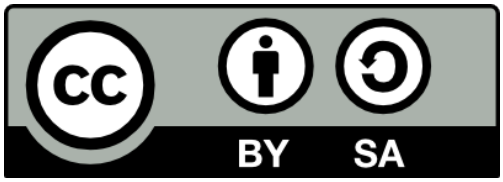
INTRODUCTORY

PHYSICS

Building Models to Describe Our World

Ryan Martin • Emma Neary • Joshua Rinaldo • 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.

Joshua Rinaldo I am a third year physics major and concurrent education student. I was first introduced to the flipped classroom approach in Ryan Martin's first year physics class, and have found that the experience shaped the way I approach education. I intend on continuing to make use of the flipped classroom approach as I move forward in my career. Being able to co-author this textbook has been an amazing opportunity for me to grow as an educator, and I look forward to applying the skills I learned while working on the textbook. Outside of physics, I enjoy making jewelry and practicing mixed martial arts.

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
Alex Hughes
Alexis Brossard
Allyson Smish
Amy Van Nest
Camren Oakes
Ceaira Heimstra
Damara Gagnier
Daniel Barake
Daniel Tazbaz
David Cutler
Emily Darling
Emily Mendelson
Emily Wener
Emma Lanciault
Erin Parson
Genevieve Fawcett
Gregory Love
Haoyuan Wang
Ian McClean
Jack Fitzgerald
James Godfrey
Jenna Vanker
Jesse Fu
Jesse Simmons
Jessica Grennan
Joanna Fu
Jonathan Abott
Kate Fenwick
Lily Dodd
Madison Facchini
Marie Vidal
Matt Routliffe
Maya Gibb
Natalie Dubas
Nathan Wilson
Neil Rajan
Nicholas Everton
Nick Brown
Nicole Gaul
Noah Rowe
Olivia Bouaban
Patrick Singal
Qiqi Zhang
Quentin Sanders
Robin Joshi
Ryan Underwood
Sam Connolly
Sara Stephens
Sarmund Mahmood
Shaundra Buelow
Shona Birkett
Stephanie Ciccone
Talia Castillo
Tamy Puniani
Thomas Faour
Troy Allen
Tashifa Imtiaz
Wei Zhuolin
Yannick Bisson
Yumian Chen
Zifeng Chen
Zoe Macmillan

Contents

1 The Scientific Method and Physics **2**
- 1.1 Science and the Scientific Method . . . 2
- 1.2 Theories, hypotheses and models . . . 5
- 1.3 Fighting intuition . . . 6
- 1.4 The scope of Physics . . . 6
  - 1.4.1 Classical Physics . . . 7
    - Mechanics . . . 7
    - Electromagnetism . . . 8
  - 1.4.2 Modern Physics . . . 8
    - Quantum mechanics and particle physics . . . 8
    - The Special and General Theories of Relativity . . . 9
    - Cosmology and astrophysics . . . 9
    - Particle astrophysics . . . 9
- 1.5 Thinking like a physicist . . . 10
- 1.6 Summary . . . 11
- 1.7 Thinking about the Material . . . 11
- 1.8 Sample problems and solutions . . . 12
  - 1.8.1 Problems . . . 12
  - 1.8.2 Solutions . . . 13

2 Comparing Model and Experiment **14**
- 2.1 Orders of magnitude . . . 14
- 2.2 Units and dimensions . . . 16
  - 2.2.1 Base dimensions and their SI units . . . 17
  - 2.2.2 Dimensional analysis . . . 18
- 2.3 Making measurements . . . 23
  - 2.3.1 Measurement uncertainties . . . 24
    - Determining the central value and uncertainty . . . 25
    - Random and systematic sources of error/uncertainty . . . 27
    - Propagating uncertainties . . . 29
  - 2.3.2 Using graphs to visualize and analyse data . . . 32
  - 2.3.3 Reporting measured values . . . 33
  - 2.3.4 Comparing model and measurement - discussing a result . . . 34
- 2.4 Summary . . . 36
- 2.5 Thinking about the material . . . 38
- 2.6 Sample problems and solutions . . . 38
  - 2.6.1 Problems . . . 38
  - 2.6.2 Solutions . . . 39

3 Describing motion in one dimension **40**
- 3.1 Motion with constant speed . . . 41

3.2 Motion with constant acceleration . . . 45
3.2.1 Visualizing motion with constant acceleration . . . 47
3.3 Using calculus to describe motion . . . 48
3.3.1 Instantaneous and average velocity . . . 48
3.3.2 Using calculus to obtain acceleration from position . . . 50
3.3.3 Using calculus to obtain position from acceleration . . . 50
3.4 Relative motion . . . 53
3.5 Summary . . . 58
3.6 Thinking about the material . . . 60
3.7 Sample Problems and Solutions . . . 61
3.7.1 Problems . . . 61
3.7.2 Solutions . . . 62

4 Describing motion in multiple dimensions 67
4.1 Motion in two dimensions . . . 67
4.1.1 Using vectors to describe motion in two dimensions . . . 67
4.1.2 Relative motion . . . 75
4.2 Motion in three dimensions . . . 77
4.3 Accelerated motion when the velocity vector changes direction . . . 78
4.4 Circular motion . . . 82
4.4.1 Period and frequency . . . 87
4.5 Summary . . . 91
4.6 Thinking about the material . . . 95
4.7 Sample problems and solutions . . . 96
4.7.1 Problems . . . 96
4.7.2 Solutions . . . 97

5 Newton's Laws 101
5.1 Newton's Three Laws . . . 101
5.1.1 Newton's First Law . . . 102
5.1.2 Newton's Second Law . . . 103
5.1.3 Newton's Third Law . . . 104
5.2 Force . . . 104
5.2.1 Types of forces . . . 105
Weight . . . 106
Normal forces . . . 107
Frictional forces . . . 107
Tension forces . . . 109
Drag forces . . . 110
Spring forces . . . 110
Inertial forces . . . 111
"Applied" forces . . . 112
5.3 Mass and inertia . . . 112
5.4 Applying Newton's Laws . . . 112
5.4.1 Identifying the forces . . . 113

5.4.2 Free body diagrams . . . 116
5.4.3 Using Newton's Second Law . . . 118
5.5 The acceleration due to gravity . . . 123
5.6 Non-inertial frames of reference and inertial forces . . . 124
5.7 Summary . . . 129
5.8 Thinking about the material . . . 132
5.9 Sample problems and solutions . . . 133
5.9.1 Problems . . . 133
5.9.2 Solutions . . . 134

6 Applying Newton's Laws 138
6.1 Statics . . . 138
6.2 Linear motion . . . 141
6.2.1 Modelling situations where forces change magnitude . . . 146
6.3 Uniform circular motion . . . 155
6.3.1 Banked curves . . . 161
6.3.2 Inertial forces in circular motion . . . 164
6.4 Non-uniform circular motion . . . 166
6.5 Summary . . . 170
6.6 Thinking about the material . . . 171
6.6.1 Problems and Solutions . . . 172
6.6.2 Solutions . . . 173

7 Work and energy 175
7.1 Work . . . 175
7.1.1 Work in one dimension. . . . 177
7.1.2 Work in one dimension - varying force . . . 178
7.1.3 Work in multiple dimensions . . . 180
7.1.4 Net work done . . . 189
7.2 Kinetic energy and the work energy theorem . . . 194
7.3 Power . . . 200
7.4 Summary . . . 203
7.5 Thinking about the material . . . 205
7.6 Sample problems and solutions . . . 206
7.6.1 Problems . . . 206
7.6.2 Solutions . . . 207

8 Potential Energy and Conservation of Energy 212
8.1 Conservative forces . . . 213
8.2 Potential energy . . . 216
8.2.1 Recovering the force from potential energy . . . 221
8.3 Mechanical energy and conservation of energy . . . 222
8.4 Energy diagrams and equilibria . . . 230
8.5 Advanced Topic: The Lagrangian formulation of classical physics . . . 234
8.6 Summary . . . 237

8.7 Thinking about the material . . . 240
8.8 Sample problems and solutions . . . 241
8.8.1 Problems . . . 241
8.8.2 Solutions . . . 243

9 Gravity **247**
9.1 Kepler's Laws . . . 247
9.1.1 Kepler's First Law . . . 248
9.1.2 Kepler's Second Law . . . 249
9.1.3 Kepler's Third Law . . . 250
9.2 Newton's Universal Theory of Gravity . . . 250
9.2.1 Weight and apparent weight . . . 255
Effects of Earth's rotation . . . 256
9.2.2 The gravitational field . . . 259
9.2.3 Gauss' Law . . . 261
9.3 Gravitational potential energy . . . 265
9.3.1 Mechanical energy with gravity . . . 267
Types of orbits . . . 270
9.4 Einstein's Theory of General Relativity . . . 271
9.5 Summary . . . 274
9.6 Thinking about the material . . . 277
9.7 Sample problems and solutions . . . 278
9.7.1 Problems . . . 278
9.7.2 Solutions . . . 279

10 Linear momentum and the centre of mass **284**
10.1 Momentum . . . 284
10.1.1 Momentum of a point particle . . . 284
10.1.2 Impulse . . . 286
10.1.3 Systems of particles: internal and external forces . . . 290
10.1.4 Conservation of momentum . . . 292
10.2 Collisions . . . 296
10.2.1 Inelastic collisions . . . 296
10.2.2 Elastic collisions . . . 300
10.2.3 Frames of reference . . . 305
10.3 The centre of mass . . . 308
10.3.1 The centre of mass for a continuous object . . . 315
10.4 Summary . . . 320
10.5 Thinking about the material . . . 324
10.6 Sample problems and solutions . . . 325
10.6.1 Problems . . . 325
10.6.2 Solutions . . . 326

11 Rotational dynamics **331**
11.1 Rotational kinematic vectors . . . 331

11.1.1 Scalar rotational kinematic quantities . . . 331
11.1.2 Vector rotational kinematic quantities . . . 333
11.2 Rotational dynamics for a single particle . . . 337
11.3 Torque . . . 341
11.4 Rotation about an axis versus rotation about a point . . . 343
11.5 Rotational dynamics for a solid object . . . 346
11.6 Moment of inertia . . . 351
11.6.1 The parallel axis theorem . . . 353
11.7 Equilibrium . . . 355
11.7.1 Static equilibrium . . . 355
11.7.2 Dynamic equilibrium . . . 358
11.8 Summary . . . 361
11.9 Thinking about the material . . . 365
11.10 Sample problems and solutions . . . 366
11.10.1 Problems . . . 366
11.10.2 Solutions . . . 367

12 Rotational energy and momentum **370**
12.1 Rotational kinetic energy of an object . . . 370
12.1.1 Work on a rotating object . . . 372
12.1.2 Total kinetic energy of an object . . . 374
12.2 Rolling motion . . . 375
12.2.1 The instantaneous axis of rotation . . . 380
12.3 Angular momentum . . . 384
12.3.1 Angular momentum of a particle . . . 384
12.3.2 Angular momentum of an object or system . . . 388
12.3.3 Conservation of angular momentum . . . 390
12.4 Summary . . . 394
12.5 Thinking about the material . . . 398
12.5.1 Reflect and research . . . 398
12.6 Sample problems and solutions . . . 399
12.6.1 Problems . . . 399
12.6.2 Solutions . . . 400

13 Simple harmonic motion **404**
13.1 The motion of a spring-mass system . . . 404
13.1.1 Description using energy . . . 405
13.1.2 Kinematics of simple harmonic motion . . . 406
13.1.3 Analogy with uniform circular motion . . . 410
13.2 Vertical spring-mass system . . . 412
13.2.1 Two-spring-mass system . . . 413
13.3 Simple harmonic motion . . . 415
13.4 The motion of a pendulum . . . 416
13.4.1 The physical pendulum . . . 418
13.5 Summary . . . 420

13.6 Thinking about the material . . . 423
13.7 Sample problems and solutions . . . 424
13.7.1 Problems . . . 424
13.7.2 Solutions . . . 426

14 Waves **430**
14.1 Characteristics of a wave . . . 430
14.1.1 Definition and types of waves . . . 430
14.1.2 Description of a wave . . . 433
14.2 Mathematical description of a wave . . . 435
14.2.1 The wave equation . . . 437
14.3 Waves on a rope . . . 438
14.3.1 A pulse on a rope . . . 439
14.3.2 Reflection and transmission . . . 440
14.3.3 The wave equation for a rope . . . 442
14.4 The speed of a wave . . . 444
14.5 Energy transported by a wave . . . 445
14.5.1 A wave as being made of simple harmonic oscillators . . . 445
14.5.2 Energy transported in a one dimensional wave . . . 446
14.5.3 Energy transported in a spherical, three-dimensional, wave . . . 447
14.6 Superposition of waves and interference . . . 450
14.7 Standing waves . . . 453
14.7.1 Mathematical description of a standing wave . . . 455
14.8 Summary . . . 459
14.9 Thinking about the material . . . 463
14.10 Sample problems and solutions . . . 464
14.10.1 Problems . . . 464
14.10.2 Solutions . . . 466

15 Fluid mechanics **471**
15.1 Pressure . . . 472
15.1.1 The effect of gravity . . . 474
15.1.2 Pascal's Principle . . . 478
15.1.3 Measuring pressure . . . 480
15.2 Buoyancy . . . 483
15.3 Hydrodynamics . . . 487
15.3.1 Continuity of flow . . . 487
15.3.2 Bernoulli's Principle . . . 489
15.3.3 Viscosity . . . 496
15.3.4 Poiseuille flow . . . 497
15.4 Summary . . . 501
15.5 Thinking about the material . . . 504
15.6 Sample problems and solutions . . . 506
15.6.1 Problems . . . 506
15.6.2 Solutions . . . 508

16 Electric charges and fields **511**
16.1 Electric charge . . . 511
16.1.1 Conductors and insulators . . . 513
16.1.2 Electrostatic induction . . . 513
16.2 The Coulomb force . . . 515
16.3 The electric field . . . 519
16.3.1 Visualizing the electric field . . . 522
16.3.2 Electric field from a charge distribution . . . 524
16.4 The electric dipole . . . 537
16.5 Summary . . . 539
16.6 Thinking about the material . . . 543
16.7 Sample problems and solutions . . . 544
16.7.1 Problems . . . 544
16.7.2 Solutions . . . 545

17 Gauss' Law **547**
17.1 Flux of the electric field. . . . 547
17.1.1 Non-uniform fields . . . 549
17.1.2 Closed surfaces . . . 551
17.2 Gauss' Law . . . 554
17.3 Charges in a conductor . . . 565
17.4 Interpretation of Gauss' Law and vector calculus . . . 568
17.5 Summary . . . 570
17.6 Thinking about the material . . . 573
17.7 Sample problems and solutions . . . 574
17.7.1 Problems . . . 574
17.7.2 Solutions . . . 575

18 Electric potential **579**
18.1 Electric potential energy . . . 579
18.1.1 Electrostatic potential energy . . . 581
18.2 Electric potential . . . 582
18.2.1 Electric potential from electric field . . . 586
18.2.2 Electric field from electric potential . . . 591
18.2.3 Equipotential surfaces . . . 592
18.3 Calculating electric potential from charge distributions . . . 594
18.4 Electric field and potential at the surface of a conductor . . . 598
18.5 Capacitors . . . 601
18.5.1 Capacitance . . . 601
18.5.2 Dielectric materials . . . 602
18.5.3 Energy stored in a capacitor . . . 603
18.6 Summary . . . 605
18.7 Thinking about the material . . . 609
18.8 Sample problems and solutions . . . 610
18.8.1 Problems . . . 610

18.8.2 Solutions . . . 611

19 Electric current **615**
19.1 Current . . . 615
19.2 Microscopic model of current . . . 618
19.3 Ohm's Law . . . 621
19.3.1 Resistivity . . . 622
19.4 Resistors . . . 623
19.4.1 Resistance . . . 624
19.4.2 Combining resistors . . . 625
19.4.3 Electrical power dissipated in resistors . . . 628
19.4.4 Superconductors . . . 630
19.5 Alternating voltages and currents . . . 630
19.6 Electrical safety . . . 631
19.7 Summary . . . 634
19.8 Thinking about the material . . . 640
19.9 Sample problems and solutions . . . 641
19.9.1 Problems . . . 641
19.9.2 Solutions . . . 642

20 Electric circuits **643**
20.1 Batteries and simple circuits . . . 643
20.1.1 The electrochemical cell . . . 643
20.1.2 The ideal battery in a circuit . . . 645
20.1.3 The real battery in a circuit . . . 649
20.2 Kirchhoff's rules . . . 652
20.2.1 Junction rule . . . 653
20.2.2 Loop rule . . . 654
20.3 Applying Kirchhoff's rule to model circuits . . . 655
20.4 Measuring current and voltage . . . 664
20.4.1 The ammeter . . . 665
20.4.2 The voltmeter . . . 666
20.5 Modelling circuits with capacitors . . . 668
20.6 Summary . . . 670
20.7 Thinking about the material . . . 673
20.8 Sample problems and solutions . . . 674
20.8.1 problems . . . 674
20.8.2 Solutions . . . 675

21 The magnetic force **678**
21.1 Magnetic fields . . . 678
21.2 The magnetic force on a moving charge . . . 681
21.3 The magnetic force on a current-carrying wire . . . 685
21.4 The torque on a current-carrying loop . . . 690
21.4.1 Magnetic dipole moment . . . 691

21.4.2 Potential energy for a magnetic moment in a magnetic field 694
21.5 The Hall Effect 694
21.6 Applications 696
21.6.1 Velocity selector and mass spectrometer 696
21.6.2 Galvanometer 697
21.6.3 Electric motor 698
21.6.4 Cathode ray tube 699
21.6.5 Loudspeaker 699
21.7 Summary 701
21.8 Thinking about the material 704
21.9 Sample problems and solutions 705
21.9.1 Problems 705
21.9.2 Solutions 706

22 Source of magnetic field **710**
22.1 The Biot-Savart Law 710
22.1.1 Magnetic field from a straight current-carrying wire 711
22.1.2 Magnetic field from a circular current-carrying wire 714
22.2 Force between two current-carrying wires 717
22.3 Ampère's Law 719
22.3.1 Interpretation of Ampère's Law and vector calculus 723
22.4 Solenoids and toroids 725
22.5 Summary 730
22.6 Thinking about the material 733
22.7 Sample problems and solutions 734
22.7.1 Problems 734
22.7.2 Solutions 735

23 Electromagnetic Induction **737**
23.1 Faraday's Law 737
23.1.1 Lenz' Law 739
23.2 Induction in a moving conductor 742
23.2.1 Motion of a bar on two parallel rails 742
23.2.2 The generator 744
23.3 Back EMF in an electric motor 747
23.4 The induced electric field and eddy currents 748
23.4.1 Magnetic braking 752
23.5 Transformers 753
23.6 Maxwell's equations and electromagnetic waves 756
23.7 Summary 760
23.8 Thinking about the material 763
23.9 Sample problems and solutions 764
23.9.1 Problems 764
23.9.2 Solutions 765

24 The theory of special relativity **766**
24.1 Introduction: The issue with Maxwell's equations . . . 766
24.2 Einstein's postulates . . . 771
24.2.1 Simultaneity . . . 772
24.3 Time dilation . . . 774
24.4 Length contraction . . . 780
24.5 Electric and magnetic fields and Special Relativity . . . 782
24.6 Lorentz transformations and space-time . . . 784
24.6.1 Four-dimensional space-time . . . 784
24.6.2 Space-time diagrams . . . 785
24.6.3 Lorentz transformations . . . 786
24.6.4 Lorentz addition of velocities . . . 793
24.7 Relativistic momentum and energy . . . 795
24.8 Closing remarks . . . 800
24.9 Summary . . . 802
24.10 Thinking about the material . . . 807
24.11 Sample problems and solutions . . . 808
24.11.1 Problems . . . 808
24.11.2 Solutions . . . 809

A Vectors **811**
A.1 Coordinate systems . . . 811
A.1.1 1D Coordinate systems . . . 811
A.1.2 2D Coordinate systems . . . 812
A.1.3 3D Coordinate systems . . . 814
A.2 Vectors . . . 816
A.2.1 Unit vectors . . . 817
A.2.2 Notations and representation of vectors . . . 818
A.3 Vector algebra . . . 819
A.3.1 Multiplication/division of a vector by a scalar . . . 819
A.3.2 Addition/subtraction of two vectors . . . 820
A.3.3 The scalar product . . . 822
A.3.4 The vector product . . . 823
A.4 Example uses of vectors in physics . . . 825
A.4.1 Kinematics and vector equations . . . 825
A.4.2 Work and scalar products . . . 827
A.4.3 Using vectors to describe rotational motion . . . 828
A.4.4 Torque and vector products . . . 829
A.5 Summary . . . 831
A.6 Thinking about the Material . . . 833
A.7 Sample problems and solutions . . . 833
A.7.1 Problems . . . 833
A.7.2 Solutions . . . 834

B Calculus **835**

B.1 Functions of real numbers . . . 835
B.2 Derivatives . . . 837
B.2.1 Common derivatives and properties . . . 839
B.2.2 Partial derivatives and gradients . . . 841
B.2.3 Common uses of derivatives in physics . . . 845
B.3 Anti-derivatives and integrals . . . 845
B.3.1 Common anti-derivative and properties . . . 851
B.3.2 Common uses of integrals in Physics - from a sum to an integral . . . 852
B.4 Summary . . . 856
B.5 Thinking about the Material . . . 857
B.6 Sample problems and solutions . . . 857
B.6.1 Problems . . . 857
B.6.2 Solutions . . . 859

C Guidelines for lab related activities **860**
C.1 The process of science and the need for scientific writing . . . 860
C.2 Scientific writing . . . 861
C.3 Guide for writing a proposal . . . 863
C.4 Guide for reviewing a proposal . . . 864
C.5 Guide for writing a lab report . . . 865
C.5.1 Guide for reviewing a lab report . . . 867
C.6 Sample proposal (Measuring g using a pendulum) . . . 868
C.7 Sample proposal review (Measuring g using a pendulum) . . . 870
C.8 Sample lab report (Measuring g using a pendulum) . . . 871
C.9 Sample lab report review (Measuring g using a pendulum) . . . 873

D The Python programming language **874**
D.1 A quick intro to programming . . . 874
D.2 Arrays . . . 875
D.3 Plotting . . . 876
D.4 The QExpy python package for experimental physics . . . 878
D.4.1 Propagating uncertainties . . . 878
D.4.2 Plotting experimental data with uncertainties . . . 879
D.5 Advanced topics . . . 881
D.5.1 Defining your own functions . . . 881
D.5.2 Using a loop to calculate an integral . . . 882

1 The theory of special relativity

In this chapter, we introduce the theory of Special Relativity, originally formulated by Albert Einstein in 1905. Along with the development of Quantum Mechanics, Special Relativity marks the start of "modern physics", and the introduction of theories to describe our world that are decidedly counter-intuitive.

Learning Objectives

- Understand the motivation for developing the Theory of Special Relativity.
- Understand Einstein's postulates and their consequences.
- Understand how to apply Einstein's postulates to describe simultaneity.
- Understand how to model length contraction and time dilation.
- Understand how to apply Lorentz transformations and make space-time diagrams.
- Understand how to model the energy and momentum of a relativistic object.

Think About It

Is it possible to time-travel into the future, so that you will be younger than people that are currently older than you?

A) Yes, it's possible.
B) No, it is impossible because it would violate causality.
C) No, it is impossible because it's a ridiculous idea.

1.1 Introduction: The issue with Maxwell's equations

In Chapter 23, we summarized our knowledge of electromagnetism using Maxwell's four equations. As far as we can tell, this is the best description that we have of classical electric and magnetic phenomena (classical in the sense that the equations do not describe the behaviour of particles that are described by Quantum Mechanics). One of the consequences of Maxwell's equations is that they describe the existence of electromagnetic waves that propagate with a speed, c, given by:

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$$

where ϵ_0 and μ_0 are the permittivity and permeability of free-space, respectively. The obvious question to ask about these electromagnetic waves is: "In what medium do these waves propagate?". In the late 1800s, it was thought that the Universe was bathed in

a substance called the "luminous ether" (or just "ether"), through which electromagnetic waves propagate. It was then thought that the speed, c, of these waves was, naturally, measured with respect to the ether. This led to the idea that there exists a special inertial frame of reference in the Universe, corresponding to that frame of reference in which light travels at a speed, c. This frame of reference would be at rest relative to the ether.

In the late 1880s, Michelson and Morley developed a clever experiment to measure the speed of the Earth relative to the ether. If the ether exists, and the Earth is moving through it, then a beam of light travelling parallel to the motion of the Earth should travel at a slightly different speed than a beam of light travelling in the perpendicular direction. However, Michelson and Morley conclusively demonstrated that this was not the case. There is no detectable motion of the Earth through a medium in which light (a electromagnetic wave) propagates. There is no ether. This was a very puzzling discovery, with strange implications for Maxwell's equations.

Let us demonstrate, through a simple example, an "issue" with the theory of electromagnetism. Rather, it is not an issue, but a very strange implication. Consider two infinitely-long wires, separated by a distance, r, each carrying a uniform charge per unit length, λ, as illustrated in Figure 24.1.

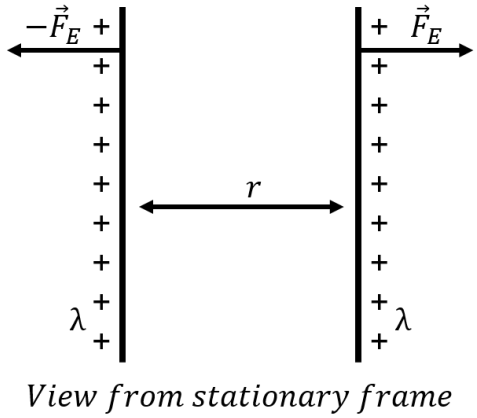


Figure 1.1: Two infinitely long charged wires exert a repulsive electric force on each other.

We can easily calculate the magnitude of the repulsive electric force, $\vec{F}_E$, exerted by one charged wire on a section of length, l, of the other wire. The magnitude of the electric field at a distance, r, from an infinitely-long wire with charge per unit length, λ, is given by:

$$E = \frac{\lambda}{2\pi\epsilon_0 r}$$

A section of length, l, of the other wire carries charge, $q = l\lambda$, so that the force on that section of wire has a magnitude:

$$F_E = qE = \lambda l \left(\frac{\lambda}{2\pi\epsilon_0 r}\right) = \frac{\lambda^2 l}{2\pi\epsilon_0 r}$$

And the force per unit length, on either one of the wires, has a magnitude:

$$\frac{F_E}{l} = \frac{\lambda^2}{2\pi\epsilon_0 r}$$

This is the only force exerted on one of the wires, and will thus allow us to completely specify the motion of that wire (we know all of the forces exerted on the wire, so we can use Newton's Second Law to determine its acceleration and describe its motion).

Consider the same two wires, each carrying charge per unit length but as viewed from a frame of reference that is moving downwards (parallel to the wires), with a speed, v. In this frame of reference, the infinite wires still have a net charge per unit length, but they also appear to have an upwards moving current, I, since we observe positive charges moving upwards through space.

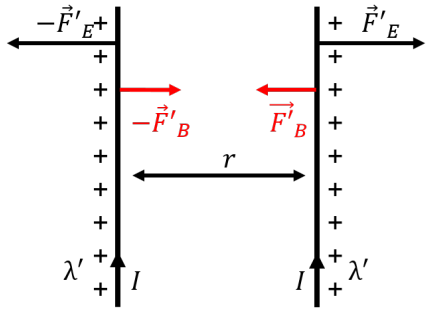


Figure 1.2: Two infinitely-long charged wires as viewed from a down-going frame of reference will appear to have upwards-going currents that will result in an attractive magnetic force between the wires.

In this new frame of reference, we see two wires with charges on them, moving upwards with speed, v. In a length of time, Δt, we see a length of wire, $\Delta x = v\Delta t$, go by, with total charge, $\Delta Q = \lambda' \Delta x$. For reasons that will be clear below, we use a different charge density, λ', in the moving frame of reference, although we *expect* that $\lambda' = \lambda$. This corresponds to a current, I, given by:

$$I = \frac{\Delta Q}{\Delta t} = \lambda' \frac{\Delta x}{\Delta t} = \lambda' v$$

Thus, in the downward going frame of reference, we see two wires with upwards current in them, and these wires must extract an attractive magnetic force between each other, with magnitude (per unit length):

$$\frac{F'_B}{l} = -\frac{\mu_0 I_1 I_2}{2\pi r} = -\frac{\mu_0 I^2}{2\pi r} = -\frac{\mu_0 \lambda'^2 v^2}{2\pi r}$$

where the prime (') on the force indicates that the force is measured in this different inertial frame of reference, and the minus sign indicates that it is in the opposite direction from the repulsive electric force.

In the downwards going frame of reference, the wires are still charged, and must still exert a repulsive force, with magnitude (per unit length):

$$\frac{F'_E}{l} = \frac{\lambda'^2}{2\pi\epsilon_0 r}$$

where, again, we used primes ('), to denote quantities that are measured in the moving frame of reference.

The description of how the wires will move should not depend on the frame of reference that we choose to model the wires (they will move under the forces exerted on them regardless of whether we are observing them from a fixed or a moving point, and indeed regardless of whether we observe them at all!). Thus, the net force (per unit length) exerted on a wire cannot depend on our frame of reference. The total repulsive electric force, F_E, calculated in the stationary frame of reference must be equal to the sum of the magnetic, F'_B, and electric force, F'_E, calculated in the moving frame of reference [1]:

$$\begin{aligned} \frac{F_E}{l} &= \frac{F'_E}{l} + \frac{F'_B}{l} \\ \frac{\lambda^2}{2\pi\epsilon_0 r} &= \frac{\lambda'^2}{2\pi\epsilon_0 r} - \frac{\mu_0 \lambda'^2 v^2}{2\pi r} \end{aligned}$$

where we recognized that the charge per unit length, λ', must be different in the moving frame of reference, or the above would give an inconsistent equation (the electric forces would cancel and we would find that the magnetic force is equal to zero). Thus, the repulsive electric force must be larger as observed in the moving frame of reference, or the net force on the wire would be different when evaluated in the two frames of reference. This is a truly bizarre conclusion, as we will see.

Before proceeding, let us clearly state our assumptions in modelling the force between the two charged wires:

1. The net force on the wire, allowing us to describe its motion, cannot depend on our frame of reference. We expect the laws of physics to be applicable from any inertial frame of reference.
2. We assume that Maxwell's equations hold in all inertial frames of reference. In particular, we assume that the constants, μ_0 and ϵ_0, are the same in all inertial reference frames.

The first assumption allows us to state that the net force in the two frames of reference must be the same. The second assumption implies that we must change the charge density, λ', in the moving frame of reference, since the constants must remain the same, and this is the only quantity that can lead to a different electric force in the moving frame of reference (which is required if the net force is to be the same, according to our first assumption). Let us determine the new charge density, λ', in terms of the charge density that is measured at

[1] This statement is generally true for Special Relativity, because the force is exerted in the direction perpendicular to that of motion.

rest. Starting with the requirement that the net force on the wire must not depend on the frame of reference, we find:

$$\begin{aligned}\frac{\lambda^2}{2\pi\epsilon_0 r} &= \frac{\lambda'^2}{2\pi\epsilon_0 r} - \frac{\mu_0\lambda'^2 v^2}{2\pi r} \\ \lambda^2 &= \lambda'^2 - \epsilon_0\mu_0\lambda'^2 v^2 \\ \lambda^2 &= \lambda'^2(1 - \epsilon_0\mu_0 v^2) \\ \therefore \lambda' &= \lambda\frac{1}{\sqrt{\epsilon_0\mu_0 - v^2}}\end{aligned}$$

Finally, recognizing that we can use the speed of light, c, to replace the combination of constants, $\epsilon_0\mu_0$, we find:

$$\lambda' = \lambda\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Thus, the charge per unit length on the wire is larger when measured from the moving frame of reference (the factor that multiplies, λ, is larger than one if $v < c$). It should be somewhat bothersome to you that the charge per unit length depends on the frame of reference in which it is measured, but this is the only way for our two assumptions to hold.

So far, this has just been some math to ensure that "things work out", namely that our description of the motion of the wire does not depend on our frame of reference. However, the consequences of what we just derived are profound. We concluded that the charge per unit length on a wire depends on our frame of reference.

Imagine drawing two lines on one of the wires, and imagine that you can actually see the charges on the wire (maybe they fluoresce or something). The charge per unit length on the wire, λ, is found by counting the number of charges between the two lines and dividing that by the distance between the two lines. Now, both an observer at rest with the wire, and one that is moving relative to the wire will agree on the number of charges contained between the two lines. They will both count the same number. Thus, if the observer moving relative to the wire is to measure a larger charge density, then the distance between the lines must be smaller for that observer! To the observer moving relative to the wire, the wire is actually shorter. It does not appear to be shorter, it IS shorter!

To summarize, by requiring that the laws of physics are the same in all inertial frames of reference, and by requiring that Maxwell's equation are the same in all inertial frames of reference, we conclude that the charge per unit length that is measured on a wire must depend on the frame of reference in which it is measured. Since it cannot be the number of charges on the wire that depends on the frame of reference, it must be the length of the wire that depends on the frame of reference. Thus, either we accept that Maxwell's equations are incorrect, or we accept that they are correct but that they imply that objects shrink in length when they are moving (regardless of whether charges are involved). It turns out that the latter choice provides a better description of nature (and one that has not been invalidated!).

As an additional consequence of accepting these implications from Maxwell's equations is that the definition of the electric and magnetic fields must depend on the frame of reference. In the example from this section, we saw that what looks like an electric field in the stationary frame of reference, can appear as the combination of a magnetic and electric fields in a moving frame of reference.

1.2 Einstein's postulates

Albert Einstein was the first to provide a complete description of how to deal with the issues that arise from Maxwell's equations when these are examined in different inertial frames of reference. The Theory of Special Relativity, is based on Einstein's two postulates:

1. The laws of physics are the same in all inertial reference frames. There is no experiment that can be performed to determine whether one is at rest or moving with constant velocity.
2. The speed of light propagating in vacuum is the same in all inertial reference frames. Any observer in an inertial frame of reference, regardless of their velocity, will measure that light has a speed of c, when it propagates in vacuum.

These postulates are equivalent to the assumptions that we made above to model the force between the two wires (we stated that the constants, ϵ_0 and μ_0, were independent of reference frame, instead of c). While the first postulate is perhaps "acceptable" to our common sense, the second one grossly defies common intuition. Consider two archers, as illustrated in Figure 24.3.

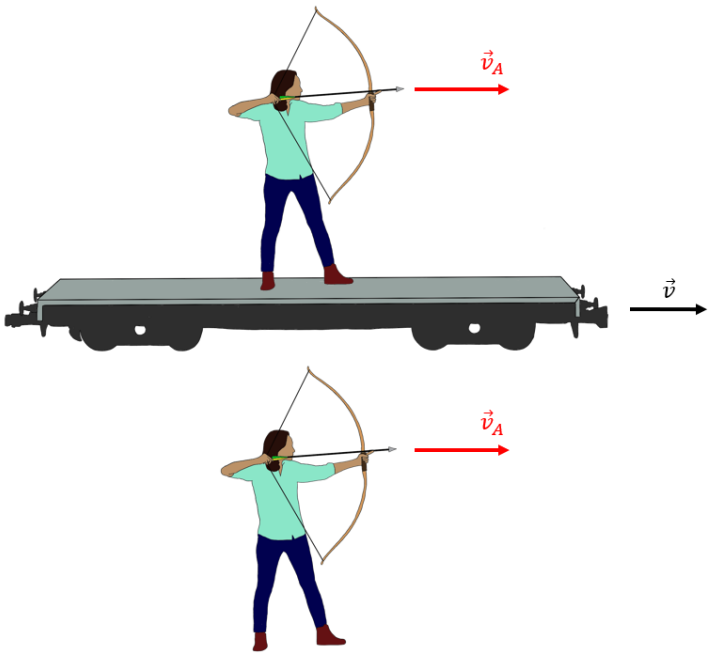


Figure 1.3: Two archers can fire an arrow with speed v_A. As measured in the frame of reference of the ground (of the target), the arrow fired from the archer that is on the train will have a higher speed.

Both archers can fire an arrow with a speed, v_A. One archer fires her arrow from the ground,

at a target on the ground, and that arrow will hit the target with a speed, v_A. The other archer is located on a train that is moving with speed v, in the same direction that she wishes to shoot her arrow. She measures her arrow to leave her bow with speed, v_A, but, as seen from the ground (and from the target), her arrow has a speed $v_A + v$, and it will hit the target with a higher speed, as expected.

Now, consider two space cops that instead fire a pulse of laser light at a target on the ground, as illustrated in Figure 24.4.

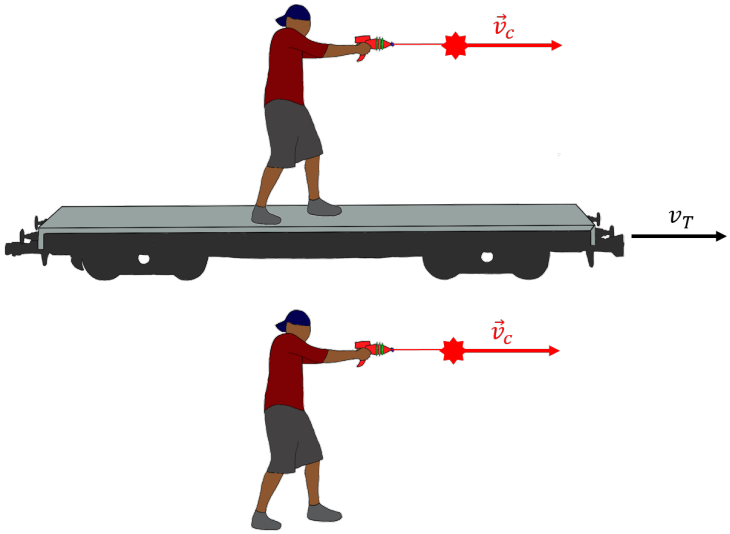


Figure 1.4: Two people fire a laser pulse. Regardless of whether the pulse of laser light was fired from a moving train or from the ground, it will have a speed of c in all frames of reference.

In this case, according to Einstein's second postulate, the speed of the pulses as measured on the ground (by the target), will be c, regardless of whether one of the pulses was fired from a moving train. This is truly strange and not compatible with our experience. Imagine that the train is moving close to the speed of light. The space cop on the train would fire a laser pulse that he would observe to move away from him at the speed of light. When observed from the ground, we will see the pulse of light moves away from him very slowly, since he is on a train going at almost the speed of light.

1.2.1 Simultaneity

As a first consequence of Einstein's postulates, let us consider the notion of simultaneity. Figure 24.5 shows Alice on the platform of a train station. Alice is midway between two clocks, A and B. Both identical clocks were configured so that they send a pulse of laser light when the time is 20 minutes past four o'clock. Since Alice is midway between the clocks, if they emit their pulses of light at the same time, then Alice will see two pulses of light arrive at her location at the same time. She signals that the two pulses of light have reached her at the same time by raising her hands.

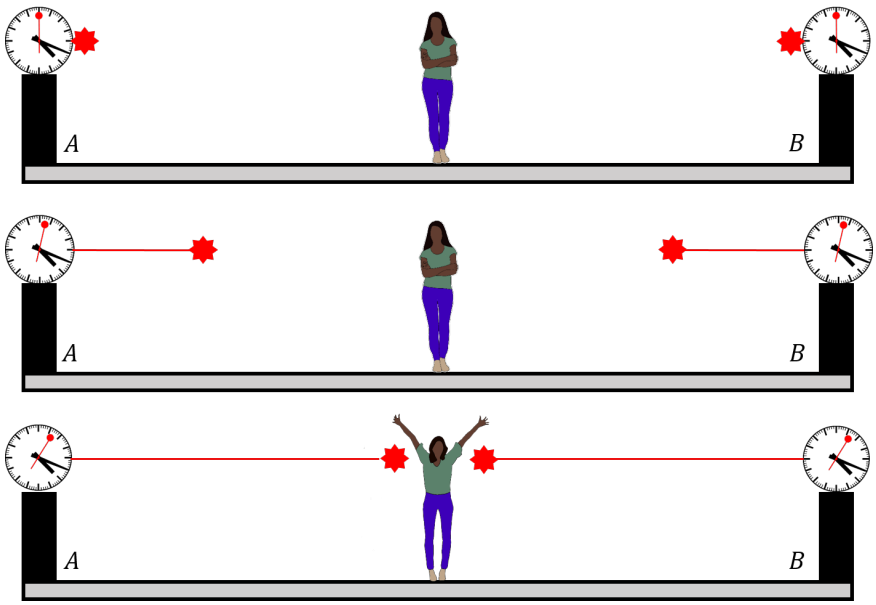


Figure 1.5: Alice is equidistant from two clocks. The clocks fire a laser pulse when the time is 20 past four, and Alice observes both pulses arriving at her location at the same time, concluding that the pulses were emitted by the clocks at the same time.

Brice is located on a train that is travelling with speed, v, in the direction from clock A to clock B, as illustrated in Figure 24.6. He sees Alice and the platform moving towards him.

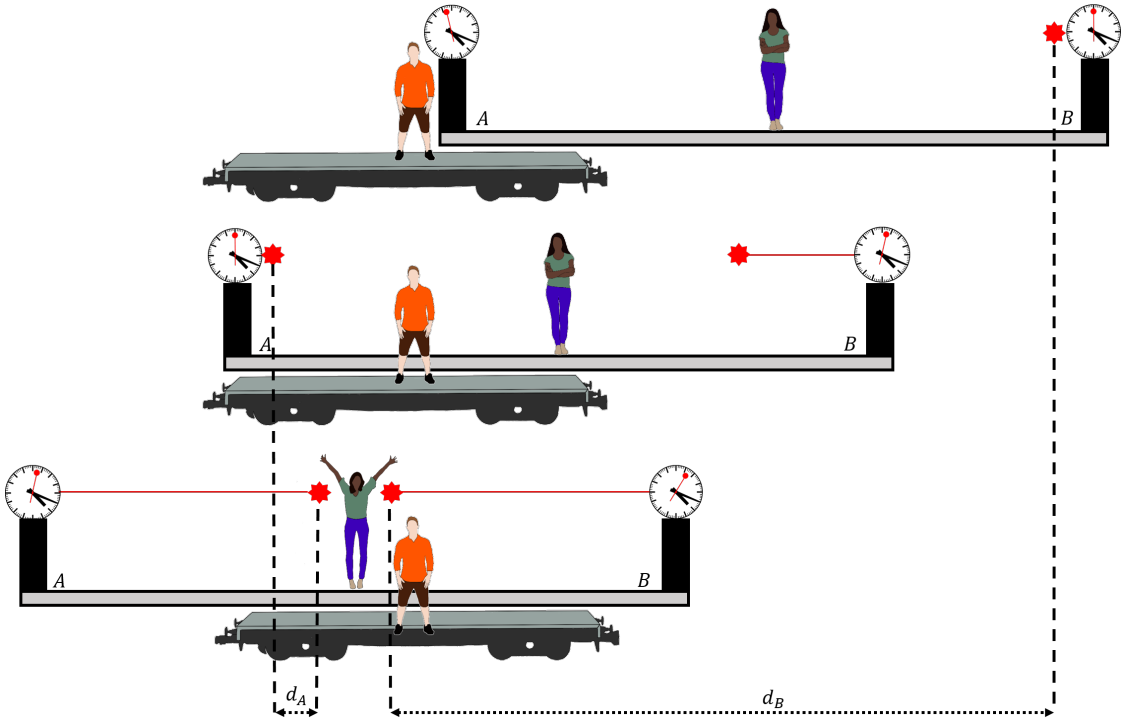


Figure 1.6: Brice is on a moving train, and, from his perspective, it is Alice and the platform that are moving towards him. Brice must conclude that the pulse from clock B was emitted earlier, since it must travel further than the pulse emitted from clock A to reach Alice at the same time.

Brice must agree that the two pulses arrived at Alice's location at the same time, since he can also see her raise her hands. In Brice's frame of reference, the two pulses of light must travel with the speed of light (Einstein's second postulate). Once the pulse of light has been emitted from clock B, Brice observes that Alice is moving away from the location of where the pulse was emitted, so that pulse must travel a large distance, d_B. On the other hand, once the pulse from clock A is emitted, Brice observes that Alice moves towards where the pulse was emitted, so it only needs to travel a shorter distance, d_A, in order to reach Alice. Thus, for both pulses to arrive at Alice at the same time and travel at the speed of light, the pulse from clock B had to be emitted first, according to Brice.

That is, while Alice measures the clocks to be synchronized and emit pulses at the same time, Brice measures that clock B is running ahead of clock A. The two observers, Alice and Brice, in different reference frames, cannot agree on whether two events are simultaneous. Even worse, if a third observer, Chloë, is located on a train going in the opposite direction from Brice's train, she will conclude that the pulse from clock A was emitted earlier than the pulse from clock B. A consequence of Einstein's postulates is that observers in different frames of reference will not agree on whether two events happen at the same time, and in some cases, as the one we illustrated, the observers will not agree on which event happened first. Think of the implications for causality!

1.3 Time dilation

Einstein was famous for his "thought experiments", which allow us to understand the consequences of a theory by performing thought experiments that would be impractical to actually carry out (such as the experiment with Alice and Brice described above, which would be impractical to carry out, since the speed of light is so high that Brice would never notice that clock A emitted the pulse slightly earlier).

Imagine that we build a clock using a pulse of light travelling (oscillating) back and forth between two mirrors, separated by a distance, L, as illustrated in Figure 24.7.

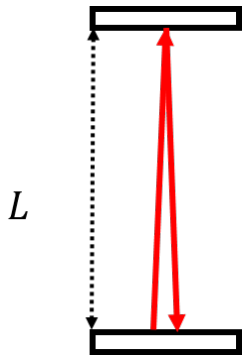


Figure 1.7: A clock is made by having a pulse of light bounce back and forth between two parallel mirrors separated by a distance, L.

Since the speed of light is, c, the time that it will take for the pulse of light to travel back

and forth between the two mirrors, namely the period of the clock, is given by:

$$\Delta t = \frac{2L}{c}$$

where the speed of light, c, is given by the total distance travelled by the pulse of light divided by the time taken to do so:

$$c = \frac{2L}{\Delta t}$$

Now, imagine placing this clock on a spaceship that travels with speed, v, perpendicular to the direction of the movement of the light. The clock is illustrated in Figure 24.8, as seen from the ground.

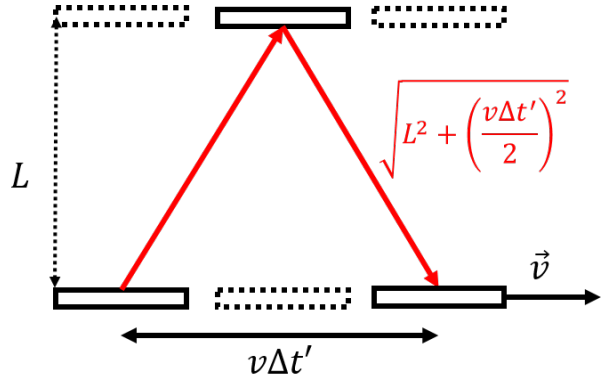


Figure 1.8: A clock is made by having a pulse of light bounce back and forth between two parallel mirrors separated by a distance, L. When the clock is placed on a spaceship moving with speed, v, the light travels a longer distance before completing a full cycle, as observed by someone not travelling with the clock.

From the perspective of a person watching the clock go by, the pulse of light travels a larger distance over one clock period, since the mirrors move to the right as the pulse of light moves up and down. However, by Einstein's second postulate, the pulse of light must still travel with the same speed, c, so it must take the pulse of light longer to bounce between the two mirrors than it did when the clock is at rest. Let us determine the relationship between the period of the clock, Δt, measured when the clock is at rest, and the period of the clock, $\Delta t'$, as measured by an observer that sees the clock go by with speed, v.

To an observer that sees the clock move by with speed, v, the speed of the pulse of light, which must also be equal to c, is given by:

$$c = \frac{2\sqrt{L^2 + \left(\frac{v\Delta t'}{2}\right)^2}}{\Delta t'}$$

where the distance in the numerator was simply found by Pythagoras' theorem, as the spaceship will travel a horizontal distance, $v\Delta t'$, as measured by the observer that is not moving with the spaceship. Squaring this relationship, we can isolate the period of the

clock, $\Delta t'$, as measured by the observer that sees the clock move with speed, v:

$$\begin{aligned} c^2 &= \frac{4L^2}{\Delta t'^2} + v^2 \\ \Delta t'^2(c^2 - v^2) &= 4L^2 \\ \therefore \Delta t' &= 2L\frac{1}{\sqrt{c^2 - v^2}} = \frac{2L}{c}\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{aligned}$$

Note that the term, $2L/c$, is simply the period of the clock as measured in a frame of reference where the clock is stationary. Thus, we can relate the two clock periods:

$$\boxed{\Delta t' = \Delta t \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}}$$

To re-iterate: the period of the clock, $\Delta t'$, as measured in a frame of reference that is moving relative to the clock is longer than the period of the clock, Δt, as measured in the "rest frame" of the clock (the reference frame where the clock is stationary). We call this effect **"time dilation"**, and it is not just some mathematical curiosity. The clock that we imagined with a pulse of light is a real clock that one could actually construct; we could use it to measure time. That clock will appear to tick slower if it is moving. **Time goes by slower in a moving reference frame**. If a person climbs on a ship that is moving, that person will age at a slower rate than a person that remained on Earth. By travelling at high speeds, you effectively travel into the future, as observed on Earth. The equation above allows us to relate the amount of time that went by in one reference frame to the amount of time that went by in a different frame of reference.

We define the time that is measured at rest as the "proper time". In our example, Δt, is the proper time (proper period) for the clock, since it is defined in a frame of reference where the clock is at rest. The "dilated time", $\Delta t'$, is measured in a frame of reference that is moving relative to the clock.

The factor by which time is dilated comes up often in Special Relativity, and is called the gamma factor:

$$\boxed{\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}}$$

As a corollary to Einstein's postulates, we will see that nothing can ever exceed the speed of light in vacuum. The gamma factor is always greater than 1, since v (the speed between the two different inertial frames of reference), must always be smaller than c. You may also recognize that the gamma factor appeared in our introductory example with the force between two wires. Here, we derived the gamma factor from kinematical considerations, whereas in the example with the two wires, it came straight out of the equations for electromagnetism.

Checkpoint 1-1

What is gamma for a speed of $v = 0.75c$?

A) 1.51
B) 0.75
C) 75
D) 1.68

Checkpoint 1-2

What speed corresponds, v, to a gamma factor of 2.5?

A) $v = 2.5c$
B) $v = 0.92c$
C) $v = 0.25c$
D) $v = 0.47c$

Time-dilation is a real effect that has been observed, for example by placing high precisions atomic clocks on an airplane to observe their period slow down. Another example of time-dilation is the fact that we observe many particles called muons at the surface of the Earth. Muons are very similar to electrons, except that they have a larger mass, and that they are unstable (they radioactively decay into an electron and neutrinos, after $2.2\,\mu$s on average). Muons are produced in large amounts when cosmic rays (high energy particles from outside our Solar System) strike the molecules in our upper atmosphere, at altitudes of tens of kilometres. As the muons travel down towards the Earth, they decay.

Suppose that muons are produced travelling at the speed of light; in that case, they would travel a distance $d = (3 \times 10^8\,\text{m/s})(2.2 \times 10^{-6}\,\text{s}) = (660\,\text{m})$, on average, before decaying. However, muons are produced tens of kilometres above the surface of the Earth, travel slower than the speed of light, and yet, we are able to detect many muons at the surface of the Earth. We would expect that all muons would have decayed before reaching the surface of the Earth.

We can understand this in terms of time dilation; in the reference frame of the muon, the muon decays after $\Delta t = 2.2\,\mu$s. In a reference frame from which the muon appears to move with speed, v, the "clock" that measures how long the muon has existed ticks slower. Thus, from the Earth, we observe that the muon takes longer than $2.2\,\mu$s to decay, giving it time to reach the surface of the Earth.

Example 1-1

A muon travels with a speed of $0.9c$ as observed from the surface of the Earth. As measured in the frame of reference of the Earth, how far has the muon travelled after $2.2\,\mu$s have elapsed **in the muon's frame of reference**?

Solution

The muon is travelling with a speed of $v = 0.9c$ relative to the Earth, thus the gamma factor is given by:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.9^2}} = 2.29$$

The amount of time that goes by in the frame of reference of the Earth, Δt, when $\Delta t' = 2.2\,\mu$s has gone by in the muon's frame of reference will be dilated by the gamma factor. $\Delta t'$ is the proper time in the muon frame's of reference, which corresponds to a longer time in Earth's frame of reference:

$$\Delta t = \gamma \Delta t' = (2.29)(2.2\,\mu\text{s}) = 5.0\,\mu\text{s}$$

In the frame of reference of the Earth, the muon has travelled a distance:

$$d' = v\Delta t' = (0.8c)(5.0\,\mu\text{s}) = 1350\,\text{m}$$

Discussion: In this example, we see that an object, such as a muon, that travels with a speed that is 90 percent of the speed of light will have a gamma factor around 2. Thus, from the Earth's frame of reference, it appears that the muons "ages" at about half of the rate at which one would observe the muon to age if moving along with the muon. This is the mechanism that allows muons to exist much longer than $2.2\,\mu$s when they are travelling relative to the Earth.

Also, in Earth's reference frame, the muons travel a distance of 1350 m in the period of time between being produced and decaying. In the reference frame of the muon, only $2.2\,\mu$s elapse as the Earth moves closer to the muon, at the same speed. In the reference frame of the muon, the Earth has travelled a distance:

$$d' = v\Delta t = (0.9c)(2.2\,\mu\text{s}) = 594\,\text{m}$$

Thus, as viewed from the muon's frame of reference, the distance that it travelled between being produced and decaying is about half the distance as measured in the Earth's reference frame. This is called "length contraction" and is a necessary consequence of time-dilation.

Example 1-2

A spaceship carrying your friend Alice speeds away at a speed of $0.99c$ towards the nearest star, Proxima Centauri, a distance of 4.2 ly (light-years) away. How much time does the trip take as measured by Alice? How far has the spaceship travelled, according to Alice?

Solution

Alice's trip is illustrated in Figure 24.9, showing the trip as viewed from Earth's and from Alice's frame of reference.

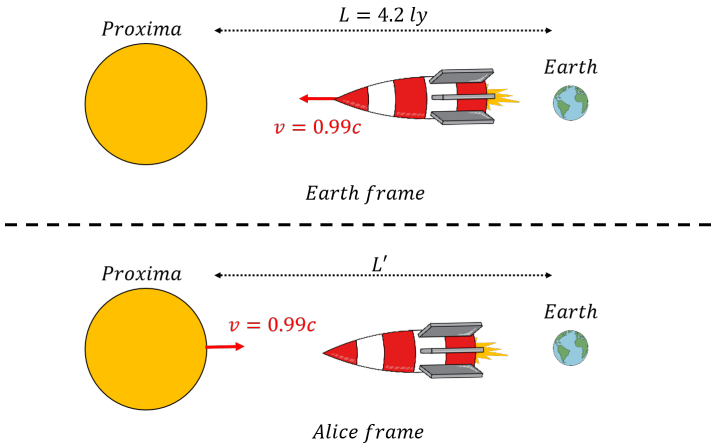


Figure 1.9: Alice travels in a spaceship from the Earth to the star Proxima Centauri. In the Earth frame of reference, the star is 4.2 ly *away.*

In Earth's frame of reference, the spaceship travels a distance of 4.2 ly at a speed of $0.99c$, which will take a time, $\Delta t'$, given by:

$$\Delta t' = \frac{(4.2\,\text{ly})}{(0.99c)} = 4.2\,\text{y}$$

which is not surprising, since Alice is travelling at almost the speed of light. This is the time that goes by on planet Earth. Since Alice's spaceship is moving, less time will go by on the spaceship, as the 4.2 y is the dilated time measured at Earth, not the proper time measured by Alice. First, we determine the gamma factor:

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = \frac{1}{\sqrt{1-0.99^2}} = 7.1$$

The proper time measured by Alice is:

$$\Delta t = \frac{\Delta t'}{\gamma} = \frac{(4.2\,\text{y})}{(7.1)} = 0.6\,\text{y}$$

That is, Alice only ages by 0.6 y (about 7 months), while everyone on Earth ages by 4.2 y!

In Alice's frame of reference, she is not moving, and Proxima Centauri moves towards her at a speed of $0.99c$. Since her trip only lasts about 7 months (0.6 y), Proxima Centauri moves towards her by a distance, L':

$$L' = (v)(\Delta t) = (0.99c)(0.6\,\text{y}) = 0.6\,\text{ly}$$

as illustrated in Figure 24.9. Thus, Alice concludes that the distance between Earth and Proxima Centauri is only 0.6 ly instead of 4.2 ly. The distance that she observes is contracted compared to the "proper distance" between Earth and Proxima (the distance measured when we are at rest relative to Earth and Proxima).

Discussion: In this example we saw, again, how the time that one measures depends on the frame of reference. In particular, if one can build spaceships that goes close to the speed of light, one can cover large distances in the Universe without ageing much. We also saw that length contraction is a necessary corollary to time-dilation. Objects appear contracted when they move, relative to their length when they are measured at rest (their "rest length" or their "proper length").

One interesting issue uncovered by Example 24-2 is the so-called "twin-paradox". Imagine that Alice has a twin brother, Brice, that remains on Earth. Alice travels to Proxima Centauri and back (return trip), and will have aged by about 14 months, whereas Brice, will have aged by about 8.4 years (using the numbers in Example 24-2). However, Einstein's first postulate implies that there are no special frames of reference that are at rest. We should be able to think about this situation from the perspective where Alice is at rest, and it is the Earth (with Brice on it), that moves away from her and then back. In this case, Alice is at rest, and she will conclude that it takes about 8.4 years for Brice to move away and come back, and that Brice would have aged by about 7 months. When Alice and Brice meet up again, clearly Alice cannot be both younger and older than Brice, so which one is it? (You will have to look this up, see associated question in the "Thinking about the material" section).

1.4 Length contraction

As we saw in the examples from the previous section, time dilation implies "length contraction". When an object is measured in a frame of reference that is at rest relative to the object, the length of the object, L, is called the "rest length" or the "proper length" of the object. If that object is moving relative to an observer, the observer will measure the object to be shorter, and have a "contracted length", L', given by:

$$\boxed{L' = L\sqrt{1 - \frac{v^2}{c^2}} = \frac{L}{\gamma}}$$

In Example 24-2, Alice measured a contracted distance between Earth and Proxima Centauri, as she was in a frame of reference that is moving relative to the Earth-Proxima Centauri reference frame. One point that is important to note is that length contraction only occurs along the direction parallel to the direction of motion.

Example 1-3

A square painting hanging in a museum has a side with a length of 1 m. If you view the stationary painting from a train moving in the horizontal direction at a speed of $0.85c$, what is the surface area of the painting that you measure?

Solution

Since your train is moving horizontally, only the horizontal dimension of the painting will be contracted. The gamma factor is given by:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{1}{\sqrt{1 - 0.85^2}} = 1.9$$

Thus, the horizontal side of the painting will have a contracted length:

$$L' = \frac{L}{\gamma} = \frac{(1\,\text{m})}{(1.9)} = 0.53\,\text{m}$$

The area of the painting, as measured in the moving frame of reference, is given by:

$$A = (1\,\text{m})(0.53\,\text{m}) = 0.53\,\text{m}^2$$

Checkpoint 1-3

What speed must an object travel in order for it to appear 1 % shorter

A) $0.01c$
B) $0.04c$
C) $0.99c$
D) $0.65c$

Length contraction also allows us to discuss a famous paradox (the "barn", or "ladder" or "barn-pole" paradox). Consider a train that has a rest length of 500 m, travelling at a speed such that $\gamma = 2.5$. As the train goes by, from Earth, it appears to have a (contracted) length:

$$L'_{train} = \frac{(500\,\text{m})}{2.5} = 200\,\text{m}$$

Suppose that there is a tunnel on Earth that is exactly 200 m long, so that the train, when contracted, will fit in the tunnel. When the train passes, an operator briefly closes (and re-opens) the doors at the ends of the tunnel, briefly "capturing" the train, and since the train is contracted, it never hits any of the doors, and all is fine.

From the train's frame of reference, the train has a proper length of 500 m, and the tunnel

is contracted to a length of:

$$L'_{tunnel} = \frac{(200\,\text{m})}{(2.5)} = 80\,\text{m}$$

Thus, from the train's perspective, if the doors of the tunnel are closed, there is no way that the 500 m long train can ever fit in the 80 m long tunnel, as illustrated in Figure 24.10. So what happens when the operator on Earth closes the doors of the tunnel to briefly "capture" the train?

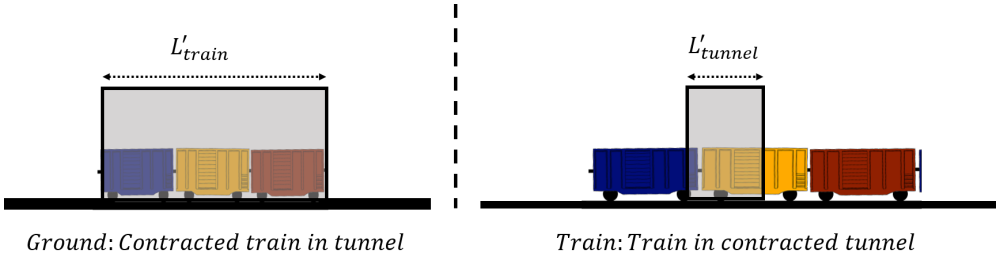


Figure 1.10: In the ground's reference frame, the contracted train appears to fit inside the tunnel. From the train, the (proper length) train will not fit in the contracted tunnel.

Clearly, people on the Earth and people on the train have to agree on whether the train was destroyed by the tunnel doors. The operator on Earth can clearly close both doors of the tunnel when the train is inside and not destroy the train. Hence, people on the train must agree that the train never collided with the doors, and that the doors were closed. The answer to this paradox lies in the fact that simultaneity is relative. The tunnel operator believes that she has closed the two doors of the tunnel at exactly the same time, precisely when the contracted train is lined up with the tunnel. However, to people on the train, in a different frame of reference, the doors did not close at the same time, since events that are simultaneous in one frame of reference are not necessarily simultaneous in a different frame of reference. To people on the train, there was never a time when the train was in the tunnel and both doors were closed at the same time!

Checkpoint 1-4

Referring to the above paradox, to people on the train, which tunnel door closes first?

A) The door at the entrance of the tunnel closes first.
B) The door at the exit of the tunnel closes first.

1.5 Electric and magnetic fields and Special Relativity

In this section, we present one more example to show how Special Relativity is connected to electromagnetism. Consider a wire that carries an electric current towards the left, and a positive charge, $+Q$, located next to the wire, as illustrated in Figure 24.11.

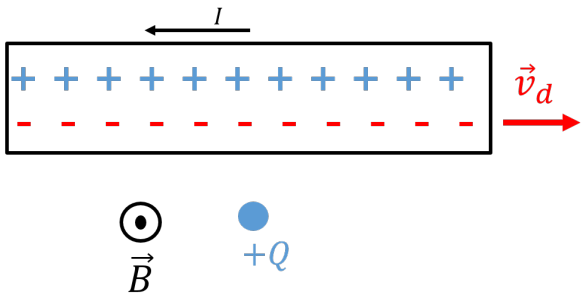


Figure 1.11: A stationary positive charge, $+Q$, near a wire carrying current towards the left. This leads to a magnetic field out of the page at the location of $Q+$.

Inside the wire, negative electrons are moving towards the right, with a drift velocity, $\vec{v}_d$, while positive ions remain stationary. Since the charge $+Q$ has a velocity of zero, it experiences no magnetic force. Furthermore, the wire appears to be neutral, with no net electric charge.

If the charge, $+Q$, has a velocity, $\vec{v}_d$, towards the right, it will experience a downwards magnetic force, as illustrated in Figure 24.12.

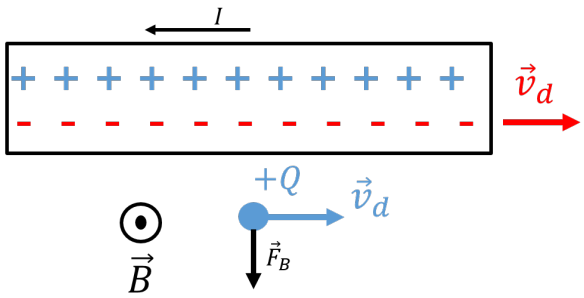


Figure 1.12: A positive charge, $+Q$, moving towards the right, near a wire carrying current towards the left, will experience a downwards magnetic force, $\vec{F}_B = Q\vec{v}_d \times \vec{B}$.

Now, consider this from the perspective of the charge, $+Q$, as illustrated in Figure 24.13. The charge Q is moving towards the right at the same speed as the electrons in the wire. In the reference frame of the charge, $+Q$, the charge has a velocity of zero, and thus will experience no magnetic force. The wire still appears to have a (different) current, I', as the positive ions move to the left, creating a magnetic field, $\vec{B}'$, out of the page.

In the "lab" frame of reference, where the electrons and the charge $+Q$ move towards the right at the same speed, v_d, the electrons appear closer together (length contracted) than they are in the frame of reference of the electrons (or of the charge $+Q$, since it is moving with the electrons). In the frame of reference of the charge $+Q$, the electrons thus appear to be spaced further apart (less dense). On the other hand, in the frame of reference of $+Q$, the positive ions, which are moving towards the left, appear closer together, as the distance between them is now contracted, as illustrated in Figure 24.13.

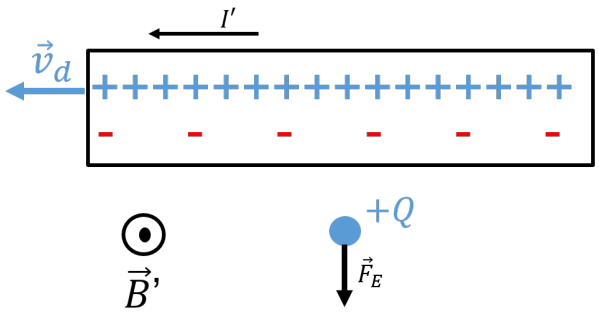


Figure 1.13: In the frame of reference of the charge $+Q$, the charge has a velocity of zero and cannot experience a magnetic force. The ions appear to move to the left, and thus appear denser, since the distance between ions is contracted. The distance between electrons, on the other hand, is larger in this frame of reference. It thus appears that the wire is positively charged, and would exert a downwards electric force on the charge, $+Q$.

In the frame of reference of the charge $+Q$, the wire no longer appears neutral, but appears to have a net positive charge. This results in an electric field away from the wire that will exert a downwards force on $+Q$. In both frames of reference, we conclude that the charge will experience a downwards force. Whether that force is magnetic or electric depends on the frame of reference! Here, we came to the conclusion by using the notion of length contraction, but, remember that it is in fact length contraction that is a consequence of Maxwell's equations holding in different frames of reference, as we illustrated at the beginning of this chapter.

In most real-world applications, we do not see the effects of Special Relativity, as the speeds involved must be very high for the gamma factor to be appreciably different from 1. However, as you recall, the drift speed of electrons in a wire is usually (much) less than mm/s, yet, when dealing with the electric and magnetic forces (fields), even the minuscule length contraction of the electrons/ions at those speeds, leads to relativistic effects. This can be thought of in terms of how strong the electric force really is; even a minute change in charge density (due to length contraction) has an appreciable relativistic effect in how we model the dynamics of a charged particle.

1.6 Lorentz transformations and space-time

1.6.1 Four-dimensional space-time

So far, we have seen that our notions of time intervals (the time between two events) and space intervals (the distance between two locations) depend on our frame of reference. We also saw how space and time are connected, for example by the fact that time-dilation must go hand-in-hand with length contraction. We also concluded that there is no absolute concept of time, and that time is relative (depends on your frame of reference).

In the context of Special Relativity, we introduce the concept of space-time. To describe the location of an object in space-time, we must specify both the location/position coordinates (x, y, z) **and** the time "coordinate", t. Since time, t, has the dimension of time, we usually specify the time coordinate by multiplying it by speed of light, ct, so that it has dimensions

of length. Thus, position in space-time is given by 4 coordinates: (x, y, z, ct).

1.6.2 Space-time diagrams

It is practically impossible to visualize situations in three dimensions, so four dimensions is hopeless! However, we can gain a lot of insight into Special Relativity models by using "space-time diagrams". In a space-time diagram, we use only one of the space coordinates (typically x) along with the time coordinate, ct, to define the two axes of a space-time diagram. Space-time diagrams are analogous to "position as a function of time" graphs that one would draw in kinematics, although they are fundamentally different in that, for a space-time diagram, the coordinates should be thought of as independent (one is not plotting a dependent variable as a function of an independent variable).

Figure 24.14 shows a space-time diagram for an object that was located at position $x = x_1$ at time $t = t_1$ (location A), and at position $x = x_2$ at time $t = t_2$ (location B). The path of an object through space-time, indicated by the line that connects A and B, is called the "world line" of the object.

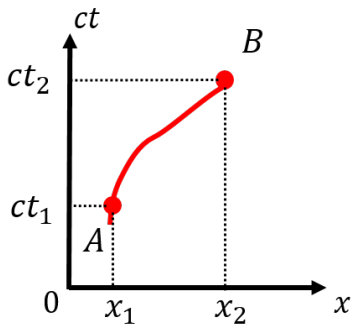


Figure 1.14: World line of an object that moved from locations A in space-time to location B in space-time.

Checkpoint 1-5

What does the world-line of a stationary particle look like?

A) A vertical line.
B) A horizontal line.
C) A point.

A pulse of light travelling in the x direction will always have a world-line that makes a 45° angle with the horizontal (space) axis (since $x = ct$). The world line of any object that travels with a speed below the speed of light must always make an angle with the horizontal axis that is greater than 45°.

A position in space-time is usually called an "**event**". We can draw a set of lines, at 45° degrees from the horizontal axis, that intersect at an event in space-time. Those lines define two "light cones" corresponding to: (1) locations in space-time in the past that could have had a causal effect on the event (the "past light cone"), and (2), locations in space-time in the future for which the event can have a causal effect (the "future light cone").

Figure 24.15 shows the light cones associated with an event, A, in space-time. The past light cone is the only region of space-time in which a different event could have had an impact on the event A. For example, the event A might be that "the object is at position $x = x_1$ at time $t = t_1$", so that the past light cone corresponds to the only locations in space-time that the object could have been in the past. Similarly, the future light-cone defines the locations in space-time upon which the event A could have an effect. For example, this could define the possible locations of the object in the future. The regions outside the light cones can never have an effect on the event A; they are not causally connected. A signal or object would need to travel faster than the speed of light in order to have an effect on something outside of its light cone. There are locations in space-time, in the future of our Universe, that we cannot influence, no matter what we do.

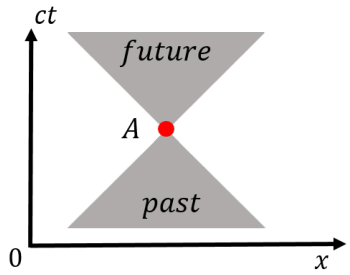


Figure 1.15: The past and future light cones associated with the space-time event, A.

When two events in space-time are within each other's light cones, we say that the space-time interval between them (the line that you draw from one event to the other) is "time-like". Time-like events are such that all observers, in any frame of reference, will agree that one event happened before the other. Thus, events that are causally related must have a time-like interval between them (they are connected by a line that makes an angle greater than 45° with the horizontal axis).

Two events that are outside of each other's light cones are said to be "space-like". Events that are connected by space-like intervals cannot be causally related (one cannot cause the other). Observers in different frames of reference will disagree on the time ordering of space-like events. For example, when Alice observed the two clocks on the platform to emit pulses of light at the same time, Brice disagreed; those two events are connected by a space-like interval.

Finally, the space-time interval between events that are on each other's light-cone (connected by a line that makes a 45° angle with the x-axis), is said to be "light-like" or "null".

1.6.3 Lorentz transformations

In this section, we consider how to transform the space-time coordinates, (x, y, z, ct), as measured in a frame of reference, S, to coordinate $(x', y', z', c't)$, as measured in a frame of reference, S', that is moving with a constant speed, v, relative to the frame, S. For simplicity, we assume that frame S' is moving with speed v in the positive x direction, as measured in frame, S, and that the origin of the two coordinate systems coincided at time $t = 0$. Figure 24.16 shows an illustration of how the two frames of reference are related

(note that these are actual coordinate systems, not space-time diagrams).

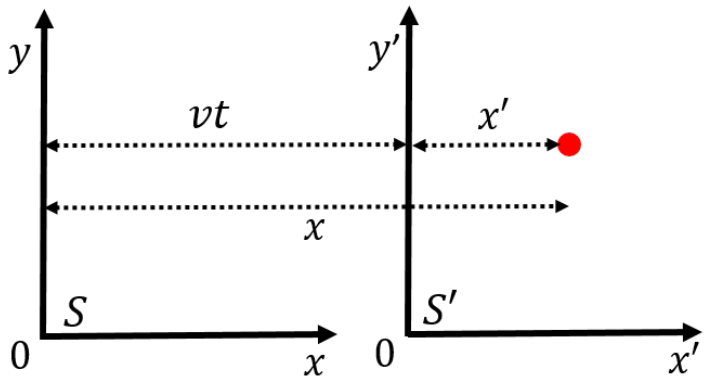


Figure 1.16: The reference frame S' is moving relative to reference frame, S, with speed v in the x direction. At time $t = 0$, the origins of the two coordinate systems coincided.

If we ignore any of Special Relativity, then the coordinates in S' are easily related to those in the S frame of reference using the "Galilean transformations":

$$\begin{aligned} x' &= x - vt \\ y' &= y \\ z' &= z \\ t' &= t \end{aligned}$$

and this corresponds to transformations that we have implicitly used before considering Special Relativity. These equations also allow us to relate the speeds measured in different frames of reference. Suppose that an object has a velocity, $\vec{u} = (u_x, u_y, u_z)$, as measured in the frame of reference, S. We can obtain the components of the velocity vector, $\vec{u}'$, as measured in the frame of reference, S', by taking the time derivatives of the above equations:

$$\begin{aligned} u'_x &= \frac{dx'}{dt'} = \frac{dx'}{dt} = \frac{d}{dt}(x - vt) = \frac{dx}{dt} - v = u_x - v \\ u'_y &= \frac{dy'}{dt'} = u_y \\ u'_z &= \frac{dz'}{dt'} = u_z \end{aligned}$$

which is trivial, since $t = t'$. The transformation above are equivalent (identical) to the rules for transforming velocity that we derived in Section 3.4 for kinematics. In Galilean relativity, time is an absolute quantity that does not depend on the frame of reference. In Special Relativity, the time coordinate is different between different frames of reference, so we cannot simply convert a time derivative in t' to a derivative in t. Instead, we must use the Lorentz transformations.

We can use the formulas for length contraction and time dilation to derive the Lorentz Transformations. Referring to Figure 24.16, x refers to the distance between a point in space-time and the origin of the x axis in the frame, S, as measured in frame, S. Similarly,

x', is the distance to the point in space-time as measured in frame S', from the origin of S'. In frame, S, the distance x' will be contracted to the length x'/γ, so that the Galilean transformation for the x coordinate is modified as follows:

$$\begin{aligned} x' &= x - vt \quad \text{(Galilean)} \\ \frac{x'}{\gamma} &= x - vt \\ \therefore x' &= \gamma(x - vt) \quad \text{(Lorentz)} \end{aligned}$$

The y and z coordinates are the same between frames of references, since all of the length contraction will take place in the direction of the relative motion between frames of reference, which we chose to be in the x direction.

We can obtain the equation for the time coordinate by considering that, in the S' frame of reference, it is the x coordinate that is contracted to x/γ. In the S' frame of reference, the distance between the origins of the two systems is vt' (note the prime on t). We can thus write the contracted distance x, in the S' frame of reference:

$$\begin{aligned} \frac{x}{\gamma} &= vt' + x' \\ t' &= \frac{1}{v}\left(\frac{x}{\gamma} - x'\right) \end{aligned}$$

We can eliminate x' from the last equation using the Lorentz transformation for x' that we just found:

$$\begin{aligned} t' &= \frac{1}{v}\left(\frac{x}{\gamma} - x'\right) \\ t' &= \frac{1}{v}\left(\frac{x}{\gamma} - \gamma x + \gamma vt\right) \\ \frac{t'}{\gamma} &= \frac{1}{v}\left(\frac{x}{\gamma^2} - x + vt\right) \\ &= \frac{1}{v}\left(x\left(1 - \frac{v^2}{c^2}\right) - x + vt\right) \\ &= \frac{1}{v}\left(-\frac{v^2}{c^2}x + vt\right) \\ &= t - \frac{vx}{c^2} \\ \therefore t' &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

where we wrote out the γ factor out explicitly in the fourth line. We can summarize the

Lorentz transformations as follows:

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

and the inverse relations are easily found:

$$\begin{aligned} x &= \gamma(x' + vt') \\ y &= y' \\ z &= z' \\ t &= \gamma\left(t' + \frac{vx'}{c^2}\right) \end{aligned}$$

Note that the Lorentz transformations reduce to the Galilean transformations when the speed, v, between frames of reference is small (so that $\gamma \sim 1$).

Example 1-4

In a frame, S, a pulse of light is emitted (at the speed of light) in the positive x direction, at $t = 0$, from the origin. The pulse is then absorbed at time t, at position $x = d$. Use the Lorentz transformation to show that, in a frame, S', moving in the positive x direction with speed v, relative to S, the pulse also travelled at the speed of light.

Solution

In order to use the Lorentz transformations, we need to define "events", with coordinates in space-time, that we can then convert from one frame of reference to another. Let A be the event that corresponds to the emission of the pulse of light, and B the event that corresponds to the absorption of the pulse. In frame, S, the coordinates of these events are:

$$\begin{aligned} x_A &= 0 \\ t_A &= 0 \\ x_B &= d \\ t_B &= \frac{d}{c} \end{aligned}$$

where in the last line, we used the fact that, in frame, S, the pulse travels at the speed of light. Applying the Lorentz transformations, we can find the coordinates of the same

events in frame, S':

$$x'_A = \gamma(x_A - vt_A) = 0$$
$$t'_A = \gamma\left(t_A - \frac{vx_A}{c^2}\right) = 0$$
$$x'_B = \gamma(x_B - vt_B) = \gamma\left(d - v\frac{d}{c}\right)$$
$$t'_B = \gamma\left(t_B - \frac{vx_B}{c^2}\right) = \gamma\left(\frac{d}{c} - \frac{vd}{c^2}\right)$$

The speed, v'_p, of the pulse of light in frame, S', is given by:

$$v'_p = \frac{(x'_B - x'_A)}{(t'_B - t'_A)} = \frac{\gamma\left(d - v\frac{d}{c}\right)}{\gamma\left(\frac{d}{c} - \frac{vd}{c^2}\right)}$$
$$= \frac{\left(d - v\frac{d}{c}\right)}{\left(\frac{d}{c} - \frac{vd}{c^2}\right)} = c\frac{\left(\frac{d}{c} - v\frac{d}{c^2}\right)}{\left(\frac{d}{c} - \frac{vd}{c^2}\right)} = c$$

which is the speed of light, as expected.

Discussion: In this example, we showed how to use the Lorentz transformations, by clearly defining "events" and their coordinates in space-time. We saw that the Lorentz transformation are consistent with Einstein's second postulate and that the speed of light is the same all frames of reference. This of course makes sense, as we derived the Lorentz transformations from time dilation and length contraction, which are consequences of the postulate.

Einstein's second postulate states that the speed of light is independent of the frame of reference. Consider two points in space-time corresponding to the emission (A) and the absorption (B) of a pulse of light. In the reference frame, S, the distance squared in space between these two events must be equal to the distance (squared) that light travelled between the time of emission and absorption:

$$(x_B - x_A)^2 + (y_B - y_A)^2 + (z_B - z_A)^2 = c^2(t_B - t_A)^2$$
$$\therefore \Delta x^2 + \Delta y^2 + \Delta z^2 = c^2\Delta t^2$$

where (x_A, y_A, z_A, ct_A) and (x_B, y_B, z_B, ct_B) are the space-time coordinates of events A and B. The above equation must hold in all frame of references (e.g. adding a prime to each coordinate), since it is a statement that the speed of light is c.

We can define, s, the "space-time interval", between events, A and B:

$$s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2\Delta t^2$$

which turns out to be "Lorentz invariant" (meaning that this value is the same in all reference frames). The space-time interval can be thought of as a "distance" in space-time that is the same in all reference frames. If the events A and B corresponds to the emission and

absorption of light, then $s = 0$, and we say that the interval between A and B is light-like or null. If $s < 0$, the events are on a time-like interval, and if $s > 0$, the events are separated by a space-like interval. Since s does not depend on the frame of reference, all observers will agree on whether events are separated by time or space-like intervals.

We can visualize the effect of Lorentz transformations on space-time diagrams, as in Figure 24.17, which shows the space-time diagrams for a reference frame, S, and a second reference frame, S', moving with speed v in the x direction relative to S.

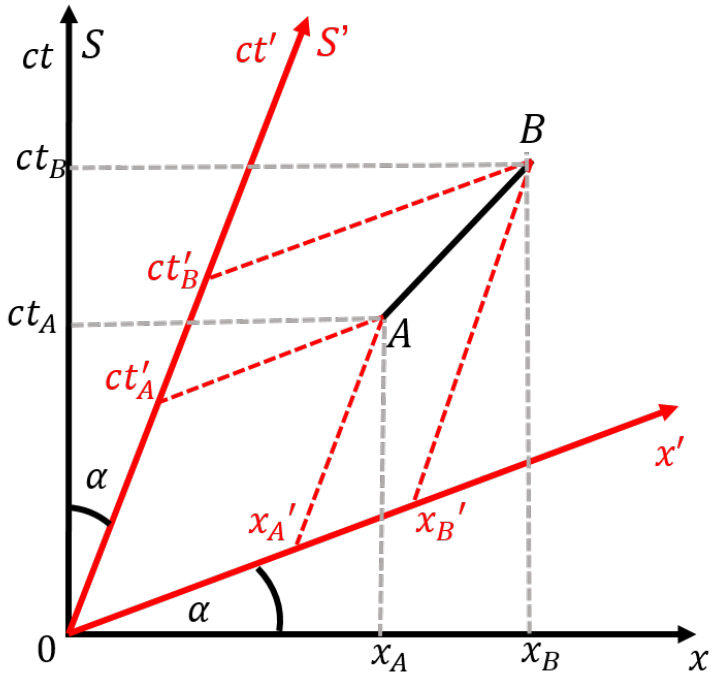


Figure 1.17: The reference frame S' is moving relative to reference frame, S, with speed v in the x direction. We can illustrate this on a space-time diagram by tilting the axes of the S' coordinate system by an angle $\tan\alpha = v/c$, as shown.

The effect of the Lorentz transformation on a space-time diagram is to tilt both the space and time axes "inwards"[2], by an angle, α, given by:

$$\tan\alpha = \frac{v}{c}$$

Figure 24.17 shows a light-like interval between two points, A and B, and how to determine the space-time coordinates in the two reference frames. You can think of space-time as the sheet of paper on which events happen. You can then draw different coordinate systems on that piece of paper to describe the position (in space and time) of different events.

Example 1-5

Use a space-time diagram to show how two events at different locations that are si-

[2]Outwards if the speed of S' is in the negative x direction relative to S.

multaneous in one frame of reference are not simultaneous in a reference frame that is moving relative to the one where the events are simultaneous. This is an illustration of the relativity of simultaneity that we uncovered at the beginning of the chapter when examining Alice on a train platform and the two pulses of light.

Solution

Let S be the frame of reference where events, A and B, are simultaneous. These events are connected by a space-like interval, since they are separated in space, but not in time. There is no way for one event to have caused the other. In the frame, S, these events are on a horizontal line in a space-time diagram.

Let us define a second reference frame, S', that is moving with speed v, relative to S. We have illustrated space-time diagrams for the the two reference frames, and the events A and B, in Figure 24.18.

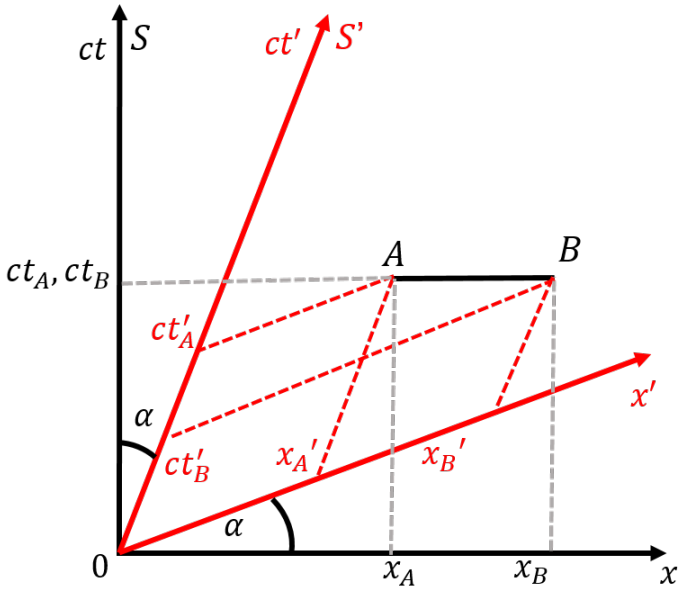


Figure 1.18: In the frame of reference, S, the events A and B occur at the same time. In frame, S', event B occurs before event A (since $t_B < t_A$). The two events are space-like, so observers in different frames of reference cannot agree on which happened first.

From the space-time diagram, it is clear that, in the frame of reference, S', the event B happened first. If the frame of reference S' was moving in the opposite (negative x) direction, event A would occur first, and the axes of S' would be tilted in the other direction (so that the opening angle between the axes is greater than 90°).

Discussion: This example illustrates how space-time diagrams can be used to qualitatively model events in space-time between two different reference frames. In particular, we showed how two events that are simultaneous in one frame (S) are not simultaneous in a different frame. For events connected by space-like intervals, there exist frames of

reference where the events are simultaneous, or where either one happened first. If two events are separated by a time-like interval, there is no frame of reference in which one will disagree in the ordering of the events (although observers in different frames of reference will still measure different lengths of time between events due to time-dilation). For time-like events, the moving frame of reference would have to go faster than the speed of light for the time ordering to be different. This would violate causality, and is a good argument as to why nothing can go faster than the speed of light!

Josh's Thoughts

This chapter is where the bubble of intuitive reality is popped, and students (like you and me) are given the oppourtunity to challenge our understanding of how the universe operates. As amazing and exciting as this is, it can also be incredibly frustrating. Many students rely on intuition to guide them as they problem solve, but hit a wall in special relativity. To avoid this issue, I suggest drawing spacetime diagrams and using Lorentz transformations. Practicing with these tools will help make the process of understanding the strange consequences of Einstein's postulates less awkward.

In addition to the practical advice I have given, I reccommend embracing the strangeness of reality. Throughout history, scientists have ventured into the unknown in attemps to discover and decode the universe. In many cases, the answer to a question has posed more questions than answers, and humanity is given the oppourunity to further understand the world we live in. As a student, you are participating in a process of understanding which allows us to continue the adventure that is scientific inquiry. Confusion can be frustrating, but don't let it discourage you, being confused means that you're only a few steps away from understanding!

1.6.4 Lorentz addition of velocities

In the previous section, we reviewed the Galilean velocity transformations, that allow us to convert a velocity, $\vec{u}$, as measured in one frame of reference, to a velocity, $\vec{u}'$, measured in a different frame of reference. We now derive the equivalent relations based on the Lorentz transformation. Again, we assume that frame, S', moves in the positive x direction with speed, v, relative to frame, S.

The x component of the velocity vector, $\vec{u}'$, for some object in the S' frame of reference is given by:

$$u'_x = \frac{d}{dt'}x'$$

In Galilean relativity, we could simply replace the derivative over t' by a derivative over t, since the two are equivalent. This is no longer the case. However, we can use the Chain

Rule and the Lorentz transformations to convert a derivative over t' to a derivative over t:

$$\begin{aligned}\frac{d}{dt'} &= \frac{dt}{dt'}\frac{d}{dt} \\ &= \frac{d}{dt'}\gamma\left(t' + \frac{vx'}{c^2}\right)\frac{d}{dt} \\ &= \gamma\left(1 + \frac{v}{c^2}\frac{dx'}{dt'}\right)\frac{d}{dt} \\ &= \gamma\left(1 + \frac{vu'_x}{c^2}\right)\frac{d}{dt}\end{aligned}$$

where we recognized that $\frac{dx'}{dt'} = u'_x$. The x component of the velocity, as measured in the S' frame of reference, is then given by:

$$\begin{aligned}u'_x = \frac{d}{dt'}x' &= \gamma\left(1 + \frac{vu'_x}{c^2}\right)\frac{d}{dt}x' \\ &= \gamma\left(1 + \frac{vu'_x}{c^2}\right)\frac{d}{dt}\gamma(x - vt) \\ &= \gamma^2\left(1 + \frac{vu'_x}{c^2}\right)(u_x - v) \\ \frac{u'_x}{\gamma^2} &= u_x - v + \frac{vu'_x u_x}{c^2} - \frac{v^2 u'_x}{c^2} \\ u'_x\left(1 - \frac{v^2}{c^2}\right) &= u_x - v + \frac{vu'_x u_x}{c^2} - \frac{v^2 u'_x}{c^2} \\ u'_x\left(1 - \frac{vu_x}{c^2}\right) &= u_x - v \\ \therefore \boxed{u'_x = \frac{u_x - v}{1 - \frac{vu_x}{c^2}}}\end{aligned}$$

where we made use of the Lorentz transformation: $x' = \gamma(x - vt)$. We can proceed in a similar way to determine the y and z components. Note that, unlike the Galilean case, all of the velocity components must transform, since the time derivative is involved for each component. Intuitively, we expect all components of velocity to be affected, since one needs to guarantee that the total speed is always below c. The velocity transformations for all components are given by the following:

$$\begin{aligned}u'_x &= \frac{u_x - v}{1 - \frac{vu_x}{c^2}} \\ u'_y &= \frac{u_y}{\gamma\left(1 - \frac{vu_x}{c^2}\right)} \\ u'_z &= \frac{u_z}{\gamma\left(1 - \frac{vu_x}{c^2}\right)}\end{aligned}$$

and the reverse transformations are given by:

$$u_x = \frac{u'_x + v}{1 + \frac{vu_x}{c^2}}$$
$$u_y = \frac{u'_y}{\gamma\left(1 + \frac{vu_x}{c^2}\right)}$$
$$u_z = \frac{u'_z}{\gamma\left(1 + \frac{vu_x}{c^2}\right)}$$

Example 1-6

An archer can shoot a very fast arrow with a speed of $0.5c$. The archer is on a train moving with speed, $v = 0.7c$, and fires an arrow in the direction of motion. What is the speed of the arrow, as measured in the frame of reference of the ground?

Solution

Let the train be the frame of reference, S', moving in the positive x direction with speed $v = 0.7c$ relative to the frame, S, which corresponds to the ground. The speed of the arrow, as seen from the train (S'), is given by:

$$u'_x = 0.5c$$

The speed of the arrow, as measured from the ground, is thus given by:

$$\begin{aligned} u_x &= \frac{u'_x + v}{1 + \frac{vu_x}{c^2}} \\ &= \frac{(0.5c) + (0.7c)}{1 + \frac{(0.7c)(0.5c)}{c^2}} \\ \text{`} &= \frac{(1.2c)}{1 + (0.7)(0.5)} = \frac{1.2}{1.35}c = 0.89c \end{aligned}$$

Discussion: By using the Lorentz transformations for velocity, we see that the arrow does not exceed the speed of light. Had we used Galilean relativity, we would have concluded that the arrow has a speed of $1.2c$ when measured from the ground.

1.7 Relativistic momentum and energy

In this section, we show how to define momentum and energy in a way that is consistent with the postulates of Special Relativity. We expect that, since time and space depend on the frame of reference of the observer, so too will the momentum and the energy of an object. Consider an object of mass m_0, moving in a frame of reference, S, with velocity, $\vec{u}$ (we reserve $\vec{v}$ to represent the speed between two inertial frames of reference), in the x

direction. At some time, t, the object will be at position, x, along the x axis. We define the relativistic momentum as:

$$p = m_0 \frac{dx}{dt'}$$

where t' is the time as measured in the rest frame of the object. By defining momentum in terms of the proper time of the object, all observers will agree on the value of t'. In the frame of reference, S, (with time t) this corresponds to:

$$p = m_0 \frac{dx}{dt'} = m_0 \frac{dt}{dt'} \frac{dx}{dt} = m_0 \frac{dt}{dt'} u$$

where u is the speed of the particle in frame, S. We can use time dilation to re-express the derivative:

$$\begin{aligned} \Delta t &= \gamma \Delta t' \\ \frac{\Delta t}{\Delta t'} &= \gamma \\ \therefore \frac{dt}{dt'} &= \gamma \end{aligned}$$

where in the last line, we simply took the limit of an infinitesimally short time interval. Therefore, the relativistic momentum of the particle, in frame, S, can be defined:

$$\boxed{\vec{p} = m_0 \gamma \vec{u} = \frac{m_0 \vec{u}}{\sqrt{1 - \frac{u^2}{c^2}}}}$$

where γ is calculated with the same speed, u, since that is the speed of the reference frame of the object relative to S. Note that as the speed, u, of the particle approaches the speed of light, the factor of γ approaches infinity. This means that an object with a mass can never reach the speed of light, as it would have an infinite momentum. In order to define momentum in a way that resembles the classic definition, one can think of the mass of the object as depending on the speed of the object. We define the rest-mass, m_0, of the object as the mass that is measured when the object is at rest. We can then model the mass of the object as increasing with its speed:

$$m(u) = \gamma m_0 = \frac{m_0}{\sqrt{1 - \frac{u^2}{c^2}}}$$

so that the relativistic momentum would be defined as:

$$\vec{p} = m(u)\vec{u}$$

In this case, we can think of the mass of the object as increasing with its speed. The object would acquire infinite mass if it were to reach the speed of light.

With the relativistic definition of momentum, Newton's Second Law can be written as:

$$\vec{F} = \frac{d\vec{p}}{dt} = \frac{d}{dt} m_0 \gamma \vec{u}$$

Example 1-7

A constant force of $1 \times 10^{-22}\,\text{N}$ is applied to an electron (with mass $m_e = 9.11 \times 10^{-31}\,\text{kg}$) in order to accelerate it from rest to a speed of $u = 0.99c$. Compare the length of time over which the force must be applied using classical and relativistic dynamics.

Solution

In both cases, we can start with Newton's Second Law:

$$\vec{F} = \frac{d\vec{p}}{dt}$$
$$\therefore \int \vec{F} dt = \Delta\vec{p} = \vec{p}$$

where $\vec{p}$ is the final momentum of the electron (which is different depending on whether we use the classical or the relativistic definition of momentum). Since the force is constant:

$$\int \vec{F} dt = \vec{F}\Delta t = \vec{p}$$
$$\therefore \Delta t = \frac{p}{F}$$

where Δt is the length of time over which the force is applied. With the classical definition of momentum, the time is given by:

$$\Delta t = \frac{p}{F} = \frac{mu}{F} = \frac{(9.11 \times 10^{-31}\,\text{kg})(0.99)(3 \times 10^{8}\,\text{m/s})}{(1 \times 10^{-22}\,\text{N})} = 2.71\,\text{s}$$

With the relativistic definition of momentum, we first need the gamma factor:

$$\gamma = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} = \frac{1}{\sqrt{1 - (0.99)^2}} = 7.1$$

We can then calculate the time over which the force needs to be applied:

$$\Delta t = \frac{p}{F} = \frac{\gamma m_0 u}{F} = \gamma \frac{m_0 u}{F} = (7.1)(2.71\,\text{s}) = 19.2\,\text{s}$$

Discussion: When using the relativistic definition of momentum, we find that the time over which the force must be applied to reach a given speed is longer. This makes sense, since it will take infinitely long to reach the speed of light. Also, note that the time that is required using relativistic dynamics is just the time-dilated time that is required in classical dynamics.

Recall how we defined kinetic energy, in Section 7.2, by defining the change in kinetic energy of an object as the net work done on that object. We use the same formalism here to redefine

kinetic energy using relativistic dynamics.

The work done by the net force, $\vec{F}$, on an object that goes from a position A to a position B, is given by

$$W = \int_A^B \vec{F} \cdot d\vec{l} = \int_0^t \left(\frac{d}{dt} m_0 \gamma \vec{u} \right) \cdot (\vec{u} dt)$$

where we recognized that a infinitesimal segment $d\vec{l}$ along the path of the object is given by $d\vec{l} = \vec{u}dt$. The time infinitesimals, dt, cancel, and we are left with:

$$\begin{aligned} W &= \int_0^t \left(\frac{d}{dt} m_0 \gamma \vec{u} \right) \cdot (\vec{u} dt) \\ &= \int d(m_0 \gamma \vec{u}) \cdot \vec{u} \end{aligned}$$

which we can integrate by parts. We can integrate this over the speed, u, and we assume that the object started with a speed of $u = 0$ at the beginning of the path and has a speed, $u = U$, at the end of the path:

$$\begin{aligned} W &= \int_0^U d(m_0 \gamma \vec{u}) \cdot \vec{u} = \Big[\gamma m_0 \vec{u} \cdot \vec{u} \Big]_0^U - \int_0^U m_0 \gamma u du \quad \text{(int. by parts)} \\ &= \gamma m_0 U^2 - m_0 \int_0^U \frac{u du}{\sqrt{1 - \frac{u^2}{c^2}}} \\ &= \gamma m_0 U^2 - m_0 \Big[c^2 \sqrt{1 - \frac{u^2}{c^2}} \Big]_0^U \\ &= \gamma m_0 U^2 - m_0 c^2 + m_0 c^2 \sqrt{1 - \frac{U^2}{c^2}} \\ &= \gamma \left(m_0 U^2 + m_0 c^2 \left(1 - \frac{U^2}{c^2} \right) \right) - m_0 c^2 \\ &= m_0 c^2 (\gamma - 1) \end{aligned}$$

Since the object started at rest (with a speed $u = 0$) the above integral corresponds to what we would call the kinetic energy of the object, with a speed, u:

$$K = m_0 c^2 (\gamma - 1) = m_0 c^2 \left(\frac{1}{\sqrt{1 - \frac{u^2}{c^2}}} - 1 \right)$$

This form for the relativistic kinetic energy of the object is not at all similar to the form that we obtained in classical physics. As the speed of the object approaches the speed of light, the γ factor approaches infinity, as does the kinetic energy. Thus, it would take an infinite amount of work to accelerate an object to the speed of light, and again, we see that it is impossible for anything with mass to ever reach the speed of light. The formula above, however, should always be correct, even in the non-relativistic limit, when $v << c$. We can approximate the gamma factor using the binomial expansion for the case where $x << 1$:

$$(1 + x)^n \sim 1 + nx + \ldots$$

So that, when $v << c$ (and $v^2/c^2 << 1$), the gamma factor is approximated by:

$$\gamma = \left(1 - \frac{u^2}{c^2}\right)^{-\frac{1}{2}} \sim 1 + \frac{1}{2}\frac{u^2}{c^2}$$

In this limit, the relativistic kinetic energy reduces to:

$$\lim_{v<<c} K = \lim_{v<<c} m_0c^2(\gamma - 1) \sim m_0c^2\left(1 + \frac{1}{2}\frac{u^2}{c^2} - 1\right) = \frac{1}{2}mu^2$$

which is the classical definition of kinetic energy. The kinetic energy is also zero when the speed is zero.

The kinetic energy has two terms in it:

$$K = m_0c^2\gamma - m_0c^2$$

The first term increases with speed and behaves as we would expect. The second term is constant, and depends only on the rest mass of the object (we call this term the rest mass energy). We can think of this in slightly different terms. Let us define the total energy, E, of the object as:

$$E = m_0c^2\gamma$$
$$\therefore \boxed{E = K + m_0c^2}$$

so that the total energy is just the rest mass energy plus the kinetic energy. This highlights a key aspect of Special Relativity. An object will have energy, E, even when it is at rest. That energy, at rest, is called the rest mass energy, and corresponds to energy that an object has by virtue of having mass. This is, of course, Einstein's famous equation:

$$\boxed{E = m_0c^2} \quad \text{(rest mass energy)}$$

This equation implies that mass can be thought of as a form of energy. Nuclear reactors function by converting a small amount of mass of uranium atoms into energy (in the form of heat), that is then used to produce high pressure steam to rotate a turbine.

Einstein's relation is often used to express the mass of subatomic particles in terms of energy. For example, an electron has a mass of $511 \times 10^3\,\text{eV/c}^2$ in these units.

Example 1-8

What is the mass of a proton, $m_p = 1.67 \times 10^{-27}\,\text{kg}$, in units of MeV/c^2 (where the M stands for "Mega", and corresponds to $1\,\text{MeV} = 1 \times 10^6\,\text{eV}$)?

Solution

We can first calculate the rest mass energy of the proton in Joules:

$$E = m_pc^2 = (1.67 \times 10^{-27}\,\text{kg})(3 \times 10^8\,\text{m/s})^2 = 1.503 \times 10^{-10}\,\text{J}$$

We can then convert from Joules to electron-volts:

$$\frac{(1.503 \times 10^{-10}\,\mathrm{J})}{(1.6 \times 10^{-19}\,\mathrm{J/eV})} = 939.4 \times 10^{6}\,\mathrm{eV} = 939.4\,\mathrm{MeV}$$

The mass of the proton can then be expressed as $m_p = 939.4\,\mathrm{MeV/c^2}$.

Finally, it is interesting to examine the relationship between the momentum and the energy of a relativistic object. Consider the quantity c^2p^2:

$$\begin{aligned} c^2p^2 &= c^2(\gamma m_0 u)^2 = c^2\gamma^2 m_0^2 u^2 = c^4\gamma^2 m_0^2 \frac{u^2}{c^2} = c^4\gamma^2 m_0^2\left(1 - \frac{1}{\gamma^2}\right) \\ &= c^4\gamma^2 m_0^2 - c^4 m_0^2 \\ &= E^2 - c^4 m_0^2 \end{aligned}$$

where we recognized that $c^4\gamma^2 m_0^2$ is simply the energy, E, squared. This is generally called the "energy-momentum" relation and written:

$$\boxed{E^2 = p^2c^2 + m_0^2c^4}$$

An interesting consequence of this relationship is that particles with no mass will still have a momentum. For example, the photon, which is a particle of light and must thus have a mass of zero (or it could not move at the speed of light), will have a momentum given by:

$$p = \frac{E}{c}$$

Thus, one can use light to impart momentum to something. This is how a solar sail, a proposed propulsion mechanism for space travel, operates.

1.8 Closing remarks

In this chapter, we introduced the first hints of how the laws of physics become counter-intuitive, and quite bizarre. One can wrap one's head around Newton's Second Law, $\vec{F}^{net} = m\vec{a}$, and develop some intuition as to how an object may behave. However, it is difficult to imagine how people age slower if they travel faster, and how cars become shorter when they are moving. However, as far as we can tell, this is the best way to describe the Universe around us.

This all goes back to our original statements about physics. The goal is to come up with rules that allow us to describe Nature. It's nice when those rules make sense, but, unfortunately, that is not a requirement. It does appear that the rules that describe Nature do not make sense, at least not based on our common experience, living in a macroscopic world where speeds are much less than the speed of light. With Special Relativity, we introduced the modern framework for modelling dynamics. We have not introduced Quantum Mechanics, which describes how elementary particles behave.

Quantum Mechanics is even less intuitive than Special Relativity, as it implies that particles act as if they are in multiple places at the same time. Even worse, Quantum Mechanics

requires us to abandon the concept of determinism that is critical in Classical Mechanics; in Quantum Mechanics, we can only ever determine probabilities. For example, we can only determine the probability that a particle will be at a particular location at a particular time, but we cannot use kinematics and dynamics to predict where it will be at some time based on the forces acting upon it.

If you decide to pursue further studies in physics, you will get to learn more about these theories, which are quite marvellous. It should not bother you that physics is not intuitive, as that is not the purpose. The exciting part of physics is that, even if Nature behaves in an exquisitely weird way, it does appear that this can all be described with a rather limited set of mathematical equations. One can argue that there is beauty in the fact that succinct mathematics can describe a large number of seemingly unrelated phenomena, as Newton's Universal Theory of Gravity was able to describe both the motion of a falling apple and the orbit of the moon.

1.9 Summary

Key Takeaways

The Theory of Special Relativity is based on Einstein's two postulates:

1. The laws of physics are the same in all inertial reference frames. There is no experiment that can be performed to determine whether one is at rest or moving with constant velocity.
2. The speed of light propagating in vacuum is the same in all inertial reference frames. Any observer in an inertial frame of reference, regardless of their velocity, will measure that light has a speed of c, when it propagates in vacuum.

These postulates are required in order for the equations from electromagnetism to be valid in all inertial frames of reference. However, they lead to very counter-intuitive results. For example, if two events, A and B, are simultaneous in one frame of reference, an observer in a different frame of reference will observe event A to happen earlier/later than event B (earlier or later will depend on the direction of motion of the moving observer).

The Theory of Special Relativity allows us to relate observations made in one inertial frame of reference, S, to observations made in a different inertial frame of reference, S', that is moving with constant velocity, $\vec{v}$, relative to S. We always choose to define the x axis in the S and S' frames of reference so that they are both co-linear with the velocity of S', $\vec{v}$, which is defined to be in the positive x direction in frame, S. Furthermore, we assume that the origin of both frames of reference coincided at time $t = 0$.

We define the gamma factor, γ, based on the speed, v, of S' relative to S:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

The gamma factor is always greater or equal to 1.

If a time interval, Δt, is measured in frame, S, then a "dilated" time interval, $\Delta t'$, will be measured in frame S':

$$\Delta t' = \gamma \Delta t$$

since $\gamma \geq 1$. We call the time that is measured in a frame of reference that we consider "at rest" to be the "proper time" in that frame of reference. For example, a muon decays in $2.2\,\mu$s when at rest. If a muon moves at high speed, in the frame of reference where the muon is moving, it will take *longer* (time dilation), for the muon to decay. The time $2.2\,\mu$s is the "proper time" for the muon decay (since it is measured when the muon is at rest).

As a consequence of time dilation, observers in different frames of reference will measure different lengths due to "length contraction". If an object has a "proper length", L, in

a frame of reference, S, that is at rest relative to the object, the object will have a contracted length, L', in a reference frame, S', moving with speed, v, relative to S:

$$L' = \frac{L}{\gamma}$$

Note that only the dimension of the object that is co-linear with the velocity vector, $\vec{v}$, is contracted.

We also noted that Special Relativity is intimately connected to electromagnetism. In particular, we described how what we model as a magnetic force in one frame of reference might be modelled as an electric force in a different frame of reference.

In order to describe the motion of objects, we found that we need to define a four-dimensional space-time, where positions in space-time are labelled by 4 "coordinates", (x, y, z, ct), instead of the usual 3 (space) position coordinates. This is a result of the fact that time is no longer absolute and depends on the frame of reference (e.g. time dilation).

In space-time, we think in terms of events that occur at specific locations in space and instants in time. We can visualise space-time using "space-time diagrams", where one axis corresponds to space (x), and the other axis corresponds to time (ct). The path of an object through space-time is called its "world line".

For a given event in space-time, we can define past and future "light cones". Only events in the past light-cone could have had a causal effect on the event. Similarly, only events in the future light-cone can ever be influenced by that event. Events that can be causally connected (within each other's light cones) are said to be "time-like". Events that are outside of each other's light cones are said to be "space-like". If two events are time-like, all observers will agree on the order in which the events happened, preserving the notion of causality. Different observers can disagree on the order in which space-like events occurred.

The Lorentz transformations allow us to convert the coordinates of events in one frame of reference, S, to those in a frames, S', moving with constant speed, v, relative to S:

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

and the inverse relations are easily found:

$$\begin{aligned} x &= \gamma(x' + vt') \\ y &= y' \\ z &= z' \\ t &= \gamma\left(t' + \frac{vx'}{c^2}\right) \end{aligned}$$

Certain quantities, which are measured to be the same in all frames of reference, are said to be "Lorentz invariant". In particular, we can define the space-time interval, s, between two events in space-time as:

$$s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2\Delta t^2$$

One can think of this as a sort of "distance" in space-time, that does not depend on the frame of reference.

If an object has a velocity vector, $\vec{u}$, as measured in frame of reference S, then its velocity, $\vec{u}'$, in a frame, S', moving with speed, v, relative to S, is given by:

$$\begin{aligned} u'_x &= \frac{u_x - v}{1 - \frac{vu_x}{c^2}} \\ u'_y &= \frac{u_y}{\gamma\left(1 - \frac{vu_x}{c^2}\right)} \\ u'_z &= \frac{u_z}{\gamma\left(1 - \frac{vu_x}{c^2}\right)} \end{aligned}$$

and the reverse transformations are given by:

$$\begin{aligned} u_x &= \frac{u'_x + v}{1 + \frac{vu_x}{c^2}} \\ u_y &= \frac{u'_y}{\gamma\left(1 + \frac{vu_x}{c^2}\right)} \\ u_z &= \frac{u'_z}{\gamma\left(1 + \frac{vu_x}{c^2}\right)} \end{aligned}$$

In order for momentum and energy to be conserved in Special Relativity, these need to be redefined. If a particles with rest mass, m_0, has a velocity, $\vec{u}$, in an inertial frame of reference, its relativistic momentum, $\vec{p}$, is defined to be:

$$\vec{p} = \gamma m_0 \vec{u}$$

where the gamma factor is evaluated using the speed, u:

$$\gamma = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}}$$

This relativistic definition of momentum is equivalent to the classical definition when $u << c$. We can think of relativistic momentum in the same way as classical momentum, if we model the mass of the object as increasing with its speed:

$$m(u) = \gamma m_0$$
$$\therefore \vec{p} = m(u)\vec{u}$$

where m_0 is the mass of the object measured when the object is at rest (its "rest mass"). An object with a rest mass can never reach the speed of light, as this would correspond to it having infinite momentum (or infinite mass).

With the relativistic definition of momentum, one can still use Newton's Second Law in the form:

$$\vec{F} = \frac{d\vec{p}}{dt}$$

We define the total energy, E, of an object as:

$$E = K + m_0 c^2$$

which has a contribution from its kinetic energy, K, and from its mass (the second term). The energy that an object has by virtue of having a mass is called "rest mass energy", which implies that mass and energy can really be thought of as the same thing; one can convert mass into energy and vice versa (as in a nuclear reactor).

The kinetic energy of an object moving with speed, u, is given by:

$$K = m_0 c^2(\gamma - 1)$$

where the gamma factor is obtained using the speed, u. This relativistic definition of kinetic energy is equivalent to the classical definition when $u << c$. The total energy of a particle can also be written as:

$$E = \gamma m_0 c^2$$

Since energy and mass are simply related by a constant, one can use units of energy to describe the mass of a particle. It is common in particle physics to express the mass of particles in units of MeV/c^2.

Finally, we saw that the relativistic momentum and energy of an object are related:

$$E^2 = p^2c^2 + m_0^2c^4$$

In particular, particles of light, which have no mass but have kinetic energy, have non-zero momentum:

$$p = \frac{E}{c}$$

Important Equations

Lorentz factor:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Time dilation

$$\Delta t' = \gamma \Delta t$$

Length contraction

$$L' = \frac{L}{\gamma}$$

Lorentz transformations:

$$\begin{aligned} x' &= \gamma(x - vt) \\ y' &= y \\ z' &= z \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned}$$

Velocity addition:

$$\begin{aligned} u'_x &= \frac{u_x - v}{1 - \frac{vu_x}{c^2}} \\ u'_y &= \frac{u_y}{\gamma\left(1 - \frac{vu_x}{c^2}\right)} \\ u'_z &= \frac{u_z}{\gamma\left(1 - \frac{vu_x}{c^2}\right)} \end{aligned}$$

The spacetime interval:

$$s^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2\Delta t^t$$

Relativistic momentum:

$$\vec{p} = \gamma m_0 \vec{u}$$

Relativistic energy:

$$E = \gamma m_0 c^2 = K + m_0 c^2$$

Relativistic kinetic energy:

$$K = (\gamma - 1)m_0 c^2$$

Newton's Second Law

$$\vec{F} = \frac{d\vec{p}}{dt}$$

Energy-momentum relation:

$$E^2 = p^2c^2 + m_0^2c^4$$

Important Definitions

Proper time: The time measured in a frame of reference considered at rest. SI units: [s]. Common variable(s): Δt.

Proper length: The length of an object as measured at rest relative to the object. SI units: [m]. Common variable(s): L.

1.10 Thinking about the material

Reflect and research

1. How did Michelson and Morley demonstrate that the ether does not exist?
2. Why is 1905 the "year of physics"?
3. Give an example of a device that you use that is affected by relativistic effects.
4. How do you resolve the twin paradox? Can you show it on a space-time diagram?
5. What did Lorentz do and when?
6. Apart from the space-time interval, s, what else is Lorentz invariant?
7. What is Cherenkov radiation?

To try at home

1. Build a particle accelerator.
2. Look up a video illustrating the barn paradox, and other relativistic effects.

To try in the lab

1. Propose an experiment to measure the speed of light.
2. Propose an experiment to test relativistic effects with electro magnetism.

1.11 Sample problems and solutions

1.11.1 Problems

Problem 1-1: The Sun is powered by nuclear fusion reactions in which, predominantly, hydrogen atoms are fused together into helium atoms. Inside the Sun, the material, mostly hydrogen, is in a form of a plasma, where the electrons are not attached to the nuclei of their atoms. Effectively, one can model the solar fusion reactions[3] as:

$$4p + 2e^- \rightarrow \alpha + 2\nu$$

where the four protons correspond to the nuclei of four hydrogen atoms, α is the nucleus of a helium atom, with two neutrons and two protons, and the two ν are neutrinos, particles with virtually zero mass. The reaction above is exothermic, and releases energy, because the total mass of particles on the right is less than the total mass on the left. Given that the mass of a proton is $m_p = 938.3\,\mathrm{MeV/c^2}$, the mass of an electron is $m_e = 0.511\,\mathrm{MeV/c^2}$, and the mass of the alpha particle is $m_\alpha = 3727.4\,\mathrm{MeV/c^2}$, how much energy (in MeV and in J) is released in each fusion reaction? (Solution)

Problem 1-2: A proton is measured by a scientist to have a total energy of $2.5 \times 10^3\,\mathrm{MeV}$. (Solution)

a) What is the speed of the proton?
b) How far does the proton travel (in the lab) when 1 s goes by in the scientist's frame of reference?
c) How far does the proton travel (in the lab) when 1 s goes by in the proton's frame of reference?

[3] In reality, there are many more reactions involved in getting from hydrogen to helium.

1.11.2 Solutions

Solution to problem 24-2: In order to determine the amount of energy released in each reaction, we need to determine the difference in mass between the two sides of the equation:

$$4p + 2e^- \rightarrow \alpha + 2\nu$$

On the left-hand side, the total mass is:

$$M_{LHS} = 4m_p + 2m_e = 4(938.3\,\mathrm{MeV/c^2}) + 2(0.511\,\mathrm{MeV/c^2}) = 3754.22\,\mathrm{MeV/c^2}$$

whereas on the right-hand side, the total mass is:

$$M_{RHS} = m_\alpha = 3727.4\,\mathrm{MeV/c^2}$$

Thus, the total energy released in each reaction is given by:

$$\begin{aligned} E = c^2\Delta M = c^2(M_{LHS} - M_{RHS}) &= c^2((3754.22\,\mathrm{MeV/c^2}) - (3727.4\,\mathrm{MeV/c^2})) \\ &= 26.8\,\mathrm{MeV} = 4.29 \times 10^{-12}\,\mathrm{J} \end{aligned}$$

where we showed the answer in both MeV and J. Although it may not seem like that much energy per reaction, keep in mind that there are of order 1×10^{38} reactions per second in the Sun, corresponding to a power output of order $4 \times 10^{26}\,\mathrm{W}$, enough to keep us warm in the summer.

Solution to problem 24-2:

a) From the total energy, we can calculate the gamma factor, which will give us the velocity of the proton (in the reference frame of the scientist):

$$\begin{aligned} E &= \gamma m_0 c^2 \\ \frac{1}{\gamma} &= \frac{m_0 c^2}{E} \\ \sqrt{1 - \frac{v^2}{c^2}} &= \frac{m_0 c^2}{E} \\ \frac{v^2}{c^2} &= 1 - \frac{m_0^2 c^4}{E^2} \\ \therefore v &= \left(\sqrt{1 - \frac{m_0 c^4}{E^2}}\right) c \\ &= \left(\sqrt{1 - \frac{(938.3\,\mathrm{MeV/c^2})^2 c^4}{(2.5 \times 10^3\,\mathrm{MeV})^2}}\right) c \\ &= \left(\sqrt{1 - \frac{(938.3\,\mathrm{MeV})^2}{(2.5 \times 10^3\,\mathrm{MeV})^2}}\right) c \\ &= 0.92c \\ &= 2.76 \times 10^8\,\mathrm{m/s} \end{aligned}$$

b) In the frame of the lab, when one second goes by, the proton will travel a distance:

$$d = vt = (2.76 \times 10^8\,\mathrm{m/s})(1\,\mathrm{s}) = 2.76 \times 10^8\,\mathrm{m}$$

c) In order to find out how far the proton travels in the lab when one second of proper time goes by in the proton's frame of reference, we need to determine how much time went by in the lab's frame of reference.
The gamma factor for the proton can be obtained from the speed that we determined in part a), or from the total energy directly:

$$\gamma = \frac{E}{m_0 c^2} = \frac{(2.5 \times 10^3\,\mathrm{MeV})}{(938.3\,\mathrm{MeV/c^2})c^2} = \frac{(2.5 \times 10^3\,\mathrm{MeV})}{(938.3\,\mathrm{MeV})} = 2.66$$

Thus, when $\Delta t = 1\,\mathrm{s}$ elapses in the proton's frame of reference, a time dilated time, $\Delta t'$, elapses in the lab frame of reference:

$$\Delta t' = \gamma \Delta t = 2.66\,\mathrm{s}$$

In the lab frame, the proton will travel a distance:

$$d = vt = (2.76 \times 10^8\,\mathrm{m/s})(2.66\,\mathrm{s}) = 7.34 \times 10^8\,\mathrm{m}$$

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).

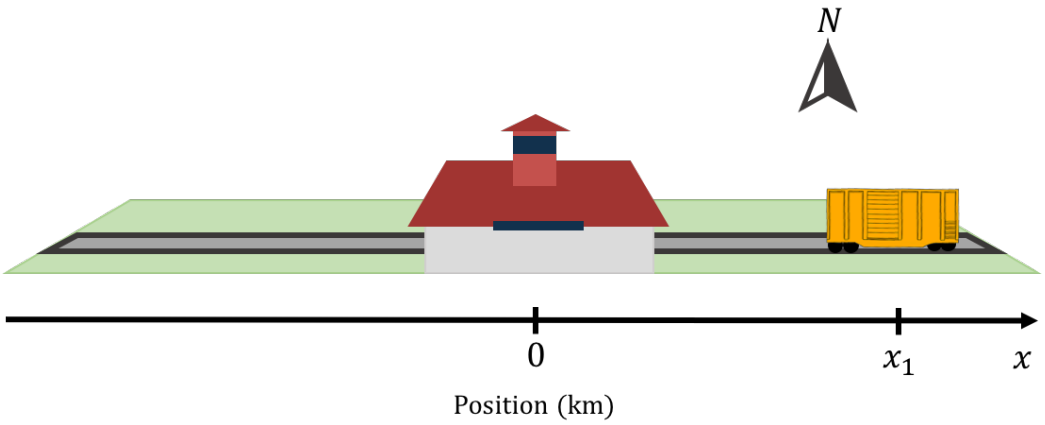


Figure A.1: A 1d coordinate system describing the position of a train. The Kingston train station is the origin and the East side of the track corresponds to positive values of x. The train is located at position x_1.

A.1.2 2D Coordinate systems

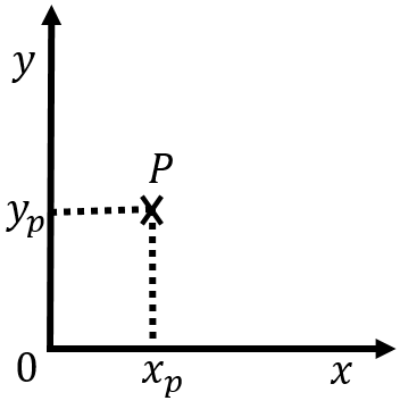


Figure A.2: Example of Cartesian coordinate system and a point P with coordinates (x_p, y_p).

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?

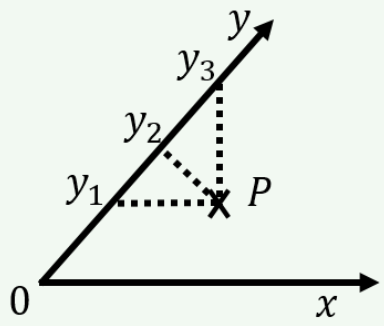


Figure A.3: A non-orthogonal coordinate system (the x and y axes are not perpendicular).

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.

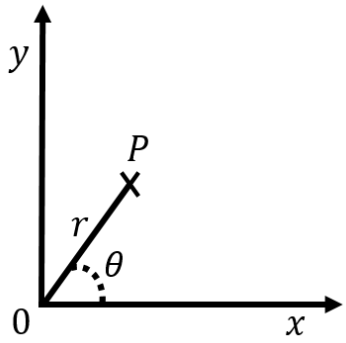


Figure A.4: Example of a polar coordinate system and a point P with coordinates (r, θ).

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.

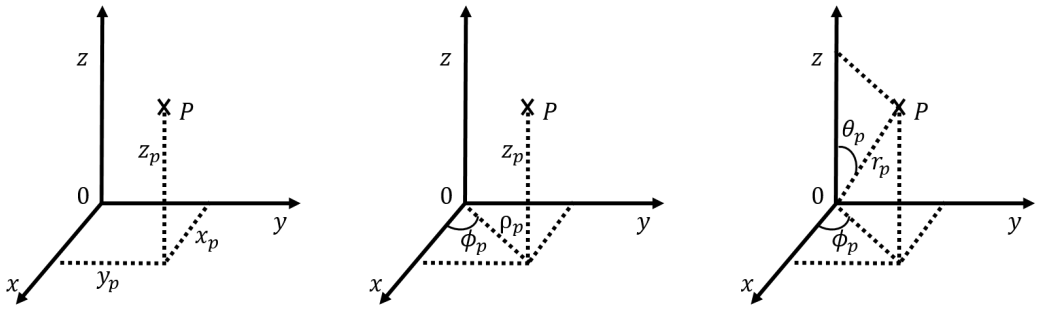


Figure A.5: Cartesian (left), cylindrical (centre) and spherical (right) coordinate systems used in three dimensions. The y and z axes are in the plane of the page, whereas the x axis comes out of the page.

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.

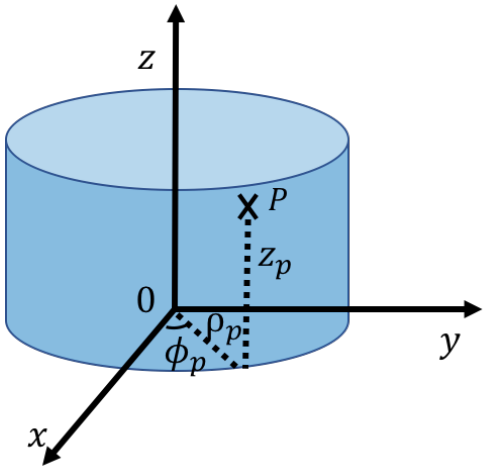


Figure A.6: Describing the position of P, located on the surface of a cylinder, in cylindrical coordinates.

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.

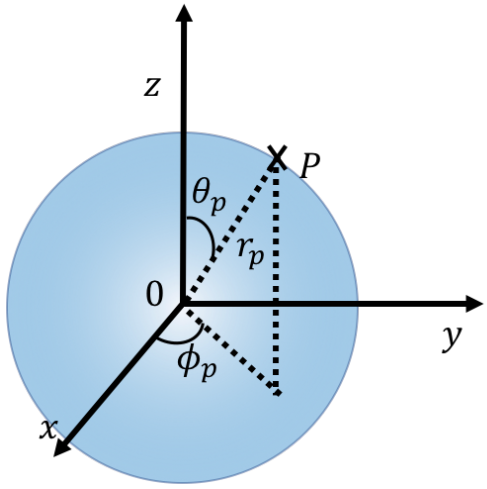


Figure A.7: Describing the position of P, located on the surface of a sphere, in spherical coordinates.

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).

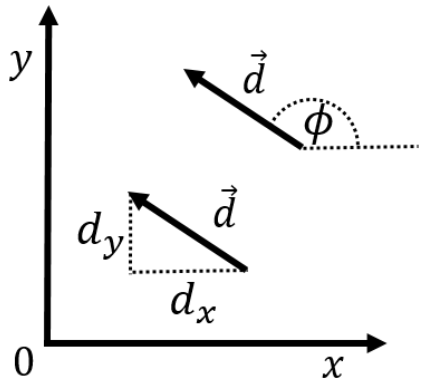


Figure A.8: A vector $\vec{d}$ shown twice, once with its Cartesian components (d_x, d_y) and once with its magnitude and direction (d, ϕ).

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:

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

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.

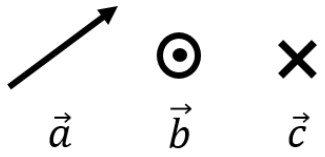


Figure A.9: Geometric representation of three vectors. The vector $\vec{a}$ lies in the plane of the page, the vector $\vec{b}$ is pointing out of the page, and the vector $\vec{c}$ is pointing into the page.

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 $\vec{b}$.

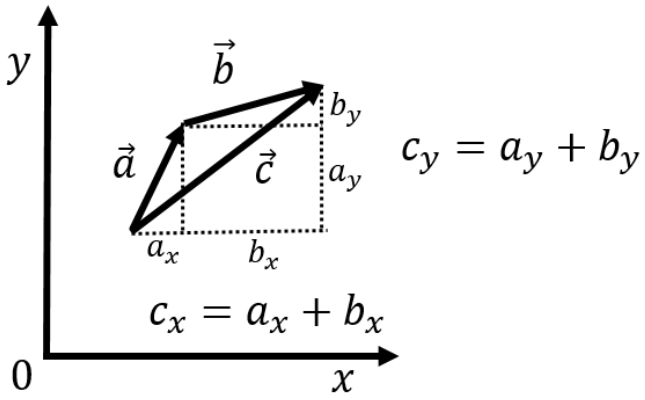


Figure A.10: Geometric addition of the vectors $\vec{a}$ and $\vec{b}$ by placing them "head to tail".

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".

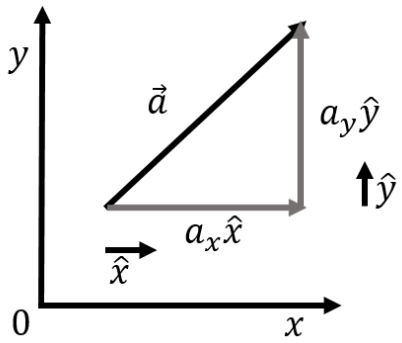


Figure A.11: Illustration that the notation $\vec{a} = a_x\hat{x} + a_y\hat{y}$ is in fact the vector addition of $a_x\hat{x}$ and $a_y\hat{y}$.

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

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

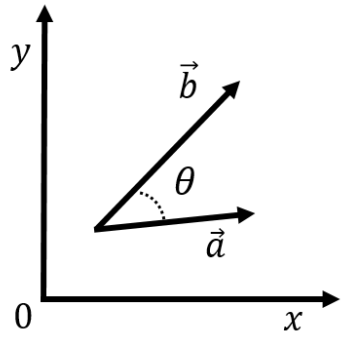


Figure A.12: Illustration of the angle between vectors $\vec{a}$ and $\vec{b}$ when these are placed tail to tail.

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.

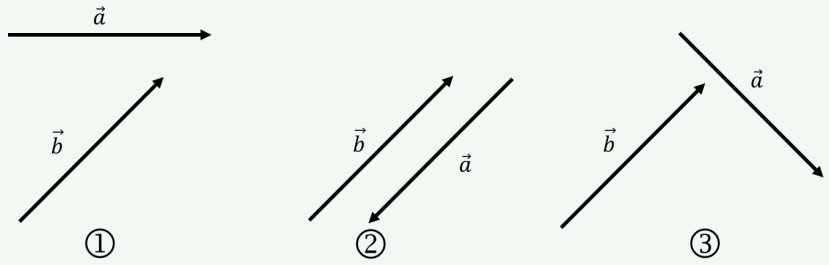


Figure A.13: Put these in order of the magnitude of their scalar product.

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.

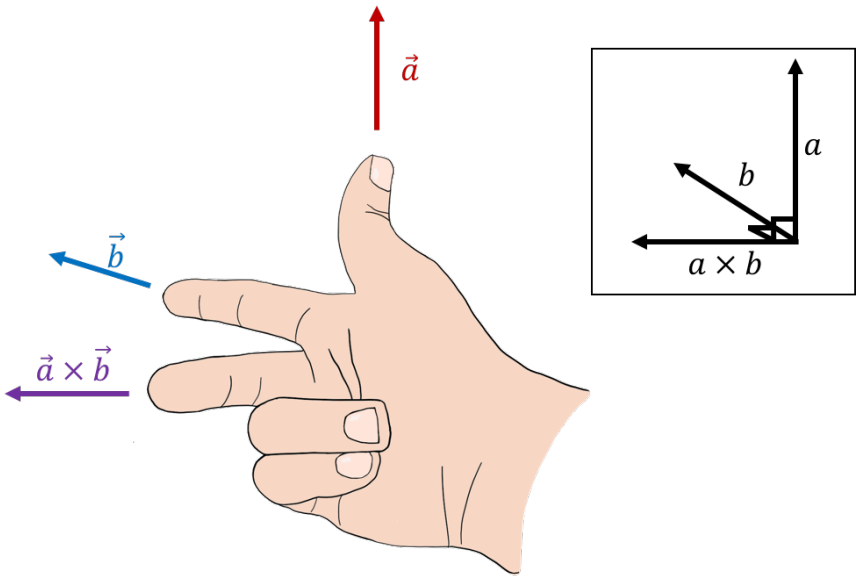


Figure A.14: Using the right hand rule for cross products to find the direction of the cross product of vectors $\vec{a}$ (upwards) and $\vec{b}$ (into the page).

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.

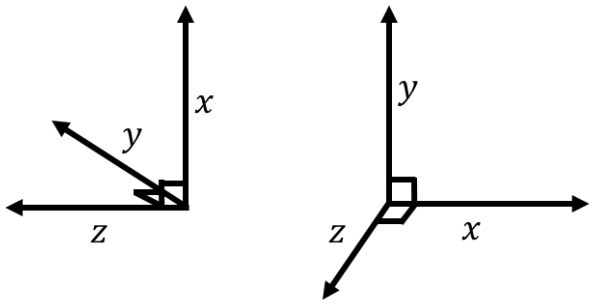


Figure A.15: Two possible orientations for a three dimensional coordinate system. You can confirm using the right hand rule that the z axis is the cross product $\vec{x} \times \vec{y}$.

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 3 and 4). 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.

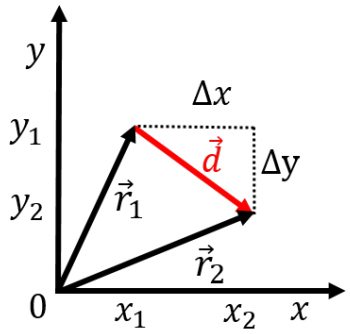


Figure A.16: Illustration of a displacement vector, $\vec{d} = \vec{r}_2 - \vec{r}_1$, for an object that was located at position $\vec{r}_1$ at time t_1 and at position $\vec{r}_2$ at time t_2.

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 7). 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 $\vec{d}$).

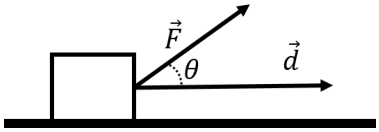


Figure A.17: Example of a force $\vec{F}$ being applied on an object as it moves along the displacement vector $\vec{d}$.

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:

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

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. If 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.

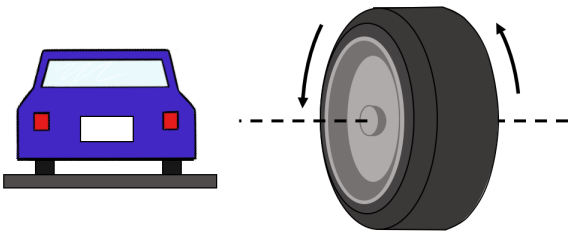


Figure A.18: The wheels on a car that is driving away from you.

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.

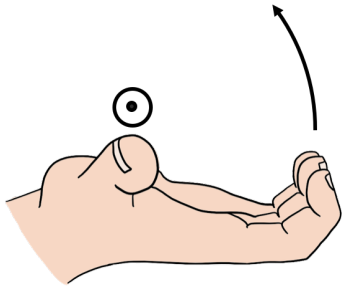


Figure A.19: Using the right hand rule for axial quantities. In this case, the direction of rotation is counter clockwise when looking at the page (the direction that the fingers curl), so the rotation vector points out of the page (the direction of the thumb).

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."

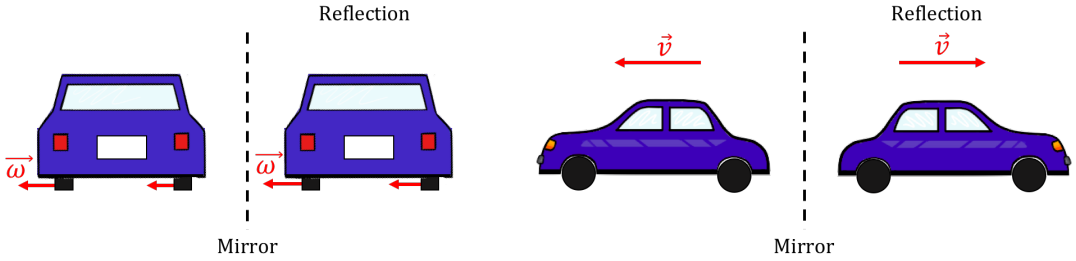


Figure A.20: Left: The angular velocity vector for the rotation of the wheels, $\vec{\omega}$, which points to the left, also points left in the reflection. Right: The velocity vector, pointing to the left, points to the right in the reflection of the car. The angular velocity vector is called an "axial" 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.

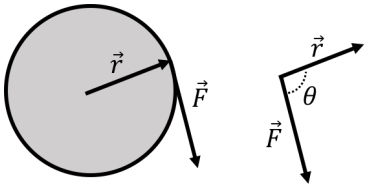


Figure A.21: A force, $\vec{F}$, is exerted in the plane of a disk at a position given by the vector $\vec{r}$ relative to the centre of the disk.

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

A.7.1 Problems

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:

$$\begin{aligned} f(x) &= x^2 \\ \therefore f(2) &= 4 \end{aligned}$$

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:

$$\begin{aligned} g(x, y) &= x^2 + 2y^2 \\ \therefore g(2, 3) &= 22 \end{aligned}$$

We can easily visualize a function of 1 variable by plotting it, as in Figure B.1.

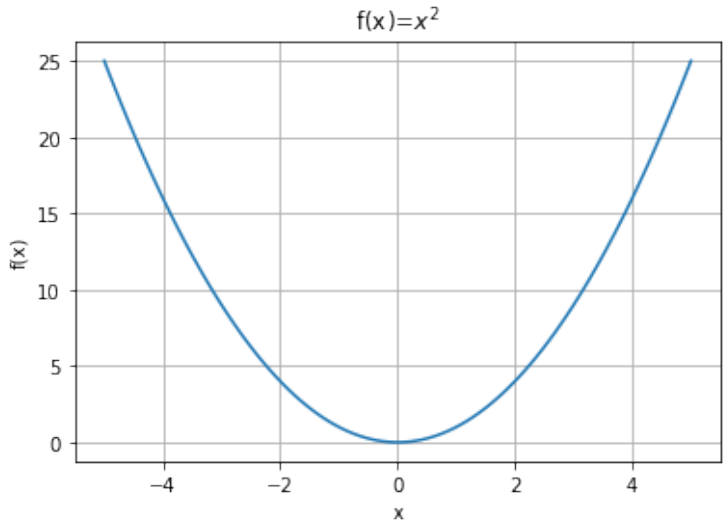


Figure B.1: $f(x) = x^2$ plotted between $x = -5$ and $= +5$.

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 $g(x, y)$). Figure B.2 shows an example of plotting a function of 2 variables.

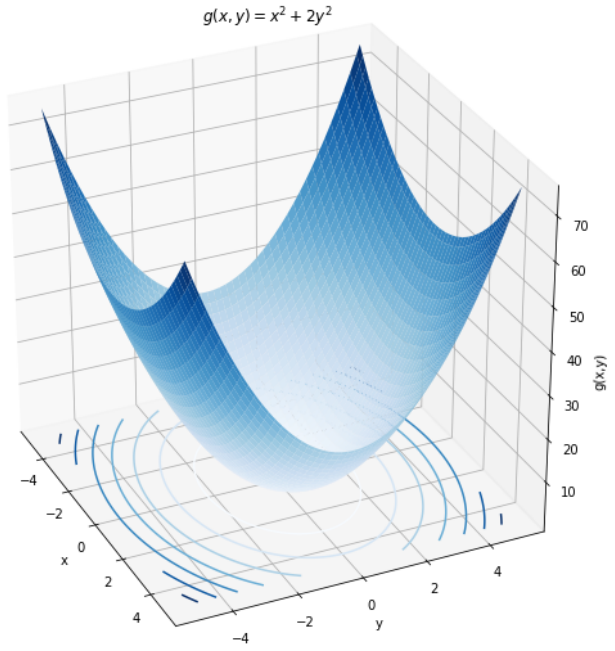


Figure B.2: $g(x, y) = x^2 + 2y^2$ plotted for x between -5 and +5 and for y between -5 and +5. A function of two variables can be visualized as a surface in three dimensions. One can also visualize the function by look at its "contours" (the lines drawn in the xy plane).

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.

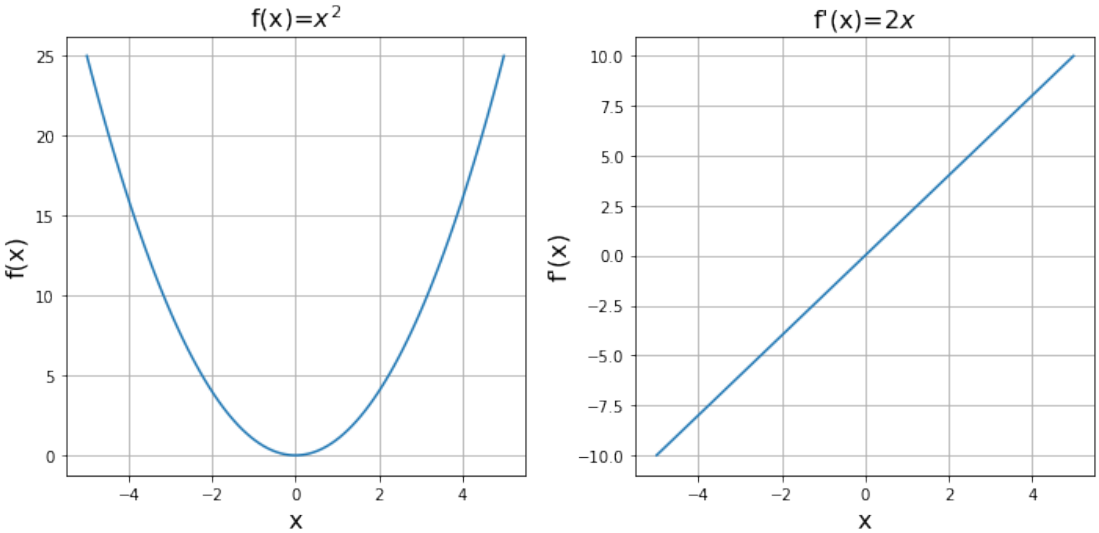


Figure B.3: $f(x) = x^2$ and its derivative, $f'(x) = 2x$ plotted for x between -5 and +5.

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 document 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.

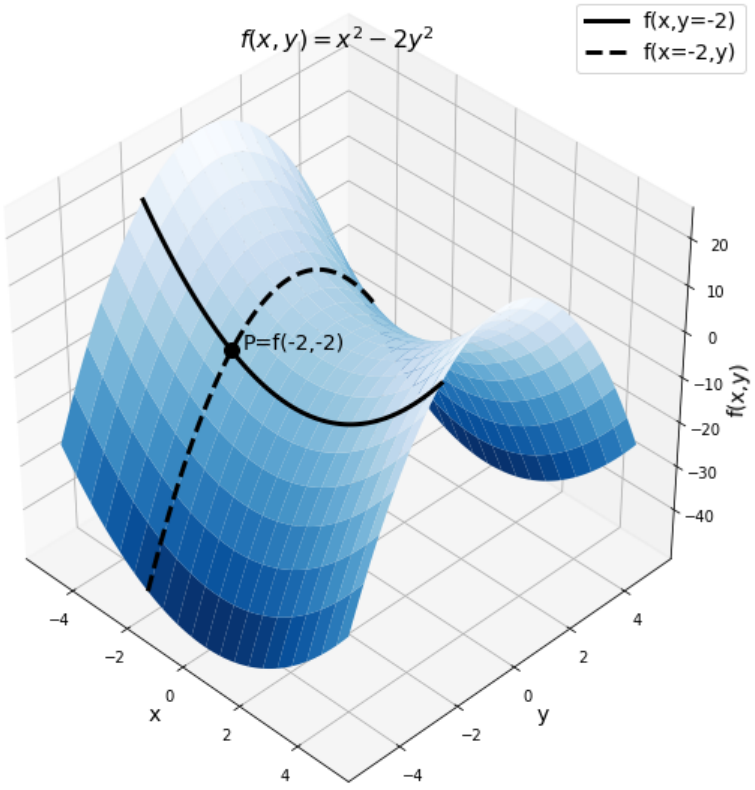


Figure B.4: $f(x, y) = x^2 - 2y^2$ plotted for x between -5 and +5 and for y between -5 and +5. The point P labelled on the figure shows the value of the function at $f(-2, -2)$. The two lines show the function evaluated when one of x or y is held constant.

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). 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

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

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, $F(x_0) = F_0$.

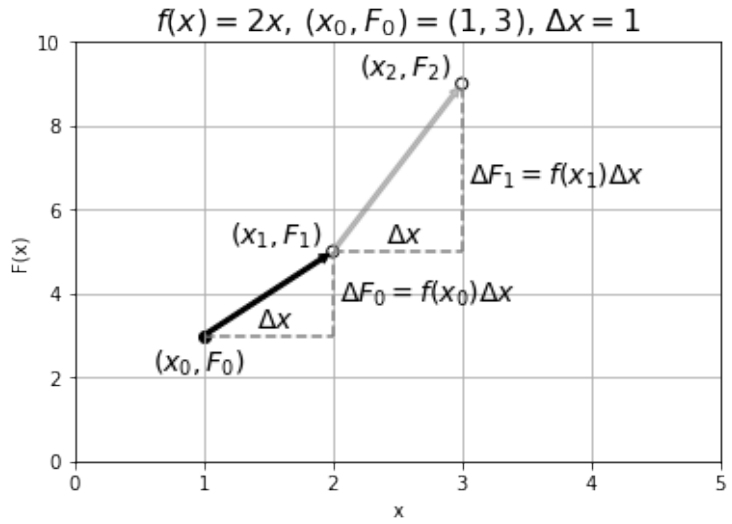


Figure B.5: Determining the anti-derivative, $F(x)$, given the function $f(x) = 2x$ and the initial condition that $F(x)$ passes through the point $(x_0, F_0) = (1, 3)$.

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).

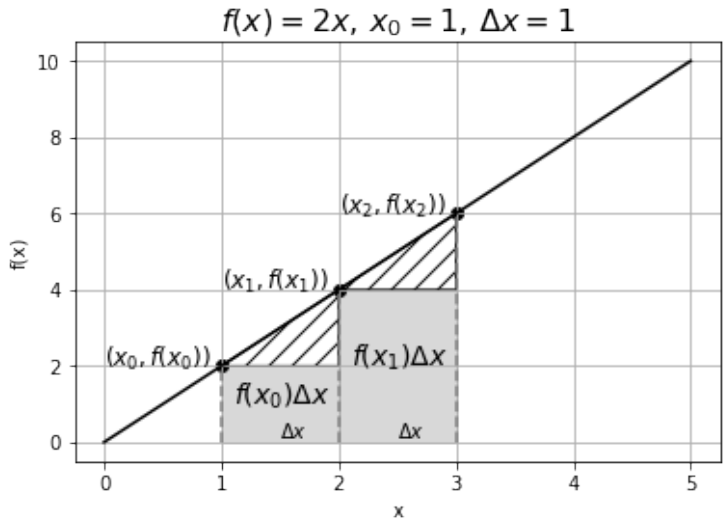


Figure B.6: The function $f(x) = 2x$ and illustration of the terms $f(x_0)\Delta x$ and $f(x_1)\Delta x$ as the area between the curve $f(x)$ and the x axis when $\Delta x \to 0$.

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:

$$\int_{x_0}^{x_N} 2xdx = F(x_N) - F(x_0) = x_N^2 - x_0^2$$

we can immediately identify the anti-derivative and the indefinite integral:

$$\begin{aligned}F(x) &= x^2 + C \\ \int 2xdx &= x^2 + C\end{aligned}$$

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.

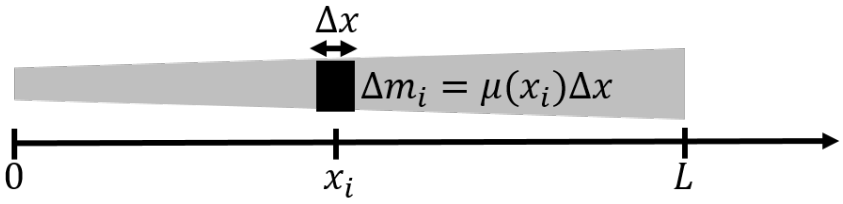


Figure B.7: A rod with a varying linear density. To calculate the mass of the rod, we consider a small mass element Δm_i of length Δx at position x_i. The total mass of the rod is found by summing the mass of the small mass elements.

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 equal 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:

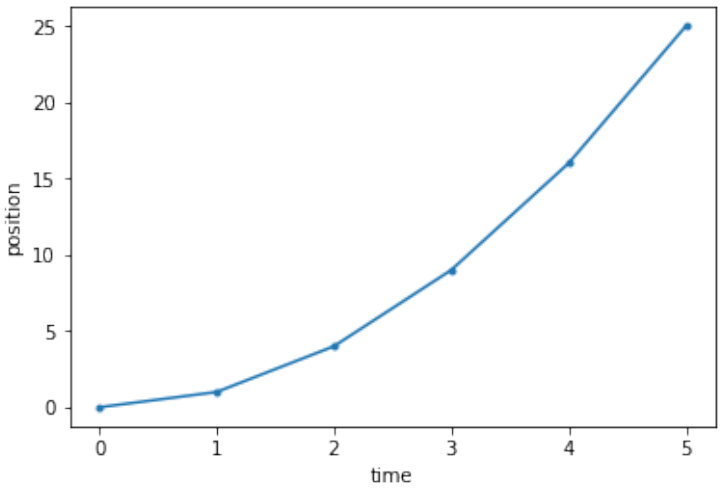


Figure D.1: Using two arrays and plotting them.

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:

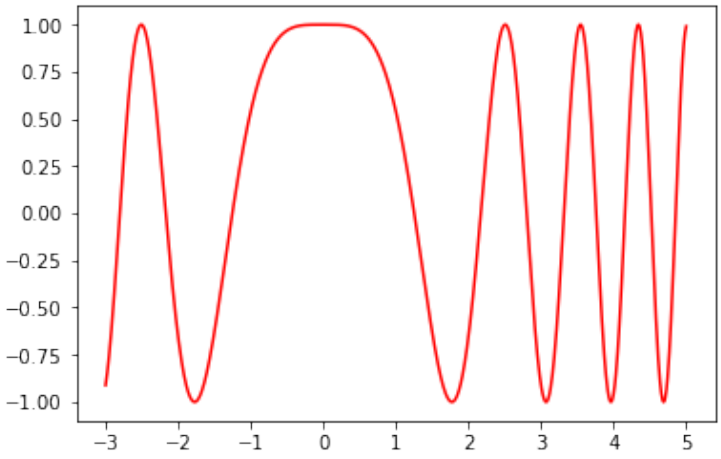


Figure D.2: Plotting a function using arrays.

D.4 The QExpy python package for experimental physics

QExpy is a Python module that was developed with students from 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 2, 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 2-7, 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 2.3.1), 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 2-6.

D.4.2 Plotting experimental data with uncertainties

In Chapter 2 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----------
```

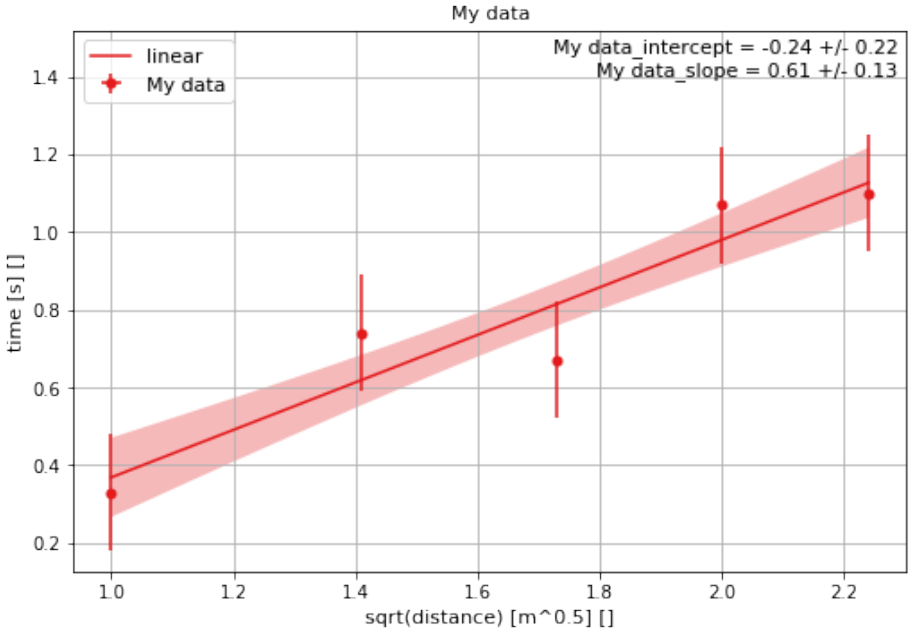


Figure D.3: QExpy plot of t versus $\sqrt{x}$ and line of best fit.

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?