aligned}$$

7 It is well beyond our scope to derive this algorithm, though the derivation is similar in flavour to that of the improved Euler method. You can find more in, for example, Wikipedia.

$$
\begin{aligned}
t_1 = 0.1 \;\; y_1 &= 3.3051708 \\
\implies k_{1,1} &= f(0.1, 3.3051708) = -2 * 0.1 + 3.3051708 = 3.1051708 \\
\implies y_1 + \tfrac{h}{2} k_{1,1} &= 3.3051708 + 0.05 * 3.1051708 = 3.4604293 \\
\implies k_{2,1} &= f(0.15, 3.4604293) = -2 * 0.15 + 3.4604293 = 3.1604293 \\
\implies y_1 + \tfrac{h}{2} k_{2,1} &= 3.3051708 + 0.05 * 3.1604293 = 3.4631923 \\
\implies k_{3,1} &= f(0.15, 3.4631923) = -2 * 0.15 + 3.4631923 = 3.1631923 \\
\implies y_1 + h k_{3,1} &= 3.3051708 + 0.1 * 3.4631923 = 3.62149 \\
\implies k_{4,1} &= f(0.2, 3.62149) = -2 * 0.2 + 3.62149 = 3.22149 \\
\implies y_2 &= 3.3051708 + \tfrac{0.1}{6}[3.1051708 + 2 * 3.1604293 + \\
&\qquad + 2 * 3.1631923 + 3.22149] = 3.6214025 \\
t_2 = 0.2 \;\; y_2 &= 3.6214025
\end{aligned}
$$

While this might look intimidating written out in full like this, one should keep in mind that it is quite easy to write a program to do this. Here is a table giving the first five steps. The intermediate data is only given to three decimal places even though the computation has been done to many more.

n	t_n	y_n	$k_{1,n}$	$y_{n,1}$	$k_{2,n}$	$y_{n,2}$	$k_{3,n}$	$y_{n,3}$	$k_{4,n}$	y_{n+1}
0	0.0	3.000	3.000	3.150	3.050	3.153	3.053	3.305	3.105	3.305170833
1	0.1	3.305	3.105	3.460	3.160	3.463	3.163	3.621	3.221	3.621402571
2	0.2	3.621	3.221	3.782	3.282	3.786	3.286	3.950	3.350	3.949858497
3	0.3	3.950	3.350	4.117	3.417	4.121	3.421	4.292	3.492	4.291824240
4	0.4	4.292	3.492	4.466	3.566	4.470	3.570	4.649	3.649	4.648720639
5	0.5	4.6487206								

As we saw at the end of Section D.1.2, the exact $y(0.5)$ is 4.6487213, to seven decimal places, Euler's method gave 4.611 and improved Euler gave 4.647.

So far we have, hopefully, motivated the Euler, improved Euler and Runge-Kutta algorithms. We have not attempted to see how efficient and how accurate the algorithms are. A first look at those questions is provided in the next section.

D.2▲ Simple ODE Solvers — Error Behaviour

We now provide an introduction to the error behaviour of Euler's Method, the improved Euler's method and the Runge-Kutta algorithm for generating approximate solutions to the initial value problem

$$
\begin{aligned}
y'(t) &= f\big(t, y(t)\big) \\
y(t_0) &= y_0
\end{aligned}
$$

Here $f(t, y)$ is a given known function, t_0 is a given initial time and y_0 is a given initial value for y. The unknown in the problem is the function $y(t)$.

Two obvious considerations in deciding whether or not a given algorithm is of any practical value are

(a) the amount of computational effort required to execute the algorithm and
(b) the accuracy that this computational effort yields.

For algorithms like our simple ODE solvers, the bulk of the computational effort usually goes into evaluating the function[8] $f(t, y)$. Euler's method uses one evaluation of $f(t, y)$ for each step; the improved Euler's method uses two evaluations of f per step; the Runge-Kutta algorithm uses four evaluations of f per step. So Runge-Kutta costs four times as much work per step as does Euler. But this fact is extremely deceptive because, as we shall see, you typically get the same accuracy with a few steps of Runge-Kutta as you do with hundreds of steps of Euler.

To get a first impression of the error behaviour of these methods, we apply them to a problem that we know the answer to. The solution to the first order constant coefficient linear initial value problem

$$\begin{aligned} y'(t) &= y - 2t \\ y(0) &= 3 \end{aligned}$$

is

$$y(t) = 2 + 2t + e^t$$

In particular, the exact value of $y(1)$, to ten decimal places, is $4 + e = 6.7182818285$. The following table lists the error in the approximate value for this number generated by our three methods applied with three different step sizes. It also lists the number of evaluations of f required to compute the approximation.

	Euler		**Improved Euler**		**Runge Kutta**	
steps	error	# evals	error	# evals	error	# evals
5	2.3×10^{-1}	5	1.6×10^{-2}	10	3.1×10^{-5}	20
50	2.7×10^{-2}	50	1.8×10^{-4}	100	3.6×10^{-9}	200
500	2.7×10^{-3}	500	1.8×10^{-6}	1000	3.6×10^{-13}	2000

Observe

- Using 20 evaluations of f worth of Runge-Kutta gives an error 90 times smaller than 500 evaluations of f worth of Euler.
- With Euler's method, decreasing the step size by a factor of ten appears to reduce the error by about a factor of ten.
- With improved Euler, decreasing the step size by a factor of ten appears to reduce the error by about a factor of one hundred.
- With Runge-Kutta, decreasing the step size by a factor of ten appears to reduce the error by about a factor of about 10^4.

8 Typically, evaluating a complicated function will take a great many arithmetic operations, while the actual ODE solver method (as per, for example, (D.1.3)) takes only an additional handful of operations. So the great bulk of computational time goes into evaulating f and we want to do it as few times as possible.

Use $A_E(h)$, $A_{IE}(h)$ and $A_{RK}(h)$ to denote the approximate value of $y(1)$ given by Euler, improved Euler and Runge-Kutta, respectively, with step size h. It looks like

Equation D.2.1.

$$A_E(h) \approx y(1) + K_E h$$
$$A_{IE}(h) \approx y(1) + K_{IE} h^2$$
$$A_{RK}(h) \approx y(1) + K_{RK} h^4$$

with some constants K_E, K_{IE} and K_{RK}.

To test these conjectures further, we apply our three methods with about ten different step sizes of the form $\frac{1}{n} = \frac{1}{2^m}$ with m integer. Below are three graphs, one for each method. Each contains a plot of $Y = \log_2 e_n$, the (base 2) logarithm of the error for step size $\frac{1}{n}$, against the logarithm (of base 2) of n. The logarithm of base 2 is used because $\log_2 n = \log_2 2^m = m$ — nice and simple.

Here is why it is a good reason to plot $Y = \log_2 e_n$ against $x = \log_2 n$. If, for some algorithm, there are (unknown) constants K and k such that

$$\text{approx value of } y(1) \text{ with step size } h = y(1) + Kh^k$$

then the error with step size $\frac{1}{n}$ is $e_n = K\frac{1}{n^k}$ and obeys

$$\log_2 e_n = \log_2 K - k \log_2 n \tag{E1}$$

The graph of $Y = \log_2 e_n$ against $x = \log_2 n$ is the straight line $Y = -kx + \log_2 K$ of slope $-k$ and y intercept $\log_2 K$.

Remark D.2.2. This procedure can still be used even if we do not know the exact value of $y(1)$. Suppose, more generally, that we have some algorithm that generates approximate values for some (unknown) exact value $\mathcal{A}$. Call A_h the approximate value with step size h. Suppose that

$$A_h = \mathcal{A} + Kh^k$$

with K and k constant (but also unknown). Then plotting

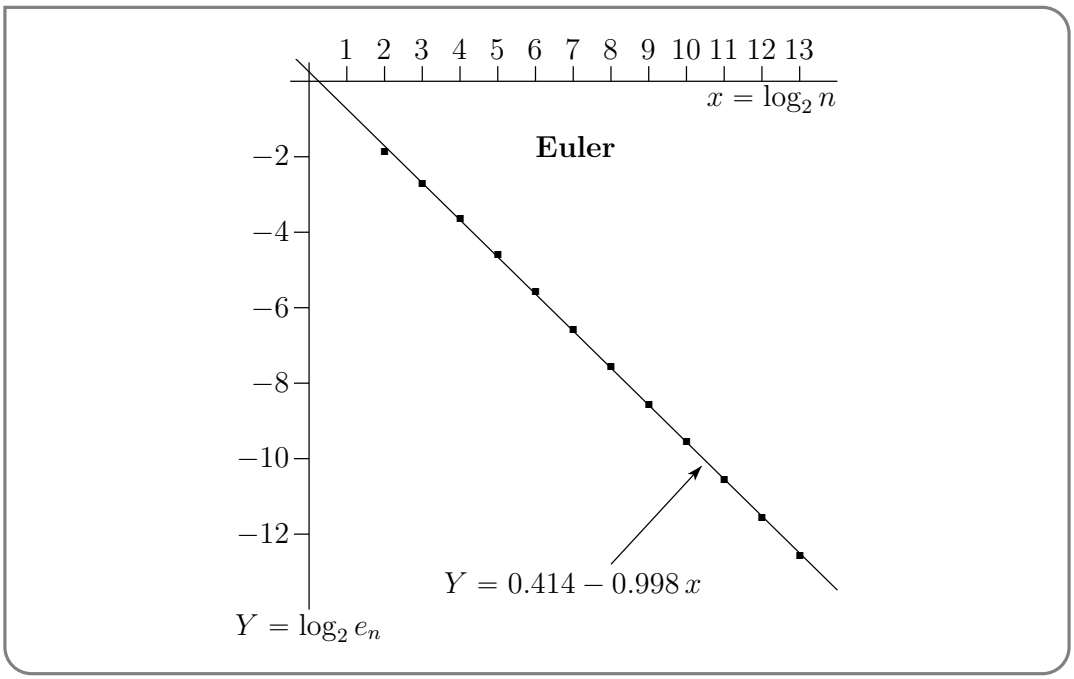
$$y = \log(A_h - A_{h/2}) = \log\left(Kh^k - K\left(\tfrac{h}{2}\right)^k\right) = \log\left(K - \tfrac{K}{2^k}\right) + k \log h$$

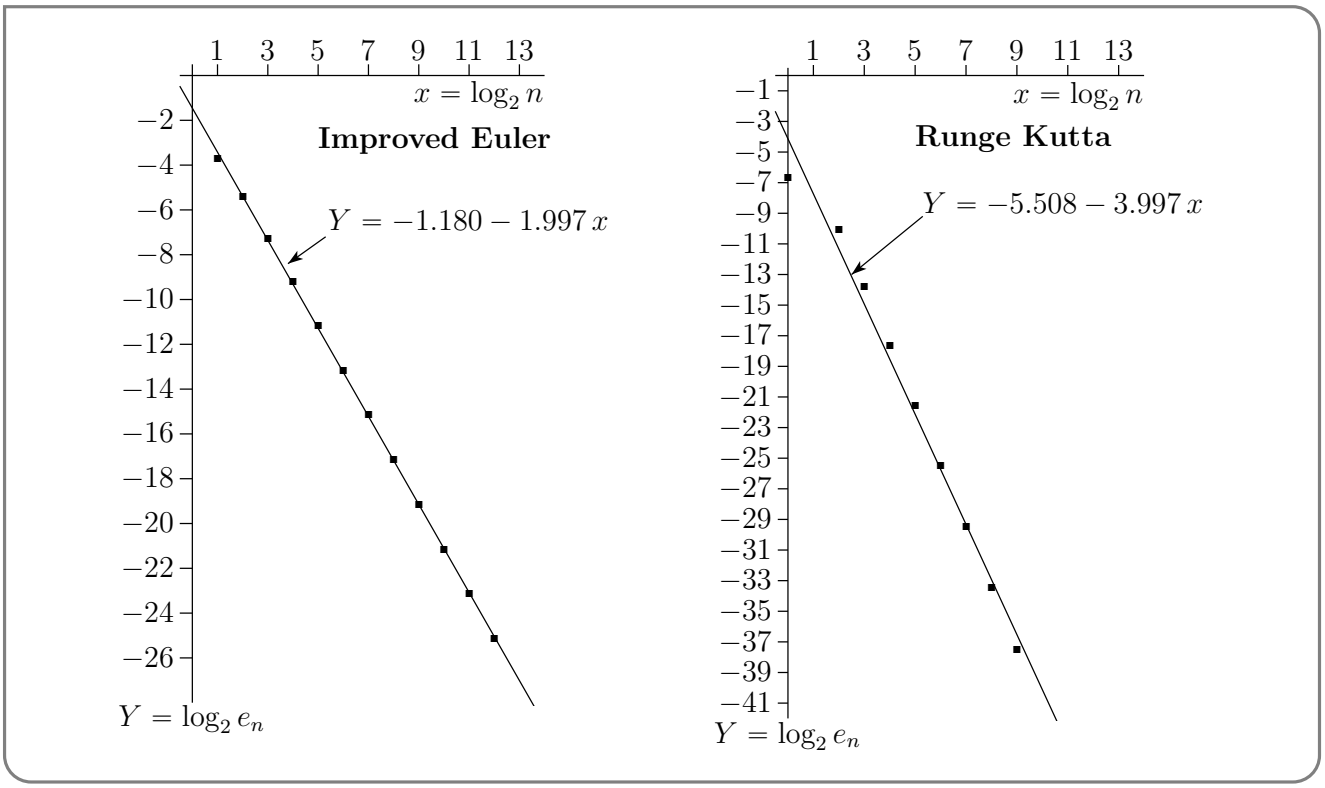
against $x = \log h$ gives the straight line $y = mx + b$ with slope $m = k$ and y intercept $b = \log\left(K - \frac{K}{2^k}\right)$. So we can

- read off k from the slope of the line and then
- compute $K = e^b \left(1 - \frac{1}{2^k}\right)^{-1}$ from the y intercept b and then
- compute[9] $A = A_h - Kh^k$.

9 This is the type of strategy used by the Richardson extrapolation of Section C.1.

Here are the three graphs — one each for the Euler method, the improved Euler method and the Runge-Kutta method. Each graph contains about a dozen data points, $(x, Y) = (\log_2 n, \log_2 e_n)$.

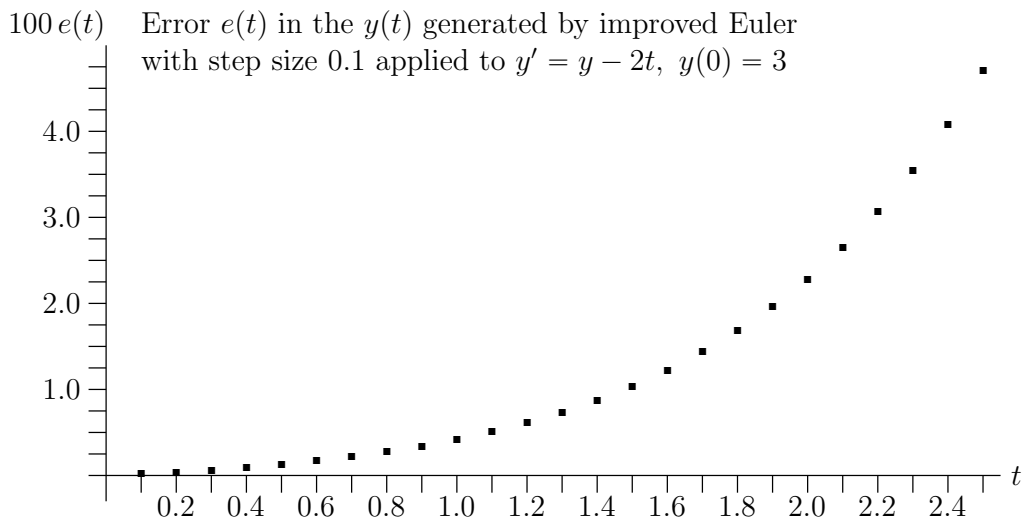




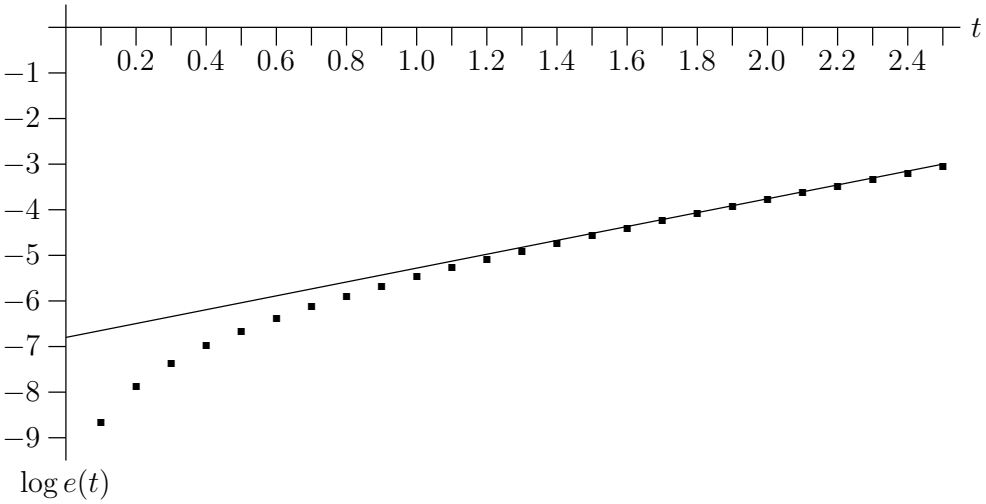
Each graph also contains a straight line, chosen by linear regression, to best fit the data. The method of linear regression for finding the straight line which best fits a given set of data points is covered in Example 2.9.11 of the CLP-3 text. The three straight lines

have slopes -0.998 for Euler, -1.997 for improved Euler and -3.997 for Runge Kutta. Reviewing (E1), it sure looks like $k = 1$ for Euler, $k = 2$ for improved Euler and $k = 4$ for Runge-Kutta (at least if k is integer).

So far we have only looked at the error in the approximate value of $y(t_f)$ as a function of the step size h with t_f held fixed. The graph below illustrates how the error behaves



as a function of t, with h held fixed. That is, we hold the step size fixed and look at the error as a function of the distance, t, from the initial point. From the graph, it appears that the error grows exponentially with t. But it is not so easy to visually distinguish exponential curves from other upward curving curves. On the other hand, it is pretty easy to visually distinguish straight lines from other curves, and taking a logarithm converts the exponential curve $y = e^{kx}$ into the straight line $Y = \log y = k\,x$. Here is a graph of the logarithm, $\log e(t)$, of the error at time t, $e(t)$, against t. We have added a straight line as an aide to your eye.



It looks like the log of the error grows very quickly initially, but then settles into a straight line. Hence it really does look like, at least in this example, except at the very beginning, the error $e(t)$ grows exponentially with t.

The above numerical experiments have given a little intuition about the error behaviour of the Euler, improved Euler and Runge-Kutta methods. It's time to try and understand what is going on more rigorously.

D.2.1 ▸▸ Local Truncation Error for Euler's Method

We now try to develop some understanding as to why we got the above experimental results. We start with the error generated by a single step of Euler's method.

Definition D.2.3 (Local truncation error).

The (signed) error generated by a single step of Euler's method, under the assumptions that we start the step with the exact solution and that there is no roundoff error, is called the *local truncation error* for Euler's method. That is, if $\phi(t)$ obeys $\phi'(t) = f\big(t, \phi(t)\big)$ and $\phi(t_n) = y_n$, and if $y_{n+1} = y_n + hf(t_n, y_n)$, then the local truncation error for Euler's method is

$$\phi\big(t_{n+1}\big) - y_{n+1}$$

That is, it is difference between the exact value, $\phi\big(t_{n+1}\big)$, and the approximate value generated by a single Euler method step, y_{n+1}, ignoring any numerical issues caused by storing numbers in a computer.

Denote by $\phi(t)$ the exact solution to the initial value problem

$$y'(t) = f(t, y) \qquad y(t_n) = y_n$$

That is, $\phi(t)$ obeys

$$\phi'(t) = f\big(t, \phi(t)\big) \qquad \phi(t_n) = y_n$$

for all t. Now execute one more step of Euler's method with step size h:

$$y_{n+1} = y_n + hf\big(t_n, y_n\big)$$

Because we are assuming that $y_n = \phi(t_n)$

$$y_{n+1} = \phi(t_n) + hf\big(t_n, \phi(t_n)\big)$$

Because $\phi(t)$ is the exact solution, $\phi'(t_n) = f\big(t_n, \phi(t_n)\big) = f(t_n, y_n)$ and

$$y_{n+1} = \phi(t_n) + h\phi'(t_n)$$

The local truncation error in y_{n+1} is, by definition, $\phi(t_{n+1}) - y_{n+1}$.

Taylor expanding (see (3.4.10) in the CLP-1 text) $\phi(t_{n+1}) = \phi(t_n + h)$ about t_n

$$\phi(t_n + h) = \phi(t_n) + \phi'(t_n)h + \tfrac{1}{2}\phi''(t_n)h^2 + \tfrac{1}{3!}\phi'''(t_n)h^3 + \cdots$$

so that

$$\begin{aligned}\phi(t_{n+1}) - y_{n+1} &= \left[\phi(t_n) + \phi'(t_n)h + \tfrac{1}{2}\phi''(t_n)h^2 + \tfrac{1}{3!}\phi'''(t_n)h^3 + \cdots\right] - \left[\phi(t_n) + h\phi'(t_n)\right] \\ &= \tfrac{1}{2}\phi''(t_n)h^2 + \tfrac{1}{3!}\phi'''(t_n)h^3 + \cdots \qquad \text{(E2)}\end{aligned}$$

Notice that the constant and h^1 terms have cancelled out. So the first term that appears is proportional to h^2. Since h is typically a very small number, the h^3, h^4, $\cdots$ terms will usually be much smaller than the h^2 term.

We conclude that the local truncation error for Euler's method is h^2 times some unknown constant (we usually don't know the value of $\frac{1}{2}\phi''(t_n)$ because we don't usually know the solution $\phi(t)$ of the differential equation) plus smaller terms that are proportional to h^r with $r \geqslant 3$. This conclusion is typically written

Equation D.2.4.

$$\text{Local truncation error for Euler's method } = Kh^2 + O(h^3)$$

The symbol $O(h^3)$ is used to designate any function that, for small h, is bounded by a constant times h^3. So, if h is very small, $O(h^3)$ will be a lot smaller than h^2.

To get from an initial time $t = t_0$ to a final time $t = t_f$ using steps of size h requires $(t_f - t_0)/h$ steps. If each step were to introduce an error[10] $Kh^2 + O(h^3)$, then the final error in the approximate value of $y(t_f)$ would be

$$\frac{t_f - t_0}{h}\left[Kh^2 + O(h^3)\right] = K(t_f - t_0)\,h + O(h^2)$$

This very rough estimate is consistent with the experimental data for the dependence of error on step size with t_f held fixed, shown on the first graph after Remark D.2.2. But it is not consistent with the experimental time dependence data above, which shows the error growing exponentially, rather than linearly, in $t_f - t_0$.

We can get some rough understanding of this exponential growth as follows. The general solution to $y' = y - 2t$ is $y(t) = 2 + 2t + c_0 e^t$. The arbitrary constant, c_0, is to be determined by initial conditions. When $y(0) = 3$, $c_0 = 1$. At the end of step 1, we have computed an approximation y_1 to $y(h)$. This y_1 is not exactly $y(h) = 2 + 2h + e^h$. Instead, it is a number that differs from $2 + 2h + e^h$ by $O(h^2)$. We choose to write the number $y_1 = 2 + 2h + e^h + O(h^2)$ as $2 + 2h + (1 + \epsilon)e^h$ with $\epsilon = e^{-h}O(h^2)$ of order of magnitude h^2. That is, we choose to write

$$y_1 = 2 + 2t + c_0 e^t\Big|_{t=h} \qquad \text{with } c_0 = 1 + \epsilon$$

If we were to make no further errors we would end up with the solution to

$$y' = y - 2t, \qquad y(h) = 2 + 2h + (1 + \epsilon)e^h$$

10 For simplicity, we are assuming that K takes the same value in every step. If, instead, there is a different K in each of the $n = (t_f - t_0)/h$ steps, the final error would be $K_1h^2 + K_2h^2 + \cdots + K_nh^2 + nO(h^3) = \bar{K}nh^2 + nO(h^3) = \bar{K}(t_f - t_0)\,h + O(h^2)$, where $\bar{K}$ is the average of $K_1, K_2, \cdots, K_n$.

which is[11]

$$\begin{aligned} y(t) &= 2 + 2t + (1+\epsilon)e^t = 2 + 2t + e^t + \epsilon e^t \\ &= \phi(t) + \epsilon e^t \end{aligned}$$

So, once as error has been introduced, the natural time evolution of the solutions to this differential equation cause the error to grow exponentially. Other differential equations with other time evolution characteristics will exhibit different t_f dependence of errors[12]. In the next section, we show that, for many differential equations, errors grow at worst exponentially with t_f.

D.2.2 ▸▸ Global Truncation Error for Euler's Method

Suppose once again that we are applying Euler's method with step size h to the initial value problem

$$\begin{aligned} y'(t) &= f(t,y) \\ y(0) &= y_0 \end{aligned}$$

Denote by $\phi(t)$ the exact solution to the initial value problem and by y_n the approximation to $\phi(t_n)$, $t_n = t_0 + nh$, given by n steps of Euler's method (applied without roundoff error).

Definition D.2.5 (Global truncation error).

The (signed) error in y_n is $\phi(t_n) - y_n$ and is called the *global truncation error* at time t_n.

The word "truncation" is supposed to signify that this error is due solely to Euler's method and does not include any effects of roundoff error that might be introduced by our not writing down an infinite number of decimal digits for each number that we compute along the way. We now derive a bound on the global truncation error.

Define

$$\varepsilon_n = \phi(t_n) - y_n$$

The first half of the derivation is to find a bound on ε_{n+1} in terms of ε_n.

$$\begin{aligned} \varepsilon_{n+1} &= \phi(t_{n+1}) - y_{n+1} \\ &= \phi(t_{n+1}) - y_n - hf(t_n, y_n) \\ &= [\phi(t_n) - y_n] + h[f(t_n, \phi(t_n)) - f(t_n, y_n)] + [\phi(t_{n+1}) - \phi(t_n) - hf(t_n, \phi(t_n))] \end{aligned} \tag{E3}$$

where we have massaged the expression into three manageable pieces.

- The first $[\cdots]$ is exactly ε_n.

11 Note that this $y(t)$ obeys both the differential equation $y' = y - 2t$ and the initial condition $y(h) = 2 + 2h + (1+\epsilon)e^h$.

12 For example, if the solution is polynomial, then we might expect (by a similar argument) that the error also grows polynomially in t_f

- The third $[\cdots]$ is exactly the local truncation error. Assuming that $|\phi''(t)| \leqslant A$ for all t of interest[13], we can bound the third $[\cdots]$ by

$$\left|\phi(t_{n+1}) - \phi(t_n) - hf\big(t_n, \phi(t_n)\big)\right| \leqslant \tfrac{1}{2}Ah^2$$

  This bound follows quickly from the Taylor expansion with remainder ((3.4.32) in the CLP-1 text),

$$\begin{aligned}\phi(t_{n+1}) &= \phi(t_n) + \phi'(t_n)h + \tfrac{1}{2}\phi''(\tilde{t})h^2\\ &= \phi(t_n) + h\,f\big(t_n, \phi(t_n)\big) + \tfrac{1}{2}\phi''(\tilde{t})h^2\end{aligned}$$

  for some $t_n < \tilde{t} < t_{n+1}$.

- Finally, by the mean value theorem, the magnitude of the second $[\cdots]$ is h times

$$\begin{aligned}|f\big(t_n, \phi(t_n)\big) - f(t_n, y_n)| &= F_{t_n}\big(\phi(t_n)\big) - F_{t_n}(y_n) \quad \text{where} \quad F_{t_n}(y) = f\big(t_n, y\big)\\ &= \left|F'_{t_n}(\tilde{y})\right| |\phi(t_n) - y_n| \quad \text{for some } \tilde{y} \text{ between } y_n \text{ and } \phi(t_n)\\ &= \left|F'_{t_n}(\tilde{y})\right| |\varepsilon_n|\\ &\leqslant B|\varepsilon_n|\end{aligned}$$

  assuming that $\left|F'_t(y)\right| \leqslant B$ for all t and y of interest[14].

Substituting into (E3) gives

$$|\varepsilon_{n+1}| \leqslant |\varepsilon_n| + Bh|\varepsilon_n| + \tfrac{1}{2}Ah^2 = (1 + Bh)|\varepsilon_n| + \tfrac{1}{2}Ah^2 \qquad (\text{E4}_n)$$

Hence the (bound on the) magnitude of the total error, $|\varepsilon_{n+1}|$, consists of two parts. One part is the magnitude of the local truncation error, which is no more than $\frac{1}{2}Ah^2$ and which is present even if we start the step with no error at all, i.e. with $\varepsilon_n = 0$. The other part is due to the combined error from all previous steps. This is the ε_n term. At the beginning of step number $n + 1$, the combined error has magnitude $|\varepsilon_n|$. During the step, this error gets magnified by no more than a factor of $1 + Bh$.

The second half of the derivation is to repeatedly apply (E4_n) with $n = 0, 1, 2, \cdots$. By definition $\phi(t_0) = y_0$ so that $\varepsilon_0 = 0$, so

$$\begin{aligned}(\text{E4}_0) &\implies |\varepsilon_1| \leqslant (1 + Bh)|\varepsilon_0| + \tfrac{A}{2}h^2 = \tfrac{A}{2}h^2\\ (\text{E4}_1) &\implies |\varepsilon_2| \leqslant (1 + Bh)|\varepsilon_1| + \tfrac{A}{2}h^2 = (1 + Bh)\tfrac{A}{2}h^2 + \tfrac{A}{2}h^2\\ (\text{E4}_2) &\implies |\varepsilon_3| \leqslant (1 + Bh)|\varepsilon_2| + \tfrac{A}{2}h^2 = (1 + Bh)^2\tfrac{A}{2}h^2 + (1 + Bh)\tfrac{A}{2}h^2 + \tfrac{A}{2}h^2\end{aligned}$$

Continuing in this way

$$|\varepsilon_n| \leqslant (1 + Bh)^{n-1}\tfrac{A}{2}h^2 + \cdots + (1 + Bh)\tfrac{A}{2}h^2 + \tfrac{A}{2}h^2 = \sum_{m=0}^{n-1}(1 + Bh)^m\tfrac{A}{2}h^2$$

13 We are assuming that the derivative $\phi'(t)$ doesn't change too rapidly. This will be the case if $f(t, y)$ is a reasonably smooth function.

14 Again, this will be the case if $f(t, y)$ is a reasonably smooth function.

This is the beginning of a geometric series, and we can sum it up by using $\sum_{m=0}^{n-1} ar^m = \frac{r^n-1}{r-1}a$ (which is Theorem 1.1.6(a)) with $a = \frac{A}{2}h^2$ and $r = 1+Bh$ gives

$$|\varepsilon_n| \leqslant \frac{(1+Bh)^n - 1}{(1+Bh)-1}\frac{A}{2}h^2 = \frac{A}{2B}\big[(1+Bh)^n - 1\big]h$$

We are interested in how this behaves as $t_n - t_0$ increases and/or h decreases. Now $n = \frac{t_n - t_0}{h}$ so that $(1+Bh)^n = (1+Bh)^{(t_n-t_0)/h}$. When h is small, the behaviour of $(1+Bh)^{(t_n-t_0)/h}$ is not so obvious. So we'll use a little trickery to make it easier to understand. Setting $x = Bh$ in

$$x \geqslant 0 \implies 1 + x \leqslant 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots = e^x$$

(the exponential series $e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \cdots$ was derived in Example 3.6.3) gives[15] $1 + Bh \leqslant e^{Bh}$. Hence $(1+Bh)^n \leqslant e^{Bhn} = e^{B(t_n-t_0)}$, since $t_n = t_0 + nh$, and we arrive at the conclusion

Equation D.2.6.

$$|\varepsilon_n| \leqslant \frac{A}{2B}\left[e^{B(t_n-t_0)} - 1\right]h$$

which is of the form $K(t_f)h^k$ with $k = 1$ and the coefficient $K(t_f)$ growing exponentially with $t_f - t_0$. If we keep h fixed and increase t_n we see exponential growth, but if we fix t_n and decrease h we see the error decrease linearly. This is just what our experimental data suggested.

D.3▲ Variable Step Size Methods

We now introduce a family of procedures that decide by themselves what step size to use. In all of these procedures the user specifies an acceptable error rate and the procedure attempts to adjust the step size so that each step introduces error at no more than that rate. That way the procedure uses a small step size when it is hard to get an accurate approximation, and a large step size when it is easy to get a good approximation.

Suppose that we wish to generate an approximation to the initial value problem

$$y' = f(t,y), \qquad y(t_0) = y_0$$

for some range of t's and we want the error introduced per unit increase[16] of t to be no more than about some small fixed number ε. This means that if $y_n \approx y(t_0 + nh)$ and $y_{n+1} \approx y(t + (n+1)h)$, then we want the local truncation error in the step from y_n to y_{n+1} to be no more than about εh. Suppose further that we have already produced the

15 When $x = Bh$ is large, it is not wise to bound the linear $1 + x$ by the much larger exponential e^x. However when x is small, $1 + x$ and e^x are almost the same.

16 We know that the error will get larger the further we go in t. So it makes sense to try to limit the error per unit increase in t.

approximate solution as far as t_n. The rough strategy is as follows. We do the step from t_n to $t_n + h$ twice using two different algorithms, giving two different approximations to $y(t_{n+1})$, that we call $A_{1,n+1}$ and $A_{2,n+1}$. The two algorithms are chosen so that

(1) we can use $A_{1,n+1} - A_{2,n+1}$ to compute an approximate local truncation error and
(2) for efficiency, the two algorithms use almost the same evaluations of f. Remember that evaluating the function f is typically the most time-consuming part of our computation.

In the event that the local truncation error, divided by h, (i.e. the error per unit increase of t) is smaller than ε, we set $t_{n+1} = t_n + h$, accept $A_{2,n+1}$ as the approximate value[17] for $y(t_{n+1})$, and move on to the next step. Otherwise we pick, using what we have learned from $A_{1,n+1} - A_{2,n+1}$, a new trial step size h and start over again at t_n.

Now for the details. We start with a very simple minded procedure.

D.3.1 ▸▸ Euler and Euler-2step (preliminary version)

Denote by $\phi(t)$ the exact solution to $y' = f(t, y)$ that satisfies the initial condition $\phi(t_n) = y_n$. If we apply one step of Euler with step size h, giving

$$A_{1,n+1} = y_n + hf(t_n, y_n)$$

we know, from (D.2.4), that

$$A_{1,n+1} = \phi(t_n + h) + Kh^2 + O(h^3)$$

The problem, of course, is that we don't know what the error is, even approximately, because we don't know what the constant K is. But we can estimate K simply by redoing the step from t_n to $t_n + h$ using a judiciously chosen second algorithm. There are a number of different second algorithms that will work. We call the simple algorithm that we use in this subsection Euler-2step[18]. One step of Euler-2step with step size h just consists of doing two steps of Euler with step size $h/2$:

$$A_{2,n+1} = y_n + \tfrac{h}{2} f(t_n, y_n) + \tfrac{h}{2} f\big(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\big)$$

Here, the first half-step took us from y_n to $y_{\text{mid}} = y_n + \frac{h}{2} f(t_n, y_n)$ and the second half-step took us from y_{mid} to $y_{\text{mid}} + \frac{h}{2} f\big(t_n + \frac{h}{2}, y_{\text{mid}}\big)$. The local truncation error introduced in the first half-step is $K(h/2)^2 + O(h^3)$. That for the second half-step is $K(h/2)^2 + O(h^3)$ with the same[19] K, though with a different $O(h^3)$. All together

$$\begin{aligned} A_{2,n+1} &= \phi(t_n + h) + \big[K\big(\tfrac{h}{2}\big)^2 + O(h^3)\big] + \big[K\big(\tfrac{h}{2}\big)^2 + O(h^3)\big] \\ &= \phi(t_n + h) + \tfrac{1}{2} Kh^2 + O(h^3) \end{aligned}$$

17 Better still, accept $A_{2,n+1}$ minus the computed approximate error in $A_{2,n+1}$ as the approximate value for $y(t_{n+1})$.

18 This name is begging for a dance related footnote and we invite the reader to supply their own.

19 Because the two half-steps start at values of t only $h/2$ apart, and we are thinking of h as being very small, it should not be surprising that we can use the same value of K in both. In case you don't believe us, we have included a derivation of the local truncation error for Euler-2step later in this appendix.

The difference is[20]

$$\begin{aligned} A_{1,n+1} - A_{2,n+1} &= \left[\phi(t_n + h) + Kh^2 + O(h^3)\right] - \left[\phi(t_n + h) - \tfrac{1}{2}Kh^2 - O(h^3)\right] \\ &= \tfrac{1}{2}Kh^2 + O(h^3) \end{aligned}$$

So if we do one step of both Euler and Euler-2step, we can estimate

$$\tfrac{1}{2}Kh^2 = A_{1,n+1} - A_{2,n+1} + O(h^3)$$

We now know that in the step just completed Euler-2step introduced an error of about $\frac{1}{2}Kh^2 \approx A_{1,n+1} - A_{2,n+1}$. That is, the current error rate is about $r = \frac{|A_{1,n+1} - A_{2,n+1}|}{h} \approx \frac{1}{2}|K|h$ per unit increase of t.

- If this $r = \frac{|A_{1,n+1} - A_{2,n+1}|}{h} > \varepsilon$, we reject[21] $A_{2,n+1}$ and repeat the current step with a new trial step size chosen so that $\frac{1}{2}|K|(\text{new } h) < \varepsilon$, i.e. $\frac{r}{h}(\text{new } h) < \varepsilon$. To give ourselves a small safety margin, we could use[22]

$$\text{new } h = 0.9\,\frac{\varepsilon}{r}\,h$$

- If $r = \frac{|A_{1,n+1} - A_{2,n+1}|}{h} < \varepsilon$ we can accept[23] $A_{2,n+1}$ as an approximate value for $y(t_{n+1})$, with $t_{n+1} = t_n + h$, and move on to the next step, starting with the new trial step size[24]

$$\text{new } h = 0.9\,\frac{\varepsilon}{r}\,h$$

That is our preliminary version of the Euler/Euler-2step variable step size method. We call it the preliminary version, because we will shortly tweak it to get a much more efficient procedure.

Example D.3.1

As a concrete example, suppose that our problem is

$$y(0) = e^{-2},\ y' = 8(1 - 2t)y,\ \varepsilon = 0.1$$

and that we have gotten as far as

$$t_n = 0.33,\ y_n = 0.75,\ \text{ trial } h = 0.094$$

20 Recall that every time the symbol $O(h^3)$ is used it can stand for a different function that is bounded by some constant times h^3 for small h. Thus $O(h^3) - O(h^3)$ need not be zero, but is $O(h^3)$. What is important here, is that if K is not zero and if h is very small, then $O(h^3)$ is much smaller than $\frac{1}{2}Kh^2$.

21 The measured error rate, r, is bigger than the desired error rate ε. That means that it is harder to get the accuracy we want than we thought. So we have to take smaller steps.

22 We don't want to make the new h too close to $\frac{\varepsilon}{r}h$ since we are only estimating things and we might end up with an error rate bigger that ε. On the other hand, we don't want to make the new h too small because that means too much work — so we choose it to be just a little smaller than $\frac{\varepsilon}{r}h$ ... say $0.9\frac{\varepsilon}{r}h$.

23 The measured error rate, r, is smaller than the desired error rate ε. That means that it is easier to get the accuracy we want than we thought. So we can make the next step larger.

24 Note that in this case $\frac{\varepsilon}{r} > 1$. So the new h can be bigger than the last h.

Then, using $E = |A_{1,n+1} - A_{2,n+1}|$ to denote the magnitude of the estimated local truncation error in $A_{2,n+1}$ and r the corresponding error rate

$$\begin{aligned} f(t_n, y_n) &= 8(1 - 2 \times 0.33)0.75 = 2.04 \\ A_{1,n+1} &= y_n + hf(t_n, y_n) = 0.75 + 0.094 \times 2.04 = 0.942 \\ y_{\text{mid}} &= y_n + \tfrac{h}{2} f(t_n, y_n) = 0.75 + \tfrac{0.094}{2} \times 2.04 = 0.846 \\ f\big(t_n + \tfrac{h}{2}, y_{\text{mid}}\big) &= 8\Big[1 - 2\big(0.33 + \tfrac{0.094}{2}\big)\Big]0.846 = 1.66 \\ A_{2,n+1} &= y_{\text{mid}} + \tfrac{h}{2} f(t_n + \tfrac{h}{2}, y_{\text{mid}}) = 0.846 + \tfrac{0.094}{2} 1.66 = 0.924 \\ E &= |A_{1,n+1} - A_{2,n+1}| = |0.942 - 0.924| = 0.018 \\ r &= \frac{|E|}{h} = \frac{0.018}{0.094} = 0.19 \end{aligned}$$

Since $r = 0.19 > \varepsilon = 0.1$, the current step size is unacceptable and we have to recompute with the new step size

$$\text{new } h = 0.9\frac{\varepsilon}{r}(\text{old } h) = 0.9\ \frac{0.1}{0.19}\ 0.094 = 0.045$$

to give

$$\begin{aligned} f(t_n, y_n) &= 8(1 - 2 \times 0.33)0.75 = 2.04 \\ A_{1,n+1} &= y_n + hf(t_n, y_n) = 0.75 + 0.045 \times 2.04 = 0.842 \\ y_{\text{mid}} &= y_n + \tfrac{h}{2} f(t_n, y_n) = 0.75 + \tfrac{0.045}{2} \times 2.04 = 0.796 \\ f\big(t_n + \tfrac{h}{2}, y_{\text{mid}}\big) &= 8\Big[1 - 2\big(0.33 + \tfrac{0.045}{2}\big)\Big]0.796 = 1.88 \\ A_{2,n+1} &= y_{\text{mid}} + \tfrac{h}{2} f(t_n + \tfrac{h}{2}, y_{\text{mid}}) = 0.796 + \tfrac{0.045}{2} 1.88 = 0.838 \\ E &= A_{1.n+1} - A_{2.n+1} = 0.842 - 0.838 = 0.004 \\ r &= \frac{|E|}{h} = \frac{0.004}{0.045} = 0.09 \end{aligned}$$

This time $r = 0.09 < \varepsilon = 0.1$, is acceptable so we set $t_{n+1} = 0.33 + 0.045 = 0.375$ and

$$y_{n+1} = A_{2,n+1} = 0.838$$

The initial trial step size from t_{n+1} to t_{n+2} is

$$\text{new } h = 0.9\frac{\varepsilon}{r}(\text{old } h) = 0.9\ \frac{0.1}{0.09}\ .045 = .045$$

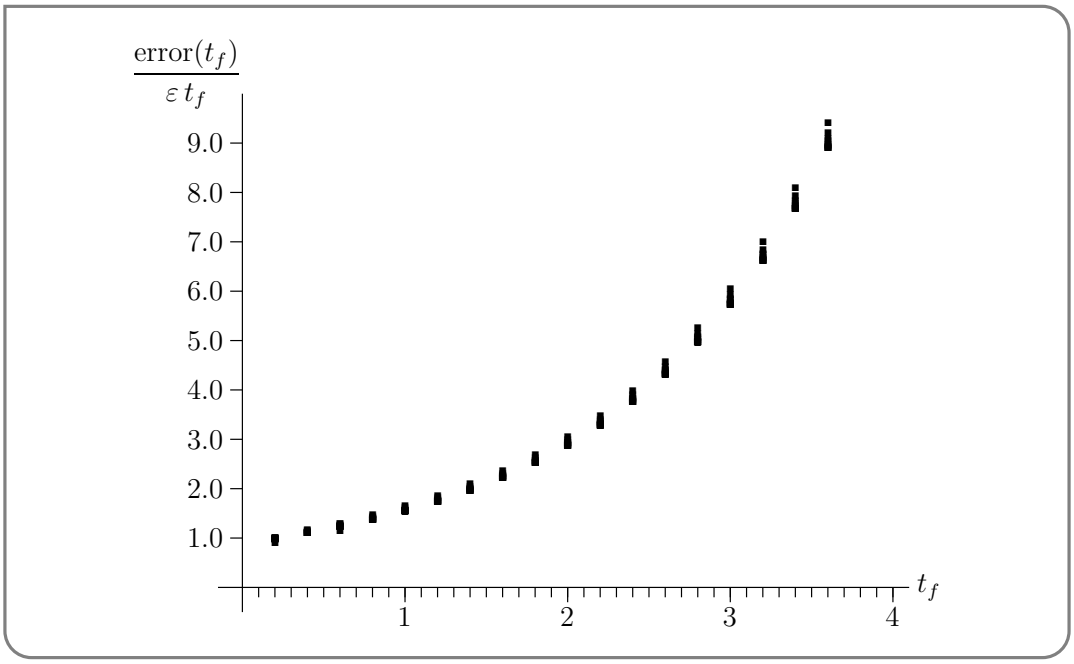
By a fluke, it has turned out that the new h is the same as the old h (to three decimal places). If r had been significantly smaller than ε, then the new h would have been signficantly bigger than the old h - indicating that it is (relatively) easy to estimate things in this region, making a larger step size sufficient.

Example D.3.1

As we said above, we will shortly upgrade the above variable step size method, that we are calling the preliminary version of the Euler/Euler-2step method, to get a much more efficient procedure. Before we do so, let's pause to investigate a little how well our preliminary procedure does at controlling the rate of error production.

We have been referring, loosely, to ε as the desired rate for introduction of error, by our variable step size method, as t advances. If the rate of increase of error were exactly ε, then at final time t_f the accumulated error would be exactly $\varepsilon(t_f - t_0)$. But our algorithm actually chooses the step size h for each step so that the estimated local truncation error in $A_{2,n+1}$ for that step is about εh. We have seen that, once some local truncation error has been introduced, its contribution to the global truncation error can grow exponentially with t_f.

Here are the results of a numerical experiment that illustrate this effect. In this experiment, the above preliminary Euler/Euler-2step method is applied to the initial value problem $y' = t - 2y$, $y(0) = 3$ for $\varepsilon = \frac{1}{16}, \frac{1}{32}, \cdots$ (ten different values) and for $t_f = 0.2, 0.4, \cdots, 3.8$. Here is a plot of the resulting $\frac{\text{actual error at } t=t_f}{\varepsilon t_f}$ against t_f. If the rate of



introduction of error were exactly ε, we would have $\frac{\text{actual error at } t=t_f}{\varepsilon t_f} = 1$. There is a small square on the graph for each different pair ε, t_f. So for each value of t_f there are ten (possibly overlapping) squares on the line $x = t_f$. This numerical experiment suggests that $\frac{\text{actual error at } t=t_f}{\varepsilon t_f}$ is relatively independent of ε and starts, when t_f is small, at about one, as we want, but grows (perhaps exponentially) with t_f.

D.3.2 ▸▸ Euler and Euler-2step (final version)

We are now ready to use a sneaky bit of arithemtic to supercharge our Euler/Euler-2step method. As in our development of the preliminary version of the method, denote by $\phi(t)$ the exact solution to $y' = f(t, y)$ that satisfies the initial condition $\phi(t_n) = y_n$. We have

seen, at the beginning of §D.3.1, that applying one step of Euler with step size h, gives

$$\begin{aligned} A_{1,n+1} &= y_n + hf(t_n, y_n) \\ &= \phi(t_n + h) + Kh^2 + O(h^3) \end{aligned} \tag{E5}$$

and applying one step of Euler-2step with step size h (i.e. applying two steps of Euler with step size $h/2$) gives

$$\begin{aligned} A_{2,n+1} &= y_n + \tfrac{h}{2} f(t_n, y_n) + \tfrac{h}{2} f\big(t_n + \tfrac{h}{2}\,,\, y_n + \tfrac{h}{2} f(t_n, y_n)\big) \\ &= \phi(t_n + h) + \tfrac{1}{2} Kh^2 + O(h^3) \end{aligned} \tag{E6}$$

because the local truncation error introduced in the first half-step was $K(h/2)^2 + O(h^3)$ and that introduced in the second half-step was $K(h/2)^2 + O(h^3)$. Now here is the sneaky bit. Equations (E5) and (E6) are very similar and we can eliminate all Kh^2's by subtracting (E5) from 2 times (E6). This gives

$$\text{2(E6)}-\text{(E5)}: \qquad 2A_{2,n+1} - A_{1,n+1} = \phi(t_n + h) + O(h^3)$$

(no more h^2 term!) or

$$\phi(t_n + h) = 2A_{2,n+1} - A_{1,n+1} + O(h^3) \tag{E7}$$

which tells us that choosing

$$y_{n+1} = 2A_{2,n+1} - A_{1,n+1} \tag{E8}$$

would give a local truncation error of order h^3, rather than the order h^2 of the preliminary Euler/Euler-2step method. To convert the preliminary version of the Euler/Euler-2step algorithm to the final version, we just replace $y_{n+1} = A_{2,n+1}$ by $y_{n+1} = 2A_{2,n+1} - A_{1,n+1}$:

Equation D.3.2 (Euler/Euler-2step Method).

Given $\varepsilon > 0$, t_n, y_n and the current step size h

- compute

$$\begin{aligned} A_{1,n+1} &= y_n + hf(t_n, y_n) \\ A_{2,n+1} &= y_n + \tfrac{h}{2} f(t_n, y_n) + \tfrac{h}{2} f\big(t_n + \tfrac{h}{2}\,,\, y_n + \tfrac{h}{2} f(t_n, y_n)\big) \\ r &= \frac{|A_{1,n+1} - A_{2,n+1}|}{h} \end{aligned}$$

- If $r > \varepsilon$, repeat the first bullet but with the new step size

$$(\text{new } h) = 0.9\,\frac{\varepsilon}{r}\,(\text{old } h)$$

- If $r < \varepsilon$ set

$$\begin{aligned} t_{n+1} &= t_n + h \\ y_{n+1} &= 2A_{2,n+1} - A_{1,n+1} \quad \text{and the new trial step size} \\ (\text{new } h) &= 0.9\,\frac{\varepsilon}{r}\,(\text{old } h) \end{aligned}$$

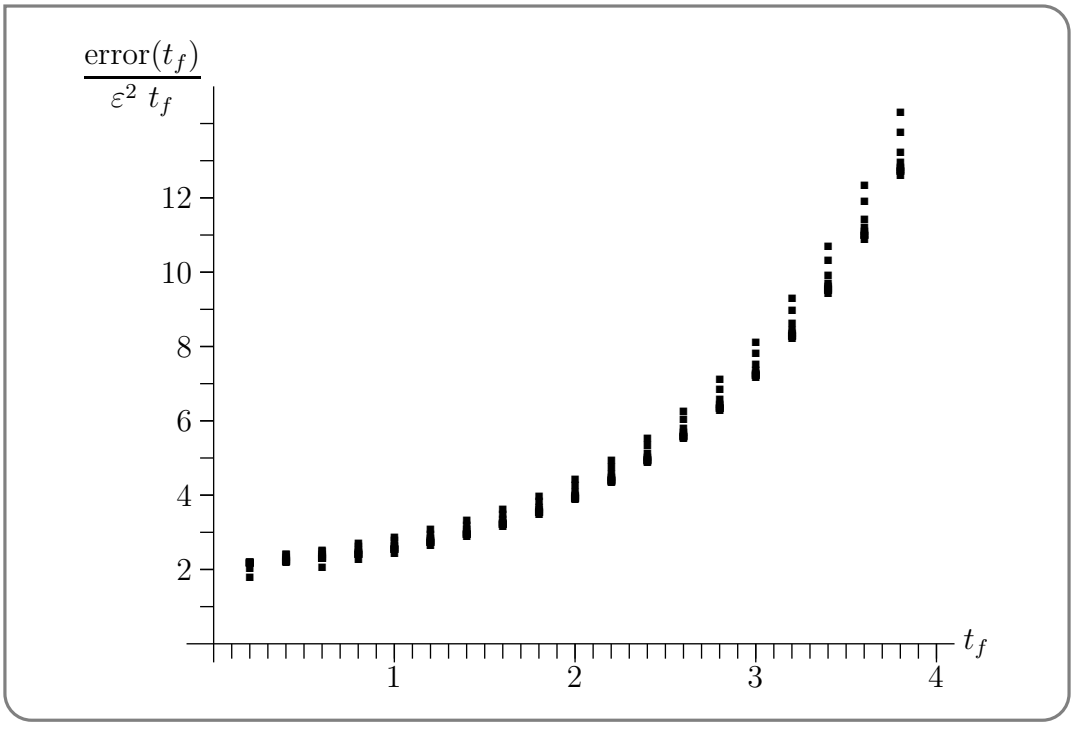
and move on to the next step.

Note that when $r < \varepsilon$, we have $\frac{r}{\varepsilon}h > h$ which indicates that the new h can be larger than the old h. We include the 0.9 to be careful not to make the error of the next step too big.

Let's think a bit about how our final Euler/Euler-2step method should perform.

- The step size here, as in the preliminary version, is chosen so that the local truncation error in $A_{2,n+1}$ per unit increase of t, namely $r = \frac{|A_{1,n+1}-A_{2,n+1}|}{h} \approx \frac{Kh^2/2}{h} = \frac{K}{2}h$, is approximately ε. So h is roughly proportional to ε.
- On the other hand, (E7) shows that, in the full method, local truncation error is being added to y_{n+1} at a rate of $\frac{O(h^3)}{h} = O(h^2)$ per unit increase in t.
- So one would expect that local truncation increases the error at a rate proportional to ε^2 per unit increase in t.
- If the rate of increase of error were exactly a constant time ε^2, then the error accumulated between the initial time $t = 0$ and the final time $t = t_f$ would be exactly a constant times $\varepsilon^2\, t_f$.
- However we have seen that, once some local truncation error has been introduced, its contribution to the global error can grow exponentially with t_f. So we would expect that, under the full Euler/Euler-2step method, $\frac{\text{actual error at } t=t_f}{\varepsilon^2 t_f}$ to be more or less independent of ε, but still growing exponentially in t_f.

Here are the results of a numerical experiment that illustrate this. In this experiment, the above final Euler/Euler-2step method, (D.3.2), is applied to the initial value problem $y' = t - 2y$, $y(0) = 3$ for $\varepsilon = \frac{1}{16}, \frac{1}{32}, \cdots$ (ten different values) and for $t_f = 0.2,\ 0.4,\ \cdots,\ 3.8$. In the following plot, there is a small square for the resulting $\frac{\text{actual error at } t=t_f}{\varepsilon^2 t_f}$ for each different pair ε, t_f.



It does indeed look like $\frac{\text{actual error at } t=t_f}{\varepsilon^2 t_f}$ is relatively independent of ε but grows (perhaps exponentially) with t_f. Note that $\frac{\text{actual error at } t=t_f}{\varepsilon^2 t_f}$ contains a factor of ε^2 in the denominator. The actual error rate $\frac{\text{actual error at } t=t_f}{t_f}$ is much smaller than is suggested by the graph.

D.3.3 ▸▸ Fehlberg's Method

Of course, in practice more efficient and more accurate methods[25] than Euler and Euler-2step are used. Fehlberg's method[26] uses improved Euler and a second more accurate method. Each step involves three calculations of f:

$$\begin{aligned} f_{1,n} &= f(t_n, y_n) \\ f_{2,n} &= f(t_n + h, y_n + h f_{1,n}) \\ f_{3,n} &= f\left(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{4}[f_{1,n} + f_{2,n}]\right) \end{aligned}$$

Once these three evaluations have been made, the method generates two approximations for $y(t_n + h)$:

$$\begin{aligned} A_{1,n+1} &= y_n + \tfrac{h}{2}\left[f_{1,n} + f_{2,n}\right] \\ A_{2,n+1} &= y_n + \tfrac{h}{6}\left[f_{1,n} + f_{2,n} + 4f_{3,n}\right] \end{aligned}$$

Denote by $\phi(t)$ the exact solution to $y' = f(t, y)$ that satisfies the initial condition $\phi(t_n) = y_n$. Now $A_{1,n+1}$ is just the y_{n+1} produced by the improved Euler's method. The local truncation error for the improved Euler's method is of order h^3, one power of h smaller than that for Euler's method. So

$$A_{1,n+1} = \phi(t_n + h) + Kh^3 + O(h^4)$$

and it turns out[27] that

$$A_{2,n+1} = \phi(t_n + h) + O(h^4)$$

So the error in $A_{1,n+1}$ is

$$\begin{aligned} E &= \left|Kh^3 + O(h^4)\right| = \left|A_{1,n+1} - \phi(t_n + h)\right| + O(h^4) \\ &= \left|A_{1,n+1} - A_{2,n+1}\right| + O(h^4) \end{aligned}$$

and our estimate for rate at which error is being introduced into $A_{1,n+1}$ is

$$r = \frac{|A_{1,n+1} - A_{2,n+1}|}{h} \approx |K|h^2$$

per unit increase of t.

25 There are a very large number of such methods. We will only look briefly at a couple of the simpler ones. The interested reader can find more by search engining for such keywords as "Runge-Kutta methods" and "adaptive step size".

26 E. Fehlberg, NASA Technical Report R315 (1969) and NASA Technical Report R287 (1968).

27 The interested reader can find Fehlberg's original paper online (at NASA!) and follow the derivation. It requires careful Taylor expansions and then clever algebra to cancel the bigger error terms.

- If $r > \varepsilon$ we redo this step with a new trial step size chosen so that $|K|(\text{new } h)^2 < \varepsilon$, i.e. $\frac{r}{h^2}(\text{new } h)^2 < \varepsilon$. With our traditional safety factor, we take

$$\text{new } h = 0.9\sqrt{\frac{\varepsilon}{r}}\,h \qquad \text{(the new } h \text{ is smaller)}$$

- If $r \leqslant \varepsilon$ we set $t_{n+1} = t_n + h$ and $y_{n+1} = A_{2,n+1}$ (since $A_{2,n+1}$ should be considerably more accurate than $A_{1,n+1}$) and move on to the next step with trial step size

$$\text{new } h = 0.9\sqrt{\frac{\varepsilon}{r}}\,h \qquad \text{(the new } h \text{ is usually bigger)}$$

D.3.4 ▸▸ The Kutta-Merson Process

The Kutta-Merson process[28] uses two variations of the Runge-Kutta method. Each step involves five calculations[29] of f:

$$\begin{aligned}
k_{1,n} &= f(t_n, y_n)\\
k_{2,n} &= f\left(t_n + \tfrac{1}{3}h, y_n + \tfrac{1}{3}hk_{1,n}\right)\\
k_{3,n} &= f\left(t_n + \tfrac{1}{3}h, y_n + \tfrac{1}{6}hk_{1,n} + \tfrac{1}{6}hk_{2,n}\right)\\
k_{4,n} &= f\left(t_n + \tfrac{1}{2}h, y_n + \tfrac{1}{8}hk_{1,n} + \tfrac{3}{8}hk_{3,n}\right)\\
k_{5,n} &= f\left(t_n + h, y_n + \tfrac{1}{2}hk_{1,n} - \tfrac{3}{2}hk_{3,n} + 2hk_{4,n}\right)
\end{aligned}$$

Once these five evaluations have been made, the process generates two approximations for $y(t_n + h)$:

$$\begin{aligned}
A_{1,n+1} &= y_n + h\left[\tfrac{1}{2}k_{1,n} - \tfrac{3}{2}k_{3,n} + 2k_{4,n}\right]\\
A_{2,n+1} &= y_n + h\left[\tfrac{1}{6}k_{1,n} + \tfrac{2}{3}k_{4,n} + \tfrac{1}{6}k_{5,n}\right]
\end{aligned}$$

The (signed) error in $A_{1,n+1}$ is $\frac{1}{120}h^5K + O(h^6)$ while that in $A_{2,n+1}$ is $\frac{1}{720}h^5K + O(h^6)$ with the same constant K. So $A_{1,n+1} - A_{2,n+1} = \frac{5}{720}Kh^5 + O(h^6)$ and the unknown constant K can be determined, to within an error $O(h)$, by

$$K = \frac{720}{5\,h^5}(A_{1,n+1} - A_{2,n+1})$$

and the approximate (signed) error in $A_{2,n+1}$ and its corresponding rate per unit increase of t are

$$\begin{aligned}
E &= \frac{1}{720}Kh^5 = \frac{1}{5}(A_{1,n+1} - A_{2,n+1})\\
r &= \frac{|E|}{h} = \frac{1}{720}|K|h^4 = \frac{1}{5\,h}\left|A_{1,n+1} - A_{2,n+1}\right|
\end{aligned}$$

28 R.H. Merson, "An operational method for the study of integration processes" , Proc. Symp. Data Processing , Weapons Res. Establ. Salisbury , Salisbury (1957) pp. 110–125.

29 Like the other methods described above, the coefficients 1/3, 1/6, 1/8 etc. are chosen so as to cancel larger error terms. While determining the correct choice of coefficients is not conceptually difficult, it does take some work and is beyond the scope of this appendix. The interested reader should search-engine their way to a discussion of adaptive Runge-Kutta methods.

- If $r > \varepsilon$ we redo this step with a new trial step size chosen so that $\frac{1}{720}|K|(\text{new h})^4 < \varepsilon$, i.e. $\frac{r}{h^4}(\text{new } h)^4 < \varepsilon$. With our traditional safety factor, we take

$$\text{new } h = 0.9\left(\frac{\varepsilon}{r}\right)^{1/4} h$$

- If $r \leqslant \varepsilon$ we set $t_{n+1} = t_n + h$ and $y_{n+1} = A_{2,n+1} - E$ (since E is our estimate of the signed error in $A_{2,n+1}$) and move on to the next step with trial step size

$$\text{new } h = 0.9\left(\frac{\varepsilon}{r}\right)^{1/4} h$$

D.3.5 ▸▸ The Local Truncation Error for Euler-2step

In our description of Euler/Euler-2step above we simply stated the local truncation error without an explanation. In this section, we show how it may be derived. We note that very similar calculations underpin the other methods we have described.

In this section, we will be using partial derivatives and, in particular, the chain rule for functions of two variables. That material is covered in Chapter 2 of the CLP-3 text. If you are not yet comfortable with it, you can either take our word for those bits, or you can delay reading this section until you have learned a little multivariable calculus.

Recall that, by definition, the local truncation error for an algorithm is the (signed) error generated by a single step of the algorithm, under the assumptions that we start the step with the exact solution and that there is no roundoff error[30] Denote by $\phi(t)$ the exact solution to

$$\begin{aligned} y'(t) &= f(t,y) \\ y(t_n) &= y_n \end{aligned}$$

In other words, $\phi(t)$ obeys

$$\begin{aligned} \phi'(t) &= f\big(t,\phi(t)\big) \qquad \text{for all } t \\ \phi(t_n) &= y_n \end{aligned}$$

In particular $\phi'(t_n) = f\big(t_n,\phi(t_n)\big) = f(t_n,y_n)$ and, carefully using the chain rule, which is (2.4.2) in the CLP-3 text,

$$\begin{aligned} \phi''(t_n) = \frac{\mathrm{d}}{\mathrm{d}t} f\big(t,\phi(t)\big)\Big|_{t=t_n} &= \Big[f_t\big(t,\phi(t)\big) + f_y\big(t,\phi(t)\big)\phi'(t)\Big]_{t=t_n} \\ &= f_t(t_n,y_n) + f_y(t_n,y_n)\, f(t_n,y_n) \end{aligned} \tag{E9}$$

Remember that f_t is the partial derivative of f with respect to t, and that f_y is the partial derivative of f with respect to y. We'll need (E9) below.

30 We should note that in serious big numerical computations, one really does have to take rounding errors into account because they can cause serious problems. The interested reader should search-engine their way to the story of Edward Lorenz's numerical simulations and the beginnings of chaos theory. Unfortunately we simply do not have space in this text to discuss all aspects of mathematics.

By definition, the local truncation error for Euler is

$$E_1(h) = \phi(t_n + h) - y_n - hf\big(t_n, y_n\big)$$

while that for Euler-2step is

$$E_2(h) = \phi(t_n + h) - y_n - \tfrac{h}{2} f(t_n, y_n) - \tfrac{h}{2} f\big(t_n + \tfrac{h}{2}, y_n + \tfrac{h}{2} f(t_n, y_n)\big)$$

To understand how $E_1(h)$ and $E_2(h)$ behave for small h we can use Taylor expansions ((3.4.10) in the CLP-1 text) to write them as power series in h. To be precise, we use

$$g(h) = g(0) + g'(0)\, h + \tfrac{1}{2} g''(0)\, h^2 + O(h^3)$$

to expand both $E_1(h)$ and $E_2(h)$ in powers of h to order h^2. Note that, in the expression for $E_1(h)$, t_n and y_n are constants — they do not vary with h. So computing derivatives of $E_1(h)$ with respect to h is actually quite simple.

$$\begin{aligned}
E_1(h) &= \phi(t_n + h) - y_n - hf\big(t_n, y_n\big) & E_1(0) &= \phi(t_n) - y_n = 0 \\
E_1'(h) &= \phi'(t_n + h) - f\big(t_n, y_n\big) & E_1'(0) &= \phi'(t_n) - f\big(t_n, y_n\big) = 0 \\
E_1''(h) &= \phi''(t_n + h) & E_1''(0) &= \phi''(t_n)
\end{aligned}$$

By Taylor, the local truncation error for Euler obeys

Equation D.3.3.

$$E_1(h) = \tfrac{1}{2}\phi''(t_n)h^2 + O(h^3) = Kh^2 + O(h^3) \qquad \text{with } K = \tfrac{1}{2}\phi''(t_n)$$

Computing arguments of $E_2(h)$ with respect to h is a little harder, since h now appears in the arguments of the function f. As a consequence, we have to include some partial derivatives.

$$E_2(h) = \phi(t_n + h) - y_n - \frac{h}{2} f(t_n, y_n) - \frac{h}{2} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)$$

$$E_2'(h) = \phi'(t_n + h) - \frac{1}{2} f(t_n, y_n) - \frac{1}{2} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big) - \frac{h}{2} \underbrace{\frac{\mathrm{d}}{\mathrm{d}h} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)}_{\text{leave this expression as is for now}}$$

$$E_2''(h) = \phi''(t_n + h) - 2 \times \frac{1}{2} \underbrace{\frac{\mathrm{d}}{\mathrm{d}h} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)}_{\text{leave this one too}} - \frac{h}{2} \underbrace{\frac{\mathrm{d}^2}{\mathrm{d}h^2} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)}_{\text{and leave this one too}}$$

Since we only need $E_2(h)$ and its derivatives at $h = 0$, we don't have to compute the $\frac{d^2 f}{dh^2}$ term (thankfully) and we also do not need to compute the $\frac{df}{dh}$ term in E_2'. We do, however, need $\frac{df}{dh}\Big|_{h=0}$ for $E_2''(0)$.

$$
\begin{aligned}
E_2(0) &= \phi(t_n) - y_n = 0 \\
E_2'(0) &= \phi'(t_n) - \frac{1}{2} f(t_n, y_n) - \frac{1}{2} f(t_n, y_n) = 0 \\
E_2''(0) &= \phi''(t_n) - \frac{d}{dh} f\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)\Big|_{h=0} \\
&= \phi''(t_n) - \frac{1}{2} f_t\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)\Big|_{h=0} \\
&\qquad\qquad - \frac{1}{2} f(t_n, y_n)\, f_y\Big(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n)\Big)\Big|_{h=0} \\
&= \phi''(t_n) - \frac{1}{2} f_t(t_n, y_n) - \frac{1}{2} f_y(t_n, y_n)\, f(t_n, y_n) \\
&= \tfrac{1}{2}\phi''(t_n) \qquad \text{by (E9)}
\end{aligned}
$$

By Taylor, the local truncation error for Euler-2step obeys

Equation D.3.4.

$$E_2(h) = \frac{1}{4}\phi''(t_n)\, h^2 + O(h^3) = \tfrac{1}{2} K h^2 + O(h^3) \qquad \text{with } K = \tfrac{1}{2}\phi''(t_n)$$

Observe that the K in (D.3.4) is identical to the K in (D.3.3). This is exactly what we needed in our analysis of Sections D.3.1 and D.3.2.