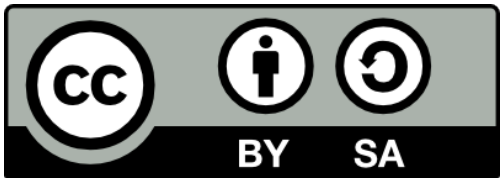
INTRODUCTORY

PHYSICS

Building Models to Describe Our World

Ryan Martin • Emma Neary • Olivia Woodman

License



Preface

About this textbook

This textbook is written to fill several needs that we believe were not already met by the many existing introductory physics textbooks. First, we wanted to ensure that the textbook is free to use for students and professors. Second, we wanted to design a textbook that is mindful of the new pedagogies being used in introductory physics, by writing it in a way that is adapted to a flipped-classroom approach where students complete readings, think about the readings, and then discuss the material in class. Third, we wanted to create a textbook that also addresses the experimental aspect of physics, by proposing experiments to be conducted at home or in the lab, as well as providing guidelines for designing experiments and reporting on experimental results. Finally, we wanted to create a textbook that is a sort of "living document", that professors can edit and re-mix for their own needs, and to which students can contribute material as well. The textbook is hosted on GitHub, which allows anyone to make suggestions, point out issues and mistakes, and contribute material.

This textbook is meant to be paired with the accompanying "Question Library", which contains many practice problems, many of which were contributed by students.

This textbook would not have been possible without the support of Queen's University and the Department of Physics, Engineering Physics & Astronomy at Queen's University, as well as the many helpful discussions with the students, technicians and professors at Queen's University.

Hello from the authors

Ryan Martin I am a professor of physics at Queen's University. My main research is in the field of particle astrophysics, particularly in studying the properties of neutrinos. I grew up in Switzerland, obtained my Bachelor's, Master's and Ph.D. at Queen's University. I was then a postdoctoral fellow at Lawrence Berkeley National Laboratory, a faculty at the University of South Dakota, before returning to Queen's. I am particularly passionate about education, and I am always seeking opportunities to involve students in helping to make education more accessible. I also like to cook and to play volleyball.

Emma Neary I am currently a second year physics major and QuARMS (Queen's University Accelerated Route to Medical School)

student, as well as a native of St. John's, Newfoundland. Uniting the perspectives of students and professors in an accessible way is important to me. I strongly believe in the importance of building physical models; whether it be in physics, medicine, sciences or the arts. It has been my goal to infuse the textbook with the theme of modelling in a creative and engaging way. Aside from doing physics, I enjoy hiking, dancing, reading and doing research in gastroenterology and neuropsychiatry.

Olivia Woodman I am a currently a third year undergraduate student at Queen's Univeristy, majoring in physics. The flipped classroom approach has been beneficial to my own learning, and I think that we have created a textbook that really complements this learning style. Throughout this book, I have shared my thoughts on various topics in physics, as well as some useful tips and tricks. I hope that students enjoy using this book and continue to contribute to it in the future. Working on this textbook has also allowed me to combine my love of physics with my love of doodling, so I hope you enjoy the drawings!

How to use this textbook

This textbook is designed to be used in a flipped-classroom approach, where students complete readings at home, and the material is then discussed in class. The material is thus presented fairly succinctly, and contains **Checkpoint Questions** throughout that are meant to be answered as the students complete the reading. We suggest including these Checkpoint Questions as part of a quiz in a reading assignment (marked based on completion, not correctness), and then using these questions as a starting point for discussions in class.

For topics that are particularly difficult, we have included **Thought Boxes** written by students that try to present the material in a different light. We are always happy if students (or professors) wish to contribute additional thought boxes.

Chapters start with a set of **Learning outcomes** and an **Opening question** to help students have a sense of the chapter contents. The chapters have **Examples** throughout, as well as additional practice problems at the end. The **Question Library** should be consulted for additional practice problems. At the end of the chapter, a **Summary** presents the key points from the chapter. We suggest that students carefully read the summaries to make sure that they understand the contents of the chapter (and potentially identify, before reading the chapter, if the content is review to them). At the end of the chapters, we also present a section to **Think about the material**. This includes questions that can be assigned in reading assignments to research applications of the material or historical context. The thinking about the material section also includes experiments that can be done at home (as part of the reading assignment) or in the lab.

Appendices cover the main background in mathematics (Calculus and Vectors), as well as present an introduction to programming in python, which we feel is a useful skill to have in science. There is also an Appendix that is intended to guide work in the lab, by providing examples of how to write experimental proposals and reports, as well as guidelines

for reviewing proposals and reports. We believe that introductory laboratories should not be be "recipe-based", but rather that students should take an approach similar to that of a researcher in designing (proposing) an experiment, conducting it, and reviewing the proposals and results of their peers.

Credits

This textbook, and especially the many questions in the Question Library would not have been possible without the many contributions from students, teaching assistants and other professors. Below is a list of the people that have contributed material that have made this textbook and Question Library possible.

Adam McCaw
Ali Pirhadi
Alexis Brossard
Amy Van Nest
Ceaira Heimstra
Damara Gagnier
Daniel Barake
Daniel Tazbaz
David Cutler
Emily Darling
Emily Mendelson
Emily Wener
Emma Lanciault
Genevieve Fawcett
Gregory Love
Haoyuan Wang
Jack Fitzgerald
James Godfrey
Jenna Vanker
Jesse Fu
Jesse Simmons
Jessica Grennan
Joanna Fu
Jonathan Abott
Josh Rinaldo
Kate Fenwick
Madison Facchini
Marie Vidal
Matt Routliffe
Maya Gibb
Nicholas Everton
Nick Brown
Nicole Gaul
Noah Rowe
Olivia Bouaban
Patrick Singal
Qiqi Zhang
Quentin Sanders
Robin Joshi
Ryan Underwood
Sam Connolly
Sara Stephens
Shona Birkett
Stephanie Ciccone
Talia Castillo
Tamy Puniani
Thomas Faour
Troy Allen
Wei Zhuolin
Yumian Chen
Zifeng Chen
Zoe Macmillan

Contents

1 Applying Newton's Laws

In this chapter, we take a closer look at how to use Newton's Laws to build models to describe motion. Whereas the previous chapter was focused on identifying the forces that are acting on an object, this chapter focuses on using those forces to describe the motion of the object.

Newton's Laws are meant to describe "point particles", that is, objects that can be thought of as a point and thus have no orientation. A block sliding down a hill, a person on a merry-go-round, a bird flying through the air can all be modelled as point particles, as long as we do not need to model their orientation. In all of these cases, we can model the forces on the object using a free-body diagram as the location of where the forces are applied on the object do not matter. In later chapters, we will introduce the tools required to apply Newton's Second Law to objects that can rotate, where we will see that the location of where a force is exerted matters.

Learning Objectives

- Understand when an object's motion can be modelled as one dimensional (linear).
- Be able to develop models for objects undergoing linear motion.
- Be able to develop models for objects undergoing circular motion.
- Be able to develop models for objects undergoing arbitrary three dimensional motion.
- Understand the forces involved in circular motion, and understand that "centripetal" and "centrifugal" forces are not really forces.

Think About It

If a person swings on a swing where the ropes are damaged, where are the ropes most likely to break?

A) at the bottom of the trajectory, when the speed is the greatest.
B) at the top of the trajectory, when the speed is zero.
C) at the point in the trajectory where the speed is one half of its maximal value.

1.1 Statics

When using Newton's Laws to model an object, one can identify two broad categories of situations: static and dynamic. In static situations, the acceleration of the object is zero. By Newton's Second Law, this means that the vector sum of the forces (and torques, as we will see in a later chapter) exerted on an object must be zero. In dynamic situations, the acceleration of the object is non-zero.

For static problems, since the acceleration vector is zero, we can choose a coordinate system in a way that results in as many forces as possible being aligned with the axes (so that we minimize the number of forces that we need to break up into components).

Example 1-1

You push horizontally with a force $\vec{F}$ on a box of mass m that is resting against a vertical wall, as shown in Figure 1.1. The coefficient of static friction between the wall and the box is μ_s. What is the minimum magnitude of the force that you must exert for the box to remain stationary?

Solution

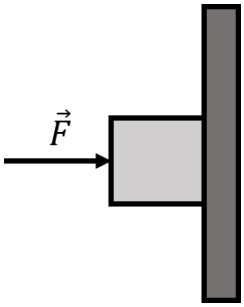


Figure 1.1: A horizontal force exerted on box that is resting against a wall.

Since the acceleration of the box is zero, the vector sum of the forces exerted on the box is zero. We start by identifying the forces exerted on the box; these are:

1. $\vec{F}$, the horizontal force that you exert on the box.
2. $\vec{F}_g$, the weight of the box, with magnitude mg.

3. $\vec{N}$, a normal force exerted by the wall on the box. The force is in the horizontal direction, in the opposite direction to $\vec{F}$.
4. $\vec{f}_s$, a vertical force of static friction between the wall and the box. The force points upwards as the "impeding motion" of the block is downwards. The force will have at most a magnitude of $f_s \leq \mu_s N$, since the force of static friction depends on the other forces exerted on the object.

The forces are shown in the free-body diagram in Figure 1.2, along with our choice of coordinate system which was chosen so that all forces are either in the x or y direction.

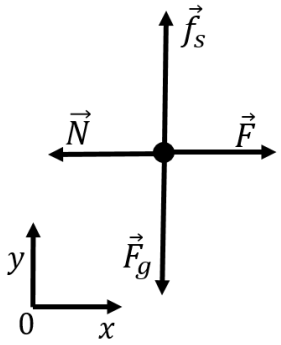


Figure 1.2: Free-body diagram of the forces exerted on the box.

The x component of Newton's Second Law is:

$$\sum F_x = F - N = 0$$
$$\therefore N = F$$

which tells us that the normal force exerted by the wall has the same magnitude as the applied force, $\vec{F}$. The y component of Newton's Second Law is:

$$\sum F_y = f_s - F_g = 0$$
$$\therefore f_s - mg = 0$$
$$\therefore f_s = mg$$

which tells us that the force of friction must have the same magnitude as the weight. This makes sense, since they are the only forces with components in the y direction, and thus, they must cancel each other out.

The force of friction will be less than or equal to $\mu_s N$, and thus less than or equal to $\mu_s F$, since $\vec{F}$ and $\vec{N}$ have the same magnitude (from the x component of Newton's

Second Law). Furthermore, since $f_s = mg$, we can write:

$$\begin{aligned} f_s &\leq \mu_s F \\ \therefore mg &\leq \mu_s F \\ \therefore \frac{mg}{\mu_s} &\leq F \end{aligned}$$

which gives us the condition that $F \geq mg/\mu_s$, and thus the minimum magnitude of F in order to keep the box from sliding down.

Although we used the lesser than or equal to sign in the above equations, we could have used an equal sign if we were confident that the force of friction has its maximal magnitude, $f_s = \mu_s N$. The maximal magnitude of the force of friction is proportional to the force that we exert (since $N = F$); if we want to exert the least amount of force F, then we need the force of friction to be equal to its maximal magnitude which needs to be equal to the weight of the box.

Discussion: This model for the minimal required force makes sense because:

- The dimension of mg/μ_s is force.
- If the mass of the box is increased, then one needs to push harder against the box to keep it up.
- If the coefficient of static friction, μ_s, is increased, one does not need to push as hard.

1.2 Linear motion

We can describe the motion of an object whose *velocity vector does not continuously change direction* as "linear" motion. For example, an object that moves along a straight line in a particular direction, then abruptly changes direction and continues to move in a straight line can be modelled as undergoing linear motion over two different segments (which we would model individually). An object moving around a circle, with its velocity vector continuously changing direction, would not be considered to be undergoing linear motion. For example, paths of objects undergoing linear and non-linear motion are illustrated in Figure 1.3.

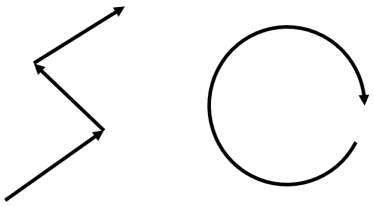


Figure 1.3: (Left:) Displacement vectors for an object undergoing three segments that can each be modelled as linear motion. (Right:) Path of an object whose velocity vector changes continuously and cannot be considered as linear motion.

When an object undergoes linear motion, we always model the motion of the object over straight segments separately. Over one such segment, the acceleration vector will be co-linear with the displacement vector of the object (parallel or anti-parallel - note that the

acceleration can change direction as it would from a spring force, but will always be co-linear with the displacement).

Example 1-2

A block of mass m is placed at rest on an incline that makes an angle θ with respect to the horizontal, as shown in Figure 1.4. The block is nudged slightly so that the force of static friction is overcome and the block starts to accelerate down the incline. At the bottom of the incline, the block slides on a horizontal surface. The coefficient of kinetic friction between the block and the incline is μ_{k1}, and the coefficient of kinetic friction between the block and horizontal surface is μ_{k2}. If one assumes that the block started at rest a distance L from the bottom of the incline, how far along the horizontal surface will the block slide before stopping?

Solution

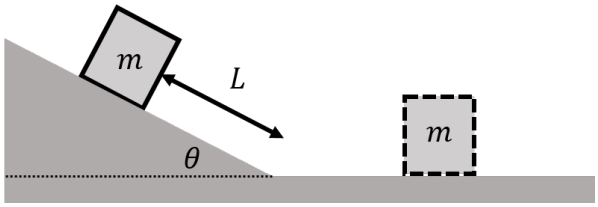


Figure 1.4: A block slides down an incline before sliding on a flat surface and stopping.

We can identify that this is linear motion that we can break up into two segments: (1) the motion down the incline, and (2), the motion along the horizontal surface. We will thus identify the forces, draw the free-body diagram for the block, and use Newton's Second Law twice, once for each segment.

It is often useful to describe the motion in words to help us identify the steps required in building a model for the block. In this case we could say that:

1. The block slides down the incline and accelerates in the direction of motion. By identifying the forces and applying Newton's Second Law, we can determine its acceleration which will be parallel to the incline.
2. The block will reach a certain speed at the bottom of the incline, which we can determine from kinematics by knowing that the block travelled a distance L, with a known acceleration and that it started at rest.
3. The block will decelerate along the horizontal surface. Again, by identifying the forces and using Newton's Second Law, we will be able to determine the acceleration of the block.
4. The block will stop after having travelled an unknown distance, which we can

find by using kinematics and knowing the acceleration of the block as well as its initial velocity at the bottom of the incline.

Our first step is thus to identify the forces on the block while it is on the incline. These are:

1. $\vec{F}_g$, its weight.
2. $\vec{N}_1$, a normal force exerted by the incline.
3. $\vec{f}_{k1}$, a force of kinetic friction exerted by the incline. The force is opposite of the direction of motion, and has a magnitude given by $f_{k1} = \mu_{k1} N_1$.

These are shown on the free-body diagram in Figure 1.5. As usual, we drew the acceleration, $\vec{a}_1$, on the free-body diagram, and chose the direction of the x axis to be parallel to the acceleration.

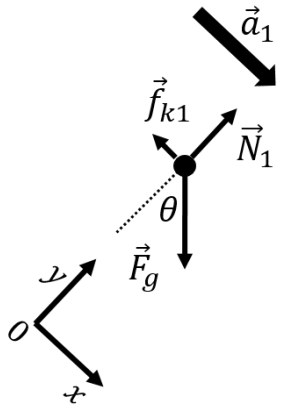


Figure 1.5: Free-body diagram for the block when it is on the incline.

Writing out the x component of Newton's Second Law, and using the fact that the acceleration is in the x direction ($\vec{a} = a_1 \hat{x}$):

$$\sum F_x = F_g \sin\theta - f_{k1} = ma_1$$
$$\therefore mg \sin\theta - \mu_{k1} N_1 = ma_1$$

where we expressed the magnitude of the kinetic force of friction in terms of the normal force exerted by the plane, and the weight in terms of the mass and gravitational field, g. The y component of Newton's Second Law can be written:

$$\sum F_y = N_1 - F_g \cos\theta = 0$$
$$\therefore N_1 = mg \cos\theta$$

which we used to express the normal force in terms of the weight. We can use this expression for the normal force by substituting it into the equation we obtained from

the x component to find the acceleration along the incline:

$$\begin{aligned} mg\sin\theta - \mu_{k1}N_1 &= ma_1 \\ mg\sin\theta - \mu_{k1}mg\cos\theta &= ma_1 \\ \therefore a_1 &= g(\sin\theta - \mu_{k1}\cos\theta) \end{aligned}$$

Now that we know the acceleration down the incline, we can easily find the velocity at the bottom of the incline using kinematics. We choose the origin of the x axis to be zero where the block started ($x_0 = 0$), so that the block is at position $x = L$ at the bottom of the incline. Using kinematics, we can find the speed, v, given that the initial speed, $v_0 = 0$:

$$\begin{aligned} v^2 - v_0^2 &= 2a_1(x - x_0) \\ v^2 &= 2a_1L \\ \therefore v &= \sqrt{2a_1L} \\ &= \sqrt{2Lg(\sin\theta - \mu_{k1}\cos\theta)} \end{aligned}$$

We can now proceed to build a model for the second segment. We first identify the forces on the block when it is on the horizontal surface; these are:

1. $\vec{F}_{g1}$, its weight.
2. $\vec{N}_2$, a normal force exerted by the horizontal surface. This is in general different than the normal force exerted when the block was on the inclined plane.
3. $\vec{f}_{k2}$, a force of kinetic friction exerted by the horizontal surface. The force is opposite of the direction of motion, and has a magnitude given by $f_{k2} = \mu_{k2}N_2$.

The forces are illustrated by the free-body diagram in Figure 1.6, where we showed the acceleration vector, $\vec{a}_2$, which we determined to be to the left since the block is decelerating. We also chose an xy coordinate system such that the x axis is anti-parallel to the acceleration, so that the motion is in the positive x direction (and the acceleration in the negative x direction).

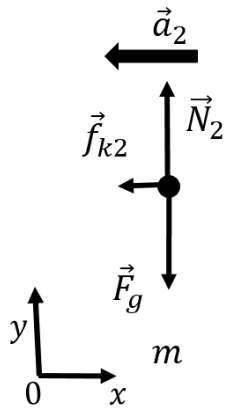


Figure 1.6: Free-body diagram for the block when it is sliding along the horizontal surface. We (arbitrarily) chose the positive x direction to be in the direction of motion and anti-parallel to the acceleration. We could easily have chosen the opposite direction.

Writing out the x component of Newton's Second Law:

$$\sum F_x = -f_{k2} = -ma_2$$
$$\therefore \mu_{k2} N_2 = ma_2$$

where we expressed the force of kinetic friction using the normal force. We have to be careful here with the sign of the acceleration; the equation that we wrote implies that a_2 is a positive number, since μ_{k2} is positive and N_2 is also positive (it is the magnitude of the normal force). a_2 is the magnitude of the acceleration, and we included the fact that the acceleration points in the negative x direction when we put a negative sign in the first line. The x component of the acceleration is $-a_2$, and the vector is given by $\vec{a}_2 = -a_2\hat{x}$.

The y component of Newton's Second Law will allow us to find the normal force:

$$\sum F_y = N_2 - F_g = 0$$
$$\therefore N_2 = mg$$

which we can substitute back into the x equation to find the magnitude of the acceleration along the horizontal surface:

$$ma_2 = \mu_{k2} N_2$$
$$\therefore a_2 = \mu_{k2} g$$

Now that we have found the acceleration along the horizontal surface, we can use kinematics to find the distance that the block travelled before stopping. We choose the origin of the x axis to be the bottom of the incline ($x_0 = 0$), the acceleration is negative $a_x = -a_2 = -mu_{k2}g$, the final speed is zero, $v = 0$, and the initial speed, v_0 is given by

our model for the first segment. Using one of the kinematic equations:

$$\begin{aligned} v^2 - v_0^2 &= 2(-a_2)(x - x_0) \\ v_0^2 &= 2a_2 x \\ \therefore x &= \frac{1}{2a_2} v_0^2 \\ &= \frac{1}{2\mu_{k2} g} 2Lg(\sin\theta - \mu_{k1}\cos\theta) \\ \therefore x &= \frac{(\sin\theta - \mu_{k1}\cos\theta)}{\mu_{k2}} L \end{aligned}$$

Discussion: The model for the distance x that it takes the block to stop makes sense because:

- All of the terms in the fraction are dimensionless, so the value of x will have the same dimension as L.
- If we make L bigger, then x will be bigger (if we release the block from higher up on the incline, it will have more time to accelerate and will slide further before stopping).
- If we make μ_{k1} bigger, then x will be smaller: if we increase friction on the incline, the block will have a smaller acceleration and smaller speed at the bottom.
- If we increase the friction with the horizontal plane (increase μ_{k2}), then x will be reduced (it won't slide as far if there is more friction on the horizontal plane).
- If we increase θ, the numerator will be larger, so x will increase (the block will accelerate more down a steeper incline and end up further).

Checkpoint 1-1

A present is placed at rest on a plane that is inclined, at a distance L from the bottom of the incline, much like the box in Example 1-2 above. At the bottom of the incline, the box is determined to have a speed v. If the box is instead released from a distance of $4L$ from the bottom of the incline, what will its speed at the bottom of the incline be?

A) v
B) $2v$
C) $4v$
D) it depends on the coefficient of friction between the present and the plane.

1.2.1 Modelling situations where forces change magnitude

So far, the models that we have considered involved forces that remained constant in magnitude. In many cases, the forces exerted on an object can change magnitude and direction. For example, the force exerted by a spring changes as the spring changes length or the force of drag changes as the object changes speed. In these case, even if the object undergoes linear motion, we need to break up the motion into many small segments over which we can assume that the forces are constant. If the forces change continuously, we will need to break up the motion into an infinite number of segments and use calculus.

Consider the block of mass m that is shown in Figure 1.7, which is sliding along a frictionless horizontal surface and has a horizontal force $\vec{F}(x)$ exerted on it. The force has a different magnitude in the three segments of length Δx that are shown. If the block starts at position $x = x_0$ axis with speed v_0, we can find, for example, its speed at position $x_3 = 3\Delta x$, after the block travelled through the three segments.

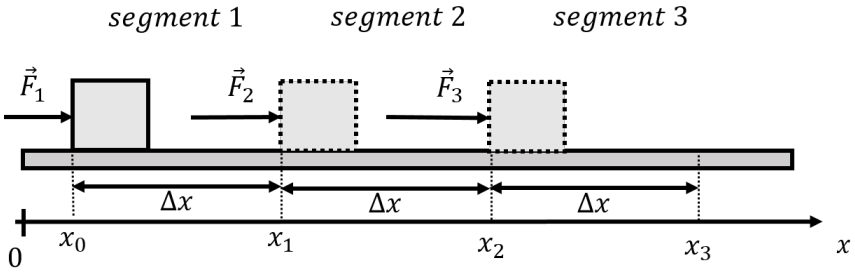


Figure 1.7: A block being pushed along a frictionless horizontal surface with a force that changes.

The horizontal force, $\vec{F}$, exerted on the block can be written as:

$$\vec{F}(x) = \begin{cases} F_1\hat{x} & x < \Delta x \quad \text{(segment 1)} \\ F_2\hat{x} & \Delta x \leq x < 2\Delta x \quad \text{(segment 2)} \\ F_3\hat{x} & 2\Delta x \leq x \quad \text{(segment 3)} \end{cases}$$

as it depends on the location of the block. To find the speed of the block at the end of the third segment, we can model each segment separately. The forces exerted on the block are the same in each segment:

1. $\vec{F}_g$, its weight, with magnitude mg.
2. $\vec{N}$, a normal force exerted by the ground.
3. $\vec{F}(x)$, an applied force that changes magnitude with position and is different in the three different segments.

The forces are illustrated in the free-body diagram show in Figure 1.8.

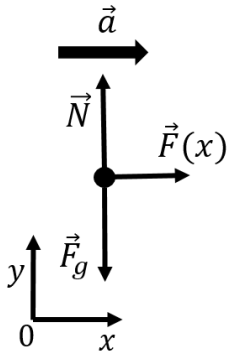


Figure 1.8: Free-body diagram for the block shown in Figure 1.7.

Newton's Second Law can be used to determine the acceleration of the block for each of the three segments, since the forces are constant within one segment. For all three segments, the y component of Newton's Second Law just tells us that the normal force exerted by the ground is equal in magnitude to the weight of the block. The x component of Newton's Second Law gives the acceleration:

$$\sum F_x = F_i = ma_i$$

where we have used the index i to indicate which segment the block is in (i can be 1, 2 or 3). The acceleration of the block in segment i is given by:

$$a_i = \frac{F_i}{m}$$

If the speed of the block is v_0 at the beginning of segment 1 ($x = x_0$), we can find its speed at the end of segment 1 ($x = x_1$), v_1, using kinematics and the fact that the acceleration in segment 1 is a_1:

$$\begin{aligned} v_1^2 - v_0^2 &= 2a_1(x_1 - x_0) \\ v_1^2 &= v_0^2 + 2a_1\Delta x \\ \therefore v_1^2 &= v_0^2 + 2\frac{F_1}{m}\Delta x \end{aligned}$$

We can now easily find the speed at the end of segment 2 ($x = x_2$), v_2, since we know the speed at the beginning of segment 2 (x_1,v_1) and the acceleration a_2:

$$\begin{aligned} v_2^2 - v_1^2 &= 2a_2(x_2 - x_1) \\ \therefore v_2^2 &= v_1^2 + 2a_2\Delta x \\ &= v_0^2 + 2\frac{F_1}{m}\Delta x + 2\frac{F_2}{m}\Delta x \end{aligned}$$

It is easy to show that the speed at the end of the third segment is:

$$v_3^2 = v_0^2 + 2\frac{F_1}{m}\Delta x + 2\frac{F_2}{m}\Delta x + 2\frac{F_3}{m}\Delta x$$

If there were N segments, with the force being different in each segment, we could use the summation notation to write:

$$v_N^2 = v_0^2 + 2\sum_{i=1}^{i=N}\frac{F_i}{m}\Delta x$$

Finally, if the magnitude of the force varied continuously as a function of x, $\vec{F}(x)$, we would model this by taking segments whose length, Δx, tends to zero (and we would need an infinite number of such segments). For example, if we wanted to know the speed of the object at position $x = X$ along the x axis, with a force that was given by $\vec{F}(x) = F(x)\hat{x}$, if the object started at position x_0 with speed v_0, we would take the following limit:

$$v^2 = v_0^2 + \lim_{\Delta x \to 0} 2\sum_{i=1}^{i=N}\frac{F(x)}{m}\Delta x$$

where $\Delta x = \frac{X}{N}$ so that as $\Delta x \to 0$, $N \to \infty$. Of course, integrals are the exact tool that allow us to evaluate the sum in this limit:

$$\lim_{\Delta x \to 0} 2 \sum_{i=1}^{i=N} \frac{F_i}{m} \Delta x = 2 \int_{x_0}^{X} \frac{F(x)}{m} dx$$

and the speed at position $x = X$ is given by:

$$v^2 = v_0^2 + 2 \int_{x_0}^{X} \frac{F(x)}{m} dx$$

Naturally, we can find the above result starting directly from calculus. If the component of the (net) force in the x direction is given by $F(x)$, then the acceleration is given by $a(x) = \frac{F(x)}{m}$. The velocity is related to the acceleration:

$$a(x) = \frac{dv}{dt}$$
$$\therefore dv = a(x)dt$$

We cannot simply integrate the last equation to find that $v = \int a(x)dt$ because the acceleration is given as a function of position, $a(x)$, and not a function of time, t. Thus, we cannot simply take the integral over t and must instead "change variables" to take the integral over x. x and t are related through velocity:

$$v = \frac{dx}{dt}$$
$$\therefore dt = \frac{1}{v} dx$$

We can thus write:

$$dv = a(x)dt = a(x)\frac{1}{v}dx$$

The equation above is called a "separable differential equation", which can also be written:

$$\frac{dv}{dx} = \frac{1}{v} a(x)$$

This is called a differential equation because it relates the derivative of a function (the derivative of v with respect to x, on the left) to the function itself (v appears on the right as well). The differential equation is "separable", because we can separate out all of the quantities that depend on v and on x on different sides of the equation:

$$vdv = a(x)dx$$

This last equation says that vdv is equal to $a(x)dx$. Remember that dx is the length of a very small segment in x, and that dv is the change in velocity over that very small segment.

Since the terms on the left and right are equal, if we sum (integrate) the quantity vdv over many segments, that sum must be equal to the sum (integral) of the quantity $a(x)dx$ over the same segments. Let us choose those segment such that for the beginning of the first interval the position and speed are x_0 and v_0, respectively, and the position and speed at the end of the last segment are X and V, respectively. We then must have that:

$$\begin{aligned}\int_{v_0}^{V} v dv &= \int_{x_0}^{X} a(x)dx \\ \frac{1}{2}V^2 - \frac{1}{2}v_0^2 &= \int_{x_0}^{X} a(x)dx \\ \therefore V^2 &= v_0^2 + 2\int_{x_0}^{X} a(x)dx\end{aligned}$$

which is the same as we found earlier. If the acceleration is constant, we recover our formula from kinematics:

$$\begin{aligned}V^2 &= v_0^2 + 2\int_{x_0}^{X} a dx \\ &= v_0^2 + 2a(X - x_0) \\ \therefore V^2 - v_0^2 &= 2a(X - x_0)\end{aligned}$$

Example 1-3

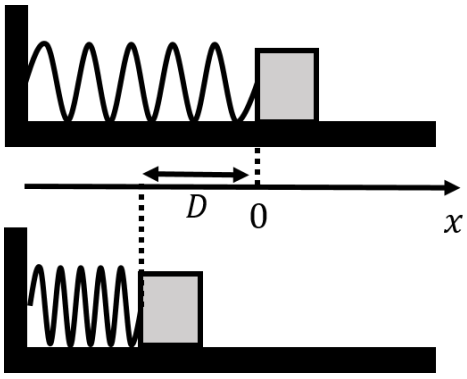


Figure 1.9: A block is launched along a frictionless surface by compressing a spring by a distance D. The top panel shows the spring when at rest, and the bottom panel shows the spring compressed by a distance D just before releasing the block.

A block of mass m can slide freely along a frictionless surface. A horizontal spring, with spring constant, k, is attached to a wall on one end, while the other end can move freely, as shown in Figure 1.9. A coordinate system is defined such that the x axis is horizontal and the free end of the spring is at $x = 0$ when the spring is at rest. The block is pushed against the spring so that the spring is compressed by a distance D. The block is then released. What speed will the block have when it leaves the spring?

Solution

As you recall, the force exerted by a spring depends on the compression or extension of the spring and is given by Hooke's Law:

$$\vec{F}(x) = -kx\hat{x}$$

where x is the position of the free end of the spring and $x = 0$ corresponds to the spring being at rest. In our case, when the edge of the block is located at $x_0 = -D$ (the spring is compressed), the force is thus in the positive x direction (since x_0 is a negative number).

The forces on the block are:

1. $\vec{F}_g$, its weight, with magnitude mg.
2. $\vec{N}$, a normal force exerted by the ground.
3. $\vec{F}(x)$, the spring force.

Since the block is not moving vertically, the magnitude of the normal force must equal the weight $N = mg$, since these are the only forces with components in the vertical direction. The x component of Newton's Second Law gives us the acceleration of the block (which depends on x):

$$\sum F_x = -kx = ma(x)$$
$$\therefore a(x) = -\frac{k}{m}x$$

Again, recall that if x is negative, then the acceleration will be in the positive direction. Since this scenario is exactly the same that we described above in the text, namely a force that varies continuously with position, we can apply the formula that we found earlier for determining the velocity after a varying force has been applied from position $x = x_0$ to position $x = X$:

$$V^2 = v_0^2 + 2\int_{x_0}^{X} a(x)dx$$

V is the final speed that we would like to find, $v_0 = 0$ because the block starts at rest, and $x_0 = -D$ is the starting position of the block. X is the position along the x axis where the block leaves the spring.

We have to think a little about what the value of X should be: when the spring is compressed and the block accelerating, the spring is pushing the block in the positive x direction. Once the block reaches $x = 0$ the spring would want to pull the block

backwards, but since it is not attached to the block, it stops exerting a force on the block at that point. The block thus leaves the spring at $x = 0$, so that the final position is $X = 0$. The speed of the block when it leaves the spring is thus:

$$\begin{aligned} V^2 &= v_0^2 + 2\int_{x_0}^{X} a(x)dx \\ &= 0 + 2\int_{-D}^{0} a(x)dx \\ &= 2\int_{-D}^{0} -\frac{k}{m}xdx \\ &= 2\left[-\frac{k}{m}\frac{1}{2}x^2\right]_{-D}^{0} \\ &= \frac{k}{m}D^2 \\ \therefore V &= \sqrt{\frac{k}{m}}D \end{aligned}$$

Discussion: This model for the speed of the block when it leaves the spring makes sense because:

- The dimension for the expression for V is correct (you should check this!).
- If the spring is compressed more (bigger value of D), then the speed will be higher.
- If the mass is bigger (more inertia), then the final speed will be lower.
- If the spring is stiffer (bigger value of k), then the final speed will be higher.

If you have studied physics before, you may have realized that the speed is easily found by conservation of energy:

$$\frac{1}{2}mV^2 = \frac{1}{2}kD^2$$

which gives the same value for V. As we will see in a later chapter, kinetic and potential energy are defined as they are, precisely because it makes using conservation of energy equivalent to using forces as we just did.

Example 1-4

An object of mass m is released from rest out of a helicopter. The drag (air-resistance) on the object can be modelled as having a magnitude given by bv, where v is the speed of the object and b is a constant of proportionality. How does the velocity of the object depend on time?

Solution

As the object falls through the air, the forces exerted on the object are:

1. F_g, its weight, with magnitude mg, exerted downwards.
2. F_d, the force of drag, with magnitude bv, exerted upwards.

Since the object will fall in a straight line, this is a one-dimensional problem, and we can choose the x axis to be vertical, with positive x pointing downwards, and the origin located where the object was released. The object will thus have a positive acceleration and move in the positive x direction with this choice of coordinate system. This is illustrated in the free-body diagram in Figure 1.10.

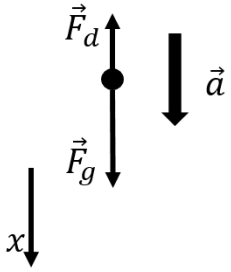


Figure 1.10: Free-body diagram for a block free-falling with drag.

Newton's Second Law for the object gives:

$$\begin{aligned} \sum F_x = F_g - F_d &= ma \\ mg - bv &= ma \\ \therefore a &= g - \frac{b}{m}v \end{aligned}$$

In this case, the acceleration depends explicitly on velocity rather than position, as we had before. However, we can use the same methodology to find how the velocity changes with time. First, we can note that the acceleration is zero if:

$$\begin{aligned} g - \frac{b}{m}v &= 0 \\ \therefore v &= \frac{mg}{b} \end{aligned}$$

That is, once the object reaches a speed of $v_{term} = mg/b$, it will stop accelerating, i.e. it will reach "terminal velocity". Note that this is the same condition as requiring that the drag force (bv) have the same magnitude as the weight (mg).

Writing the acceleration as $a = \frac{dv}{dt}$, we can write:

$$\frac{dv}{dt} = \left(g - \frac{b}{m}v\right)$$

which again, is a separable differential equation, in which we can write the terms that depend on v and those that depend on t on separate sides of the equal sign:

$$\frac{dv}{g - \frac{b}{m}v} = dt$$
$$\frac{dv}{v - \frac{mg}{b}} = -\frac{b}{m}dt$$

where we re-arranged the equation in the second line so that it would be easier to integrate in the next step. We can find the velocity, $v(t)$, at some time, t, by stating that $v = 0$ at $t = 0$ and taking the integrals (sum) on both sides. Again, we are modelling the motion as being made up of a large number of very small segments where the quantities on both sides of the equation are the same. Thus, if we sum (integrate) those quantities over all of the same segments, the left and right hand side of the equations will still be equal to each other:

$$\int_0^{v(t)} \frac{dv}{v - \frac{mg}{b}} = -\int_0^t \frac{b}{m}dt$$
$$\left[\ln\left(v - \frac{mg}{b}\right)\right]_0^{v(t)} = -\frac{b}{m}t$$
$$\ln\left(v(t) - \frac{mg}{b}\right) - \ln\left(-\frac{mg}{b}\right) = -\frac{b}{m}t$$
$$\ln\left(\frac{v(t) - \frac{mg}{b}}{-\frac{mg}{b}}\right) = -\frac{b}{m}t$$

where, in the last line, we used the property that $\ln(a) - \ln(b) = \ln(a/b)$. By taking the exponential on either side of the equation ($e^{\ln(x)} = x$), we can find an expression for the velocity as a function of time:

$$\frac{v(t) - \frac{mg}{b}}{-\frac{mg}{b}} = e^{-\frac{b}{m}t}$$
$$v(t) - \frac{mg}{b} = -\frac{mg}{b}e^{-\frac{b}{m}t}$$
$$\therefore v(t) = \frac{mg}{b} - \frac{mg}{b}e^{-\frac{b}{m}t}$$
$$= \frac{mg}{b}\left(1 - e^{-\frac{b}{m}t}\right)$$

Discussion: This equation tells us that the velocity increases as a function of time, but the rate of increase decreases exponentially with time. At time $t = 0$, the velocity is zero, as expected. As t approaches infinity, v approaches, $\frac{mg}{b}$, which is the terminal velocity. The time dependence of the velocity is illustrated in Figure 1.11.

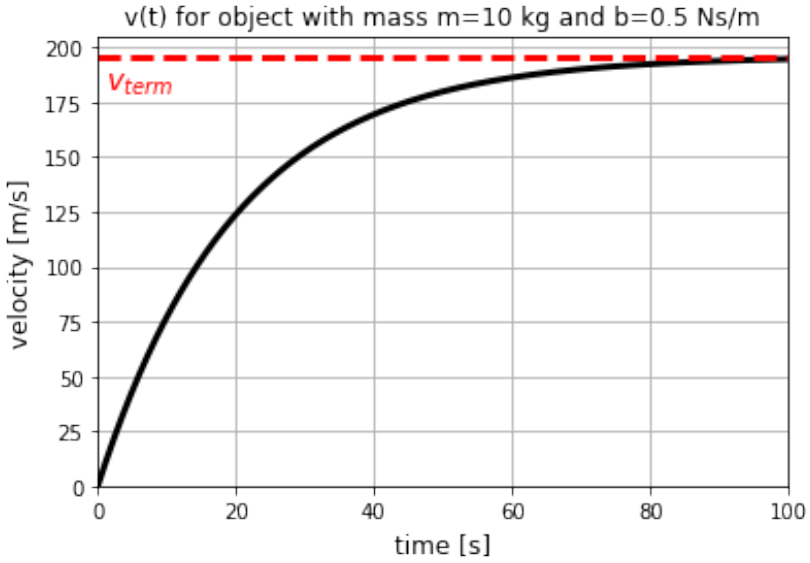


Figure 1.11: Velocity as a function of time for an object of mass $m = 10\,\text{kg}$ which is free-falling from rest with a drag coefficient $b = 0.5\,\text{Ns/m}$.

1.3 Uniform circular motion

As we saw in Chapter **??**, "uniform circular motion" is defined to be motion along a circle with constant speed. This may be a good time to review Section **??** for the kinematics of motion along a circle. In particular, for the uniform circular motion of an object around a circle of radius R, you should recall that:

- The velocity vector, $\vec{v}$, is always tangent to the circle.
- The acceleration vector, $\vec{a}$, is always perpendicular to the velocity vector, because the magnitude of the velocity vector does not change.
- The acceleration vector, $\vec{a}$, always points towards the centre of the circle.
- The acceleration vector has magnitude $a = v^2/R$.
- The angular velocity, ω, is related to the magnitude of the velocity vector by $v = \omega R$ and is constant.
- The angular acceleration, α, is zero for uniform circular motion, since the angular velocity does not change.

In particular, you should recall that even if the speed is constant, the acceleration vector is always non-zero in uniform circular motion because the **velocity changes direction**. According to Newton's Second Law, this implies that there **must be a net force on the**

object that is directed towards the centre of the circle[1] (parallel to the acceleration):

$$\sum \vec{F} = m\vec{a}$$

where the acceleration has a magnitude $a = v^2/R$. Because the acceleration is directed towards the centre of the circle, we sometimes call it a "radial" acceleration (parallel to the radius), a_R, or a "centripetal" acceleration (directed towards the centre), a_c.

Consider an object in uniform circular motion in a horizontal plane on a frictionless surface, as depicted in Figure 1.12.

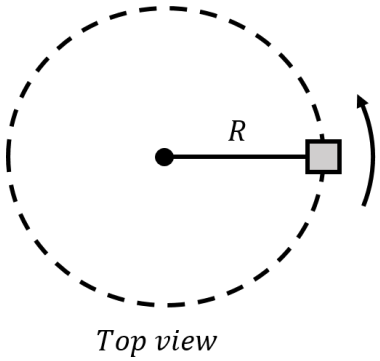


Figure 1.12: An object undergoing uniform circular motion on a frictionless surface, as seen from above.

The only way for the object to undergo uniform circular motion as depicted is if the net force on the object is directed towards the centre of the circle. One way to have a force that is directed towards the centre of the circle is to attach a string between the center of the circle and the object, as shown in Figure 1.12. If the string is under tension, the force of tension will always be towards the centre of the circle. The forces on the object are thus:

1. $\vec{F}_g$, its weight with magnitude mg.
2. $\vec{N}$, a normal forced exerted by the surface.
3. $\vec{T}$, a force of tension exerted by the string.

The forces are depicted in the free-body diagram shown in Figure 1.13 (as viewed from the side), where we also drew the acceleration vector. Note that this free-body diagram is only "valid" at a particular instant in time since the acceleration vector continuously changes direction and would not always be lined up with the x axis.

[1]The sum of the forces is often called the "net force" on an object, and in the specific case of uniform circular motion, that net force is sometimes called the "centripetal force" - however, it is not a force in and of itself and it is always the sum of the forces that points towards the centre of the circle.

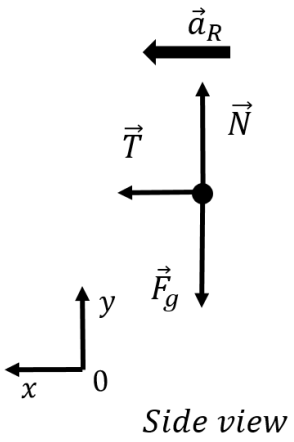


Figure 1.13: Free-body diagram (side view) for the object from Figure 1.13 undergoing uniform circular motion.

Writing out the x and y components of Newton's Second Law:

$$\sum F_x = T = ma_R$$
$$\sum F_y = N - F_g = 0$$

The y component just tells us that the normal force must have the same magnitude as the weight because the object is not accelerating in the vertical direction. The x component tells us the relation between the magnitudes of the tension in the string and the radial acceleration. Using the speed of the object, we can also write the relation between the tension and the speed:

$$T = ma_R = m\frac{v^2}{R}$$

Thus, we find that the tension in the string increases with the square of the speed, and decreases with the radius of the circle.

Checkpoint 1-2

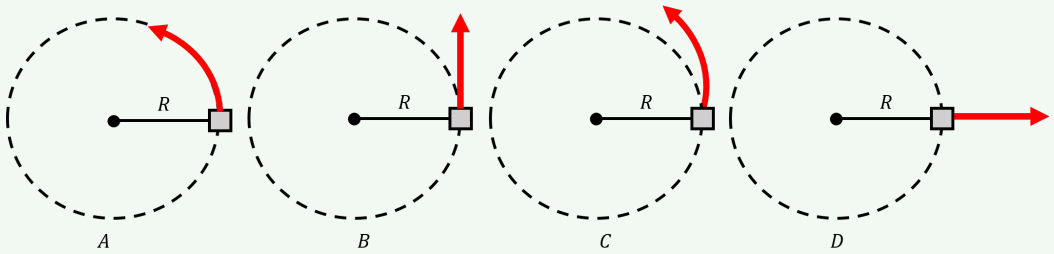


Figure 1.14: Possible trajectories (in red) that the block will follow if the string breaks.

An object is undergoing uniform circular motion in the horizontal plane, when the string connecting the object to the centre of rotation suddenly breaks. What path will the block take after the string broke?

A) A
B) B
C) C
D) D

Example 1-5

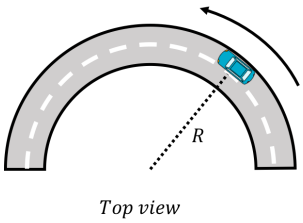


Figure 1.15: A car going around a curve that can be approximated as the arc of a circle of radius R.

A car goes around a curve which can be approximated as the arc of a circle of radius R, as shown in Figure 1.15. The coefficient of static friction between the tires of the car and the road is μ_s. What is the maximum speed with which the car can go around the curve without skidding?

Solution

If the car is going at constant speed around a circle, then the sum of the forces on the car must be directed towards the centre of the circle. The only force on the car that could be directed towards the centre of the circle is the force of friction between the

tires and the road. If the road were perfectly slick (think driving in icy conditions), it would not be possible to drive around a curve since there could be no force of friction. The forces on the car are:

1. $\vec{F}_g$, its weight with magnitude mg.
2. $\vec{N}$, a normal force exerted upwards by the road.
3. $\vec{f}_s$, a force of static friction between the tires and the road. This is static friction, because the surface of the tire does not move relative to the surface of the road if the car is not skidding. The force of static friction has a magnitude that is at most $f_s \leq \mu_s N$.

The forces on the car are shown in the free-body diagram in Figure 1.16.

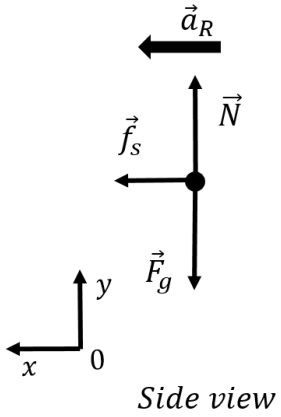


Figure 1.16: Free-body diagram for the car as seen looking at the car from the back (the centre of the curve is towards the left).

The y component of Newton's Second Law tells us that the normal force exerted by the road must equal the weight of the car:

$$\sum F_y = N - F_g = 0$$
$$\therefore N = mg$$

The x component relates the force of friction to the radial acceleration (and thus to the speed):

$$\sum F_x = f_s = ma_R = m\frac{v^2}{R}$$
$$\therefore f_s = m\frac{v^2}{R}$$

The force of friction must be less than or equal to $f_s \leq \mu_s N = \mu_s mg$ (since $N = mg$ from the y component of Newton's Second Law), which gives us a condition on the

speed:

$$f_s = m\frac{v^2}{R} \leq \mu_s mg$$
$$v^2 \leq \mu_s gR$$
$$\therefore v \leq \sqrt{\mu_s gR}$$

Thus, if the speed is less than $\sqrt{\mu_s gR}$, the car will not skid and the magnitude of the force of static friction, which results in an acceleration towards the centre of the circle, will be smaller or equal to its maximal possible value.

Discussion: The model for the maximum speed that the car can travel around the curve makes sense because:

- The dimension of $\sqrt{\mu_s gR}$ is speed.
- The speed is larger if the radius of the curve is larger (one can go faster around a wider curve without skidding).
- The speed is larger if the coefficient of friction is large (if the force of friction is larger, a larger radial acceleration can be sustained).

Example 1-6

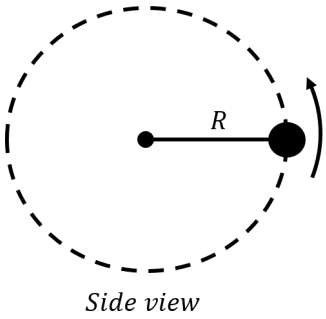


Figure 1.17: A ball attached to a string undergoing circular motion in a vertical plane.

A ball is attached to a mass-less string and executing circular motion along a circle of radius R that is in the vertical plane, as depicted in Figure 1.17. Can the speed of the ball be constant? What is the minimum speed of the ball at the top of the circle if it is able to make it around the circle?

Solution

The forces that are acting on the ball are:

1. $\vec{F}_g$, its weight with magnitude mg.
2. $\vec{T}$, a force of tension exerted by the string.

Figure 1.18 shows the free-body diagram for the forces on the ball at three different locations along the path of the circle.

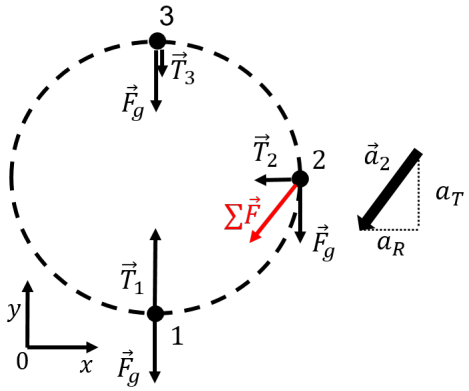


Figure 1.18: A ball attached to a string undergoing circular motion in a vertical plane.

In order for the ball to go around in a circle, there must be at least a component of the net force on the ball that is directed towards the centre of the circle at all times. In the bottom half of the circle (positions 1 and 2), only the tension can have a component directed towards the centre of the circle.

Consider in particular the position labelled 2, when the string is horizontal and the tension is equal to $\vec{T}_2$. The free-body diagram in Figure 1.18 also shows the vector sum of the weight and tension at position 2 (the red arrow labelled $\sum \vec{F}$), which points downwards and to the left. It is thus clearly impossible for the acceleration vector to point towards the centre of the circle, and the acceleration will have components that are both tangential (a_T) to the circle and radial (a_R), as shown by the vector $\vec{a}_2$ in Figure 1.18.

The radial component of the acceleration will change the direction of the velocity vector so that the ball remains on the circle, and the tangential component will reduce the magnitude of the velocity vector. According to our model, it is thus impossible for the ball to go around the circle at constant speed, and the speed must decrease as it goes from position 2 to position 3, no matter how one pulls on the string (you can convince yourself of this by drawing the free-body diagram at any point between points 2 and 3).

The minimum speed for the ball at the top of the circle is given by the condition that the tension in the string is zero just at the top of the trajectory (position 3). The ball can still go around the circle because, at position 3, gravity is towards the centre of

the circle and can thus give an acceleration that is radial, even with no tension. The y component of Newton's Second Law, at position 3 gives:

$$\sum F_y = -F_g = ma_y$$
$$\therefore a_y = -g$$

The magnitude of the acceleration is the radial acceleration, and is thus related to the speed at the top of the trajectory:

$$a_R = -a_y = g = m\frac{v^2}{R}$$
$$\therefore v_{min} = \sqrt{\frac{gR}{m}}$$

which is the minimum speed at the top of the trajectory for the ball to be able to continue along the circle. The tension in the string would change as the ball moves around the circle, and will be highest at the bottom of the trajectory, since the tension has to be bigger than gravity so that the net force at the bottom of the trajectory is upwards (towards the centre of the circle).

Discussion: The model for the minimum speed of the ball at the top of the circle makes sense because:

- $\sqrt{\frac{gR}{m}}$ has the dimension of speed.
- The minimum velocity is larger if the circle has a larger radius (try this with a mass attached at the end of a string).
- The minimum velocity is larger if the mass is bigger (again, try this at home!).

Checkpoint 1-3

Consider a ball attached to a string, being spun in a vertical circle (such as the one depicted in figure 1.17). If you shortened the string, how would the minimum angular velocity (measured at the top of the trajectory) required for the ball to make it around the circle change?

A) It would decrease
B) It would stay the same
C) It would increase

1.3.1 Banked curves

As we saw in Example 1-5, there is a maximum speed with which a car can go around a curve before it starts to skid. You may have noticed that roads, highways especially, are banked where there are curves. Racetracks for cars that go around an oval (the boring kind of car races) also have banked curves. As we will see, this allows the speed of vehicles to be higher when going around the curve; or rather, it makes the curves safer as the speed at which vehicles *would* skid is higher. In Example 1-5, we saw that it was the force of static friction between the tires of the car and the road that provided the only force with a

component towards the centre of the circle. The idea of using a banked curve is to change the direction of the normal force between the road and the car tires so that it, too, has a component in the direction towards the centre of the circle.

Consider the car depicted in Figure 1.19 which is seen from behind making a left turn around a curve that is banked by an angle θ with respect to the horizontal and can be modelled as an arc from a circle of radius R.

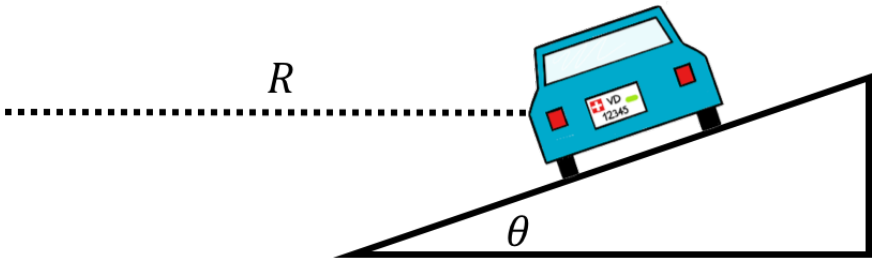


Figure 1.19: A car moving into the page and going around a banked curved so that it is turning towards the left (the centre of the circle is to the left).

The forces exerted on the car are the same as in Example 1-5, except that they point in different directions. The forces are:

1. $\vec{F}_g$, its weight with magnitude mg.
2. $\vec{N}$, a normal force exerted by the road, perpendicular to the surface of the road.
3. $\vec{f}_s$, a force of static friction between the tires and the road. This is static friction, because the surface of the tire does not move relative to the surface of the road if the car is not skidding. The force of static friction has a magnitude that is at most $f_s \leq \mu_s N$ and is perpendicular to the normal force. The force could be either upwards or downwards, *depending on the other forces on the car.*

A free-body diagram for the forces on the car is shown in Figure 1.20, along with the acceleration (which is in the radial direction, towards the centre of the circle), and our choice of coordinate system (choosing x parallel to the acceleration). The direction of the force of static friction is not known *a priori* and depends on the speed of the car:

- If the speed of the car is zero, the force of static friction is upwards. With a speed of zero, the radial acceleration is zero, and the sum of the forces must thus be zero. The impeding motion of the car would be to slide down the banked curve (just like a block on an incline).
- If the speed of the car is very large, the force of static friction is downwards, as the impeding motion of the car would be to slide up the bank. The natural motion of the car is to go in a straight line (Newton's First Law). If the components of the normal force and of the force of static friction directed towards the centre of the circle are too small to allow the car to turn, then the car would slide up the bank (so the impeding motion is up the bank and the force of static friction is downwards).

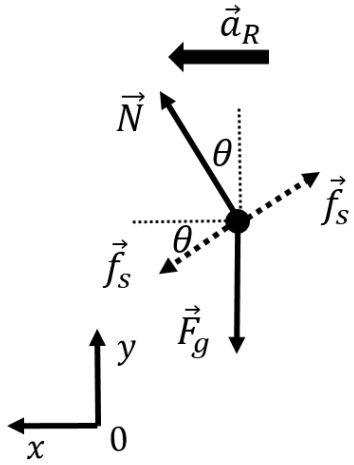


Figure 1.20: Free-body diagram for the forces on the car. The direction of the force of static friction cannot be determined, as it depends on the acceleration of the car, so it is shown twice (with dotted lines).

There is thus an "ideal speed" at which the force of static friction is precisely zero, and the x component of the normal force is responsible for the radial acceleration. At higher speeds, the force of static friction is downwards and increases in magnitude to keep the car's acceleration towards the centre of the circle. At some maximal speed, the force of friction will reach its maximal value, and no longer be able to keep the car's acceleration pointing towards the centre of the circle. At speeds lower than the ideal speed, the force of friction is directed upwards to prevent the car from sliding down the bank. If the coefficient of static friction is too low, it is possible that at low speeds, the car would start to slide down the bank (so there would be a minimum speed below which the car would start to slide down).

Let us model the situation where the force of static friction is identically zero so that we can determine the ideal speed for the banked curve. The only two forces on the car are thus its weight and the normal force. The x and y component of Newton's Second Law give:

$$\sum F_x = N\sin\theta = ma_R = m\frac{v^2}{R}$$
$$\therefore N\sin\theta = m\frac{v^2}{R} \tag{1.1}$$

$$\sum F_y = N\cos\theta - F_g = 0$$
$$\therefore N\cos\theta = mg \tag{1.2}$$

We can divide Equation 1.1 by Equation 1.2, noting that $\tan\theta = \sin\theta/\cos\theta$, to obtain:

$$\tan\theta = \frac{v^2}{gR}$$
$$\therefore v_{ideal} = \sqrt{gR\tan\theta}$$

At this speed, the force of static friction is zero. In practice, one would use this equation to determine which bank angle to use when designing a road, so that the ideal speed is around the speed limit or the average speed of traffic. We leave it as an exercise to determine the maximal speed that the car can go around the curve before sliding out.

1.3.2 Inertial forces in circular motion

As you sit in a car that is going around a curve, you will feel pushed outwards, away from the centre of the circle that the car is going around. This is because of your inertia (Newton's First Law), and your body would go in a straight line if the car were not exerting a net force on you towards the centre of the circle. You are not so much feeling a force that is pushing you outwards as you are feeling the effects of the car seat pushing you inwards; if you were leaning against the side of the car that is on the outside of the curve, you would feel the side of the car pushing you inwards towards the centre of the curve, even if it "feels" like you are pushing outwards against the side of the car.

If we model your motion looking at you from the ground, we would include a force of friction between the car seat (or the side of the car, or both) and you that is pointing towards the centre of the circle, so that the sum of the forces exerted on you is towards the centre of the circle. We can also model your motion from the non-inertial frame of the car. As you recall, because this is a non-inertial frame of reference, we need to include an additional inertial force, $\vec{F}_I$, that points opposite of the acceleration of the car, with magnitude $F_I = ma_R$ (if the net acceleration of the car is a_R). Inside the non-inertial frame of reference of the car, your acceleration (relative to the reference frame, i.e. the car) is zero. This is illustrated by the diagrams in Figure 1.21.

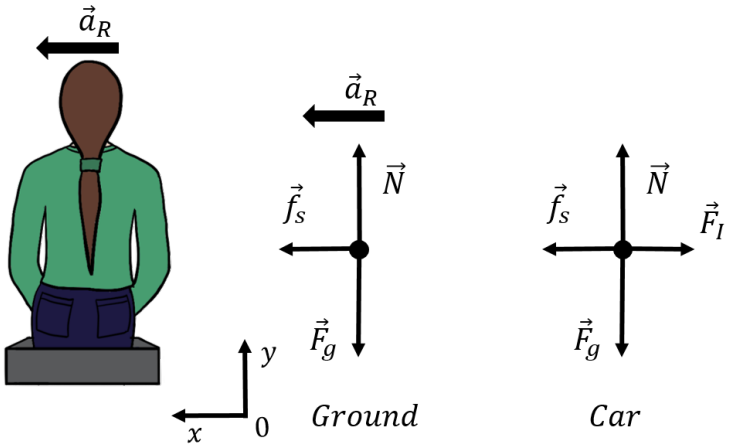


Figure 1.21: (Left:) A person sitting on a car seat in a car turning towards the left. (Centre:) Free-body diagram for the person as modelled in the inertial reference frame of the ground. (Right:) Free-body diagram for the person as modelled in the non-inertial frame of reference of the car, including an additional inertial force.

The y component of Newton's Second Law in both frames of reference is the same:

$$\sum F_y = N - F_g = 0$$
$$\therefore N = mg$$

and simply tells us that the normal force is equal to the weight. In the reference frame of the ground, the x component of Newton's Second Law gives:

$$\sum F_x = f_s = ma_R$$
$$\therefore f_s = m\frac{v^2}{R}$$

In the frame of reference of the car, where your acceleration is zero and an inertial force of magnitude $F_I = mv^2/R$ is exerted on you, the x component of Newton's Second Law gives:

$$\sum F_x = f_s - F_I = 0$$
$$\therefore f_s - m\frac{v^2}{R} = 0$$

which of course, mathematically, is exactly equivalent. The inertial force is not a real force in the sense that it is not exerted by anything. It only comes into play because we are trying to use Newton's Laws in a non-inertial frame of reference. However, it does provide a good model for describing the sensation that we have of being pushed outwards when the car goes around a curve. Sometimes, people will refer to this force as a "centrifugal" force, which means "a force that points away from the centre". You should however remember that this is not a real force exerted on the object, but is the result of modelling motion in a non-inertial frame of reference.

Checkpoint 1-4

Jamie is driving his tricycle around a circular pond. Jamie feels a centrifugal force with magnitude F_I. If Jamie pedals twice as fast, what will be the magnitude of the centrifugal force that he experiences?

A) $\sqrt{2}F_I$
B) $\frac{1}{2}F_I$
C) $2F_I$
D) $4F_I$

1.4 Non-uniform circular motion

In non-uniform circular motion, an object's motion is along a circle, but the object's speed is not constant. In particular, the following will be true

- The object's velocity vector is always tangent to the circle.
- The speed and angular speed of the object are not constant.
- The angular acceleration of the object is not zero.
- The acceleration vector will not point towards the centre of the circle.

Since the acceleration vector does not point towards the centre of the circle, it is usually convenient to break up the acceleration vector into two components: a_R, a component that is radial (towards the centre of the circle), and a_T, a component that is tangent to the circle (and perpendicular to to the radial component). The **radial component is "responsible" for the change in direction of the velocity** such that the object goes in a circle. the

magnitude of the radial acceleration is the same as it is for uniform circular motion:

$$a_R = \frac{v^2}{r}$$

where the speed is no longer constant in time. The tangential component of the acceleration is responsible for changing the magnitude of the velocity of the object:

$$a_T = \frac{dv}{dt}$$

Example 1-7

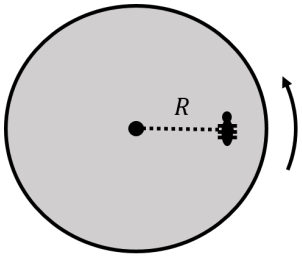


Figure 1.22: An ant on a horizontal turntable that is starting to spin, as seen from above.

A small ant is sleeping on a turntable just as the turntable starts to spin from rest, with an angular acceleration $\alpha = 1\,\text{rad/s}^2$ that is small enough so that, initially, the ant remains on the turntable. The ant is a distance $R = 0.1\,\text{m}$ from the centre of the turntable, as shown in Figure 1.22 and the coefficient of static friction between the ant's "feet" and the turntable is $\mu_s = 0.5$. After how much time will the ant slide off from the turntable?

Solution

As the turntable accelerates, the force of static friction between the turntable and the ant will keep the ant moving with the turntable. Once the turntable is going fast enough, the force of friction will no longer be large enough to provide the total acceleration that is required to keep the ant moving with the turntable (with a constant tangential component of the acceleration and an increasing radial component of the acceleration).

The forces on the ant are:

1. $\vec{F}_g$, its weight, with magnitude mg.
2. $\vec{N}$, a normal force exerted by the turntable on the ant.
3. $\vec{f}_s$, a force of static friction exerted by the turntable on the ant. The force of friction will be such that it has both radial and tangential components.

A free-body diagram for the forces on the ant is shown in Figure 1.23, as seen from above and from the side, for some point in time. We have chosen the point in time to be just when the ant is about to slide off of the turntable, when the force of static friction makes an unknown angle θ with the x axis. We have placed the origin of the coordinate system at the centre of the turntable and chosen the x axis such that the ant is located on the positive x axis with its velocity in the positive y direction. We used a three dimensional coordinate system where the weight and normal force are exerted in the z (vertical) direction since the acceleration vector of the ant will have both radial (x) and tangential (y) components.

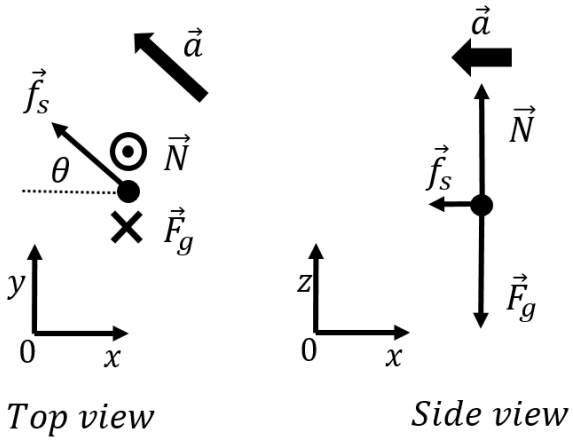


Figure 1.23: (Left:) Forces on the ant as seen from above. The normal force is out of the page ($\odot$), whereas the weight is into the page ($\times$). (Right:) Forces on the ant as seen from the side. Note that the acceleration vector and force of static friction also have components in the y direction, which is why their magnitude is shown as being smaller than in the top view.

Newton's Second Law has to be written out in three components. The z component relates the weight and normal force:

$$\sum F_z = N - F_g = 0$$
$$\therefore N = mg$$

The x component of Newton's Second Law is such that the x component of the acceleration is its radial component:

$$\sum F_x = -f_s \cos\theta = -ma_R = -m\frac{v^2}{R}$$
$$\therefore f_s \cos\theta = m\frac{v^2}{R}$$

The y component of Newton's Second relates the tangential component of the force of

static friction to the tangential component of the acceleration:

$$\begin{aligned} \sum F_y &= f_s \sin\theta = ma_T \\ \therefore f_s \sin\theta &= m\alpha R \end{aligned}$$

where we used the fact that the (linear) tangential acceleration, a_T, is related to the angular acceleration, α, by:

$$a_T = \alpha R$$

Summarizing the three equations that we obtained from the three components of Newton's Second Law:

$$\begin{aligned} f_s \cos\theta &= m\frac{v^2}{R} \\ f_s \sin\theta &= m\alpha R \\ N &= mg \end{aligned}$$

Also, note that the speed, $v(t)$ at some time t is given by simple kinematics:

$$v(t) = v_0 + a_T t = (0) + \alpha R t$$

The ant will start to slip when the force of friction reaches its maximal amplitude, $f_s = \mu_s N = \mu_S mg$. The x of Newton's Second Law can be used to find an expression for the time at which force of friction reaches its maximal value (in terms of the unknown angle θ):

$$\begin{aligned} f_s \cos\theta &= m\frac{v^2}{R} \\ \mu_s g \cos\theta &= R\alpha^2 t^2 \\ \therefore t &= \sqrt{\frac{\mu_s g \cos\theta}{R\alpha^2}} \end{aligned}$$

We can use the y component to determine the angle θ:

$$\begin{aligned} f_s \sin\theta &= m\alpha R \\ \mu_s g \sin\theta &= \alpha R \\ \therefore \sin\theta &= \frac{\alpha R}{\mu_s g} \\ \therefore \theta &= \sin^{-1}\left(\frac{\alpha R}{\mu_s g}\right) = \sin^{-1}\left(\frac{(1\,\mathrm{rad/s^2})(0.1\,\mathrm{m})}{(0.5)(9.8\,\mathrm{N/kg})}\right) \\ &= 1.17^\circ \end{aligned}$$

The angle is very small, and we see that the force of friction is mostly directed towards the centre of the circle. The radial acceleration is thus much larger than the tangential acceleration. We can then use the angle to find the time using the expression we derived above:

$$\begin{aligned} t &= \sqrt{\frac{\mu_s g \cos\theta}{R\alpha^2}} = \sqrt{\frac{(0.5)(9.8\,\mathrm{N/kg})\cos(1.17^\circ)}{(0.1\,\mathrm{m})(1\,\mathrm{rad/s^2})^2}} \\ &= 7.0\,\mathrm{s} \end{aligned}$$

1.5 Summary

Key Takeaways

When the velocity of an object does not change direction continuously ("linear motion"), we can model its motion independently over several segments in such a way that the motion is one dimensional in each segment. This allows us to choose a coordinate system in each segment where the acceleration vector is co-linear with one of the axes.

When the forces on an object changes continuously, we need to use calculus to determine the motion of the object. If the velocity vector for an object changes direction continuously, we need to model the motion in each dimension independently.

If an object undergoes uniform circular motion, the acceleration vector and the sum of the forces always point towards the centre of the circle. In the radial direction, Newton's Second Law gives

$$\sum \vec{F} = ma_R = m\frac{v^2}{R}$$

If an object's speed is changing as it moves around a circle the acceleration vector will have a component that is towards the centre of the circle (the radial component) and a component that is tangential to the circle. The tangential component is responsible for the change in speed, whereas the radial component is responsible for the change in direction of the velocity.

In a reference frame that is rotating about a circle, an inertial force, sometimes called the centrifugal force, appears to push all objects co-moving with the reference frame towards the outside of the circle.

1.8 Thinking about the material

Reflect and research

1. Is there a maximum speed with which an object can spin? (Something about the thing eventually flying apart if it rotates too fast, as the atoms can not be held together at some point - maybe there is a cool video to look up?)

To try at home

1. Spin a mass on a string in a vertical circle, what is the tension in the string when the mass is at the top for it to barely make it around?
2. Spin a mass on a string in a vertical circle, how does the minimum speed at the top of the circle to barely make it around depend on the radius of the circle or the mass?
3. Spin a mass on a string in a vertical circle, describe the motion if the mass does not have the minimum speed to make it around the circle. If it makes it to the top, does it automatically make it all the way around the circle?

To try in the lab

1. Build a conical pendulum and determine whether the opening angle of the cone is related to the speed of the bob, in the way that you expect it to be.

1.6.1 Problems and Solutions

Problem 1-1: Consider a conical pendulum with a mass m, attached to a string of length L. The mass executes uniform circular motion in the horizontal plane, about a circle of radius R, as shown in Figure 1.24. One can think of the horizontal circle and the point where the string is attached to as forming a cone. The circular motion is such that the (constant) angle between the string and the vertical is θ. (Solution)

a) Derive an expression for the tension in the string.
b) Derive an expression for the speed of the mass.
c) Derive an expression for the period of the motion.

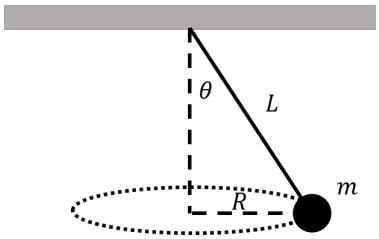


Figure 1.24: The conical pendulum.

Problem 1-2: Barb and Kenny are going to the amusement park. Barb insists on riding the giant roller coaster, but Kenny is scared that they will fall out of the roller coaster at the top of the loop. Barb reassures Kenny by asking the roller coaster technician for more information. The technician says that they will be travelling at 15 m/s when upside down, and that the roller coaster loop has a radius of 22 m. Kenny is still sceptical. Is he correct in being sceptical? (Solution)

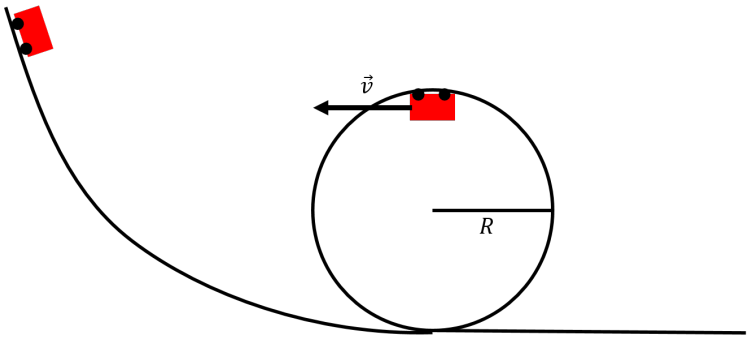


Figure 1.25: The roller coaster

1.6.2 Solutions

Solution to problem 1-1:

a) We start by identifying the forces that are acting on the mass. These are:
- $\vec{F}_g$, its weight, with a magnitude mg.
- $\vec{F}_T$, a force of tension exerted by the string.

The forces are illustrated in Figure 1.26, along with our choice of coordinate system and the direction of the acceleration of the mass (towards the centre of the circle).

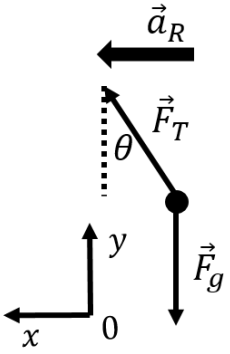


Figure 1.26: Forces acting on the conical pendulum

The y component of Newton's Second law gives the relation between the tension in the string, the weight, and the angle θ

$$\begin{aligned} \sum F_y &= 0 \\ F_T \cos\theta - F_g &= 0 \\ F_T \cos\theta &= mg \\ \therefore F_T &= \frac{mg}{\cos\theta} \end{aligned}$$

b) In order for the mass to move in a circle, the net force must be directed towards the centre of the circle at all times. The x component of Newton's Second Law, combined with our expression for the magnitude of the tension, F_T, allows us to determine the speed of the mass:

$$\begin{aligned} \sum F_x &= ma_r \\ F_T \sin\theta &= m\frac{v^2}{R} \\ \left(\frac{mg}{\cos\theta}\right)\sin\theta &= m\frac{v^2}{R} \\ g\tan\theta &= \frac{v^2}{R} \\ \therefore v &= \sqrt{gR\tan\theta} \end{aligned}$$

c) Now that we know the speed, we can easily find the period, T, of the motion:

$$\begin{aligned} T &= \frac{2\pi R}{v} \\ &= \frac{2\pi R}{\sqrt{gR\tan\theta}} = 2\pi\sqrt{\frac{R}{g\tan\theta}} \end{aligned}$$

Solution to problem 1-2: We need to determine if the speed of Barb and Kenny is large enough for them to go around the circle. The minimum speed that they must have at the top of the loop is such that their weight (the only force acting on them) provides the centripetal (net) force required to go around the loop.

Writing Newton's Second Law in the vertical direction, for the case where only the weight acts on Barb or Kenny (mass m), when they are going at speed v

$$\begin{aligned} mg = ma_R &= m\frac{v^2}{R} \\ \therefore v = \sqrt{gR} &= \sqrt{(9.8\,\mathrm{m/s^2})(22\,\mathrm{m})} = 14.68\,\mathrm{m/s} \end{aligned}$$

This corresponds to the minimum speed that they must have at the top of the loop to make it around. If they go faster, the normal force from their seat (downwards, since they are upside-down), would result in a larger net force towards the centre of the circle. This situation corresponds to the normal force from their seat just barely reaching 0 at the top of the loop. Since the roller coaster is quoted as having a speed of 15 m/s at the top of the loop, they will just barely make it. However, this is way too close to the minimal speed to not fall out of the roller coaster, so Kenny is correct in being sceptical! The engineers designing the roller coaster should include a much bigger safety margin!

2 Gauss' Law

In this chapter, we take a detailed look at Gauss' Law applied in the context of the electric field. We have already encountered Gauss' Law briefly in Section **??** when we examined the gravitational field. Since the electric force is mathematically identical to the gravitational force, we can apply the same tools, including Gauss' Law, to model the electric field as we do the gravitational field. Many of the results from this chapter are thus equally applicable to the gravitational force.

Learning Objectives

- Understand the concept of flux for a vector field.
- Understand how to calculate the flux of a vector field through an open and a closed surface.
- Understand how to apply Gauss' Law quantitatively to determine an electric field.
- Understand how to apply Gauss' Law qualitatively to discuss charges on a conductor.

Think About It

A neutral spherical conducting shell encloses a point charge, Q, located at the centre of the shell. Due to separation of charge, the outer surface of the shell will acquire a net positive charge. What is the magnitude of that charge?

A) less than Q.
B) exactly Q.
C) more than Q.

2.1 Flux of the electric field

Gauss' Law makes use of the concept of "flux". Flux is always defined based on:

- A surface.
- A vector field (e.g. the electric field).

and can be thought of as a measure of the number of field lines from the vector field that

cross the given surface. For that reason, one usually refers to the "flux of the electric field through a surface". This is illustrated in Figure 2.1 for a uniform horizontal electric field, and a flat surface, whose normal vector, $\vec{A}$, is shown. If the surface is perpendicular to the field (left panel), and the field vector is thus parallel to the vector, $\vec{A}$, then the flux through that surface is maximal. If the surface is parallel to the field (right panel), then no field lines cross that surface, and the flux through that surface is zero. If the surface is rotated with respect to the electric field, as in the middle panel, then the flux through the surface is between zero and the maximal value.

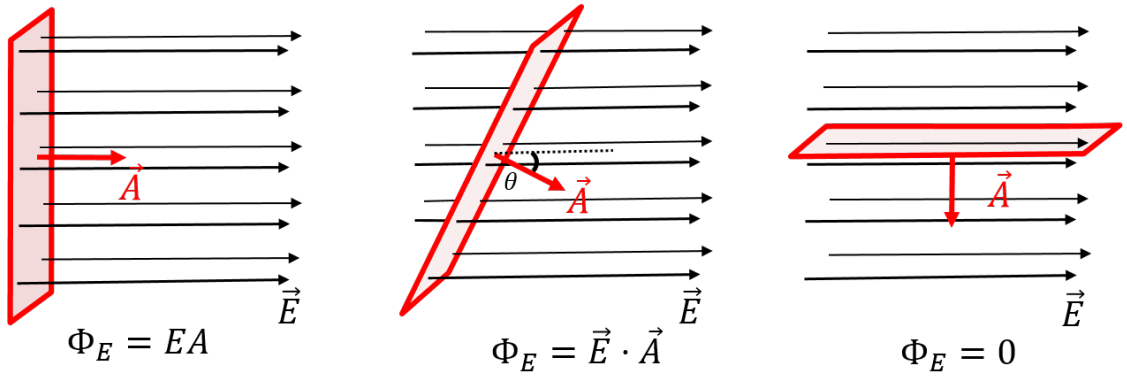


Figure 2.1: Flux of an electric field through a surface that makes different angles with respect to the electric field. In the leftmost panel, the surface is oriented such that the flux through it is maximal. In the rightmost panel, there are no field lines crossing the surface, so the flux through the surface is zero.

We define a vector, $\vec{A}$, associated with the surface such that the magnitude of $\vec{A}$ is equal to the area of the surface, and the direction of $\vec{A}$ is such that it is perpendicular to the surface, as illustrated in Figure 2.1. We define the flux, Φ_E, of the electric field, $\vec{E}$, through the surface represented by vector, $\vec{A}$, as:

$$\Phi_E = \vec{E} \cdot \vec{A} = EA\cos\theta$$

since this will have the same properties that we described above (e.g. no flux when $\vec{E}$ and $\vec{A}$ are perpendicular, flux proportional to number of field lines crossing the surface). Note that the flux is only defined up to an overall sign, as there are two possible choices for the direction of the vector $\vec{A}$, since it is only required to be perpendicular to the surface. By convention, we usually choose $\vec{A}$ so that the flux is positive.

Checkpoint 2-1

What are the units of electric flux?

A) Nm/C
B) Vm
C) V/m
D) The units of flux depend on the dimensions of the charged object

Example 2-1

A uniform electric field is given by: $\vec{E} = E\cos\theta\hat{x} + E\sin\theta\hat{y}$ throughout space. A rectangular surface is defined by the four points $(0,0,0)$, $(0,0,H)$, $(L,0,0)$, $(L,0,H)$. What is the flux of the electric field through the surface?

Solution

The surface that is defined corresponds to a rectangle in the xz plane with area $A = LH$. Since the rectangle lies in the xz plane, a vector perpendicular to the surface will be along the y direction. We choose the positive y direction, since this will give a positive number for the flux (as the electric field has a positive component in the y direction). The vector $\vec{A}$ is given by:

$$\vec{A} = A\hat{y} = LH\hat{y}$$

The flux through the surface is thus given by:

$$\begin{aligned}\Phi_E &= \vec{E}\cdot\vec{A} = (E\cos\theta\hat{x} + E\sin\theta\hat{y})\cdot(LH\hat{y}) \\ &= ELH\sin\theta\end{aligned}$$

where one should note that the angle θ, in this case, is not the angle between $\vec{E}$ and $\vec{A}$, but rather the complement of that angle.

Discussion: In this example, we calculated the flux of a uniform electric field through a rectangle of area, $A = LH$. Since we knew the components of both the electric field vector, $\vec{E}$, and the surface vector, $\vec{A}$, we used their scalar product to determine the flux through the surface. In some cases, it is easier to work with the magnitude of the vectors and the angle between them to determine the scalar product (although note that in this example, the angle between $\vec{E}$ and $\vec{A}$ is $90° - \theta$).

2.1.1 Non-uniform fields

So far, we have considered the flux of a uniform electric field, $\vec{E}$, through a surface, S, described by a vector, $\vec{A}$. In this case, the flux, Φ_E, is given by:

$$\Phi_E = \vec{E}\cdot\vec{A}$$

However, if the electric field is not constant in magnitude and/or in direction over the entire surface, then we divide the surface, S, into many infinitesimal surfaces, dS, and sum together (integrate) the fluxes from those infinitesimal surfaces:

$$\boxed{\Phi_E = \int \vec{E}\cdot d\vec{A}}$$

where, $d\vec{A}$, is the normal vector for the infinitesimal surface, dS. This is illustrated in Figure 2.2, which shows, in the left panel, a surface for which the electric field changes magnitude

along the surface (as the field lines are closer in the lower left part of the surface), and, in the right panel, a scenario in which the direction (and magnitude) of the electric field vary along the surface.

In order to calculate the flux through the total surface, we first calculate the flux through an infinitesimal surface, dS, over which we assume that $\vec{E}$ is constant in magnitude and direction, and then, we sum (integrate) the fluxes from all of the infinitesimal surfaces together. Remember, the flux through a surface is related to the number of field lines that cross that surface; it thus makes sense to count the lines crossing an infinitesimal surface, dS, and then adding those together over all the infinitesimals surfaces to determine the flux through the total surface, S.

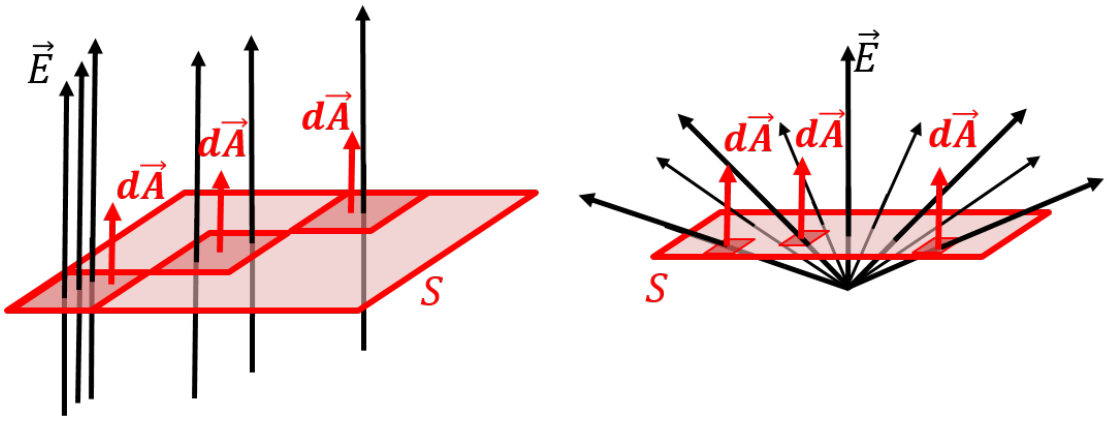


Figure 2.2: Examples of surfaces that need to be sub-divided in order to determine the net flux through them. The surface on the left must be subdivided because the electric field changes magnitude over the surface, whereas the one on the right needs to be subdivided because the angle between $\vec{E}$ and $d\vec{A}$ is not constant (and the magnitude of $\vec{E}$ also changes along the surface).

Example 2-2

An electric field points in the z direction everywhere in space. The magnitude of the electric field depends linearly on the x position in space, so that the electric field vector is given by: $\vec{E} = (a - bx)\hat{z}$, where, a, and, b, are constants. What is the flux of the electric field through a square of side, L, that is located in the xy plane?

Solution

We need to calculate the flux of the electric field through a square of side L in the xy plane. The electric field is always in the z direction, so the angle between $\vec{E}$ and $d\vec{A}$ (the normal vector for any infinitesimal area element) will remain constant.

We can calculate the flux through the square by dividing up the square into thin strips of length L in the y direction and infinitesimal width dx in the x direction, as illustrated in

Figure 2.3. In this case, because the electric field does not change with y, the dimension of the infinitesimal area element in the y direction is finite (L). If the electric field varied both as a function of x and y, we would start with area elements that have infinitesimal dimensions in both the x and the y directions.

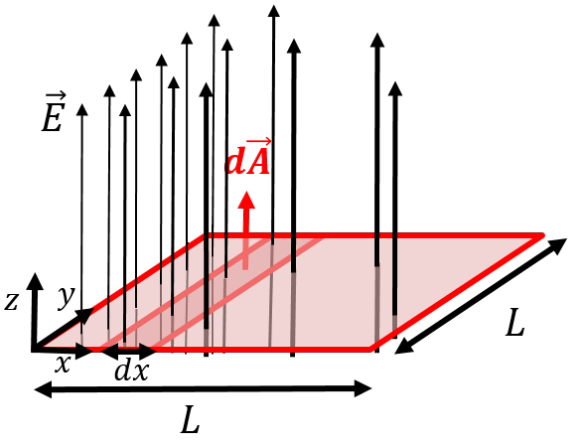


Figure 2.3: Dividing a square in the xy plane into thin strips of length L and width dx.

As illustrated in Figure 2.3, we first calculate the flux through a thin strip of area, $dA = Ldx$, located at position x along the x axis. Choosing, $d\vec{A}$, in the direction to give a positive flux, the flux through the strip that is illustrated is given by:

$$d\Phi_E = \vec{E} \cdot d\vec{A} = EdA = (ax - b)Ldx$$

where $\vec{E} \cdot d\vec{A} = EdA$, since the angle between $\vec{E}$ and $\vec{A}$ is zero. Summing together the fluxes from the strips, from $x = 0$ to $x = L$, the total flux is given by:

$$\Phi_E = \int d\Phi_E = \int_0^L (ax - b)Ldx = \frac{1}{2}aL^3 - bL^2$$

Discussion: In this example, we showed how to calculate the flux from an electric field that changes magnitude with position. We modelled a square of side, L, as being made of many thin strips of length, L, and width, dx. We then calculated the flux through each strip and added those together to obtain the total flux through the square.

2.1.2 Closed surfaces

One can distinguish between a "closed" surface and an "open" surface. A surface is closed if it completely defines a volume that could, for example, be filled with a liquid. A closed surface has a clear "inside" and an "outside". For example, the surface of a sphere, of a cube, or of a cylinder are all examples of closed surfaces. A plane, a triangle, and a disk are, on the other hand, examples of "open surfaces".

For a closed surface, one can unambiguously define the direction of the vector $\vec{A}$ (or $d\vec{A}$) as the direction that it is perpendicular to the surface and **points towards the outside**. Thus, the sign of the flux out of a closed surface is meaningful. The flux will be positive if there is a net number of field lines exiting the surface (since $\vec{E}$ and $\vec{A}$ will be parallel on

average) and the flux will be negative if there is a net number of field lines entering the surface (as $\vec{E}$ and $\vec{A}$ will be anti-parallel on average). The flux through a closed surface is thus zero if the number of field lines that enter the surface is the same as the number of field lines that exit the surface.

When calculating the flux over a closed surface, we use a different integration symbol to show that the surface is closed:

$$\Phi_E = \oint \vec{E} \cdot d\vec{A}$$

which is the same integration symbol that we used for indicating a path integral when the initial and final points are the same (see for example Section **??**).

Example 2-3

A negative electric charge, $-Q$, is located at the origin of a coordinate system. Calculate the flux of the electric field through a spherical surface of radius, R, that is centred at the origin.

Solution

Figure 2.4 shows the spherical surface of radius, R, centred on the origin where the charge $-Q$ is located.

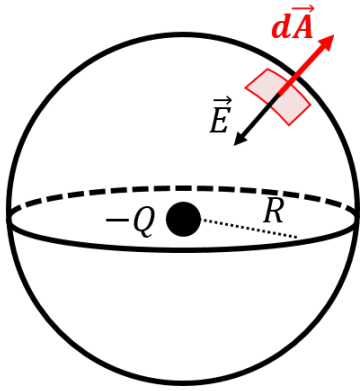


Figure 2.4: Calculating the flux through a spherical surface.

At all points along the surface, the electric field has the same magnitude:

$$E = \frac{1}{4\pi\epsilon_0}\frac{Q}{R^2}$$

as given by Coulomb's law for a point charge. Although the vector, $\vec{E}$, changes direction everywhere along the surface, it always makes the same angle ($-180°$) with the corresponding vector, $d\vec{A}$, at any particular location. Indeed, for a point charge, the

electric field points in the radial direction (inwards for a negative charge) and is thus perpendicular to the spherical surface at all points. Since the surface is closed, the vector, $d\vec{A}$, points outwards anywhere on the surface. Thus, at any point on the surface, we can evaluate the flux through an infinitesimal area element, $d\vec{A}$:

$$d\Phi_E = \vec{E} \cdot d\vec{A} = EdA\cos(-180^\circ) = -EdA$$

where the overall minus sign comes from the fact that, $\vec{E}$, and, $d\vec{A}$, are anti-parallel. The total flux through the spherical surface is obtained by summing together the fluxes through each area element:

$$\Phi_E = \oint d\Phi_E = \oint -EdA = -E\oint dA = -E(4\pi R^2)$$

where we factored, E, out of the integral, since the magnitude of the electric field is constant over the entire surface (a constant distance R from the charge). In the last equality, we recognized that, $\oint dA$, simply means "sum together all of the areas, dA, of the surface elements", which gives the total surface area of the sphere, $4\pi R^2$. The flux through the spherical surface is negative, because the charge is negative, and the field lines point towards $-Q$.

Using the value that we obtained for the magnitude of the electric field from Coulomb's Law, the total flux is given by:

$$\Phi_E = -E(4\pi R^2) = -\frac{1}{4\pi\epsilon_0}\frac{Q}{R^2}(4\pi R^2) = -\frac{Q}{\epsilon_0}$$

which, surprisingly, is independent of the radius of the spherical surface. Note that we used ϵ_0 instead of Coulomb's constant, k, since the result is cleaner without the extra factor of 4π.

Discussion: In this example, we calculated the flux of the electric field from a negative point charge through a spherical surface concentric with the charge. We found the flux to be negative, which makes sense, since the field lines go towards a negative charge, and there is thus a net number of field lines entering the spherical surface. Perhaps surprisingly, we found that the total flux through the surface does not depend on the radius of the surface! In fact, that statement is precisely Gauss' Law: the net flux out of a closed surface depends only on the amount of charge enclosed by that surface (and the constant, ϵ_0). Gauss' Law is of course more general, and applies to surfaces of any shape, as well as charges of any shape (whereas Coulomb's Law only holds for point charges).

2.2 Gauss' Law

Gauss' Law is a relation between the net flux through a closed surface and the amount of charge, Q^{enc}, in the volume enclosed by that surface:

$$\boxed{\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}}$$

In particular, note that Gauss' Law holds true for **any** closed surface, and the shape of that surface is not specified in Gauss' Law. That is, we **can always choose the surface to use** when calculating the flux. For obvious reasons, we often call the surface that we choose a "gaussian surface". But again, this surface is simply a mathematical tool, there is no actual property that makes a surface "gaussian"; it simply means that we chose that surface in order to apply Gauss' Law. In Example 2-3 above, we confirmed that Gauss' Law is compatible with Coulomb's Law for the case of a point charge and a spherical gaussian surface.

Physically, Gauss' Law is a statement that field lines must begin or end on a charge (electric field lines original from positive charges and terminate on negative charges). Recall, flux is a measure of the net number of lines coming out of a surface. If there is a net number of lines coming out of a closed surface (a positive flux), that surface must enclose a positive charge from where those field lines originate. Similarly, if there are the same number of field lines entering a closed surface as there are lines exiting that surface (a flux of zero), then the surface encloses no charge. Gauss' Law simply states that the number of field lines exiting a closed surface is proportional to the amount of charge enclosed by that surface.

Primarily, Gauss' Law is a useful tool to determine the magnitude of the electric field from a given charge, or charge distribution. We usually have to use symmetry to determine the direction of the electric field vector. In general, the integral for the flux is difficult to evaluate, and Gauss' Law can only be used analytically in cases with a high degree of symmetry. Specifically, the integral for the flux is easiest to evaluate if:

1. **The electric field makes a constant angle with the surface**. When this is the case, the scalar product can be written in terms of the cosine of the angle between $\vec{E}$ and $d\vec{A}$, which can be taken out of the integral if it is constant:
$$\oint \vec{E} \cdot d\vec{A} = \oint E \cos\theta dA = \cos\theta \oint E dA$$
Ideally, one has chosen a surface such that this angle is 0 or 180°.
2. **The electric field is constant in magnitude along the surface**. When this is the case, the integral can be simplified further by factoring out, E, and simply becomes an integral over dA (which corresponds to the total area of the surface, A):
$$\oint \vec{E} \cdot d\vec{A} = \cos\theta \oint E dA = E \cos\theta \oint dA = EA\cos\theta$$

Ultimately, the points above should dictate the choice of gaussian surface **so that** the integral for the flux is easy to evaluate. The choice of surface will depend on the symmetry of the problem. For a point (or spherical) charge, a spherical gaussian surface allows the flux to easily be calculated (Example 2-3). For a line of charge, as we will see, a cylindrical surface results is a good choice for the gaussian surface. Broadly, the steps for applying Gauss' Law to determine the electric field are as follows:

1. Make a diagram showing the charge distribution.

2. Use symmetry arguments to determine in which way the electric field vector points.
3. Choose a gaussian surface that goes through the point for which you want to know the electric field. Ideally, the surface is such that the electric field is constant in magnitude and always makes the same angle with the surface, so that the flux integral is straightforward to evaluate.
4. Calculate the flux, $\oint \vec{E} \cdot d\vec{A}$.
5. Calculate the amount of charge located within the volume enclosed by the surface, Q^{enc}.
6. Apply Gauss' Law, $\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$.

Example 2-4

An insulating sphere of radius, R, contains a total charge, Q, which is uniformly distributed through out its volume. Determine an expression for the electric field as a function of distance, r, from the centre of the sphere.

Solution

Note that this is identical, mathematically, as the derivation that is done in Section **??** for the case of gravity.

When applying Gauss' Law, we first need to think about symmetry in order to determine the direction of the electric field vector. We also need to think about all possible regions of space in which we need to determine the electric field. In particular, for this case, we need to determine the electric field both inside ($r \leq R$) and outside ($r \geq R$) of the charged sphere.

Figure 2.5 shows the charged sphere of radius R. If we consider the direction of the electric field outside the sphere (where $\vec{E}_{out}$ is drawn), we realize that it can only point in the radial direction (towards or away from the centre of the sphere), as this is the only choice that preserves the symmetry of the sphere. Being a sphere, the charge looks the same from all angles; thus, the electric field must also look the same from all angles, otherwise, there would be a preferred orientation for the sphere. The same argument holds for the electric field vector inside the sphere (drawn as $\vec{E}_{in}$).

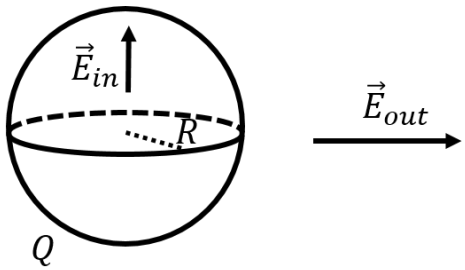


Figure 2.5: For a spherical charge distribution, the electric field inside and outside must point in the radial direction, by symmetry.

We now need to choose a gaussian surface that will make the flux integral easy to evaluate. Ideally, we can find a surface over which the electric field makes the same angle with the surface and over which the electric field is constant in magnitude. Again, based on the symmetry of the charge distribution, it is clear that a spherical surface of radius, r, will satisfy these properties.

We start by applying Gauss' Law outside the charge (with $r \geq R$) to determine the electric field, $\vec{E}_{out}$. Figure 2.6 shows our choice of spherical gaussian surface (labelled S) of radius, r, which is concentric with the spherical charge distribution of radius, R, and total charge, $+Q$.

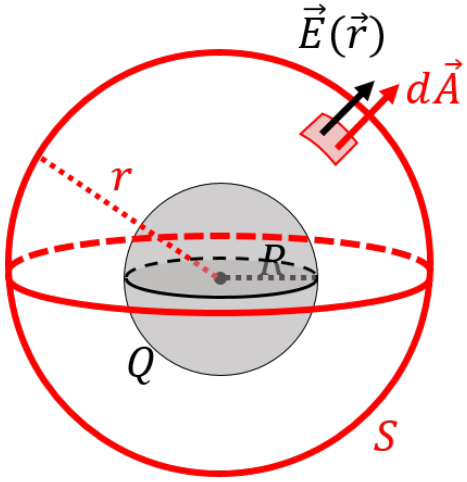


Figure 2.6: A spherical gaussian surface to determine the electric field outside a sphere of radius, R, holding charge, $+Q$.

In order to apply Gauss' Law, we need to calculate:

- the net flux through the surface.
- the charge in the volume enclosed by the surface.

The net flux through the surface is found identically as in Example 2-3, and is given by:

$$\Phi_E = \oint \vec{E} \cdot d\vec{A} = \oint E dA = E \oint dA = E(4\pi r^2)$$

where our choice of spherical surface led to $\vec{E} \cdot d\vec{A} = EdA$, since $\vec{E}$ and $d\vec{A}$ are always parallel. Furthermore, by symmetry, the electric field must be constant in magnitude along the whole surface, or the spherical symmetry would be broken. This allowed us

to factor the E out of the integral, leaving us with, $\oint dA$, which is simply the area of our gaussian spherical surface, $4\pi r^2$.

The gaussian surface with $r \geq R$ encloses the whole charged sphere, so the charge enclosed is simply the charge of the sphere, $Q^{inc} = Q$. Applying Gauss' Law allows us to determine the magnitude of the electric field:

$$\begin{aligned} \oint \vec{E} \cdot d\vec{A} &= \frac{Q^{enc}}{\epsilon_0} \\ E(4\pi r^2) &= \frac{Q^{enc}}{\epsilon_0} \\ \therefore E &= \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2} \end{aligned}$$

which is the same as the electric field a distance r from a point charge. Thus, from the outside, a spherical charge distribution leads to the same electric field as if the charge were concentrated at the centre of the sphere.

Next, we determine the magnitude of the electric field inside the charged sphere. In this case, we choose a spherical gaussian surface of radius $r \leq R$, that is concentric with the sphere, as illustrated by the surface labelled, S, that is shown in Figure 2.7.

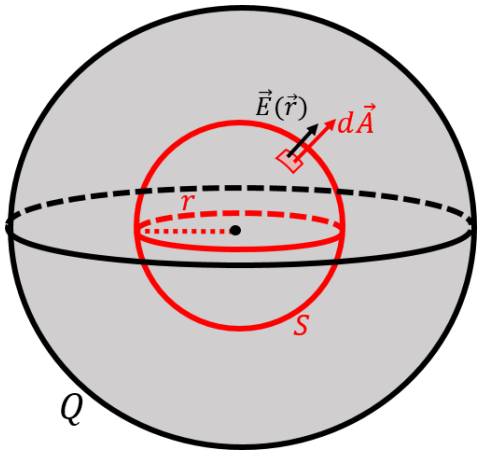


Figure 2.7: A spherical gaussian surface to determine the electric field outside a sphere of radius, R, holding charge, +Q.

The flux integral is trivial again, since the electric field always makes the same angle with the gaussian surface, and the magnitude of the electric field is constant in magnitude along the surface:

$$\Phi_E = \oint \vec{E} \cdot d\vec{A} = \oint E dA = E \oint dA = E(4\pi r^2)$$

In this case, however, the charge in the volume enclosed by the gaussian surface is less than Q, since the whole charge is not enclosed. We are told that the charge is distributed uniformly throughout the spherical volume of radius R. We can thus define a volume charge density, ρ, (charge per unit volume) for the sphere:

$$\rho = \frac{Q}{V} = \frac{Q}{\frac{4}{3}\pi R^3}$$

The volume enclosed by the gaussian surface is $\frac{4}{3}\pi r^3$, thus, the charge, Q^{enc}, contained in that volume is given by:

$$Q^{enc} = \frac{4}{3}\pi r^3 \rho = \frac{4}{3}\pi r^3 \frac{Q}{\frac{4}{3}\pi R^3} = Q\frac{r^3}{R^3}$$

Finally, we apply Gauss' Law to find the magnitude of the electric field inside the sphere:

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$$
$$E(4\pi r^2) = \frac{Q^{enc}}{\epsilon_0} = \frac{Q}{\epsilon_0}\frac{r^3}{R^3}$$
$$\therefore E = \frac{Q}{4\pi\epsilon_0 R^3} r$$

Note that the electric field increases linearly with radius inside of the charge sphere, and then decreases with radius squared outside of the sphere. Also, note that at the centre of the sphere, the electric field has a magnitude of zero, as expected from symmetry.

Discussion: In this example, we showed how to use Gauss' Law to determine the electric field inside and outside of a uniformly charged sphere. We recognized the spherical symmetry of the charge distribution and chose to use a spherical surface in order to apply Gauss' Law. This, in turn, allowed the flux to be easily calculated. We found that outside the sphere, the electric field decreases in magnitude with radius squared, just as if the entire charge were concentrated at the centre of the sphere. Inside the sphere, we found that the electric field is zero at the centre, and increases linearly with radius.

Example 2-5

An infinitely long straight wire carries a uniform charge per unit length, λ. What is the electric field at a distance, R, from the wire?

Solution

We start by making a diagram of the charge distribution, as in Figure 2.8, so that we

can use symmetry arguments to determine the direction of the electric field vector. At any point in space, the electric field vector must be radial (point to/from the centre of the wire) and in the plane perpendicular to the wire. If this were not the case, one would be able to look at the electric field to determine a preferred direction (either around the wire, if the field were not radial, or upwards/downwards, if the field were not perpendicular to the wire).

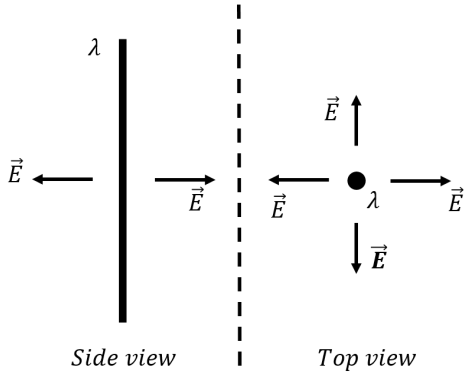


Figure 2.8: An infinite line of charge carrying uniform charge per unit length, λ. The left panel shows a side view and the right panel a view from above. The electric field must be in the radial direction or there would be a preferred direction.

Next, we need to choose a gaussian surface in order to apply Gauss' Law. A convenient choice is a cylinder (a "pill box") of radius, R, and length, L, as shown in Figure 2.9, as this goes through a point that is a distance, R, from the wire (where we are asked for the electric field). At all points on the cylindrical surface, the electric field vector is either perpendicular or parallel to the surface.

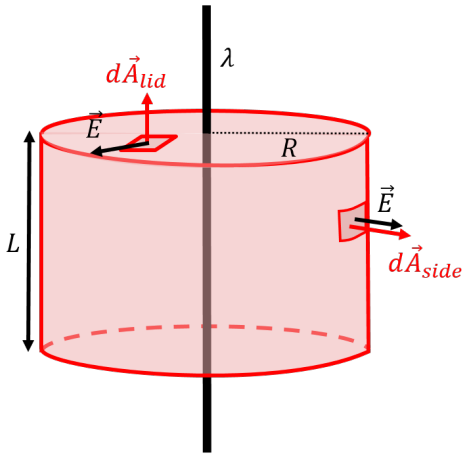


Figure 2.9: A cylindrical gaussian surface is used to calculate the flux from an infinite line of charge.

We can think of the cylindrical surface as being composed of three surfaces: 2 disks

on either end (the lids of the pill box), and the curved surface that makes up the side of the cylinder. The flux through the entire cylindrical surface will be the sum of the fluxes through the two lids plus the flux through the side:

$$\oint \vec{E} \cdot d\vec{A} = \int_{side} \vec{E} \cdot d\vec{A} + \int_{lid} \vec{E} \cdot d\vec{A} + \int_{lid} \vec{E} \cdot d\vec{A}$$

where you should note that the closed integral ($\oint$) was separated into three normal integrals ($\int$) corresponding to the three "open" surfaces that make up the closed surface. Again, remember that the flux is proportional to the net number of field lines exiting/entering the closed surface, so it make sense to count those lines over the three open surfaces and add them together to get the total number for the closed surface.

The flux through the lids is identically zero, since the electric field is perpendicular to $d\vec{A}$ everywhere on the lids. The total flux is thus equal to the flux through the curved side surface, for which the electric field vector is always parallel to $d\vec{A}$, and for which the electric field vector is constant in magnitude:

$$\oint \vec{E} \cdot d\vec{A} = \int_{side} \vec{E} \cdot d\vec{A} = \int_{side} E dA = E \int_{side} dA = E(2\pi RL)$$

where we recognized that the side surface can be unfolded into a rectangle of height, L, and width, $2\pi R$, corresponding to the circumference of the cylinder, so that the area is given by $A = 2\pi RL$.

Next, we determine the charge inside the volume enclosed by the surface. Since the cylinder encloses a length, L, of wire, the enclosed charge is given by:

$$Q^{enc} = \lambda L$$

where λ is the charge per unit length on the wire. Putting this altogether into Gauss' Law gives us the electric field at a distance, R, from the wire:

$$\begin{aligned} \oint \vec{E} \cdot d\vec{A} &= \frac{Q^{enc}}{\epsilon_0} \\ E(2\pi RL) &= \frac{\lambda L}{\epsilon_0} \\ \therefore E &= \frac{\lambda}{2\pi\epsilon_0 R} \end{aligned}$$

Note that this is the same result that we obtained in Example **??** when we took the limit of the finite line of charge having infinite length.

Discussion: In this example, we applied Gauss' Law to determine the electric field at a distance from an infinitely long charged wire. We used symmetry to argue that the field should be radial and in the plane perpendicular to the wire, and recognized that

a cylindrical gaussian surface would exploit the symmetry so that the flux can easily be calculated. We obtained the same result as we did from integrating Coulomb's Law in Example **??**. However, using Gauss' Law was much less work than integrating Coulomb's Law.

Example 2-6

Determine the electric field above an infinitely large plane of charge with uniform surface charge per unit area, σ.

Solution

Figure 2.10 shows a portion of the infinite plane. The electric field vector must be perpendicular to the plane or a preferred direction could otherwise be inferred from the direction of the electric field. We can also argue that the horizontal components of the electric field will cancel everywhere above the plane, since the plane is infinite. The electric field will point away from (towards) the plane, if the charge is positive (negative).

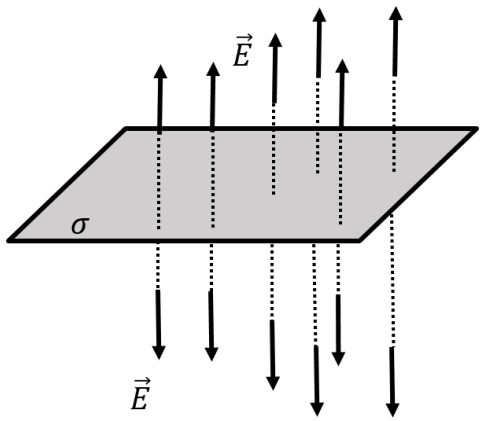


Figure 2.10: The electric field above an infinite plane with uniform charge per unit area, σ, must be perpendicular to the plane.

A cylindrical or box-shaped gaussian surface would both lead to the flux integral being easy to calculate, as illustrated in Figure 2.11. Indeed, since the electric field is perpendicular to the plane, only the parts of the surface that are parallel to the plane (the lids on the cylinder, the two horizontal planes in the box) will have a net flux through them.

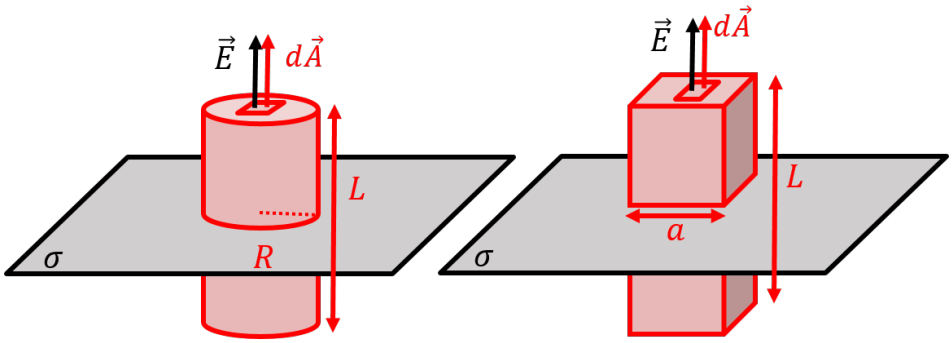


Figure 2.11: A cylindrical surface or a box are both good choices for a gaussian surface above a plane, since only the parts of the surface parallel to the plane will have net flux through them.

Let us choose a box (right panel of Figure 2.11) of length, L, with a square cross-section of side, a. We place the box such that the plane intersects the centre of the box (although this is not required, since we already know that the electric field will not depend on distance from the plane). The flux through the box is simply the flux through the two horizontal planes (of area a^2):

$$\oint \vec{E} \cdot d\vec{A} = \int_{top} EdA + \int_{bottom} EdA = 2Ea^2$$

The box encloses a section of the plane with area a^2, so that the net charge enclosed by the surface is:

$$Q^{enc} = \sigma a^2$$

Applying Gauss' Law allows us to determine the magnitude of the electric field:

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$$
$$2Ea^2 = \frac{\sigma a^2}{\epsilon_0}$$
$$\therefore E = \frac{\sigma}{2\epsilon_0}$$

which is the same result that we found in Example **??**.

Discussion: In this example, we used Gauss' Law to determine the electric field above an infinite plane. We found that we had a choice of gaussian surfaces (cylinder, box) that allowed us to apply Gauss' Law. We found the same result that we had found in Example **??** where we had integrated Coulomb's Law (twice, once for a ring of charge, then for a disk, then took the limit of the disk radius going to infinity). Again, we see that in configurations with a high degree symmetry, Gauss' Law can be very straightforward to apply.

2.3 Charges in a conductor

We can use Gauss' Law to understand how charges arrange themselves on a conductor. Consider (again) an infinite plane that carries a total charge per unit area, σ, similar to what we considered in Example 2-6. In this case, we explicitly consider the plane to be a conductor and to have a finite thickness. If we zoom into the plane, we will see that the charges will migrate to the surface of the plane, as illustrated in Figure 2.12, where the plane is seen edge on. Thus, the **charge density at the surface is half of the total charge density** of the plane.

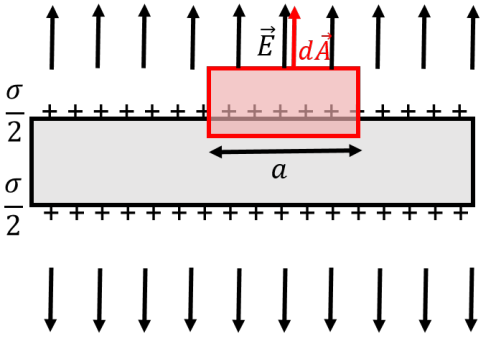


Figure 2.12: Cross-section of a conducting plane where the charges migrate to the surface. A box-shaped gaussian surface is also shown as seen from the side (the third dimension of the box is perpendicular to the plane of the page).

To determine the electric field near the plane, we choose a gaussian surface that is a box (as in Example 2-6), but require the lower end of the box to go through the plane, as illustrated in Figure 2-6. With this choice of gaussian surface, only the top surface (area a^2) will have flux through it, since the **electric field inside a conductor must be zero**[1]. The total flux is given by:

$$\oint \vec{E} \cdot d\vec{A} = \int_{top} E dA = Ea^2$$

The charge enclosed is given by:

$$Q^{enc} = \frac{\sigma}{2} a^2$$

where we used the fact that only half of the charges are inside the volume enclosed by our gaussian surface, so that the charge per unit area is half ($\frac{\sigma}{2}$) of that for the entire plane. Applying Gauss' Law, we find that the electric field is given by:

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$$

$$Ea^2 = \frac{\sigma a^2}{2\epsilon_0}$$

$$\therefore E = \frac{\sigma}{2\epsilon_0} \quad \text{(Field above an infinite plane)}$$

[1] Since charges can freely move in a conductor, they will move until there is no reason to move. Eventually, the charges accumulate in such a way that the net field in the conductor is zero. For a plane, this means that half of the charges will move to each side, as illustrated.

as before, but the factor of 2 now came from the charge density, rather than from the fact that two of the faces of the box had non-zero flux (as was the case in Example 2-6). We can generalize this result to determine the electric field near the surface of any conductor. Very close to the surface of any object, one can consider the surface as being similar to an infinite plane. If that surface carries charge per unit area, σ, then the electric field just above the surface is given by:

$$E = \frac{\sigma}{\epsilon_0} \quad \text{(Field near a conducting surface)}$$

In this case, there is no factor of 2, because the charge density in this equation is the charge density at the surface of the conductor. In the previous equation, the charge density on the surface of the conducing plane was $\frac{\sigma}{2}$.

Consider, now, a neutral spherical conducting shell, as shown from the side in the left panel of Figure 2.13. When a charge, $+Q$, is placed at the centre of the shell (right panel), charges inside the shell will move until the field in the shell is identically zero. The negative charges will move towards the inner surface (as they are attracted to $+Q$) and positive charges will be repelled onto the outer surface, under the influence of the electric field created by $+Q$ (shown in the diagram as $\vec{E}_Q$). Eventually, the separation of charges will lead to an electric field (shown in the diagram as $\vec{E}_\sigma$) in the opposite direction. The charges will stop moving once the total electric field in the conductor is zero (when the two fields cancel exactly everywhere in the conductor).

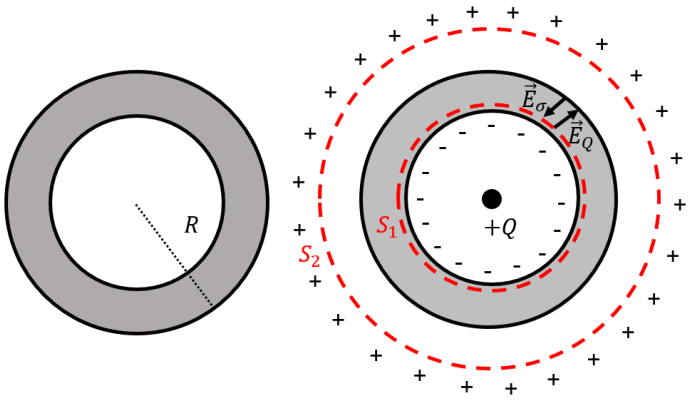


Figure 2.13: Left: a neutral conducting spherical shell (seen edge on). Right: A positive charge, $+Q$, placed at the centre of the shell. Charges in the shell will separate in order to keep the electric field inside the conductor zero.

We can use Gauss' Law to determine the amount of charge that has accumulated on the inner surface. Consider the gaussian spherical surface, S_1, in Figure 2.13, that is concentric with the shell and has a radius such that the surface is just inside the shell. Since the electric field is zero inside the shell, the flux out of the gaussian surface must be zero. By Gauss' Law, the amount of charge enclosed by the surface must also be zero. Thus, a total charge, $-Q$, will have accumulated on the inner surface of the conductor (since $Q^{enc} = -Q + Q = 0$). Because one cannot just create charge from nothing, there must be

an equal amount of opposite charge, $+Q$, on the outer surface of the shell. This is true of any conducting material with a cavity inside of it: if you place a charge $+Q$ in the cavity, a charge, $-Q$ will accumulated on the inner surface and a charge, $+Q$, will accumulate on the outer surface.

If we now consider the flux out of the surface, S_2, outside of the shell, the net charge enclosed will be $Q^{enc} = +Q - Q + Q = +Q$. The flux out of the spherical surface of radius, say, r, is then given by:

$$\oint \vec{E} \cdot d\vec{A} = E(4\pi r^2)$$

and the electric field, from Gauss' Law, is simply that of a point charge, $+Q$:

$$E = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}$$

and the shell has no effect on the field in regions where there is no conducting material from the shell. Right at the surface of the shell (outer radius, R), the surface charge density is given by:

$$\sigma = \frac{Q}{4\pi r^2}$$

Above, we found the electric field at the surface of a conductor that carries charge per unit area, σ, to be:

$$E = \frac{\sigma}{\epsilon_0}$$

which is clearly the same result that we obtained using the spherical surface, S_2:

$$E = \frac{\sigma}{\epsilon_0} = \frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}$$

Note that we found the electric field using Gauss' Law only in this last case, and found it to be equal to the electric field that one obtains from Coulomb's law. Thus, Gauss' Law only works if the field has an "inverse square law" dependence. If Gauss' Law does not provide the correct electric field, then the force does not depend on $1/r^2$. Gauss' Law can be used to make extremely stringent tests of whether the force goes as $1/r^2$ or deviates from this model.

2.4 Interpretation of Gauss' Law and vector calculus

In this section, we provide a little more theoretical background and intuition on Gauss' Law, as well as its connection to vector calculus. Very generally, Gauss' Law is a statement that connects a property of a vector field to the "source" of that field. We think of mass as the source for the gravitational field, and we think of charge as the source for the electric field. The property of the field that we considered in this case was its "flux out of a closed surface".

Recall that determining the flux of a field out of a closed surface is equivalent to counting the net number of field lines that exit that closed surface. Field lines must start on a positive charge and must end on a negative charge. Thus, if there is a net number of field lines exiting the surface, there must be a positive charge in the volume defined by the surface (a "source" of field lines). If there is a net number of field lines entering the surface, then the volume defined by the surface must enclose a negative charge (a "sink" of field lines). Gauss' Law is simply a statement that the number of field lines entering/exiting a closed surface is proportional to the amount of charge enclosed in that volume.

The flux out of a closed surface is tightly connected to the vector calculus concept of "divergence", which describes whether field lines are diverging (spreading out or getting closer together). When a point charge is present, field lines will emanate radially from that point charge; in other words, they will diverge. We say that the electric field has non-zero divergence if there is a source of the electric field in that position of space. The key difference between the concept of divergence and that of "flux out of a closed surface", is that divergence is a local property of the field (it is true at a point), whereas the flux out of a surface must be calculated using a finite volume and makes it challenging to define the field at a specific position. Gauss's Law defined using flux is thus not as useful for describing how the field changes at specific positions, and is usually limited to situations with a high degree of symmetry.

The divergence, $\nabla \cdot \vec{E}$, of a vector field, $\vec{E}$, at some position is defined as:

$$\nabla \cdot \vec{E} = \frac{\partial E}{\partial x} + \frac{\partial E}{\partial y} + \frac{\partial E}{\partial z}$$

and corresponds to the sum of three partial derivatives evaluated at that position in space. Gauss' Theorem (also called the Divergence Theorem) states that:

$$\int_V \nabla \cdot \vec{E} = \oint_S \vec{E} \cdot d\vec{A}$$

where the V (S) on the integral indicate whether the sum (integral) should be carried out over a volume, V, or over a closed surface, S, as we have practised in this chapter. While it is not important at this level to understand the theorem in detail, the point is that one can convert a "flux over a closed surface" into an integral of the divergence of the field. In other words, we can convert a global property (flux) to a local property (divergence). Gauss' Law in terms of divergence can be written as:

$$\boxed{\nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0}} \quad \text{(Local version of Gauss' Law)}$$

where ρ is the charge per unit volume at a specific position in space. This is the version of Gauss' Law that is usually seen in advanced textbooks and in Maxwell's unified theory of electromagnetism. This version of Gauss's Law relates a local property of the field (its divergence) to a local property of charge at that position in space (the charge per unit volume at that position in space). If we integrate both sides of the equation over volume,

we recover the original formulation of Gauss' Law: the left hand side, by the Divergence Theorem, leads to flux when integrated over volume, whereas on the right hand side, the integral over volume of charge per unit volume, ρ, will give the total charge enclosed in that volume, Q^{enc}:

$$\int_V \left(\nabla \cdot \vec{E}\right) dV = \int_V \left(\frac{\rho}{\epsilon_0}\right) dV$$
$$\oint_S \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$$

2 [illegible]

Key Takeaways

We can define the **flux** of a uniform and constant vector field, $\vec{E}$, through a flat surface, as:

$$\Phi_E = \vec{E} \cdot \vec{A} = EA\cos\theta$$

where, $\vec{A}$, is a vector that is perpendicular to the surface with a magnitude equal to the area of that surface, and, θ, is the angle between, $\vec{A}$ and $\vec{E}$. The flux of a field through a surface is proportional to the number of field lines that cross that surface. If the surface is parallel to the field ($\vec{A}$ and $\vec{E}$ are thus perpendicular), the flux through that surface is zero (no field lines cross the surface, the scalar product is zero).

If $\vec{E}$ and $\vec{A}$ change over the surface ($\vec{E}$ and/or $\vec{A}$ change magnitude and/or direction along the surface), then we treat the surface as being made of infinitesimal surface elements over which the two vectors are constant. We define a vector $d\vec{A}$ to be perpendicular to the surface element with an infinitesimal area, dA. The total flux is then obtained by summing the fluxes through each surface element:

$$\Phi_E = \int \vec{E} \cdot d\vec{A} = \int EdA\cos\theta$$

Note that the direction of the vector $d\vec{A}$ (or $\vec{A}$) is ambiguous, as one can choose either of two directions perpendicular to a surface. Usually, one chooses the direction of $\vec{A}$ so that the flux is positive (i.e. $\vec{A}$ has a component parallel to $\vec{E}$). However, if the surface is "closed" (that is, it defines a volume), then we always choose the direction of $d\vec{A}$ so that it points outwards from the surface (since the surface encloses a volume, one can define an "inside" and an "outside").

In the case of the electric field, Gauss' Law relates the flux of the electric field from a closed surface to the amount of charge, Q^{enc}, contained in the volume enclosed by that surface:

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$$

Physically, Gauss' Law is a statement that field lines must begin or end on a charge (electric field lines original from positive charges and terminate on negative charges). If there is a net number of lines coming out of a closed surface (a positive flux), that surface must enclose a positive charge from where those field lines originate. Similarly, if there are the same number of field lines entering a closed surface as there are lines exiting that surface (a flux of zero), then the surface encloses no charge. Gauss' Law states that the number of field lines exiting a closed surface is proportional to the amount of charge enclosed by that surface.

Gauss' Law is useful to determine the electric field. However, this can only be done analytically for charge distributions with a very high degree of symmetry. This is because the flux integral is not usually easy to evaluate unless:

1. **The electric field makes a constant angle with the surface**. When this is the case, the scalar product can be written in terms of the cosine of the angle between $\vec{E}$ and $d\vec{A}$, which can be taken out of the integral if it is constant:

$$\oint \vec{E} \cdot d\vec{A} = \oint E \cos\theta dA = \cos\theta \oint E dA$$

2. **The electric field is constant in magnitude along the surface**. When this is the case, the integral can be simplified further by factor out E, and simply becomes an integral over dA (which corresponds to the total area of the surface, A):

$$\oint \vec{E} \cdot d\vec{A} = \cos\theta \oint E dA = E \cos\theta \oint dA = EA \cos\theta$$

Note that Gauss' Law does not specify a closed surface over which to calculate the flux; it holds for any surface. We can thus choose a surface that will make the flux integral easy to evaluate - we call this choice a "gaussian surface" (not because it has some special property, but because we chose that surface to apply Gauss' Law). A procedure for applying Gauss' Law to determine the electric field at some point in space can be written as:

1. Make a diagram showing the charge distribution.
2. Use symmetry arguments to determine in which way the electric field vector points.
3. Choose a gaussian surface that goes through the point for which you want to know the electric field. Ideally, the surface is such that the electric field is constant in magnitude and always makes the same angle with the surface, so that the flux integral is straightforward to evaluate.
4. Calculate the flux, $\oint \vec{E} \cdot d\vec{A}$.
5. Calculate the amount of charge in the volume enclosed by the surface, Q^{enc}.
6. Apply Gauss' Law, $\oint \vec{E} \cdot d\vec{A} = \frac{Q^{enc}}{\epsilon_0}$.

We showed how Gauss' Law can be used to understand and quantify how charges arrange themselves on a conductor, in such a way that the electric field is zero everywhere in the conductor. Finally, we briefly introduced a more modern version of Gauss' Law that uses divergence instead of flux:

$$\nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0}$$

This last version has the advantage that it relates a local property of the field (divergence) to a local property of charge (charge density at some position in space).

Important Equations

Momentum of a point particle:

$$\vec{p} = m\vec{v}$$

$$\frac{d}{dt}\vec{p} = \sum \vec{F} = \vec{F}^{net}$$

Position of the Centre of Mass of a system:

$$\vec{r}_{CM} = \frac{1}{M}\sum_i m_i \vec{r}_i$$

2.8 Thinking about the material

Reflect and research

1. Explain

To try at home

1. Try

To try in the lab

1. Propose an experiment

2.7 Sample problems and solutions

2.7.1 Problems

Problem 2-1: Consider a sphere which has a charge density of $\rho = ar^2$ and has a radius R. What is the electric field at the distances $0 \leq d \leq r$ and $r < d$ from the centre of the sphere? (Solution)

Problem 2-2: Consider two conducting shells. Both shells have a hollow circle of radius r at their centre. One shell is a square on the outside and the other is a triangle on the outside, both of the outside shapes have a side length of L. A point charge is placed at the centre of the hollowed out circle of both shells. (Solution)

a) What is the electric field outside of the shells?
b) What is the average linear charge density on the inner and outter surfaces of the shells?

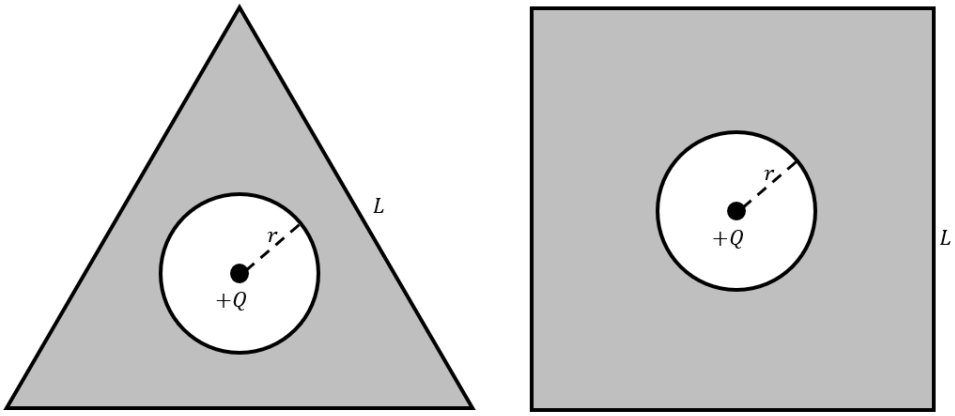


Figure 2.14: A triangular and square shell, both with a hollowed out circular centre and a point charge.

2.7.2 Solutions

Solution to problem 2-2:

First, we must find the total charge of the sphere. To do this, we will need to integrate ρ over the volume of the sphere:

$$\begin{aligned} Q &= \int_0^R \rho dV \\ Q &= \int_0^R 4\pi r^4 dr \\ Q &= \frac{4}{5}\pi R^5 \end{aligned}$$

Now that we have our charge, Q, we can find the enclosed charge at any distance r by taking the ratio of the volumes, then multiplying it by the total charge of the sphere:

$$\begin{aligned} Q_{enc} &= \frac{Q\frac{4}{3}\pi r^3}{\frac{4}{3}\pi R^3} \\ Q_{enc} &= \frac{Qr^3}{R^3} \\ Q_{enc} &= \frac{4\pi r^3 R^2}{5} \end{aligned}$$

Now that we have our enclosed charge for any distance r, we must apply Gauss' law:

$$\begin{aligned} \Phi = EA &= \frac{Q_{enc}}{\epsilon_0} \\ E(4\pi r^2) &= \frac{4\pi r^3 R^2}{5\epsilon_0} \\ E &= \frac{rR^2}{5\epsilon_0} \end{aligned}$$

Which gives our answer for the magnitude of the electric field felt at a distance r from the centre of the shell when $0 \leq r \leq R$.

To find the electric field at a distance r from the centre of the sphere when $R < r$, we will apply the same technique, but will instead set the enclosed charge to be constant at $r = R$:

$$\Phi = EA = \frac{Q_{enc}}{\epsilon_0}$$
$$E(4\pi r^2) = \frac{4\pi R^5}{5\epsilon_0}$$
$$E = \frac{\pi R^5}{5\epsilon_0 r^2}$$

Which gives our final answer.

Solution to problem 2-2:

1. The conducting shells have no net charge, so the only charge in the system is the point charge Q, which means that the electric field outside of the shells is simply $E = \frac{kQ}{r^2}$.
2. Let's begin with the shell that has a triangle on the outside. We will use Gauss' law to determine the charge density of the inner and outer shells. To do this, we will draw a circle within the shell, S_1 and a triangle outside of the outer shell, S_2:

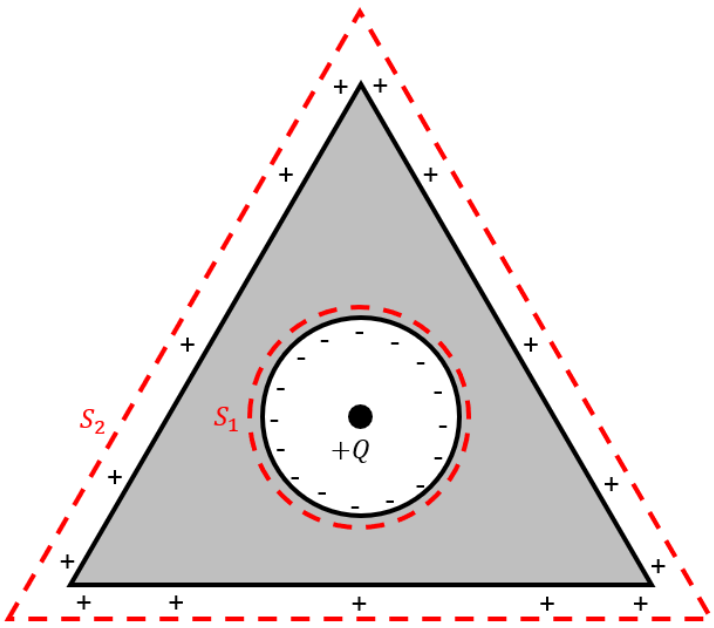


Figure 2.15: A solution to the triangular conducting shell

When considering S_1 know that the electric field inside of the shell is 0, so we also know that the flux will be zero. This means that the point charge on the inside of the shell will be equal and opposite to the sum of the surface charges on the inner shell. From here, we divide the net charge by the circumference of the inner shell to determine the linear charge density:

$$\lambda_{circle} = \frac{-Q}{2\pi r}$$

When considering S_2, we know that the $Q_{enc} = +Q$, which means that the total linear charge on the outer triangle will be $+Q$ such that it cancels the $-Q$ along the inner circle, leaving the point charge, $+Q$. The sum of charges would be $Q_{enc} = Q_{point} + Q_{triangle} - Q_{circle}$. Knowing this, we must divide the total charge on the outter shell by the sum of the length of each of the triangle's sides in order to find the average linear charge density:

$$\lambda_{circle} = \frac{Q}{3L}$$

Now, we must solve for the inner and outer linear charge densities of the conducting shell with a square outter surface. We will begin this process by drawing a circle within the shell, S_1, and a square outside of the shell, S_2

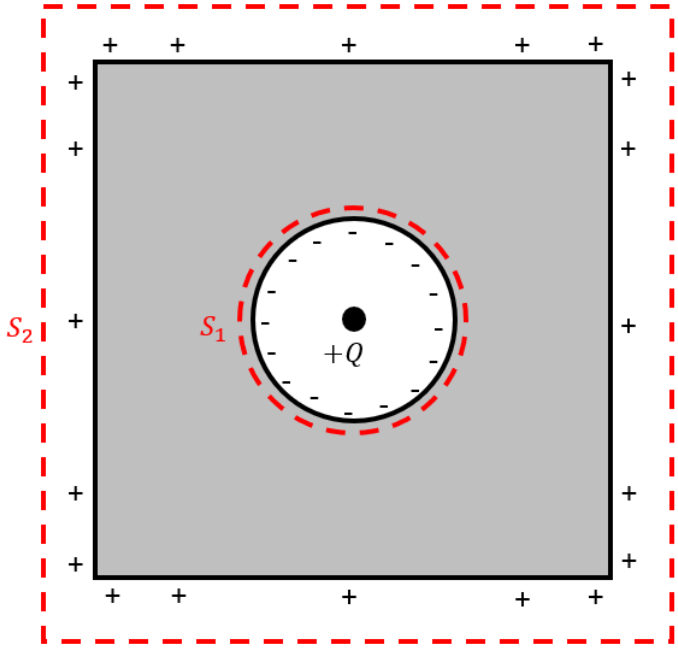


Figure 2.16: A solution to the square conducting shell

For S_1, the circle is treated as it was while solving the trianglular shell. The electric field is also 0 within the square conducting shell, so we know that the average linear charge density is $\frac{-Q}{2\pi r}$.
When considering S_2, we know that Q_{enc} is $+Q$, so we know that the total charge on the square surface of the shell will be $+Q$. This leave us with the following average linear charge density:

$$\lambda_{square} = \frac{Q}{4L}$$

A Vectors

This appendix gives a very brief introduction to coordinate systems and vectors.

Learning Objectives

- Understand the definition of a coordinate system
- Understand the definition of a vector and of a scalar
- Be able to perform algebra with vectors (addition, scalar products, vector products)

A.1 Coordinate systems

Coordinate systems are used to describe the position of an object in space. A coordinate system is an artificial mathematical tool that we construct in order to describe the position of a real object.

A.1.1 1D coordinate systems

The easiest coordinate system to construct is one that we can use to describe the location of objects in one dimensional space. For example, we may wish to describe the location of a train along a straight section of track that runs in the East-West direction. In order to do so, we must first define an "origin", which is the reference point of our coordinate system. For example, the origin for our train track may be the Kingston train station (Figure A.1).

We can describe the position of the train by specifying how far it is from the train station (the origin), using a single real number, say x. If the train is at position $x = 0$, then we know that it is at the Kingston station. If the object is not at the origin, then we need to be able to specify on which side (East or West in our train example) of the origin the object is located. We do this by choosing a direction for our one dimensional coordinate x. For example, we may choose that the East side of the track corresponds to positive values of x and that the West side of the track correspond to the negative values of x. Thus, in order to fully specify a one-dimensional coordinate system we need to choose:

- the location of the origin.
- the direction in which the coordinate, x, increases.

- the units in which we wish to express x.

In one dimension, it is common to use the variable x to define the position along the "x-axis". The x-axis *is* our coordinate system in one dimension, and we represent it by drawing a line with an arrow in the direction of increasing x and indicate where the origin is located (as in Figure A.1).

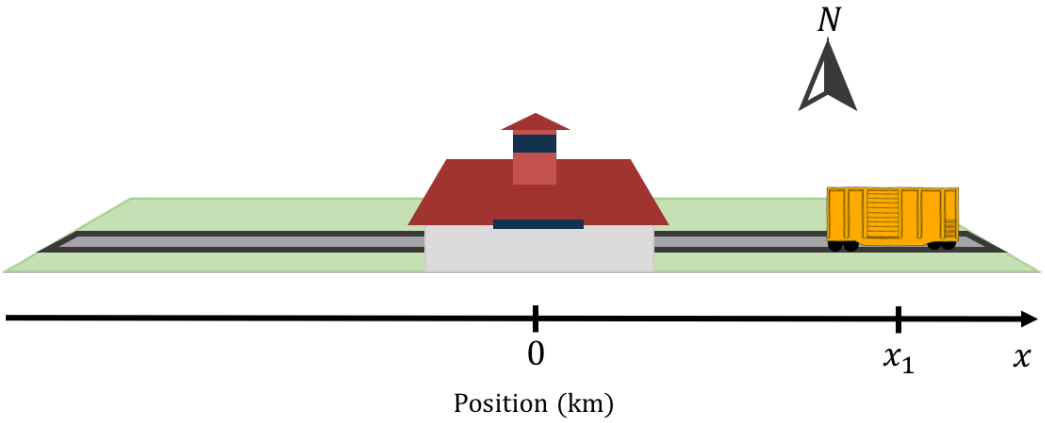


Figure A.1: A 1d coordinate system describing the position of a train. The Kingston train station is the origin and the East side of the track corresponds to positive values of x. The train is located at position x_1.

A.1.2 2D Coordinate systems

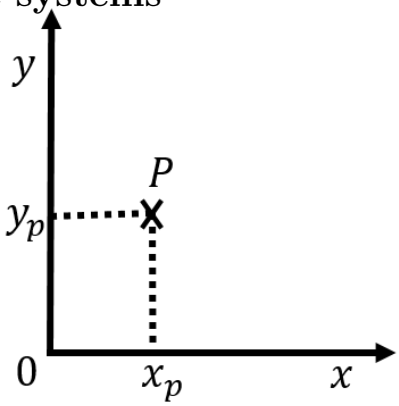


Figure A.2: Example of Cartesian coordinate system and a point P with coordinates (x_p, y_p).

To describe the position of an object in two dimensions (e.g. a marble rolling on a table), we need to specify two numbers. The easiest way to do this is to define two axes, x and y, whose origin and direction we must define. Figure A.2 shows an example of such a coordinate system. Although it is not necessary to do so, we chose x and y axes that are perpendicular to each other. The origin of the coordinate system is where the two axes intersect. One is free to choose any two directions for the axes (as long as they are not parallel). However, choosing axes that are perpendicular (a "Cartesian" coordinate system) is usually the most convenient.

To fully describe the position of an object, we must specify both its position along the x and y axes. For example, point P in Figure A.2 has two **coordinates**, x_p and y_p, that define its position. The x coordinate is found by drawing a line through P that is parallel to the y axis and is given by the intersection of that line with the x axis. The y coordinate is found by drawing a line through point P that is parallel to the x axis and is given by the intersection of that line with the y axis.

Checkpoint A-1

Figure A.3 shows a coordinate system that is not orthogonal (where the x and y axes are not perpendicular). Which value on the figure correctly indicates the y coordinate of point P?

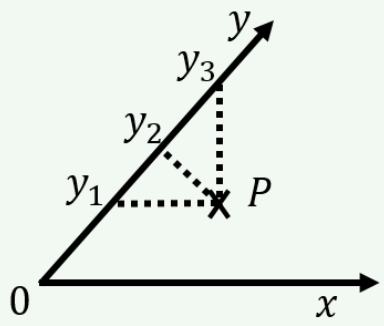


Figure A.3: A non-orthogonal coordinate system (the x and y axes are not perpendicular).

A) y_1
B) y_2
C) y_3

The most common choice of coordinate system in two dimensions is the Cartesian coordinate system that we just described, where the x and y axes are perpendicular and share a common origin, as shown in Figure A.2. When applicable, by convention, we usually choose the y axis to correspond to the vertical direction.

Another common choice is a "polar" coordinate system, where the position of an object is specified by a distance to the origin, r, and an angle, θ, relative to a specified direction, as shown in Figure A.4. Often, a polar coordinate system is defined alongside a Cartesian system, so that r is the distance to the origin of the Cartesian system and θ is the angle with respect to the x axis.

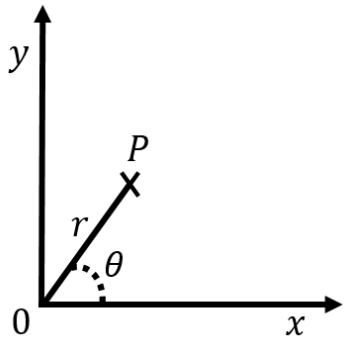


Figure A.4: Example of a polar coordinate system and a point P with coordinates (r, θ).

One can easily convert between the two Cartesian coordinates, x and y, and the two corresponding polar coordinates, r and θ:

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta) \\ r &= \sqrt{x^2 + y^2} \\ \tan(\theta) &= \frac{y}{x} \end{aligned}$$

Polar coordinates are often used to describe the motion of an object moving around a circle, as this means that only one of the coordinates (θ) changes with time (if the origin of the coordinate system is chosen to coincide with the centre of the circle).

A.1.3 3D Coordinate systems

In three dimensions, we need to specify three numbers to describe the position of an object (e.g. a bird flying in the air). In a three dimensional Cartesian coordinate system, we simply add a third axis, z, that is mutually perpendicular to both x and y. The position of an object can then be specified by using the three coordinates, x, y, and z. By convention, we use the z axis to be the vertical direction in three dimensions.

Two additional coordinate systems are common in three dimensions: "cylindrical" and "spherical" coordinates. All three systems are illustrated in Figure A.5 superimposed onto the Cartesian system.

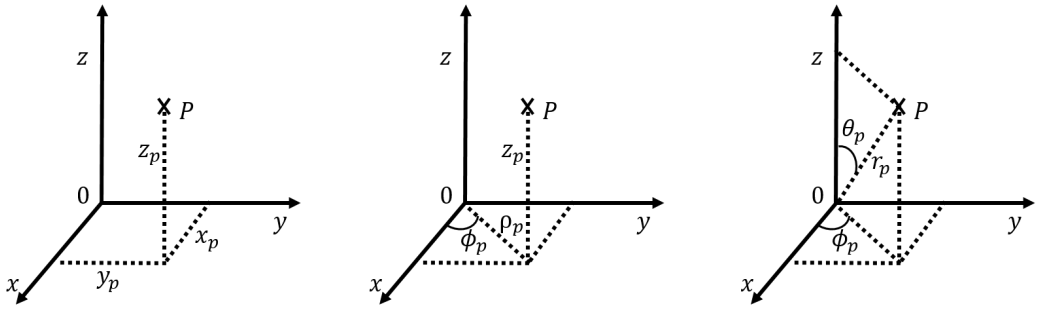


Figure A.5: Cartesian (left), cylindrical (centre) and spherical (right) coordinate systems used in three dimensions. The y and z axes are in the plane of the page, whereas the x axis comes out of the page.

Cylindrical coordinates can be thought of as an extension of the polar coordinates. We keep the same Cartesian coordinate z to indicate the height above the xy plane, however, we use the *azimuthal angle*, ϕ, and the radius, ρ, to describe the position of the projection of a point onto the xy plane. ϕ is the angle between the x axis and the line from the origin to the projection of the point in the xy plane and ρ is the distance between the point and the z axis. Thus, cylindrical coordinates are very similar to the polar coordinate system introduced in two dimensions, except with the addition of the z coordinate. Cylindrical coordinates are useful for describing situations with azimuthal symmetry, such as the motion along the surface of a cylinder. For example, consider point P in Figure A.6. Point P is located a distance ρ from the z axis, as it is located on the surface of the cylinder (the circular end of the cylinder has a radius ρ). Point P is a height z above the xy plane, and a line from the z axis to point P makes an angle ϕ with the x axis.

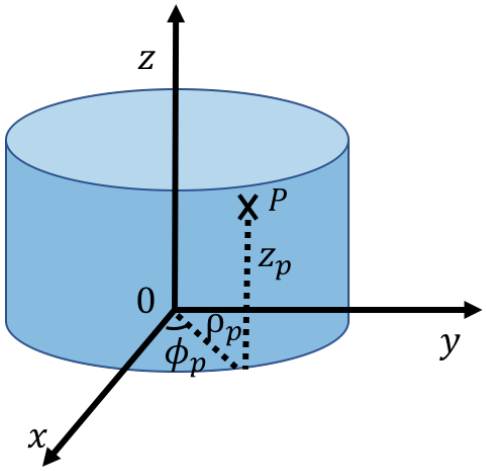


Figure A.6: Describing the position of P, located on the surface of a cylinder, in cylindrical coordinates.

The cylindrical coordinates are related to the Cartesian coordinates by:

$$\begin{aligned} \rho &= \sqrt{x^2 + y^2} \\ \tan(\phi) &= \frac{y}{x} \\ z &= z \end{aligned}$$

In spherical coordinates, a point P is described by the radius, r, the *polar angle* θ, and the *azimuthal angle*, ϕ. The radius is the distance between the point and the origin. The polar angle is the angle with the z axis that is made by the line from the origin to the point. The azimuthal angle is defined in the same way as in polar coordinates. Note that the value of ϕ must be between 0 and 2π, whereas the value of θ must be between 0 and π.

Spherical coordinates are useful for describing situations that have spherical symmetry, such as a person walking on the surface of the Earth, since the radial coordinate will not change. For example, this is shown with Point P in Figure A.7, located on the surface of a sphere

of radius r.

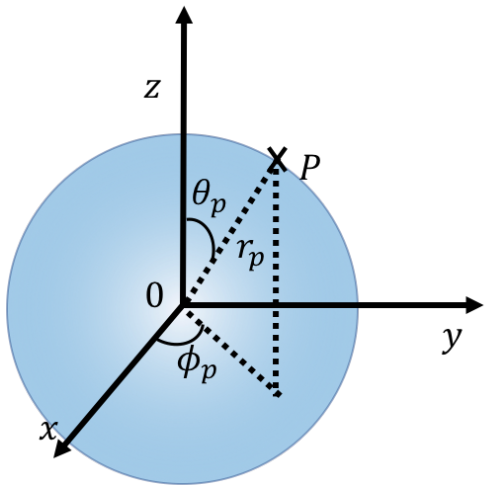


Figure A.7: Describing the position of P, located on the surface of a sphere, in spherical coordinates.

The spherical coordinates are related to the Cartesian coordinates by:

$$r = \sqrt{x^2 + y^2 + z^2}$$
$$\cos(\theta) = \frac{z}{r} = \frac{z}{\sqrt{x^2 + y^2 + z^2}}$$
$$\tan(\phi) = \frac{y}{x}$$

A.2 Vectors

So far, we have seen how to use a coordinate system to describe the position of a single point in space relative to an origin. In this section, we introduce the notion of a "vector", which allows us to describe quantities that have a **magnitude** and a **direction**. For example, you can use a vector to describe the fact that you walked 5 km in the North direction. A vector can be visualized by an arrow. The length of the arrow is the magnitude that we wish to describe, and the direction of the arrow corresponds to the direction that we would like to describe.

Unlike a point in space, vectors **have no location**. That is, vectors are simply an arrow, and you can choose to draw that arrow anywhere you like. In two dimensional space, one requires two numbers to completely define a vector. In three dimensional space, one requires three numbers to completely define a vector. Figure A.8 shows a two dimensional vector, $\vec{d}$, twice. Because both arrows in the figure have the same magnitude and direction, they represent the *same* vector. When we refer to quantities that are vectors, we usually draw an arrow on top of the quantity ($\vec{d}$) to indicate that they are vectors. We use the word "scalar" to refer to numbers that are not vectors (a regular number is thus also called a scalar to distinguish it from a quantity that is a vector).

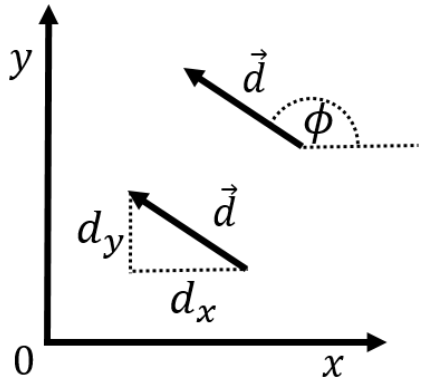


Figure A.8: A vector $\vec{d}$ shown twice, once with its Cartesian components (d_x, d_y) and once with its magnitude and direction (d, ϕ).

In analogy with coordinate systems, we have multiple ways to choose the numbers that we use to describe the vector. The most convenient choice is usually to use the "Cartesian components" of the vector which correspond to the length of the vector when projected onto a Cartesian coordinate system. For example, in Figure A.8, the Cartesian components of the vector $\vec{d}$ are labelled as (d_x, d_y) indicating that the vector has a length of d_x in the x direction and d_y in the y direction. Furthermore, the number d_x is negative, since the vector points in the negative x direction. Another common choice is to use the length of the vector, which we label d (the name of the vector without the arrow on top), and the angle, ϕ that the vector makes with the x-axis, as illustrated in Figure A.8. In terms of the two dimensional Cartesian components, the magnitude of the vector is given by:

$$d = ||\vec{d}|| = \sqrt{d_x^2 + d_y^2}$$

where we also introduced the notation that placing two vertical bars around a vector ($||\vec{d}||$) is used to indicated its magnitude. Note that in three dimensions, it is usually not convenient to specify the direction unless the vector lies in one of the planes defined by the coordinate system (e.g the xy plane). In three dimensions, it is usually most convenient to specify the three Cartesian components.

A.2.1 Unit vectors

A special category of vectors is "unit vectors", which are simply vectors that have a length (magnitude) of 1 (in whichever units the coordinate system is defined). Unit vectors are particularly useful for indicating direction. For example, in Figure A.8, we may be interested in indicating the direction of the vector $\vec{d}$. Unit vectors are denoted by using a "hat" instead of an arrow. Thus, the vector $\hat{d}$, is the vector of length 1 that points in the same direction as $\vec{d}$. The (Cartesian) components of $\hat{d}$ are easily found by dividing the corresponding

components of $\vec{d}$ by d (the magnitude):

$$(\hat{d})_x = \frac{d_x}{d} = \frac{d_x}{\sqrt{d_x^2 + d_y^2}}$$
$$(\hat{d})_y = \frac{d_y}{d} = \frac{d_y}{\sqrt{d_x^2 + d_y^2}}$$
$$\therefore d = ||\hat{d}|| = \sqrt{(\hat{d})_x^2 + (\hat{d})_y^2} = \sqrt{\frac{d_x^2}{d_x^2 + d_y^2} + \frac{d_y^2}{d_x^2 + d_y^2}} = 1$$

A specific type of unit vector is the units vectors that are parallel to the axes of the coordinate system. Those vectors are denoted $\hat{x}$, $\hat{y}$, $\hat{z}$ (and sometimes $\hat{i}$, $\hat{j}$, $\hat{k}$ or $\hat{e}_x$, $\hat{e}_y$, $\hat{e}_z$) for the x, y, and z axes, respectively. Thus, the vector $d\hat{x}$, is the vector of length d that points in the positive x direction.

A.2.2 Notations and representation of vectors

There are multiple notations for describing a vector using its components. The following are all equivalent ways to write down the vector $\vec{d}$ in terms of its components d_x and d_y:

$$\begin{aligned}
\vec{d} &= (d_x, d_y) && \text{row vector} \\
&= \begin{pmatrix} d_x \\ d_y \end{pmatrix} && \text{column vector} \\
&= d_x\hat{x} + d_y\hat{y} && \text{using } \hat{x},\ \hat{y} \\
&= d_x\hat{i} + d_y\hat{j} && \text{using } \hat{i},\ \hat{j}
\end{aligned}$$

The vectors $\hat{x}$ ($\hat{i}$) and $\hat{y}$ ($\hat{j}$) are unit vectors in x and y directions respectively. For example, the unit vector $\hat{y}$ can be written down as (0,1) in two dimensions or (0,1,0) in three dimensions, using the row notation.

Checkpoint A-2

What is the magnitude (the length) of the vector $5\hat{x} - 2\hat{y}$?

A) 3.0
B) 5.4
C) 7.0
D) 10.0

Illustrating a vector graphically in two dimensions is straightforward, but difficult in three dimensions. To help remedy this, a notation is introduced in order to draw vectors that point in or out of the page (perpendicular to the plane of the page). The notation comes from imagining that the vector is an archery arrow. If the vector is coming out of the page (at you!), then you would see the head of the arrow, which we represent as a circle with a dot (the dot is the point of the arrow, the circle is the base of the conically shaped arrowhead). If instead, the vector points into the page, then you would see the back of the

arrow, which we represent as a cross (the cross being the feathers in the tail of the arrow). This is illustrated in Figure A.9.

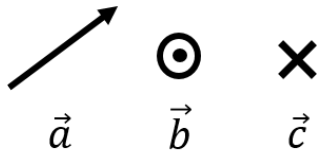


Figure A.9: Geometric representation of three vectors. The vector $\vec{a}$ lies in the plane of the page, the vector $\vec{b}$ is pointing out of the page, and the vector $\vec{c}$ is pointing into the page.

A.3 Vector algebra

In this section, we describe the various algebraic operations that can be performed using vectors.

A.3.1 Multiplication/division of a vector by a scalar

One can multiply (or divide) a vector by a scalar (a number). Suppose that we are given a vector $\vec{v} = (v_x, v_y, v_z)$ and a scalar a. The multiplication $a\vec{v}$ is defined to be a new vector, say $\vec{w}$, whose components are the components of $\vec{v}$ multiplied by a:

$$\vec{w} = a\vec{v} = (av_x, av_y, av_z)$$

Similarly, the division of a vector by a scalar is defined analogously by dividing each Cartesian component by the scalar::

$$\vec{w} = \frac{\vec{v}}{a} = \left(\frac{v_x}{a}, \frac{v_y}{a}, \frac{v_z}{a}\right)$$

Checkpoint A-3

What happens to the length of a vector if the vector is multiplied by 2 (a scalar)?

A) The length doubles
B) The length is halved
C) The length is quadrupled
D) It depends on the direction of the vector

In particular, this makes it easy to determine the unit vector, $\hat{v}$, that points in the same direction as $\vec{v}$:

$$\hat{v} = \frac{\vec{v}}{v}$$

where v is the (scalar) magnitude of $\vec{v}$.

A.3.2 Addition/subtraction of two vectors

The sum of two vectors, $\vec{a}$ and $\vec{b}$, is found by adding the components of the two vectors. Similarly, the difference between two vectors is found by subtracting the components. For

example, if $\vec{c} = \vec{a} + \vec{b}$, the components of $\vec{c}$ are given by:

$$\vec{c} = \vec{a} + \vec{b} = \begin{pmatrix} a_x \\ a_y \end{pmatrix} + \begin{pmatrix} b_x \\ b_y \end{pmatrix}$$
$$\therefore \begin{pmatrix} c_x \\ c_y \end{pmatrix} = \begin{pmatrix} a_x + b_x \\ a_y + b_y \end{pmatrix}$$

where we chose to use the "column vector" notation. The column vector notation highlights the fact that the algebra (addition, subtraction) is performed independently on the x and y components. We can thus use write this sum equivalently as two scalar equations, one for each coordinate:

$$c_x = a_x + b_x$$
$$c_y = a_y + b_y$$

Vectors can thus be used as a short-hand notation for representing multiple equations (one equation per component). When we use vectors to write an equation such as:

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

we really mean that there is one scalar equation per component of the vectors:

$$F_x = ma_x$$
$$F_y = ma_y$$
$$F_z = ma_z$$

Example A-1

Given two vectors, $\vec{a} = 2\hat{x} + 3\hat{y}$, and $\vec{b} = 5\hat{x} - 2\hat{y}$, calculate the vector $\vec{c} = 2\vec{a} - 3\vec{b}$.

Solution

This can easily be solved algebraically by collecting terms for each component, $\hat{x}$ and $\hat{y}$:

$$\begin{aligned} \vec{c} &= 2\vec{a} - 3\vec{b} \\ &= 2(2\hat{x} + 3\hat{y}) - 3(5\hat{x} - 2\hat{y}) \\ &= (4\hat{x} + 6\hat{y}) - (15\hat{x} - 6\hat{y}) \\ &= (4 - 15)\hat{x} + (6 + 6)\hat{y} \\ &= -11\hat{x} + 12\hat{y} \end{aligned}$$

We can think of these operations as being performed independently on the components:

$$c_x = 2a_x - 3b_x = -11$$
$$c_y = 2a_y - 3b_y = 12$$

Geometrically, one can easily visualize the addition and subtraction of vectors. This is illustrated in Figure A.10 for the case of adding vectors $\vec{a}$ and $\vec{b}$ to get the vector $\vec{c}$. Geometrically, the sum of the vectors $\vec{a}$ and $\vec{b}$ (sometimes also called the "resultant") can be found by:

1. Placing the "tail" of vector $\vec{b}$ at the "head" of $\vec{a}$ (think of an arrow, the pointy part is the head and the feathery part is the tail)
2. Drawing the vector that goes from the tail of vector $\vec{a}$ to the head of vector $\vec{b}$.

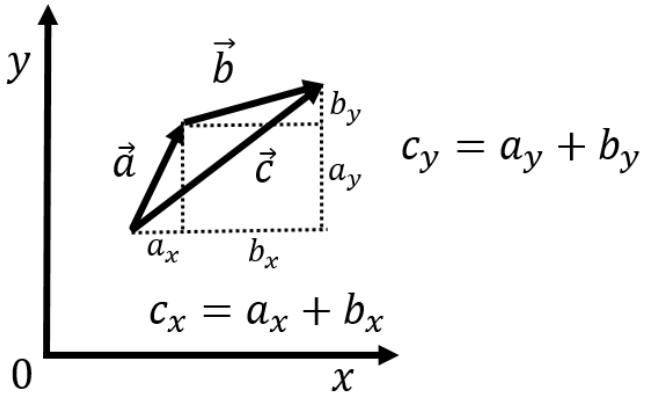


Figure A.10: Geometric addition of the vectors $\vec{a}$ and $\vec{b}$ by placing them "head to tail".

Subtracting two vectors geometrically is done in the same way as addition. For example, the vector $\vec{c}$, given by $\vec{c} = \vec{a} - \vec{b}$ can also be expressed as $\vec{c} = \vec{a} + (-1)\vec{b}$. That is, first multiply the vector $\vec{b}$ by minus 1 (which just reverses its direction), then add that vector, "head to tail", to the vector $\vec{a}$.

Now that we know how to add vectors, we can better understand the notation $\vec{a} = a_x\hat{x} + a_y\hat{y}$. This is not simply a notation, but is in fact algebraically correct. It means: "multiply the vector $\hat{x}$ by a_x (thus giving it a length of a_x) and then add a_y times the vector $\hat{y}$". This is illustrated in Figure A.11, which shows the unit vectors, $\hat{x}$ and $\hat{y}$, which are then multiplied by a_x and a_y, respectively, and then added together "head to tail".

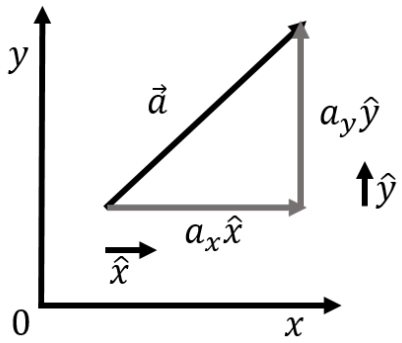


Figure A.11: Illustration that the notation $\vec{a} = a_x\hat{x} + a_y\hat{y}$ is in fact the vector addition of $a_x\hat{x}$ and $a_y\hat{y}$.

A.3.3 The scalar product

There are two ways to multiply vectors: the "scalar product" and the "vector product". The scalar product (or "dot product") takes two vectors and results in a scalar (a number). The vector product (or "cross product") takes two vectors and results in a third vector.

The scalar product, $\vec{a} \cdot \vec{b}$, of two vectors $\vec{a}$ and $\vec{b}$, is defined as the following:

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y$$

That is, one multiplies the individual components of the two vectors and then adds those products for each component. This is easily extended to the three dimensional case by adding a term $a_z b_z$ to the sum. The scalar product is also related to the angle between the two vectors when the vectors are placed "tail to tail", as in Figure A.12

$$\vec{a} \cdot \vec{b} = ab\cos\theta$$

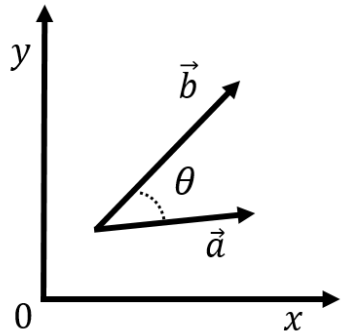


Figure A.12: Illustration of the angle between vectors $\vec{a}$ and $\vec{b}$ when these are placed tail to tail.

The scalar product between two vectors of a fixed length will be maximal when the two vectors are parallel ($\cos\theta = 1$) and zero when the vectors are perpendicular ($\cos\theta = 0$). The scalar product is thus useful when we want to calculate quantities that are maximal when two vectors are parallel.

Checkpoint A-4

The vectors $\vec{a}$ and $\vec{b}$ in the three diagrams below have the same magnitude. Order the diagrams from the one that gives the smallest scalar product $\vec{a} \cdot \vec{b}$ to the largest scalar product.

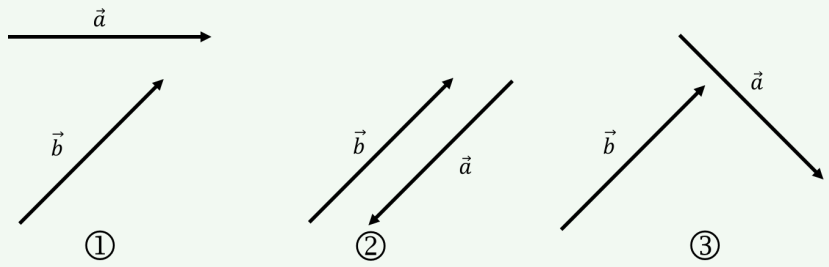


Figure A.13: Put these in order of the magnitude of their scalar product.

A.3.4 The vector product

The vector (or cross) product takes two vectors to produce a third vector that is **mutually perpendicular** to both vectors. The vector product only has meaning in three dimensions. Two vectors that are not co-linear, meaning they can not be arranged so that they lie along the same line, can always be used to define a plane in three dimensions. The cross product of those two vectors will give a third vector that is perpendicular to the plane (making it perpendicular to both vectors).

Algebraically, the three components of the vector product, $\vec{a} \times \vec{b}$, of vectors $\vec{a}$ and $\vec{b}$ are found as follows:

$$\vec{a} \times \vec{b} = \begin{pmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{pmatrix} \tag{A.1}$$

One important property to note is that $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$; that is, the cross product is not commutative (the order matters). The magnitude of the vector obtained by a cross product is given by:

$$||\vec{a} \times \vec{b}|| = ab \sin\theta \tag{A.2}$$

where θ is the angle between the vectors $\vec{a}$ and $\vec{b}$ when these are placed tail to tail (Figure A.12). The vector resulting from a cross product will be null (have a zero length) if the vectors $\vec{a}$ and $\vec{b}$ are parallel, and will have a maximal length when these are perpendicular. The cross product is useful to determine quantities that are maximal when two vectors are perpendicular.

Geometrically, one can determine the direction of the cross product of two vectors by using a "right hand rule". To distinguish it from another right hand rule (see Section A.4.3), we

will call it "the right hand rule for the cross product"). This is done by using your right hand, aligning your thumb with the first vector and your index with the second vector. The cross product will point in the direction of your middle finger (when you hold your middle finger perpendicular to the other two fingers). This is illustrated in Figure A.14. Thus, you can often avoid using equation A.1 and instead use the right hand rule to determine the direction of the cross product and equation A.2 to find its magnitude.

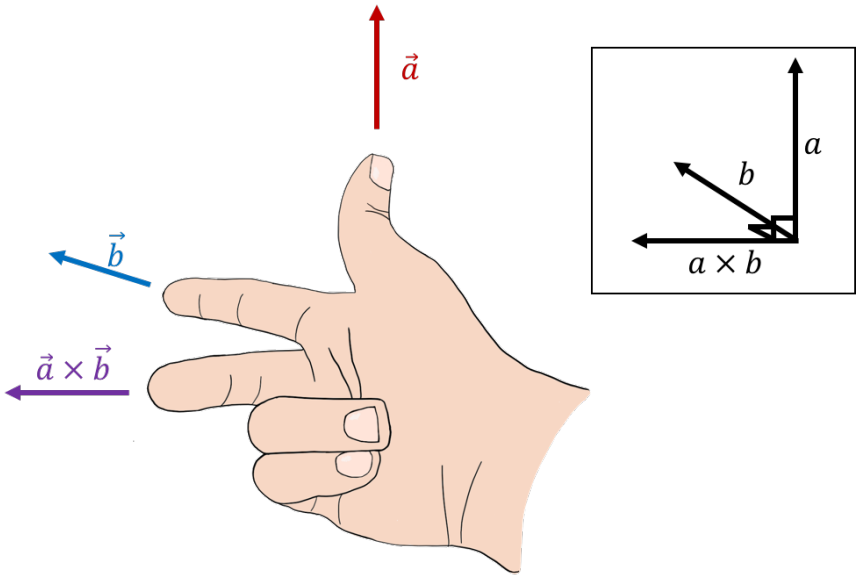


Figure A.14: Using the right hand rule for cross products to find the direction of the cross product of vectors $\vec{a}$ (upwards) and $\vec{b}$ (into the page).

The unit vectors that define a coordinate system have the following properties relative to the cross product:

$$\vec{x} \times \vec{y} = \vec{z}$$
$$\vec{y} \times \vec{z} = \vec{x}$$
$$\vec{z} \times \vec{x} = \vec{y}$$

For these properties to be correct, it should be noted that the direction of the z axis in three dimensions is specified by the choice of x and y axes. That is, one can freely choose the direction of the x and y axes, which then define a plane to which the z axis will be perpendicular. The direction of the z axis must be chosen so that $\vec{x} \times \vec{y} = \vec{z}$ (this guarantees that the coordinate system is "right handed"), as in Figure A.15.

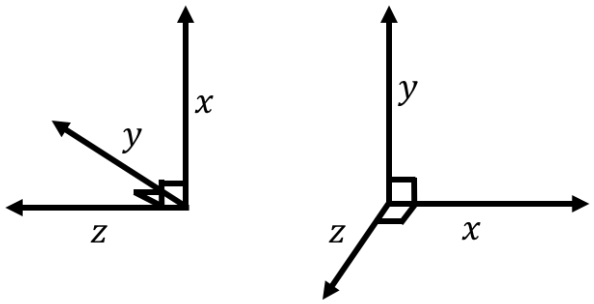


Figure A.15: Two possible orientations for a three dimensional coordinate system. You can confirm using the right hand rule that the z axis is the cross product $\vec{x} \times \vec{y}$.

A.4 Example uses of vectors in physics

This section gives a quick overview of some applications of vectors in physics.

A.4.1 Kinematics and vector equations

Kinematics is the description of the position and motion of an object (Chapters **??** and **??**). The laws of physics are the principles that ultimately allow us to determine how the position of an object changes with time. For example, Newton's Laws are a mathematical framework that introduce the concepts of force and mass in order to model and determine how an object will move through space.

We often use a **position vector**, $\vec{r}(t)$, to describe the position of an object as a function of time. Because the object can move, the position vector is a function of time. A position vector is a special vector in the sense that it should be considered to be fixed in space; the position vector for an object points from the origin of a coordinate system to the location of the object.

The three components of the position vector in Cartesian coordinates, are the x, y, and z coordinates of the object:

$$\vec{r}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}$$

where the three coordinates of the object are functions of time if the object can move. Suppose that the object was initially at position $\vec{r}_1 = (x_1, y_1, z_1)$ at some time $t = t_1$, and that later, at time $t = t_2$, the object was at as second position, $\vec{r}_2 = (x_1, y_1, z_1)$. We can define the **displacement vector**, $\vec{d}$, as the vector from position $\vec{r}_1$ to position $\vec{r}_2$:

$$\vec{d} = \vec{r}_2 - \vec{r}_1 = \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \\ z_2 - z_1 \end{pmatrix} = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix}$$

The displacement vector is such that one can add the vector $\vec{d}$ to the vector $\vec{r}_1$ to describe

the new position of the object at time t_2:

$$\vec{d} = \vec{r}_2 - \vec{r}_1$$
$$\therefore \vec{r}_2 = \vec{r}_1 + \vec{d}$$

The components of the displacement vector, Δx, Δy, and Δz correspond to the displacements (the distance travelled) along the x, y, and z axes, respectively. This is illustrated for the two dimensional case in Figure A.16.

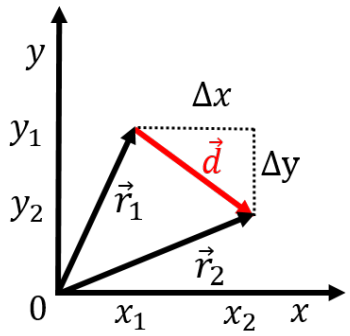


Figure A.16: Illustration of a displacement vector, $\vec{d} = \vec{r}_2 - \vec{r}_1$, for an object that was located at position $\vec{r}_1$ at time t_1 and at position $\vec{r}_2$ at time t_2.

The velocity vector of the object, $\vec{v} = (v_x, v_y, v_z)$, is defined to be the displacement vector, $\vec{d}$, divided by the amount of time (a scalar) that elapsed, $\Delta t = t_2 - t_1$, while the object moved by the corresponding displacement:

$$\vec{v} = \frac{\vec{d}}{\Delta t} = \begin{pmatrix} \frac{\Delta x}{\Delta t} \\ \frac{\Delta y}{\Delta t} \\ \frac{\Delta z}{\Delta t} \end{pmatrix}$$

We used the property that dividing a vector by a scalar (Δt) is defined as dividing each component by the scalar. If we write the components of the velocity vector out explicitly, we have:

$$\begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = \begin{pmatrix} \frac{\Delta x}{\Delta t} \\ \frac{\Delta y}{\Delta t} \\ \frac{\Delta z}{\Delta t} \end{pmatrix}$$

That is, we can think of each row in this "vector equation" as an independent equation. That is, when we write the vector equation:

$$\vec{v} = \frac{\vec{d}}{\Delta t}$$

we are really just using a shorthand notation for writing the three **independent** equations that are true for each individual component of the vectors:

$$v_x = \frac{\Delta x}{\Delta t}$$
$$v_y = \frac{\Delta y}{\Delta t}$$
$$v_z = \frac{\Delta z}{\Delta t}$$

Whenever we write an equation using vectors, we are really writing out multiple equations all at once, one for each component. Newton's Second Law:

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

thus corresponds to the three (scalar) equations:

$$F_x = ma_x$$
$$F_y = ma_y$$
$$F_z = ma_z$$

A.4.2 Work and scalar products

As we will see, work is a scalar quantity that allows us to determine the change in the speed (squared) of an object that results from a force exerted over a particular displacement (Chapter **??**). Both force and the displacement are vector quantities (a force has a magnitude and is exerted in a particular direction). The work, W, done by a force, $\vec{F}$, over a displacements, $\vec{d}$, is defined as:

$$W = \vec{F} \cdot \vec{d}$$

The work energy theorem tells us that this work is related to the change in speed squared of the object as it moves along the displacement vector d. If the work is zero, the object has the same speed at the beginning and end of the displacement. If the work is positive, the object is moving faster at the end of the displacement (and slower if the work is negative). A one dimensional example is shown in Figure A.17, which shows a force $\vec{F}$ being applied to a block as it slides along the ground over a distance d (represented by the displacement vector $\vec{d}$).

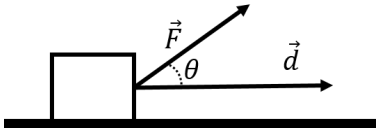


Figure A.17: Example of a force $\vec{F}$ being applied on an object as it moves along the displacement vector $\vec{d}$.

Intuitively, it makes sense that only the horizontal component of the force would contribute to changing the speed of the object as it moves along the horizontal trajectory defined by the vector $\vec{d}$. The vertical component of the force does not contribute to changing the speed of the object. Thus, the work (the change in speed), should only depend on the component of the force that is parallel to the displacement vector. The scalar product allows us to formalize this in an equation. The scalar product is given by:

$$\vec{F} \cdot \vec{d} = Fd\cos\theta = F_{\parallel} d$$

where we introduced $F_{\parallel} = F\cos\theta$ as the component of $\vec{F}$ that is parallel to $\vec{d}$ (see Figure A.17). The scalar product thus "picks out" the component of $\vec{F}$ that is parallel to $\vec{d}$, which is exactly what we need to in order for work to make sense.

A.4.3 Using vectors to describe rotational motion

Often, we need to describe rotational motion in physics. When an object is rotating, one must specify:

1. The axis about which the object is rotating
2. The direction about that axis in which the object is rotating (e.g. clockwise or counter-clockwise)
3. How fast the object is rotating

We introduce a new type of vector, an "axial vector", to describe this kind of rotational motion. We choose the direction of the vector to be co-linear with the axis of rotation and the magnitude of the vector to represent the speed with which the object is rotating. We are thus left with two choices for the direction of the vector. For example, consider the wheels on a car that is moving away from you (Figure A.18, the car is moving into the page). The axis of rotation is the axis of the wheel, so we know that the vector describing the wheel's rotation (the angular velocity vector) must point either to the left or to the right.

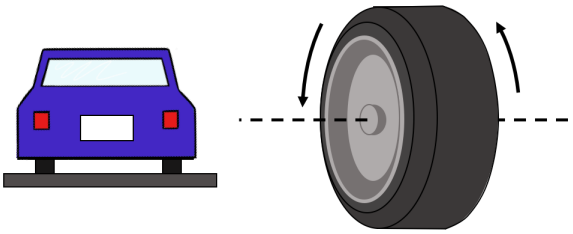


Figure A.18: The wheels on a car that is driving away from you.

We choose the direction of the vector by using another right hand rule. We will refer to this as "the right hand rule for axial vectors" to distinguish it from the right hand rule for the cross product. When using the right hand rule for axial vectors, the vector points in the direction of your thumb when you curl your fingers in the direction of rotation, as in Figure A.19. For the car moving away from you, the wheels will be turning such that the closest point to you is moving up and the furthest point is moving down. Using the right hand rule, we find that the rotation vector points to the left.

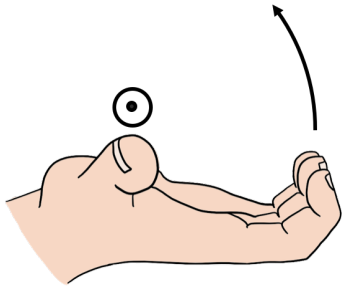


Figure A.19: Using the right hand rule for axial quantities. In this case, the direction of rotation is counter clockwise when looking at the page (the direction that the fingers curl), so the rotation vector points out of the page (the direction of the thumb).

We have to distinguish axial vectors from "true" vectors because they do not behave like true vectors in all cases. For instance, imagine that there is a giant mirror that runs parallel to the road (Figure A.20). When the car is reflected in the mirror, the reflected car will also be moving away from you. This means that the wheels will be turning in the same direction as before, so the rotation vector still points to the left. Now consider a true vector, like a velocity vector. If the velocity vector initially pointed to the left (i.e. if the car was moving to the left), the reflected car would be moving to the *right*. So, we expect a true vector to change directions when it is reflected in this way. Since the rotation vector does not always behave like a true vector, we call it an axial vector or a "pseudovector."

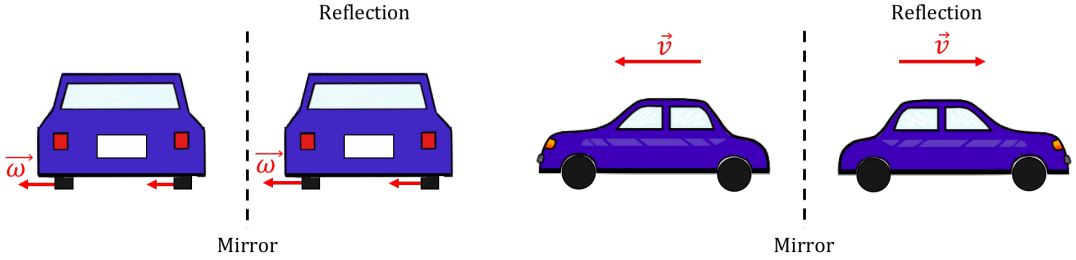


Figure A.20: Left: The angular velocity vector for the rotation of the wheels, $\vec{\omega}$, which points to the left, also points left in the reflection. Right: The velocity vector, pointing to the left, points to the right in the reflection of the car. The angular velocity vector is called an "axial" or "pseudo" vector because it does not change direction under a reflection.

A.4.4 Torque and vector products

We will introduce the concept of a torque in order to describe how a force can cause an object to rotate. Consider the disk illustrated in Figure A.21 that is free to rotate about an axis that goes through its centre and that is perpendicular to the plane of the page. A force $\vec{F}$ is applied at the edge of the disk (imagine pulling on a string attached to the edge of the disk), at a position that is displaced from the axis of rotation by the vector $\vec{r}$. The torque, $\vec{\tau}$, of the force about the centre of the disk is defined to be:

$$\vec{\tau} = \vec{r} \times \vec{F}$$

and represents how much the force $\vec{F}$ will contribute to making the disk rotate about its axis. If the force vector were parallel to the vector $\vec{r}$, the disk would not rotate; if you pull outwards on a disk, it will not rotate about its centre. However, if the force is perpendicular

to the vector $\vec{r}$ (i.e. tangent to the circumference of the disk), then it will maximally cause the disk to rotate. The magnitude of the torque (cross-product) is given by:

$$\tau = rF\sin\theta = F_\perp r = Fr_\perp$$

where θ is the angle between the vectors when placed tail to tail, as in the right side of Figure A.21. In the last two equalities, we have defined $F_\perp = F\sin\theta$ or $r_\perp = r\sin\theta$ to refer to the part of the vector $\vec{F}$ that is perpendicular to the vector $\vec{r}$ or the part of the vector $\vec{r}$ that is perpendicular to the vector $\vec{F}$. That is, the vector product "picks out" the part of a vector that is perpendicular to the other, which is exactly the property that we need for the physical quantity of torque.

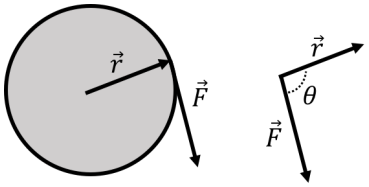


Figure A.21: A force, $\vec{F}$, is exerted in the plane of a disk at a position given by the vector $\vec{r}$ relative to the centre of the disk.

Checkpoint A-5

Referring to Figure A.21, in which direction does the torque vector point?

A) to the right
B) to the left
C) out of the page
D) into the page

A.5 Summary

Key Takeaways

Cartesian coordinate systems can be defined using an origin, and mutually perpendicular axes that specify a direction in which each corresponding coordinate increases. The position of a point is described by the coordinates of the point (one coordinate per axis). Polar, cylindrical and spherical coordinate systems can be defined relative to a Cartesian coordinate system and sometimes facilitate the description of situations with cylindrical (azimuthal) or spherical symmetry.

Vectors can be represented by arrows and are quantities that have both a magnitude and a direction, as opposed to "scalars", which are simply numbers. Vectors are not fixed in space, so two vectors are equal if they have the same magnitude and direction, regardless of where they are drawn. We place a little arrow above a variable, $\vec{d}$, to indicate that it is a vector. There are several, equivalent, notations to indicate the components of a vector:

$$\begin{aligned}
\vec{d} &= (d_x, d_y, d_z) && \text{row vector} \\
&= \begin{pmatrix} d_x \\ d_y \\ d_z \end{pmatrix} && \text{column vector} \\
&= d_x\hat{x} + d_y\hat{y} + d_z\hat{z} && \text{using } \hat{x},\ \hat{y},\ \hat{z} \\
&= d_x\hat{i} + d_y\hat{j} + d_z\hat{j} && \text{using } \hat{i},\ \hat{j},\ \hat{k}
\end{aligned}$$

If we multiply (divide) a vector by a scalar, we multiply (divide) each component of the vector individually by that quantity. As a result, the magnitude of the vector will also be multiplied (divided) by that quantity:

$$a\vec{d} = \begin{pmatrix} ad_x \\ ad_y \\ ad_z \end{pmatrix}$$

In particular, we can define a unit vector, $\hat{d}$, to be a vector of length 1 in the same direction as $\vec{d}$, by simply dividing $\vec{d}$ by its magnitude, d:

$$\hat{d} = \frac{\vec{d}}{d}$$

where the magnitude of the vector, $||\vec{d}|| = d$, expressed in Cartesian coordinates, is

given by:

$$||\vec{d}|| = d = \sqrt{d_x^2 + d_y^2 + d_z^2}$$

We can add two vectors by independently adding the individual components of the vectors:

$$\begin{aligned} \vec{c} &= \vec{a} + \vec{b} \\ \therefore c_x &= a_x + b_x \\ \therefore c_y &= a_y + b_y \\ \therefore c_z &= a_z + b_z \end{aligned}$$

Graphically, this corresponds to adding vectors "head to tail". This also highlights that an equation written using vectors (as the first line above) really represents three independent equations, one for each coordinate of the vectors (or two in two dimensions). Subtraction of vectors is treated in the same way as addition (but using minus signs where appropriate).

One can define the scalar (or dot) product between two vectors, as a scalar quantity that is obtained from the two vectors:

$$\vec{a} \cdot \vec{b} = a_x b_x + a_y b_y + a_z b_z$$

The scalar product is also related to the angle, θ, between the two vectors when these are placed "tail to tail":

$$\vec{a} \cdot \vec{b} = ab \cos\theta$$

In particular, the scalar product between two vectors is zero if the two vectors are perpendicular to each other ($\cos\theta = 0$), and maximal when these are parallel to each other.

The vector (or cross) product between two vectors is a vector that is mutually perpendicular to both vectors and is defined as the following:

$$\vec{a} \times \vec{b} = \begin{pmatrix} a_y b_z - a_z b_y \\ a_z b_x - a_x b_z \\ a_x b_y - a_y b_x \end{pmatrix}$$

The vector product can only be defined in three dimensions, since it must be mutually perpendicular to the vectors. The magnitude of the vector product is given by:

$$||\vec{a} \times \vec{b}|| = ab \sin\theta$$

where θ is the angle between the two vectors when these are placed tail to tail. In particular, the vector product between two vectors is zero if the two vectors are parallel to each other (and maximal when these are perpendicular). The direction of the vector product is given by the right-hand rule for the cross product.

An axial vector can be used to describe a quantity that is related to rotation. The direction of the axial vector is co-linear with the axis of rotation, its magnitude is given by the magnitude of the rotational quantity (e.g. angular speed), and its direction is defined using the right-hand rule for axial vectors.

A.6 Thinking about the Material

Reflect and research

1. What are some quantities that need to be represented by a vector?
2. Can a vector in three dimensions be represented using spherical coordinates? How would you calculate the scalar product between two vectors represented in spherical coordinates?

A.7 Sample problems and solutions

Problem A.1: (Solution)

a) What is the displacement vector from position $(1, 2, 3)$ to position $(4, 5, 6)$?
b) What angle does that displacement vector make with the x axis?

A.7.2 Solutions

Solution to problem A-1:

a) The displacement vector is given by:

$$\vec{d} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}$$

b) We can find the angle that this vector makes with the x axis by taking the scalar product of the displacement vector and the unit vector in the x direction (1,0,0):

$$\hat{x} \cdot \vec{d} = (1)(3) + (0)(3) + (0)(3) = 3$$

This is equal to the product of the magnitude of $\hat{x}$ and $\vec{d}$ multiplied by the cosine of the angle between them:

$$\begin{aligned} \hat{x} \cdot \vec{d} &= ||\hat{x}||||\vec{d}|| \cos\theta = (1)(\sqrt{3^2 + 3^2 + 3^2}) \cos\theta = \sqrt{27} \cos\theta \\ 3 &= \sqrt{27} \cos\theta \\ \therefore \cos\theta &= \frac{3}{\sqrt{27}} = \frac{1}{\sqrt{3}} \\ \theta &= 54.7^\circ \end{aligned}$$

B Calculus

This appendix gives a very brief introduction to calculus with a focus on the tools needed in physics.

Learning Objectives

- Understand how to determine a derivative and that it measures a rate of change.
- Understand how to determine partial derivatives and gradients.
- Understand how to determine anti-derivatives and that integrals are sums.

B.1 Functions of real numbers

In calculus, we work with functions and their properties, rather than with variables as we do in algebra. We are usually concerned with describing functions in terms of their slope, the area (or volumes) that they enclose, their curvature, their roots (when they have a value of zero) and their continuity. The functions that we will examine are a mapping from one or more *independent* real numbers to one real number. By convention, we will use x, y, z to indicate independent variables, and $f()$ and $g()$, to denote functions. For example, if we say:

$$f(x) = x^2$$
$$\therefore f(2) = 4$$

we mean that $f(x)$ is a function that can be evaluated for any real number, x, and the result of evaluating the function is to square the number x. In the second line, we evaluated the function with $x = 2$. Similarly, we can have a function, $g(x, y)$ of multiple variables:

$$g(x, y) = x^2 + 2y^2$$
$$\therefore g(2, 3) = 22$$

We can easily visualize a function of 1 variable, for example by plotting it in python (see Appendix D):

Python Code B.1: Plotting a function of 1 variable

```
#import pacakges for creating arrays of values and for plotting
import numpy as np #arrays
import pylab as pl #plotting

#define the function:
def f(x):
    return x*x

#create 100 values of x between -5 and +5
xvals = np.linspace(-5,5,100)

#Plot the function evaluated at the values of x against the values of x:
pl.plot(xvals,f(xvals))
pl.xlabel('x')
pl.ylabel('f(x)')
pl.title('f(x)=$x^2$')
pl.grid()
pl.show()
```

Output B.1:

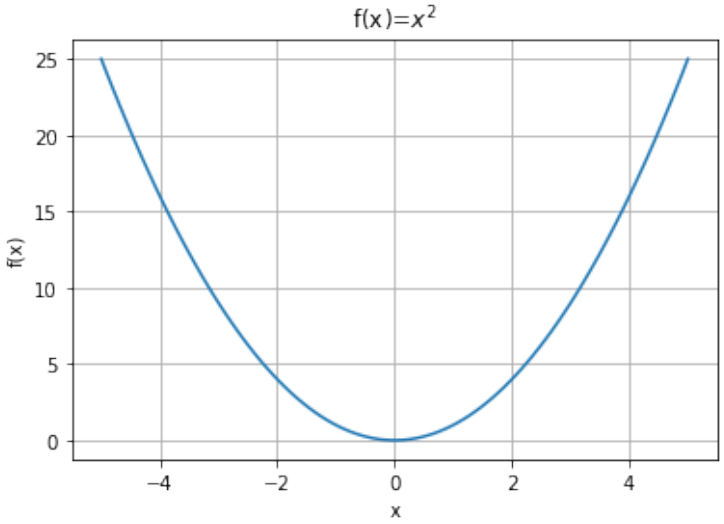


Figure B.1: $f(x) = x^2$ plotted between $x = -5$ and $= +5$.

Plotting a function of 2 variables is a little trickier, since we need to do it in three dimensions (one axis for x, one axis for y, and one axis for $g(x, y)$). This can be done in python with a little more work:

Python Code B.2: Plotting a function of 2 variables

```
#import pacakges for creating arrays of values and for plotting
import numpy as np #arrays
import pylab as pl #plotting
#import package for handling 3D graphs:
from mpl_toolkits.mplot3d import Axes3D
```

```
#define the function:
def g(x,y):
    return x*x+2*y*y

#create 100 values of x and y between -5 and +5
xvals = np.linspace(-5,5,100)
yvals = np.linspace(-5,5,100)
#create a grid with the values of x and y:
X,Y = np.meshgrid(xvals, yvals)
#evaluate the function everywhere on the grid
gvals = g(X,Y)

#Plot the function as a surface (create a figure, add 3D, plot it):
fig = pl.figure(figsize=(10,10))
ax = fig.add_subplot(111, projection='3d')
ax.plot_surface(X,Y,gvals,cmap="Blues")
#show contours for the surface, projected on xy plane:
ax.contour(X, Y, gvals, offset=-1,cmap="Blues")
#add some labels
ax.set_xlabel('x')
ax.set_ylabel('y')
ax.set_zlabel('g(x,y)')
ax.set_title("$g(x,y)=x^2+2y^2$")
#choose the view point:
ax.view_init(elev=30, azim=-25)
pl.show()
```

Output B.2:

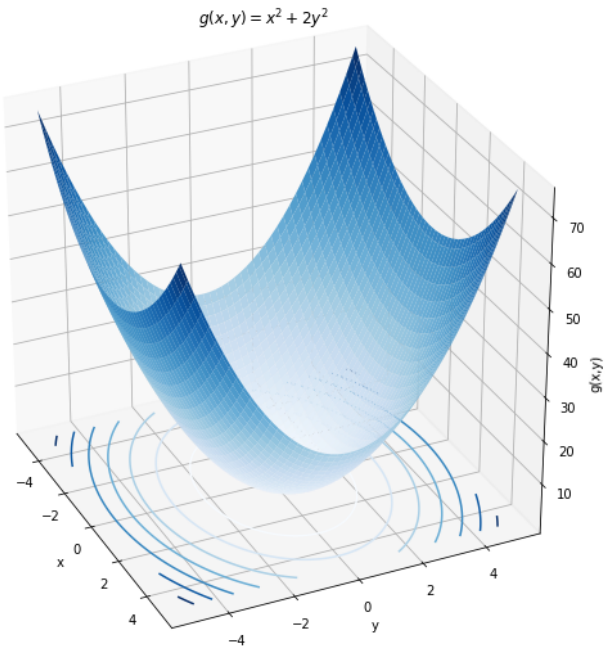


Figure B.2: $g(x, y) = x^2 + 2y^2$ plotted for x between -5 and +5 and for y between -5 and +5. A function of two variables can be visualized as a surface in three dimensions. One can also visualize the function by look at its "contours" (the lines drawn in the xy plane).

Unfortunately, it becomes difficult to visualize functions of more than 2 variables, although one can usually look at projections of those functions to try and visualize some of the features (for example, contour maps are 2D projections of 3D surfaces, as shown in the xy plane of Figure B.2). When you encounter a function, it is good practice to try and visualize it if you can. For example, ask yourself the following questions:

- Does the function have one or more maxima and/or minima?
- Does the function cross zero?
- Is the function continuous everywhere?
- Is the function always defined for any value of the independent variables?

B.2 Derivatives

Consider the function $f(x) = x^2$ that is plotted in Figure B.1. For any value of x, we can define the slope of the function as the "steepness of the curve". For values of $x > 0$ the function increases as x increases, so we say that the slope is positive. For values of $x < 0$, the function decreases as x increases, so we say that the slope is negative. A synonym for the word slope is "derivative", which is the word that we prefer to use in calculus. The derivative of a function $f(x)$ is given the symbol $\frac{df}{dx}$ to indicate that we are referring to the slope of $f(x)$ when plotted as a function of x.

We need to specify which variable we are taking the derivative with respect to when the function has more than one variable but only one of them should be considered *independent*. For example, the function $f(x) = ax^2 + b$ will have different values if a and b are changed, so we have to be precise in specifying that we are taking the derivative with respect to x. The following notations are equivalent ways to say that we are taking the derivative of $f(x)$ with respect to x:

$$\frac{df}{dx} = \frac{d}{dx}f(x) = f'(x) = f'$$

The notation with the prime ($f'(x), f'$) can be useful to indicate that the derivative itself is *also* a function of x.

The slope (derivative) of a function tells us how rapidly the value of the function is changing when the independent variable is changing. For $f(x) = x^2$, as x gets more and more positive, the function gets steeper and steeper; the derivative is thus increasing with x. The sign of the derivative tells us if the function is increasing or decreasing, whereas its absolute value tells how quickly the function is changing (how steep it is).

We can approximate the derivative by evaluating how much $f(x)$ changes when x changes by a small amount, say, Δx. In the limit of $\Delta x \to 0$, we get the derivative. In fact, this is the formal definition of the derivative:

$$\boxed{\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}} \tag{B.1}$$

where Δf is the small change in $f(x)$ that corresponds to the small change, Δx, in x. This makes the notation for the derivative more clear, dx is Δx in the limit where $\Delta x \to 0$, and df is Δf, in the same limit of $\Delta x \to 0$.

As an example, let us determine the function $f'(x)$ that is the derivative of $f(x) = x^2$. We start by calculating Δf:

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

We now calculate $\frac{\Delta f}{\Delta x}$:

$$\begin{aligned}\frac{\Delta f}{\Delta x} &= \frac{2x\Delta x + \Delta x^2}{\Delta x} \\ &= 2x + \Delta x\end{aligned}$$

and take the limit $\Delta x \to 0$:

$$\begin{aligned}\frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} \\ &= \lim_{\Delta x \to 0} (2x + \Delta x) \\ &= 2x\end{aligned}$$

We have thus found that the function, $f'(x) = 2x$, is the derivative of the function $f(x) = x^2$. This is illustrated in Figure B.3. Note that:

- For $x > 0$, $f'(x)$ is positive and increasing with increasing x, just as we described earlier (the function $f(x)$ is increasing and getting steeper).
- For $x < 0$, $f'(x)$ is negative and decreasing in magnitude as x increases. Thus $f(x)$ decreases and gets less steep as x increases.
- At $x = 0$, $f'(x) = 0$ indicating that, at the origin, the function $f(x)$ is (momentarily) flat.

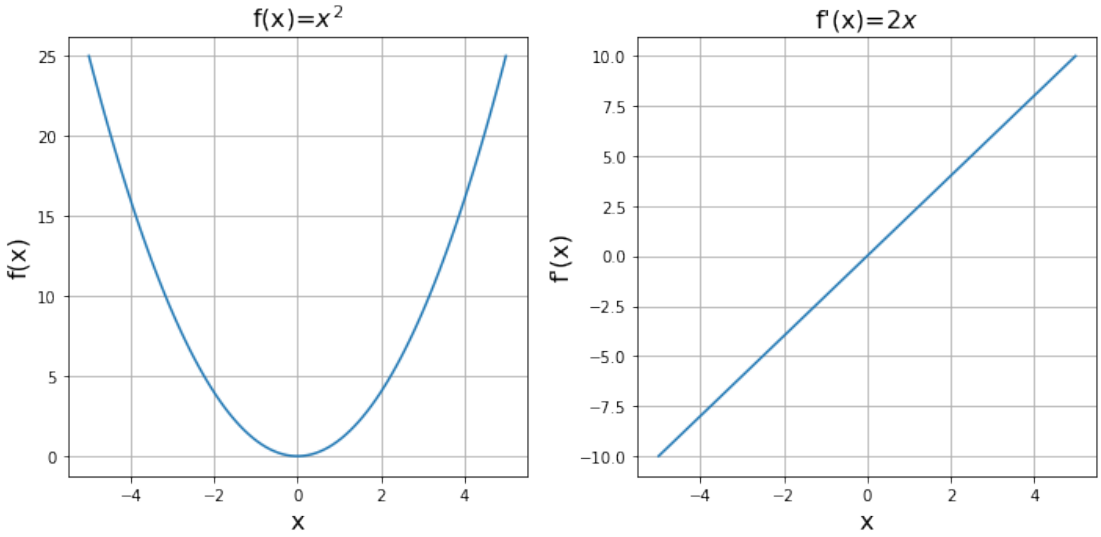


Figure B.3: $f(x) = x^2$ and its derivative, $f'(x) = 2x$ plotted for x between -5 and +5.

Checkpoint B-1

When a function has a maximum, its derivative at that point

A) also has a maximum
B) is zero
C) has a minimum
D) is infinite

B.2.1 Common derivatives and properties

It is beyond the scope of this textbook to derive the functional form of the derivative for any function using equation B.1. Table B.1 below gives the derivatives for common functions. In all cases, x is the independent variable, and all other variables should be thought of as constants:

Function, $f(x)$	**Derivative,** $f'(x)$
$f(x) = a$	$f'(x) = 0$
$f(x) = x^n$	$f'(x) = nx^{n-1}$
$f(x) = \sin(x)$	$f'(x) = \cos(x)$
$f(x) = \cos(x)$	$f'(x) = -\sin(x)$
$f(x) = \tan(x)$	$f'(x) = \frac{1}{\cos^2(x)}$
$f(x) = e^x$	$f'(x) = e^x$
$f(x) = \ln(x)$	$f'(x) = \frac{1}{x}$

Table B.1: Common derivatives of functions.

If two functions of 1 variable, $f(x)$ and $g(x)$, are combined into a third function, $h(x)$, then

there are simple rules for finding the derivative, $h'(x)$, based on the derivatives $f'(x)$ and $g'(x)$. These are summarized in Table B.2 below.

Function, $h(x)$	**Derivative,** $h'(x)$
$h(x) = f(x) + g(x)$	$h'(x) = f'(x) + g'(x)$
$h(x) = f(x) - g(x)$	$h'(x) = f'(x) - g'(x)$
$h(x) = f(x)g(x)$	$h'(x) = f'(x)g(x) + f(x)g'(x)$ (The product rule)
$h(x) = \frac{f(x)}{g(x)}$	$h'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$ (The quotient rule)
$h(x) = f(g(x))$	$h'(x) = f'(g(x))g'(x)$ (The Chain Rule)

Table B.2: Derivatives of combined functions.

Example B-1

Use the properties from Table B.2 to show that the derivative of $\tan(x)$ is $\frac{1}{\cos^2(x)}$

Solution

Since $\tan(x) = \frac{\sin(x)}{\cos(x)}$, we can write:

$$\begin{aligned} h(x) &= \frac{f(x)}{g(x)} \\ f(x) &= \sin(x) \\ g(x) &= \cos(x) \end{aligned}$$

Using the fourth row in Table B.2, and the common derivatives from Table B.1, we have:

$$\begin{aligned} f'(x) &= \cos(x) \\ g'(x) &= -\sin(x) \\ g^2(x) &= \cos^2(x) \\ h'(x) &= \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)} \\ &= \frac{\cos(x)\cos(x) - \sin(x)(-\sin(x))}{\cos^2} \\ &= \frac{\cos^2(x) + \sin^2(x)}{\cos^2} \\ &= \frac{1}{\cos^2(x)} \end{aligned}$$

as required.

Example B-2

Use the properties from Table B.2 to calculate the derivative of $h(x) = \sin^2(x)$

Solution

To calculate the derivative of $h(x)$, we need to use the Chain Rule. $h(x)$ is found by first taking $\sin(x)$ and then taking that result squared. We can thus identify:

$$\begin{aligned} h(x) &= \sin^2(x) = f(g(x)) \\ f(x) &= x^2 \\ g(x) &= \sin(x) \end{aligned}$$

Using the common derivatives from Table B.1, we have:

$$\begin{aligned} f'(x) &= 2x \\ g'(x) &= \cos(x) \end{aligned}$$

Applying the Chain Rule, we have:

$$\begin{aligned} h'(x) &= f'(g(x))g'(x) \\ &= 2\sin(x)g'(x) \\ &= 2\sin(x)\cos(x) \end{aligned}$$

where $f'(g(x))$ means apply the derivative of $f(x)$ to the function $g(x)$. Since the derivative of $f(x)$ is $f'(x) = 2x$, when we apply it to $g(x)$ instead of $2x$, we get $2g(x) = 2\cos(x)$.

B.2.2 Partial derivatives and gradients

So far, we have only looked at the derivative of a function of a single independent variable and used it to quantify how much the function changes when the independent variable changes. We can proceed analogously for a function of multiple variables, $f(x, y)$, by quantifying how much the function changes along the direction associated with a particular variable. This is illustrated in Figure B.4 for the function $f(x, y) = x^2 - 2y^2$, which looks somewhat like a saddle.

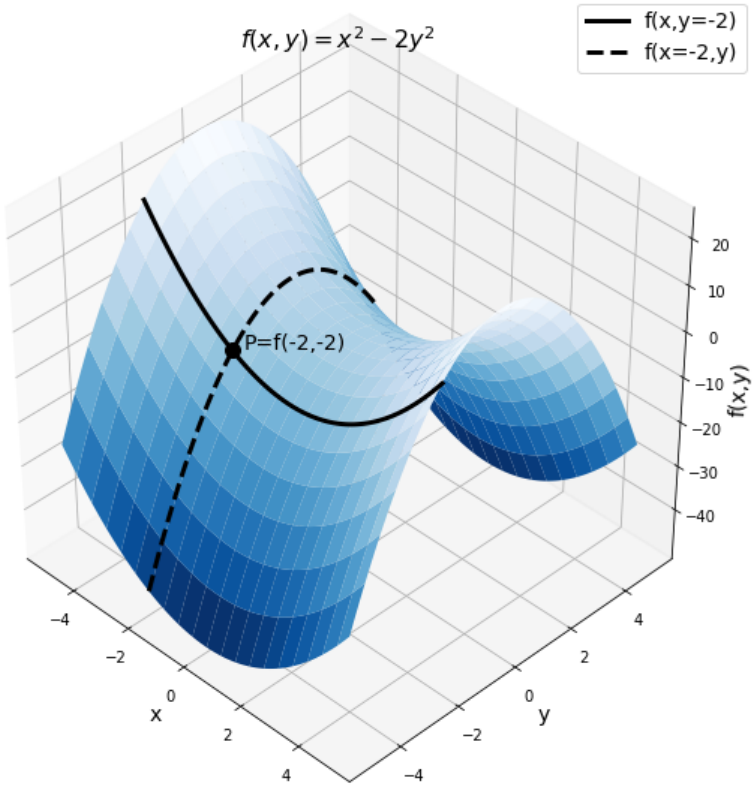


Figure B.4: $f(x, y) = x^2 - 2y^2$ plotted for x between -5 and +5 and for y between -5 and +5. The point P labelled on the figure shows the value of the function at $f(-2, -2)$. The two lines show the function evaluated when one of x or y is held constant.

Suppose that we wish to determine the derivative of the function $f(x)$ at $x = -2$ and $y = -2$. In this case, it does not make sense to simply determine the "derivative", but rather, we must specify *in which direction* we want the derivative. That is, we need to specify in which direction we are interested in quantifying the rate of change of the function.

One possibility is to quantify the rate of change in the x direction. The solid line in Figure B.4 shows the part of the function surface where y is fixed at -2, that is, the function evaluated as $f(x, y = -2)$. The point P on the figure shows the value of the function when $x = -2$ and $y = -2$. By looking at the solid line at point P, we can see that as x increases, the value of the function is gently decreasing. The derivative of $f(x, y)$ with respect to x when y is held constant and evaluated at $x = -2$ and $y = -2$ is thus negative. Rather than saying "The derivative of $f(x, y)$ with respect to x when y is held constant" we say "The **partial derivative** of $f(x, y)$ with respect to x".

Since the partial derivative is different than the ordinary derivative (as it implies that we are holding independent variables fixed), we give it a different symbol, namely, we use ∂

instead of d:

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x} f(x, y) \text{ (Partial derivative of f with respect to x)}$$

Calculating the partial derivative is very easy, as we just treat all variables as constants except for the variable with respect to which we are differentiating[1]. For the function $f(x, y) = x^2 - 2y^2$, we have:

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(x^2 - 2y^2) = 2x$$
$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(x^2 - 2y^2) = -4y$$

At $x = -2$, the partial derivative of $f(x, y)$ is indeed negative, consistent with our observation that, along the solid line, at point P, the function is decreasing.

A function will have as many partial derivatives as it has independent variables. Also note that, just like a normal derivative, a partial derivative is still a function. The partial derivative with respect to a variable tells us how steep the function is in the direction in which that variable increases and whether it is increasing or decreasing.

Example B-3

Determine the partial derivatives of $f(x, y, z) = ax^2 + byz - \sin(z)$.

Solution

In this case, we have three partial derivatives to evaluate. Note that a are b constants and can be thought of as numbers that we do not know.

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(ax^2 + byz - \sin(z)) = 2ax$$
$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(ax^2 + byz - \sin(z)) = bz$$
$$\frac{\partial f}{\partial z} = \frac{\partial}{\partial z}(ax^2 + byz - \sin(z)) = by - \cos(z)$$

Since the partial derivatives tell us how the function changes in a particular direction, we can use them to find the direction in which the function changes *the most rapidly*. For example, suppose that the surface from Figure B.4 corresponds to a real physical surface and that we place a ball at point P. We wish to know in which direction the ball will roll. The direction that it will roll in is the opposite of the direction where $f(x, y)$ increases the most rapidly (i.e. it will roll in the direction where $f(x, y)$ decreases the most rapidly).

[1] To take the derivative is to "differentiate"!

The direction in which the function increases the most rapidly is called the "gradient" and denoted by $\nabla f(x, y)$.

Since the gradient is a direction, it cannot be represented by a single number. Rather, we use a "vector" to indicate this direction. Since $f(x, y)$ has two independent variables, the gradient will be a vector with two components. The components of the gradient are given by the partial derivatives:

$$\nabla f(x, y) = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y}$$

where $\hat{x}$ and $\hat{y}$ are the unit vectors in the x and y directions, respectively (sometimes, the unit vectors are denoted $\hat{i}$ and $\hat{j}$). The direction of the gradient tells us in which direction the function increases the fastest, and the magnitude of the gradient tells us how much the function increases in that direction.

Example B-4

Determine the gradient of the function $f(x, y) = x^2 - 2y^2$ at the point $x = -2$ and $y = -2$.

Solution

We have already found the partial derivatives that we need to evaluate at $x = -2$ and $y = -2$:

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2x \\ \frac{\partial f}{\partial y} &= -4y \\ \therefore \nabla f(x, y) &= \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} \\ &= 2x\hat{x} - 4y\hat{y} \end{aligned}$$

Evaluating the gradient at $x = -2$ and $y = -2$:

$$\begin{aligned} \nabla f(x, y) &= 2x\hat{x} - 4y\hat{y} \\ &= -4\hat{x} + 8\hat{y} \\ &= 4(-\hat{x} + 2\hat{y}) \end{aligned}$$

The gradient vector points in the direction $(-1, 2)$. That is, the function increases the most in the direction where you would take 1 pace in the negative x direction and 2 paces in the positive y direction. You can confirm this by looking at point P in Figure

B.4 and imagining in which direction you would have to go to climb the surface to get the steepest climb.

The gradient is itself a function, but it is not a real function (in the sense of a real number), since it evaluates to a vector. It is a mapping from real numbers x, y to a vector. As you take more advanced calculus courses, you will eventually encounter "vector calculus", which is just the calculus for functions of multiple variables to which you were just introduced. The key point to remember here is that the gradient can be used to find the vector that points in the direction of maximal increase of the corresponding multi-variate function. This is precisely the quantity that we need in physics to determine in which direction a ball will roll when placed on a surface (it will roll in the direction opposite to the gradient vector).

Checkpoint B-2

The gradient of a function of one variable, $f(x)$, is

A) undefined
B) zero
C) equal to its derivative
D) infinite

B.2.3 Common uses of derivatives in physics

The simplest case of using a derivative is to describe the speed of an object. If an object covers a distance Δx in a period of time Δt, it's "average speed", v_{avg}, is defined as the distance covered by the object divided by the amount of time it took to cover that distance:

$$v_{avg} = \frac{\Delta x}{\Delta t}$$

If the object changes speed (for example it is slowing down) over the distance Δx, we can still define its "instantaneous speed", v, by measuring the amount of time, Δt, that it takes the object to cover a *very small distance*, Δx. The instantaneous speed is defined in the limit where $\Delta x \to 0$:

$$v = \lim_{\Delta x \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt}$$

which is precisely the derivative of $x(t)$ with respect to t. $x(t)$ is a function that gives the position, x, of the object along some x axis as a function of time. The speed of the object is thus the rate of change of its position.

Similarly, if the speed is changing with time, then we can define the "acceleration", a, of an object as the rate of change of its speed:

$$a = \frac{dv}{dt}$$

B.3 Anti-derivatives and integrals

In the previous section, we were concerned with determining the derivative of a function $f(x)$. The derivative is useful because it tells us how the function $f(x)$ varies as a function of x. In physics, we often know how a function varies, but we do not know the actual

function. In other words, we often have the opposite problem: we are given the derivative of a function, and wish to determine the actual function. For this case, we will limit our discussion to functions of a single independent variable.

Suppose that we are given a function $f(x)$ and we know that this is the derivative of some other function, $F(x)$, which we do not know. We call $F(x)$ the **anti-derivative** of $f(x)$. The anti-derivative of a function $f(x)$, written $F(x)$, thus satisfies the property:

$$\frac{dF}{dx} = f(x)$$

Since we have a symbol for indicating that we take the derivative with respect to x ($\frac{d}{dx}$), we also have a symbol, $\int dx$, for indicating that we take the anti-derivative with respect to x:

$$\int f(x)dx = F(x)$$
$$\therefore \frac{d}{dx}\left(\int f(x)dx\right) = \frac{dF}{dx} = f(x)$$

Earlier, we justified the symbol for the derivative by pointing out that it is like $\frac{\Delta f}{\Delta x}$ but for the case when $\Delta x \to 0$. Similarly, we will justify the anti-derivative sign, $\int f(x)dx$, by showing that it is related to a sum of $f(x)\Delta x$, in the limit $\Delta x \to 0$. The $\int$ sign looks like an "S" for sum.

While it is possible to exactly determine the derivative of a function $f(x)$, the anti-derivative can only be determined up to a constant. Consider for example a different function, $\tilde{F}(x) = F(x) + C$, where C is a constant. The derivative of $\tilde{F}(x)$ with respect to x is given by:

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

Hence, the function $\tilde{F}(x) = F(x) + C$ is also an anti-derivative of $f(x)$. The constant C can often be determined using additional information (sometimes called "initial conditions"). Recall the function, $f(x) = x^2$, shown in Figure B.3 (left panel). If you imagine shifting the whole function up or down, the derivative would not change. In other words, if the origin of the axes were not drawn on the left panel, you would still be able to determine the derivative of the function (how steep it is). Adding a constant, C, to a function is exactly the same as shifting the function up or down, which does not change its derivative. Thus, when you know the derivative, you cannot know the value of C, unless you are also told that the function must go through a specific point (a so-called initial condition).

In order to determine the derivative of a function, we used equation B.1. We now need to derive an equivalent prescription for determining the anti-derivative. Suppose that we have the two pieces of information required to determine $F(x)$ completely, namely:

1. the function $f(x) = \frac{dF}{dx}$ (its derivative).
2. the condition that $F(x)$ must pass through a specific point, $F(x_0) = F_0$.

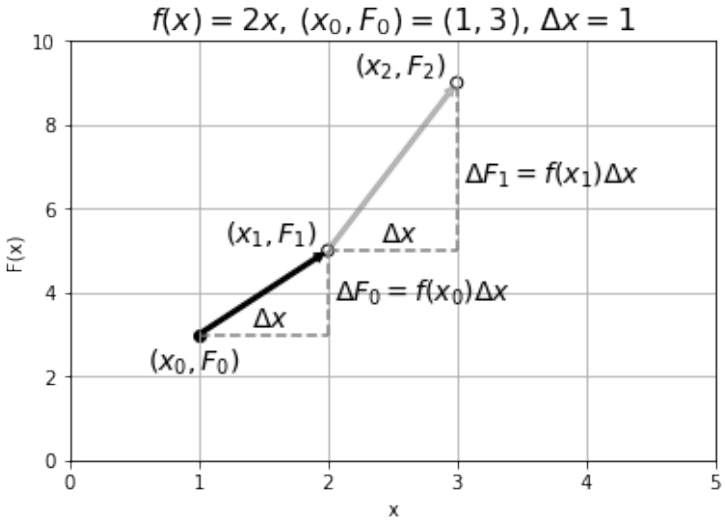


Figure B.5: Determining the anti-derivative, $F(x)$, given the function $f(x) = 2x$ and the initial condition that $F(x)$ passes through the point $(x_0, F_0) = (1, 3)$.

The procedure for determining the anti-derivative $F(x)$ is illustrated above in Figure B.5. We start by drawing the point that we know the function $F(x)$ must go through, (x_0, F_0). We then choose a value of Δx and use the derivative, $f(x)$, to calculate ΔF_0, the amount by which $F(x)$ changes when x changes by Δx. Using the derivative $f(x)$ evaluated at x_0, we have:

$$\begin{aligned} \frac{\Delta F_0}{\Delta x} &\approx f(x_0) \quad (\text{in the limit} \Delta x \to 0) \\ \therefore \Delta F_0 &= f(x_0)\Delta x \end{aligned}$$

We can then estimate the value of the function $F_1 = F(x_1)$ at the next point, $x_1 = x_0 + \Delta x$, as illustrated by the black arrow in Figure B.5

$$\begin{aligned} F_1 &= F(x_1) \\ &= F(x + \Delta x) \\ &\approx F_0 + \Delta F_0 \\ &\approx F_0 + f(x_0)\Delta x \end{aligned}$$

Now that we have determined the value of the function $F(x)$ at $x = x_1$, we can repeat the procedure to determine the value of the function $F(x)$ at the next point, $x_2 = x_1 + \Delta x$. Again, we use the derivative evaluated at x_1, $f(x_1)$, to determine ΔF_1, and add that to F_1

to get $F_2 = F(x_2)$, as illustrated by the grey arrow in Figure B.5:

$$\begin{aligned} F_2 &= F(x_1 + \Delta x) \\ &\approx F_1 + \Delta F_1 \\ &\approx F_1 + f(x_1)\Delta x \\ &\approx F_0 + f(x_0)\Delta x + f(x_1)\Delta x \end{aligned}$$

Using the summation notation, we can generalize the result and write the function $F(x)$ evaluated at any point, $x_N = x_0 + N\Delta x$:

$$F(x_N) \approx F_0 + \sum_{i=1}^{i=N} f(x_{i-1})\Delta x$$

The result above will become exactly correct in the limit $\Delta x \to 0$:

$$F(x_N) = F(x_0) + \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x \tag{B.2}$$

Let us take a closer look at the sum. Each term in the sum is of the form $f(x_{i-1})\Delta x$, and is illustrated in Figure B.6 for the same case as in Figure B.5 (that is, Figure B.6 shows $f(x)$ that we know, and Figure B.5 shows $F(x)$ that we are trying to find).

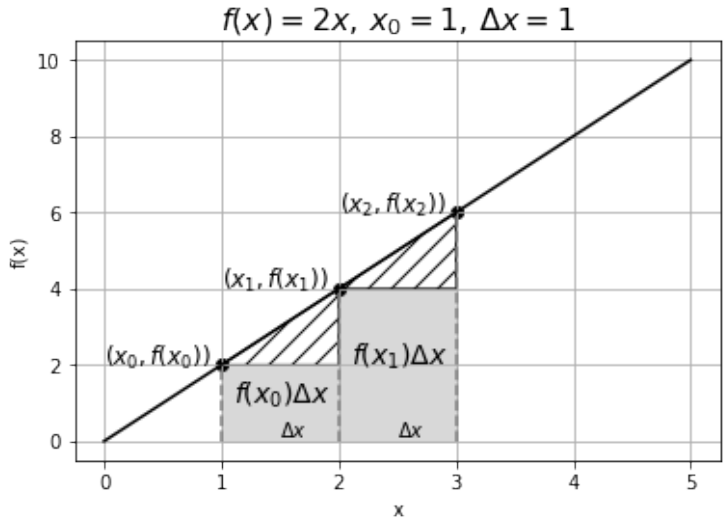


Figure B.6: The function $f(x) = 2x$ and illustration of the terms $f(x_0)\Delta x$ and $f(x_1)\Delta x$ as the area between the curve $f(x)$ and the x axis when $\Delta x \to 0$.

As you can see, each term in the sum corresponds to the area of a rectangle between the function $f(x)$ and the x axis (with a piece missing). In the limit where $\Delta x \to 0$, the missing pieces (shown by the hashed areas in Figure B.6) will vanish and $f(x_i)\Delta x$ will become exactly the area between $f(x)$ and the x axis over a length Δx. The sum of the rectangular areas will thus approach the area between $f(x)$ and the x axis between x_0 and x_N:

$$\lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x = \text{Area between f(x) and x axis from } x_0 \text{ to } x_N$$

Re-arranging equation B.2 gives us a prescription for determining the anti-derivative:

$$F(x_N) - F(x_0) = \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x$$

We see that if we determine the area between $f(x)$ and the x axis from x_0 to x_N, we can obtain the difference between the anti-derivative at two points, $F(x_N) - F(x_0)$

The difference between the anti-derivative, $F(x)$, evaluated at two different values of x is called the **integral** of $f(x)$ and has the following notation:

$$\boxed{\int_{x_0}^{x_N} f(x)dx = F(x_N) - F(x_0) = \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x} \tag{B.3}$$

As you can see, the integral has labels that specify the range over which we calculate the area between $f(x)$ and the x axis. A common notation to express the difference $F(x_N) - F(x_0)$ is to use brackets:

$$\int_{x_0}^{x_N} f(x)dx = F(x_N) - F(x_0) = \Big[F(x)\Big]_{x_0}^{x_N}$$

Recall that we wrote the anti-derivative with the same $\int$ symbol earlier:

$$\int f(x)dx = F(x)$$

The symbol $\int f(x)dx$ without the limits is called the **indefinite integral**. You can also see that when you take the (definite) integral (i.e. the difference between $F(x)$ evaluated at two points), any constant that is added to $F(x)$ will cancel. Physical quantities are always based on definite integrals, so when we write the constant C it is primarily for completeness and to emphasize that we have an indefinite integral.

As an example, let us determine the integral of $f(x) = 2x$ between $x = 1$ and $x = 4$, as well as the indefinite integral of $f(x)$, which is the case that we illustrated in Figures B.5 and B.6. Using equation B.3, we have:

$$\begin{aligned}\int_{x_0}^{x_N} f(x)dx &= \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x \\ &= \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} 2x_{i-1}\Delta x\end{aligned}$$

where we have:

$$\begin{aligned}x_0 &= 1 \\ x_N &= 4 \\ \Delta x &= \frac{x_N - x_0}{N}\end{aligned}$$

Note that N is the number of times we have Δx in the interval between x_0 and x_N. Thus, taking the limit of $\Delta x \to 0$ is the same as taking the limit $N \to \infty$. Let us illustrate the sum for the case where $N = 3$, and thus when $\Delta x = 1$, corresponding to the illustration in Figure B.6:

$$\begin{aligned}\sum_{i=1}^{i=N=3} 2x_{i-1}\Delta x &= 2x_0\Delta x + 2x_1\Delta x + 2x_2\Delta x \\ &= 2\Delta x(x_0 + x_1 + x_2) \\ &= 2\frac{x_3 - x_0}{N}(x_0 + x_1 + x_2) \\ &= 2\frac{(4) - (1)}{(3)}(1 + 2 + 3) \\ &= 12\end{aligned}$$

where in the second line, we noticed that we could factor out the $2\Delta x$ because it appears in each term. Since we only used 4 points, this is a pretty coarse approximation of the integral, and we expect it to be an underestimate (as the missing area represented by the hashed lines in Figure B.6 is quite large).

If we repeat this for a larger value of N, $N = 6$ ($\Delta x = 0.5$), we should obtain a more accurate answer:

$$\begin{aligned}\sum_{i=1}^{i=6} 2x_{i-1}\Delta x &= 2\frac{x_6 - x_0}{N}(x_0 + x_1 + x_2 + x_3 + x_4 + x_5) \\ &= 2\frac{4 - 1}{6}(1 + 1.5 + 2 + 2.5 + 3 + 3.5) \\ &= 13.5\end{aligned}$$

Writing this out again for the general case so that we can take the limit $N \to \infty$, and factoring out the $2\Delta x$:

$$\begin{aligned}\sum_{i=1}^{i=N} 2x_{i-1}\Delta x &= 2\Delta x \sum_{i=1}^{i=N} x_{i-1} \\ &= 2\frac{x_N - x_0}{N} \sum_{i=1}^{i=N} x_{i-1}\end{aligned}$$

Now, consider the combination:

$$\frac{1}{N} \sum_{i=1}^{i=N} x_{i-1}$$

that appears above. This corresponds to the arithmetic average of the values from x_0 to x_{N-1} (sum the values and divide by the number of values). In the limit where $N \to \infty$,

then the value $x_{N-1} \approx x_N$. The average value of x in the interval between x_0 and x_N is simply given by the value of x at the midpoint of the interval:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{i=N} x_{i-1} = \frac{1}{2}(x_N + x_0)$$

Putting everything together:

$$\begin{aligned}
\lim_{N\to\infty} \sum_{i=1}^{i=N} 2x_{i-1}\Delta x &= 2(x_N + x_0) \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{i=N} x_{i-1} \\
&= 2(x_N - x_0)\frac{1}{2}(x_N + x_0) \\
&= x_N^2 - x_0^2 \\
&= (4)^2 - (1)^2 = 15
\end{aligned}$$

where in the last line, we substituted in the values of $x_0 = 1$ and $x_N = 4$. Writing this as the integral:

$$\int_{x_0}^{x_N} 2x dx = F(x_N) - F(x_0) = x_N^2 - x_0^2$$

we can immediately identify the anti-derivative and the indefinite integral:

$$\begin{aligned}
F(x) &= x^2 + C \\
\int 2x dx &= x^2 + C
\end{aligned}$$

This is of course the result that we expected, and we can check our answer by taking the derivative of $F(x)$:

$$\frac{dF}{dx} = \frac{d}{dx}(x^2 + C) = 2x$$

We have thus confirmed that $F(x) = x^2 + C$ is the anti-derivative of $f(x) = 2x$.

Checkpoint B-3

The quantity $\int_a^b f(t)dt$ is equal to

A) the area between the function $f(t)$ and the f axis between $t = a$ and $t = b$
B) the sum of $f(t)\Delta t$ in the limit $\Delta t \to 0$ between $t = a$ and $t = b$
C) the difference $f(b) - f(a)$.

B.3.1 Common anti-derivative and properties

Table B.3 below gives the anti-derivatives (indefinite integrals) for common functions. In all cases, x, is the independent variable, and all other variables should be thought of as constants:

Function, $f(x)$	**Anti-derivative,** $F(x)$
$f(x) = a$	$F(x) = ax + C$
$f(x) = x^n$	$F(x) = \frac{1}{n+1}x^{n+1} + C$
$f(x) = \frac{1}{x}$	$F(x) = \ln(\|x\|) + C$
$f(x) = \sin(x)$	$F(x) = -\cos(x) + C$
$f(x) = \cos(x)$	$F(x) = \sin(x) + C$
$f(x) = \tan(x)$	$F(x) = -\ln(\|\cos(x)\|) + C$
$f(x) = e^x$	$F(x) = e^x + C$
$f(x) = \ln(x)$	$F(x) = x\ln(x) - x + C$

Table B.3: Common indefinite integrals of functions.

Note that, in general, it is much more difficult to obtain the anti-derivative of a function than it is to take its derivative. A few common properties to help evaluate indefinite integrals are shown in Table B.4 below.

Anti-derivative	**Equivalent anti-derivative**
$\int (f(x) + g(x))dx$	$\int f(x)dx + \int g(x)dx$ (sum)
$\int (f(x) - g(x))dx$	$\int f(x)dx - \int g(x)dx$ (subtraction)
$\int af(x)dx$	$a\int f(x)dx$ (multiplication by constant)
$\int f'(x)g(x)dx$	$f(x)g(x) - \int f(x)g'(x)dx$ (integration by parts)

Table B.4: Some properties of indefinite integrals.

B.3.2 Common uses of integrals in Physics - from a sum to an integral

Integrals are extremely useful in physics because they are related to sums. If we assume that our mathematician friends (or computers) can determine anti-derivatives for us, using integrals is not that complicated.

The key idea in physics is that **integrals are a tool to easily performing sums**. As we saw above, integrals correspond to the area underneath a curve, which is found by *summing* the (different) areas of an infinite number of infinitely small rectangles. In physics, it is often the case that we need to take the sum of an infinite number of small things that keep varying, just as the areas of the rectangles.

Consider, for example, a rod of length, L, and total mass M, as shown in Figure B.7. If the rod is uniform in density, then if we cut it into, say, two equal pieces, those two pieces will weigh the same. We can define a "linear mass density", μ, for the rod, as the mass per unit

length of the rod:

$$\mu = \frac{M}{L}$$

The linear mass density has dimensions of mass over length and can be used to find the mass of any length of rod. For example, if the rod has a mass of $M = 5\,\text{kg}$ and a length of $L = 2\,\text{m}$, then the mass density is:

$$\mu = \frac{M}{L} = \frac{(5\,\text{kg})}{(2\,\text{m})} = 2.5\,\text{kg/m}$$

Knowing the mass density, we can now easily find the mass, m, of a piece of rod that has a length of, say, $l = 10\,\text{cm}$. Using the mass density, the mass of the 10 cm rod is given by:

$$m = \mu l = (2.5\,\text{kg/m})(0.1\,\text{m}) = 0.25\,\text{kg}$$

Now suppose that we have a rod of length L that is not uniform, as in Figure B.7, and that does not have a constant linear mass density. Perhaps the rod gets wider and wider, or it has a holes in it that make it not uniform. Imagine that the mass density of the rod is instead given by a function, $\mu(x)$, that depends on the position along the rod, where x is the distance measured from one side of the rod.

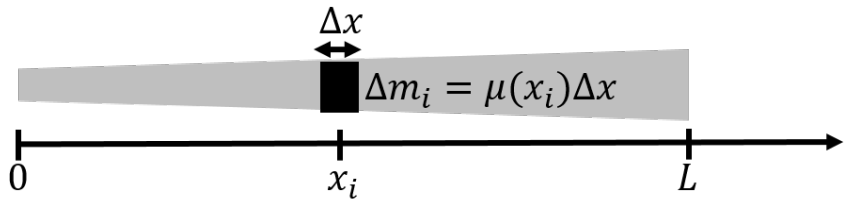


Figure B.7: A rod with a varying linear density. To calculate the mass of the rod, we consider a small mass element Δm_i of length Δx at position x_i. The total mass of the rod is found by summing the mass of the small mass elements.

Now, we cannot simply determine the mass of the rod by multiplying $\mu(x)$ and L, since we do not know which value of x to use. In fact, we have to use all of the values of x, between $x = 0$ and $x = L$.

The strategy is to divide the rod up into N pieces of length Δx. If we label our pieces of rod with an index i, we can say that the piece that is at position x_i has a tiny mass, Δm_i. We assume that Δx is small enough so that $\mu(x)$ can be taken as constant over the length of that tiny piece of rod. Then, the tiny piece of rod at $x = x_i$, has a mass, Δm_i, given by:

$$\Delta m_i = \mu(x_i)\Delta x$$

where $\mu(x_i)$ is evaluated at the position, x_i, of our tiny piece of rod. The total mass, M, of

the rod is then the sum of the masses of the tiny rods, in the limit where $\Delta x \to 0$:

$$\begin{aligned} M &= \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} \Delta m_i \\ &= \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} \mu(x_i) \Delta x \end{aligned}$$

But this is precisely the definition of the integral (equation B.2), which we can easily evaluate with an anti-derivative:

$$\begin{aligned} M &= \lim_{\Delta x \to 0} \sum_{i=1}^{i=N} \mu(x_i) \Delta x \\ &= \int_0^L \mu(x) dx \\ &= G(L) - G(0) \end{aligned}$$

where $G(x)$ is the anti-derivative of $\mu(x)$.

Suppose that the mass density is given by the function:

$$\mu(x) = ax^3$$

with anti-derivative (Table B.3):

$$G(x) = a\frac{1}{4}x^4 + C$$

Let $a = 5\,\mathrm{kg/m^4}$ and let's say that the length of the rod is $L = 0.5\,\mathrm{m}$. The total mass of the rod is then:

$$\begin{aligned} M &= \int_0^L \mu(x) dx \\ &= \int_0^L ax^3 dx \\ &= G(L) - G(0) \\ &= \left[a\frac{1}{4}L^4\right] - \left[a\frac{1}{4}0^4\right] \\ &= 5\,\mathrm{kg/m^4}\frac{1}{4}(0.5\,\mathrm{m})^4 \\ &= 78\,\mathrm{g} \end{aligned}$$

With a little practice, you can solve this type of problem without writing out the sum explicitly. Picture an *infinitesimal* piece of the rod of length dx at position x. It will have an *infinitesimal* mass, dm, given by:

$$dm = \mu(x) dx$$

The total mass of the rod is the then the sum (i.e. the integral) of the mass *elements*

$$M = \int dm$$

and we really can think of the $\int$ sign as a sum, when the things being summed are *infinitesimally* small. In the above equation, we still have not specified the range in x over which we want to take the sum; that is, we need some sort of index for the mass elements to make this a meaningful definite integral. Since we already know how to express dm in terms of dx, we can substitute our expression for dm using one with dx:

$$M = \int dm = \int_0^L \mu(x)dx$$

where we have made the integral definite by specifying the range over which to sum, since we can use x to "label" the mass elements.

One should note that coming up with the above integral is physics. Solving it is math. We will worry much more about writing out the integral than evaluating its value. Evaluating the integral can always be done by a mathematician friend or a computer, but determining which integral to write down is the physicist's job!

B.4 Summary

Key Takeaways

The derivative of a function, $f(x)$, with respect to x can be written as:

$$\frac{d}{dx}f(x) = \frac{df}{dx} = f'(x)$$

and measures the rate of change of the function with respect to x. The derivative of a function is generally itself a function. The derivative is defined as:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Graphically, the derivative of a function represents the slope of the function, and it is positive if the function is increasing, negative if the function is decreasing and zero if the function is flat. Derivatives can always be determined analytically for any continuous function.

A partial derivative measures the rate of change of a multi-variate function, $f(x, y)$, with respect to one of its independent variables. The partial derivative with respect to one of the variables is evaluated by taking the derivative of the function with respect to that variable while treating all other independent variables as if they were constant. The partial derivative of a function (with respect to x) is written as:

$$\frac{\partial f}{\partial x}$$

The gradient of a function, $\nabla f(x, y)$, is a vector in the direction in which that function is increasing most rapidly. It is given by:

$$\nabla f(x, y) = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y}$$

Given a function, $f(x)$, its anti-derivative with respect to x, $F(x)$, is written:

$$F(x) = \int f(x)dx$$

$F(x)$ is such that its derivative with respect to x is $f(x)$:

$$\frac{dF}{dx} = f(x)$$

The anti-derivative of a function is only ever defined up to a constant, C. We usually write this as:

$$\int f(x)dx = F(x) + C$$

since the derivative of $F(x) + C$ will also be equal to $f(x)$. The anti-derivative is also called the "indefinite integral" of $f(x)$.

The definite integral of a function $f(x)$, between $x = a$ and $x = b$, is written:

$$\int_a^b f(x)dx$$

and is equal to the difference in the anti-derivative evaluated at $x = a$ and $x = b$:

$$\int_a^b f(x)dx = F(b) - F(a)$$

where the constant C no longer matters, since it cancels out. Physical quantities only ever depend on definite integrals, since they must be determined without an arbitrary constant.

Definite integrals are very useful in physics because they are related to a sum. Given a function $f(x)$, one can relate the sum of terms of the form $f(x_i)\Delta x$ over a range of values from $x = a$ to $x = b$ to the integral of $f(x)$ over that range:

$$\lim_{\Delta x \to 0} \sum_{i=1}^{i=N} f(x_{i-1})\Delta x = \int_{x_0}^{x_N} f(x)dx = F(x_N) - F(x_0) =$$

B.5 Thinking about the Material

Reflect and research

1. When was calculus first discovered, and by whom?
2. What is an example of a physical quantity that is given by a derivative (other than speed or acceleration)?
3. What is a case when you would need to perform an integral to evaluate a physical quantity?

B.6 Sample problems and solutions

B.6.1 Problems

Problem B-1: You find that the number of customers in your store as a function of time is given by:

$$N(t) = a + bt - ct^2$$

where a, b and c are constants. At what time does your store have the most customers, and what will the number of customers be? (Give the answer in terms of a, b and c). (Solution)

Problem B-2: You measure the speed, $v(t)$, of an accelerating train as function of time,

t, to be given by:

$$v(t) = at + bt^2$$

where a and b are constants. How far does the train move between $t = t_0$ and $t = t_1$? (Solution)

B.6.2 Solutions

Solution to problem B-1: We need to find the value of t for which the function $N(t)$ is maximal. This will occur when its derivative with respect to t is zero:

$$\frac{dN}{dt} = b - 2ct = 0$$
$$\therefore t = \frac{b}{2c}$$

At that time, the number of customers will be:

$$N\left(t = \frac{b}{2c}\right) = a + bt - ct^2$$
$$= a + \frac{b^2}{2c} - \frac{b^2}{4c} = a + \frac{3b^2}{4c}$$

Solution to problem B-2: We are given the speed of the train as a function of time, which is the rate of change of its position:

$$v(t) = \frac{dx}{dt}$$

We need to find how its position, $x(t)$, changes with time, given the speed. In other words, we need to find the anti-derivative of $v(t)$ to get the function for the position as a function of time, $x(t)$:

$$x(t) = \int v(t)dt = \int (at + bt^2)dt$$
$$= \frac{1}{2}at^2 + \frac{1}{3}bt^3 + C$$

where C is an arbitrary constant. The distance covered, Δx, between time t_0 and time t_1 is simply the difference in position at those two times:

$$\Delta x = x(t_1) - x(t_0)$$
$$= \frac{1}{2}at_1^2 + \frac{1}{3}bt_1^3 + C - \frac{1}{2}at_0^2 + \frac{1}{3}bt_0^3 - C$$
$$= \frac{1}{2}a(t_1^2 - t_0^2) + \frac{1}{3}b(t_1^3 - t_0^3)$$

C Guidelines for lab related activities

This chapter introduces the skills that are necessary for thinking about how to design an experiment and to report on its results.

Learning Objectives

- Develop skills in general scientific writing.
- Learn to write scientific proposals and experimental reports.
- Learn to review others' scientific proposals and experimental reports.

C.1 The process of science and the need for scientific writing

Conducting experiments that test a scientific theory is integral to the advancement of science and to the refining of scientific theories. In practice, scientists do not have a lab full of equipment ready to go and to be used for testing whichever theory suits their fancy. Instead, they need to write a "proposal" for conducting a particular experiment to a funding source (e.g. a funding agency). That funding source will then select a panel of experts in the field to review whether the proposal is feasible and useful in advancing science, to decide whether it should be funded. If the scientist is awarded with funds, they are then expected to carry out their experiment and report on the results in a peer-reviewed scientific journal. Again, before the results are published, the scientific journal will ask a panel of experts to review the results to ensure that they are scientifically valid and interesting.

In order for a proposal to be funded, it must thus propose an experiment that is well-thought out and feasible. For example, the reviewers will want to make sure that the proposed experiment is designed in the best possible way to test a theory. Often, this means that thought has been put into designing an experiment that minimizes the uncertainty on the result, so that the test of the theory is as stringent as possible.

A proposal needs to be well-written and precise. We generally call this type of writing "scientific writing", and it is a style of writing that takes some practice. Similarly, when reporting on the results of an experiment, the report will need to be clear and precise as

well. For example, in scientific writing, one avoids giving opinions or using sentences that do not add necessary information or that are not factual.

This chapter provides some guidelines for scientific writing, writing proposals, and writing reports. In addition to this, guidelines for reviewing others' proposals and reports are also presented. Not only is it important to develop the ability to critically evaluate others' work, but it is also helpful in learning to reflect and improve on one's own work.

C.2 Scientific writing

Scientific writing is important in communicating with other scientists. Think of scientific writing as a style of writing where **every word counts**. It makes for rather "dry" reading, but it is important for clearly and precisely communicating factual information. The main guidelines for scientific writing are **be concise, precise, factual, and clear**. Below are some tips to help with scientific writing:

- Avoid subjective/imprecise terms: avoid using subjective and imprecise terms, stick to factual statements and avoid opinions. Instead of saying "our calculated value of g was much greater than the expected value", say "our calculated value of g was greater than the expected value". Your opinion that it was "much greater" does not communicate anything and is imprecise (much greater in relation to what?).
- Definitive statements: avoid attributing definitive causes to your experimental outcomes. You can never prove a theory to be correct, so at most, your results will be consistent with a theory. For example, instead of saying "as the data exhibit, we have detected the Purple Particle", you should state that "the data are consistent with the detection of the Purple Particle".
- Data is the plural of datum. "This data shows" is incorrect, rather, "these data show", or "this set of data shows".
- Active vs. passive voice: when writing scientific papers, it is recommended to use the third person, passive voice. For example, this would mean saying "the drop time for balls at various heights was measured" rather than "we measured the drop time for balls at various heights". However, both passive and active voices are acceptable in scientific writing, as long as it is consistent throughout the text.
- Tense. Generally, for a proposal, you would use the future tense, and you would use the past tense for reporting on your results.

Emma's Thoughts

Writing and editing - how can I be more concise? We've all felt that our writing was lacking at some point or another. Here are some general tips to avoid overall "wordiness" and to increase ease of reading when writing scientifically:

- What would you want to read? Let's say that you wanted to know the strength of Earth's magnetic field, and how it was found, so you decide to do a literature search. Would you choose a brief, succinct article, or a wordy Magnetic Field Manifesto?
- The kindergarten test: If you had to explain your concept to a six year old cousin, how would you break it down in a way that they could understand it? If you can't break it down enough to explain to a six year old, perhaps you need to revisit your own understanding of the concept before writing about it scientifically.
- Avoid unnecessary adjectives: while this might be ok in a creative writing class, in scientific writing, the goal is to get your point across as succinctly as possible. Using "big" words might be ok (as long as they properly describe what you are trying to say), but it is important to communicate your message in the simplest manner.
- Think about it: every time you use a comma, dash or even an "and", you should reconsider the brevity of your statement. In scientific writing, commas are carefully placed, and semicolons are rare.
- Cut it in half: For every word you read, think of another that you can cut. For every sentence that you read, think of three sentences that communicate the same idea. Pick the sentence that is the shortest and most concise.
- Proofread - the more, the better.

The following sections provide basic outlines for writing a proposal and a lab report, as well as rubrics for evaluating/reviewing proposals and reports. Additionally, samples of a proposal, proposal review, report, and report review for the experiment "Measuring g using a pendulum" are provided. In the sample proposal and lab report, errors are purposefully included and addressed in the reviews. It is important to entirely read the rest of this section to capture the common proposal/lab mistakes and their corresponding corrections. That is, do not take the sample proposal as a "perfect proposal", but rather, consider it in the light of the corresponding review.

C.3 Guide for writing a proposal

Summary and Goal

Write a few short sentences briefly summarizing the aim of your experiment, how it will be conducted, and how precise of a result you expect to obtain.

Method and equipment

Clearly describe, in as much detail as required, the method/procedure that you will use to carry out your experiment, and how you will analyse the results. Justify the choices that you made (no need to say you chose to use a ruler because you will need to measure a distance, but perhaps say why you need to measure a given distance, or that you chose to measure something in a particular way as it would reduce the corresponding uncertainty). Provide a list of the equipment that you will need. Also, propose a method of assessing whether or not your project was successful.

Consider the following questions:

- What theory are you testing and through what model?
- How precisely do you estimate that you will be able to make your measurement? Estimate the uncertainty that you will obtain with the proposed experiment. Use this in guiding the design of your experiment.
- What materials, equipment and/or tools are necessary in making your measurements?
- What are the cost of these materials? Can they be easily obtained?
- Where should this experiment be conducted?
- Are there any safety concerns?
- How will you make your measurements? How many times will you make them?
- How will you record your measurements?
- How will you maximize the precision of your experiments?
- How will you determine uncertainties?
- How will you analyse the data?
- What issues could arise in your experiment? How do you plan to resolve these issues?

Timeline and Team

Provide the names of team members, and assign relevant duties to each member. Give a rough outline of the timeline to conduct the experiment, to analyse the data, and to report on the results.

C.4 Guide for reviewing a proposal

Summary

Summarize your overall evaluation of the proposal in 2-3 sentences. Focus on the experiment's methods and goals. For example, "The authors wish to drop balls from different heights to determine the value of g". You don't need to go into the specific details, just give a high level summary of the proposal and your opinion on whether this is a strong proposal. If the proposal is unclear, specify this.

Review

This is where you give your detailed review of the proposal. Consider the following questions:

- Is the proposed experiment well thought-out and feasible?
- Is the experimental procedure clear and concise? Could you could carry out the experiment without asking the authors for additional information? Do the authors specify what instruments to use to measure different quantities and how to determine the associated uncertainties?
- Does the experimental design minimize uncertainties?
- Is it possible to complete the experiment in a reasonable period of time?
- Is it possible to obtain the equipment/materials to conduct the experiment?
- Do the authors describe how to analyse the data (correctly)?
- Does the plan incorporate a mechanism to assess success?
- Is a troubleshooting plan in place, in case of unexpected difficulties?

Overall Rating of the Experiment

Give the proposal an overall score, based on the criteria described above. Use one of the following to rate the proposal and include a sentence to justify your choice.

- Excellent
- Good
- Satisfactory
- Needs work
- Incomplete

C.5 Guide for writing a lab report

Abstract

Write a few short sentences briefly summarizing what you did, how you did it, what you found and whether anything went wrong in your experiment.

Procedure

Describe relevant theories that relate to your experiment here, and the steps to carry out your procedure.

Consider the following questions:

- What are the relevant theories/principles that you used?
- What equations did you use? Show how you modelled your experiment.
- What materials, equipment and/or tools were necessary in making your measurements?
- Where was this experiment conducted?
- How did you make your measurements? How many times did you make them?
- How did you record your measurements?
- How did you determine and minimize the uncertainties in your measurements? Why did you choose to measure a specific quantity in a certain way?

Prediction It can be useful to predict the value (and uncertainty) that you expect to measure before conducting the measurement. You should report on this initial prediction in order to help you better understand the data from your experiment.

Consider the following questions:

- Predict your measured values and uncertainties. How precise do you expect your measurements to be?
- What assumptions did you have to make to predict your results?
- Have these predictions influenced how you should approach your procedure? Make relevant adjustments to the procedure based on your predictions.

Data and Analysis

Present your data. Include relevant tables/graphs. Describe in detail how you analysed the data, including how you propagated uncertainties. If the data do not agree with your model prediction (or the prediction from your proposal), examine whether you can improve your model.

Consider the following questions:

- How did you obtain the "final" measurement/value from your collected data?

- How did you propagate uncertainties? Why did you do it that way?
- What is the relative uncertainty on your value(s)?

Discussion and Conclusion

Summarize your findings, and address whether or not your model described the data. Discuss possible reasons why your measured value is not consisted with your model expectation (is it the model? is it the data?).

Consider the following questions:

- Were there any systematic errors that you didn't consider?
- Did you learn anything that you didn't previously know? (eg. about the subject of your experiment, about the scientific method in general)
- If you could redo this experiment, what would you change (if anything)?

C.5.1 Guide for reviewing a lab report

Summary

Summarize your overall evaluation of the report in 2-3 sentences. Focus on the experiment's method and its result. For example, "The authors dropped balls from different heights to determine the value of g". You don't need to go into the specific details, just give a high level summary of the report. If the report is unclear, specify this.

Review

Consider the following questions:

- Is the the procedure well thought-out, clearly and concisely described?
- Do you have sufficient information that you could repeat this experiment?
- Does the report clearly describe how different quantities were measured and how the uncertainties were determined?
- Does the report motivate why the specific procedure was chosen? (e.g. to minimize uncertainties).
- Does the experiment clearly state how uncertainties were propagated and how the data were analysed?
- Do you believe their result to be scientifically valid?

Overall Rating of the Experiment

Give the report an overall score, based on the criteria described above. Use one of the following to rate the proposal and include a sentence to justify your choice.

- Excellent
- Good
- Satisfactory
- Needs work
- Incomplete

C.6 Sample proposal (Measuring g using a pendulum)

Summary and Goal

One can measure the gravitational constant, g, by measuring the period of a pendulum of a known length, requiring only a string, mass, ruler and timer. Because the experimental design can be easily adjusted and the experiment is simple, the experiment has a high chance of success.

Method and equipment

The period of a pendulum of length L is easily shown to be given by:

$$T = 2\pi\sqrt{\frac{L}{g}}$$

Thus, by measuring the period, T, of a pendulum as well as its length, one can determine the value of g:

$$g = \frac{4\pi^2 L}{T^2}$$

One can carry out the experiment using the following materials:

- a mass
- inextensible string
- a metre stick
- stand to attach string
- cell-phone with timer and slow-motion camera

The materials listed above are all inexpensive and can be easily obtained. It is recommended that the experiment be completed indoors at room temperature, in order to minimize any environmental effects.

One should tie the string to the mass at one end and the stand at the other, and measure the length, L, of the string from the point on the stand to the centre of mass of the mass.

The period of the pendulum is measured by timing how long it takes the pendulum to complete 20 oscillations and dividing that time by 20. This will be more precise than trying to time the period of a single oscillation.

The pendulum should be released from 90°. When releasing the pendulum, the string should

be pulled taught, and the team member's eye that is measuring the angle should be situated parallel to the measuring device.

A slow-motion video will be taken of the pendulum to track the time of the oscillation in order to minimize error due to reaction time. The team member in charge of taking the video will start the video shortly before the pendulum is released. After releasing the pendulum, the team should record 20 oscillations before stopping the pendulum and the video. Data from the video should be entered into a Jupyter Notebook. It is recommended that this measurement be repeated at least 5 times.

The uncertainty in the time should be taken as half of the smallest division of the cell-phone timer, and the uncertainty in the length of the pendulum as half the smallest division of the metre stick used to measure the length of the pendulum.

Foreseeable issues in this experiment may arise when trying to find a string that is optimally inextensible, as any extensibility will cause error in the results. Additionally, being able to measure exactly 90° as the drop-angle for the pendulum could be difficult. In order to correct for this, the team member who is dropping the pendulum must stand directly parallel to the measuring device, minimizing parallax error.

The measure of success will be determined by the uncertainty and precision of the measured value of g. If the measured value of g has a relative uncertainty that is less than 10 %, and is consistent with the accepted value, then one can consider the experiment to have been carried out successfully.

Team and timeline

One should be able to complete the experiment and analysis in approximately 1 hour and 30 minutes with the data being collected in the first 30 minutes. The remainder of the time should be spent processing the data and writing the experimental report. Following the strengths of the members of the team, the following people should be responsible for leading the following tasks, while everyone participates:

- Alice: building the pendulum
- Brice: taking the measurements
- Chloë: analysing the data
- Dennis: writing and formatting

C.7 Sample proposal review (Measuring g using a pendulum)

Summary and Goal

The authors propose to measure the value of g to within 10% by measuring the period of a simple pendulum, using the SHM equations and theory. The proposal is reasonably clear, but lacks some details in how to measure the initial angle of the pendulum. The authors propose to use a an amplitude of 90° for the pendulum, but at such a large angle, the motion is not expected to be SHM, since it is only so at small angles. By using a smaller angle, the experiment has a good chance of being successful in the proposed timeline.

Review

The experimental methods are described clearly and succinctly, with most information clearly stated. For the materials list, it is stated that "a mass" must be used. Here, it should be stated that a small, solid, non-deformable mass should be used to minimize drag and to act as a point mass. The authors refer to a "measuring device" when determining the amplitude of the pendulum, but this is not described. Anyhow, the amplitude of the oscillations in irrelevant for a pendulum in SHM, as long as the amplitude is small.

Most equations are described in the theory section, but it is incorrectly assumed that the period of a pendulum is independent of the drop angle for all angles. The small angle approximation is not expected to apply with an oscillation amplitude of 90°.

No justification is provided for the use of 20 oscillations prior to measuring the period - it may be necessary to iterate on the reason why 20 oscillations was chosen.

The equipment can be easily obtained and is fairly inexpensive. Adequate resources are available to the group to perform this experiment. A clear troubleshooting plan is described and a method for evaluating success is included.

Timeline and team

This experiment is fairly simple and the equipment/setup is not difficult to handle. The proposed team should be qualified to perform this experiment in the proposed amount of time, although I worry a little bit about Dennis, as he seems to be a bit of a menace.

Overall Rating of the Proposal

Good - this proposal was clearly explained and is scientifically sound, apart from the use of a large angle for the oscillations. It was succinctly written, and most components of the experiment were clearly described. A little more detail in the justification for using 20 oscillations is necessary.

C.8 Sample lab report (Measuring g using a pendulum)

Abstract

In this experiment, we measured g by measuring the period of a pendulum of a known length. We measured $g = (7.650 \pm 0.378)\,\mathrm{m/s^2}$. This correspond to a relative difference of 22% with the accepted value ($9.8\,\mathrm{m/s^2}$), and our result is not consistent with the accepted value.

Theory

A pendulum exhibits simple harmonic motion (SHM), which allowed us to measure the gravitational constant by measuring the period of the pendulum. The period, T, of a pendulum of length L undergoing simple harmonic motion is given by:

$$T = 2\pi\sqrt{\frac{L}{g}}$$

Thus, by measuring the period of a pendulum as well as its length, we can determine the value of g:

$$g = \frac{4\pi^2 L}{T^2}$$

We assumed that the frequency and period of the pendulum depend on the length of the pendulum string, rather than the angle from which it was dropped.

Predictions

We built the pendulum with a length $L = (1.0000 \pm 0.0005)\,\mathrm{m}$ that was measured with a ruler with 1 mm graduations (thus a negligible uncertainty in L). We plan to measure the period of one oscillation by measuring the time to it takes the pendulum to go through 20 oscillations and dividing that by 20. The period for one oscillation, based on our value of L and the accepted value for g, is expected to be $T = 2.0\,\mathrm{s}$. We expect that we can measure the time for 20 oscillations with an uncertainty of 0.5 s. We thus expect to measure one oscillation with an uncertainty of 0.025 s (about 1% relative uncertainty on the period). We thus expect that we should be able to measure g with a relative uncertainty of the order of 1%

Procedure

The experiment was conducted in a laboratory indoors.

1. Construction of the pendulum

We constructed the pendulum by attaching a inextensible string to a stand on one end and to a mass on the other end. The mass, string and stand were attached together with knots. We adjusted the knots so that the length of the pendulum was $(1.0000 \pm 0.0005)\,\mathrm{m}$. The uncertainty is given by half of the smallest division of the ruler that we used.

2. Measurement of the period

The pendulum was released from 90° and its period was measured by filming the pendulum with a cell-phone camera and using the phone's built-in time. In order to minimize the uncertainty in the period, we measured the time for the pendulum to make 20 oscillations, and divided that time by 20. We repeated this measurement five times. We transcribed the measurements from the cell-phone into a Jupyter Notebook.

Data and Analysis

Using a 100 g mass and 1.0 m ruler stick, the period of 20 oscillations was measured over 5 trials. The corresponding value of g for each of these trials was calculated. The following data for each trial and corresponding value of g are shown in the table below.

Trial	Angle (Degrees)	Measured Period (s)	Value of g (m/s^2)
1	90	2.24	7.87
2	90	2.37	7.03
3	90	2.28	7.59
4	90	2.26	7.73
5	90	2.22	8.01

Our final measured value of g is $(7.650 \pm 0.378)\,\mathrm{m/s^2}$. This was calculated using the mean of the values of g from the last column and the corresponding standard deviation. The relative uncertainty on our measured value of g is 4.9% and the relative difference with the accepted value of $9.8\,\mathrm{m/s^2}$ is 22%, well above our relative uncertainty.

Discussion and Conclusion

In this experiment, we measured $g = (7.650 \pm 0.378)\,\mathrm{m/s^2}$. This has a relative difference of 22% with the accepted value and our measured value is not consistent with the accepted value. All of our measured values were systematically lower than expected, as our measured periods were all systematically higher than the §2.0s that we expected from our prediction. We also found that our measurement of g had a much larger uncertainty (as determined from the spread in values that we obtained), compared to the 1% relative uncertainty that we predicted.

We suspect that by using 20 oscillations, the pendulum slowed down due to friction, and

this resulted in a deviation from simple harmonic motion. This is consistent with the fact that our measured periods are systematically higher. We also worry that we were not able to accurately measure the angle from which the pendulum was released, as we did not use a protractor.

If this experiment could be redone, measuring 10 oscillations of the pendulum, rather than 20 oscillations, could provide a more precise value of g. Additionally, a protractor could be taped to the top of the pendulum stand, with the ruler taped to the protractor. This way, the pendulum could be dropped from a near-perfect 90° rather than a rough estimate.

C.9 Sample lab report review (Measuring g using a pendulum)

Summary

The authors measured the period of a pendulum to determine g. They measured g to be $(7.650 \pm 0.378)\,\mathrm{m/s^2}$ which is inconsistent with the accepted value. The authors were incorrect in assuming that the pendulum would undergo simple harmonic motion in the conditions that they used.

Review

The experimental procedure was clearly written and one could mostly reproduce this experiment with the given description.

The authors thought about minimizing uncertainties by measuring the period over several oscillations, although it appears that 20 was perhaps too large, as friction was likely to have an effect. The authors should have taken more care in determining the number of oscillations to use so that the uncertainty in the time is minimized while also keeping the effects of friction negligible. Ultimately, the authors did not specify the uncertainty in the time that they measured.

The authors also claim to have measured the length of the pendulum with a precision of 0.5 mm, but did not specify the length of the ruler that they used. I would not expect the measurement to be that precise unless they used a very precise ruler that is longer than 1 m. However, the authors made the length of the pendulum as long as possible so as to minimize the uncertainty in the length.

The authors did not describe the mass that was attached at the end of the pendulum, and whether its size would be expected to cause significant air drag.

The authors made a mistake in assuming that a pendulum would undergo simple harmonic motion with an amplitude of 90°, as the small angle approximation used to determine the period does not apply in this case.

The experimental procedure was scientifically sound, other than the choices for the number of oscillations and their amplitude.

Overall rating of the Experiment

Satisfactory - The experiment was well described, but the authors should have paid more attention to their choice of 20 oscillations, and they made a mistake in assuming that their pendulum would exhibit simple harmonic oscillation with a large amplitude.

D The Python programming language

This appendix gives a very brief introduction to programming in python and is primarily aimed at introducing tools that are useful for the experimental side of physics.

Learning Objectives

- Be able to perform simple algebra using python.
- Be able to plot a function in python.
- Be able to propagate uncertainties in python.
- Be able to plot and fit data to a straight line.
- Understand how to use Python to numerically calculate *any* integral.

In this textbook, we will encourage you to use computers to facilitate making calculations and displaying data. We will make use of a popular programming language called Python, as well as several "modules" from Python that facilitate working with numbers and data. Do not worry if you do not have any programming experience; we assume that you have none and hope that by the end of this book, you will have some capability to decrease your workload by using computer programming.

The only way to become proficient at programming is through practice. If you want to effectively learn from this chapter, it is important that you take the time to actually type the commands into a Python environment rather than simply reading through the chapter. Reading through the chapter will at least give you a sense of what is possible and some terminology, but it will not teach you programming!

D.1 A quick intro to programming

In Python, as in other programming languages, the equals sign is called the **assignment operator**. Its role is to *assign* the value on its right to the variable on its left. The following code does the following:

- *assigns* the value of `2` to the variable `a`
- *assigns* the values of `2*a` to the variable `b`
- prints out the value of the variable `b`

Python Code D.1: Declaring variables in Python

```
#This is a comment, and is ignored by Python
a = 2
b = 2*a
print(b)
```

Output D.1:

```
4
```

Note that any text that follows a pound sign (#) is intended as a comment and will be ignored by Python. Inserting comments in your code is very important for being able to understand your computer program in the future or if you are sharing your code with someone who would like to understand it. In the above example, we called the `print()` **function** and passed to it the variable `b` as an **argument**; this allowed us to print (display) the value of the variable `b` and verify that it was indeed equal to the number 4.

In Python, if you want to have access to "functions", which are a more complex series of operations, then you typically need to load the *module* that defines those operations.

A large number of functions are provided in Python. Most of these functions need to be "imported" from "modules". For example, if you want to be able to take the square root of a number, then you need to load (import) the "math module" which contains the square root function, as in the following example:

Python Code D.2: Using functions from modules

```
#First, we load (import) the math module
import math as m
a = 9
b = m.sqrt(a)
print(b)
```

Output D.2:

```
3
```

In the above code, we loaded the math module (and renamed it `m`); this then allows us to use the functions that are part of that module, including the square root function (`m.sqrt()`).

D.2 Arrays

It is often the case that we need to represent a series of numbers. For example, imagine that you have measured the position of an object as a function of time. **Arrays** are a convenient way to hold a series of numbers that are all alike, for example, all of the values of the position and corresponding time values for the trajectory of the object. In Python, we can define variables that hold arrays instead of a single value (arrays are called "lists" in Python):

Python Code D.3: Arrays in python

```
#define an array of values for the position of the object
position = [0,1,4,9,16,25]
#define an array of values for the corresponding times
time = [0,1,2,3,4,5]
```

D.3 Plotting

Several modules are available in python for plotting. We will show here how to use the `pylab` module (which is equivalent to the `matplotlib` module). For example, we can easily plot the data in the two arrays from the previous section in order to plot the position versus time for the object:

Python Code D.4: Plotting two arrays

```
#import the pylab module
import pylab as pl

#define an array of values for the position of the object
position = [0,1,4,9,16,25]
#define an array of values for the corresponding times
time = [0,1,2,3,4,5]

#make the plot showing points and the line (.-)
pl.plot(time, position, '.-')
#add some labels:
pl.xlabel("time") #label for x-axis
pl.ylabel("position") #label for y-axis
#show the plot
pl.show()
```

Output D.4:

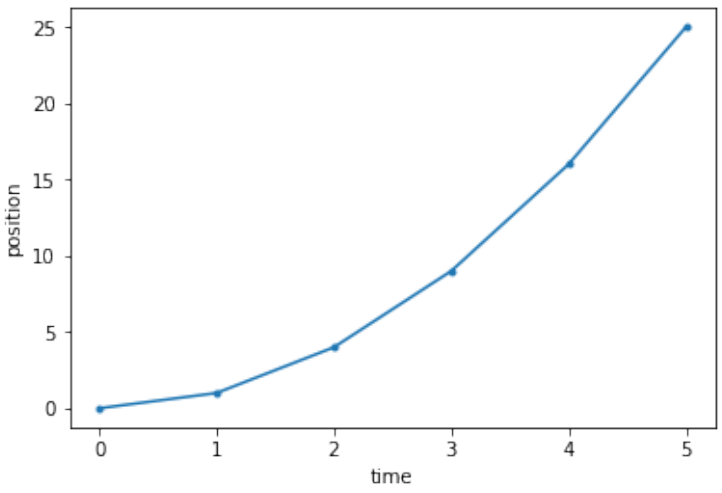


Figure D.1: Using two arrays and plotting them.

Checkpoint D-1

How would you modify the Python code above to show only the points, and not the line?

We can use Python to plot any mathematical function that we like. It is important to realize that computers do not have a representation of a continuous function. Thus, if we would like to plot a continuous function, we first need to evaluate that function at many points, and then plot those points. The `numpy` module provides many useful features for working with arrays of numbers and applying functions directly to those arrays.

Suppose that we would like to plot the function $f(x) = cos(x^2)$ between $x = -3$ and $x = 5$. In order to do this in Python, we will first generate an array of many values of x between -3 and 5 using the `numpy` package and the function `linspace(min,max,N)` which generates N linearly spaced points between min and max. We will then evaluate the function at all of those points to create a second array. Finally, we will plot the two arrays against each other:

Python Code D.5: Plotting a function of 1 variable

```
#import the pylab and numpy modules
import pylab as pl
import numpy as np

#Use numpy to generate 1000 values of x between -3 and 5.
#xvals is an array with 1000 values in it:
xvals =np.linspace(-3,5,1000)

#Now, evaluate the function for all of those values of x.
#We use the numpy version of cos, since it allows us to take the cos
#of all values in the array.
#fvals will be an array with the 1000 corresponding cosines of the xvals
    squared
fvals = np.cos(xvals**2)

#make the plot showing only a line, and color it
pl.plot(xvals, fvals, color='red')
#show the plot
pl.show()
```

Output D.5:

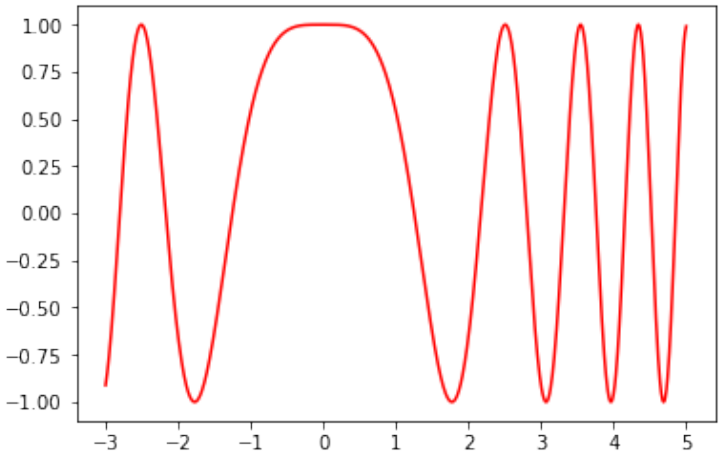


Figure D.2: Plotting a function using arrays.

D.4 The QExpy python package for experimental physics

QExpy is a Python module that was developed with students in Queen's University to handle all aspects of undergraduate physics laboratories. In this section, we look at how to use QExpy to propagate uncertainties and to plot experimental data.

D.4.1 Propagating uncertainties

In Chapter ??, we saw how to use the "derivative method" to propagate the uncertainty from measurements into the uncertainty in a value that depended on those measurements. In Example **??**, we propagated the uncertainties $x = (3.00 \pm 0.01)\,\text{m}$ and $t = (0.76 \pm 0.15)\,\text{s}$ to the quantity $k = \frac{t}{\sqrt{x}}$. We show below how easily this can be done with QExpy:

Python Code D.6: QExpy to propagate uncertainties

```
#First, we load the QExpy module
import qexpy as q
#Now define our measurements with uncertainties:
t = q.Measurement(0.76, 0.15) # 0.76 +/- 0.15
x = q.Measurement(3,0.1) # 3 +/- 0.1
#Now define k, which depends on t and x:
k = t/q.sqrt(x) # use the QExpy version of sqrt() since x is of type
    Measurement
#Print the result:
print(k)
```

Output D.6:

```
0.44 +/- 0.09
```

which is the result that we obtained when manually applying the derivative method. Note that we used the square root function from the QExpy module, as it "knows" how to take the square root of a value with uncertainty (a "Measurement" in the language of QExpy).

We also saw that when we had repeated measurements of the same quantity (Section **??**), one could define a central value and uncertainty for that quantity by using the mean and standard deviations of the measurements. QExpy can easily take a set of measurements (an array of values) and convert them into a single quantity (a "Measurement") with a central value and uncertainty that correspond to the mean and standard deviation of the set of measurements:

Python Code D.7: QExpy to calculate mean and standard deviation

```
#First, we load the QExpy module
import qexpy as q
#We define $t$ as an array of values (note the square brackets):
t = q.Measurement([1.01, 0.76, 0.64, 0.73, 0.66])
#Choose the number of significant figures to print:
q.set_sigfigs(2)
#Print the result:
print("t = ",t)
```

Output D.7:

```
t = 0.76 +/- 0.15
```

By using QExpy, we do not need to tediously calculate the mean and standard deviation,

as we had in Example ??.

D.4.2 Plotting experimental data with uncertainties

In Chapter ?? we had presented the data in Table D.1 which corresponded to our measurements of how long it took (t) for an object to drop a certain distance, x. We had also introduced Chloë's Theory of gravity that predicted that the data should be described by the following model:

$$t = k\sqrt{x}$$

where k was an undetermined constant of proportionality.

x [m]	**t** [s]	$\sqrt{x}$ [$\mathrm{m}^{\frac{1}{2}}$]	**k** [$\mathrm{s\,m}^{-\frac{1}{2}}$]
1.00	0.33	1.00	0.33
2.00	0.74	1.41	0.52
3.00	0.67	1.73	0.39
4.00	1.07	2.00	0.54
5.00	1.10	2.24	0.49

Table D.1: Measurements of the drop times, t, for a bowling ball to fall different distances, x. We have also computed $\sqrt{x}$ and the corresponding value of k.

The easiest way to visualize and analyse those data is to plot them. In particular, if we plot (graph) t versus $\sqrt{x}$, we expect that the points will fall on a straight line that goes through zero, with a slope of k (if the data are described by Chloë's Theory). We can use QExpy to graph the data as well as determine ("fit") for the slope of the line that best describes the data, since we expect that the slope will correspond to the value of k. When plotting data and fitting them to a line (or other function), it is important to make sure that the values have at least an uncertainty in the quantity that is being plotted on the y axis. In this case, we have assumed that all of the measurements of time have an uncertainty of 0.15 s and that the measurements of the distance have no (or negligible) uncertainties. The python code below shows how to use QExpy to plot and fit the data to a straight line.

Python Code D.8: Using QExPy to plot and fit linear data

```
#First, we load the QExpy module:
import qexpy as q

#Use matplotlib as the plot engine (try using 'bokeh' instead of 'mpl')
q.plot_engine = 'mpl'

#Set the number of significant figures to 2:
q.set_sigfigs(2)

#Then we enter the data:
#start with the values for the square root of height:
sqx = [1. , 1.41, 1.73, 2., 2.24]
#and then, the corresponding times:
t = [ 0.33,  0.74,  0.67,  1.07,  1.1 ]

#Let us attribute an uncertainty of 0.15 to each measured values of t:
```

```
terr = 0.15

#We now make the plot. First, we create the plot object with the data
#Note that x and y refer to the x and y axes
fig = q.MakePlot( xdata = sqx, xname = "sqrt(distance) [m^0.5]",
                  ydata = t, yerr = terr, yname = "time [s]",
                  data_name = "My data")

#Ask QExpy to also determine the line of best fit
fig.fit("linear")

#Then, we show it:
fig.show()
```

Output D.8:

```
------------------Fit results-----------------------
Fit of My data to linear
Fit parameters:
My data_linear_fit0_fitpars_intercept = -0.24 +/- 0.22,
My data_linear_fit0_fitpars_slope = 0.61 +/- 0.13

Correlation matrix:
[[ 1.     -0.968]
 [-0.968   1.   ]]

chi2/ndof = 2.04/2
------------------End fit results-------------------
```

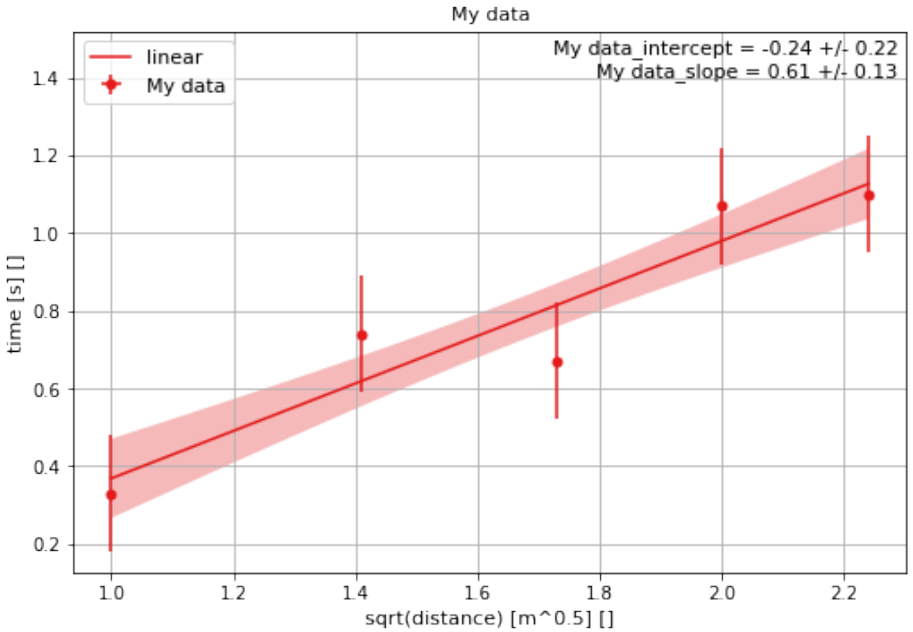


Figure D.3: QExpy plot of t versus $\sqrt{x}$ and line of best fit.

The plot in Figure D.3 shows that the data points are consistent with falling on a straight line, when their error bars are taken into account. We've also asked QExpy to show us

the line of best fit to the data, represented by the line with the shaded area. When we asked for the line of best fit, QExpy not only drew the line, but also gave us the values and uncertainties for the slope and the intercept of the line. The shaded area around the line corresponds to other possible lines that one would obtain using different values of the slope and intercept within their corresponding uncertainties. The output also provides a line that tells us that `chi2/ndof = 2.04/2`; although you do not need to understand the details, this is a measure of how well the data are described by the line of best fit. Generally, the fit is assumed to be "good" if this ratio is close to 1 (the ratio is called "the reduced chi-squared"). The "correlation matrix" tells us how the best fit value of the slope is linked to the best fit value of the intercept, which you do not need to worry about here.

Since we expect the slope of the data to be k, this provides us a method to determine k from the data as $(0.61 \pm 0.13)\,\mathrm{s\,m^{-\frac{1}{2}}}$. **Performing a linear fit of the data is the best way to determine a constant of proportionality between the measurements**. Finally, we expect the intercept to be equal to zero according to our model. The best fit line from QExpy has an intercept of $(-0.24 \pm 0.22)\,\mathrm{s}$, which is slightly below, but consistent, with zero. From these data, we would conclude that the measurements are consistent with Chloë's Theory.

D.5 Advanced topics

This section introduces a few more advanced topics that allow you to use computer programming to simplifying many tasks. In this section, we will show you how you can write your own program to numerically estimate the value of an integral of any function.

D.5.1 Defining your own functions

Although Python provides many modules and functions, it is often useful to be able to define your own functions. For example, suppose that you would like to define a function that calculates $\frac{1}{3}x^2 + \frac{1}{4}x^3 + \cos(2x)$, for a given value of x. This is done easily using the `def` keyword in Python:

Python Code D.9: Defining a function

```
#import the math module in order to use cos
import math as m

#define our function and call it myfunction:
def myfunction(x):
  return x**2 / 3 + x**3 / 4 + m.cos(2*x)

#Test our function by printing out the result of evaluating it at x = 3
print( myfunction(3) )
```

Output D.9:

```
10.710170286650365
```

A few things to note about the code above:

- Functions are defined using the `def` keyword followed by the name that we choose for the function (in our case, `myfunction`)
- If functions take arguments, those are specified in parenthesis after the name of the function (in our case, we have one argument that we chose to call `x`)
- After the name of the function and the arguments, we place a colon

- The code that belongs to the function, after the colon, must be indented (this allows Python to know where the code for the function ends)
- The function can "return" a value; this is done by using the `return` keyword.
- We used the "operator" `**` to take the power of a number (`x**2`), and the operator `*`, to multiply numbers. Python would not understand something like `2x`; you need to use the multiplication operator, i.e. `2*x`.

In the example above, we wrote a Python function to represent a mathematical function. However, one can write a function to execute any set of tasks, not just to apply a mathematical function. Python functions are very useful in order to avoid having to repeatedly type the same code.

Recall that the `numpy` module allows us to apply functions to arrays of numbers, instead of a single number. We can modify the code above slightly so that, if the argument to the function, `x`, is an array, the function will gracefully return an array of numbers to which the function has been applied. This is done by simply replacing the call to the `math` version of the `cos` function by using the `numpy` version:

Python Code D.10: Defining a function that works on an array

```
#import the numpy module in order to use cos to an array
import numpy as np

#define our function and call it myfunction:
def myfunction(x):
  return x**2 / 3 + x**3 / 4 + np.cos(2*x)

#Test our function by printing out the result of evaluating it at x = 3 (same
    as before)
print( myfunction(3) )

#Test it with an array
xvals = np.array([1,2,3])
print ( myfunction(xvals) )
```

Output D.10:

```
10.710170286650365
[ 0.1671865   2.67968971 10.71017029]
```

where we created the array `xvals` using the `numpy` module.

D.5.2 Using a loop to calculate an integral

The ability to define our own functions in Python allows us to easily simplify complex tasks. Using "loops" is another way that computer programming can greatly simplify calculations that would otherwise be very tedious. In a loop, one is able to repeat the same task many times. The example below simply prints out a statement five times:

Python Code D.11: A simple loop

```
#A loop to print out a statement 5 times:

for i in range(5):
  print("The value of i is ",i)
```

Output D.11:

```
The value of i is  0
The value of i is  1
The value of i is  2
The value of i is  3
The value of i is  4
```

A few notes on the code above:

- The loop is defined by using the keywords `for ... in`
- The value after the keyword `for` is the "iterator" variable and will have a different value each time that the code inside of the loop is run (in our case, we called the variable `i`)
- The value after the keyword `in` is an array of values that the iterator will take
- The `range(N)` function returns an array of `N` integer values between 0 and `N-1` (in our case, this returns the five values 0,1,2,3,4)
- The code to be executed at each "iteration" of the loop is preceded by a colon and indented (in the same way as the code for a function also follows a colon and is indented)

We now have all of the tools to evaluate an integral numerically. Recall that the integral of the function $f(x)$ between x_a and x_b is simply a sum:

$$\int_{x_a}^{x_b} f(x)dx = \lim_{\Delta x \to 0} \sum_{i=0}^{i=N-1} f(x_i)\Delta x$$
$$\Delta x = \frac{x_b - x_a}{N}$$
$$x_i = x_a + i\Delta x$$

The limit of $\Delta x \to 0$ is equivalent to the limit $N \to \infty$. Our strategy for evaluating the integral is:

1. Define a Python function for $f(x)$.
2. Create an array, `xvals`, of N values of x between x_a and x_b.
3. Evaluate the function for all those values and store those into an array, `fvals`.
4. Loop over all of the values in the array `fvals`, multiply them by Δx, and sum them together.

Let's use Python to evaluate the integral of the function $f(x) = 4x^3 + 3x^2 + 5$ between $x = 1$ and $x = 5$:

Python Code D.12: Numerical integration of a function

```
#import numpy to work with arrays:
import numpy as np
```

```
#define our function
def f(x):
  return 4*x**3 + 3*x**2 + 5

#Make N and the range of integration variables:
N = 1000
xmin = 1
xmax = 5

#create the array of values of x between xmin and xmax
xvals = np.linspace(xmin, xmax, N)

#evaluate the function at all those values of x
fvals = f(xvals)

#calculate delta x
deltax = (xmax - xmin) / N

#initialize the sum to be zero:
sum = 0

#loop over the values fvals and add them to the sum
for fi in fvals:
  sum = sum + fi*deltax

#print the result:
print("The integral between {} and {} using {} steps is {:.2f} ".format(xmin,
    xmax, N, sum))
```

Output D.12:

```
The integral between 1 and 5 using 1000 steps is 768.42
```

One can easily integrate the above function analytically and obtain the exact result of 768. The numerical answer will approach the exact answer as we make N bigger. Of course, the power of numerical integration is to use it when the function cannot be integrated analytically.

Checkpoint D-2

What value of N should you use above in order to get within 0.01 of the exact analytic answer?