



CONTENTS

1	Introduction to the Kontinua Sequence	21
1.1	Matter and Energy Introduction	22
1.2	Chemical Reaction	24
1.3	Mass and Acceleration	26
1.4	Mass and Gravity	27
1.5	Mass and Weight	28
2	Atomic and Molecular Mass	31
2.1	Molar Mass	35
2.2	Heavy atoms aren't stable	35
3	Work and Energy	37
3.1	Forms of Energy	38
3.1.1	Heat	39
3.1.2	Electricity	39
3.1.3	Chemical Energy	39
3.1.4	Kinetic Energy	40
3.1.5	Gravitational Potential Energy	40
3.2	Conservation of Energy	41
3.3	Efficiency	41
4	Units and Conversions	43
4.1	Conversion Factors	45
4.2	Conversion Factors and Ratios	46
4.3	When Conversion Factors Don't Work	47

5 Simple Machines	49
5.1 Levers	50
5.2 Ramps	51
5.3 Gears	53
5.4 Hydraulics	54
6 Buoyancy	57
6.1 The Mechanism of Buoyancy: Pressure	59
6.2 The Mechanism of Buoyancy: Density	60
7 Heat	63
7.1 Specific Heat Capacity	63
7.2 Getting to Equilibrium	65
7.3 Specific Heat Capacity Details	66
8 Cognitive Biases 1	67
8.1 Fundamental Attribution Error	68
8.2 Self-Serving Bias	68
8.3 In-group favoritism	69
8.4 The Bandwagon Effect and Groupthink	69
8.5 The Curse of Knowledge	70
8.6 False Consensus	71
8.7 The Spotlight Effect	71
8.8 The Dunning-Kruger Effect	72
8.9 Confirmation Bias	73
8.10 Survivorship bias	75
9 Friction	77
9.1 Static vs Kinetic Friction Coefficients	79
9.2 Skidding and Anti-Lock Braking Systems	80
10 The Greek Alphabet	83
11 Basic Statistics	85
11.1 Mean	86
11.2 Variance	87
11.3 Median	88
11.4 Histograms	89
11.5 Root-Mean-Squared	91
12 Basic Statistics in Spreadsheets	93

12.1 Your First Spreadsheet	93
12.2 Formatting	95
12.3 Comma-Separated Values	96
12.4 Statistics in Spreadsheets	97
12.5 Histogram	98
13 Introduction to Electricity	101
13.1 Units	103
13.2 Circuit Diagrams	104
13.3 Ohm's Law	105
13.4 Power and Watts	105
13.5 Another great use of RMS	106
13.6 Electricity Dangers	106
14 DC Circuit Analysis	109
14.1 Resistors in Series	110
14.2 Resistors in Parallel	112
15 Charge	115
15.1 Lightning	116
15.2 But...	117
16 Fertilizer	119
16.1 The Nitrogen Cycle	120
16.2 The Haber-Bosch Process	121
16.3 Other nutrients	121
17 Concrete	123
17.1 Steel reinforced concrete	124
17.2 Recycling concrete	124
18 Metals	125
18.1 Steel	125
18.2 What metal for what task?	126
19 Angles	129
20 Introduction to Triangles	133
20.1 Equilateral and Isosceles Triangles	133
20.2 Interior Angles of a Triangle	134

21 Pythagorean Theorem	139
21.1 Distance between Points	141
21.2 Distance in 3 Dimensions	142
22 Congruence	143
22.1 Triangle Congruency	144
23 Parallel and Perpendicular	149
24 Circles	151
24.1 Tangents	154
25 Functions and Their Graphs	157
25.1 Graphs of Functions	158
25.2 Can this be expressed as a function?	159
25.3 Inverses	160
25.4 Graphing Calculators	162
26 Volumes of Common Solids	165
26.1 Cylinders	166
26.2 Volume, Area, and Height	168
27 Conic Sections	175
27.1 Definitions	175
27.1.1 Circle	175
27.1.2 Ellipse	176
27.1.3 Hyperbola	176
27.1.4 Parabola	176
28 Falling Bodies	179
28.1 Calculating the Velocity	181
28.2 Calculating Position	182
28.3 Quadratic functions	184
28.4 Simulating a falling body in Python	184
28.4.1 Graphs and Lists	186
29 Solving Quadratics	191
29.1 The Traditional Quadratic Formula	194
30 Complex Numbers	195
30.1 Definition	195

30.2 Why Are Complex Numbers Necessary?	195
30.2.1 Roots of Negative Numbers	196
30.2.2 Polynomial Equations	196
30.2.3 Physics and Engineering	196
30.3 Adding Complex Numbers	196
30.4 Multiplying Complex Numbers	196
31 Vectors	199
31.1 Adding Vectors	200
31.2 Multiplying a vector with a scalar	202
31.3 Vector Subtraction	204
31.4 Magnitude of a Vector	204
31.5 Vectors in Python	205
31.5.1 Formatting Floats	206
32 Momentum	209
33 The Dot Product	213
33.1 Properties of the dot product	214
33.2 Cosines and dot products	215
33.3 Dot products in Python	216
33.4 Work and Power	216
34 Boats	219
34.1 Basic Terminology	219
34.2 Why Boats Float Upright	220
34.2.1 Center of Buoyancy	220
34.2.2 Center of Mass	221
34.3 Center of Lateral Resistance	221
34.4 Steering with a Rudder	221
34.5 Boat Length and Resistance	222
35 Sailboats	223
35.1 Magnitude of the Wind Force	223
35.2 Direction and Location of the Wind Force	224
35.3 Beam Reach	225
35.4 Apparent Wind	226
35.5 Close Reach	226
35.6 Shaping the Sail	226
35.7 Tacking into the Wind	227

35.8 Heeling	227
36 Introduction to Spreadsheets	229
36.1 Solving It Symbolically	230
36.2 Solving It Numerically (with a spreadsheet)	231
36.3 Graphing	233
36.4 Other Things You Should Know About Spreadsheets	234
36.5 Challenge: Make a spreadsheet	235
37 Compound Interest	237
37.1 An example with annual interest payments	237
37.2 Exponential Growth	238
37.3 Sensitivity to interest rate	239
38 Introduction to Data Visualization	241
38.1 Common Types of Data Visualizations	241
38.1.1 Bar Chart	242
38.1.2 Line Graph	243
38.1.3 Pie Chart	244
38.1.4 Scatter Plot	245
38.2 Make Bar Graph	246
39 Atmospheric Pressure	249
39.1 Altitude and Atmospheric Pressure	251
39.2 How a Drinking Straw Works	252
39.2.1 The Longest Usable Straw	253
39.2.2 Millimeters Mercury	255
39.3 How Siphon Works	256
39.4 How a Toilet Works	258
40 Exponents	261
40.1 Identities for Exponents	262
41 Exponential Decay	265
41.1 Radioactive Decay	266
41.2 Model Exponential Decay	268
42 Logarithms	269
42.1 Logarithms in Python	270
42.2 Logarithm Identities	270
42.3 Changing Bases	271

42.4 Natural Logarithm	271
42.5 Logarithms in Spreadsheets	272
43 Trigometric Functions	273
43.1 Graphs of sine and cosine	275
43.2 Plot cosine in Python	275
43.3 Derivatives of sine and cos	276
43.4 A weight on a spring	277
43.5 Integral of sine and cosine	280
44 Transforming Functions	281
44.1 Translation up and down	282
44.2 Translation left and right	283
44.3 Scaling up and down in the y direction	283
44.4 Scaling up and down in the x direction	284
44.5 Order is important!	285
45 Sound	289
45.1 Pitch and frequency	290
45.2 Chords and harmonics	292
45.3 Making waves in Python	293
45.3.1 Making a sound file	295
46 Alternating Current	297
46.1 Power of AC	298
46.2 Power Line Losses	299
46.3 Transformers	300
46.4 Phase and 3-phase power	301
47 Drag	303
47.1 Wind resistance	304
47.2 Initial velocity and acceleration due to gravity	304
47.3 Simulating artillery in Python	305
47.4 Terminal velocity	307
48 Vector-valued Functions	309
48.1 Finding the velocity vector	310
48.2 Finding the acceleration vector	311
49 Circular Motion	313
49.1 Velocity	315

49.2 Acceleration	316
49.3 Centripetal force	317
50 Orbits	319
50.1 Astronauts are <i>not</i> weightless	321
50.2 Geosynchronous Orbits	322
51 Simulation with Vectors	325
51.1 Force, Acceleration, Velocity, and Position	325
51.2 Simulations and Step Size	326
51.3 Make a Text-based Simulation	327
51.4 Graph the Paths of the Moons	329
51.5 Conservation of Momentum	333
51.6 Animation	335
51.7 Challenge: The Three-Body Problem	339
52 Longitude and Latitude	341
52.1 Nautical Mile	344
52.2 Haversine Formula	344
53 Tides and Eclipses	347
53.1 Leap Years	348
53.2 Phases of the Moon	348
53.3 Eclipses	351
53.4 The Far Side of the Moon	352
53.5 Tides	352
53.5.1 Computing the Forces	353
53.5.2 Solar Tidal Forces	358
54 Electromagnetic Waves	359
54.1 The greenhouse effect	360
55 How Cameras Work	365
55.1 The Light That Shines On the Cow	366
55.2 Light Hits the Cow	368
55.3 Pinhole camera	369
55.4 Lenses	370
55.5 Sensors	372
56 How Eyes Work	373
56.1 Eye problems	374

56.1.1	Glaucoma	374
56.1.2	Cataracts	375
56.1.3	Nearsightedness, farsightedness, and astigmatism	375
56.2	Seeing colors	376
56.3	Pigments	378
57	Reflection	379
57.1	Reflection	380
57.2	Curved Mirrors	383
57.2.1	The Reflective Properties of Circles and Spheres	383
57.2.2	Ellipses and Ellipsoids	384
57.2.3	Elliptical Orbits	388
57.2.4	Ellipsoids	388
57.2.5	Parabolas and Parabolic Reflectors	390
58	Refraction	395
59	Lenses	397
59.1	Focal Length	398
59.2	Refractive Index	398
60	Images in Python	401
60.1	Adding color	402
60.2	Using an existing image	405
61	Introduction to Polynomials	407
62	Python Lists	411
62.1	Evaluating Polynomials in Python	412
62.2	Walking the list backwards	413
62.3	Plot the polynomial	415
63	Adding and Subtracting Polynomials	419
63.1	Subtraction	420
63.2	Adding Polynomials in Python	421
63.3	Scalar multiplication of polynomials	423
64	Multiplying Polynomials	425
64.1	Multiplying a monomial and a polynomial	426
64.2	Multiplying polynomials	427

65	Multiplying Polynomials in Python	431
65.1	Something surprising about lists	434
66	Differentiating Polynomials	435
67	Python Classes	439
67.1	Making a Polynomial class	440
68	Common Polynomial Products	445
68.1	Difference of squares	445
68.2	Powers of binomials	448
69	Factoring Polynomials	451
69.1	How to factor polynomials	452
70	Practice with Polynomials	455
71	Graphing Polynomials	457
71.1	Leading term in graphing	459
72	Interpolating with Polynomials	461
72.1	Interpolating polynomials in python	464
73	Limits	467
74	Rational Functions	475
75	Differentiation	483
75.1	Differentiability	485
75.2	Using the definition of derivative	486
76	Derivatives	487
76.1	Definition	487
76.2	Applications in Physics	488
76.2.1	Velocity and Acceleration	488
76.2.2	Force and Momentum	488
77	Rules for Finding Derivatives	489
77.1	Constant Rule	489
77.2	Power Rule	489
77.3	Product Rule	490
77.4	Quotient Rule	490

77.5	Chain Rule	490
77.6	Conclusion	490
78	Optimization	491
78.1	Optimization Problems	491
78.2	Types of Optimization Problems	492
78.3	Applications	492
79	Implicit Differentiation	493
79.1	Implicit Differentiation Procedure	493
79.2	Example	494
80	Related Rates	495
80.1	Steps to solve related rates problems	495
80.1.1	Step 1: Understand the problem	495
80.1.2	Step 2: Draw a diagram	495
80.1.3	Step 3: Write down what you know	496
80.1.4	Step 4: Write an equation	496
80.1.5	Step 5: Differentiate both sides of the equation	496
80.1.6	Step 6: Substitute the known rates and solve for the unknown	496
80.2	Example	496
81	Multivariate Functions	499
82	Partial Derivatives and Gradients	501
83	Vectors and Matrices	503
83.1	Applications of Matrix-Vector Multiplication	504
83.1.1	Computer Graphics	504
83.1.2	Data Analysis	504
83.1.3	Economics	504
83.1.4	Engineering	504
83.1.5	Image Processing	505
83.2	Vector-Matrix Multiplication	505
83.2.1	Vector-Matrix Multiplication in Python	507
83.3	Where to Learn More	507
84	Linear Combinations	509
84.1	Weighted Averages of Vectors	511
84.2	Weighted Averages of Vectors in Python	512

85 Vector Spans and Independence	513
85.1 Vector Independence	514
85.1.1 Dependent Vectors	514
85.1.2 Independent Vectors	515
85.2 Checking for Linear Independence Using Python	516
85.3 Determinants	518
85.4 Where to Learn More	519
86 Matrices	521
86.1 Types of Matrices	522
86.1.1 Symmetric Matrices	522
86.1.2 Creating Matrices in Python	523
86.1.3 Creating Special Matrices in Python	525
87 Projections	527
87.1 Projections in Python	530
87.2 Where to Learn More	530
88 The Gram-Schmidt Process	531
88.1 The Process	532
88.2 Example Calculation	532
88.3 The Gram-Schmidt Process in Python	534
88.4 Where to Learn More	535
89 Eigenvectors and Eigenvalues	537
89.1 Definition	538
89.2 Finding Eigenvalues and Eigenvectors	539
89.3 Example	539
89.4 Eigenvalues and Eigenvectors in Python	541
89.5 Where to Learn More	541
90 Singular Value Decomposition	543
90.1 Definition	543
90.2 Applications of SVD	544
90.3 Calculating SVD Manually	544
90.4 Singular Value Decomposition with Python	547
90.5 Sign Ambiguity	548
90.6 SVD Applied to Image Compression	549
90.7 Where to Learn More	550

91 Data Tables and pandas	551
91.1 Data types	552
91.2 pandas	552
91.3 Reading a CSV with pandas	553
91.4 Looking at a Series	554
91.5 Rows and the index	555
91.6 Changing data	556
91.7 Derived columns	557
92 Data tables in SQL	559
92.1 Using SQL from Python	562
93 Representing Natural Numbers	565
94 Making Web Requests with HTTP	567
94.1 HTTP Requests	567
94.2 Using HTTP with Web-Based APIs	568
95 Using and Creating APIs	569
96 Data Compression and Decompression	571
96.1 Data Compression and Decompression	571
96.2 Entropy	572
96.3 Entropy and Compression	572
97 Dealing with JSON and XML	573
98 HTML	575
98.1 HTML Elements	575
98.2 HTML Document Structure	575
99 Introduction to Text	577
99.1 Newlines and Carriage Returns	577
99.2 ASCII	578
99.3 UTF-8	578
100 Stop Words	579
101 Stemming and Lemmatization	581
101.0.1 Stemming	581
101.0.2 Lemmatization	582

102 Alphabets and Accents	583
103 Making Plots with matplotlib	585
104 Geographical Data	587
105 Geocoding and Reverse Geocoding	589
105.1 Geocoding	589
105.2 Reverse Geocoding	590
106 Making a Map	593
107 Introduction to Discrete Probability	595
107.1 The Probability of All Possibilities is 1.0	596
107.2 Independence	596
107.3 Why 7 is the most likely sum of two dice	597
107.4 Random Numbers and Python	598
107.4.1 Making a bar graph	602
108 Beginning Combinatorics	605
108.0.1 Choose	607
109 Permutations and Sorting	609
109.1 Notation	610
109.1.1 Challenge	611
109.2 Sorting in Python	611
109.3 Inverses	611
109.4 Cycles	612
110 Conditional Probability	615
110.1 Marginalization	616
110.2 Conditional Probability	617
110.3 Chain Rule for Probability	618
111 Bayes' Theorem	619
111.1 Bayes Theorem	620
111.2 Using Bayes' Theorem	621
111.3 Confidence	622
112 Definite Integrals	623
112.1 Definition	623

113 Antiderivatives	625
114 The Fundamental Theorem of Calculus	627
114.1 First Part	627
114.2 Second Part	628
115 Continuous Probability Distributions	629
115.1 Cumulative Distribution Function	630
115.2 Probability Density Function	632
115.3 The Continuous Uniform Distribution	633
115.4 Continuous Distributions In Python	635
116 The Physics of Gases	641
116.0.1 A Statistical Look At Temperature	641
116.0.2 Absolute Zero and Degrees Kelvin	643
116.1 Temperature and Volume	643
116.2 Pressure and Volume	646
116.3 The Ideal Gas Law	647
116.4 Molecules Like To Stay Close to Each Other	648
117 Kinetic Energy and Temperature of a Gas	651
117.1 Molecule Shape and Molar Heat Capacity	652
117.2 Kinetic Energy and Temperature	653
117.3 Why is $C_{V,m}$ different from $C_{P,m}$?	654
117.4 Work of Creating Volume Against Constant Pressure	654
117.5 Why does a gas get hotter when you compress it?	655
117.6 How much hotter?	656
117.7 How an Air Conditioner Works	661
118 Phases of Matter	663
118.1 Thinking Microscopically About Phase	664
118.2 Phase Changes and Energy	665
118.3 How a Rice Cooker Works	667
118.4 Thinking Statistically About Phase Change	669
118.4.1 Evaporative Cooling Systems	670
118.4.2 Humidity and Condensation	670
119 The Piston Engine	673
119.1 Parts of the Engine	673
119.2 The Four-Stroke Process	674

119.3 Dealing with Heat	676
119.4 Dealing with Friction	677
119.5 Challenges	677
119.6 How We Measure Engines	677
119.7 The Ford Model T and Ethanol	678
119.8 Compression Ratio	680
119.9 The Choke and Direct Fuel Injection	680
120 u-Substitution	681
121 Differential Equations	683
121.1 Ordinary Differential Equations	683
121.2 Partial Differential Equations	684
122 Population Proportion Statistics	685
122.1 Sample Probabilities from Population Proportion	685
122.2 Population Proportion from Sample	687
122.3 From Likelihood to Probability Density Function	688
122.4 Beta Distribution	690
123 The Normal Distribution	693
123.1 Defining the Normal Distribution	693
123.2 Importance of the Normal Distribution	694
124 Change of Variables	695
124.1 Making a Probability Density Function	697
124.2 Decreasing Conversions	701
125 Poisson and Exponential Probability Distributions	703
126 Multiple Integrals	705
127 Multivariate Distributions	707
128 The Multivariate Normal Distribution	709
128.1 Multivariate Normal Distribution	709
128.2 Covariance Matrix	710
129 Sets and Logic	711
129.1 Sets	712
129.1.1 And and Or	713
129.1.2 How simple are sets?	714

129.1.3 Subsets	714
129.1.4 Union and Intersection of Sets	715
129.1.5 Venn Diagrams	715
129.2 Logic	717
129.3 Implies	717
129.4 If and Only If	717
129.5 Not	718
129.6 Cardinality	719
129.7 Complement of a Set	719
129.8 Subtracting Sets	720
129.9 Power Sets	721
129.10 Booleans in Python	721
129.11 The Contrapositive	722
129.12 The Distributive Property of Logic	722
129.13 Exclusive Or	723
130 Linked Lists	725
131 Trees	727
132 Searching Trees	729
133 Hash Tables	731
133.1 Structure of a Hash Table	731
133.2 Inserting and Retrieving Data	731
133.3 Handling Collisions	732
133.4 Time Complexity	732
134 Sorting Algorithms	733
135 Introduction to Graphs	735
135.1 Finding Good Paths	737
135.2 Graphs in Python	738
136 Dijkstra's Algorithm	743
136.1 Algorithm Description	743
136.2 Implementation	745
136.3 Making it faster	748
137 Binary Search	751
137.1 A Naive Implementation of the Priority Queue	752

137.2 Using the Priority Queue	752
137.3 Binary Search	755
137.4 Algorithm	755
138 Other Graph Algorithms	757
138.1 Depth-First Search	757
138.2 Bellman-Ford Algorithm	758
139 Bayesian Networks	759
139.1 Components	759
139.2 Inferences	760
139.3 Learning	760
140 Introduction to Classification and Regression	761
140.1 Classification Systems	761
140.2 Regression Systems	762
140.3 Algorithms	762
140.4 Performance Metrics	762
141 Simple Linear Regression	763
141.0.1 The model behind simple linear regression	764
142 Simple Logistic Regression	765
143 Standardizing Data	767
143.1 Why Do We Standardize Data?	768
143.1.1 Homogeneity of Variances	768
143.1.2 Interpreting Coefficients	768
143.1.3 Algorithm Convergence	768
143.1.4 Comparing Variables	768
143.1.5 Preventing Numerical Instabilities	768
144 One-Hot Encoding	769
144.1 Why One-Hot Encoding?	769
144.2 How does One-Hot Encoding Work?	770
145 Multiple Logistic Regression	771
145.1 Multiple Logistic Regression	771
145.2 Divide by 4 Rule	772
146 The Training/Validation/Testing Process	773

146.0.1 Training Set	773
146.0.2 Validation Set	774
146.0.3 Testing Set	774
147 Evaluating Classification Systems	775
147.1 Definition of a Confusion Matrix	775
147.2 Performance Metrics	776
148 Evaluation Binary Classifiers	777
148.0.1 Binary Classification	777
148.0.2 Accuracy	778
148.0.3 Precision and Recall	778
148.0.4 F1 Score	779
149 The k-Nearest Neighbor Classifier	781
149.1 The k-NN Algorithm	781
149.2 Choosing the Right 'k'	782
149.3 Considerations	782
150 Naive Bayes Classifier	783
150.1 Bayes' Theorem	783
150.2 The Naivety of Naive Bayes	784
150.3 Working of Naive Bayes Classifier	784
151 Evaluating the Fit of a Linear Regression Model	785
151.1 Residuals	785
151.2 R-Squared (R^2)	786
151.3 Root Mean Squared Error (RMSE)	786
152 Linear Regression and Gradient Descent	787
152.1 Standardizing Inputs	788
153 Generalized Linear Models	789
153.1 Components of a Generalized Linear Model	789
153.2 Formulation of a Generalized Linear Model	790
153.3 Fitting a Generalized Linear Model	790
153.4 Examples of Generalized Linear Models	790
154 Link Functions	791
155 Decision Trees for Classification	793

155.1 Decision Trees for Classification	793
155.2 Gini Impurity	793
155.3 How Gini Impurity is Used	794
156 Bagging and Random Forests	795
156.1 Bagging	795
156.2 Random Forests	796
157 Boosting	797
157.1 AdaBoost	797
157.2 Gradient Boosted Trees	798
158 Clustering using k-Means	799
158.1 The K-Means Algorithm	799
158.2 Choosing K	800
159 Neural Nets for Regression	801
159.1 Neural Networks	801
159.2 Neural Networks for Regression	802
160 Neural Networks for Classification	803
160.1 Neural Networks for Classification	803
161 Deep Learning	805
161.1 Deep Learning	805
161.2 Chain Rule	806
161.3 Backpropagation	806
A Answers to Exercises	807
Index	849



CHAPTER 1

Introduction to the Kontinua Sequence

This book will start you on the long and difficult trek to becoming a modern problem solver. Along the path, you will learn how to use the tools of math, computers, and science.

Why should you bother? There are big problems in this world that will require expert problem solvers. Those people will make the world a better place while enjoying interesting and lucrative careers. We are talking about engineers, scientists, doctors, computer programmers, architects, actuaries, and mathematicians. Right now, those occupations represent about 6% of all the jobs in the United States. Soon, that number is expected to rise above 10%. On average, people in that 10% of the population are expected to have salaries twice that of their non-technical counterparts.

Solving problems is difficult. At some point on this journey, you will see people who are better at solving problems than you are. You, like every other person who has gone on this journey, will think “I have worked so hard on this, but that person is better at it than I am. I should quit.” Don’t.

First, solving problems is like a muscle. The more you do, the better you get at it. It is OK to say “I am not good at this yet.” That just means you need more practice.

Second, you don’t need to be the best in the world. 10 million people your age can be better at solving problems than you, *and you can still be in the top 10% of the world*. If you complete this journey, there will be problems for you to solve and a job where your problem-solving skills will be appreciated.

So where do we start?

1.1 Matter and Energy Introduction

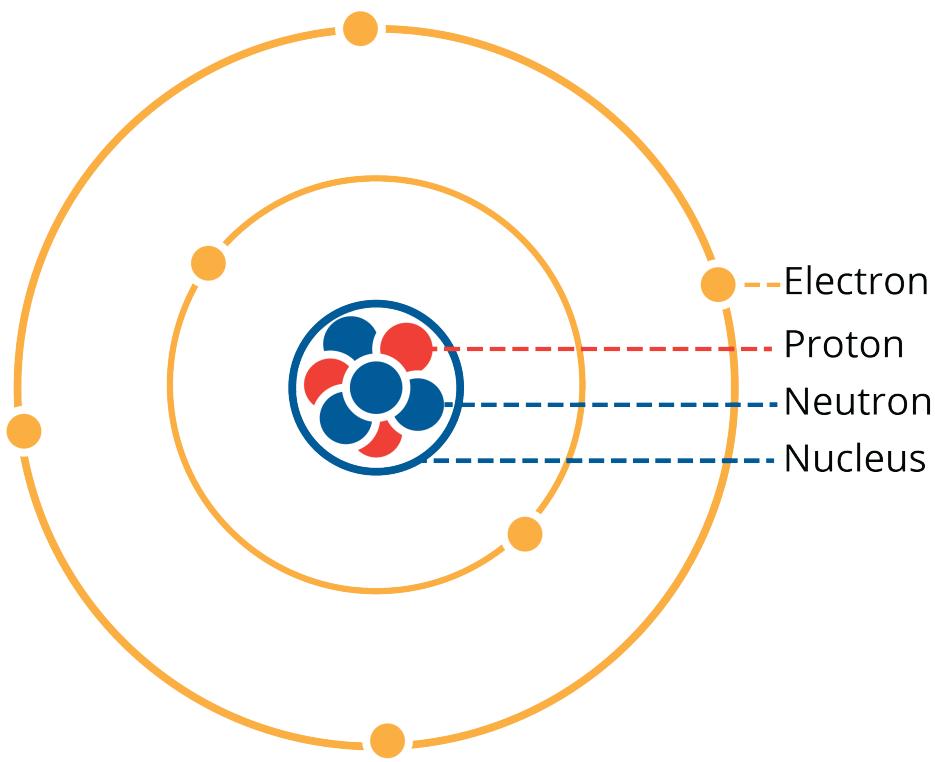
The famous physicist Richard Feynman once asked this question: “If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence was passed on to the next generation of creatures, what statement would contain the most information in the fewest words?”

His answer was “All things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.”

That seems like a good place to start.

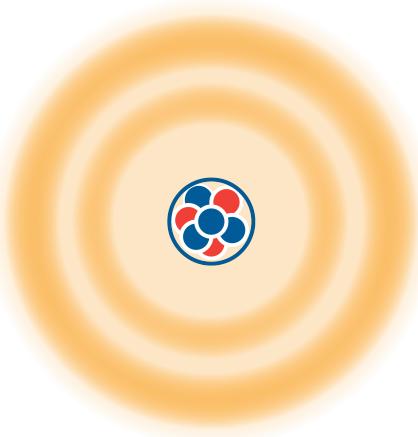
All things (including the air around you) are made of atoms. Atoms are very tiny – there are more atoms in a drop of water than there are drops of water in all the oceans.

Every atom has a nucleus that contains protons and neutrons. There is also a cloud of electrons flying around the nucleus. However, the mass of the atom comes mainly from the protons and neutrons, which are exponentially heavier than electrons.



Watch **Elements and atoms** from Khan Academy at https://youtu.be/IFKnq9QM6_A.

The previous graphic is slightly untrue. While it is a convenient model for thinking about atoms, in reality electrons don't neatly orbit the nucleus. Scientists don't know exactly where an electron will be in relation to the nucleus, but they do know where it's most likely to be. They use a cloud that is thicker in the center but fades out at the edges to represent an electron's position.



We classify atoms by the numbers of protons they have. An atom with one proton is a hydrogen atom, an atom with two protons is a helium atom, and so forth (refer to periodic table on pg.). We say that hydrogen and helium are *elements* because the classification of elements is based on proton number. And we give each element an atomic symbol. Hydrogen gets H. Helium gets He Oxygen gets O. Carbon gets C, etc.

Often two hydrogen atoms will attach to an oxygen atom. The result is a water molecule. Why do they cluster together? because they share electrons in their clouds.

A molecule is described by the elements it contains. Water is H_2O because it has two hydrogen atoms and one oxygen atom.

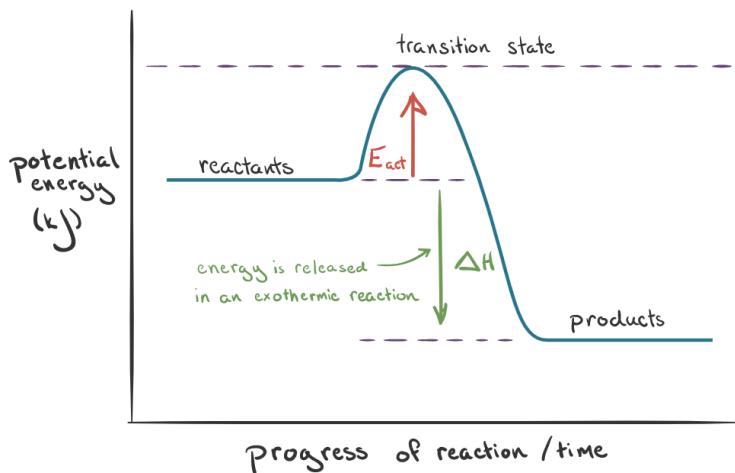
There are many kinds of molecules. You know a few:

- Table salt is crystals made of NaCl molecules: a sodium atom attached to a chlorine atom.
- Baking soda, or sodium bicarbonate, is NaHCO_3 .
- Vinegar is a solution including acetic acid (CH_3COOH).
- O_2 is the oxygen molecules that you breathe out of the air (Air, a blend of gases, is mostly N_2).

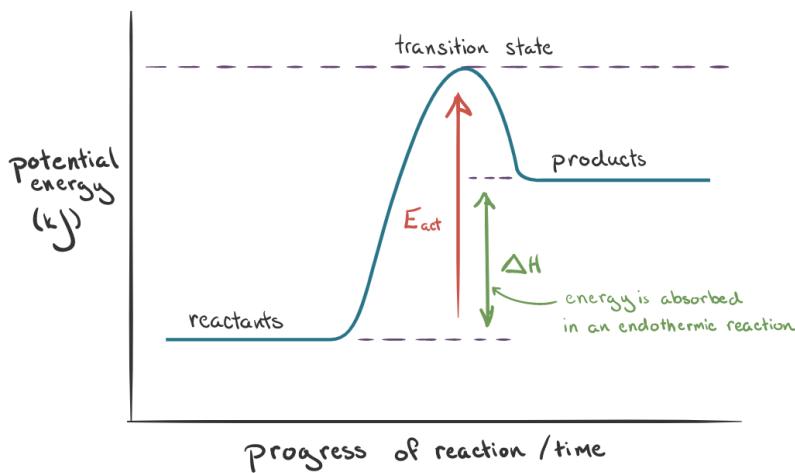
1.2 Chemical Reaction

Sometimes two hydrogen atoms form a molecule (H_2). Sometimes two oxygen atoms form a molecule (O_2). If you mix these together and light a match, they will rearrange themselves into water molecules. This is called a *chemical reaction*. In any chemical reaction, the atoms are rearranged into new molecules.

Some chemical reactions (like the burning of hydrogen gas described above) are *exothermic* – that is, they give off energy. Burning hydrogen gas happens quickly and gives off a lot of energy. If you have enough, it will make quite an explosion.



Other chemical reactions are *endothermic* – that is they consume energy. Photosynthesis, the process by which plants consume energy from the sun to make sugar from CO_2 and H_2O requires an endothermic chemical reaction.



In a chemical reaction, the transition state is the point where there is a maximum value of energy. This energy is called the activation energy.

Here's an overview of chemical reactions: https://simple.wikipedia.org/wiki/Chemical_reaction

1.3 Mass and Acceleration

Each atom has a mass, so everything that is made up of atoms has a mass, which is pretty much everything. We measure mass in grams. A paper clip is about 1 gram of steel. An adult human can weigh 70,000 grams, so for larger things we often talk about kilograms. A kilogram is 1000 grams.

The first interesting thing about mass is that objects with more mass require more force to accelerate. For example, pushing a bicycle so that it accelerates from a standstill to jogging speed in 2 seconds requires a lot less force than pushing a train so that it accelerates at the same rate.

You will probably find it useful to watch Khan Academy's summary of Newton's second law of motion: <https://youtu.be/ou9YMWlJgkE>

Newton's Second Law of Motion

The force necessary to accelerate an object of mass m is given by:

$$F = ma$$

That is the force is equal to the mass times the acceleration.

What are the units here? We already know that mass is measured in kilograms. We can measure velocity in meters per second, but that is different from acceleration. Acceleration is the rate of change in velocity. So if we want to go from 0 to 5 meters per second (that's jogging speed) in two seconds. That is a change in velocity of 2.5 meters per second every second. We would say this acceleration is 2.5 m/s^2 .

What about measuring force? Newton decided to name the unit after himself: The force necessary to accelerate one kilogram at 1 m/s^2 is known as *a newton*.

Exercise 1 Acceleration**Working Space**

While driving a bulldozer, you come across a train car (with no brakes and no locomotive) on a track in the middle of a city. The train car has a label telling you that it weighs 2,400 kg. There is a bomb welded to the interior of the train car, and the timer tells you that you can safely push the train car for 120 seconds. To get the train car to where it can explode safely, you need to accelerate it to 20 meters per second. Fortunately, the track is level and the train car's wheels have almost no rolling resistance.

With what force, in newtons, do you need to push the train for those 120 seconds?

Answer on Page 807

1.4 Mass and Gravity

The second interesting thing about mass is that masses are attracted to each other by the force we call *gravity*. The force of attraction between two objects is proportional to the product of their masses. As objects get farther away, the force decreases. That is why you are more attracted to the earth than you are to distant stars, which have much more mass than the earth.

Newton's Law of Universal Gravitation

Two masses (m_1 and m_2) that are a distance of r from each other, are attracted toward each other with a force of magnitude:

$$F = G \frac{m_1 m_2}{r^2}$$

where G is the universal gravitational constant. If you measure the mass in kilograms and the distance in meters, G is about 6.674×10^{-11} . That will get you the force of the attraction in newtons.

Exercise 2 **Gravity****Working Space**

The earth's mass is about 6×10^{24} kilograms.

Your spacecraft's mass is 6,800 kilograms.

Your spacecraft is also about 100,000 km from the center of the earth. (For reference, the moon is about 400,000 km from the center of the earth.)

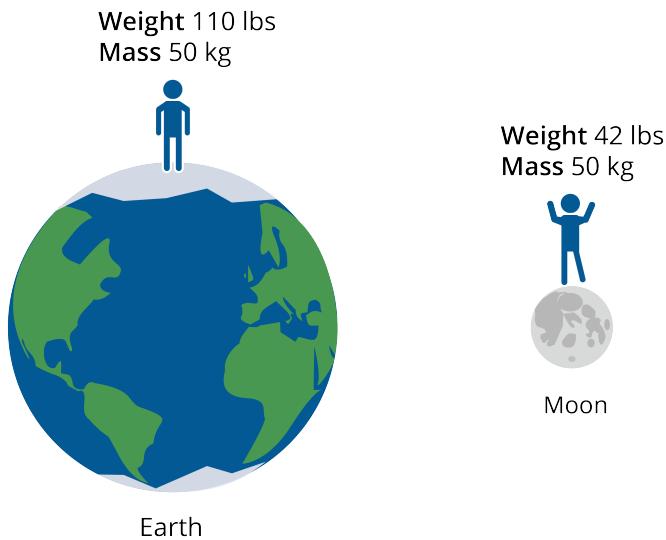
What is the force of gravity that is pulling your spacecraft and the earth toward each other?

Answer on Page 807

1.5 Mass and Weight

Gravity pulls on things proportional to their mass, so we often ignore the difference between mass and weight.

The weight of an object is the force due to the object's mass and gravity. When we say, "This potato weighs 1 pound," we actually mean "This potato weighs 1 pound on earth." That same potato would weigh about one-fifth of a pound on the moon.



But that potato has a mass of 0.45 kg anywhere in the universe.

FIXME Global layout note: Let's discuss adding Title's and Captions to all graphics.

For example:

TITLE: Mass versus Weight

CAPTION: Human Earth weight: 150lbs / Moon weight: ??lbs

Potato Earth weight: .25lbs / Moon weight: ??lbs

FIXME: Allison thinks it would be funny if the person in the graphic were holding a potato and we also added the weight and mass of the potato to the caption. No worries if this type of edit isn't in the budget!

FIXME: What are your thoughts about using the metric system consistently – in which case we'll replace pounds here with kilos.



CHAPTER 2

Atomic and Molecular Mass

A proton and a neutron have about the same mass. An electron, on the other hand, has much less mass: One neutron weighs about the same amount as 2000 electrons. Thus, the mass of any object comes mostly from the protons and neutrons in the nucleus of its atoms.

We know how many protons an atom has by what element it is, but how do we know the number neutrons?

If you fill a balloon with helium, it will have two different kinds of helium atoms: Most of the helium atoms will have 2 neutrons, but a few will have only 1 neutron. We say that these are two different *isotopes* of helium. We call them helium-4 (or ${}^4\text{He}$) and helium-3 (or ${}^3\text{He}$). Isotopes are named for the sum of protons and neutrons the atom has: helium-3 has 2 protons and 1 neutron.

Watch Khan Academy's **Atomic mass, number, and isotopes** at <https://www.khanacademy.org/science/chemistry/atomic-structure-and-properties/introduction-to-the-atom/v/atomic-number-mass-number-and-isotopes>

A hydrogen atom nearly always has just 1 proton and no neutrons. A helium atom nearly

always has 2 protons and 2 neutrons. So, if you have a 100 hydrogen atoms and 100 helium atoms, the helium will have about 4 times more mass than the hydrogen. We say "Hydrogen is about 1 atomic mass unit(amu), and helium-4 is about 4 atomic mass units."

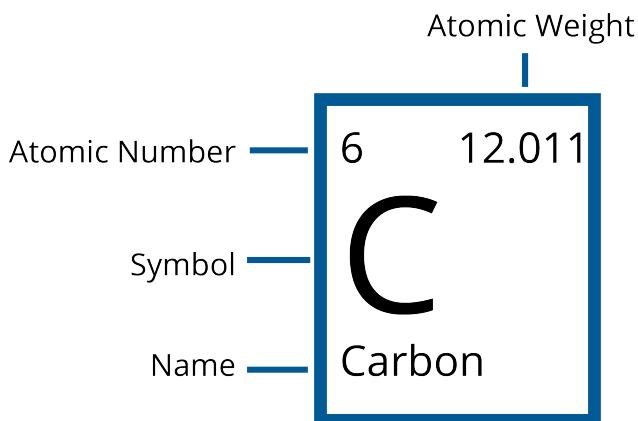
What, precisely, is an atomic mass unit? It is defined as 1/12 of the mass of a carbon-12 atom. Scientists have measured the mass of helium-4, and it is about 4.0026 atomic mass units. (By the way, an atomic mass unit is also called a *dalton*.)

Now you are ready to take a good look at the periodic table of elements. Here is the version from Wikipedia:

IA		VIIA																	
H Hydrogen 1.01		He Helium 4.00																	
Li Lithium 6.94		Ne Neon 20.18																	
Na Sodium 22.99		Ar Argon 39.95																	
K Potassium 39.10		Kr Krypton 83.80																	
Ca Calcium 40.08		Br Bromine 179.90																	
Sc Scandium 44.96		Kr Krypton 83.80																	
Ti Titanium 47.87		Br Bromine 179.90																	
V Vanadium 50.94		Br Bromine 179.90																	
Cr Chromium 52.00		Br Bromine 179.90																	
Mn Manganese 54.94		Br Bromine 179.90																	
Fe Iron 55.85		Br Bromine 179.90																	
Co Cobalt 58.93		Br Bromine 179.90																	
Ni Nickel 58.73		Br Bromine 179.90																	
Cu Copper 63.54		Br Bromine 179.90																	
Zn Zinc 65.41		Br Bromine 179.90																	
Ga Gallium 69.72		Br Bromine 179.90																	
Ge Germanium 72.03		Br Bromine 179.90																	
As Arsenic 74.92		Br Bromine 179.90																	
Se Selenium 78.97		Br Bromine 179.90																	
Br Bromine 179.90		Kr Krypton 83.80																	
Ta Tantalum 183.84		Br Bromine 179.90																	
W Tungsten 186.21		Br Bromine 179.90																	
Re Rhenium 190.23		Br Bromine 179.90																	
Os Osmium 192.22		Br Bromine 179.90																	
Ir Iridium 195.08		Br Bromine 179.90																	
Pt Platinum 196.97		Br Bromine 179.90																	
Au Gold 198.97		Br Bromine 179.90																	
Hg Mercury 200.59		Br Bromine 179.90																	
Tl Thallium 204.38		Br Bromine 179.90																	
Pb Lead 207.20		Br Bromine 179.90																	
Bi Bismuth 208.98		Br Bromine 179.90																	
Po Polonium (209)		Br Bromine 179.90																	
At Astatine (210)		Br Bromine 179.90																	
Rn Radon (222)		Br Bromine 179.90																	
Fr Francium (223)		Br Bromine 179.90																	
Ra Radium (226)		Br Bromine 179.90																	
Db Dubnium (260)		Br Bromine 179.90																	
Sg Seaborgium (270)		Br Bromine 179.90																	
Mt Mendelevium (276)		Br Bromine 179.90																	
Cn Copernicium (285)		Br Bromine 179.90																	
Nh Nihonium (284)		Br Bromine 179.90																	
Fl Flerovium (289)		Br Bromine 179.90																	
Lv Livermorium (293)		Br Bromine 179.90																	
Ts Tennessine (294)		Br Bromine 179.90																	
Og Oganesson (294)		Br Bromine 179.90																	
La	Ce	Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Yttrium	Thulium	Ytterbium	Lu	Lawrencium	
Lanthanum	Cerium	Praseodymium	Neodymium	Europium	Terbium	Dysprosium	Holmium	Erbium	Thulium	Ytterbium	Yttrium	Ytterbium	Lu	Yttrium	Thulium	Ytterbium	Lu	Lawrencium	
138.91	140.12	140.91	144.24	150.36	151.96	157.25	158.53	164.93	167.26	168.93	173.05	174.97	179.90	180.18	182.01	183.90	187.22	190.23	
Ac	Th	Pa	U	No	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lu	Yttrium	Thulium	Ytterbium	Lu	
Actinides	Thorium	Protactinium	Uranium	Neptunium	Plutonium	Americium	Curium	Berkelium	Californium	Einsteinium	Fermium	Mendelevium	Nobelium	Lawrencium	Yttrium	Thulium	Ytterbium	Lu	
(227)	(232)	(231)	(234)	(237)	(244)	(243)	(247)	(247)	(251)	(252)	(257)	(259)	(259)	(262)	(263)	(263)	(263)	(262)	

There is a square for each element. In the middle, you see the atomic symbol and the name of the element. In the upper right corner is the atomic number – the number of protons in the atom.

In the upper left corner is the atomic mass in atomic mass units.



Look at the atomic mass of boron. About 80% of all boron atoms have six neutrons. The other 20% have only 5 neutrons. So most boron atoms have a mass of about 11 atomic mass units, but some have a mass of about 10 atomic mass units. The atomic mass of boron is equivalent to the average mass of a boron atom: 10.811.

Exercise 3 Mass of a Water Molecule

Working Space

Using the periodic table, what is the average mass of one water molecule in atomic mass units?

Answer on Page 808

2.1 Molar Mass

An atomic mass unit is a very, very, very small unit; we would much rather work in grams. It turns out that $6.02214076 \times 10^{23}$ atoms equal 1 mole (a standard measure for chemistry). Scientists use this number so much that they gave it a name: *the Avogadro constant* or *Avogadro's number*.

Watch Khan Academy's discussion of the mole at <https://www.khanacademy.org/science/ap-chemistry-beta/x2eef969c74e0d802:atomic-structure-and-properties/x2eef969c74e0d802:moles-and-molar-mass/v/the-mole-and-avogadro-s-number>

If you have 12 doughnuts, that's a dozen doughnuts. If you have $6.02214076 \times 10^{23}$ doughnuts, you have a *mole* of doughnuts. (Note: it isn't practical to measure doughnuts this way: A mole of doughnuts would be about the size of the earth. We use moles for small things like molecules.)

Let's say you want to know how much a mole of NaCl weighs. From the periodic table, you see that Na has an atomic mass of 22.98976 atomic mass units. And Cl has 35.453 atomic mass units. One atom of NaCl has a mass of $22.98976 + 35.453 = 58.44276$ atomic mass units. Then a mole of NaCl has a mass of 58.44276 grams. Handy, right?

Exercise 4 Burning Methane

Working Space

Natural gas is mostly methane (CH_4). When one molecule of methane burns, two oxygen molecules (O_2) are consumed. One molecule of H_2O and one molecule of CO_2 are produced.

If I need 200 grams of water, how many grams of methane do I need to burn?
(This is how the hero in "The Martian" made water for his garden.)

Answer on Page 808

2.2 Heavy atoms aren't stable

When you look at the periodic table, there are a surprisingly large number of elements. You might be told to "Drink milk so that you can get the calcium you need." However, no

one has told you “You should eat kale so that you get enough copernicium in your diet.”

Copernicium, with 112 protons and 173 neutrons, has only been observed in a lab. It is highly radioactive and unstable(meaning it decays): a copernicium atom usually lives for less than a minute before decaying.

The largest stable element is lead, which has 82 protons and between 122 and 126 neutrons. Elements with lower atomic numbers than lead, have at least one stable isotope. Elements with higher atomic numbers than lead don’t.

Bismuth, with an atomic number of 83, is *almost* stable. In fact, most bismuth atoms will live for billions of years before decaying.

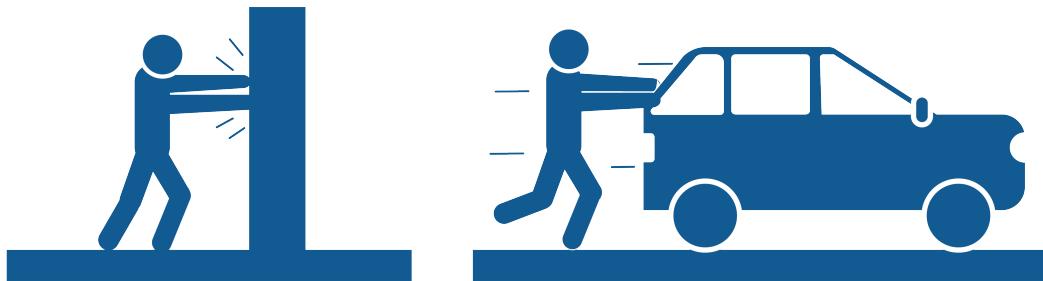


CHAPTER 3

Work and Energy

In this chapter, we are going to talk about how engineers define work and energy. It frequently takes force to get work done. Let's start with thinking about the relationship between force and energy. As we learned earlier, Force is measured in newtons, and one newton is equal to the force necessary to accelerate one kilogram at a rate of 1m/s^2 .

When you lean on a wall, you are exerting a force on the wall, but you aren't doing any work. On the other hand, if you push a car for a mile, you are clearly doing work. Work, to an engineer, is the force you apply to something, as well as the distance that it moves, in the direction of the applied force. We measure work in *joules*. A joule is one newton of force over one meter.



For example, if you push a car uphill with a force of 10 newtons for 12 meters, you have done 120 joules of work.

Work is how energy is transferred from one thing to another. When you push the car, you also burn sugars(energy of the body) in your blood. That energy is then transferred to the car: after it has been pushed uphill.

Thus, we measure the energy something consumes or generates in units of work: joules, kilowatt-hours, horsepower-hours, foot-pounds, BTUs(British Thermal Unit), and calories.

Let's go over a few different forms that energy can take.

Watch Khan Academy's **Changes in energy** at <https://www.khanacademy.org/science/ms-physics/x1baed5db7c1bb50b:energy/x1baed5db7c1bb50b:changes-in-energy/a/changes-in-energy>

3.1 Forms of Energy

In this section we are going to learn about several different types of energy:

- Heat
- Chemical Energy
- Kinetic Energy
- Gravitational Potential Energy

3.1.1 Heat

When you heat something, you are transferring energy to it. The BTU is a common unit for heat: One BTU is the amount of heat required to raise the temperature of one pound of water, by one degree. One BTU is about 1,055 joules. In fact, when you buy and sell natural gas as fuel, it is priced by the BTU.

3.1.2 Electricity

Electricity is the movement of electrons. When you push electrons through a space that resists their passage (like a light bulb), energy is transferred from the power source (a battery) into the source of the resistance.

Let's say your lightbulb consumes 60 watts of electricity, and you leave it on for 24 hours. We would say that you have consumed 1.44 kilowatt hours or 3,600,000 joules.

Watch Khan Academy's **Introduction to charge** at <https://www.khanacademy.org/science/in-in-class10th-physics/in-in-electricity/in-in-electric-current-circuit/v/intro-to-charge>

3.1.3 Chemical Energy

As mentioned early, some chemical reactions consume energy and some produce energy. Thus, energy can be stored in the structure of a molecule. When a plant uses photosynthesis to rearrange water and carbon dioxide into a sugar molecule, it converts the energy in the sunlight(solar energy) into chemical energy. Remember photosynthesis is a process that releases energy. Therefore, the sugar molecule has more chemical energy than the carbon dioxide and water molecules that were used in its creation.

In our diet, we measure this energy in *kilocalories*. A calorie is the energy necessary to raise one gram of water one degree Celsius: it is about 4.19 joules. This is a very small unit: an apple has about 100,000 calories(100 kilocalories), so people working with food started measuring everything in kilocalories.

Here is where things get confusing: People who work with food got tired of saying "kilocalories", so they just started using "Calorie" to mean 1,000 calories. This has created terrible confusion over the years. So if the C is capitalized, "Calorie" probably means kilocalorie.

3.1.4 Kinetic Energy

A mass in motion has energy. For example, if you are in a moving car and you slam on the breaks, the energy from the motion of the car will be converted into heat in the breaks and under the tires.

How much energy does the car have?

Formula for Kinetic Energy

$$E = \frac{1}{2}mv^2$$

where E is the energy in joules, m is the mass in kilograms, and v is the speed in meters per second.

3.1.5 Gravitational Potential Energy

Watch Khan Academy's **Potential energy** at <https://youtu.be/oGzwVYPxKjg>

When you lift something heavy onto a shelf, you are giving it *potential energy*. The amount of energy that you transferred to it is proportional to its weight and the height that you lifted it.

On the surface of the earth, gravity will accelerate a heavy object downward at a rate of 9.8m/s^2 .

Formula for Gravitational Potential Energy

On earth, then, gravitational potential energy is given by

$$E = (9.8)mh$$

where E is the energy in joules, m is the mass of the object you lifted, and h is the height that you lifted it.

There are other kinds of potential energy. For example, when you draw a bow, you have given that bow potential energy. When you release it, the potential energy is transferred to the arrow, which expresses it as kinetic energy.

3.2 Conservation of Energy

The first law of thermodynamics says “Energy is neither created nor destroyed.”

Energy can change forms: Your cells consume chemical energy to give gravitational potential energy to a car you push up a hill. However, the total amount of energy in a closed system stays constant.

Exercise 5 The Energy of Falling

Working Space

A 5 kg cannonball falls off the top of a 3 meter ladder. Just before it hits the floor, all of its gravitational potential energy has been converted into kinetic energy. How fast is the cannonball going when it hits the floor?

Answer on Page 808

3.3 Efficiency

Watch Khan Academy’s **Laws of thermodynamics** at <https://www.khanacademy.org/science/ap-biology/cellular-energetics/cellular-energy/a/the-laws-of-thermodynamics>

Although energy is always conserved as it moves through different forms, scientists aren’t always that good at controlling it.

For example, a car engine consumes the chemical energy in gasoline. Only about 20% of the energy consumed is used to turn the wheels. Most of the energy is actually lost as heat. If you run a car for a while, the engine gets very hot and the exhaust going out the tailpipe turns hot.

A human is about 25% efficient. Most of the loss is in the heat produced during the chemical reactions that turns food into motion.

In general, if you are trying to increase efficiency in any system, the solution is usually easy to identify because heat is produced. Reduce heat, Increase efficiency.

Light bulbs are an interesting case. To get the light of a 60 watt incandescent bulb, you can

use an 8 watt LED or a 16 watt fluorescent light. Thus, we say that the LED light is much more efficient: If you run both, the incandescent bulb will consume 1.44 kilowatt-hours. The LED will consume only 0.192 kilowatt-hours.

Besides light, the incandescent bulb is producing a lot of heat. If it is inside your house, what happens to the heat? It warms your house.

In the winter, when you want light and heat, the incandescent bulb is 100% efficient!

In the summer, if you are running the air conditioner, the incandescent bulb is worse than just “inefficient at making light” – it is actually counteracting the air conditioner!



CHAPTER 4

Units and Conversions

Accurate measurements are at the heart of good data and good problem solving. Engineers need to be able to describe many different types of phenomena – distance, sound, light, force, and so on.

At this point, you are working with a lot of units: grams for weight, joules for energy, newtons for force, meters for distance, seconds for time, etc. For each type of measurement, there are several different units; for example, distance can be measured in feet, miles, and light-years.

Some Equalencies

Distance	
1 mile	1.6093 kilometers
1 foot	0.3048 meters
1 inch	2.54 centimeters
1 light-year	9.461×10^{12} kilometers
Volume	
1 milliliter	1 cubic centimeter
1 quart	0.9461 liters
1 gallon	3.7854 liters
1 fluid ounce	29.6 milliliters
Mass	
1 pound	0.4535924 kilograms
1 ounce	0.4535924 grams
1 metric ton	1000 kilograms
Force	
1 newton	1 kilogram meter per sec ²
Pressure	
1 pascal	1 newton per square meter
1 bar	0.98692 atmosphere
1 pound per square inch	6897 pascals
Energy	
1 joule	1 newton meter
1 calorie	4.184 joules
1 kilowatt-hour	3.6×10^6 joules

(You don't need to memorize these! Just remember that this page is here.)

In the metric system, prefixes are often used to express a multiple. Here are the common prefixes:

Common Prefixes for Metric Units

giga	$\times 10^9$
mega	$\times 10^6$
kilo	$\times 10^3$
milli	$\div 10^3$
micro	$\div 10^6$
nano	$\div 10^9$

(These are worth memorizing. Here's a mnemonic: "King Henry Doesn't Usually Drink Chocolate Milk." Or Kilo, Hecto, Deca, Unit (for example: gram), Deci, Centi, Mili.

4.1 Conversion Factors

Here is a really handy trick to remembering how to do conversions between units.

Often, you will be given a table like the one above, and someone will ask you “How many miles are in 0.23 light-years?” You know that 1 mile = 1.6093 kilometers and that 1 light-year is 9.461×10^{12} kilometers. How do you do the conversion?

The trick is to treat the two parts of the equality as a fraction that equals 1. That is, you think:

$$\frac{1 \text{ miles}}{1.6093 \text{ km}} = \frac{1.6093 \text{ km}}{1 \text{ miles}} = 1$$

and

$$\frac{1 \text{ light-years}}{9.461 \times 10^{12} \text{ km}} = \frac{9.461 \times 10^{12} \text{ km}}{1 \text{ light-years}} = 1$$

We call these fractions *conversion factors*.

Now, your problem is

$$0.23 \text{ light-years} \times \text{Some conversion factors} = ? \text{ miles}$$

Note that when you multiply fractions together, things in the numerators can cancel with things in the denominator:

$$\left(\frac{31\pi}{47}\right) \left(\frac{11}{37\pi}\right) = \left(\frac{31\pi}{47}\right) \left(\frac{11}{37\pi}\right) = \left(\frac{31}{47}\right) \left(\frac{11}{37}\right)$$

When working with conversion factors, you will do the same with the units:

$$0.23 \text{ light-years} \left(\frac{9.461 \times 10^{12} \text{ km}}{1 \text{ light-years}} \right) \left(\frac{1 \text{ miles}}{1.6093 \text{ km}} \right) = \\ 0.23 \cancel{\text{light-years}} \left(\times \frac{9.461 \times 10^{12} \text{ km}}{\cancel{1 \text{ light-years}}} \right) \left(\frac{1 \text{ miles}}{1.6093 \text{ km}} \right) = \frac{(0.23)(9.461 \times 10^{12})}{1.6093} \text{ miles}$$

Exercise 6 Simple Conversion Factors*Working Space*

How many calories are in 4.5 kilowatt-hours?

*Answer on Page 808***4.2 Conversion Factors and Ratios**

Conversion factors also work on ratios. For example, if you are told that a bug is moving 0.5 feet every 120 milliseconds. What is that in meters per second?

The problem then is

$$\frac{0.5 \text{ feet}}{120 \text{ milliseconds}} = \frac{\text{? m}}{\text{second}}$$

So you will need conversion factors to replace the “feet” with “meters” and to replace “milliseconds” with “seconds”:

$$\left(\frac{0.5 \text{ feet}}{120 \text{ milliseconds}} \right) \left(\frac{0.3048 \text{ meters}}{1 \text{ foot}} \right) \left(\frac{1000 \text{ milliseconds}}{1 \text{ second}} \right) = \frac{(0.5)(0.3048)(1000)}{120} \text{ m/second}$$

Exercise 7 Conversion Factors*Working Space*

The hole in the bottom of the boat lets in 0.1 gallons every 2 minutes. How many milliliters per second is that?

Answer on Page 809

4.3 When Conversion Factors Don't Work

Conversion factors only work when the units being converted are proportional to each other. Gallons and liters, for example, are proportional to each other: If you have n gallons, you have $n \times 3.7854$ liters.

Degrees celsius and degrees farenheit are *not* proportional to each other. If your food is n degrees celsius, it is $n \times \frac{9}{5} + 32$ degrees farenheit. You can't use conversion factors to convert celsius to farenheit.

Watch Khan Academy's video on this at <https://www.khanacademy.org/test-prep/sat/x0a8c2e5f:untitled-652/x0a8c2e5f:problem-solving-and-data-analysis-lessons-by-skill/a/gtp--sat-math--article--units--lesson>



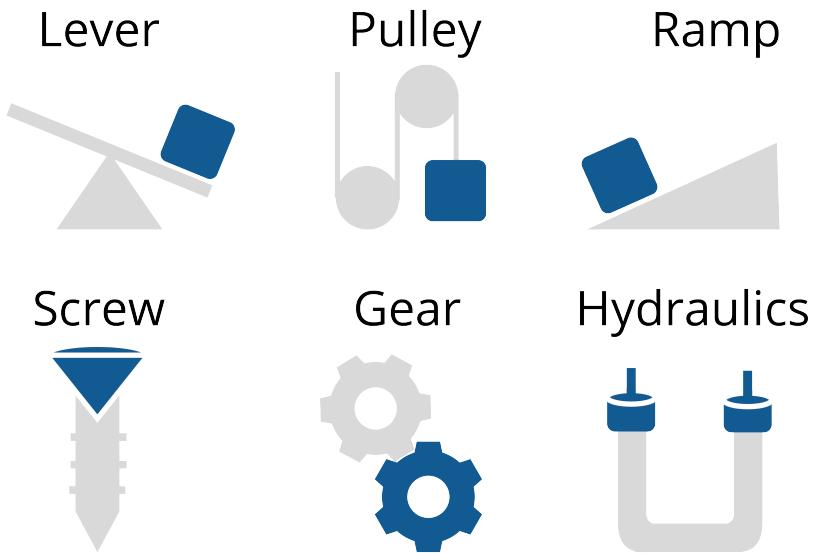
CHAPTER 5

Simple Machines

As mentioned earlier, physicists define work to be the force applied times the distance it is applied over. So, if you pushed your car 100 meters with 17 newtons of force, you have done 1700 joules of work.

Humans have always had to move really heavy things, so many centuries ago we developed simple machines to decrease the amount of force necessary to execute those tasks. These include things like:

- Levers
- Pulleys
- Ramps
- Gears
- Hydraulics
- Screws



While these machines can decrease the force needed, they don't change the amount of work that must be done. So if the force is decreased to a third, the distance that you must apply the force is increased by a factor of three.

"Mechanical gain" is what we call the increase in force.

5.1 Levers

A lever rotates on a fulcrum. To decrease the necessary force, the load is placed nearer to the fulcrum than where the force is applied.

In particular, physicists talk about the *torque* created by a force. When you push on a lever, the torque is the product of the force you exert and the distance from the point of rotation.

Torque is typically measured in newton-meters.

To balance two torques, the products must be the same. So, assuming that the forces are applied in the proper direction,

$$R_L F_L = R_A F_A$$

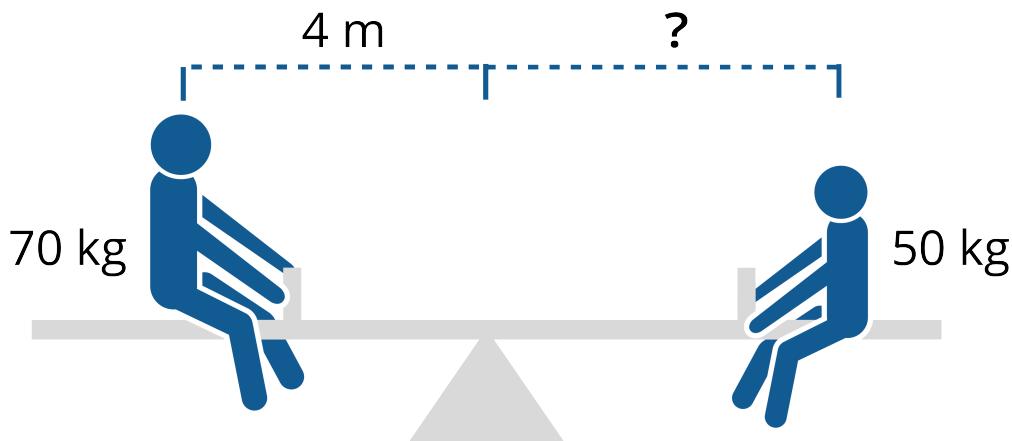
where R_L and R_A are the distance from the fulcrum to the where the load's force and the applied force (respectively) are applied, and F_L and F_A are the amounts of the forces.

Exercise 8 Lever**Working Space**

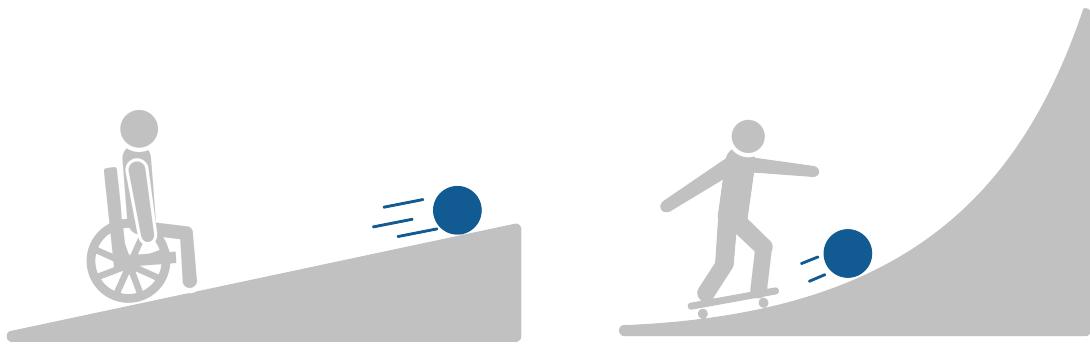
Paul, who weighs 70 kilograms, sits on a see-saw 4 meters from the fulcrum. Jan, who weighs 50 kilograms, wants to balance. How far should Jan sit from the fulcrum?

Answer on Page 809

Watch Khan Academy's video on levers: <https://www.khanacademy.org/science/physics/discoveries/simple-machines-explorations/a/lever>

**5.2 Ramps**

Ramps, or incline planes, let you roll or slide objects up to a higher level. Steeper ramps give you less mechanical gain. For example, it is much easier to roll a ball up a wheelchair ramp than on a skateboard ramp.



Assuming the ramp has a constant steepness, the mechanical gain is equal to the ratio of the length of the ramp divided by the amount that it rises.

If you assume there is no friction, the force that you push a weight up the ramp will be:

$$F_A = \frac{V}{L} F_G$$

Where F_A is the force you need to push. L is the length of the ramp, V is the amount of vertical gain and F_G is the force of gravity on the mass.

(We haven't talked about the sine function yet, but in case you already know about it:
Note that

$$\frac{V}{L} = \sin \theta$$

where θ is the angle between the ramp and level.)

Exercise 9 Ramp

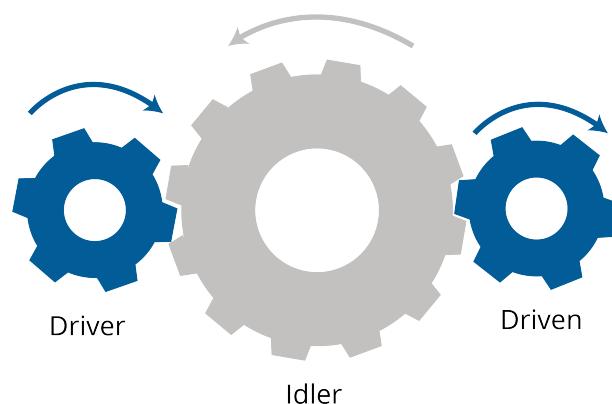
A barrel of oil weighs 136 kilograms. You can push with a force of up to 300 newtons. You have to get the barrel onto a platform that is 2 meters. What is the shortest board that you can use as a ramp?

Working Space

Answer on Page 809

5.3 Gears

Gears (which might have a chain connecting them like on a bicycle) have teeth and come in pairs. You apply torque to one gear, and it applies torque to another. The torque is increased or decreased based on the ratio between the teeth on the gears.



If N_A is the number of teeth on the gear you are turning with a torque of T_A , and N_L is the number of teeth on the gear it is turning, the resulting torque is:

$$T_L = \frac{N_A}{N_L} T_A$$

Exercise 10 Gears

Working Space

The bicycle is an interesting case because we are not trying to get mechanical gain. We want to spin the pedals slower with more force.

You like to pedal your bike at 70 revolutions per minute. The chainring that is connected to your pedals has 53 teeth. The circumference of your tire is 2.2 meters. You wish to ride a 583 meters per minute.

How many teeth should the rear sprocket have?

Answer on Page 809

Watch Khan Academy's introduction to simple machines here: <https://www.khanacademy.org/science/physics/discoveries/simple-machines-explorations/a/simple-machines-and-how-to-use-them>

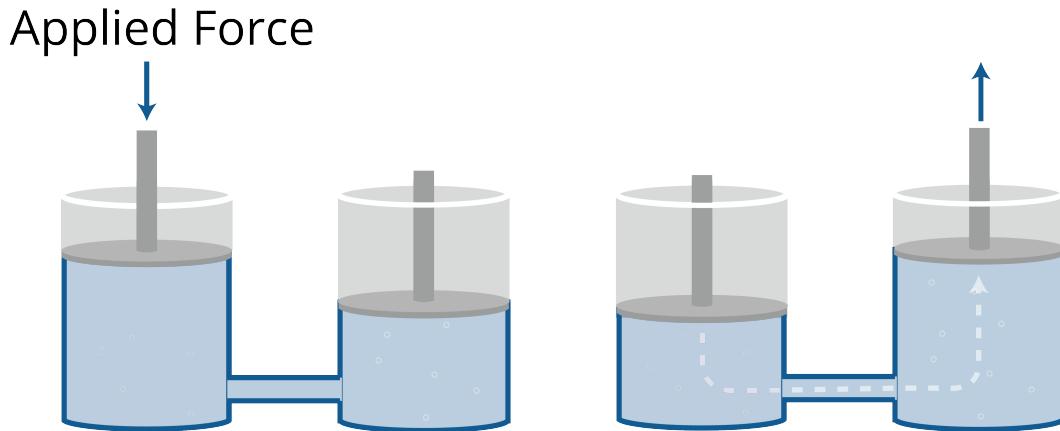
Further Reading: **The Way Things Work Now**
By David Macaulay

https://www.academia.edu/39339685/The_Way_Things_Work_Now_by_David_Macaulay

FIX ME: Aaron, what are your thoughts about adding a few, mega-classic books to the resources?

5.4 Hydraulics

In a hydraulic system, like the braking system of a car, you exert force on a piston filled with fluid. The fluid carries that pressure into another cylinder. The pressure of the fluid pushes the piston in that cylinder out.



The pressure in the hose can be measured in pounds per square inch (PSI) or newtons per square meter (Pascals or Pa). We will use Pascals.

To figure out how much pressure you create, you divide the force by the area of the piston head you are pushing.

To figure out how much force that creates on the other end, you multiply the pressure times the area of the piston head that is pushing the load.

Exercise 11 **Hydraulics**

Working Space

Your car has disc brakes. When you put 2,500,000 pascals of pressure on the brake fluid, the car stops quickly. As the car designer, you would like that to require 12 newtons of force from the driver's foot. What should the radius of the master cylinder (the one the driver is pushing on) be?

Answer on Page 810



CHAPTER 6

Buoyancy

The word buoyancy probably brings to mind images of floating in water. Before we dive in, let's zoom out for a moment and consider that the study of buoyancy is about much more than just boats and water. You might be thinking: I want to be a computer programmer, why do I need to know about buoyancy? This topic is much bigger than it might seem at first glance. Buoyancy concerns how all liquids and gasses interact with gravity. The concept of buoyancy is connected to fundamental concepts about how things work in the universe. The *buoyant force*, as it's known in engineering, is an important concept that has wide ranging applications. A big part of engineering is moving stuff around, and understanding buoyancy helps us solve problems where we need to move things in and through fluids. Even if you don't have plans to build a robotic submarine, these are super useful ideas to be familiar with. We'll start exploring the topic with familiar scenarios around boats and water.

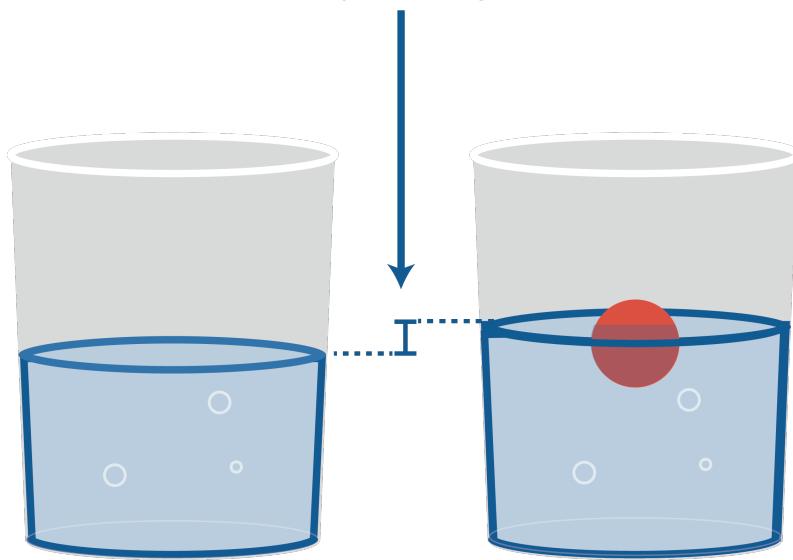
When you put a boat into water, it will sink into the water until the mass of the water it displaces is equal to the mass of the boat. We think of this in terms of forces. Gravity pulls the mass of the boat down. The *buoyant force* pushes the boat up. A boat dropped into the water will bob up and down a bit before reaching an *equilibrium* where the two forces are equal.

Watch Khan Academy's introduction to buoyancy at <https://www.khanacademy.org/science/in-in-class9th-physics-india/in-in-gravity/in-in-pressure-in-liquids-archimedes-principle/archimedes-principle-buoyancy-fluids-physics-khan-academy>

The buoyant force pushes things up – against the force of gravity. The force is equal to the weight of the fluid being replaced. So, for example, a cubic meter of freshwater has a mass of about 1000kg. If you submerge anything with a volume of one meter in freshwater on earth, the buoyant force will be about 9800 newtons.

For some things, like a block of styrofoam, this buoyant force will be sufficient to carry it to the surface. Once it reaches the surface, it will continue to rise (displacing less water) until the mass of the water it displaces is equal to its mass. And then we say “It floats!”

Water displaced by the ball



For some things, like a block of lead, the buoyant force is not sufficient to lift it to the surface, and thus we say “It sinks!”

This is why a helium balloon floats through the air. The air that it displaces weighs more than the balloon and the helium itself. (It is easy to forget that air has a mass, but it does.)

Exercise 12 Buoyancy

You have an aluminum box that has a heavy base, so it will always float upright. The box and its contents weigh 10 kg. Its base is 0.3 m x 0.4 m. It is 1m tall. When you drop it into freshwater (1000kg/m^3), how far will it sink before it reaches equilibrium.

Working Space

Answer on Page 810

6.1 The Mechanism of Buoyancy: Pressure

As you dive down in the ocean, you will experience greater and greater pressure from the water. And if you take a balloon with you, you will gradually see it get smaller as the water pressure compresses the air in the balloon.

Let's say you are 3 meters below the surface of the water. What is the pressure in Pascals (newtons per square meter)? You can think of the water as a column of water crushing down upon you. The pressure over a square meter is the weight of 3 cubic meters of water pressing down.

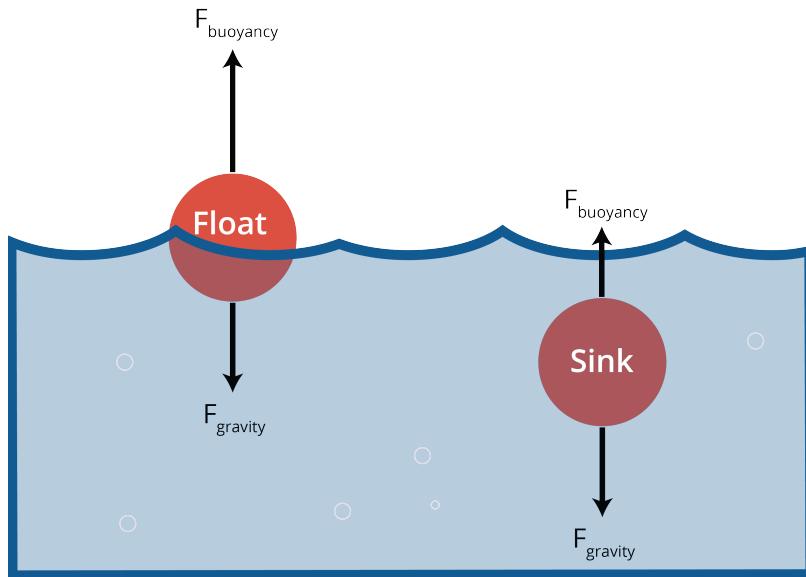
$$p = (3)(1000)(9.8) = 29,400 \text{ Pa}$$

This is called *hydrostatic pressure*. The general rule for hydrostatic pressure in Pascals p is

$$p = dgh$$

Where d is the density of the fluid in kg per cubic meter, g is the acceleration due to gravity in m/s^2 , and h is the height of the column of fluid above you.

So, where does buoyant force come from? Basically, the pressure pushing up on the deepest part of the object is higher than the pressure pushing down on the shallowest part of the object. That is where buoyancy comes from.



Exercise 13 Hydrostatic Pressure

Working Space

You dive into a tank of olive oil on Mars. How much more hydrostatic pressure does your body experience at 5 meters deep than it did at the surface?

The density of olive oil is about 900 kg per square meter. The acceleration due to gravity on Mars is 3.721 m/s^2 .

Answer on Page 810

6.2 The Mechanism of Buoyancy: Density

Notice that although the pressure is increasing as you go deeper, the buoyant force will *not increase* because the buoyant force is always equal to the weight of the fluid that is displaced, regardless if that is 1 meter or 100 meters underwater.

Also, saltwater is denser than freshwater. That is why people float better in the sea than they do in a river.

And, lipids, like fats and oils, are less dense than water. This is why oil floats in a glass of water. Animals like Polar bears and seals pack on the fat to keep them warm and it also helps them float.

Here is a real world example of a buoyancy study with implications for engineers designing ventilation systems: <https://www.scientificamerican.com/podcast/episode/snot-clouds-achieve-unexpected-buoyancy/>





CHAPTER 7

Heat

Let's say you put a 1 kg aluminum pan that is 80° C into 3 liters of water that is 20° C . Energy, in the form of heat, will be transferred from the pan to the water until they are at the same temperature. (We call this "thermal equilibrium.")

What will the temperature of the water be?

7.1 Specific Heat Capacity

If you are heating something, the amount of energy you need to transfer to it depends on three things: the mass of the thing you are heating, the amount of temperature change you want, and the *specific heat capacity* of that substance.

Energy in Heat Transfer

The energy moved in a heat transfer is given by

$$E = mc\Delta T$$

where m is the mass, ΔT is the change in temperature, and c is the specific heat capacity of the substance.

(Note that this assumes no phase change. For example, this formula works nicely on warming liquid water, but it gets more complicated if you warm the water past its boiling point.)

Can we guess the specific heat capacity of a substance? It is very, very difficult to guess the specific heat of a substance, so we determine it by experimentation.

For example, someone determined that it took about 0.9 joules to raise the temperature of solid aluminum one degree Celsius. So we say “The specific heat capacity of aluminum is 0.9 J/g °C.”

The specific heat capacity of liquid water is about 4.2 J/g °C.

To answer the question, then, the amount of energy given off by the pan must equal the amount of energy absorbed by the water. And they need to be the same temperature at the end. Let T be the final temperature of both.

Watch Khan Academy’s discussion of heat capacity at <https://www.khanacademy.org/science/ap-chemistry-beta/x2eef969c74e0d802:thermodynamics/x2eef969c74e0d802:heat-capacity-and-calorimetry/v/heat-capacity>

Three liters of water weighs 3,000 grams, so the change in energy in the water will be:

$$E_W = mc\Delta T = (3000)(4.2)(T - 20) = 12600T - 252000 \text{ joules}$$

The pan weighs 1000 grams, so the change in energy in the pan will be::

$$E_P = mc\Delta T = (1000)(0.9)(T - 80) = 900T - 72000 \text{ joules}$$

Total energy stays the same so $E_W + E_P = 0$. So you need to solve

$$(12600T - 252000) + (900T - 72000) = 0$$

And find that the temperature at equilibrium will be

$$T = 24^\circ\text{C}$$

Exercise 14 Thermal Equilibrium

Working Space

Just as you put the aluminium pan in the water as described above, someone also puts a 1.2 kg block of copper cooled to 10°C . The specific heat of solid copper is about $0.4 \text{ J/g } ^\circ\text{C}$.

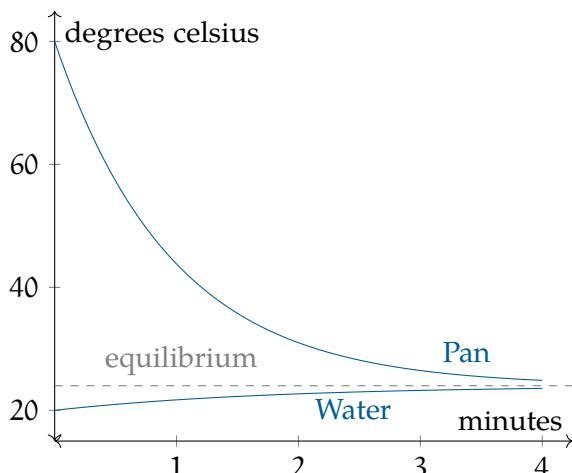
What is the new temperature at equilibrium?

Answer on Page 810

7.2 Getting to Equilibrium

When two objects with different temperatures are touching, the speed at which they exchange heat is proportional to the differences in their temperatures. Thus, as their temperatures get closer together, the heat exchange slows down.

In our example, the pan and the water will get close to equilibrium quickly, but they may never actually reach equilibrium.



Exercise 15 Cooling Your Coffee*Working Space*

You have been given a ridiculously hot cup of coffee and a small pitcher of chilled milk.

You need to start chugging your coffee in three minutes, and you want it as cool as possible at that time. When should you add the milk to the coffee?

Answer on Page 811

7.3 Specific Heat Capacity Details

For any given substance, the specific heat capacity often changes a lot when the substance changes state. For example, ice is 2.1 J/g °C, whereas liquid water is 4.2 J/g°C.

Watch Khan Academy's discussion of the specific heat of water: <https://www.khanacademy.org/science/biology/water-acids-and-bases/water-as-a-solid-liquid-and-gas/v/specific-heat-of-water>

Even within a given state, the specific heat capacity varies a bit based on the temperature and pressure. If you are trying to do these sorts of calculations with great accuracy, you will want to find the specific heat capacity that matches your situation. For example, I might look for the specific heat capacity for water at 22°C at 1 atmosphere of pressure(atm).



CHAPTER 8

Cognitive Biases 1

In this section we're going to take a look at research findings about *cognitive biases*. These are universal quirks found in the human thought process. Cognitive biases aren't biases like racial biases. Everyone, regardless of nationality, race or gender is subject to these cognitive traps. You might be wondering, why do I need to learn about cognitive science in order to be an engineer? The most important tool we have as problem solvers is our own minds. We're going to be looking at ways that our minds can trip us up.

Our brains were designed over millions of years by the evolutionary process. The resulting mind is an amazing and powerful tool, however not flawless. The human brain has tendencies (or biases) that nudge us toward bad judgment and poor decisions.

When someone first gave you a hammer they handed it over with a warning, "don't hit your thumb!" No matter how careful you are with the hammer, at some point you'll hit your thumb. It's the same with cognitive biases. In the course of life all of us will fall prey to these cognitive biases. Knowing about them is the first step in protecting ourselves.

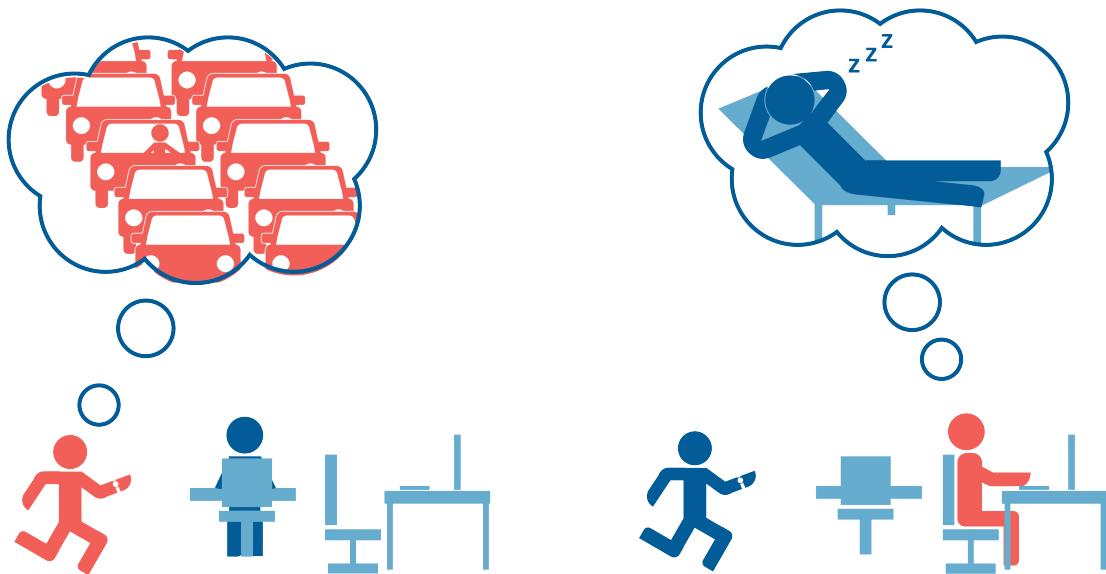
It would be irresponsible to teach you powerful ideas without also teaching you about the cognitive biases that follow them. There are about 50 that you should know about, but let's start with only a few.

8.1 Fundamental Attribution Error

You tend to attribute the mistakes of another person to their character, but attribute your own mistakes to the situation.

If someone asks you “Why were you late for work today!” You are likely to have an excuse, ‘I got stuck in a crazy traffic jam.’

If you notice your coworker is late for work, you are likely to say “My coworker is lazy.”



The solution? Cut people some slack. You probably don't know the whole story, so assume that their character is as strong as yours.

Or maybe you also need to hold yourself to a higher standard? Do you find yourself frequently rationalizing your bad judgment, lateness, or rudeness? This could be an opportunity for you to become a better person whose character is stronger regardless of the situation.

8.2 Self-Serving Bias

Self-serving bias is when you blame the situation for your failures, but attribute your successes to your strengths.

For example, when asked “Why did you lose the match?” you are likely to answer “The referee wasn’t fair.” When you are asked “Why did you win the match?” you are likely to answer “Because I have been training for weeks, and I was very focused.”

This bias tends to make us feel better about ourselves, but it makes it difficult for us to be objective about our strengths and weaknesses.

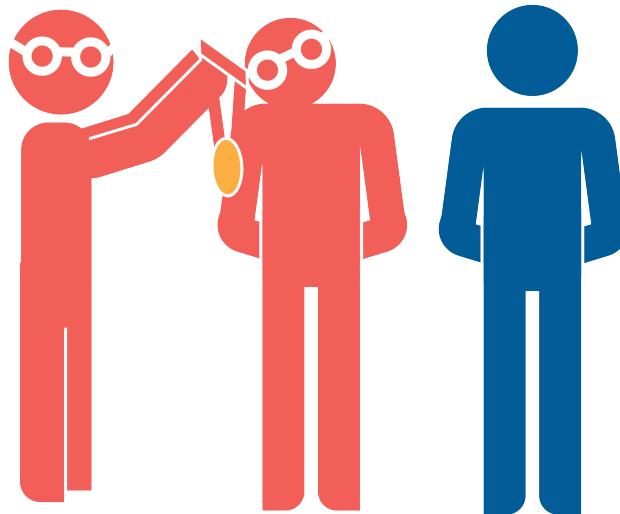
8.3 In-group favoritism

In-group favoritism: We tend to favor people who are in a group with us over people who are not in groups with us.

When asked “Who is the better goalie, Ted or John?”

If Ted is a Star Trek fan like you, you are likely to think he is also a good goalie.

As you might imagine, this unconscious tendency is the source of a lot of subtle discrimination based on race, gender, age, and religion.



8.4 The Bandwagon Effect and Groupthink

The bandwagon effect is our tendency to believe the same things that the people around us believe. This is how fads spread so quickly: one person buys in, and then the people they know have a strong tendency to buy in as well.

Groupthink is similar: To create harmony with the people around us, we go along with things we disagree with.

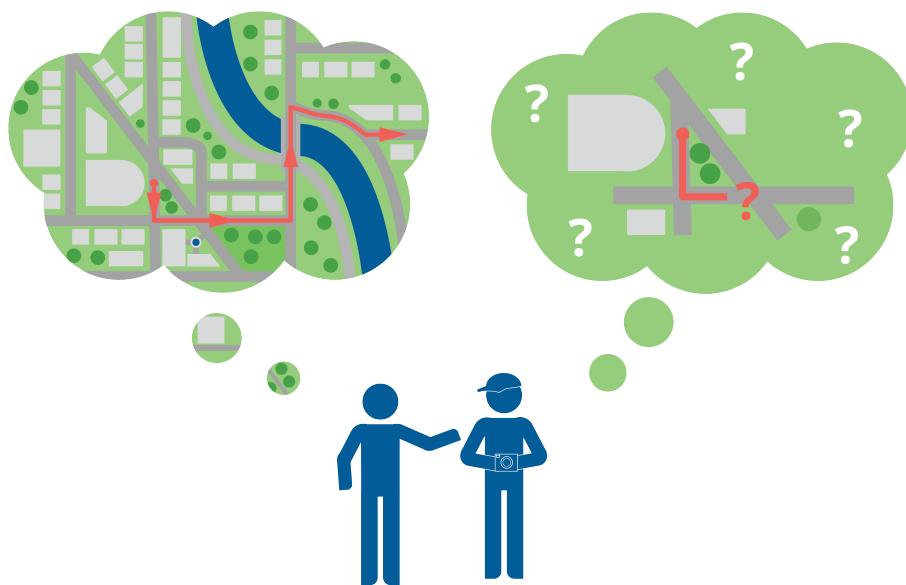
It takes a lot of perspective to recognize when those around us are wrong. And it takes even more courage to openly disagree with them.

8.5 The Curse of Knowledge

Once you know something, you tend to assume everyone else knows it too.

This is why teaching is sometimes difficult: a teacher will assume that everyone in the audience already knows the same things the teacher knows.

For example, imagine a local who has lived in a city for years giving directions to a tourist. The local has an in-depth understanding of the city, and gives overly quick and detailed instructions. The tourist politely smiles and nods, but stopped following after the local began listing unfamiliar street names and landmarks.



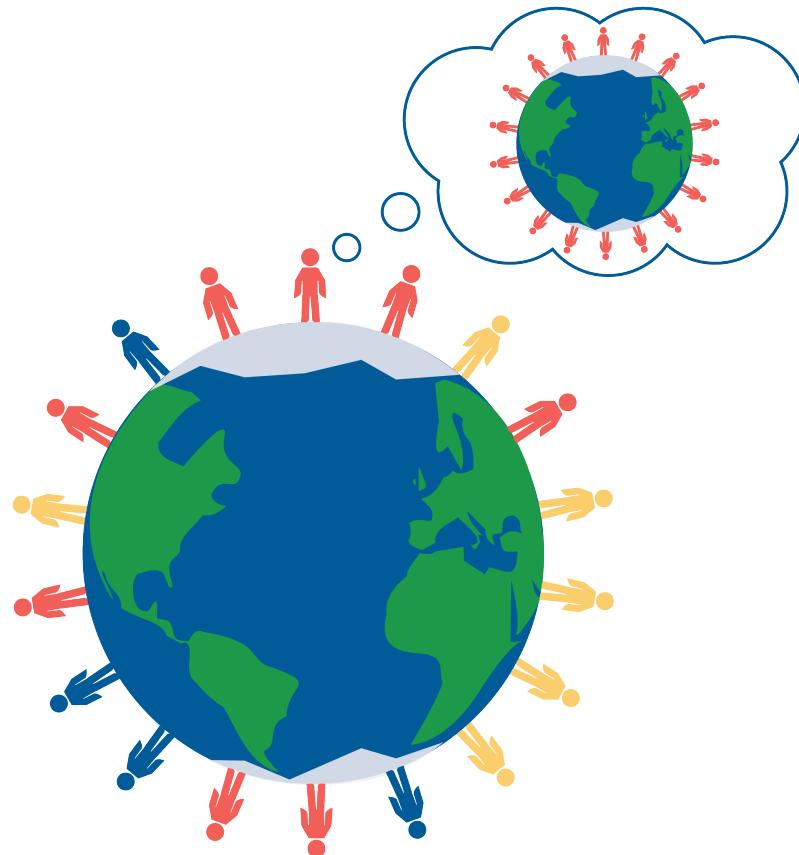
Also, when we learn that a friend doesn't know something that we know, we are often very surprised. This surprise can sometimes manifest as hurtful behavior.

When I find a gap in a friend's knowledge, I try to remind myself that the friend certainly knows many things that I don't. I also try to imagine how it would feel if they teased me for my ignorance.

8.6 False Consensus

We tend to believe that more people agree with us than is actually the case. For example, if you are a member of a particular religion, you tend to overestimate the percentage of people in the world who are members of that religion.

When people vote in elections, they are often surprised when their preferred candidate loses. "Everyone, and I mean EVERYONE, voted for Smith!" they yell. "There must have been a mistake in counting the votes."

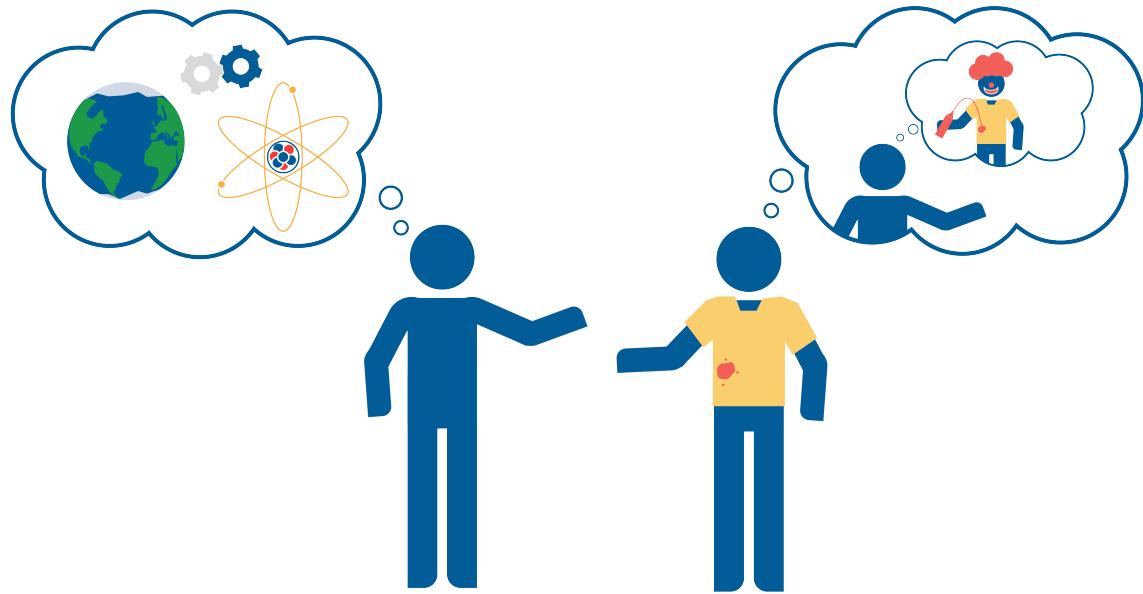


8.7 The Spotlight Effect

You tend to overestimate how much other people are paying attention to your behavior and appearance.

Think of six people that you talked to today. Can you even remember what shoes most of them were wearing? Do you care? Do you think any of them remember which shoes you wore today?

There is an old saying “You would worry a lot less about how people think of you, if you realized how rarely they do.”



8.8 The Dunning-Kruger Effect

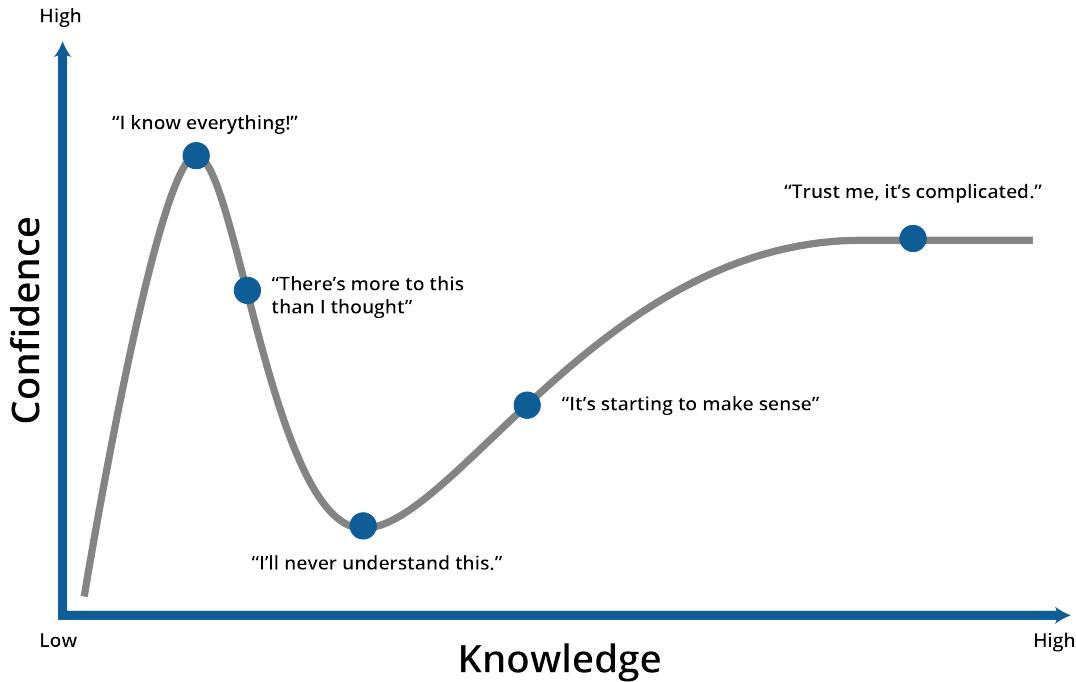
The less you know, the more confident you are.

When a person doesn't know all the nuance and context in which a question is asked, the question seems simple. Thus the person tends to be confident in their answer. As they learn more about the complexity of the space in which the question lives, they often realize the answer is not nearly so obvious.

For example, a lot of people will confidently proclaim “Taxes are too high! We need to lower taxes.” An economist who has studied government budgets, deficits, history, and monetary policy, might say something like “Maybe taxes *are* too high. Or maybe they are too low. Or maybe we are taxing the wrong things. It is a complex question.”

When I am talking with people about a particular topic, I do my best to defer to the person in the conversation who I think has the most knowledge in the area. If I disagree with the person, I try to figure out why our opinions are different.

Similarly, you should assume that any opinion that is voiced, specifically, in an internet discussion is wildly over-simplified. If you really care about the subject, read a book by a respected expert. Yes, a whole book – there are few interesting topics that can be legitimately explained in less than 100 pages.

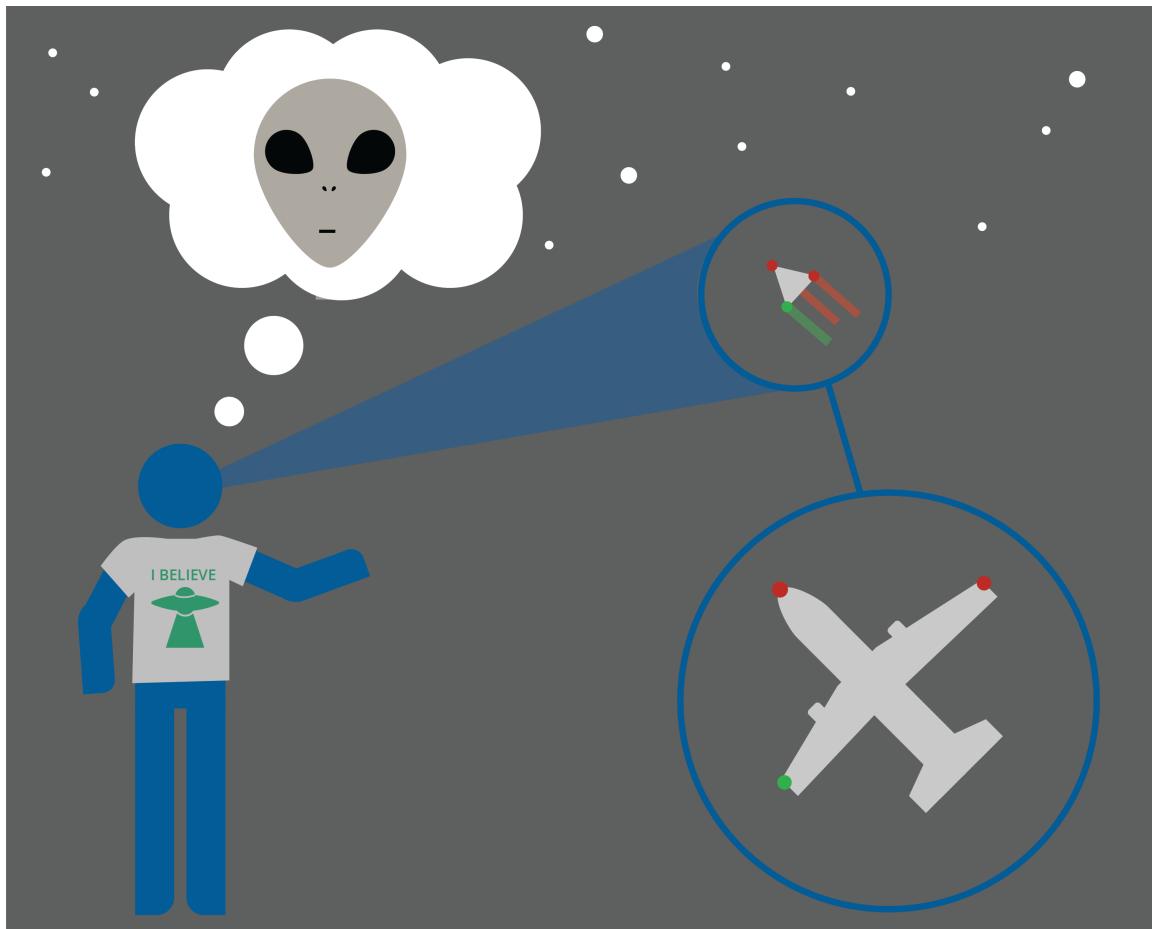


8.9 Confirmation Bias

You tend to find and remember information that supports beliefs you already have. You tend to avoid and dismiss information that contradicts your beliefs.

If you believe that intelligent creatures have visited from other planets, you will tend to look for data to support your beliefs. When you find data that shows that it is just too far for any creature to travel, you will try to find a reason why the data is incorrect.

Confirmation bias is one reason why people don't change their beliefs more often.



Confirmation bias wrecks many, many studies. The person doing the study often has a hypothesis that they believe and very much want to prove true. It is very tempting to discard data that doesn't support the hypothesis. Or maybe the person throws all the data away and experiments again and again until they get the result they want.

When you design an experiment, you must describe it explicitly before you start. You must tell someone: "If the hypothesis I love is incorrect, the results will look like this. If the hypothesis I love is correct, the results will look like that. And if the results look any other way, I have neither proved nor disproved the hypothesis."

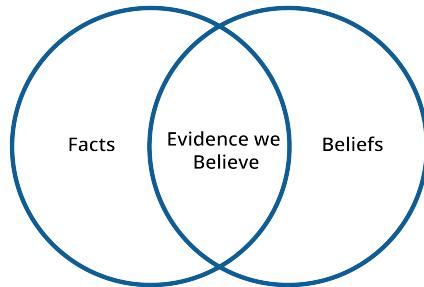
Once the experiment is underway, you must not change the plan and you must not discard any data.

This is scientific integrity. You should demand it from yourself, and you should expect it from others.

Watch a TED Talk and Learn more about Confirmation Bias: What shapes our perceptions (and misperceptions) about science? In an eye-opening talk, meteorologist J. Marshall Shepherd explains how confirmation bias, the Dunning-Kruger effect and cognitive

dissonance impact what we think we know – and shares ideas for how we can replace them with something much more powerful: knowledge.

https://www.ted.com/talks/j_marshall_shepherd_3_kinds_of_bias_that_shape_your_worldview



8.10 Survivorship bias

You will pay more attention to those that survived a process than those who failed.

After looking at a lot of old houses, you might say “In the 1880s, they built great houses.” However, you haven’t seen the houses that were built in the 1880s and didn’t survive. Which houses tended to survive for a long time? Only the great houses – you are basing your opinion on a very skewed sample.





CHAPTER 9

Friction

Imagine there is a large and heavy steel box resting in the middle of a large floor. Imagine you push it hard enough to get it moving. If you stop pushing, will it continue to glide gracefully across the floor?

Probably not. Unless the floor is very slippery for some reason, the box will come to a halt immediately after you stop pushing. We would say that it is stopped by the force of *friction*.

What's really happening? The kinetic energy of the box is being converted into heat between the bottom of the box and floor. As the bottom of the box and the floor get warmer, the speed of the box decreases.

The amount of friction is proportional to the force with which the box is pressing against the floor – so you should expect a box that is twice as heavy to experience twice as much frictional force.

That is, the frictional force is proportional to the normal force. (FIXME: picture here)

The amount of friction is also determined by the materials that are sliding against each

other. For example, if the floor is ice, the frictional force will be less than if the floor is made of wood.

If you are pushing the box with a force of F and it is moving but neither accelerating nor decelerating, then the force you are applying is exactly balanced by the frictional force. If the box is pressing against the floor with a force of N , then we say the *coefficient of friction* between the steel box and the floor is given by

$$\mu = \frac{F}{N}$$

Exercise 16 Bicycle Stopping

Working Space

You are riding your bicycle at 11 meters per second when you slam on the brakes and lock up the wheels.

You weigh 55 kg.

When any piece of rubber is skidding across a dry road, the coefficient of friction will be about 0.7.

Answer the following questions:

- How much kinetic energy do you have when you engage the brakes?
- As you skid, how much frictional force is decelerating you?
- For how many meters will you slide?

Answer on Page 811

Notice that the force of friction is not determined by how much of the tire is touching the ground. The coefficient of friction of the two materials and the normal force all all you need to compute the friction.

9.1 Static vs Kinetic Friction Coefficients

Once again, imagine the box resting on the floor. As you start to push it, it will sit still until your force is greater than the force of friction. However, once it starts moving, the force of friction seems to be less.

Between two materials, there is actually 2 different friction coefficients:

- Kinetic friction coefficient: The coefficient you use once the box is sliding against the floor.
- Static friction coefficient: The coefficient you use to figure out how much force you need to get the box to start to move.

The kinetic friction coefficient is always less than the static friction coefficient:

- *Kinetic*, μ_k : For a car skidding on a dry road, the friction coefficient is about 0.7.
- *Static*, μ_s : When the car is parked with its brakes on, it has a friction coefficient of about 1.0.

Exercise 17 Rocket Sled**Working Space**

You are built a rocket sled with steel runners on a flat, level wooden floor. The sled weighs 50 kg and you weigh 55 kg. Before you get on the sled, you push it around the floor some. You find that you can get it to move from a standstill if you push it with a force of 270 N. Once it is moving, you can keep it moving at the same speed using a force of 220 N.

What are μ_s and μ_k of your sled's runners on your wooden floor?

Now you get on the sled and gradually increase the thrust of the rocket mounted on the sled until it starts to move. Then you keep the thrust constant.

How much force was the rocket exerting on you and the sled when it started to move?

How how fast do you accelerate now that the sled is moving?

Answer on Page 811

9.2 Skidding and Anti-Lock Braking Systems

When a car goes through a curve, the friction of the tire on the road is what changes the direction of the car's travel. Even though the wheel is turning, this is the static friction coefficient because the surface of the tire is not sliding across the road.

If you go into the curve too fast, the tire may not have enough friction to turn the car. In this case the car will start to slide sideways. Now the friction between the tire and road uses the kinetic coefficient. That is, you have significantly less friction than you had before you started to skid.

When you are driving a car, the force of friction that your tires create is your friend. It lets you steer, accelerate, and stop.

In older cars, if you would panicked and slammed on the brakes, you would probably lock up the wheels: they would stop turning suddenly. And the surface of the tire would begin to slide across the pavement. At that moment, two problems occurred:

- You don't stop as quickly because now the friction between your tires and the road is based on the kinetic friction coefficient instead of the static friction coefficient.
- You can't steer the car. Steering only happens because the wheels are turning in a particular direction.

To prevent this problem, car companies developed the anti-lock brake system or ABS.

FIXME: More here.



CHAPTER 10

The Greek Alphabet

If you do anything involving math or physics, you will use a lot of Greek letters. Here is a table for your reference:

Capital	Lower	Pronounced	Capital	Lower	Pronounced
Α	α	Alpha	Ν	ν	Nu
Β	β	Beta	Ξ	ξ	Xi ("ku-ZY")
Γ	γ	Gamma	Ο	ο	Omicron
Δ	δ	Delta	Π	π	Pi
Ε	ε	Epsilon	Ρ	ρ	Rho
Ζ	ζ	Zeta	Σ	σ	Sigma
Η	η	Eta	Τ	τ	Tau
Θ	θ	Theta	Υ	υ	Upsilon
Ι	ι	Iota	Φ	φ	Phi
Κ	κ	Kappa	Χ	χ	Chi ("Kai")
Λ	λ	Lambda	Ψ	ψ	Psi ("Sigh")
Μ	μ	Mu	Ω	ω	Omega



CHAPTER 11

Basic Statistics

You live near a freeway, and someone asks you, "How fast do cars on that freeway drive?"

You say "Pretty fast."

And they say, "Can you be more specific?"

And you point your radar gun at a car, and say "That one is going 32.131 meters per second."

And they say, "I don't want to know about that specific car. I want to know about all the cars."

So, you spend the day beside the freeway measuring the speed of every car that goes by. And you get a list of a thousand numbers. Here is part of the list:

30.462 m/s	29.550 m/s	29.227 m/s
37.661 m/s	27.899 m/s	28.113 m/s
24.382 m/s	35.668 m/s	43.797 m/s
31.312 m/s	37.637 m/s	30.891 m/s

There are 12 numbers here. We say that there are 12 *samples*.

11.1 Mean

We often talk about the *average* of a set of samples, which is the same as the *mean*. To get the mean, sum up the samples and divide that number by the number of samples.

The numbers in that table sum to 388.599. If you divide that by 12, you find that the mean of those samples is 32.217 m/s.

We typically use the greek letter μ ("mu") to represent the mean.

Definition of Mean

If you have a set of samples x_1, x_2, \dots, x_n , the mean is:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

This may be the first time you are seeing a summation (\sum). The equation above is equivalent to:

$$\mu = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

Exercise 18 Mean Grade

Working Space

Teachers often use the mean for grading. For example, if you took six quizzes in a class, your final grade might be the mean of the six scores. Find the mean of these six grades: 87, 91, 98, 65, 87, 100.

Answer on Page 812

If you tell your friend "I measured the speed of 1000 cars, and the mean is 31.71 m/s", your friend will wonder "Are most of the speeds clustered around 31.71? Or are they all

over the place and just happen to have a mean of 31.71?" To answer this question we use variance.

11.2 Variance

Definition of Variance

If you have n samples x_1, x_2, \dots, x_n that have a mean of μ , the *variance* is defined to be:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

That is, you figure out how far each sample is from the median, you square that, and then you take the mean of all those squared distances.

x	$x - \mu$	$(x - \mu)^2$
30.462	-1.755	3.079
29.550	-2.667	7.111
29.227	-2.990	8.938
37.661	5.444	29.642
27.899	-4.318	18.642
28.113	-4.104	16.839
24.382	-7.835	61.381
35.668	3.451	11.912
43.797	11.580	134.106
31.312	-0.905	0.818
37.637	5.420	29.381
30.891	-1.326	1.757
$\sum x = 386.599$		$\sum (x - \mu)^2 = 323.605$
mean = 32.217		variance = 26.967

Thus, the variance of the 12 samples is 26.967. The bigger the variances, the farther the samples are spread apart; the smaller the variances, the closer samples are clustered around the mean.

Notice that most of the data points deviate from the mu by 1 to 5 m/s. Isn't it odd that the variance is a big number like 26.967? Remember that it represents the average of the squares. Sometimes, to get a better feel for how far the samples are from the mean, we use the square root of the variance, which is called the *standard deviation*.

The standard deviation of your 12 samples would be $\sqrt{26.967} = 5.193$ m/s.

The standard deviation is used to figure out a data point is an outlier. For example, if you

are asked “That car that just sped past. Was it going freakishly fast?” You might respond, “No, it was within a standard deviation of the mean.” or “Yes, its speed was 2 standard deviations more than the mean. They will probably get a ticket.”

A singular μ usually represents the mean. σ usually represents the standard deviation. So σ^2 represents the variance.

Exercise 19 Variance of Grades

Working Space

Now find the variance for your six grades.

As a reminder, they were: 87, 91, 98, 65,
87, 100.

What is your standard deviation?

Answer on Page 812

11.3 Median

Sometimes you want to know where the middle is. For example, you want to know the speed at which half the cars are going faster and half are going slower. To get the median, you sort your samples from smallest to largest. If you have an odd number of samples, the one in the middle is the median. If you have an even number of samples, we take the mean of the two numbers in the middle.

In our example, you would sort your numbers and find the two in the middle:

24.382
27.899
28.113
29.227
29.550
30.462
30.891
31.312
35.668
37.637
37.661
43.797

You take the mean of the two middle numbers: $(30.462 + 30.891)/2 = 30.692$. The median speed would be 30.692 m/s.

Medians are often used when a small number of outliers majorly skew the mean. For example, income statistics usually use the median income because a few hundred billionaires raise the mean a lot.

Exercise 20 Median Grade

Working Space

Find the median of your six grades: 87, 91, 98, 65, 87, 100.

Answer on Page 812

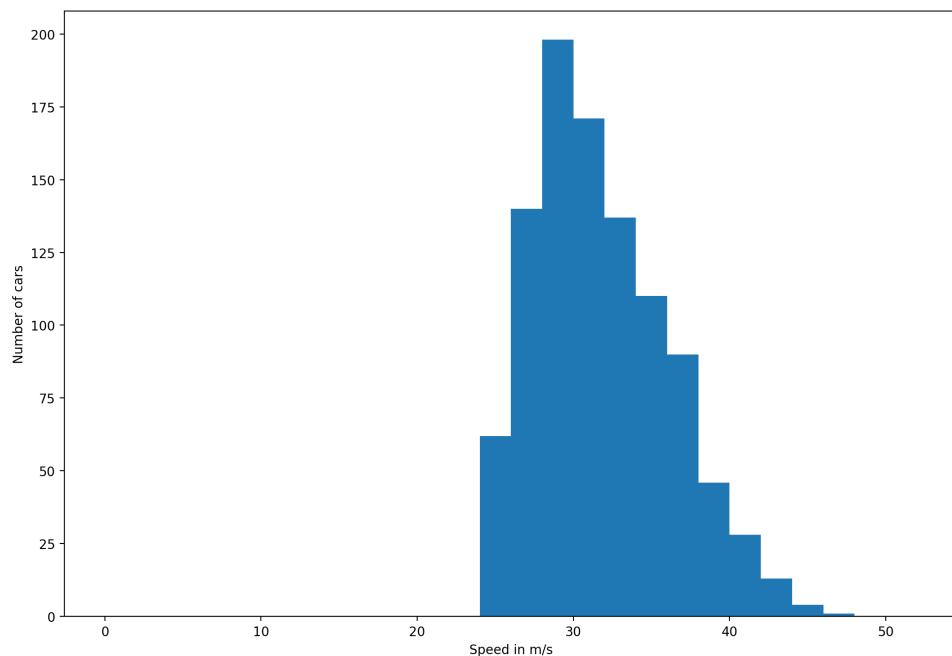
11.4 Histograms

A histogram is a bar chart that shows how many samples are in each group. In our example, we group cars by speed. Maybe we count the number of cars going between 30 and 32 m/s. And then we count the cars going between 32 and 34 m/s. And then we make a bar chart from that data.

Your 1000 cars would break up into these groups:

0 - 2 m/s	0 cars
2 - 4 m/s	0 cars
4 - 6 m/s	0 cars
...	...
20 - 22 m/s	0 cars
22 - 24 m/s	0 cars
24 - 26 m/s	65 cars
26 - 28 m/s	160 cars
28 - 30 m/s	175 cars
30 - 32 m/s	168 cars
32 - 34 m/s	150 cars
34 - 36 m/s	114 cars
36 - 38 m/s	79 cars
38 - 40 m/s	52 cars
40 - 42 m/s	20 cars
42 - 44 m/s	12 cars
44 - 46 m/s	4 cars
46 - 48 m/s	1 cars
48 - 50 m/s	0 cars

Now we make a bar chart from that:



Often a histogram will tell the story of the data. Here, you can see that no one is going

less than 24 m/s, but a lot of people travel at 30 m/s. There are a few people who travel over 40 m/s, but there are also a couple of people who drive a lot faster than anyone else.

11.5 Root-Mean-Squared

Scientists have a mean-like statistic that they love. It is named quadratic mean, but most just calls it Root-Mean-Squared or RMS.

Definition of RMS

If you have a list of numbers x_1, x_2, \dots, x_n , their RMS is

$$\sqrt{\frac{1}{n} (x_1^2 + x_2^2 + \dots + x_n^2)}$$

You are taking the square root of the mean of squares of the samples, thus the name Root-Mean-Squared.

Using your 12 samples:

x	x^2
30.462	927.933
29.550	873.203
29.227	854.218
37.661	1418.351
27.899	778.354
28.113	790.341
24.382	594.482
35.668	1272.206
43.797	1918.177
31.312	980.441
37.637	1416.544
30.891	954.254
Mean of x^2	1064.875
RMS	32.632

Why is RMS useful? Let's say that all cars had the same mass m , and you need to know what the average kinetic energy per car is. If you know the RMS of the speeds of the cars is v_{rms} , the average kinetic energy for each car is

$$k = \frac{1}{2}mv_{rms}^2$$

(You don't believe me? Let's prove it. Substitute in the RMS:

$$k = \frac{1}{2}m\sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \dots + x_n^2)}^2$$

The square root and the square cancel each other out:

$$k = \frac{1}{2}m\frac{1}{n}(x_1^2 + x_2^2 + \dots + x_n^2)$$

Use the distributive property:

$$k = \frac{1}{n}\left(\frac{1}{2}mx_1^2 + \frac{1}{2}mx_2^2 + \dots + \frac{1}{2}mx_n^2\right)$$

That is all the kinetic energy divided by the number of cars, which is the mean kinetic energy per car. Quod erat demonstrandum! (That is a Latin phrase that means "which is what I was trying to demonstrate". You will sometimes see "QED" at the end of a long mathematic proof.))

Now you are ready for the punchline: kinetic energy and heat are the same thing. Instead of cars, heat is the kinetic energy of molecules moving around. More on this soon.

Video: Mean, Median, Mode: <https://www.youtube.com/watch?v=5C9LBF3b65s>



CHAPTER 12

Basic Statistics in Spreadsheets

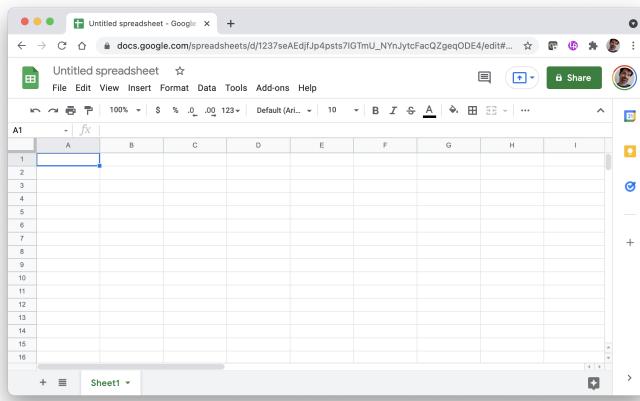
When you completed the problems in the last section, you probably noticed how long it took to compute statistics like the mean, the median, and variance by hand. Luckily, computers were designed to free us from these sorts of tedious tasks. The most basic tool for automating calculations is the spreadsheet program.

There are lots of spreadsheet programs including Microsoft's Excel and Apple's Numbers. Any spreadsheet program will work; they are all very similar. The instructions and screenshots here will be from Google Sheets – a free spreadsheet program you use through your web browser.

12.1 Your First Spreadsheet

In whatever spreadsheet program you are using, create a new spreadsheet document.

A spreadsheet is essentially a grid of cells. In each cell you can put data (like numbers or text) and formulas.



Let's put some labels in the column:

- Select the first cell (A1) and type “A number”.
- Select the cell below it (A2) and type “Another number”.
- Select the cell below that one (A3) and type “Their product”.
- In the next column, type the number 5 in B1 and 7 in B2.

It should look like this:

	A	B
1	A number	5
2	Another number	7
3	Their product	
4		

Now put a formula in cell B3. Select B3, and type “= B1 + B2”. The spreadsheet knows this is a formula because it starts with '='. It will look like this as you type:

	A	B
1	A number	5
2	Another number	7
3	Their product	= B1 * B2
4		

When you press Return or Tab, the spreadsheet will remember the formula, but display its value:

	A	B	
1	A number	5	
2	Another number	7	
3	Their product	35	
4			
5			

If you change the values of cell B1 or B2, the cell B3 will automatically be recalculated. Try it.

12.2 Formatting

Every spreadsheet lets you change the formatting of your columns and cells. They are all a little different, so play with your spreadsheet a little now. Try to do the following:

- Set the background of the first column to light gray.
- Right-justify the text in the first column.
- Make the text in the first column bold.
- Make the numbers in the second column have one digit after the decimal point.

It should look something like this:

	A	B	
1	A number	5.0	
2	Another number	7.0	
3	Their product	35.0	
4			
5			

That's a spreadsheet. You have a grid of cells. Each cell can hold a value or a formula that uses values from other cells. The cells with formulas automatically update as you edit the values in the other cells.

12.3 Comma-Separated Values

A lot of data is exchanged in a file format called *Comma-Separated Values* or just CSV. Each CSV file holds one table of data. It is a text file, and each line of text corresponds to one row of data in the table. The data in each column is separated by a comma. The first line of a CSV is usually the names of the columns. A CSV might look like this:

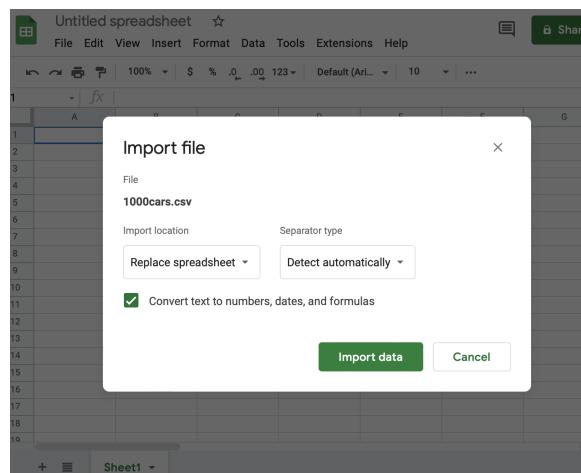
```
studentID,firstName,lastName,height,weight
1,Marvin,Sumner,260,45.3
2,Lucy,Harris,242,42.2
3,James,Boyd,261,44.2
```

In your digital resources for this module, you should have a file called `1000cars.csv`. It is a CSV with only one column called “speed”. The first few lines look like this:

```
speed
33.8000
29.9920
34.8699
27.9936
```

There is a title line and 1000 data lines.

Import this CSV into your spreadsheet program. In Google Sheets, it looks like this:



You should see a long, long column of data appear. (Mine goes from cell A2 through A1001.)

	A	B	C
1	speed		
2	33.8		
3	29.992		
4	34.8699		
5	27.9936		
6	26.2875		
7	31.6701		
8	27.3347		

12.4 Statistics in Spreadsheets

Let's take the mean of all 1000 numbers. In cell B2, type in a label: "Mean". (Feel free to format your labels as you wish. Bolding is recommended.)

In cell C2, enter the formula “=AVERAGE(A2:A1001)”. When you press return, the cell will show the mean: 31.70441, if done correctly.

	A	B	C
1	speed		
2	33.8	Mean	31.7044106
3	29.992		
4	34.8699		
5	27.9936		

Notice that by specifying that the function AVERAGE was to be performed on a range of cells: cells A2 through A1001.

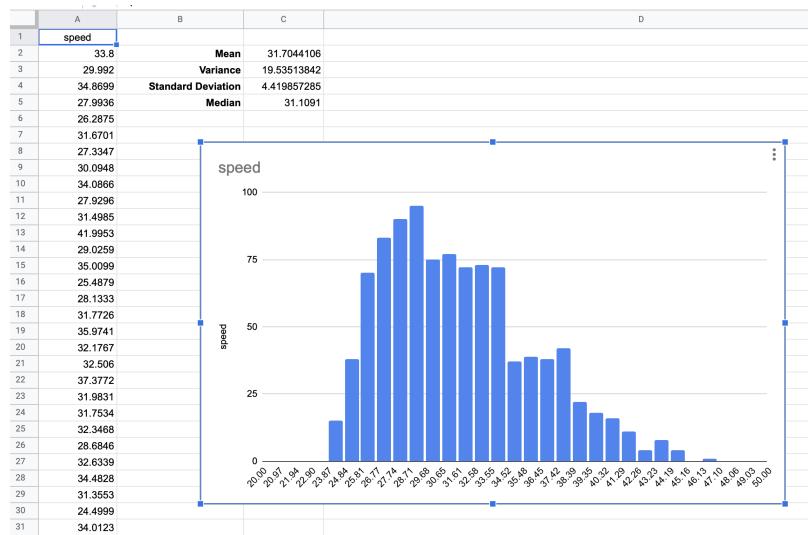
Do the calculations for variance, standard deviation, and median.

- The function for variance is VAR.
- The function for standard deviation is STDEV.
- The function for median is MEDIAN.

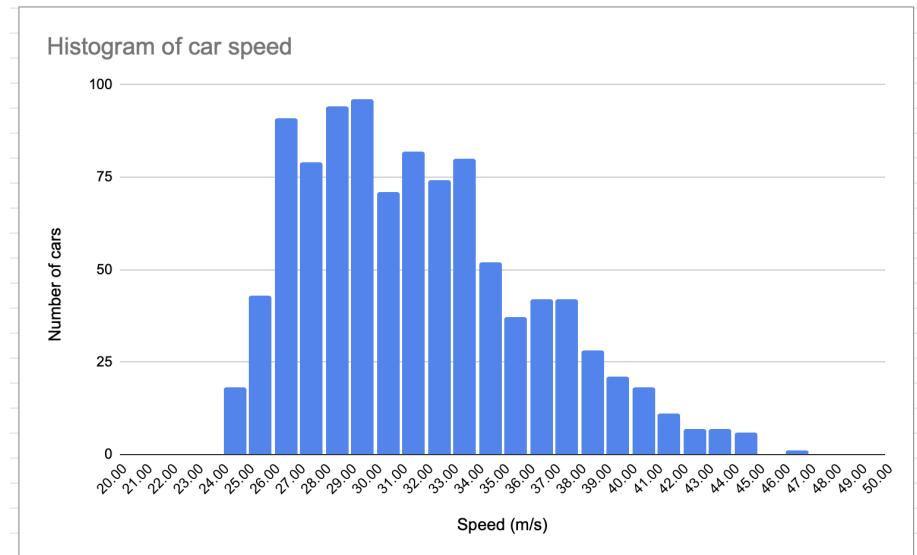
	A	B	C
1	speed		
2	33.8	Mean	31.7044106
3	29.992	Variance	19.53513842
4	34.8699	Standard Deviation	4.419857285
5	27.9936	Median	31.1091
6	26.2875		
7	31.6701		

12.5 Histogram

Most spreadsheets have the ability to create a histogram. In Google Sheets, you select the entire range A2:A1001 by selecting the first cell and then shift-clicking the last. Then you choose Insert→Chart. In the inspector, change the type of the chart to a histogram. This will get you a basic histogram.



Play with the formatting to see how unique you can make data. Here is an example:



Exercise 21 RMS

In your spreadsheet, calculate the quadratic mean (the root-mean-squared) of the speeds. You will need the following three functions:

- SUMSQ returns the sum of the squares of a range of cells.
- COUNT returns the number of cells in a range that contains numbers.
- SQRT returns the square root of a number.

Working Space

Answer on Page 813

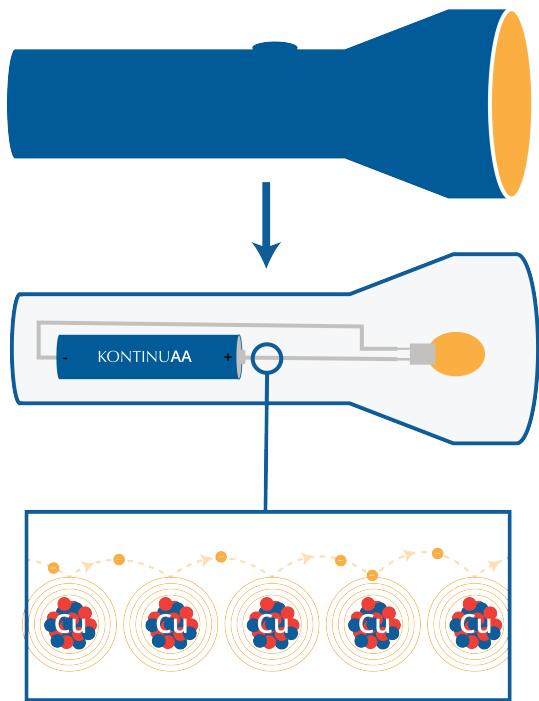


CHAPTER 13

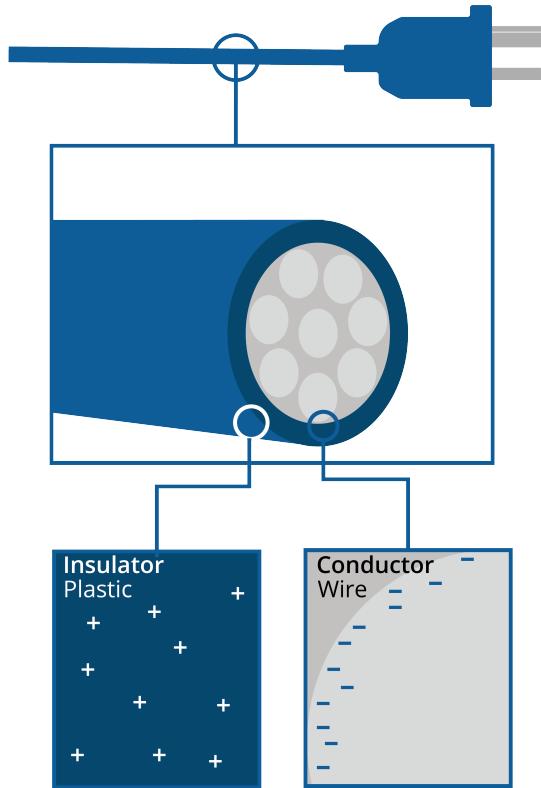
Introduction to Electricity

What happens when you turn on a flashlight? The battery in the flashlight acts as an electron pump. The electrons flow through the wires to the lightbulb (or LED). As the electrons pass through the lightbulb, they excite the molecules within, which gives off light and heat. (LEDs also give off light and heat, but they give off a lot less heat.) Then the electrons return to the battery to be pumped around again.

When electricity is flowing through a copper wire, the protons and neutrons of the copper stay put while the electrons jump between the atoms on their way from the battery to the lightbulb and back again.



In some materials, like copper and iron, electrons are loosely bound to their nuclei, forming a sea of electrons, which allows energy to flow. These are good *electrical conductors*. In other materials, like glass and plastic, electrons don't leave their nuclei easily. Thus, they are terrible electrical conductors – we call them *electrical insulators*. For example, the plastic around a wire is electrical insulation.



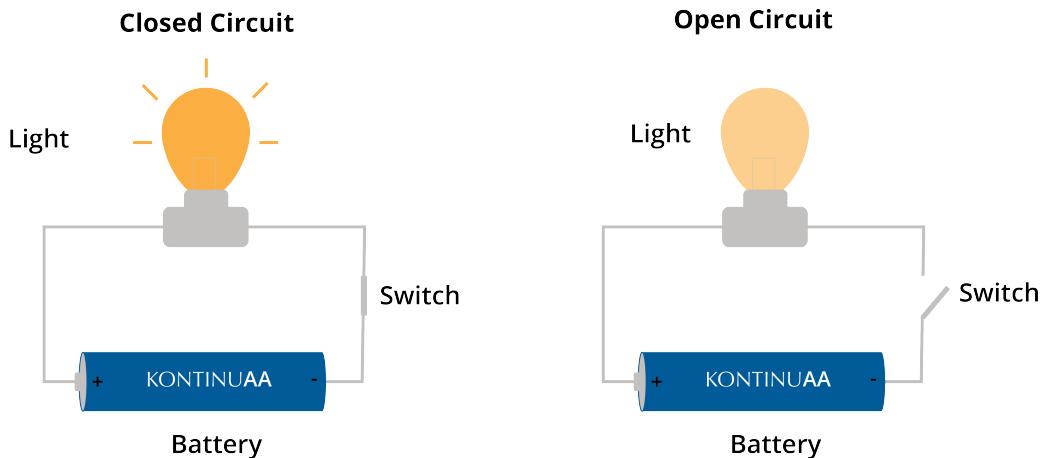
13.1 Units

Electrons are very small, so to study them, scientists came up with a unit that represents *a lot* of electrons. 1 *coulomb* is about 6,241,509,074,460,762,608 electrons. When 5 coulombs enter one end of the wire every second (and simultaneously 5 coulombs exit the other end), we say “This wire is carrying 5 amperes of current.”

(Truthfully, we usually shorten ampere to just “amp”. This is sometimes a little awkward because we often shorten the word “amplifier” to “amp”. You should be able to tell which is which from the context.)

If you look at the circuit breakers or fuses for your home’s electrical system, you’ll see that each one is rated in amps. For example, maybe the circuit that supplies power to your kitchen has a 10 amp circuit breaker. If for some reason, more than 10 amps tries to pass through that wire, the circuit breaker will turn off the whole circuit.

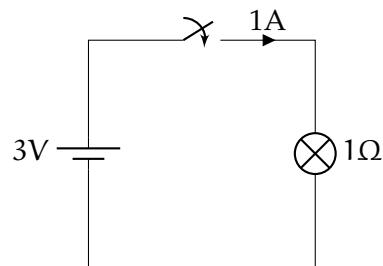
When it is on, your flashlight pushes about 1 amp of current through the lightbulb(When it is off, there is no current in the lightbulb).



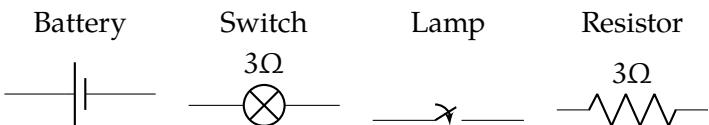
The lightbulb creates *Resistance* that the current pushes through. Think of it like plumbing: The current is the amount of water passing through a pipe. The resistance is something that tries to stop the current – like a ball of hair. The battery is what allows the current to push through the resistance; we call that pressure *voltage*.

13.2 Circuit Diagrams

Here is a circuit diagram of your flashlight:



The lines are wires. The symbols that we will use:



The battery pushes the electrons from one end and pulls them back in at the other, so the circuit must go around in a circle for the current to flow. This is why the current stops

flowing when the switch breaks the circuit.

You can think of a switch as having zero resistance when it is closed and infinite resistance when it is open.

For our purposes, a lamp is just a resistor that gives off light.

13.3 Ohm's Law

Resistance is measured in *ohms*, and we use a Greek capital omega for that: Ω

Voltage is measured in *volts*.

Ohm's Law

Whenever a voltage V is pushing a current I through a resistance of R , the following is true:

$$V = IR$$

where V is in volts, I is in amps, and R is in ohms.

13.4 Power and Watts

Joule's Law

When a current I is passing through a resistance R , the power consumed is

$$W = I^2R$$

where W is in watts, I is in amps, and R is in ohms.

Of course $V = IR$, so we can extend this to:

$$W = I^2R = IV = \frac{V^2}{R}$$

Your flashlight's batteries provide about 3 volts. How much battery power is the flashlight using when it is on? The power (in watts) produced by the battery is the product of the voltage (in volts) and the current (in amps). So your flashlight is giving off $3\text{volts} \times 1\text{amp} =$

3watts of power. Some of that power is given off as light, some as heat.

A watt is 1 joule of energy per second. We say that a watt is a measure of *power*.

When we talk about how much energy is stored in a battery, we use a unit like a kilowatt-hour. A kilowatt-hour is equivalent to 3.6 million joules.

13.5 Another great use of RMS

In many electrical problems, the voltage fluctuates a lot. For example, the fluctuations in voltage makes the sound that comes out of an audio speaker.

You can use the root-mean-squared of the voltage to figure out the average power your speaker is consuming.

Let's say that the RMS of the voltage you are sending to the speaker is V_{rms} and the resistance of the speaker is R ohms, then the power consumed by the speaker is:

$$P = \frac{V_{rms}^2}{R}$$

Similarly, if you know the RMS of the current you are pushing through the speaker is I_{rms} , then the power consumed by the speaker is:

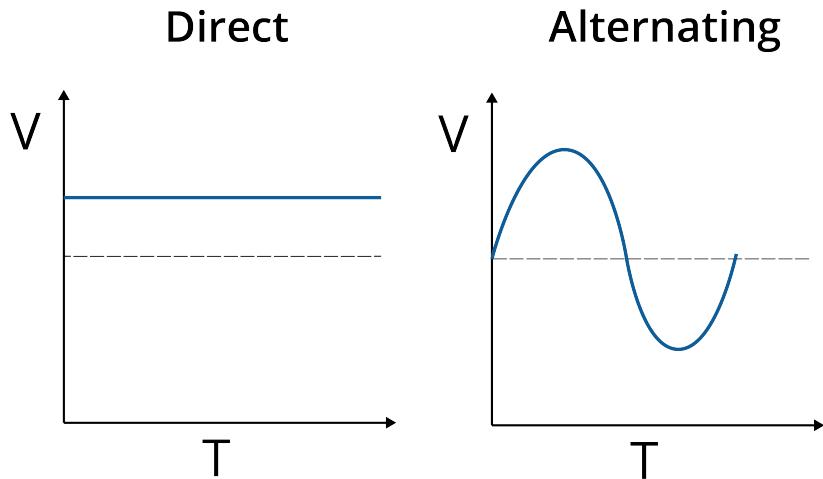
$$P = I_{rms} R$$

13.6 Electricity Dangers

Large amounts of electricity moving through your body can hurt or even kill you. You must be careful around electricity.

However, your body is not a very good conductor, so low-voltage systems (like a flashlight) don't have enough voltage to move significant amounts of current through your body.

However, the electricity in a power outlet has much more voltage. The voltage in these outlets is fluctuating between positive and negative, so we call it *Alternating Current* or AC.



In most countries, the RMS of the voltage between 110 and 240 V. (The peak voltage is always $\sqrt{2}$ times the RMS value. In the US, for example, people say “Our outlets supply 120 V.” They mean that the RMS of the voltage difference between the wire and the earth is 120V. The peak voltage is almost 170V.)

How much current can a human handle? Not much. You can barely feel 1 mA moving through your body, but at 16 mA, your muscles will clench and you won’t be able to relax them – many people die from electrocution because they grab a wire which pushes enough current through their body to prevent them from letting go of the wire. At 20 mA, a human’s respiratory muscles become paralyzed.

The fuse breaker in a house will often allow 20 A to flow through the circuit before it shuts off the power: Always, always, always shut off the power before touching any of the wiring in your house.

While water is actually a mediocre conductor, it can still deliver enough current to kill you. If you see a wire in a puddle, you should not touch the puddle. Interestingly, because of the salt, sea water is more than 100 times better at conducting electricity than the water you drink.

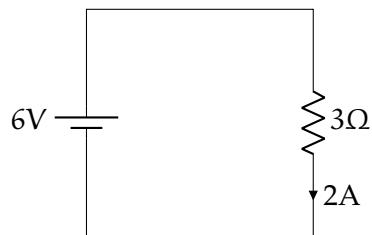
If you hold a wire in each hand, how many Ohms of resistance will your body have? Once it gets past your skin, you will look like a bag of salt water to the electricity. After the skin, your body will have a resistance of about 300Ω . However, the skin is a pretty good insulator. If you have dry, calloused hands, your skin may add a $100,000\Omega$ to the resistance.



CHAPTER 14

DC Circuit Analysis

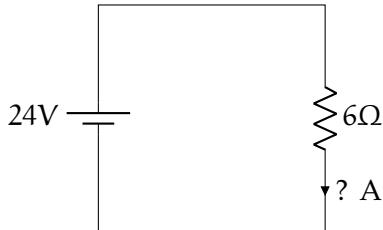
In the most basic circuit, you have only a battery and a resistor:



In this case, you only need Ohm's Law: $V = IR$. In this case, $6V = 3\Omega \times 2A$.

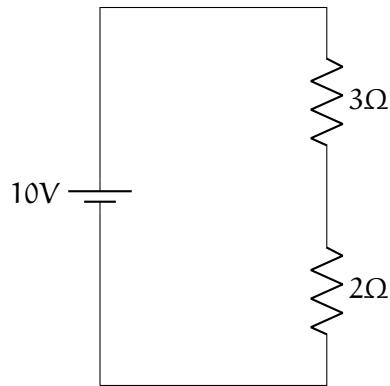
Exercise 22 Ohm's Law*Working Space*

How many amps are going around the circuit?

*Answer on Page 813***14.1 Resistors in Series**

When you have two resistors wired together in a long line, we say they are “in series”. If you have two resistors R_1 and R_2 wired in series, the total resistance is $R_1 + R_2$.

In this diagram, for example, the total resistance is 5Ω .



The current flowing through the circuit, then, is $10/4 = 2A$.

By Ohm's law, the voltage drop across the upper resistor is $IR = 2A \times 3\Omega = 6V$.

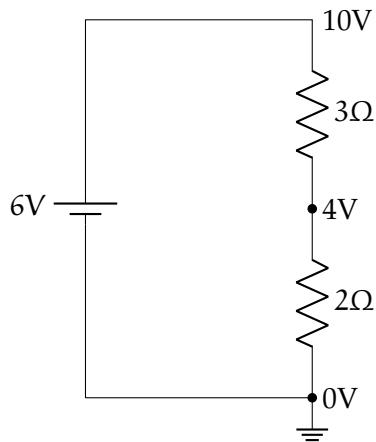
The voltage drop across the lower resistor is $IR = 2A \times 2\Omega = 4V$.

Notice that the battery pumps the voltage up to 10V, then the two resistors drop it by exactly 10V. This is known as "Kirchhoff's Voltage Law":

Kirchhoff's Voltage Law

As you make a loop around a circuit, the sum of the voltage increase must equal the sum of the voltage decrease.

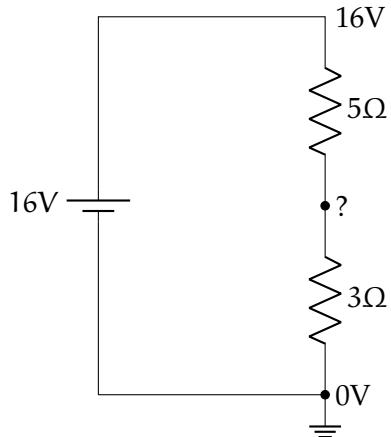
The negative end of the battery is connected to "ground" (it has zero voltage), then we can draw a diagram with the voltages (That symbol in the lower right represents a connection to ground).



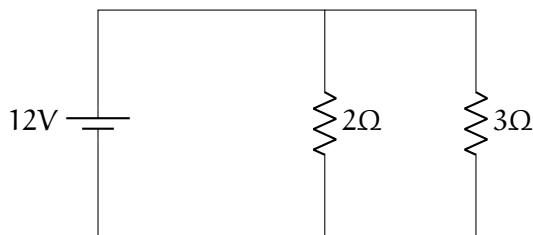
Exercise 23 Resistors In Series*Working Space*

What is the current going around the circuit?

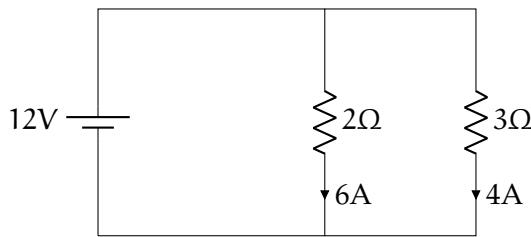
What is the voltage drop across each resistor?

*Answer on Page 813***14.2 Resistors in Parallel**

Look at this circuit. Note that the current can go two different paths.



There is 12 volts pushing current through both resistors. So 6A will go through the 2Ω resistor and 4A will go through the 3Ω resistor.



Thus, a total of 10 A will be going through the battery.

Imagine you are a battery. You can't see that you have two resistors. What does it feel like to you? $\frac{V}{I} = R$, and $V = 12$ and $I = 10$. So the effective resistance of the two resistors in parallel is $\frac{12}{10}$ or $\frac{6}{5}\Omega$.

Resistance in Parallel

If you have several resistances R_1, R_2, \dots, R_n wired in parallel, their effective resistance R_t is given by

$$\frac{1}{R_t} = \frac{1}{R_1} + \frac{1}{R_2} + \dots + \frac{1}{R_n}$$

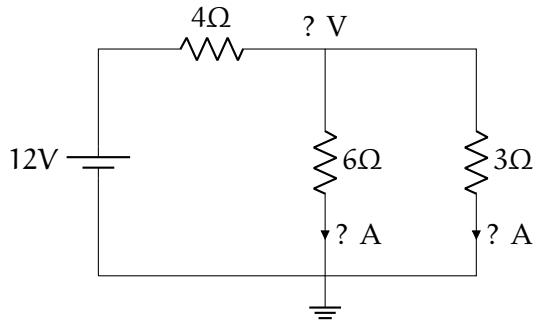
In our example:

$$\frac{1}{R_t} = \frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

Thus $R_t = \frac{6}{5}\Omega$.

Exercise 24 Resistors In Parallel*Working Space*

What is the current going through the battery? What is the drop over the 4Ω resistor? What is the current in each branch?

*Answer on Page 813*



CHAPTER 15

Charge

If you rub a balloon against your hair and then place it next to a wall it will stick. We say that it has gotten an *electrical charge*. It stole some electrons from your hair, and now the balloon has slightly more electrons than protons. We say that it has a negative electrical charge.

Objects with slightly more protons than electrons have a positive charge.

This charge is measured in coulombs. The charge of a single proton is about 1.6×10^{-19} coulombs.

An object with a negative charge and an object with a positive charge will be attracted to each other. Two objects with the same charge will be repelled by each other.

Coulomb's Law

If two objects with charge q_1 and q_2 (in coulombs) are r meters from each other, the force of attraction or repulsion is given by

$$F = K \frac{|q_1 q_2|}{r^2}$$

where F is in newtons and K is Coulomb's constant: about 8.988×10^9 .

Exercise 25 Coulomb's Law

Working Space

Two balloons are charged with an identical quantity and type of charge: -5×10^{-9} coulombs. They are held apart at a separation distance of 12 cm. Determine the magnitude of the electrical force of repulsion between them.

Answer on Page 814

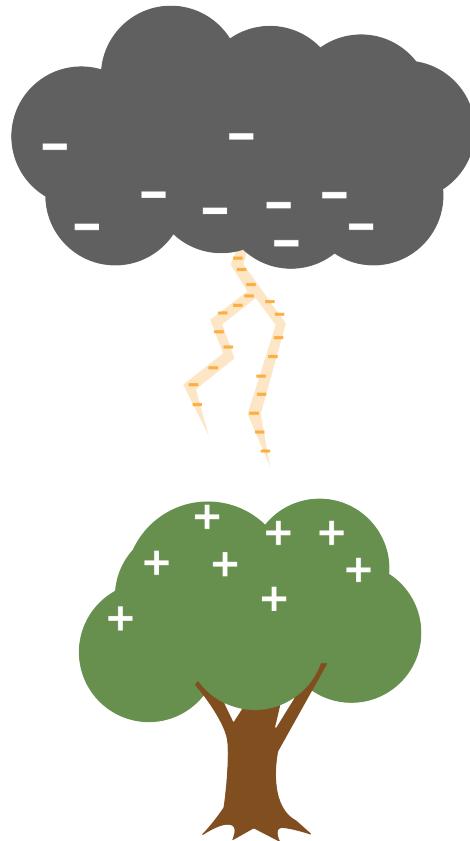
At this point, you might ask "If the wall has zero charge, why is the balloon attracted to it?" The answer: the electrons in the wall move away from the balloon. The negative charge on the balloon pushes electrons into the wall, so the surface of the wall gets a mild positive charge. The surface is close to the balloon, so the attraction is stronger than the repulsion.



15.1 Lightning

A cloud is a cluster of water droplets and ice particles. These droplets and ice particles are always moving up and down through the cloud. In this process, electrons get stripped off and end up on the water droplets at the bottom of the cloud (water droplets collect at the bottom because they are denser). The air between the droplets is a pretty good insulator,

and thus the electrons are reluctant to jump anywhere. However, eventually, the charge gets so strong that even the insulating properties of the air is not enough to prevent the jump, causing lightning.



A lot of lightning moves within a cloud or between clouds. However, a few jump to the earth. These bolts of lightning vary in the amount of electrons they carry, but the average is about 15 coulombs.

And thunder occurs because the electrons heat the air they pass through, causing the air to expand suddenly, and the resulting shockwave is the sound we know as thunder.

15.2 But...

This idea that opposite charges attract creates some heavy questions that you do not yet have the tools to work with. So the answer is basically “Don’t ask that question now!”

However, you probably have these questions, so I will point you in the direction of the answers.

The first is “In any atom bigger than hydrogen, there are multiple protons in the nucleus. Why don’t the protons push each other out of the nucleus?”

We aren’t ready to talk about it, but there is a force called *the nuclear force* which pulls the protons and neutrons in the nucleus of the atom toward each other. At very, very small distances it is strong enough to overpower the repulsive force due to the protons’ charges.

Another question is “Why do the electrons whiz around in a cloud so far from the nucleus of the atom? Negatively charged electrons should cling to the protons in the center, right?”

We aren’t ready to talk about it, but quantum mechanics tells us that electrons like to live in a certain specific energy level. Hugging protons isn’t one of those levels.



CHAPTER 16

Fertilizer

FIXME First, Allison has learned she does not need a colon after FM

Here are some thoughts on expanding the introduction to the Fertilizer Chapter. This might be a good moment to discuss the multidisciplinary nature of Kontinua. In a regular science class I'm guessing you wouldn't get electricity and Fertilizer in the same textbook. Why is there a chapter on Fertilizer here? Are you introducing us to the major ways science has given us more power and allowed population to grow? Can we discuss your thoughts on why problem solvers need a basic understanding of Fertilizer and I'll write up a new introduction to this chapter based on the discussion?

What do you think about adding a conclusion that talks about the connection between fertilizer (nitrogen), dynamite and the origin of the Nobel peace prize? I know you don't want too much history and philosophy but this seems like a great moment to add a little narrative spice.

Can we work POTATOES into this chapter!?!

Chapter text starts here:

One of the biggest problems humans face is: how can we get enough food to feed every-

one? In 1950, there were 2.5 billion people on the planet, and about 65% were malnourished. In 2019, there were 7.7 billion people on the planet, and only 15% are malnourished. How did crop yields increase so much? There were several factors: better crop varieties, reliable irrigation, increased mechanization, and affordable fertilizers.

When a plant grows, it takes molecules out of the soil and uses them to build proteins. It primarily needs the elements nitrogen (N), phosphorus (P), and potassium (K).

When you buy a bag of fertilizer at the store, it typically has three numbers on the front. For example, you might buy a bag of "24-22-4". This means that 24% of the mass of the bag is nitrogen, 22% is phosphorus, and 4% is potassium.

Potassium comes as potassium carbonate (K_2CO_3), potassium chloride (KCl), potassium sulfate (K_2SO_4), and potassium nitrate (KNO_3). Any blend of these chemicals is known as "potash". Potash is dug up out of mines.

Phosphorus is also mined, but is refined into phosphoric acid (H_3PO_4) before it is put into fertilizer.

Nitrogen is an especially interesting case for 2 reasons:

- Worldwide farmers apply more nitrogen to their soil than potassium or phosphorous combined.
- 78% of the air we breathe is nitrogen in the form of N_2 , but neither plants nor animals can utilize nitrogen in that form.

16.1 The Nitrogen Cycle

Converting the N_2 in the air into a form that a plant can use is known as *nitrogen fixation*. For billions of years, there were only two ways that nitrogen fixation occurred on earth:

- The energy from lightning causes N_2 and H_2O to reconfigure as ammonia (NH_3) and nitrate (NO_3). This accounts for about 10% of all naturally occurring nitrogen fixation.
- Cyanobacteria are responsible for the rest. They convert N_2 into ammonia.

Let's say that you are eating soybeans. There is a cyanobacteria called *rhizobia* that has a symbiotic relationship with soybean plants. Rhizobia fixes nitrogen for the soybean plant. The soybean plant performs photosynthesis and gives sugars to the rhizobia.

The proteins in the soybeans contain nitrogen from the rhizobia. When you eat them, you use some of the nitrogen to build new proteins. You probably don't use all the nitrogen, so your cells release ammonia into your blood.

Ammonia likes to react with things, so your liver combines the ammonia with carbon dioxide to make urea ($\text{CO}(\text{NH}_2)_2$). Your kidneys take the urea out of your blood and mix it with a bunch of water and salts in your bladder. When you urinate, the urea leaves your body.

If you urinate on the ground, the nearby plants can take the nitrogen out of the urea.

When you die, the nitrogen in your proteins will return to the soil as ammonia and nitrate.

For centuries, farms got their nitrogen from urine, feces, and rotting organic material. There were two challenges with this:

- Human pathogens had to be kept away from human food.
- There was simply not enough to support 7.7 billion people.

So we had to figure out how to do nitrogen fixation at an industrial level.

16.2 The Haber-Bosch Process

During World War I, two German scientists, Fritz Haber and Carl Bosch figured out how to make ammonia from N_2 and H_2 using high temperatures and pressures. This is how nearly all nitrogen fertilizer is created today.

Where do we get the H_2 ? From methane (CH_4) in natural gas. Today, 3-5% of the world's natural gas production is consumed in the Haber-Bosch process.

The ammonia is converted into ammonium nitrate (NH_4NO_3) or urea before it is shipped to farms.

16.3 Other nutrients

Healthy plants require several other elements that are sometimes applied as fertilizer: calcium, magnesium, and sulfur.

Finally, tiny amounts of copper, iron, manganese, molybdenum, zinc, and boron are sometimes needed.



CHAPTER 17

Concrete

To make concrete, you mix cement with water and an aggregate (sand or rock). The cement is usually only about 10 to 15 percent of the mixture. The cement reacts with the water, and the resulting solid binds the aggregate together. In 2019, the world consumed 4.5 billion tons of cement.

Concrete is hard and durable. The mortar between the pyramids at Giza is concrete – it is now 5000 years old. Today we use concrete to build many structures including buildings, bridges, airport runways, and dams.

There are many kinds of cement, but the most common is Portland cement. It is made by heating limestone (calcium carbonate) with clay (for silicon) in a kiln. Two things come out of the kiln: Carbon dioxide and a hard substance called “clinker”. The clinker is ground up with some gypsum before it is sent to market.

The carbon dioxide is released into the atmosphere. Cement manufacture is responsible for about 8% of the world's CO₂ emissions; it is a major contributor to climate change.

Really hard concrete, like that used in a nuclear power plant, can support 3,000 kg per centimeter without being crushed. However, if you pull on two ends of a piece of concrete

it comes apart pretty easily. We say that concrete can handle a lot of *compressive stress*, but not much *tensile* stress.

17.1 Steel reinforced concrete

Many places where we use concrete (like in a bridge), we need both compressive and tensile stress. Often the top of a beam is undergoing compression and the bottom of the beam is undergoing tension.

FIXME Picture here

Steel has tremendous tensile strength, but not as much compressive strength as concrete. To get both tensile *and* compressive strength, we often bury steel bars or cables inside the concrete. This is known as *steel-reinforced concrete*. The concrete generally does a very good job protecting the steel, which keeps it from rusting.

You may have heard of *rebar*. That is just short for “reinforcing bar”. Typically rebar has bumps and ridges that keep the bar and the concrete from moving independently.

17.2 Recycling concrete

A lot of concrete structures only last about 100 years. When they are demolished, the concrete can be reused as aggregate in other projects. Often the concrete bits are mixed with cement and made into concrete again.

If the concrete to be reused is reinforced with steel, the steel has to be removed and recycled separately. Then the concrete is crushed into small pieces.



CHAPTER 18

Metals

Elements that transmit electricity well, even at low temperatures, are called *metals*. Here are some metals that you are probably familiar with: aluminum, iron, copper, tin, gold, silver, and platinum. Aluminum and iron are particularly common; together they make up about 14% of the earth's crust.

An *alloy* is a mixture of elements that includes at least one metal. Brass, for example, is an alloy of copper and zinc. Bronze is an alloy of copper and tin.

18.1 Steel

One of the most common alloys is steel, an alloy of iron and carbon. In pure iron, the molecules slip easily past each other, so pure iron is relatively soft and easily deformed. The carbon in steel prevents that slipping, thus steel is much, much harder than iron.

How much carbon? If you put less than 0.002% by weight, you end up with something very much like pure iron. As you increase the carbon, it gets harder and harder. Once it gets above about 2%, the result is very brittle.

If you add about 11% chromium to steel, you get *stainless steel* which resists rusting.

Exercise 26 Tensile Strength

Working Space

The tensile strength of steel is usually between 400 MPa and 1200 MPa. A Mega Pascal (MPa) is the strength necessary to hold 1,000,000 newtons of force with a cable that has a 1 square meter cross section. Or, equivalently, to hold 1 newton of force with a cable that has a 1 square millimeter cross section.

If you have are buying a round cable that has a tensile strength of 700 Mpa and must hold a 100 kg man aloft, what the diameter of the smallest cable you can use?

Answer on Page 814

Here are some approximate tensile strengths of other materials:

Material	Tensile strength (MPa)
Iron	3
Concrete	4
Rubber	16
Glass	33
Wood	40
Nylon	100
Human hair	200
Aluminum	300
Steel	700
Spider webs	1000
Carbon fiber	4000

18.2 What metal for what task?

You will see copper used a lot for electrical wires in your house and appliances because it is very efficient at moving electricity (very little power is lost as heat). It is also very good

a transmitting heat, so you will often see copper pots and pans.

Aluminum is less dense than copper, and is still a pretty good conductor of electricity. Thus, the overhead wires in a power system are often made of aluminum.

Aluminum is not as strong as steel, but considerably lighter. It is often used structurally where weight is a concern: skyscrapers, cars, airplanes, and ships.

Titanium is about as strong as steel, but it weights about half as much. Titanium is very difficult to work with, so it is used in places where weight and strength are very important and cost is not: airplanes and bicycles.

(Carbon fiber, which is light, strong, and very easy to work with, is replacing aluminum and titanium in many applications. 20 years ago, many expensive bicycles were made of titanium. These days the vast majority are made with carbon fiber.)

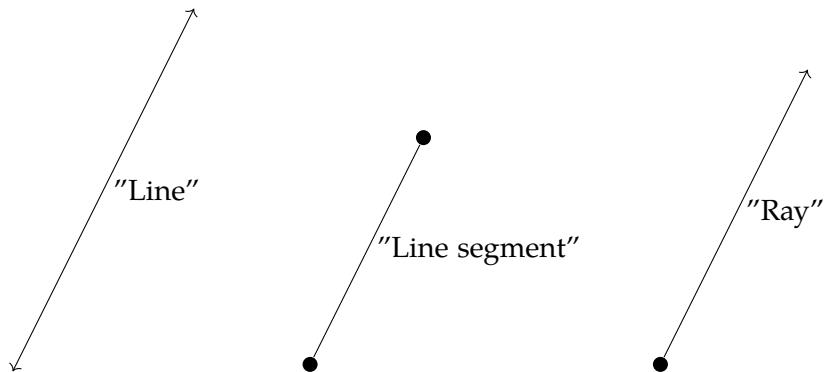
Zinc and tin are very resistant to corrosion, so they are often used as a coating to prevent steel from rusting. They are also used in many alloys for the same reason. In the United States, the penny is 97.5% zinc and only 2.5% copper.



CHAPTER 19

Angles

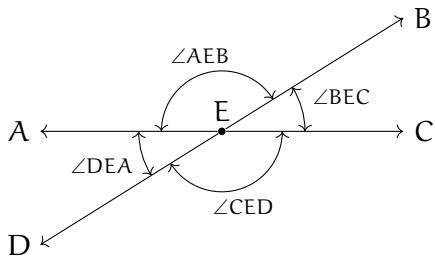
In the following recommend videos, the narrator talks about lines, line segments, and rays. When mathematicians talk about *lines*, they mean a straight line that goes forever in two directions. And if you pick any two points on that line; the space between them is a *line segment*. If you take any line, pick a point on that line and discard all the points on one side of the point, that is a *ray*. All three have no width.



Watch the following videos from Khan Academy:

- Introduction to angles: <https://youtu.be/H-de6Tkxej8>
- Measuring angles in degrees: <https://youtu.be/92aLiyeQj0w>

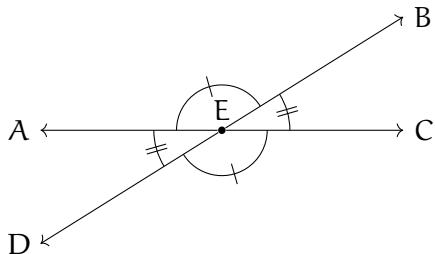
When two lines cross, they form four angles:



What do we know about those angles?

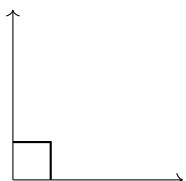
- The sum of any two adjacent angles add to be 180° . So, for example, $m\angle AEB + m\angle BEC = 180^\circ$. We use the phrase “add to be 180° ” so often that we have a special word for it: *supplementary*.
- The sum of all four angles is 360° .
- Angles opposite each other are equal. So, for example, $m\angle AEB = m\angle CED$.

In a diagram, to indicate that two angles are equal we often put hash marks in the angle:



Here the two angles with a single hash mark are equal and the two angles with double hash marks are equal.

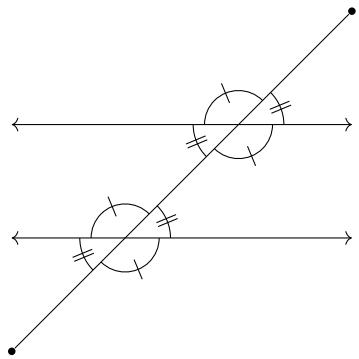
When two lines are perpendicular, the angle between them is 90° and we say they meet at a *right angle*. When drawing diagrams, we indicate right angles with an elbow:



When an angle is less than 90° , it is said to be *acute*. When an angle is more than 90° , it is said to be *obtuse*.



If two lines are parallel, line segments that intersect both lines, form the same angles with each line:

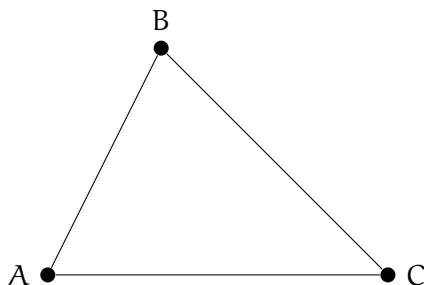




CHAPTER 20

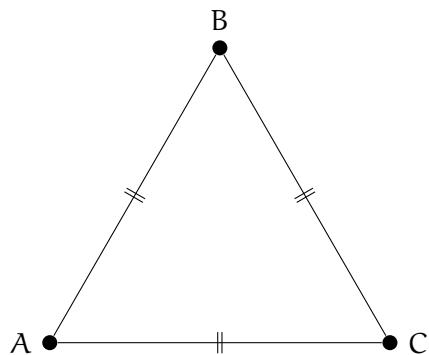
Introduction to Triangles

Connecting any three points with three line segments will get you a triangle. Here is the triangle ABC which was created by connecting three points A, B, and C:

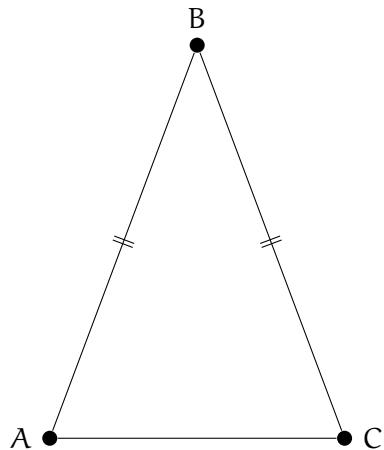


20.1 Equilateral and Isosceles Triangles

We talk a lot about the length of the sides of triangles. If all three sides of the triangle are the same length, we say it is an *equilateral triangle*:

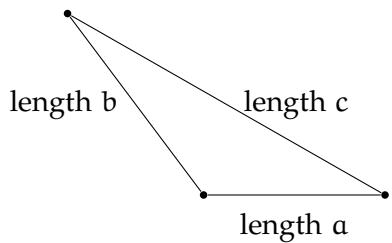


If only two sides of the triangle are the same length, we say it is an *isosceles triangle*:



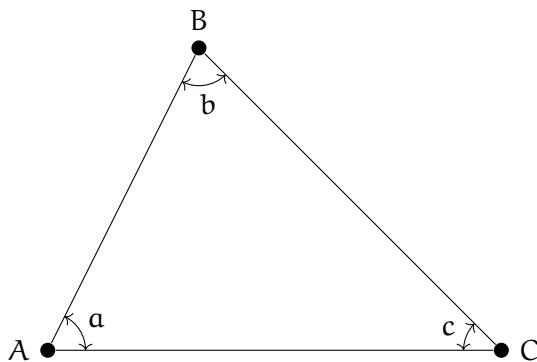
The shortest distance between two points is always the straight line between them. Thus, you can be certain that the length of one side will *always* be less than the sum of the lengths of the remaining two sides. This is known as the *triangle inequality*.

For example, in this diagram c must be less than $a + b$.

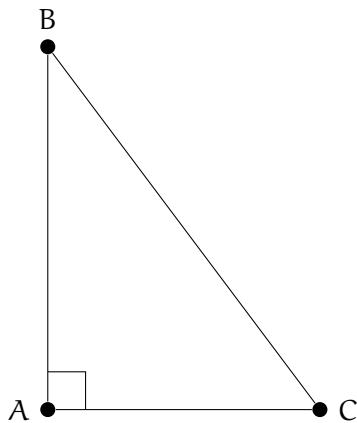


20.2 Interior Angles of a Triangle

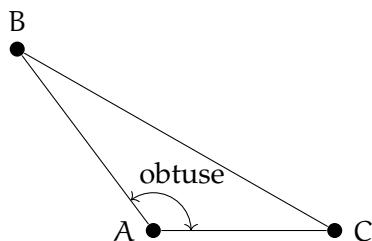
We also talk a lot about the interior angles of a triangle:



A triangle where one of the interior angles is a right angle is said to be a *right triangle*:



If a triangle has an obtuse interior angle, it is said to be an *obtuse triangle*:



If all three interior angles of a triangle are less than 90° , it is said to be an *acute triangle*.

The measures of the interior angles of a triangle always add up to 180° . For example, if we know that a triangle has interior angles of 37° and 56° , we know that the third interior angle is 87° .

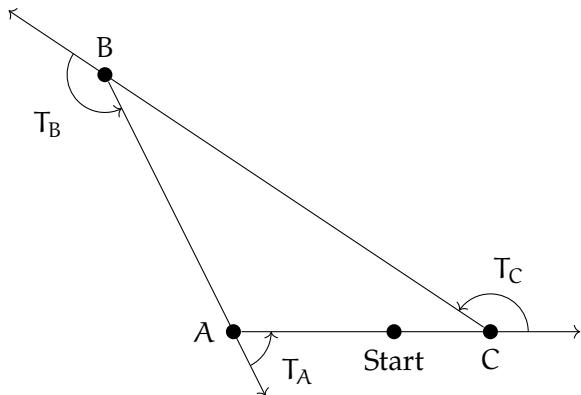
Exercise 27 Missing Angle

One interior angle of a triangle is 92° . The second angle is 42° . What is the measure of the third interior angle?

Working Space

Answer on Page 814

How can you know that the sum of the interior angles is 180° ? Imagine that you started on the edge of a triangle and walked all the way around to where you started. (going counter-clockwise.) You would turn three times to the left:



After these three turns, you would be facing the same direction that you started in. Thus $T_A + T_B + T_C = 360^\circ$. The measures of the interior angles are a , b , and c . Notice that a and T_A are supplementary. So we know that:

- $T_A = 180 - a$
- $T_B = 180 - b$
- $T_C = 180 - c$

So we can rewrite the equation above as

$$(180 - a) + (180 - b) + (180 - c) = 360^\circ$$

Which is equivalent to

$$a + b + c = 360^\circ$$

Exercise 28 Interior Angles of a Quadrilateral

Any four-sided polygon is a *quadrilateral*. Using the same “walk around the edge” logic, what is the sum of the interior angles of any quadrilateral?

Working Space

Answer on Page 814

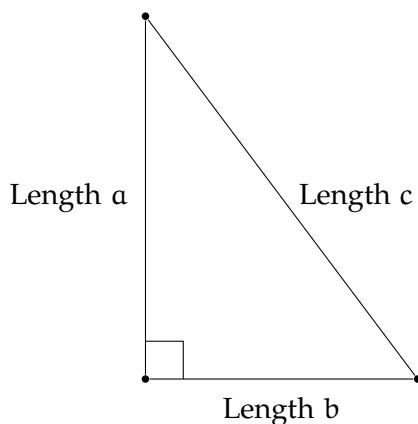


CHAPTER 21

Pythagorean Theorem

Watch Khan Academy's Intro to the Pythagorean Theorem video at <https://youtu.be/AA6RfgP-AHU>.

If you have a right triangle, the edges that touch the right angle are called *the legs*. The third edge, which is always the longest, is known as *the hypotenuse*. The Pythagorean Theorem gives us the relationship between the length of the legs and the length of the hypotenuse.



The Pythagorean Theorem tells us that $a^2 + b^2 = c^2$.

For example, if one leg has a length of 3 and the other has a length of 4, then $a^2 + b^2 = 3^2 + 4^2 = 25$. Thus c^2 must equal 25. So you know the hypotenuse must be of length 5.

(In reality, it rarely works out to be such a tidy number. For example, what is the length of the hypotenuse if the two legs are 3 and 6? $a^2 + b^2 = 3^2 + 6^2 = 45$. The length of the hypotenuse is the square root of that: $\sqrt{45} = \sqrt{9 \times 5} = 3\sqrt{5}$, which is approximately 6.708203932499369.)

Exercise 29 Find the Missing Length

What is the missing measure?

Working Space

Leg 1 = 6, Leg 2 = (It should be a
8, Hypotenuse = ? whole number.)

(It should be a Leg 1 = 3, Leg 2 =
whole number.) 3, Hypotenuse = ?

Leg 1 = 5, Leg 2 = (It is an irrational
?, Hypotenuse = 13 number. Give the
(It should be a exact answer and
whole number.) then use a calcu-

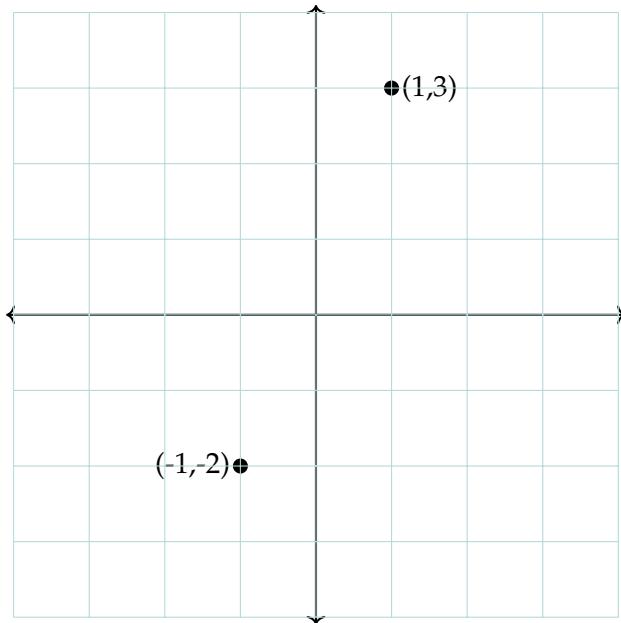
Leg 1 = ?, Leg 2 = lator to get an ap-
15, Hypotenuse = proximation.)

17

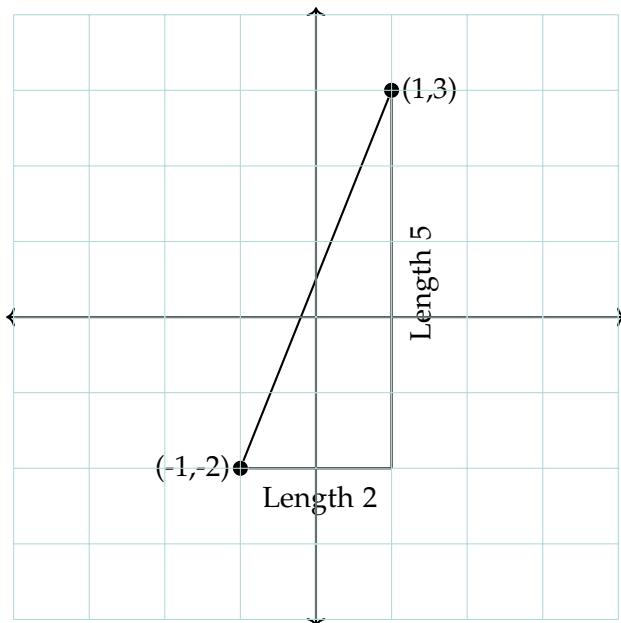
Answer on Page 814

21.1 Distance between Points

What is the distance between these two points?



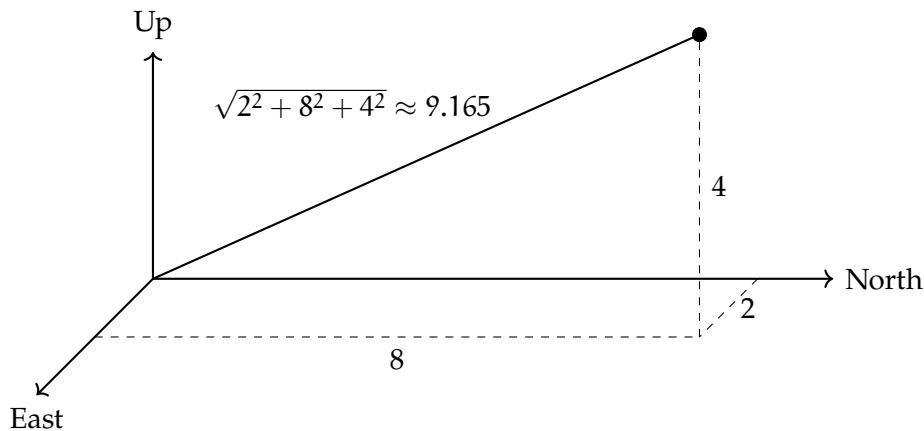
We can draw a right triangle and use the Pythagorean Theorem:



The distance between the two points is $\sqrt{2^2 + 5^2} = \sqrt{29} \approx 5.385165$. That is, you square the change in x and add it to the square of the change in y . The distance is the square root of that sum.

21.2 Distance in 3 Dimensions

What if the point is in three-dimensional space? That is, you move 2 meters East, 8 meters North, and 4 meters up in the air. How far are you from where you started? You just square each, sum them, and take the square root: $\sqrt{2^2 + 8^2 + 4^2} = \sqrt{84} = 2\sqrt{21} \approx 9.165$ meters.

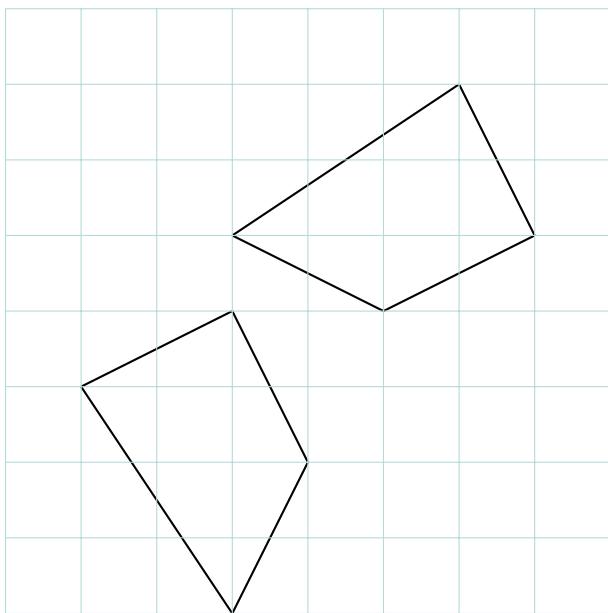




CHAPTER 22

Congruence

Look at this picture of two geometric figures.



They are the same shape, right? If you cut one out with scissors, it would lay perfectly on top of the other. In geometry, we say they are *congruent*.

What is the official definition of “congruent”? Two geometric figures are congruent if you can transform one into the other using only rigid transformations.

You might be wondering now, what are rigid transformations? A transformation is *Rigid* if it doesn’t change the distances between the points or the measure of the angles between the lines, they form. These are all rigid transformations:

- Translations
- Rotations
- Reflections

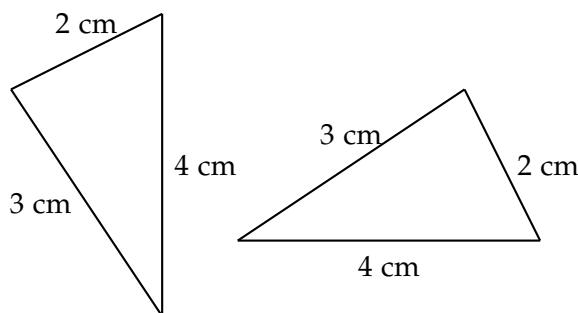
Once again imagine cutting out one figure with scissors and trying to match it with the second figure, your actions are rigid transformations:

- Translations - sliding the cutout left and right and up and down
- Rotations - rotating the cutout clockwise and counterclockwise
- Reflection - flipping the piece of paper over

A transformation is rigid if it is some combination of translations, rotations, and reflections.

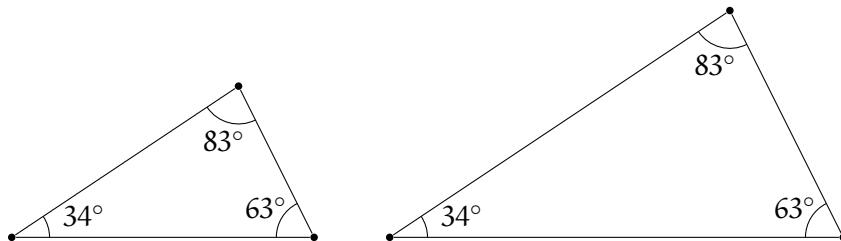
22.1 Triangle Congruency

If the sides of two triangles have the same length, the triangles must be congruent:



To be precise, the Side-Side-Side Congruency Test says that two triangles are congruent if three sides in one triangle are the same length as the corresponding sides in the other. We usually refer to this as the SSS test.

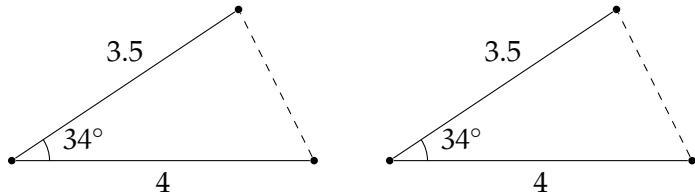
Note that two triangles with all three angles equal are not necessarily congruent. For example, here are two triangles with the same interior angles, but they are different sizes:



These triangles are not congruent, but they are *similar*. Meaning they have the same shape, but are not necessarily the same size.

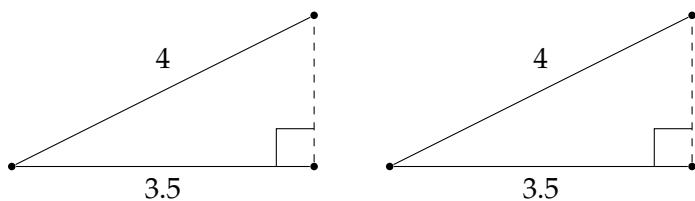
Therefore, if you know two angles of a triangle, you can calculate the third. So it makes sense to say "If two triangles have two angles that are equal, they are similar triangles." And if two similar triangles have one side that is equal in length, they must be the same size – so they are congruent. Thus, the Side-Angle-Angle Congruency Test says that two triangles are congruent if two angles and one side match.

What if you know that two triangles have two sides that are the same length and that the angle between them is also equal?



Yes, they must be congruent. This is the Side-Angle-Side Congruency Test.

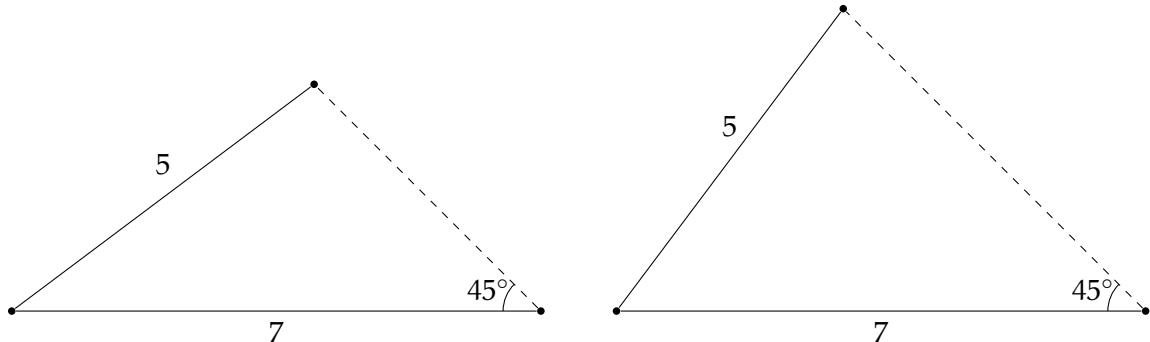
What if the angle isn't the one between the two known sides? If it is a right angle, you can be certain the two triangles are congruent. (How do I know? Because the Pythagorean Theorem tells us that we can calculate the length of the third side. There is only one possibility, thus all three sides must be the same length.)



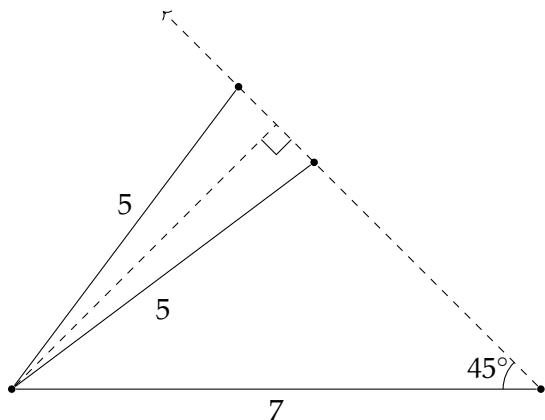
In this case, the third side of each triangle must be $\sqrt{4^2 - 3.5^2} \approx 1.9$.

What if the known angle is less than 90°? The triangles are not necessarily congruent. For

example, let's say that there are two triangles with sides of length 5 and 7 and that the corresponding angle (at the end of the side of length 7) on each is 45° . Two different triangles satisfy this:



Let's see this another way by laying one triangle on top of the other:



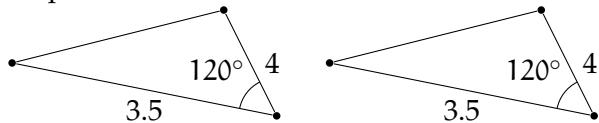
So there is *not* a general Side-Side-Angle Congruency Test.

Here, then, is the list of common congruency tests:

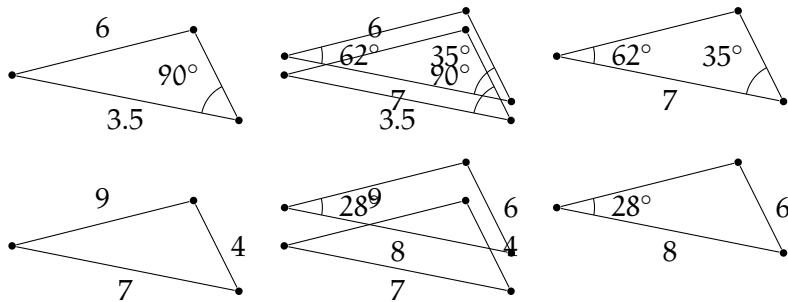
- Side-Side-Side: All three sides have the same measure
- Side-Angle-Angle: Two angles and one side have the same measure
- Side-Angle-Side: Two sides and the angle between them have the same measure
- Side-Side-Right: They are right triangles and have two sides have the same measure

Exercise 30 Congruent Triangles

Ted is terrible at drawing triangles: he always draws them exactly the same. Fortunately, he has marked these diagrams with the sides and angles that he measured. For each pair of triangles, write if you know them to be congruent and which congruency test proves it. For example:



(These drawings are clearly not accurate, but you are told the measurements are.)
The answer is "Congruent by the Side-Angle-Side test."



Working Space

Answer on Page 815



CHAPTER 23

Parallel and Perpendicular

Two vectors are said to be parallel if they have the same or opposite direction. In simpler terms, if two vectors are pointing in the same direction (even if their magnitudes differ), they are considered parallel. For example, imagine you have a vector representing the direction and speed of a car moving north. If you have another vector representing the direction and speed of a different car also moving north, these vectors are parallel.

On the other hand, if two vectors point in completely opposite directions, they are still considered parallel. For instance, if one vector represents a car moving north and the other represents a car moving south, these vectors are parallel but in opposite directions.

Perpendicular vectors, as the name suggests, are vectors that intersect each other at a right angle, forming a 90-degree angle. If we imagine a sheet of paper, drawing a horizontal vector and a vertical vector on that paper would create perpendicular vectors. In this case, the horizontal vector represents left-right direction, while the vertical vector represents up-down direction. Perpendicular vectors are often seen in geometric shapes, such as squares and rectangles, where their sides intersect at right angles.

A fundamental property of perpendicular vectors is that their dot product is zero. The dot product is a mathematical operation that measures the extent to which two vectors

align with each other. When two vectors are perpendicular, their dot product is always zero. This property provides a useful tool for determining whether two given vectors are perpendicular.

Understanding parallel and perpendicular vectors is essential in various areas of mathematics and physics. For example, in geometry, knowledge of perpendicular vectors helps us determine whether lines are perpendicular or parallel. In physics, vectors can represent forces, velocities, or displacements, and identifying parallel or perpendicular vectors aids in analyzing motion and forces acting on objects.

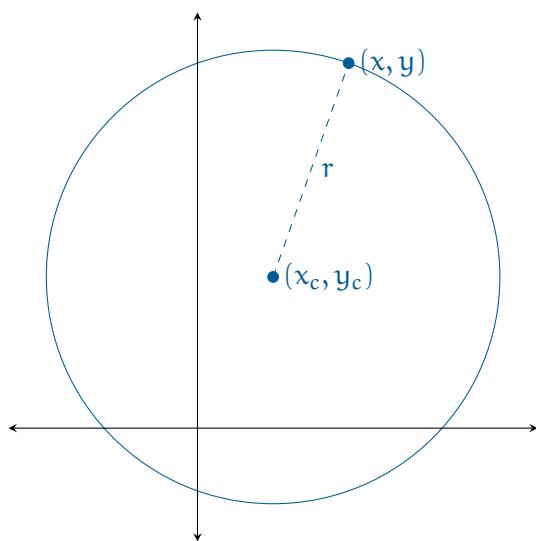
In summary, parallel vectors have the same or opposite direction, while perpendicular vectors intersect at a right angle. Recognizing these relationships between vectors enables us to solve problems involving geometry, physics, and many other fields. As you delve deeper into the exciting world of vectors, keep an eye out for parallel and perpendicular relationships, as they often hold valuable insights and solutions.



CHAPTER 24

Circles

A circle is the set of points (x, y) that are a particular distance r from a particular point (x_c, y_c) . We say that r is the *radius* and (x_c, y_c) is the *center*



Area and Radius

If the radius of a circle is r , the area of its interior (a) is given by

$$a = \pi r^2$$

Exercise 31 Area of a Circle**Working Space**

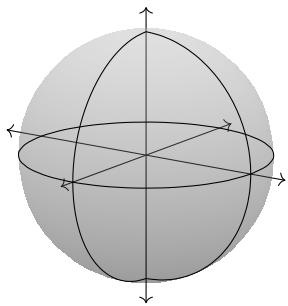
The paint you have says “One liter covers 6 square meters.”

You are painting the top of a circular table with a radius of 3 meters.

How much paint will you need?

Answer on Page 815

Note that a circle lives in a particular plane. The points (x, y, z) that are a particular distance r from a particular point (x_c, y_c, z_c) are a sphere:



The distance all the way across the middle of a circle (or a sphere) is its *diameter*. The diameter is always twice the radius.

For the rest of the chapter, we are talking about circles, points, and lines *in a plane*.

Circumference and Diameter

The circumference (c) of a circle is the distance around the circle. If the diameter is d ,

$$c = \pi d$$

Exercise 32 Circumference

Using a tape measure, you figure out that the circumference of a tree in your yard is 64 cm.

Assuming the trunk is basically circular, what is its diameter?

Working Space

Answer on Page 815

Exercise 33 Splitting a Pie

A pie has a radius of 13 cm. 7 friends all want equal sized wedges. You have a tape measure.

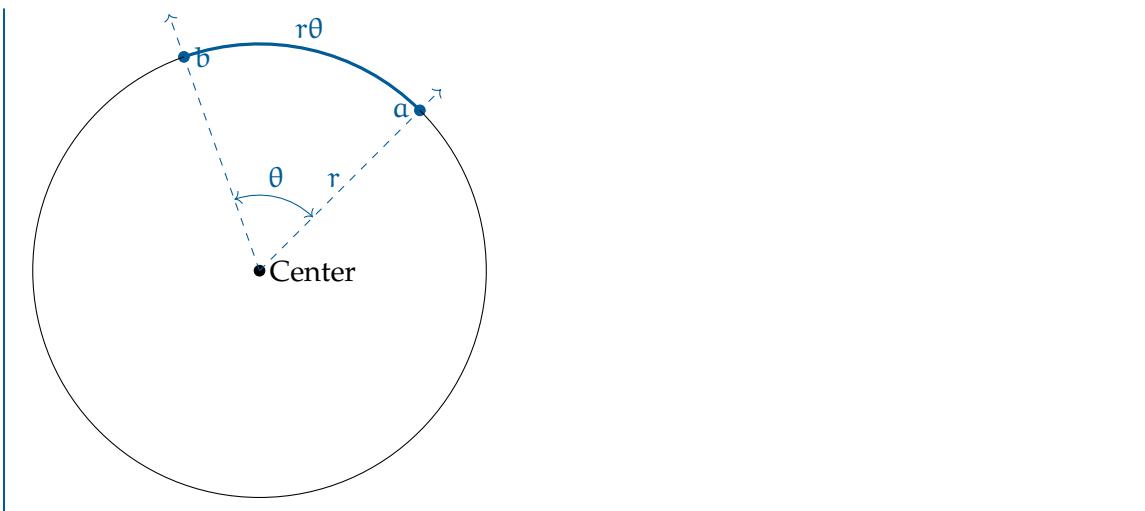
How many centimeters will each outer crust be?

Working Space

Answer on Page 815

Length of an Arc

If you have two points a and b on a circle, the ray from the center through a and the ray from the center through b form an angle. If θ is the angle in radians and r is the radius of the circle, the distance from a to b on the circle is $r\theta$.



Exercise 34 Arc Length

Working Space

You have been asked to find the radius of a very large cylindrical tank. You have a tape measure, but it is only 15 meters long and doesn't reach all the way around the tank.

However, you have a compass. So you stick one end of the tape measure to the side of the tank and measure the orientation of the wall at that point. Then you walk the 15 meters and measure the orientation of the wall there.

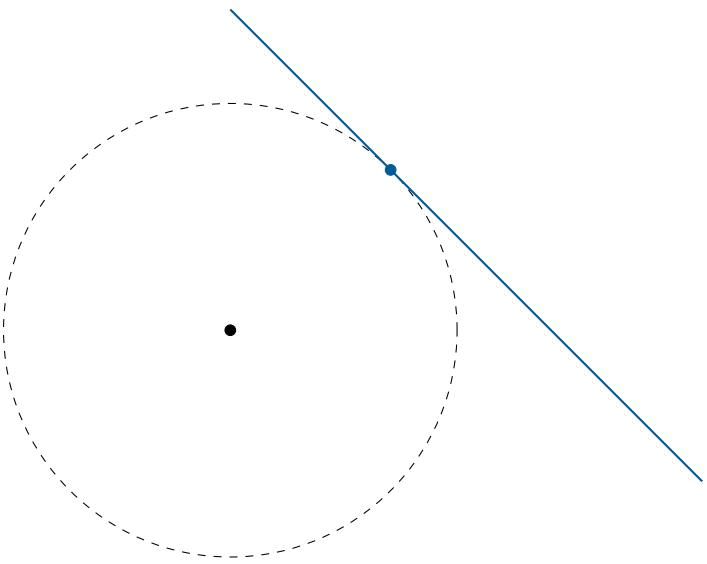
You find that 15 meters represents 72 degrees of arc.

What is the radius of the tank in meters?

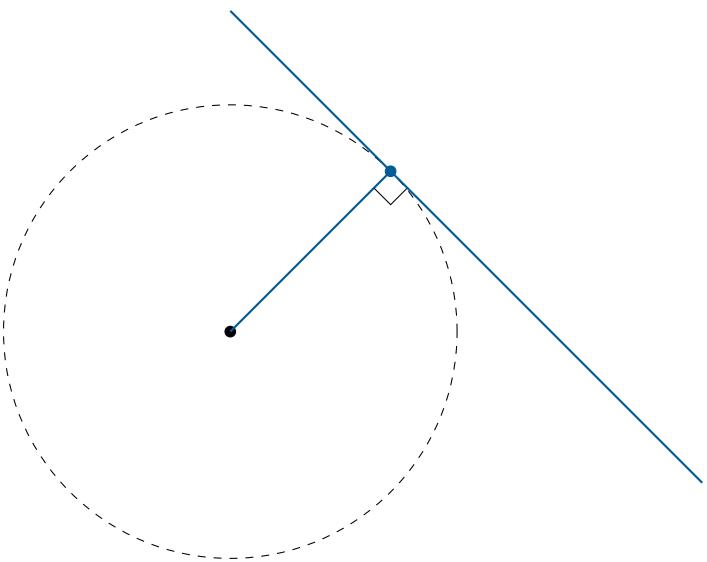
Answer on Page 816

24.1 Tangents

A line that is *tangent* to a circle touches it at exactly one point:



The tangent line is always perpendicular to the radius to the point of tangency:



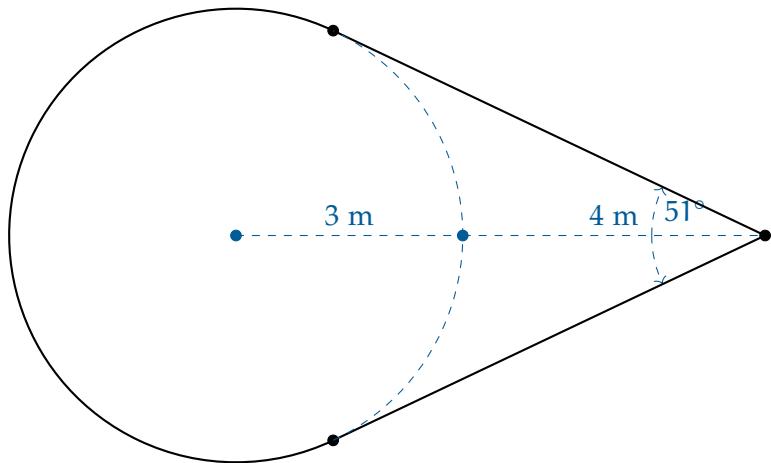
Exercise 35 Painting a Comet*Working Space*

You have been asked to paint a comet and its tail in yellow on the floor of a gymnasium.

A liter of yellow paint covers 6 square meters.

First you draw a circle with a radius of 3 meters. Then you mark a point D on the floor 7 meters from the center of the circle. Then you draw two tangent lines that pass through D.

You use a protractor to measure the angle at which the tangent lines meet: about 51° .



Before you paint the area contained by the circle and the two tangent lines, how much paint will you need?

Answer on Page 816



CHAPTER 25

Functions and Their Graphs

You can think of a function as a machine: you put something into the machine, it processes it, and out comes something else, a product. Just as we often use the variable x to stand in for a number, we often use the variable f to stand in for a function.

For example, we might ask, “Let the function f be defined like this:

$$f(x) = -5x^2 + 12x + 2$$

What is the value of $f(3)$?”

You would run the number 3 through “the machine”: $-5(3^2) + 12(3) + 2 = -7$. The answer would be “ $f(3)$ is 7”.

However, Some functions are not defined for every possible input. For example:

$$f(x) = \frac{1}{x}$$

This is defined for any x except 0, because you can't divide 1 by 0. The set of values that a function can process is called its *domain*.

Exercise 36 Domain of a function

Working Space

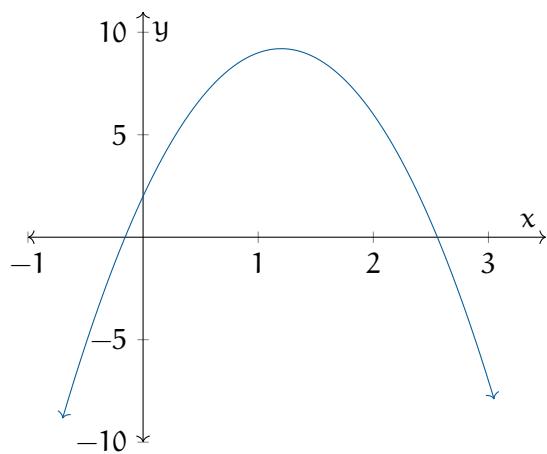
Let the function f be given by $f(x) = \sqrt{x - 3}$. What is its domain?

Answer on Page 818

25.1 Graphs of Functions

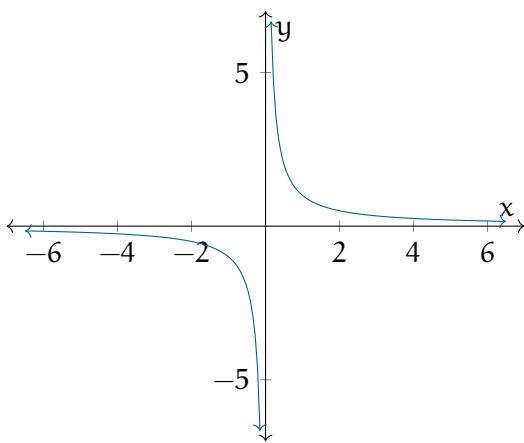
If you have a function, f , its graph is the set of pairs (x, y) such that $y = f(x)$. We usually draw a picture of this set, called a *graph*. The graph not only includes the picture, but also the values of x and y used to create it.

Here is the graph of the function $f(x) = -5x^2 + 12x + 2$:



(Note this is just part of the graph: it goes infinitely in both directions, remember your vectors.)

Here is the graph of the function $f(x) = \frac{1}{x}$:

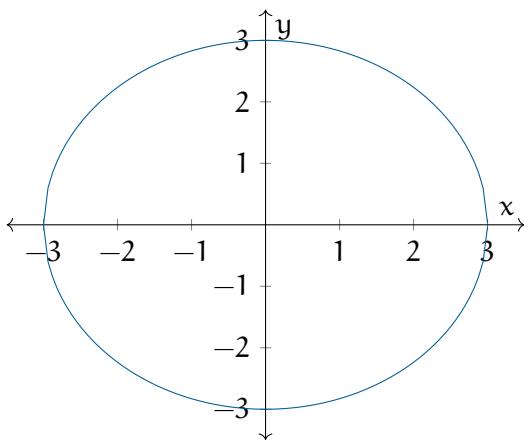

Exercise 37 Draw a graph
Working Space

Let the function f be given by $f(x) = -3x + 3$. Sketch its graph.

Answer on Page 818

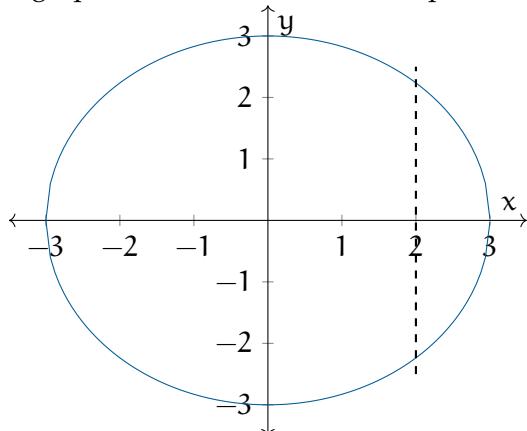
25.2 Can this be expressed as a function?

Note that not all sets can be expressed as graphs of functions. For example, here is the set of points (x, y) such that $x^2 + y^2 = 9$:



This cannot be the graph of a function because what would $f(0)$ be? 3 or -3? This set fails

what we call “the vertical line test”: If any vertical line contains more than one point from the set, it isn’t the graph of a function. For example, the vertical line $x = 2$ would cross



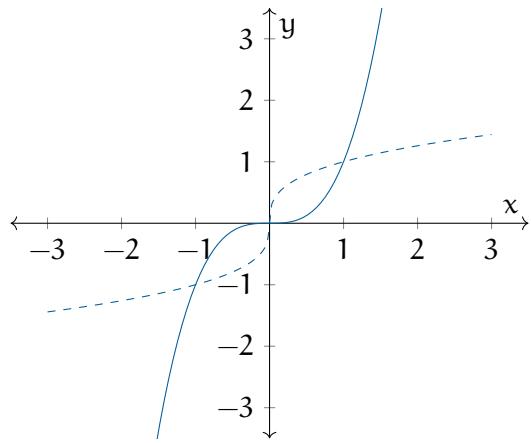
the graph twice:

25.3 Inverses

Some functions have inverse functions. If a function f is a machine that turns number x into y , the inverse (usually denoted f^{-1}) is the machine that turns y back into x .

For example, let $f(x) = 5x + 1$. Its inverse is $f^{-1}(x) = (x - 1)/5$. (Spot check it: $f(3) = 16$ and $f^{-1}(16) = 3$)

Does the function $f(x) = x^3$ have an inverse? Yes, $f^{-1}(x) = \sqrt[3]{x}$. Let’s plot the function (solid line) and its inverse (dashed):

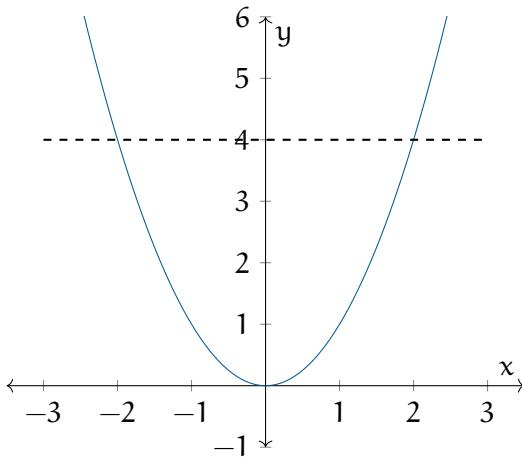


The inverse is the same as the function, just with its axes swapped. This tells us how to solve for an inverse: We swap x and y and solve for y .

For example, if you are given the function $f(x) = 5x + 1$, its graph is all (x, y) such that $y = 5x + 1$. The graph of its inverse is all (x, y) such that $x = 5y + 1$. So you solve for y :

$$y = (x - 1)/5.$$

Not every function has an inverse. For example, $f(x) = x^2$. Note that $f(2) = f(-2) = 4$. What would $f^{-1}(4)$ be? 2 or -2? This implies the “horizontal line test”: If any horizontal line contains more than one point of a function’s graph, that function has no inverse.



In some problems, you need an inverse and you don’t need the whole domain, so you trim the domain to a set you can define an inverse on. This allows you to make claims such as “If we restrict the domain to the nonnegative numbers, the function $f(x) = x^2 - 5$ has an inverse: $f^{-1}(x) = \sqrt{x + 5}$.

This begs the question: What is the domain of the inverse function f^{-1} ?

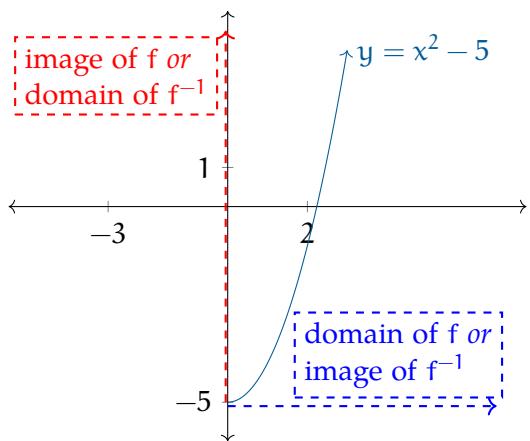
If we let X be the domain of f , we can run every member of X through “the machine” and gather them in a set on the other side. This set would be the *image* of the f “machine”. (This is the *range* of f .)

What is the image of $f(x) = x^2 - 5$? It is the set of all real numbers greater than or equal to -5. We write this

$$\{x \in \mathbb{R} | x \geq -5\}$$

Now we can say: **The image of the function is the domain of the inverse function.**

In our example, we can use any number greater than or equal to -5 as input into the inverse function.



Exercise 38 Find the inverse

Working Space

Let $f(x) = (x - 3)^2 + 2$. Sketch the graph. Using all the real numbers as a domain, does this function have an inverse? How would you restrict the domain to make the function invertible? What is the inverse of that restricted function? What is the domain of the inverse?

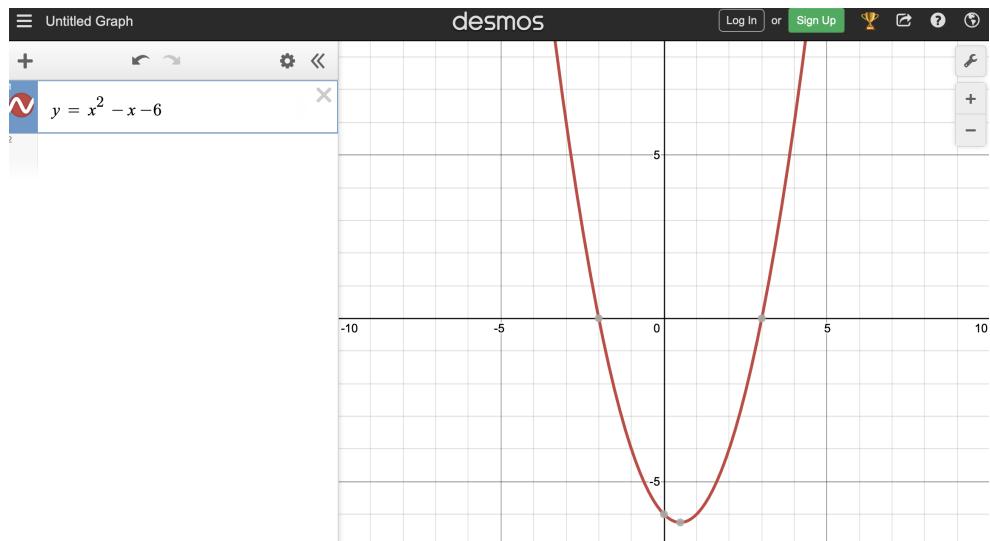
Answer on Page 818

25.4 Graphing Calculators

One really easy way to understand your function better is to use a graphing calculator. Desmos is a great, free online graphing calculator.

In a web browser, go to Desmos: <https://www.desmos.com/calculator>

In the field on the left, enter the function $y = x^2 - x - 6$. (For the exponent, just prefix it with a caret symbol: “ x^2 ”.)

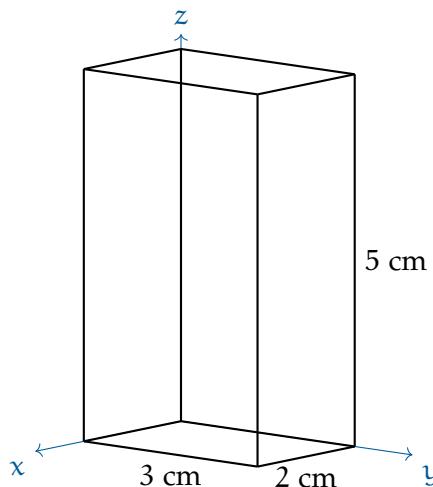




CHAPTER 26

Volumes of Common Solids

The volume of a rectangular solid is the product of its three dimensions. So if a block of ice is 5 cm tall, 3 cm wide and, 2 cm deep, its volume is $5 \times 3 \times 2 = 30$ cubic centimeters.



A cubic centimeter is the same as a milliliter. A milliliter of ice weighs about 0.92 grams.

So the block of ice would have a mass of $30 \times 0.92 = 27.6$ grams.

Volume of a Sphere

A sphere with a radius of r has a volume of

$$v = \frac{4}{3}\pi r^3$$

(For completeness, the surface area of that sphere would be

$$a = 4\pi r^2$$

Note that a circle of radius r is one quarter of this: πr^2 .)

Exercise 39 Flying Sphere

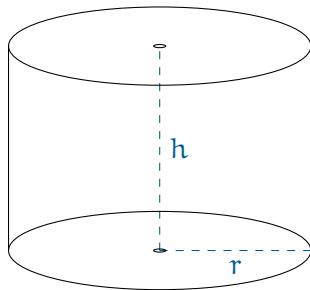
Working Space

An iron sphere is traveling at 5 m/s. (It is not spinning.) The sphere has a radius of 1.5 m. Iron has a density of 7,800 kg per cubic meter. How much kinetic energy does the sphere have?

Answer on Page 819

26.1 Cylinders

The base and the top of a right cylinder are identical circles. The circles are on parallel planes. The sides are perpendicular to those planes.



Volume of a cylinder

The volume of the a right cylinder of radius r and height h is given by:

$$v = \pi r^2 h$$

That is, it is the area of the base times the height.

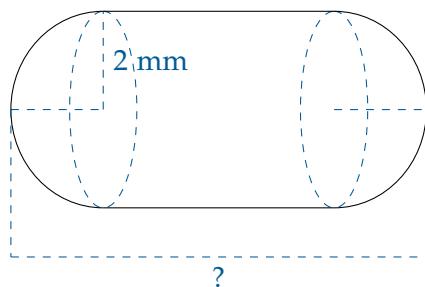
Exercise 40 Tablet

Working Space

A drug company has to create a tablet with volume of 90 cubic millimeters.

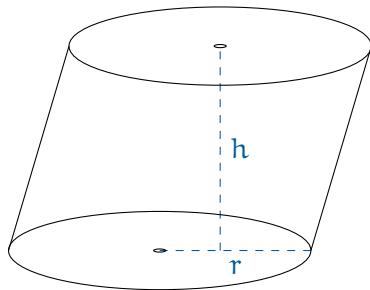
The tablet will be a cylinder with half spheres on each end. The radius will be 2mm.

How long do they need to make the tablet to be?



Answer on Page 820

What if the base and top are identical, but the sides aren't perpendicular to the base? This is called *oblique cylinder*.

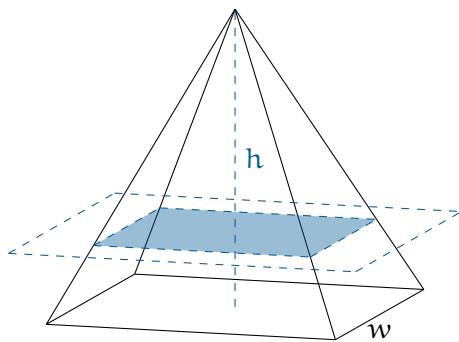


The volume is still the height times the area of the base. Note, however, that the height is measured perpendicular to the bottom and top.

Why?

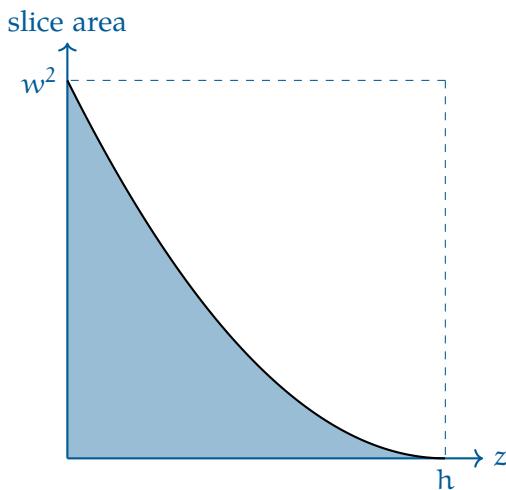
26.2 Volume, Area, and Height

On a solid with a flat base, the line that we use to measure height is always perpendicular to the plane of the base. We can take slices through the solid that are parallel to that base plane. For example, if we have a pyramid with a square base, each slice will be a square – small squares near the top, larger squares near the bottom.



We can figure out the area of the slice at every height z . For example, at $z = 0$ the slice would have area w^2 . At $z = h$, the slice would have zero area. What about an arbitrary z in between? The edge of the square would be $w(1 - \frac{z}{h})$. So the area of the slice would be $w^2(1 - \frac{z}{h})^2$

The graph of this would look like this:



The volume is given by the area under the curve and above the axis. Once you learn integration, you will be really good at finding the area under the curve. In this case, I will just tell you that in the picture, the colored region is one third of the rectangle.

Thus, the area of a square-based pyramid is $\frac{1}{3}hw^2$.

In fact:

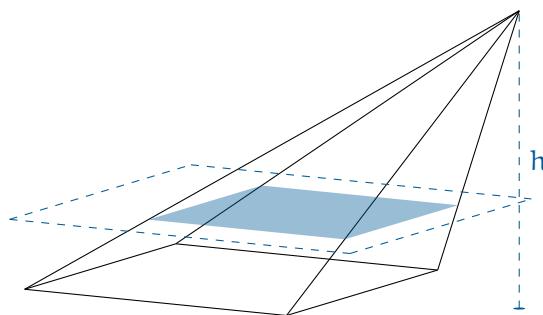
Volume of a pyramid

The volume of pyramid whose base has an area of b and height h is given by:

$$V = \frac{1}{3}hb$$

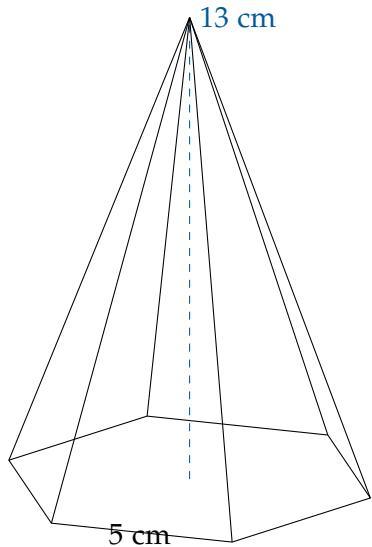
Regardless of the shape of the base.

Note that this is true even for oblique pyramids:

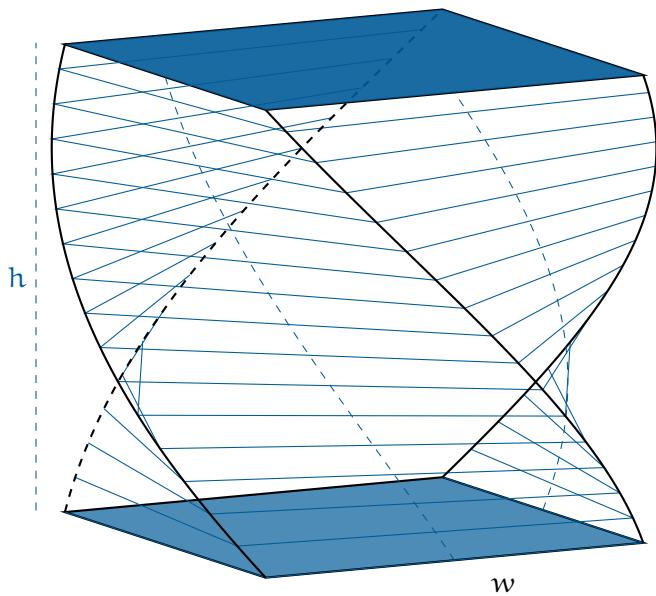


Exercise 41 Hexagon-based Pyramid*Working Space*

There is a pyramid with a regular hexagon for a base. Each edge is 5 cm long. The pyramid is 13 cm tall. What is its volume?

*Answer on Page 820*

Note that plotting the area of each slice and finding the area under the curve will let you find the area of many things. For example, let's say that you have a four-sided spiral, where each face has the same width w :



Every slice still has an area of w^2 , thus this figure has a volume of hw^2 .

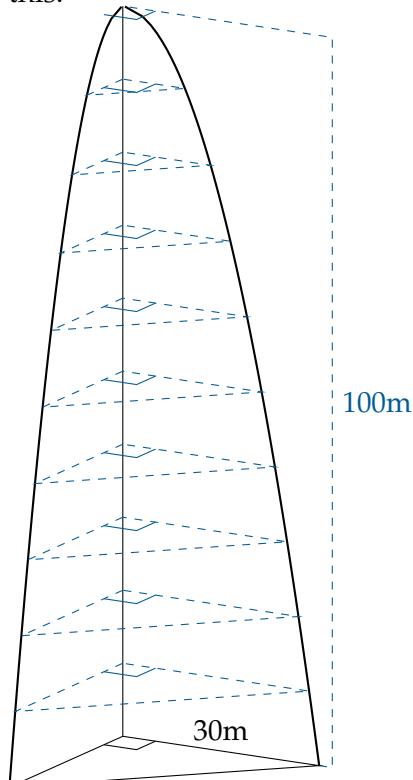
Exercise 42 Volume of a building*Working Space*

An architect is designing a hotel with a right triangular base; the base is 30 meters on each leg. The building gets narrower as you get closer to the top, and finally shrinks to a point. The spine of the building is where the right angle is. That spine is straight and perpendicular to the ground.

Each floor has a right isosceles triangle as its floor plan. The length of each leg is given by this formula:

$$w = 30 \sqrt{1 - \frac{z}{100}}$$

So the width of the building is 30 meters at height $z = 0$. At 100 meters, the building comes to a point. It will look like this:



What is the volume of the building in cubic meters?

Answer on Page 821



CHAPTER 27

Conic Sections

In mathematics, conic sections (or simply conics) are curves obtained as the intersection of the surface of a cone with a plane. The three types of conic section are the hyperbola, the parabola, and the ellipse; the circle is a special case of the ellipse, though historically it was sometimes called a fourth type.

27.1 Definitions

Each type of conic sections can be defined as follows:

27.1.1 Circle

A circle is the set of all points in a plane that are at a given distance (the radius) from a given point (the center). The standard equation for a circle with center (h, k) and radius r is:

$$(x - h)^2 + (y - k)^2 = r^2 \quad (27.1)$$

27.1.2 Ellipse

An ellipse is the set of all points such that the sum of the distances from two fixed points (the foci) is constant. The standard equation for an ellipse centered at the origin with semi-major axis a and semi-minor axis b is:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (27.2)$$

27.1.3 Hyperbola

A hyperbola is the set of all points such that the absolute difference of the distances from two fixed points (the foci) is constant. The standard equation for a hyperbola centered at the origin is:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad (27.3)$$

or

$$\frac{y^2}{b^2} - \frac{x^2}{a^2} = 1 \quad (27.4)$$

depending on the orientation of the hyperbola.

27.1.4 Parabola

A parabola is the set of all points that are equidistant from a fixed point (the focus) and a fixed line (the directrix). The standard equation for a parabola that opens upwards or downwards is:

$$y = a(x - h)^2 + k \quad (27.5)$$

and that opens leftwards or rightwards is:

$$x = a(y - k)^2 + h \quad (27.6)$$

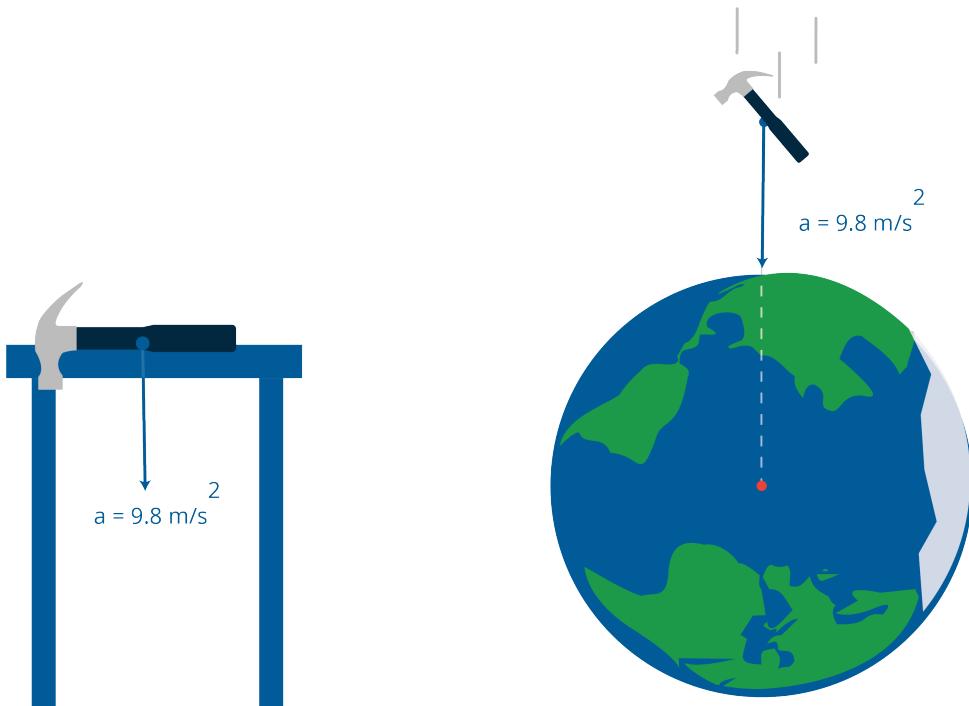
where (h, k) is the vertex of the parabola.



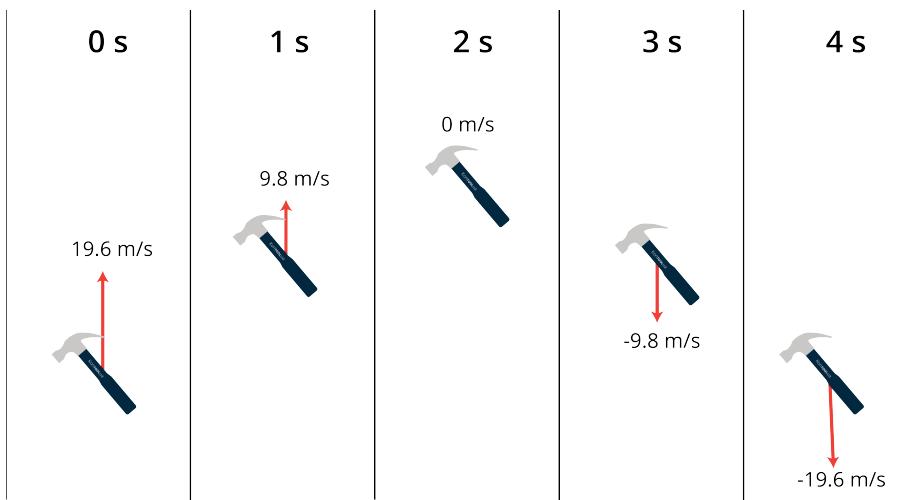
CHAPTER 28

Falling Bodies

Because of gravity, if you throw a hammer straight up in the air, from the moment it leaves your hand until it hits the ground, it is accelerating toward the center of the earth at a constant rate.



Acceleration can be defined as change in velocity. If the hammer leaves your hand with a velocity of 12 meters per second upward, one second later it will be rising, and its velocity will have slowed to 2.2 meters per second. One second after that, the hammer will be falling at a rate of 7.6 meters per second. Every second the hammer's velocity is changing by 9.8 meters per second, and that change is always toward the center of the earth. When the hammer is going up, gravity is slowing it down by 9.8 meters per second, each second it is in the air. When the hammer is coming down, gravity is speeding it up by 9.8 meters per second.



Acceleration due to gravity on earth is a constant negative 9.8 meters per second per second:

$$a = -9.8$$

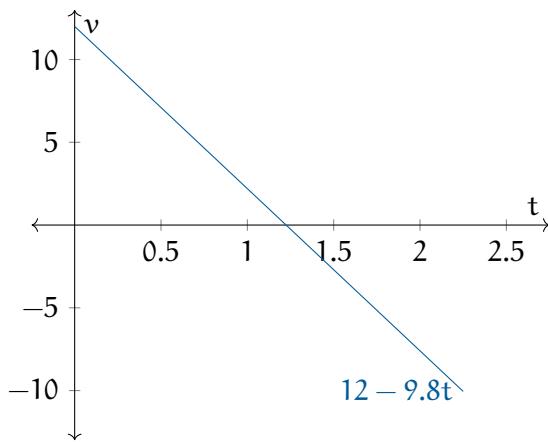
(Why is it negative? We are talking about height, which increases as you go away from the center of the earth. Acceleration is changing the velocity in the opposite direction.)

28.1 Calculating the Velocity

Given that the acceleration is constant, it makes sense that the velocity is a straight line. Assuming once again that the hammer leaves your hand at 12 meters per second, then the upwards velocity at time t is given by:

$$v = 12 - 9.8t$$

Note that the velocity of the hammer is being given as a function. Here is its graph:



Exercise 43 When is the apex of flight?

Given the hammer's velocity is given by $12 - 9.8t$, at what time (in seconds) does it stop rising and begin to fall?

Working Space

Answer on Page 822

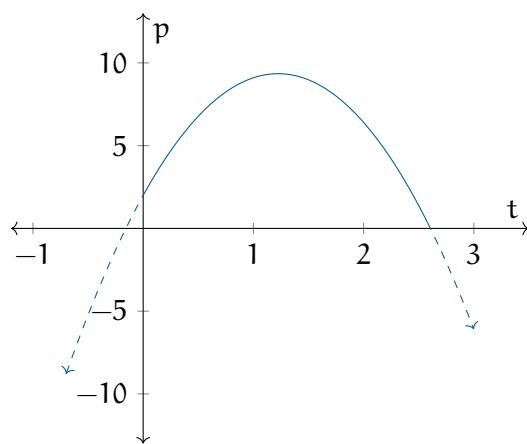
At this point, we need to acknowledge air resistance. Gravity is not the only force on the hammer; as it travels through the air, the air tries to slow it down. This force is called *air resistance*, and for a large, fast-moving object (like an airplane) it is GIGANTIC force. For a dense object (like a hammer) moving at a slow speed (what you generate with your hand), air resistance doesn't significantly affect acceleration.

28.2 Calculating Position

If you let go of the hammer when it is 2 meters above the ground, the height of the hammer is given by:

$$p = -\frac{9.8}{2}t^2 + 12t + 2$$

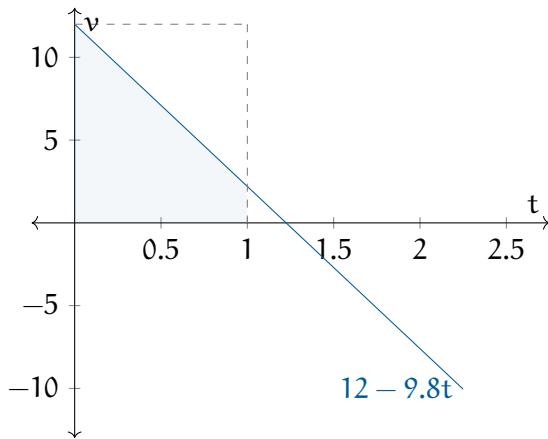
Here is a graph of this function:



How do we know? **The change in position between time 0 and any time t is equal to the area under the velocity graph between $x = 0$ and $x = t$.**

Let's use the velocity graph to figure out how much the position has changed in the first

second of the hammer's flight. Here's the velocity graph with the area under the graph for the first second filled in:



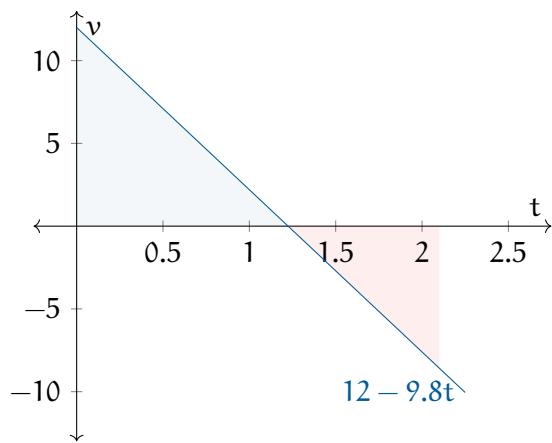
The blue filled region is the area of the dashed rectangle minus that empty triangle in its upper left. The height of the rectangle is twelve and its width is the amount of time the hammer has been in flight (t). The triangle is t wide and $9.8t$ tall. Thus, the area of the blue region is given by $12t - \frac{1}{2}9.8t^2$.

That's the change in position. Where was it originally? 2 meters off the ground. So the height is given by $p = 2 + 12t - \frac{1}{2}9.8t^2$. We usually write terms so that the exponent decreases, so:

$$p = -\frac{1}{2}9.8t^2 + 12t + 2$$

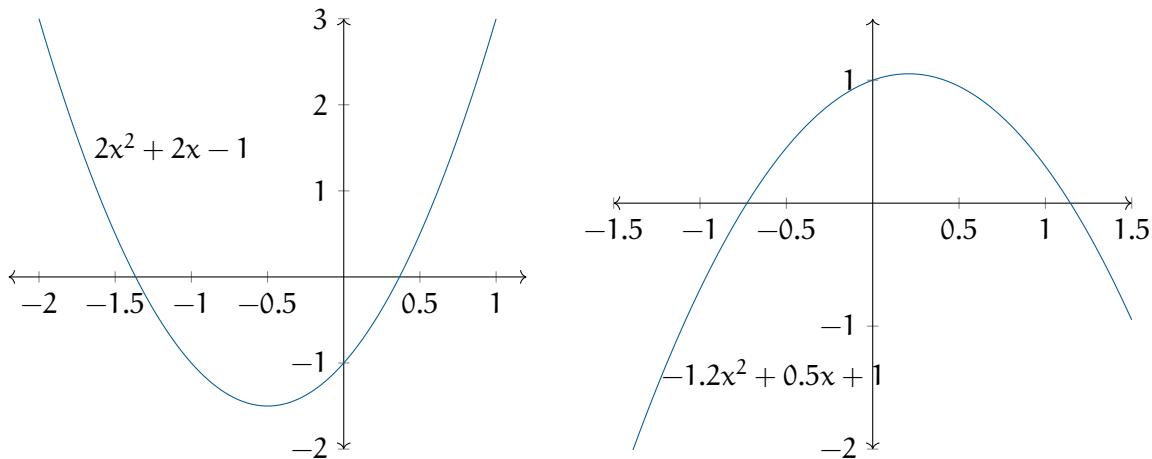
Finding the area under the curve like this is called *integration*. We say "To find a function that gives the change in position, we just integrate the velocity function." A lot of the study of calculus is learning to integrate different sorts of functions.

One important note about integration: Any time the curve drops under the x -axis, the area is considered negative. (Which makes sense, right? If the velocity is negative, the hammer's position is decreasing.)



28.3 Quadratic functions

Functions of the form $f(x) = ax^2 + bx + c$ are called *quadratic functions*. If $a > 0$, the ends go up. If $a < 0$, the ends go down.



The graph of a quadratic function is a *parabola*.

28.4 Simulating a falling body in Python

Now you are going to write some Python code that simulates the flying hammer. First, we are just going to print out the position, speed, and acceleration of the hammer for every 1/100th of a second after it leaves your hand. (Later we will make a graph.)

Create a file called `falling.py` and type this into it:

```
# Acceleration on earth
acceleration = -9.8 # m/s/s

# Size of time step
time_step = 0.01 # seconds

# Initial values
speed = 12 # m/s upward
height = 2 # m above the ground
current_time = 0.0 # seconds after release

# Is the hammer still aloft?
while height > 0.0:

    # Show the values
    print(f"current_time:.2f} s:")
    print(f"\acceleration: {acceleration:.2f} m/s/s")
    print(f"\tspeed: {speed:.2f} m/s")
    print(f"\theight: {height:.2f} m")

    # Update height
    height = height + time_step * speed

    # Update speed
    speed = speed + time_step * acceleration

    # Update time
    current_time = current_time + time_step

print("Hit the ground: Complete")
```

When you run it, you will see something like this:

```
0.00 s:
    acceleration: -9.80 m/s/s
    speed: 12.00 m/s
    height: 2.00 m
0.01 s:
    acceleration: -9.80 m/s/s
    speed: 11.90 m/s
    height: 2.12 m
0.02 s:
    acceleration: -9.80 m/s/s
    speed: 11.80 m/s
```

```
height: 2.24 m
0.03 s:
    acceleration: -9.80 m/s/s
    speed: 11.71 m/s
    height: 2.36 m
...
2.60 s:
    acceleration: -9.80 m/s/s
    speed: -13.48 m/s
    height: 0.20 m
2.61 s:
    acceleration: -9.80 m/s/s
    speed: -13.58 m/s
    height: 0.07 m
Hit the ground: Complete
```

Note that the acceleration isn't changing at all, but it is changing the speed, and the speed is changing the height.

We can see that the hammer in our simulation hits the ground just after 2.61 seconds.

28.4.1 Graphs and Lists

Now, we are going to graph the acceleration, speed, and height using a library called `matplotlib`. However, to make the graphs, we need to gather all the data into lists.

For example, we will have a list of speeds, and the first three entries will be 12.0, 11.9, and 11.8.

We create an empty list and assign it to a variable like this:

```
x = []
```

Then we can add items like this:

```
x.append(3.14)
```

To get the first time back, we can ask for the object at index 0.

```
y = x[0]
```

Note that the list starts at 0. So if you have 32 items in the list, the first item is at index 0. The last item is at index 31.

Duplicate the file falling.py and name the new copy falling_graph.py

We are going to make a plot of the height over time. At the start of the program, you will import the matplotlib library. At the end of the program, you will create a plot and show it to the user.

In falling_graph.py, add the bold code:

```
import matplotlib.pyplot as plt

# Acceleration on earth
acceleration = -9.8 # m/s/s

# Size of time step
time_step = 0.01 # seconds

# Initial values
speed = 12 # m/s upward
height = 2 # m above the ground
current_time = 0.0 # seconds after release

# Create empty lists
accelerations = []
speeds = []
heights = []
times = []

# Is the hammer still aloft?
while height > 0.0:

    # Add the data to the lists
    times.append(current_time)
    accelerations.append(acceleration)
    speeds.append(speed)
    heights.append(height)

    # Update height
    height = height + time_step * speed

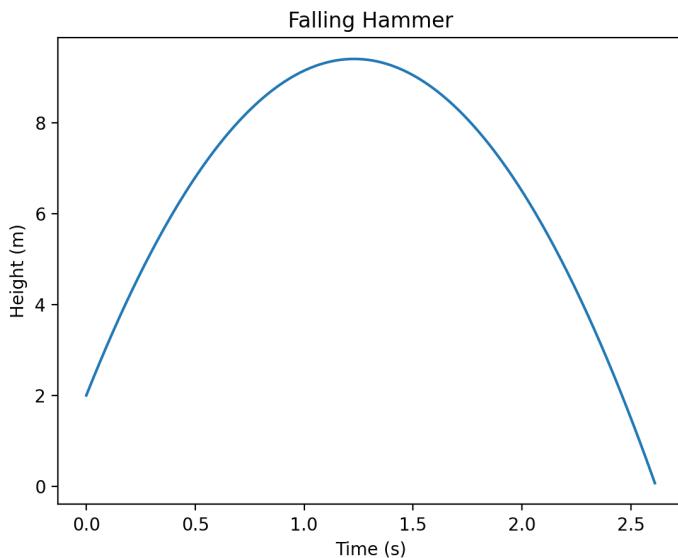
    # Update speed
    speed = speed + time_step * acceleration

    # Update time
```

```
current_time = current_time + time_step

# Make a plot
fig, ax = plt.subplots()
fig.suptitle("Falling Hammer")
ax.set_xlabel("Time (s)")
ax.set_ylabel("Height (m)")
ax.plot(times, heights)
plt.show()
```

When you run the program, you should see a graph of the height over time.



It is more interesting if we can see all three: acceleration, speed, and height. So lets make three stacked plots. Change the plotting code in `falling_graph.py` to:

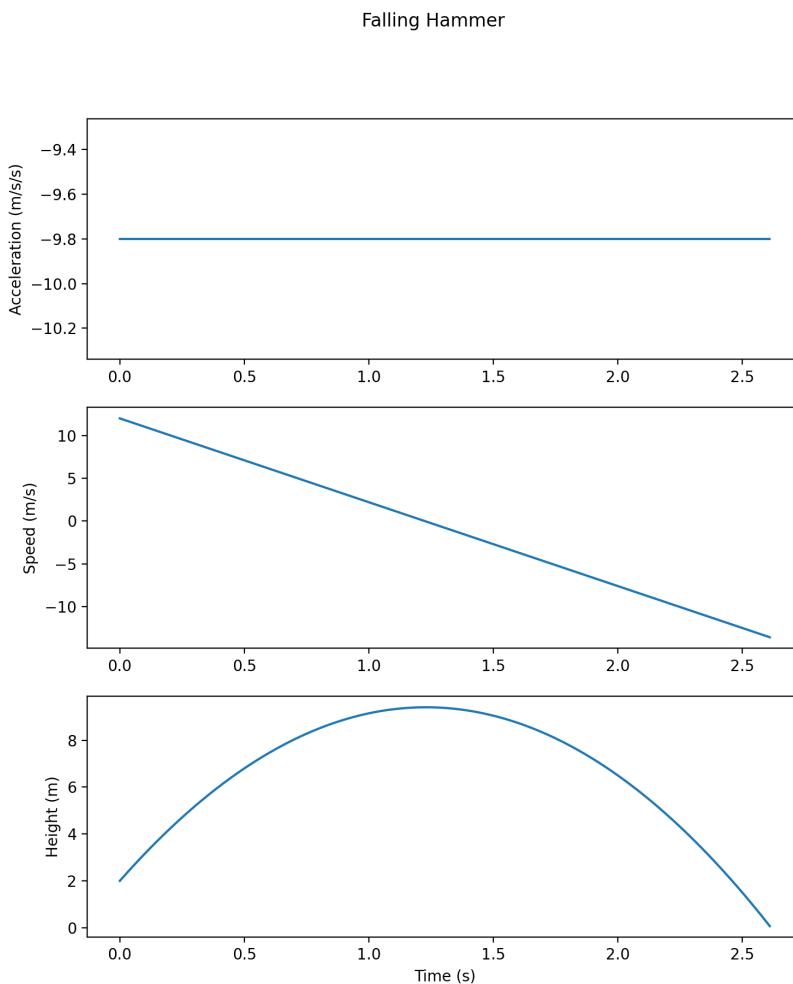
```
# Make a plot with three subplots
fig, ax = plt.subplots(3,1)
fig.suptitle("Falling Hammer")

# The first subplot is acceleration
ax[0].set_ylabel("Acceleration (m/s/s)")
ax[0].plot(times, accelerations)

# Second subplot is speed
ax[1].set_ylabel("Speed (m/s)")
ax[1].plot(times, speeds)
```

```
# Third subplot is height
ax[2].set_xlabel("Time (s)")
ax[2].set_ylabel("Height (m)")
ax[2].plot(times, heights)
plt.show()
```

Now you will get plots of all three variables:



This is what we expected, right? The acceleration is a constant negative number. The speed is a straight line with a negative slope. The height is a parabola.

A natural question at this point is “When exactly will the hammer hit the ground?” That is, when does height = 0? The values of t where a function is zero are known as its *roots*. Height is given by a quadratic function. In the next chapter, you will get the trick for finding the roots of any quadratic function.



CHAPTER 29

Solving Quadratics

A quadratic function has three terms: $ax^2 + bx + c$. a , b , and c are known as the *coefficients*. The coefficients can be any constant, except that a can never be zero. (If a is zero, it is a linear function, not a quadratic.)

When you have an equation with a quadratic function on one side and a zero on the other, you have a quadratic equation. For example:

$$72x^2 - 12x + 1.2 = 0$$

How can you find the values of x that will make this equation true?

You can always reduce a quadratic equation so that the first coefficient is 1, so that your equation looks like this:

$$x^2 + bx + c = 0$$

For example, if you are asked to solve $4x^2 + 8x - 19 = -2x^2 - 7$

$$4x^2 + 8x - 19 = -2x^2 - 7$$

$$6x^2 + 8x - 12 = 0$$

$$x^2 + \frac{4}{3}x - 2 = 0$$

Here, $b = \frac{4}{3}$ and $c = -2$.

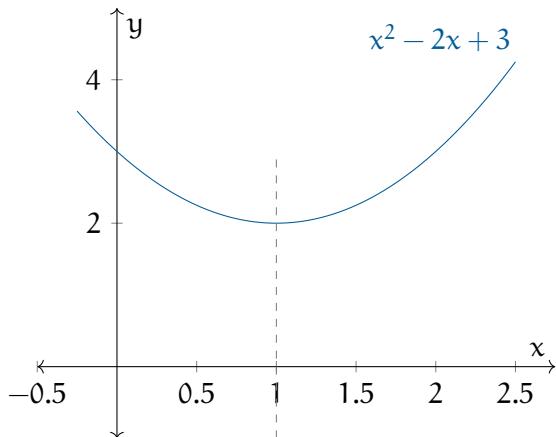
$$x^2 + bx + c = 0 \text{ when}$$

$$x = -\frac{b}{2} \pm \frac{\sqrt{b^2 - 4c}}{2}$$

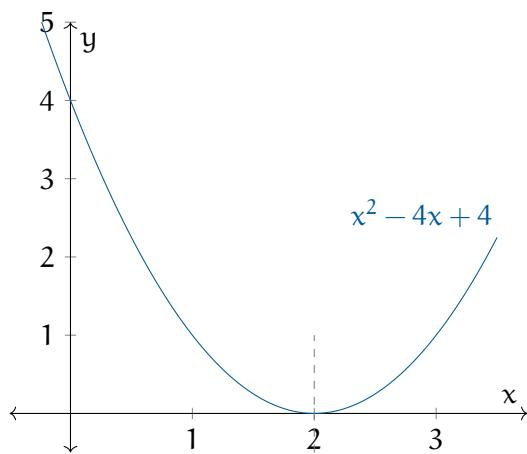
What does this mean?

For any b and c , the graph of $x^2 + bx + c$ is a parabola that goes up on each end. Its low point is at $x = -\frac{b}{2}$.

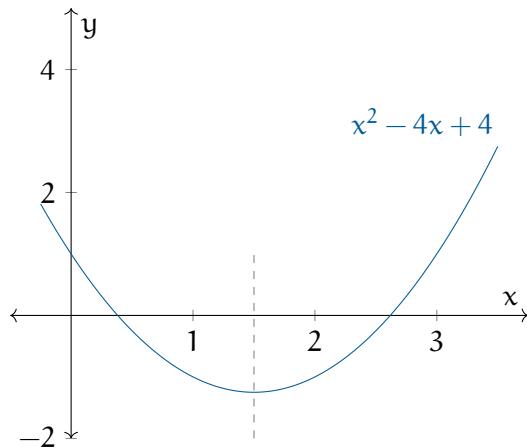
If there are no real roots ($b^2 - 4c < 0$), which means the parabola never gets low enough to cross the x -axis:



If there is one real root ($b^2 - 4c = 0$), it means that the parabola just touches the x -axis.



If there are two real roots ($b^2 - 4c > 0$), it means that the parabola crosses the x-axis twice as it dips below and then returns:



Exercise 44 Roots of a Quadratic*Working Space*

In the last chapter, you found that the function for the height of your flying hammer is:

$$p = -\frac{1}{2}9.8t^2 + 12t + 2$$

At what time will the hammer hit the ground?

*Answer on Page 822***29.1 The Traditional Quadratic Formula**

If the last explanation was a little tricky to understand the quadratic formula is a nifty tool.

The Quadratic Formula

$$ax^2 + bx + c = 0 \text{ when}$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$



CHAPTER 30

Complex Numbers

Complex numbers are an extension of the real numbers, which in turn are an extension of the rational numbers. In mathematics, the set of complex numbers is a number system that extends the real number line to a full two dimensions, using the imaginary unit which is denoted by i , with the property that $i^2 = -1$.

30.1 Definition

A complex number is a number of the form $a + bi$, where a and b are real numbers, and i is the imaginary unit, with the property that $i^2 = -1$. The real part of the complex number is a , and the imaginary part is b .

30.2 Why Are Complex Numbers Necessary?

Complex numbers are essential to many fields of science and engineering. Here are a few reasons why:

30.2.1 Roots of Negative Numbers

In the real number system, the square root of a negative number does not exist because there is no real number that you can square to get a negative number. The introduction of the imaginary unit i , which has the property that $i^2 = -1$, allows us to take square roots of negative numbers and gives rise to complex numbers.

30.2.2 Polynomial Equations

The fundamental theorem of algebra states that every non-constant polynomial equation with complex coefficients has a complex root. This theorem guarantees that polynomial equations of degree n always have n roots in the complex plane.

30.2.3 Physics and Engineering

In physics and engineering, complex numbers are used to represent waveforms, in control systems, in quantum mechanics, and many other areas. Their properties make many mathematical manipulations more convenient.

30.3 Adding Complex Numbers

The addition of complex numbers is straightforward. If we have two complex numbers $z_1 = a + bi$ and $z_2 = c + di$, their sum is defined as:

$$z_1 + z_2 = (a + c) + (b + d)i \quad (30.1)$$

In other words, you add the real parts to get the real part of the sum, and add the imaginary parts to get the imaginary part of the sum.

30.4 Multiplying Complex Numbers

The multiplication of complex numbers is a bit more involved. If we have two complex numbers $z_1 = a + bi$ and $z_2 = c + di$, their product is defined as:

$$z_1 \cdot z_2 = (a + bi) \cdot (c + di) = ac + adi + bci - bd = (ac - bd) + (ad + bc)i \quad (30.2)$$

Note the last term comes from $i^2 = -1$. You multiply the real parts and the imaginary parts just as you would in a binomial multiplication, and remember to replace i^2 with -1 .

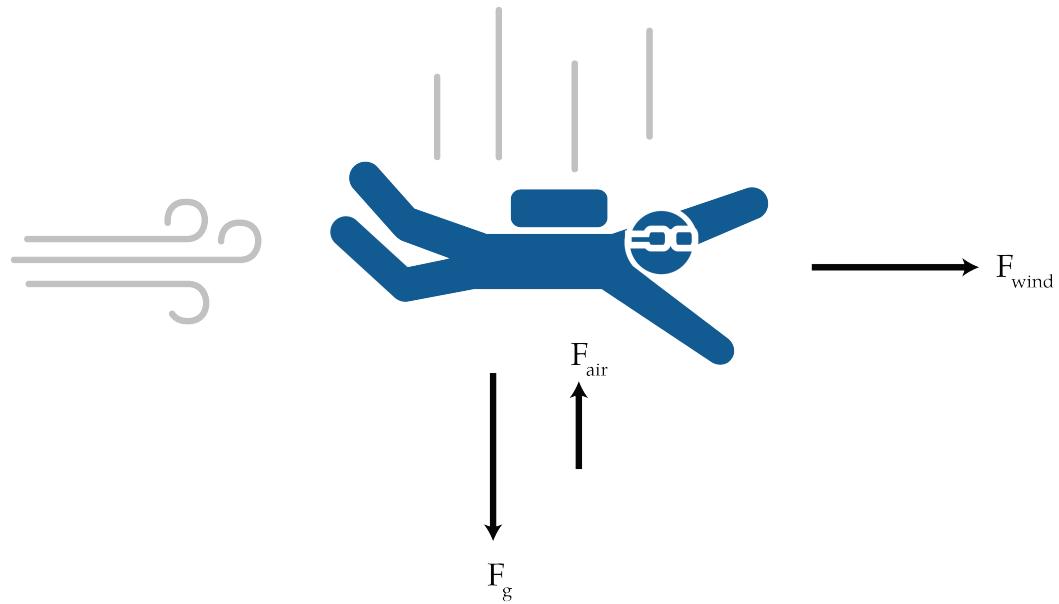


CHAPTER 31

Vectors

We have talked a some about forces, but in the calculations that we have done, we have only talked about the magnitude of a force. It is equally important to talk about its direction. To do the math on things with a magnitude and a direction (like forces), we need vectors.

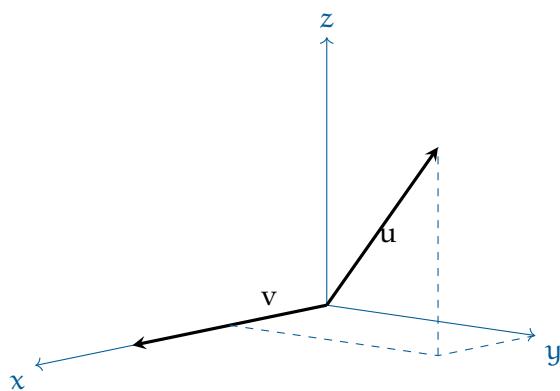
For example, if you jump out of a plane (hopefully with a parachute), several forces with different magnitudes and directions will be acting upon you. Gravity will push you straight down. That force will be proportional to your weight. If there were a wind from the west, it would push you toward the east. That force will be proportional to the square of the speed of the wind and approximately proportional to your size. Once you are falling, there will be resistance from the air that you are pushing through – that force will point in the opposite direction from the direction you are moving and will be proportional to the square of your speed.



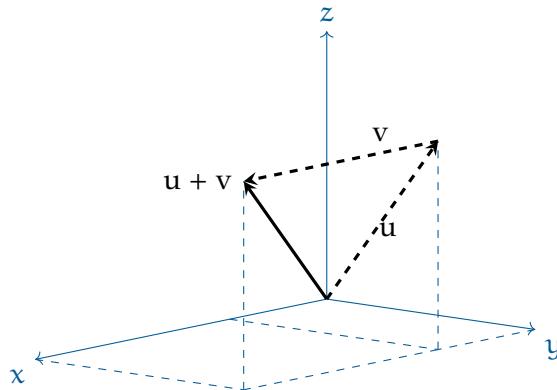
To figure out the net force (which will tell us how we will accelerate), we will need to add these forces together. So we need to learn to do math with vectors.

31.1 Adding Vectors

A vector is typically represented as a list of numbers, with each number representing a particular dimension. For example, if I am creating a 3-dimensional vector representing a force, it will have three numbers representing the amount of force in each of the three axes. For example, if a force of one newton is in the direction of the x -axis, I might represent the vector as $v = [1, 0, 0]$. Another vector might be $u = [0.5, 0.9, 0.7]$



Thinking visually, when we add two vectors, we put the starting point second vector at the ending point of the first vector.



If you know the vectors, you will just add them element-wise:

$$\mathbf{u} + \mathbf{v} = [0.5, 0.9, 0.7] + [1.0, 0.0, 0.0] = [1.5, 0.9, 0.7]$$

These vectors have 3 components, so we say they are *3-dimensional*. Vectors can have any number of components. For example, the vector $[-12.2, 3, \pi, 10000]$ is 4-dimensional.

You can only add two vectors if they have the same dimension.

$$[12, -4] + [-1, 5] = [11, 1]$$

Addition is commutative: If you have two vectors \mathbf{a} and \mathbf{b} , then $\mathbf{a} + \mathbf{b}$ is the same as $\mathbf{b} + \mathbf{a}$.

Addition is also associative: If you have three vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} , it doesn't matter which order you add them in. That is, $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$.

A 1-dimensional vector is just a number. We say it is a *scalar*, not a vector.

Exercise 45 Adding vectors

Add the following vectors:

Working Space

- $[1, 2, 3] + [4, 5, 6]$
- $[-1, -2, -3, -4] + [4, 5, 6, 7]$
- $[\pi, 0, 0] + [0, \pi, 0] + [0, 0, \pi]$

Answer on Page 823

Exercise 46 Adding Forces

Working Space

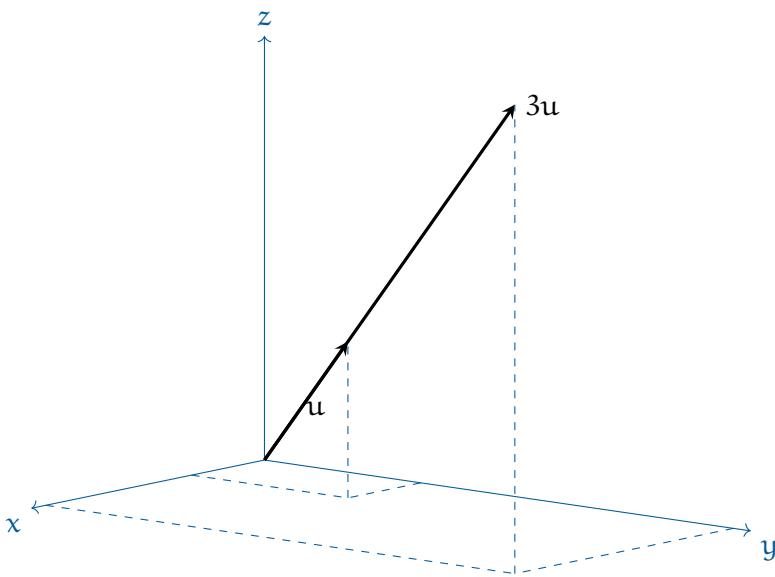
You are adrift in space. You are near two different stars. The gravity of one star is pulling you towards it with a force of $[4.2, 5.6, 9.0]$ newtons. The gravity of the other star is pulling you towards it with a force of $[-100.2, 30.2, -9.0]$ newtons. What is the net force?

Answer on Page 823

31.2 Multiplying a vector with a scalar

It is not uncommon to multiply a vector by a scalar. For example, a rocket engine might have a force vector v . If you fire 9 engines in the exact same direction, the resulting force vector would be $9v$.

Visually, when we multiply a vector u by a scalar a , we get a new vector that goes in the same direction as u but has a magnitude a times as long as u .



When you multiply a vector by a scalar, you just multiply each of the components by the scalar:

$$3 \times [0.5, 0.9, 0.7] = [1.5, 2.7, 3.6]$$

Exercise 47 Multiplying a vector and a scalar

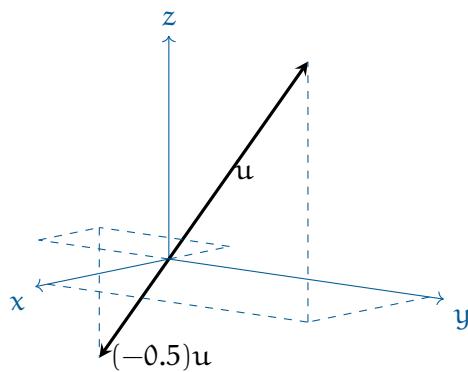
Simplify the following expressions:

Working Space

- $2 \times [1, 2, 3]$
- $[-1, -2, -3, -4] \times -2$
- $\pi[\pi, 2\pi, 3\pi]$

Answer on Page 823

Note that when you multiply a vector times a negative number, the new vector points in the opposite direction.



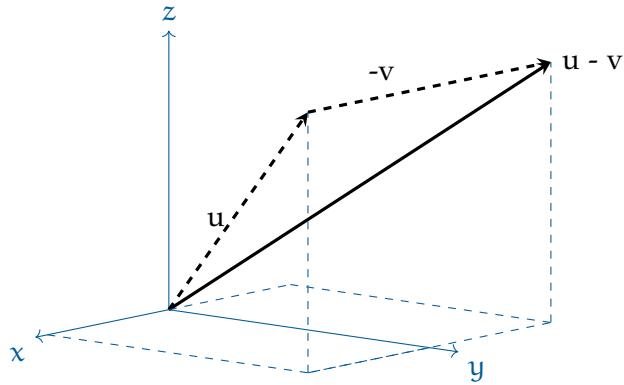
31.3 Vector Subtraction

As you might guess, when you subtract one vector from another, you just do element-wise subtraction:

$$[4, 2, 0] - [3, -2, 9] = [1, 4, -9]$$

So, $u - v = u + (-1v)$.

So visually, you reverse the one that is being subtracted:



31.4 Magnitude of a Vector

The *magnitude* of a vector is just its length. We write the magnitude of a vector v as $|v|$.

We compute the magnitude using the pythagorean theorem. If $v = [3, 4, 5]$, then

$$|v| = \sqrt{3^2 + 4^2 + 5^2} = \sqrt{50} \approx 7.07$$

(You might notice that the notation for the magnitude is exactly like the notation for absolute value. If you think of a scalar as a 1-dimensional vector, the absolute value and the magnitude are the same. For example, the absolute value of -5 is 5. If you take the magnitude of the one-dimensional vector $[-5]$, you get $\sqrt{25} = 5$.)

Notice that if you scale up a vector, its magnitude scales by the same amount. For example:

$$|7[3, 4, 5]| = 7\sqrt{50} \approx 7 \times 7.07$$

The rule then is: If you have any vector v and any scalar a :

$$|av| = |a||v|$$

Exercise 48 Magnitude of a Vector

Find the magnitude of the following vectors:

Working Space

- $[1, 1, 1]$
- $[-5, -5, -5]$ (that is the same as $-5 \times [1, 1, 1]$)
- $[3, 4, -4] + [-2, -3, 5]$

Answer on Page 823

31.5 Vectors in Python

NumPy is a library that allows you to work with vectors in Python. You might need to install it on your computer. This is done with pip. pip3 installs things specifically for Python 3.

```
pip3 install NumPy
```

We can think of a vector as a list of numbers. There are also grids of numbers known as *matrices*. NumPy deals with both in the same way, so it refers to both of them as arrays.

The study of vectors and matrices is known as *Linear Algebra*. Some of the functions we need are in a sublibrary of NumPy called `linalg`.

As a convention, everyone who uses NumPy, imports it as `np`.

Create a file called `first_vectors.py`:

```
import NumPy as np

# Create two vectors
v = np.array([2,3,4])
u = np.array([-1,-2,3])
print(f"u = {u}, v = {v}")

# Add them
w = v + u
print(f"u + v = {w}")

# Multiply by a scalar
w = v * 3
print(f"v * 3 = {w}")

# Get the magnitude
# Get the magnitude
mv = np.linalg.norm(v)
mu = np.linalg.norm(u)
print(f"\|v\| = {mv}, \|u\| = {mu}")
```

When you run it, you should see:

```
> python3 first_vectors.py
u = [-1 -2  3], v = [2 3 4]
u + v = [1 1 7]
v * 3 = [ 6  9 12]
\|v\| = 5.385164807134504, \|u\| = 3.7416573867739413
```

31.5.1 Formatting Floats

The numbers 5.385164807134504 and 3.7416573867739413 are pretty long. You probably want it rounded off after a couple of decimal places.

Numbers with decimal places are called *floats*. In the placeholder for your float, you can specify how you want it formatted, including the number of decimal places.

Change the last line to look like this:

```
print(f"\n|v| = {mv:.2f}, |u| = {mu:.2f}")
```

When you run the code, it will be neatly rounded off to two decimal places:

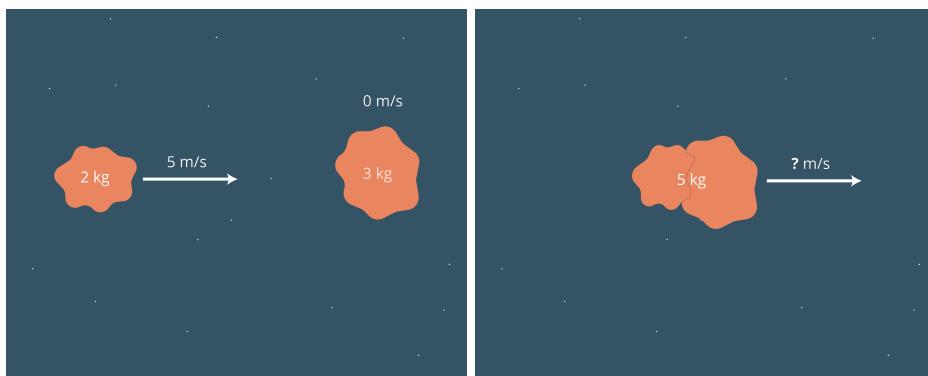
```
|v| = 5.39, |u| = 3.74
```




CHAPTER 32

Momentum

Let's say a 2 kg block of putty is flying through space at 5 meters per second, and it collides with a larger 3 kg block of putty that is not moving at all. When the two blocks deform and stick to each other, how fast will the resulting big block be moving?



Every object has *momentum*. The momentum is a vector quantity: It points in the direction that the object is moving and has a magnitude equal to its mass times its speed.

Given a set of objects that are interacting, we can sum all their momentum vectors to get

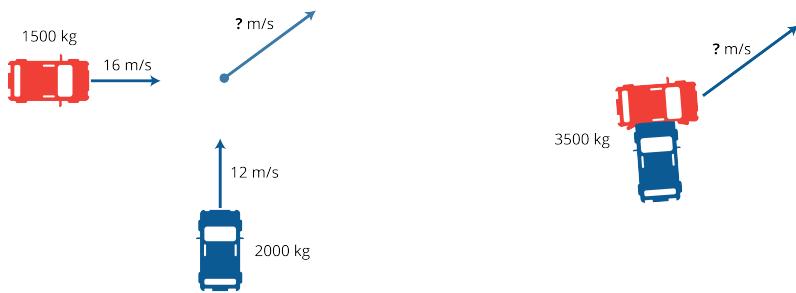
the total momentum. In such a set, the total momentum will stay constant.

So, in our example, one object has a momentum vector of magnitude of 10 kg m/s , the other has a momentum of magnitude 0 . Once they have merged, they have a combined mass of 5 kg . Thus, the velocity vector must have magnitude 2 m/s and pointing in the same direction that the first mass was moving.

Exercise 49 Cars on Ice

A car weighing 1000 kg is going north at 12 m/s . Another car weighing 1500 kg is going east at 16 m/s . They both hit a patch of ice (with zero friction) and collide. Steel is bent and the two objects become one. How what is the velocity vector (direction and magnitude) of the new object sliding across the ice?

Working Space



Answer on Page 824

Notice that kinetic energy ($(1/2)mv^2$) is *not* conserved here. Before the collision, the moving putty block has $(1/2)(2)(5^2) = 25$ joules of kinetic energy. Afterward, the big block has $(1/2)(5)(2^2) = 10$ joules of kinetic energy. What happened to the energy that was lost? It was used up deforming the putty.

What if the blocks were marble instead of putty? Then there would be very little deforming, so kinetic energy *and* momentum would be conserved. The two blocks would end up having different velocity vectors.

Let's assume for a moment that they strike each other straight on, so there is motion in only one direction, both before and after the collision. Can we solve for the speeds of the first block (v_1) and the second block (v_2)?

We end up with two equations. Conservation of momentum says:

$$2v_1 + 3v_2 = 10$$

Conservation of kinetic energy says:

$$(1/2)(2)(v_1^2) + (1/2)(3)(v_2^2) = 25$$

Using the first equation, we can solve for v_1 in terms of v_2 :

$$v_1 = \frac{10 - 3v_2}{2}$$

Substituting this into the second equation, we get:

$$\left(\frac{10 - 3v_2}{2}\right)^2 + \frac{3v_2^2}{2} = 25$$

Simplifying, we get:

$$v_2^2 - 4v_2 + 0 = 0$$

This quadratic has two solutions: $v_2 = 0$ and $v_2 = 4$. $v_2 = 0$ represents the situation before the collision. Substituting in $v_2 = 4$:

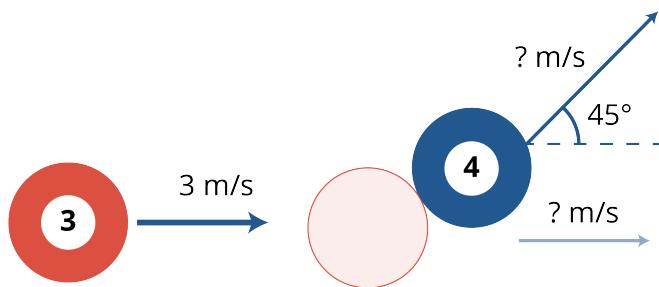
$$v_1 = \frac{10 - 3(4)}{2} = -1$$

Thus, if the blocks are hard enough that kinetic energy is conserved, after the collision, the smaller block will be heading in the opposite direction at 1 m/s. The larger block will be moving at 4 m/s in the direction of the original motion.

Exercise 50 Billiard Balls*Working Space*

A billiard ball weighing 0.4 kg and traveling at 3 m/s hits a billiard ball (same weight) at rest. It strikes obliquely so that the ball at rest starts to move at a 45 degree angle from the path of the ball that hit it.

Assuming all kinetic energy is conserved. How what is the velocity vector of each ball after the collision?

*Answer on Page 824*



CHAPTER 33

The Dot Product

If you have two vectors $u = [u_1, u_2, \dots, u_n]$ and $v = [v_1, v_2, \dots, v_n]$, we define the *dot product* $u \cdot v$ as

$$u \cdot v = (u_1 \times v_1) + (u_2 \times v_2) + \cdots + (u_n \times v_n)$$

So, for example,

$$[2, 4, -3] \cdot [5, -1, 1] = 2 \times 5 + 4 \times -1 + -3 \times 1 = 3$$

This may not seem like a very powerful idea, but dot products are *incredibly* useful. The enormous GPUs(Graphics Processing Unit) that let video games render scenes so quickly? They primarily function by computing huge numbers of dot products at mind-boggling speeds.

Exercise 51 Basic dot products

Compute the dot product of each pair of vectors:

Working Space

- $[1, 2, 3], [4, 5, -6]$
- $[\pi, 2\pi], [2, -1]$
- $[0, 0, 0, 0], [10, 10, 10, 10]$

Answer on Page 825

33.1 Properties of the dot product

Sometimes we need an easy way to say “The vector of appropriate length is filled with zeros.” We use the notation $\vec{0}$ to represent this. Then, for any vector v , this is true:

$$v \cdot \vec{0} = 0$$

The dot product is commutative:

$$v \cdot u = u \cdot v$$

The dot product of a vector with itself is its magnitude squared:

$$v \cdot v = |v|^2$$

If you have a scalar a then:

$$(v) \cdot (au) = a(v \cdot u)$$

So, if v and w are vectors that go in the same direction,

$$\mathbf{v} \cdot \mathbf{w} = |\mathbf{v}||\mathbf{w}|$$

If \mathbf{v} and \mathbf{w} are vectors that go in opposite directions,

$$\mathbf{v} \cdot \mathbf{w} = -|\mathbf{v}||\mathbf{w}|$$

if \mathbf{v} and \mathbf{w} are vectors that are perpendicular to each other, their dot product is zero:

$$\mathbf{v} \cdot \mathbf{w} = 0$$

33.2 Cosines and dot products

Furthermore, dot products' interaction with cosine makes them even more useful is what makes them so useful: If you have two vectors \mathbf{v} and \mathbf{u} ,

$$\mathbf{v} \cdot \mathbf{u} = |\mathbf{v}||\mathbf{u}| \cos \theta$$

where θ is the angle between them.

So, for example, if two vectors \mathbf{v} and \mathbf{u} are perpendicular, the angle between them is $\pi/2$. The cosine of $\pi/2$ is 0: The dot product of any two perpendicular vectors is always 0. In fact, if the dot product of two non-zero vectors is 0, the vectors *must be* perpendicular.

Exercise 52 Using dot products

What is the angle between these each pair of vectors:

- [1, 0], [0, 1]
- [3, 4], [4, 3]

Working Space

Answer on Page 826

If you have two non-zero vectors \mathbf{v} and \mathbf{u} , you can always compute the angle between

them:

$$\theta = \arccos\left(\frac{\mathbf{v} \cdot \mathbf{u}}{|\mathbf{v}||\mathbf{u}|}\right)$$

33.3 Dot products in Python

NumPy will let you do dot products using the symbol @. Open `first_vectors.py` and add the following to the end of the script:

```
# Take the dot product
d = v @ u
print("v @ u =", d)

# Get the angle between the vectors
a = np.arccos(d / (mv * mu))
print(f"The angle between u and v is {a * 180 / np.pi:.2f} degrees")
```

When you run it you should get:

```
v @ u = 4
The angle between u and v is 78.55 degrees
```

33.4 Work and Power

Earlier, we mentioned that mechanical work is the product of the force you apply to something and the amount it moves. For example, if you push a train with a force of 10 newtons as it moves 5 meters, you have done 50 joules of work.

What if you try to push the train sideways? That is, it moves down the track 5 meters, but you push it as if you were trying to derail it – perpendicular to its motion. You have done no work because the train didn't move at all in the direction you were pushing.

Now that you know about dot products: The work you do is the dot product of the force vector you apply and the displacement vector of the train. (The displacement vector is the vector that tells how the train moved while you pushed it.)

Similarly, we mentioned that power is the product of the force you apply and the velocity of the mass you are applying it to. It is actually the dot product of the force vector and the velocity vector.

For example, if you are pushing a sled with a force of 10 newtons and it is moving 2 meters per second, but your push is 20 degrees off, you aren't transferring 20 watts of power to the sled. You are transferring $10 \times 2 \times \cos(20 \text{ degrees}) \approx 18.8$ watts of power.



CHAPTER 34

Boats

For centuries, engineers have been building boats. It is through boat design that humanity learned the lessons that made airplanes and rockets possible. You should know something about boats before we go any further.

34.1 Basic Terminology

The front of a boat is called *the bow*. (It is pronounced exactly the same as "bough.") The back of the boat is called *the stern*.

The underside of the boat is called *the hull*. The top of the boat is called *the deck*.

If you are standing at the stern and looking toward the bow, everything on your left is the *port* side. Everything on your right is the *starboard* side.

There are several different ways that boats are propelled:

- A human pushes the water with a stick. If the stick is attached to the boat with a

pivot (as in a rowboat) it is an *oar*. If the blade is not attached to the boat (as in a canoe), it is a *paddle*.

- A motor turns a screw in the water, as in a motorboat. The screw is known as a *propeller*.
- The wind pushes the boat, as in a sailboat. The sails are held up by a *mast*.
- Some boats have a big fan that pushes the boat. These are called *airboats*. Airboats are not the most efficient boats, but they can travel on waterways with water just a couple of inches deep.

In the terms of physics, each of these method provide a *thrust vector* which is applied to the boat at a particular place and in a particular direction.

The speed of a boat usually measured in *knots*. 1 knot is 1 nautical mile per hour or 1.852 km per hour.

34.2 Why Boats Float Upright

Early in this sequence, we discussed buoyancy as a quantity: The magnitude of the buoyant force is equivalent to the weight of the liquid displaced.

We can also talk about the direction of the buoyant force: buoyancy pushes in the opposite direction as gravity.

How do we design boats so that they don't flip over?

34.2.1 Center of Buoyancy

Let's say you have a rowboat. If you push down on point on the floor near the front, the front of the boat will go down and the back of the boat will rise – that is, besides sinking in the water a little, the boat will rotate in that direction. If you push on the floor near the back of the boat, the back will sink a little lower and the front will rise. But there is a place, near the center of the boat, where if you push down, the boat will not rotate at all, it will just sink a little lower in the water. That point is known as the *center of buoyancy*.

How can we calculate the center of buoyancy? Imagine the shape of the water that was displaced by the boat. Now imagine that shape filled with water. The center of mass of that water is the center of buoyancy of the boat.

34.2.2 Center of Mass

Your boat and everything in it can be thought of as one object. That object has a center of mass. If you found the center of mass, you could balance the whole boat on it.

In a boat, if you move your body from the center of the boat to once side, you will have moved the center of mass. The boat will lean in that direction, which will change the center of buoyancy.

If you imagine a line is parallel to the force of gravity that passes through the center of mass of your boat, the boat will continue to increase its lean until the center of buoyancy is on that line.

If water comes over the sides of the boat before the center of gravity and center of buoyancy align, your boat will sink.

34.3 Center of Lateral Resistance

It isn't enough for a boat to float – for a boat to be useful, it must also be able to travel in a straight line.

Imagine that you are standing knee-deep in a lake next to a canoe. If you push the front of the canoe away from you, it will rotate – the back end will actually swing toward you. There is a point near the middle of the canoe where if you push it will not rotate in either direction – the boat will just slide sideways. This point is known as the *center of lateral resistance*.

The trick to making a boat travel in a straight line is to make sure that the line that contains the thrust vector passes through the center of lateral resistance.

An outboard motor allows you to direct the thrust vector: when the line of thrust passes the center of lateral resistance on the starboard side of the center of lateral resistance, the boat turns toward the port side.

34.4 Steering with a Rudder

While outboard motors and airboats let you direct the thrust vector, most boats have a *rudder*. The rudder is a blade on a pivot near the back of the boat. The angle of the rudder can be adjusted so that water rushing past it gets pushed to one side or the other.

According to Newton's third law, when the water gets pushed to the left, the back of the boat gets pushed (with the same force) to the right. This causes the boat to rotate around its center of lateral resistance.

Note that a rudder only works when the boat passing through the water.

34.5 Boat Length and Resistance

FIXME: Write about wave length, boat length, and Froude number.



CHAPTER 35

Sailboats

Imagine that you have a canoe, and you are about to paddle from one island to another that is directly east of where you are standing. Imagine, also, that there is a steady wind coming from the west, and you have a big piece of plywood. You might be inspired to use it as a sail.

This situation is the most simple form of sailing: Wind comes from behind the boat and hits the sail which generates a force that pushes the boat in the direction of the wind.

The sail has two sides: The *windward* side is the one that is getting hit with the wind. The *leeward* side is the side away from the wind.

35.1 Magnitude of the Wind Force

The first natural question is: How much force will I have pushing my canoe through the water?

Wind Force

When the sail is perpendicular to the wind, the force of the wind on the sail in newtons (F_w) will be given by:

$$F_w = A \frac{dv^2}{2}$$

where A is the area of the sail in square meters d is the density of the gas in kg per cubic meter, and v is the wind speed in meters per second.

For air at STP, d is about 1.225 kg per cubic meter.

We call $\frac{dv^2}{2}$ the *wind pressure*. It is the amount of pressure that the windward side of plywood is experiencing that is above the pressure that the leeward side of the plywood is experiencing. (The leeward side might experience some turbulence, but the pressure it is experiencing is approximately 1 atmosphere.)

Let's say your canoe is standing still and the wind is 0.5 m/s. Then the wind pressure is

$$P = \frac{1.225(0.5^2)}{2} = 0.153125 \text{ newtons per square meter}$$

Let's say your plywood sail is 2 meters tall and 1.5 meters wide. What will be the force of the wind?

$$F_w = AP = (3)(0.153125) \approx 0.46 \text{ newtons}$$

This is a very intuitive idea: There is a difference between the pressure on the windward side and the pressure on the leeward side, and the plywood experiences a force that pushes the boat through the water.

35.2 Direction and Location of the Wind Force

If there is low pressure on one side of the sail and high pressure on the other, the force vector will be perpendicular to the sail.

Where is this force vector applied? We can think of the force as being applied at the geometric center of the sail. This is called *the center of effort*. In this case, the center of effort is the exact center of the rectangular plywood.

The mast on a windsurfing board can be tilted from side to side. When the center of effort is over the center of lateral resistance, the board goes straight. To steer, the sailor moves the mast to one side of the center of lateral resistance which rotates the board.

35.3 Beam Reach

When you are sailing in the same direction as the wind, sailors say you are *running*. What if you want to go east and now the wind (still 0.5 m per second) is coming from the south? Sailing perpendicular to the wind is known as a *beam reach*.

To do a beam reach, instead of mounting the plywood perpendicular to the boat's direction of travel, you would mount it at a 45 degree angle. The wind pressure will build on the windward side of the plywood, and the plywood will experience a force pushing it at a 45° angle to the boat.

We can think of this force as having two components:

- One component pushes the boat forward (Yay!)
- One component pushes the boat sideways (Ugh.)

To minimize the effect of the sideways force, sailboats typically have a keel – a long fin on its underside that slows its sideways sliding.

Notice also that the "wind shadow" of the plywood is smaller when it is at a 45° angle to the wind. How much smaller? The effective area of your plywood has gone from 3 square meters to $\frac{3}{\sqrt{2}} \approx 2.12$ square meters. So the force generated will be smaller, and some of it will be wasted pushing the boat sideways.

If we assume that the wind pressure is still 0.153125 newtons per square meter. The force on the plywood will be about

$$F_w = AP = \frac{3}{\sqrt{2}}(0.154125) \approx 0.325 \text{ newtons}$$

However, the direction of that force is not all in the direction you want to go, so the effective force is:

$$F = F_w \frac{1}{\sqrt{2}} = \frac{3}{2}(0.154125) = 0.2311875 \text{ newtons}$$

Notice that we got twice as much effective force when we were running with the wind as

when we are on a beam reach. However, any sailor will tell you that you can go much faster on a beam reach than you can running. Why?

35.4 Apparent Wind

When you are running, you can never go faster than the wind. As you go faster and faster, the wind that the boat experience decreases. For example, if you are going 0.2 m/s in a wind of 0.5 m/s, you (and your sail) will only experience wind at 0.3 m/s. We call the wind as experienced by the boat the *apparent wind*. The wind as observed by a stationary observer is called the *actual wind*.

If you are running with the wind, as you approach the speed of the wind, the force of that wind will decrease towards zero.

On a beam reach, as you go faster, the direction of the wind seems to change. If you are going 0.2 m/s east and the actual wind is 0.5 m/s from the south, the direction of the apparent wind will seem to come from about 22 degrees east of true south. The speed of the apparent wind will be about 0.54 m/s.

35.5 Close Reach

What if you want to go east, and the wind is coming from 40 degrees east of south? This would mean that you were sailing just 50 degrees away from straight into the wind. Is this possible?

If you put your sail at a 25 degree angle, you will still catch some wind and create some pressure on one side of the sail. Most of the resulting force would be trying to push your boat sideways, but some of it would be in the direction you were trying to travel.

Picking an angle for your sail that creates high pressure that makes a desirable force is known as the "angle of attack".

This is a non-intuitive result: A boat can sail into the wind!? The boat can't sail directly into the wind – with each degree that the boat gets closer to straight into the wind, the force pushing it forward decreases and the force pushing it back increases. However, most boats can get within 45% if they have a well-shaped sail.

35.6 Shaping the Sail

Most of the power of the wind can be captured with a piece of flat plywood. The wind hits it and creates a high pressure on the windward side. What about the other side of the plywood?

It turns out that if we can get the wind to travel smoothly over the back side of the plywood, the pressure on that side will be a little lower than if we had turbulence there. (I'm not going to go into this too deeply into why. If you want to know more, look up the Coanda effect.)

For example, if we were on a close reach, the very best sail we could have would gently pull the wind along its backside. It would look like this:

Of course, for the sail to work on either side of the boat, this asymmetrical design would not work. (Although, we should note that this design works great for airplane wings.)

When we make a sail out of cloth, we give it some curve known as *camber*. Slow winds require just a little camber, fast winds require more.

Some newer sailboats have wing sails that have two pieces that can be arranged to redirect the most air possible.

Note: when running with the wind, the turbulence on the leeward side of the sail is unavoidable. But when traveling perpendicular to the wind or on a close reach, the air should move smoothly over the leeward side of the sail. Many sailors have a piece of yarn taped on each side of the sail so they can see if the air is moving smoothly.

Many sailboats also have multiple sails. Besides the increase in the sail area, each sail also redirects the wind to pass smoothly over the leeward surface of the sail behind it.

35.7 Tacking into the Wind

What if the wind is coming from the east and you really need to go directly into the wind?

The boat will not travel straight into the wind, instead you will travel on a close reach with the wind coming from one side of the boat. Then you will turn into the wind and continue turning until you are on a close reach with the wind coming from the other side. This is known as *tacking*.

35.8 Heeling

The center of effort is near the center of your sail, which is usually pretty high in the air. Yes, the force generated will push your boat, but it will also rotate your boat. Sailors call the resulting lean *heeling*. Heeling too far is problematic – the rudder gets pulled out of the water, which makes it hard to steer the boat.

If a boat is heeling too far, sailors will move their weight to the windward side of the boat – some even wear harnesses that them push their weight out beyond the edge of the hull.

If that doesn't work, they will reduce the sideways force on the sail.

Can the boat flip-over? (The word sailors use is *capsize*.) Most larger boats should not be able to capsize – as the sail gets pushed down toward the water, it loses power. So putting some weight in the keel will ensure that the boat doesn't turn upside-down. Small boats (think 3 or 4 meters long) can capsize, but they are small enough that the sailor can usually get them upright again without assistance.

There are boats with multiple hulls. A catamaran has two identical hulls that are side-by-side. As a result, the catamaran will not heel much at all. However, if a big gust of wind comes up, one hull can be pulled completely out of the water. Catamarans can be capsized.

If a boat is running when a big gust comes up, that rotating force will try to push the bow underwater. If the boat is going very fast, this can result in a somersault as the front of the boat slows and dips suddenly and the back of the boat is pitched up over it.



CHAPTER 36

Introduction to Spreadsheets

For many real-world problems, spreadsheets are the perfect tool. In this chapter, you will be introduced to how to use a spreadsheet. There are numerous spreadsheet programs: Google Sheets, Microsoft Excel, Apple Numbers, OpenOffice Calc, etc. All of them are very similar. This instruction will use Google Sheets, but if you are using one of the others, you should be able to follow along.

The first spreadsheet program (VisiCalc) was introduced in 1979 as a tool for finance people to play “what if” games. For example, a company might make a spreadsheet that told them how much more profit they would make if they changed from using an expensive metal to using a cheaper alloy.

In honor of its history, let’s start by studying a business question: I have a friend who dreams of quitting her job to become a cooper. (A cooper makes barrels that are used for aging wine and whiskey.) She says:

- It costs \$45 dollars in materials to build one barrel.
- A barrel sells for \$100 dollars.

- The workshop/warehouse she wants to rent costs \$2000 per month.
- Taxes take 20% of her profits.
- She needs to make \$4000 monthly after taxes.

She has asked you, "How many barrels do I need to make each month?"

36.1 Solving It Symbolically

Many problems can be solved two ways: symbolically or numerically. To solve this problem symbolically, you would write out the facts as equations or inequalities and then do symbol manipulations until you ended up with an answer. In this case, you would let b be the number of barrels and create the following inequality:

$$(1.0 - 0.2)(b(100 - 45) - 2000) \geq 4000$$

You would simplify it:

$$(0.8)(55b - 2000) \geq 4000$$

And simplify it more:

$$44b - 1600 \geq 4000$$

If that is true, then:

$$44b \geq 5600$$

And if that is true, then:

$$b \geq \frac{1400}{11}$$

$1400/11$ is about 127.27, so she needs to make and sell 128 barrels each month.

That is a perfect answer, and we didn't need a spreadsheet at all. Two things:

- As problems get larger and more realistic, it gets much more difficult to solve them symbolically.
- As soon as you say “Yes, you need to make and sell 128 barrels each month.” Your friend will ask “What if I make and sell 200 barrels? How much money will I make then?”

So we use a spreadsheets to solve the problem numerically.

36.2 Solving It Numerically (with a spreadsheet)

Let’s get back to our example. Put labels in the A column:

- Barrels produced (per month)
- Materials cost (per barrel)
- Sale price (per barrel)
- Pre-tax earnings (per month)
- Taxes (per month)
- Take home pay (per month)

Format them any way you like. It should look something like this:

A
1 Barrels Produced (per month)
2 Materials cost (per barrel)
3 Sale price (per barrel)
4 Rent (per month)
5 Pretax Earnings (per month)
6 Taxes (per month)
7 Take home pay (per month)
8

In the B column, the first four cells are values (not formulas):

- 115 formatted as a number with no decimal point
- 45 formatted as currency
- 100 formatted as currency

- 2000 formatted as currency

It should look something like this:

	A	B
1	Barrels Produced (per month)	115
2	Materials cost (per barrel)	\$45.00
3	Sale price (per barrel)	\$100.00
4	Rent (per month)	\$2,000.00

The next three cells in the B column will have formulas:

- $B1 * (B3 - B2) - B4$
- $0.2 * B5$
- $B5 - B6$

It should look something like this:

	A	B
1	Barrels Produced (per month)	115
2	Materials cost (per barrel)	\$45.00
3	Sale price (per barrel)	\$100.00
4	Rent (per month)	\$2,000.00
5	Pretax Earnings (per month)	\$4,325.00
6	Taxes (per month)	\$865.00
7	Take home pay (per month)	\$3,460.00
8		

Now you can share this spreadsheet with your friend and she can put different values into the cells for what-if games. Like “If I can get my materials cost down to \$42 per barrel, what happens to my take home pay?”

Sometimes it is nice to show a range of values for a variable or two. In this case, it might be nice to show your friend what the numbers look like if she produces 115, 120, 125, 130, 135, or 140 barrels per month.

We have one column, and now we need six. How do we duplicate cells?

1. Click B1 to select it and then shift-click on B7 to select all seven cells.
2. Copy them. (There is probably a menu item for this.)
3. Click C1 to select it
4. Paste them.

	A	B	C
1	Barrels Produced (per month)	115	115
2	Materials cost (per barrel)	\$45.00	\$45.00
3	Sale price (per barrel)	\$100.00	\$100.00
4	Rent (per month)	\$2,000.00	\$2,000.00
5	Pretax Earnings (per month)	\$4,325.00	\$4,325.00
6	Taxes (per month)	\$865.00	\$865.00
7	Take home pay (per month)	\$3,460.00	\$3,460.00

We want the first cell in the new column to be 120. You could just type in 120, but let's do something more clever. Put a formula into that cell: = B1 + 5. Now the cell should show 120.

Why did we put in a formula? When we duplicate this column, this cell will always have 5 more barrels than the cell to its left.

Now let's duplicate the second column a few times. The easy way to do this is to select the cells as you did before and drag the lower-right corner to the right until column G is in the selection. When you end the drag, the copies will appear:

	A	B	C	D	E	F	G
1	Barrels Produced (per month)	115	120	125	130	135	140
2	Materials cost (per barrel)	\$45.00	\$45.00	\$45.00	\$45.00	\$45.00	\$45.00
3	Sale price (per barrel)	\$100.00	\$100.00	\$100.00	\$100.00	\$100.00	\$100.00
4	Rent (per month)	\$2,000.00	\$2,000.00	\$2,000.00	\$2,000.00	\$2,000.00	\$2,000.00
5	Pretax Earnings (per month)	\$4,325.00	\$4,600.00	\$4,875.00	\$5,150.00	\$5,425.00	\$5,700.00
6	Taxes (per month)	\$865.00	\$920.00	\$975.00	\$1,030.00	\$1,085.00	\$1,140.00
7	Take home pay (per month)	\$3,460.00	\$3,680.00	\$3,900.00	\$4,120.00	\$4,340.00	\$4,560.00

Nice, right? Now your friend can easily see how many barrels correspond to how much take-home pay. Do you know what would be really helpful? A graph.

36.3 Graphing

Graphing is a little different on every different platform. Here is what you want the graph to look like.

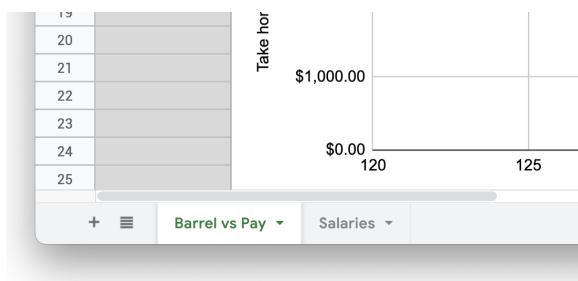


On Google Sheets:

1. Select cells B7 through G7.
2. Choose the menu item Insert -> Chart.
3. Choose the chart type (Line)
4. Add the X-axis to be B1 through G1.
5. Under the Customize tab, Set the label for the X-axis to be "Barrels Made and Sold".
6. Delete the chart title (which is the same as the Y-axis label).

36.4 Other Things You Should Know About Spreadsheets

Your spreadsheet document can have several “Sheets”. Each has its own grid of cells. The sheet has a name; usually, you call it something like “Salaries”. When you need to use a value from the “Salaries” sheet in another sheet, you can specify “Salaries!A2” – that is, cell A2 on sheet “Salaries”. To flip between the sheets there is usually a tab for each at the bottom of the document.



By default, the cell references are relative. That is when you write a formula in cell H5 that references the value in cell G4, the cell remembers “The cell that is one up and one to the left of me.” Thus, if you copy that formula into B9, now that formula reads the value from A8.

If you want an absolute reference, you use \$. If H5 references \$G\$4, G4 will be used no matter where on the sheet the formula is copied to.

You can use the \$ on the row or column. In \$A4, the column is absolute and the row is relative. In A\$4, the row is absolute and the column is relative.

36.5 Challenge: Make a spreadsheet

You have a company that bids on painting jobs. Make a spreadsheet to help you do bids. Here are the parameters:

- The client will tell you how many square meters of wall needs to be painted.
- Paint costs \$0.02 per square meter of wall
- On average, a square meter of wall takes 0.02 hours to paint.
- You can hire painters at \$15 per hour.
- You add 20% to your estimated costs for a margin of error and profit.

Make a spreadsheet such that when you type in the square meters to be painted, the spreadsheet tells you how much you will spend on paint and labor. It also tells you what your bid should be.



CHAPTER 37

Compound Interest

When you loan money to someone, you typically charge them some sort of interest. The most common loan of this sort is what the bank calls a “savings account”. Any money you put in the account is loaned to the bank. The bank then lends it to someone else, who pays interest to the bank. And the bank gives some of that interest to you. However, what if you leave the interest in your account? And you start making *interest on the interest*? This is known as *compound interest*.

37.1 An example with annual interest payments

Let’s say that you put \$1000 in a savings account that pays 6% interest every year. How much money would you have after 12 years? Let’s make a spreadsheet.

	A	B	C
1	Interest Rate	6.00%	
2			
3	After year:	Interest	Balance
4	0	\$0.00	1000
5	1	\$60.00	\$1,060.00

Create a new spreadsheet and edit the cells to look like this. All the cells in rows 1 - 4 are just values: just type in what you see.

The fifth row is all formulas:

After year	Interest	Balance
= A4 + 1	= B\$1 * C4	= C4 + B5

The interest rate field should be formatted as a percentage. One thing to know when dealing with percentages in the spreadsheet: if the field says "600%", its value is 6.

The cells in the Interest and Balance column should be formatted as currency.

You are about to make a bunch of copies of the cells in the fifth row, so make sure they look right.

Click on A5 and shift-click on C5 to select all three cells. Drag the lower-right corner down to fill the rows 6 - 15.

A5:C16			
	A	B	C
1	Interest Rate	6.00%	
2			
3	After year	Interest	Balance
4	0	\$0.00	1000
5	1	\$60.00	\$1,060.00
6	2	\$63.60	\$1,123.60
7	3	\$67.42	\$1,191.02
8	4	\$71.46	\$1,262.48
9	5	\$75.75	\$1,338.23
10	6	\$80.29	\$1,418.52
11	7	\$85.11	\$1,503.63
12	8	\$90.22	\$1,593.85
13	9	\$95.63	\$1,689.48
14	10	\$101.37	\$1,790.85
15	11	\$107.45	\$1,898.30
16	12	\$113.90	\$2,012.20
17			

Look at the numbers. The first interest payment is \$60, but the last is \$113.90. Your balance has more than doubled!

37.2 Exponential Growth

We figured this out numerically by repeatedly multiplying the balance by the interest rate. What if you wanted to know what the balance would be n years after investing P_0 dollars

with an annual interest rate of r ? (Note that r in our example would be 0.06, not 6.0.)

Each year, the balance is multiplied by $1+r$, so after one year, P_0 would become $P_0 \times (1+r)$. The next year you would multiply this number by $(1+r)$ again: $P_0 \times (1+r) \times (1+r)$. The next year? $P_0 \times (1+r) \times (1+r) \times (1+r)$ See the pattern? P_n is this balance after n years, then

$$P_n = P_0(1+r)^n$$

Because n is an exponent, we call this *exponential growth*. And there are few things as terrifying to a scientist as the phrase “The population is undergoing exponential growth”.

37.3 Sensitivity to interest rate

For most people, the first surprising thing about compound interest is how quickly your money grows after a few years. The second thing that is surprising is how much difference a small change in the percentage rate makes.

Let's add another set of columns that shows what happens to your money if you convince the bank to pay you 8% instead of 6%.

Copy everything from columns B and C:

	A	B	C	D	E	
1	Interest Rate	6.00%		6.00%		
2						
3	After year	Interest	Balance	Interest	Balance	
4	0	\$0.00	1000	\$0.00	1000	
5	1	\$60.00	\$1,060.00	\$60.00	\$1,060.00	
6	2	\$63.60	\$1,123.60	\$63.60	\$1,123.60	
7	3	\$67.42	\$1,191.02	\$67.42	\$1,191.02	
8	4	\$71.46	\$1,262.48	\$71.46	\$1,262.48	
9	5	\$75.75	\$1,338.23	\$75.75	\$1,338.23	
10	6	\$80.29	\$1,418.52	\$80.29	\$1,418.52	
11	7	\$85.11	\$1,503.63	\$85.11	\$1,503.63	
12	8	\$90.22	\$1,593.85	\$90.22	\$1,593.85	
13	9	\$95.63	\$1,689.48	\$95.63	\$1,689.48	
14	10	\$101.37	\$1,790.85	\$101.37	\$1,790.85	
15	11	\$107.45	\$1,898.30	\$107.45	\$1,898.30	
16	12	\$113.90	\$2,012.20	\$113.90	\$2,012.20	
17						

Now edit the second interest rate to be 8%:

240 Chapter 37. COMPOUND INTEREST

	A	B	C	D	E	
1	Interest Rate	6.00%		8.00%		
2						
3	After year	Interest	Balance	Interest	Balance	
4	0	\$0.00	1000	\$0.00	1000	
5	1	\$60.00	\$1,060.00	\$80.00	\$1,080.00	
6	2	\$63.60	\$1,123.60	\$86.40	\$1,166.40	
7	3	\$67.42	\$1,191.02	\$93.31	\$1,259.71	
8	4	\$71.46	\$1,262.48	\$100.78	\$1,360.49	
9	5	\$75.75	\$1,338.23	\$108.84	\$1,469.33	
10	6	\$80.29	\$1,418.52	\$117.55	\$1,586.87	
11	7	\$85.11	\$1,503.63	\$126.95	\$1,713.82	
12	8	\$90.22	\$1,593.85	\$137.11	\$1,850.93	
13	9	\$95.63	\$1,689.48	\$148.07	\$1,999.00	
14	10	\$101.37	\$1,790.85	\$159.92	\$2,158.92	
15	11	\$107.45	\$1,898.30	\$172.71	\$2,331.64	
16	12	\$113.90	\$2,012.20	\$186.53	\$2,518.17	
17						



CHAPTER 38

Introduction to Data Visualization

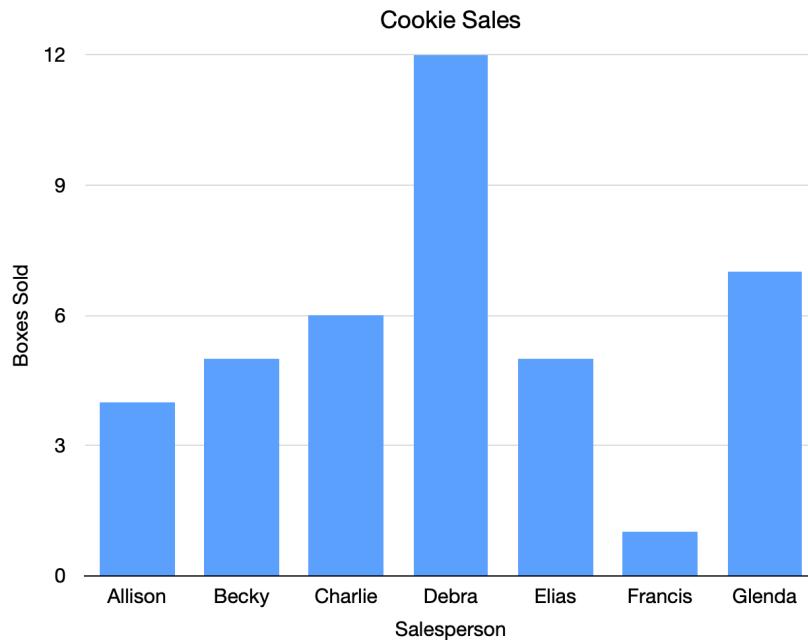
It is difficult for the human mind to look at a list of numbers and identify the patterns in them, so we often make pictures with the numbers. These pictures are called *graphs*, *charts*, or *plots*. Often the right picture can make the meaning in the data obvious. *Data visualization* is the process of making pictures from numbers.

38.1 Common Types of Data Visualizations

Depending on the type of data and what you are trying to demonstrate about it, you will use different types of data visualizations. How many types of data visualizations are there? Hundreds, but we will concentrate on just four: The bar chart, the line graph, the pie chart, and the scatter plot.

38.1.1 Bar Chart

Here is an example of a bar chart.



Each bar represents the cookie sales of one person. For example, Charlie has sold 6 boxes of cookies, so the bar goes over Charlie's name and reaches to the number 6.

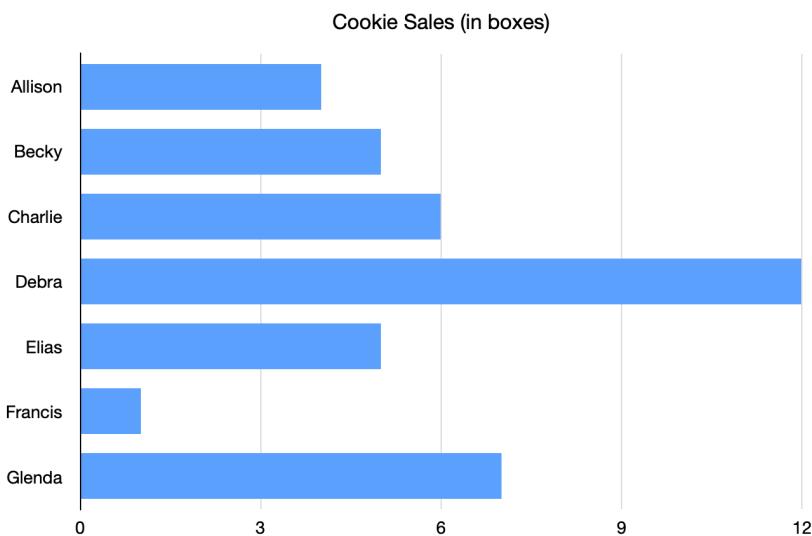
Looking at this chart, you probably think, “Wow, Debra has sold a lot more cookies than anyone else, and Francis has sold a lot fewer.”

The same data could be in a table like this:

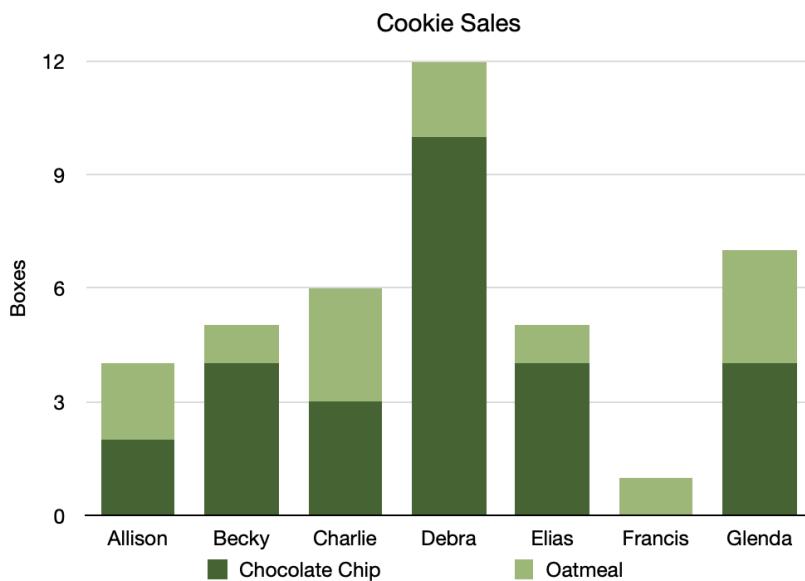
Salesperson	Boxes Sold
Allison	4
Becky	5
Charlie	6
Debra	12
Elias	5
Francis	1
Glenda	7

The table (especially a large table) is often just a bunch of numbers. A chart helps our brains understand what the numbers mean.

Bar charts can also go horizontally.



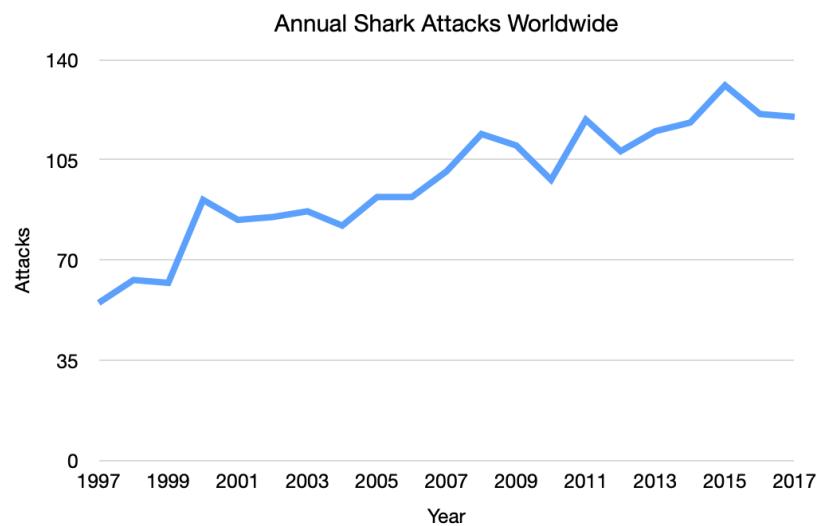
Sometimes we use colors to explain what contributed to the number.



This tells us that Becky sold more boxes of chocolate chip cookies than boxes of oatmeal cookies.

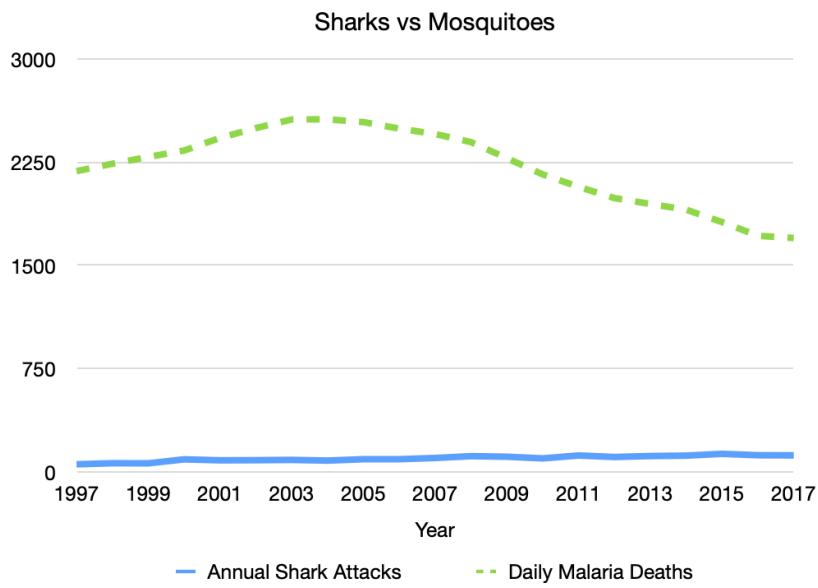
38.1.2 Line Graph

Here is a line graph.



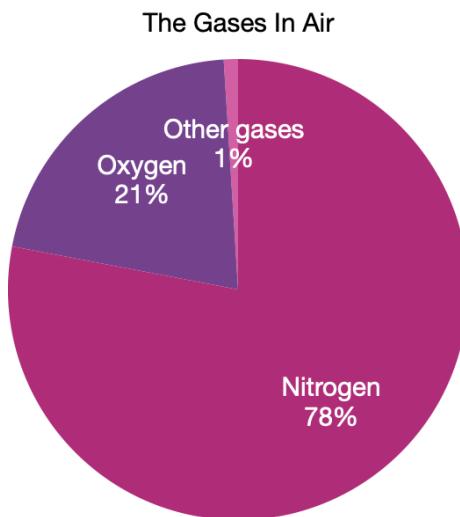
These are often used to show trends over time. Here, for example, you can see that the number of shark attacks has been increasing over time.

You can have more than one line on a graph.



38.1.3 Pie Chart

You use a pie chart when you are looking at the comparative size of numbers.



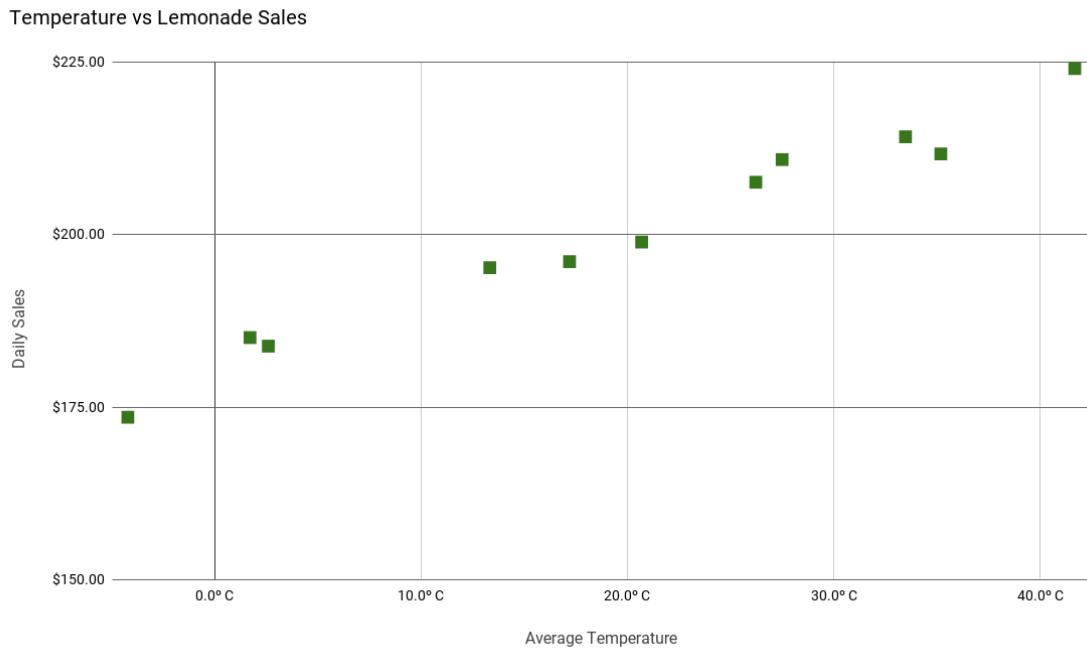
38.1.4 Scatter Plot

Sometimes you have a bunch of data points with two values and you are looking for a relationship between them. For example, maybe you write down the average temperature and the total sales for your lemonade stand on the 15th of every month:

Date	Avg. Temp.	Total Sales
15 January 2022	2.6° C	\$183.85
15 February 2022	-4.2° C	\$173.56
15 March 2022	13.3° C	\$195.22
15 April 2022	26.2° C	\$207.61
15 May 2022	27.5° C	\$210.88
15 June 2022	31.3° C	\$214.18
15 July 2022	33.5° C	\$215.23
15 Aug 2022	41.7° C	\$224.07
15 September 2022	20.7° C	\$198.94
15 October 2022	17.2° C	\$196.10
15 November 2022	1.7° C	\$185.10
15 December 2022	0.2° C	\$188.70

And you think “I wonder if I sell more lemonade on hotter days?”

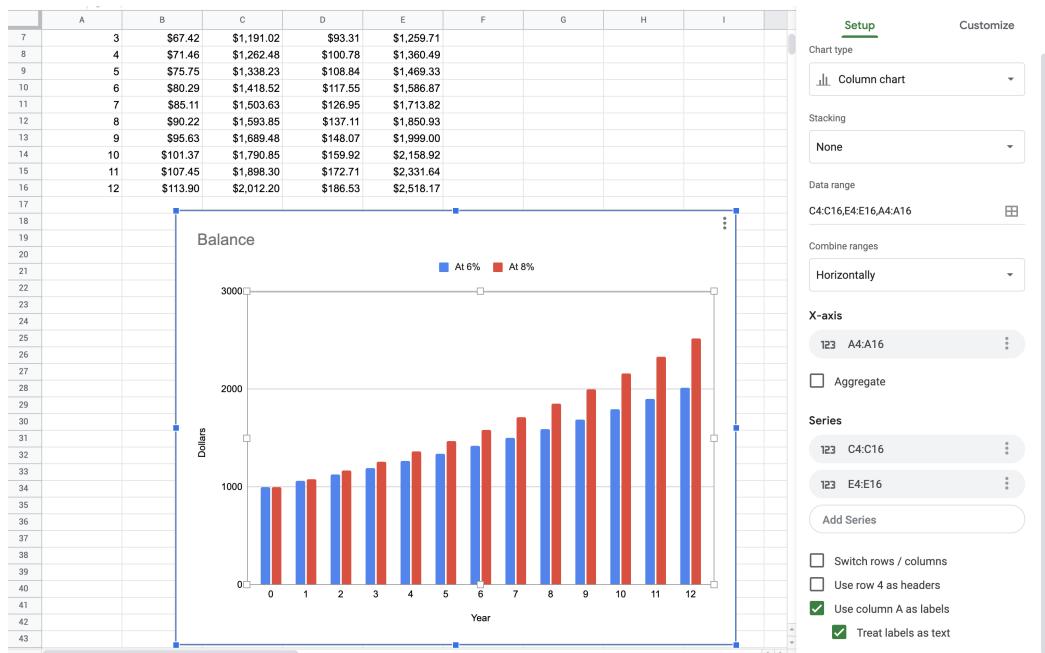
You might create a scatter plot. For each day, you put a mark that represents that temperature and the sales that day:



From this scatter plot, you can easily see that you do sell more lemonade as the temperature goes up.

38.2 Make Bar Graph

Go back to your compound interest spreadsheet and make a bar graph that shows both balances over time:



The year column should be used as the x-axis. There are two series of data that come from C4:C16 and E4:E16. Tidy up the titles and legend as much as you like.

Looking at the graph, you can see the balances start the same, but balance of the account with the larger interest rate quickly pulls away from the account with the smaller interest rate.



CHAPTER 39

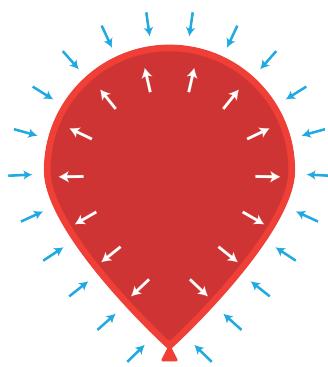
Atmospheric Pressure

The air you breathe is a blend of gases:

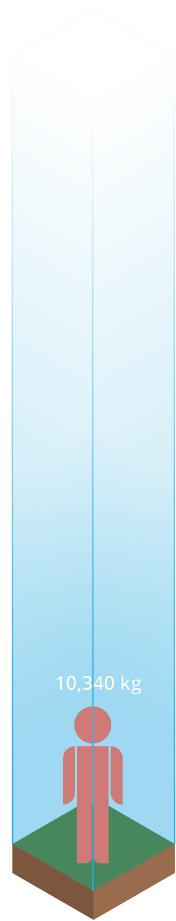
1. 78% nitrogen in the form of N_2
2. 21% oxygen in the form O_2
3. 1% other gases (mostly argon)

If you fill a balloon with helium (He), the helium will push against the interior of the balloon with some pressure. The pressure is the same at every point in the interior of the balloon. Pressure, then, is force spread over some area. Force is commonly measured in newtons. Pressure is measured in *pascals*. A pascal is 1 newton per square meter.

We don't usually think about it, but the air outside the balloon is also pushing against the exterior of the balloon. We call this *barometric pressure* or *atmospheric pressure* and it is caused by gravity pulling on the gas molecules above the balloon.



Imagine a square meter on the ground at sea level. Now imagine the column of air above it – reaching all the way to the top of the atmosphere.



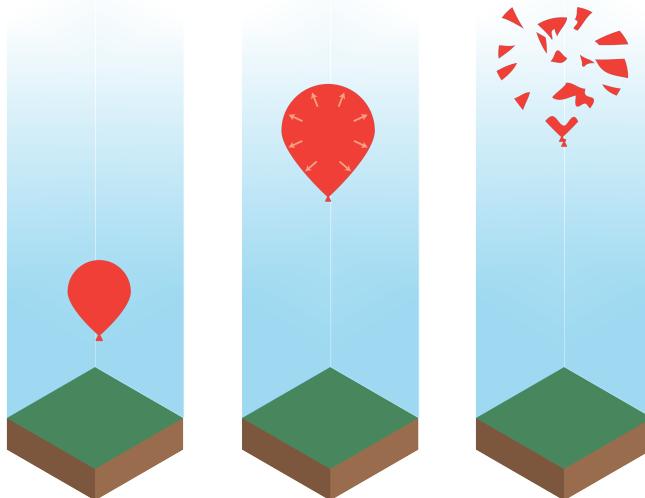
The air inside that column has a mass of about 10,340 kg. One kilogram on the earth experiences a gravitational force of 9.8 N. So the atmospheric pressure all around you is about 101,332 pascals. When dealing with such large numbers, we often use kilopascals. We'd say the barometric pressure at sea level is about 101.3 kPa.

That's a lot of pressure! Why doesn't your ribcage collapse crushing your lungs? The air *inside* your lungs is the same pressure as the air push on the outside of your rib cage.

And thus we live pretty much oblivious to this huge force that is all around us, but you can see it sometimes. For example, if you suck the air out of a plastic bottle, the bottle will be crushed by the barometric pressure.

39.1 Altitude and Atmospheric Pressure

If you let go of the balloon, as it rises through this column there will be less and less air mass above it, and thus less and less atmospheric pressure on the outside of the balloon.



What would be the atmospheric pressure at h meters above sea level? Here is a handy formula for that:

$$p = 101,332 \times \left(1 - \left(2.25577 \times 10^{-5} \times h\right)\right)^{5.25588}$$

where p is the atmospheric pressure in pascals.

Exercise 53 Atmospheric Pressure*Working Space*

You are thinking about riding your bicycle to the top of Mount Everest. You are worried when the atmospheric pressure outside the tire drops, the tire will fail. (I have had a tire fail before; It is very, very loud.)

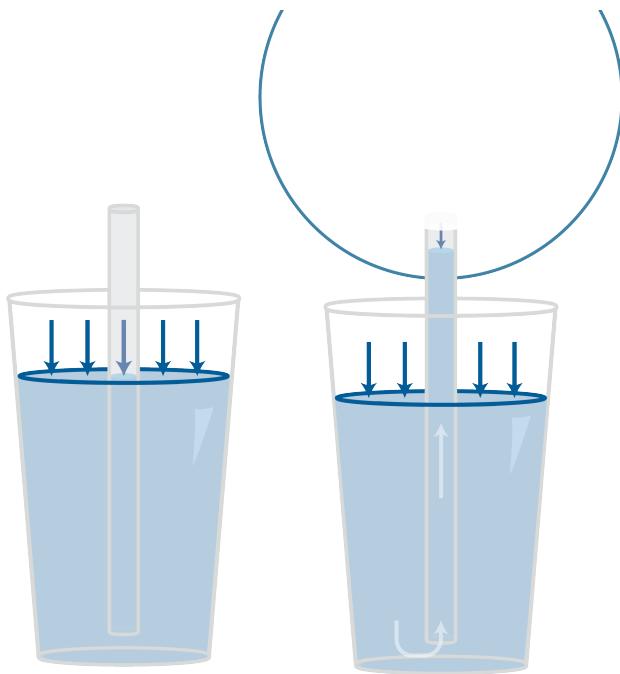
Calculate the atmospheric pressure at the top of Mount Everest (9,144 meters above sea level).

*Answer on Page 826***39.2 How a Drinking Straw Works**

When you suck on a drinking straw, why does the beverage rise? It is actually pushed by atmospheric pressure.

Before you put your mouth on the straw, the atmospheric pressure is pressing on the entire surface of the liquid (even inside the straw) evenly. Gravity pulls on the liquid making the surface level.

When you suck some air out of the straw, the pressure on the surface inside the straw drops. The atmospheric pressure on the surface outside the straw pushes into the straw and the beverage rises.

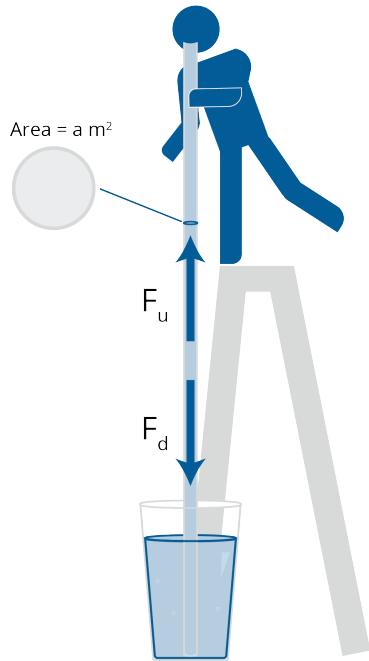


Of course, gravity is still trying to pull the liquid inside the straw back down. And for every inch that you lift the liquid in the straw, the force of gravity gets greater, demanding more suction.

39.2.1 The Longest Usable Straw

Assuming you are drinking water in a place with 100 kPa of atmospheric pressure, how high could you suck water with a perfect vacuum? That is, given a very, very long and very, very stiff drinking straw, if you created a pressure of 0 Pa inside, how far above the surface of the glass could you get the water?

Let's say a cross-section of the straw has an area of a square meters and the very top of the column of water is h meters above the surface in the glass.



With how many newtons of force is the atmosphere pushing the water upward? 100 kPa = 100,000 newtons per square meter. So:

$$F_u = (100,000)a$$

With many newtons of force is gravity pulling the water in the straw downward? The volume of the water is ah. A cubic meter of liquid water weights 1000 kg. The force of gravity is 9.8 Newtons per kg.

$$F_d = (ah)(1000)(9.8)$$

The water will stop rising when F_u = F_d. So to find h we substitute in:

$$(100,000)a = (ah)(1000)(9.8)$$

Notice that we can divide both sides by a getting:

$$h = \frac{100,000}{9,800} = 10.2 \text{ meters}$$

A perfect vacuum would only be able to drag the water up 10.2 meters.

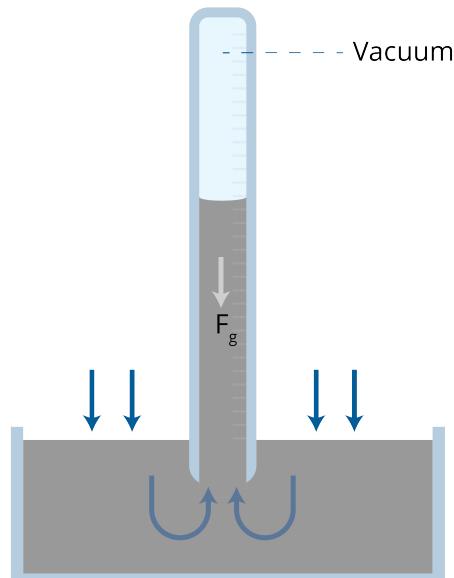
39.2.2 Millimeters Mercury

The density of mercury is 13,500 kg per cubic meter. How far up a straw would a perfect vacuum pull mercury?

$$h = \frac{100,000}{(9.8)(13,500)} \approx 0.756 \text{ meters}$$

That is, when the atmospheric pressure is 100kPa, the mercury will rise 756 mm into a vacuum.

This is actually how scientists measure atmospheric pressure. They have a long glass tube filled with mercury. One end is closed off and pointed into the sky (exactly opposite the direction of gravity). The other end is placed into a dish of mercury. There are millimeter marks on the glass tube.



We use fluctuations in the atmospheric pressure to help us predict the weather. You might hear a weather nerd with a barometer in his house say, "Wow, the barometer has gone from 752 to 761 millimeters mercury in the last hour. A high-pressure system is moving

in."

39.3 How Siphon Works

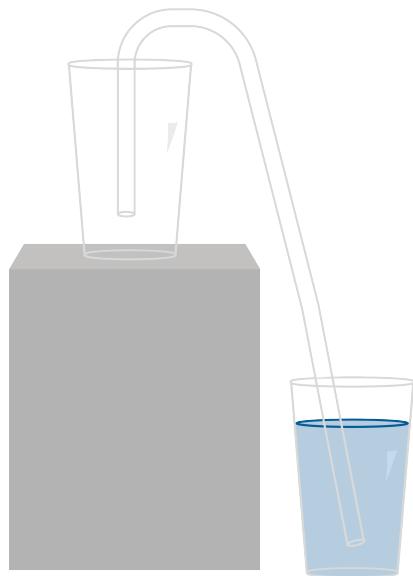
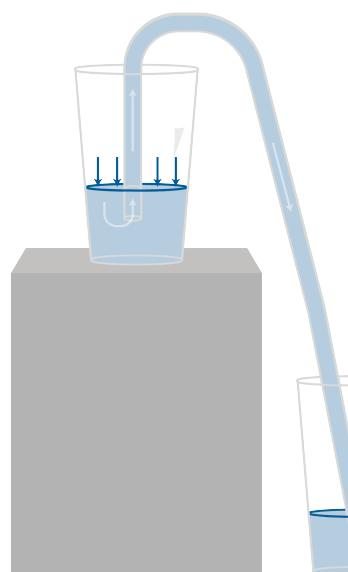
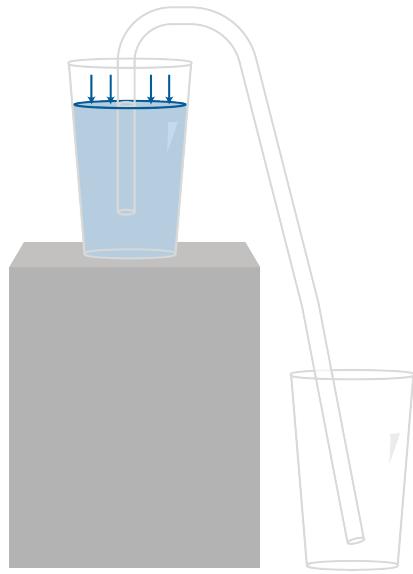
Let's say you had two cups on a table: One is filled with water, one is empty. And you connected them with an empty U-shaped straw. Water will not crawl up the empty straw: the pressure on each end of the straw is the same, and crawling the straw (against gravity) would require energy.

But, what if the straw were filled with water? Then the force of gravity is pulling the water on each side of the hump in different directions. However, the side going into the empty glass pulls a little harder. That is sufficient to create enough suction to pull the water in the other side up and over the hump.

And that will pull more water. Water will continue to flow from the full glass to the empty one until their surfaces are at the same level. At this point, the pull of gravity is the same on each side of the tube.

This is known as a *siphon*. Notice that atmospheric pressure makes the siphon possible: when the water on one side is pulled down by gravity, the atmospheric pressure pushes the other side up. If you were on a planet with plenty of gravity, but no atmosphere, a siphon wouldn't work.

A siphon is really useful when you want to get liquid out of a container that is too big to pour. For example, if you wanted to take the gasoline out of a car, you could use any flexible tubing to make a siphon. You would put the hose in the gas tank, suck enough gas up into the hose to get the siphon going, and then put the hose into your jug. (If you ever do this, be really careful not to suck any of the gasoline into your mouth: ingesting even a little bit of gasoline can make you really sick or even kill you.)



There are two rules to siphons:

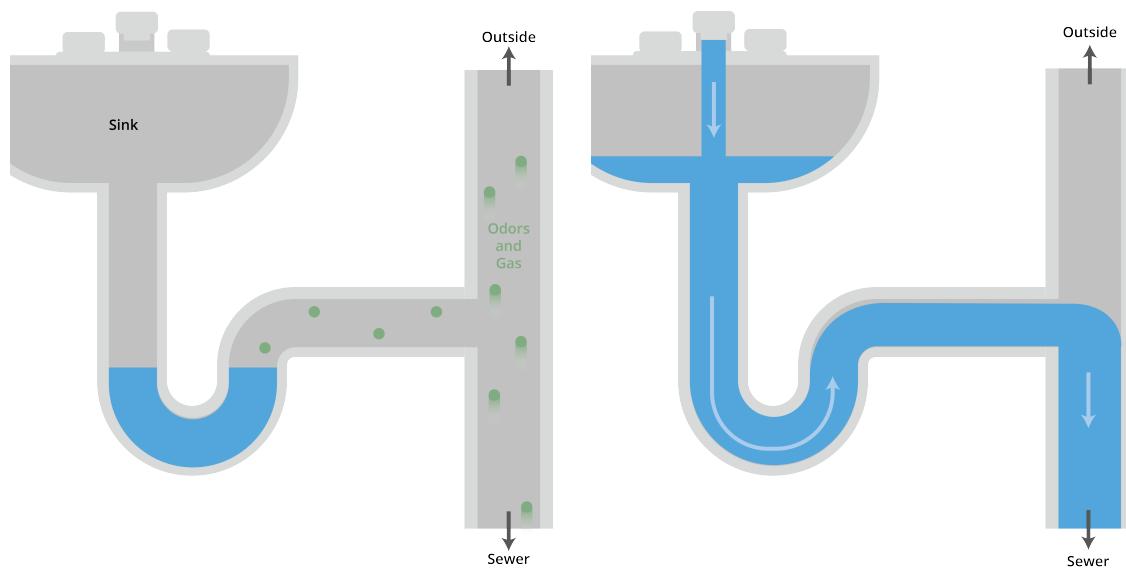
- The peak of the siphon has to be low enough for atmospheric pressure to push the liquid that high. For water at sea-level, for example, the peak of the siphon can't be more than 10.2 meters above the surface of the source liquid.
- The tube has to carry the liquid to a lower level than the surface of the source liquid. If the destination end of the siphon is submerged, the surface of the liquid it is submerged in has to be lower than the surface of the source liquid. If the

destination end of the siphon is not submerged, its opening has to be lower than the surface of the source liquid.

As long as you follow these two rules, your siphons can be very creative. For example, every toilet has a siphon in it.

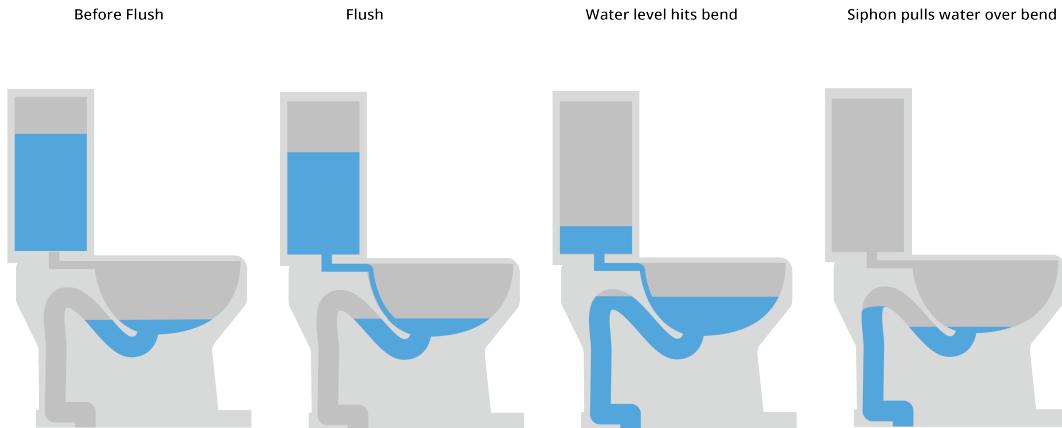
39.4 How a Toilet Works

Before we talk about toilets, you should know about P-traps. The drain from every sink, shower, and toilet in your house curves up and then down. This is known as a *P-trap*. The P-trap should always have some water in it. That keeps stinky (and flammable!) gases in the sewer from coming up and into your house.



If one of your fixtures (especially one that hasn't been used in a while) smells like raw sewage, run some water to ensure that the P-trap is full.

Now the toilet: The drain in the bottom of the toilet is connected to a siphon into the sewer. The siphon is filled with air most of the time. However, when you flush the toilet, water rushes from the tank, into the bowl, and it fills the siphon. Once the siphon is filled, it pulls water out of the toilet until air starts to enter the siphon. At that point, the water stops flowing.



The toilet tank is pretty simple: it has a float and valve that opens with the float is too low. So, anytime the water-level is too low, the value is open and slowly filling the tank. So the tank is nearly always filled with a precise amount of water.

When you flush, a small door in the bottom of the tank is opened and the water rushes into the bowl. When the water is out, the door closes again so the tank can refill.



CHAPTER 40

Exponents

Let's quickly review exponents. Ancient scientists started coming up with a lot of formulas that involved multiplying the same number several times. For example, if they knew that a sphere was r centimeters in radius, its volume in milliliters was

$$V = \frac{4}{3} \times \pi \times r \times r \times r$$

They did two things to make the notation less messy. First, they decided that if two numbers were written next to each other, the reader would assume that meant "multiply them". Second, they came up with the exponent, a little number that was lifted off the baseline of the text, that meant "multiply it by itself". For example 5^3 was the same as $5 \times 5 \times 5$.

Now the formula for the volume of a sphere is written

$$V = \frac{4}{3}\pi r^3$$

Tidy, right? In an exponent expression like this, we say that r is *the base* and 3 is *the exponent*.

40.1 Identities for Exponents

What about exponents of exponents? What is $(5^3)^2$?

$$(5^3)^2 = (5 \times 5 \times 5)^2 = (5 \times 5 \times 5)(5 \times 5 \times 5) = 5^6$$

In general, for any a , b , and c :

$$(a^b)^c = a^{(bc)}$$

If you have $(5^3)(5^4)$ that is just $5 \times 5 \times 5 \times 5 \times 5 \times 5 \times 5$ or 5^7

The general rule is, for any a , b , and c

$$(a^b)(a^c) = a^{(b+c)}$$

Mathematicians *love* this rule, so we keep extending the idea of exponents to keep this rule true. For example, at some point, someone asked “What about 5^0 ?” According to the rule, 5^2 must equal $5^{(2+0)}$ which must equal $(5^2)(5^0)$. Thus, 5^2 must be 1. So mathematicians declared “Anything to the power of 0 is 1”.

We don’t typically assume that $0^0 = 1$. It is just too weird. So we say, that for any a not equal to zero,

$$a^0 = 1$$

What about $5^{(-2)}$? By our beloved rule, we know that $(5^{-2})(5^5)$ must be equal to 5^3 , right? So 5^{-2} must be equal to $\frac{1}{5^2}$.

We say, for any a not equal to zero and any b ,

$$a^{-b} = \frac{1}{a^b}$$

This makes dividing one exponential expression by another (with the same base) easy:

$$\frac{a^b}{a^c} = a^{(b-c)}$$

We often say “cancel out” for this. Here I can “cancel out” x^2 :

$$\frac{x^5}{x^2} = x^3$$

What about $5^{\frac{1}{3}}$? By the beloved rule, we know that $5^{\frac{1}{3}}5^{\frac{1}{3}}5^{\frac{1}{3}}$ must equal 5^1 . Thus $5^{\frac{1}{3}} = \sqrt[3]{5}$.

We say, for any a and b not equal to zero and any c greater than zero,

$$a^{\frac{b}{c}} = a^b \sqrt[c]{a}$$

Before you go on to the exercises, note that the beloved rule demands a common base.

- We can combine these: $(5^2)(5^4) = 5^6$
- We cannot combine: $(5^2)(3^5)$

With that said, we note for any a, b , and c :

$$(ab)^c = (a^c)(b^c)$$

So, for example, if I were asked to simplify $(3^4)(6^2)$, I would note that $6 = 2 \times 3$, so

$$(3^4)(6^2) = (3^4)(3^2)(2^2) = (3^6)(2^2)$$

If these ideas are new to you (or maybe they have been forgotten), watch the Khan Academy’s **Intro to rational exponents** video at <https://youtu.be/lZfXc4nHooo>.



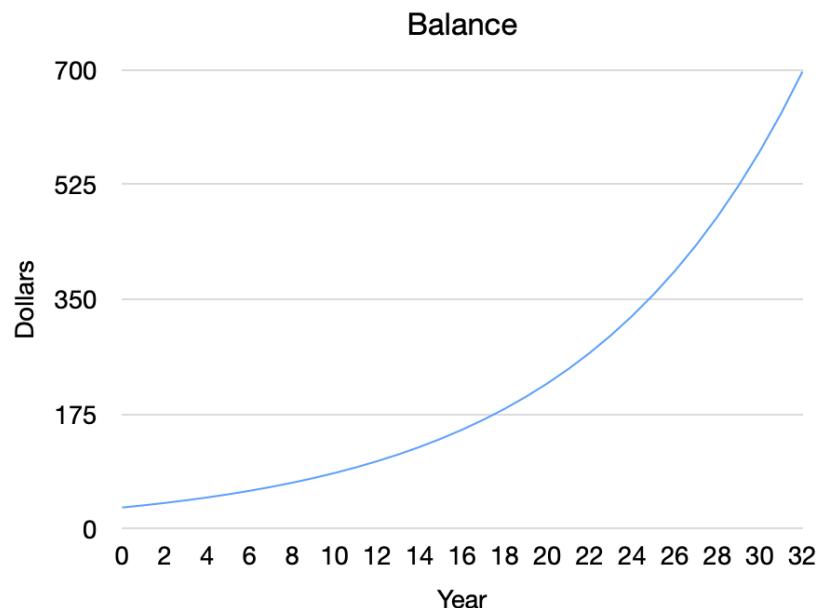
CHAPTER 41

Exponential Decay

In a previous chapter, we saw that an investment of P getting compound interest with an annual interest rate of r , grows exponentially. At the end of year t , your balance would be

$$P(1 + r)^t$$

Because r is positive, this number grows as time passes. You get a nice exponential growth curve that looks something like this:



This is \$30 invested with a 10% annual interest rate. So the formula for the balance after t years would be

$$(30)(1.1)^t$$

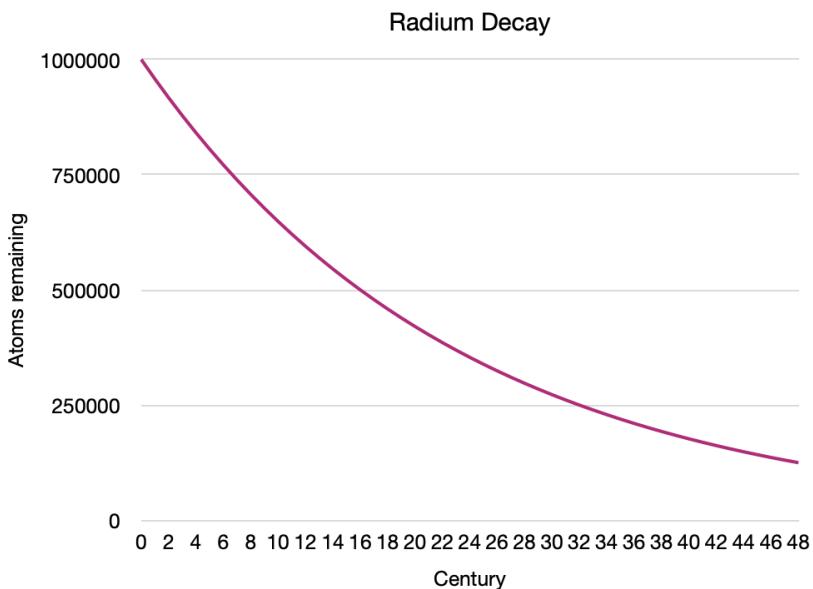
What if r were negative? This would be *exponential decay*.

41.1 Radioactive Decay

Until around 1970, there were companies making watches whose faces and hands were coated with radioactive paint. The paint usually contained radium. When a radium atom decays, it gives off some energy, loses two protons and two neutrons, and becomes a different element (radon). Some of the energy given off is visible light. Thus, these watches glow in the dark.

How many of the radium atoms in the paint decay each century? About 4.24%.

Notice the quantity of atoms lost is proportional to the number of atoms you have. This is exponential decay. If we assume that we start with a million radium atoms, the number of atoms decreases over time like this:

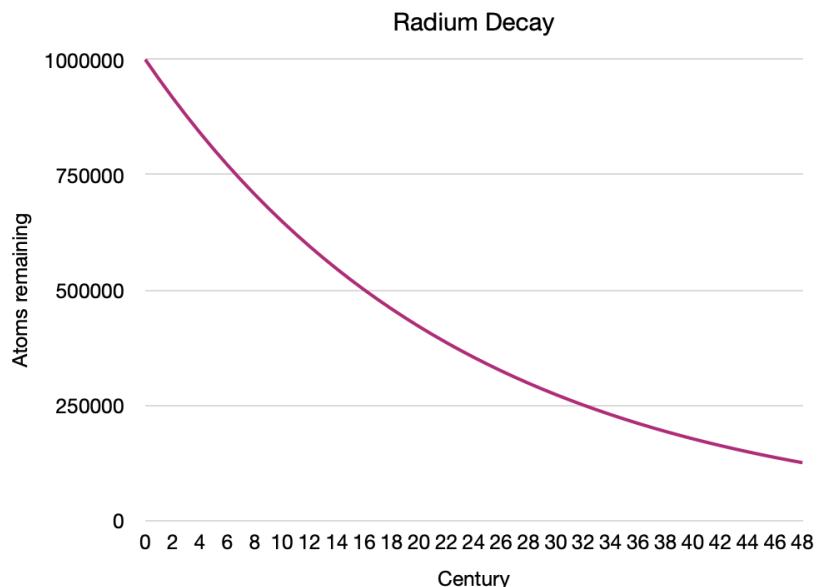


- We start with 1,000,000 atoms.
- At 16 centuries, we have only 500,000 (half as many) left.
- 16 centuries after that, we have only 250,000 (half again) left.
- 16 centuries after that, we have only 125,000 (half again) left.

A nuclear chemist would say that radium has a *half-life* of 1,600 years. Note that this means that if you bought a watch with glowing hands in 1960, it will be glowing half as brightly in the year 3560.

How do we calculate the amount of radium left at the end of century t ? If you start with P atoms, at the end of the t -th century you will have

$$P(1 - 0.0424)^t$$



This is exponential decay.

41.2 Model Exponential Decay

Let's say you get hired to run a company with 480,000 employees. Each year $1/8$ of your employees leave the company for some reason (retirement, quitting, etc.). For some reason, you never hire any new employees.

Make a spreadsheet that indicates how many of the original 480,000 employees will still be around at the end of each year for the next 12. Then make a bar graph from that data.



CHAPTER 42

Logarithms

After the world had created exponents, it needed the opposite. We could talk about the quantity $? = 2^3$, that is, “What is the product of 2 multiplied by itself three times?” We needed some way to talk about $2^? = 8$, that is “2 to the what is 8?” So we developed the logarithm.

Here is an example:

$$\log_2 8 = 3$$

In English, you would say “The logarithm base 2 of 8 is 3.”

The base (2, in this case) can be any positive number. The argument (8, in this case) can also be any positive number.

Try this one: What is the logarithm base 2 of 1/16?

You know that $2^{-4} = \frac{1}{16}$, so $\log_2 \frac{1}{16} = -4$.

42.1 Logarithms in Python

Most calculators have pretty limited logarithm capabilities, but python has a nice `log` function that lets you specify both the argument and the base. Start python, import the `math` module, and try taking a few logarithms:

```
>>> import math  
>>> math.log(8,2)  
3.0  
>>> math.log(1/16, 2)  
-4.0
```

Let's say that a friend offers you 5% interest per year on your investment for as long as you want. And you wonder, "How many years before my investment is 100 times as large?" You can solve this problem with logarithms:

```
>>> math.log(100, 1.05)  
94.3872656381287
```

If you leave your investment with your friend for 94.4 years, the investment will be worth 100 times what you put in.

42.2 Logarithm Identities

The logarithm is defined this way:

$$\log_b a = c \iff b^c = a$$

Notice that the logarithm of 1 is always zero, and $\log_b b = 1$.

The logarithm of a product:

$$\log_b ac = \log_b a + \log_b c$$

This follows from the fact that $b^{a+c} = b^a b^c$. What about a quotient?

$$\log_b \frac{a}{c} = \log_b a - \log_b c$$

Exponents?

$$\log_b(a^c) = c \log_b a$$

Notice that because logs and exponents are the opposite of each other, they can cancel each other out:

$$b^{\log_b a} = a$$

and

$$\log_b(b^a) = a$$

42.3 Changing Bases

I mentioned that most calculators have pretty limited logarithm capabilities. Most calculators don't allow you to specify what base you want to work with. All scientific calculators have a button for "log base 10". So you need to know how to use that button to get logarithms for other bases. Here is the change-of-base identity:

$$\log_b a = \frac{\log_c a}{\log_c b}$$

So, for example, if you wanted to find $\log_2 8$, you would ask the calculator for $\log_{10} 8$ and then divide that by $\log_{10} 2$. You should get 3.

42.4 Natural Logarithm

When you learn about circles, you are told that the circumference of a circle is about 3.141592653589793 times its diameter. Because we use this unwieldy number a lot, we give it a name: We say "The circumference of a circle is π times its diameter."

There is a second unwieldy number that we will eventually use a lot in solving problems. It is about 2.718281828459045 (but the digits actually go on forever, just like π). We call this number e . (I'm not going to tell you why e is special now, but soon...)

Most calculators have a button labeled "ln". That is the *natural logarithm* button. It takes the log in base e .

Similarly, in python, if you don't specify a base, the logarithm is done in base e :

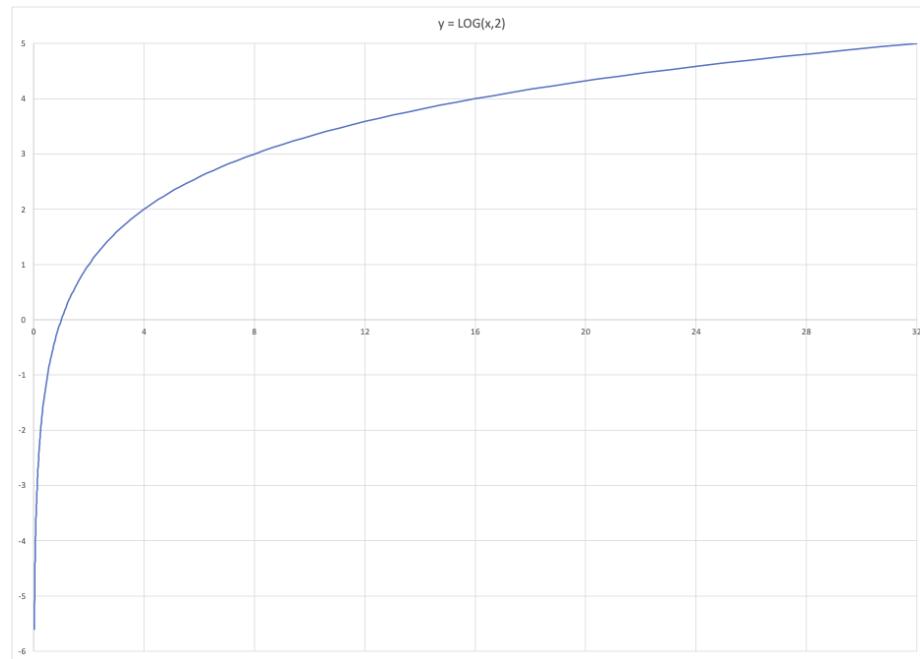
```
>>> math.log(10)
2.302585092994046
>>> math.log(math.e)
1.0
```

42.5 Logarithms in Spreadsheets

Spreadsheets have three log functions:

- LOG takes both the argument and the base. LOG(8,2) returns 3.
- LOG10 takes just the argument and uses 10 as the base.
- LN takes just the argument and uses e as the base.

Here is a plot from a spreadsheet of a graph of $y = \text{LOG}(x, 2)$.



Spreadsheets also have the function EXP(x) which returns e^x . For example, EXP(2) returns 7.38905609893065.

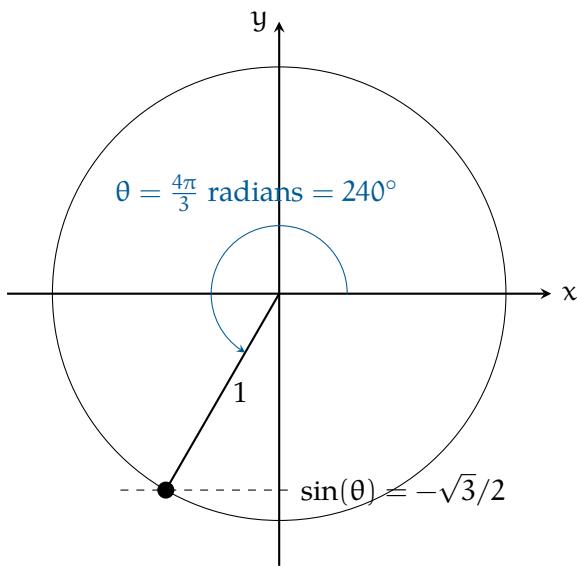


CHAPTER 43

Trigometric Functions

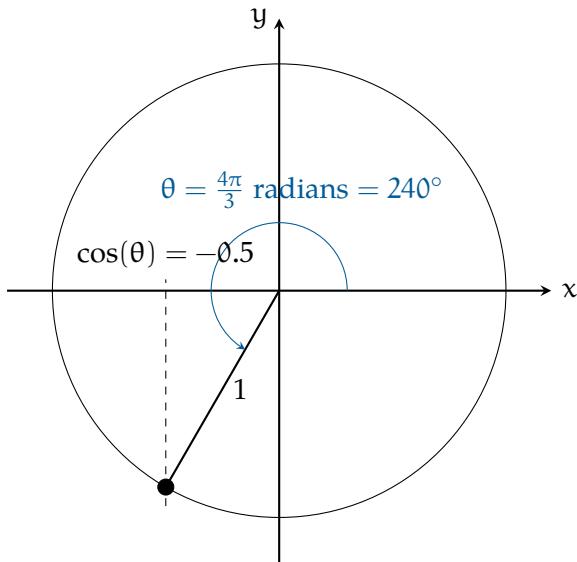
As mentioned earlier, in a right triangle where one angle is θ , the sine of θ is the length of the side opposite θ divided by the length of the hypotenuse.

The sine function is defined for any real number. We treat that real number θ as an angle, we draw a ray from the origin out to the unit circle. The y value of that point is the sine. So, for example, the $\sin(\frac{4\pi}{3})$ is $-\sqrt{3}/2$



(Note that in this section, we will be using radians instead of degrees unless otherwise noted. While degrees are more familiar to most people, engineers and mathematicians nearly always use radians when solving problems. Your calculator should have a radians mode and a degrees mode. You want to be in radians mode.)

Similarly, we define cosine using the unit circle: to find the cosine of θ , we draw a ray from the origin at the angle θ . The x component of the point where the ray intersects the unit circle is the cosine of θ .



From this description, it is easy to see why $\sin(\theta)^2 + \cos(\theta)^2 = 1$. They are the legs of a right triangle with a hypotenuse of length 1.

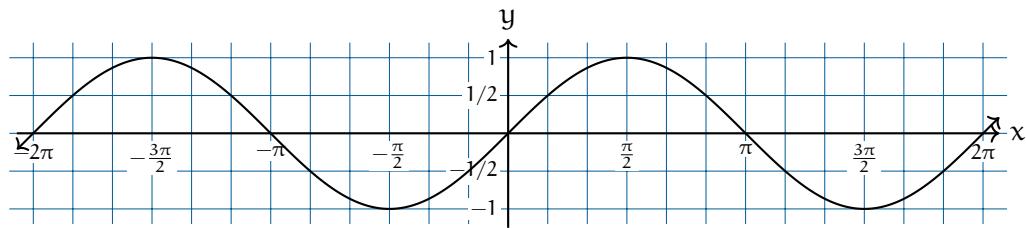
It should also be easy to see why $\sin(\theta) = \sin(\theta + 2\pi)$: Each time you go around the circle,

you come back to where you started.

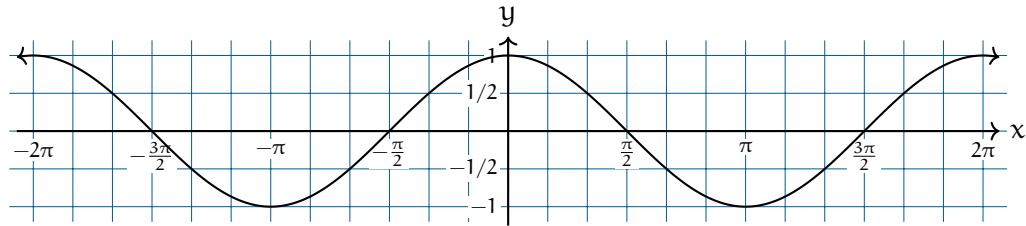
Can you see why $\cos(\theta) = \sin(\theta + \pi/2)$? Turn the picture sideways.

43.1 Graphs of sine and cosine

Here is a graph of $y = \sin(x)$:



It looks like waves, right? It goes forever to the left and right. Remembering that $\cos(\theta) = \sin(\theta + \pi/2)$, we can guess what the graph of $y = \cos(x)$ looks like:



43.2 Plot cosine in Python

Create a file called `cos.py`:

```
import numpy as np
import matplotlib.pyplot as plt

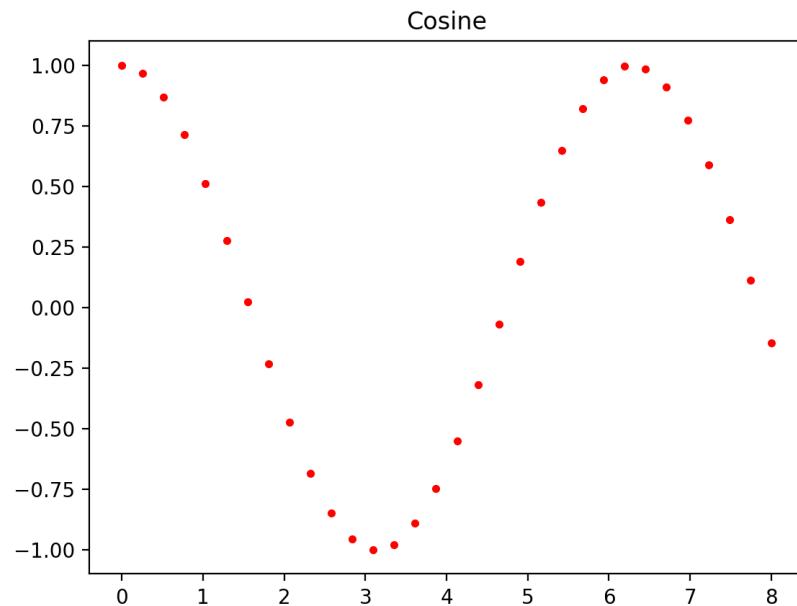
until = 8.0

# Make a plot of cosine
thetas = np.linspace(0, until, 32)
cosines = []
for theta in thetas:
    cosines.append(np.cos(theta))

# Plot the data
plt.plot(thetas, cosines)
plt.show()
```

```
fig, ax = plt.subplots()
ax.plot(thetas, cosines, 'r.', label="Cosine")
ax.set_title("Cosine")
plt.show()
```

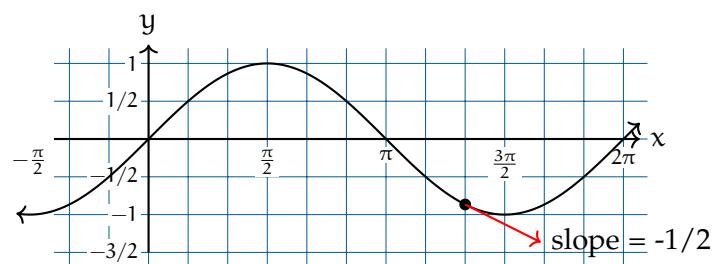
This will plot 32 points on the cosine wave between 0 and 8. When you run it, you should see something like this:



43.3 Derivatives of sine and cos

Here is a wonderful property of sine and cosine functions: At any point θ , the slope of the sine graph at θ equals $\cos(\theta)$.

For example, we know that $\sin(4\pi/3) = -(1/2)\sqrt{3}$ and $\cos(4\pi/3) = -1/2$. If we drew a line tangent to the sine curve at this point, it would have a slope of $-1/2$:



We say “The derivative of the sine function is the cosine function.”

Can you guess the derivative of the cosine function? For any θ , the slope of the graph of the $\cos(\theta)$ is $-\sin(\theta)$.

43.4 A weight on a spring

Let’s say you fill a rollerskate with heavy rocks and attach it to the wall with a stiff spring. If you push the skate toward the wall and release it, it will roll back and forth. Engineers would say “The skate will oscillate.”

Intuitively, you can probably guess:

- If the spring is stronger, the skate will oscillate more times per minute.
- If the rocks are lighter, the skate will oscillate more times per minute.

The force that the spring exerts on the skate is proportional to how far its length is from its relaxed length. When you buy a spring, the manufacturer advertises its “spring rate”, which is in pounds per inch or newtons per meter. If a spring has a rate of 5 newtons per meter, which means that if stretch or compress it 10 cm, it will push back with a force of 0.5 newtons. If you stretch or compress it 20 cm, it will push back with a force of 1 newton.

Let’s write a simulation of the skate-on-a-spring. Duplicate `cos.py`, and name the new copy `spring.py`. Add code to implement the simulation:

```
import numpy as np
import matplotlib.pyplot as plt

until = 8.0

# Constants
mass = 100 # kg
spring_constant = -1 # newtons per meter displacement
time_step = 0.01 # s

# Initial state
displacement = 1.0 # height above equilibrium in meters
velocity = 0.0
time = 0.0 # seconds

# Lists to gather data
```

```
displacements = []
times = []

# Run it for a little while
while time <= until:
    # Record data
    displacements.append(displacement)
    times.append(time)

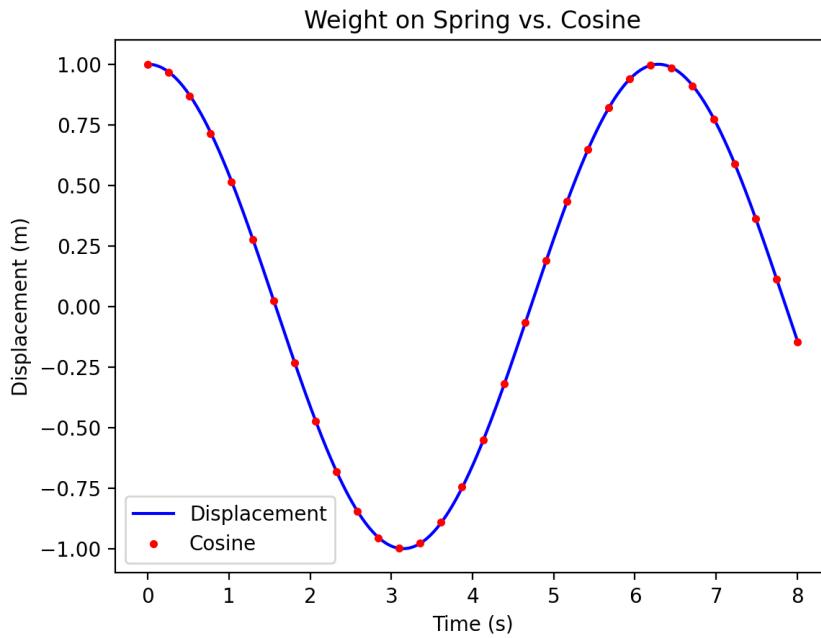
    # Calculate the next state
    time += time_step
    displacement += time_step * velocity
    force = spring_constant * displacement
    acceleration = force / mass
    velocity += acceleration

# Make a plot of cosine
thetas = np.linspace(0, until, 32)
cosines = []
for theta in thetas:
    cosines.append(np.cos(theta))

# Plot the data
fig, ax = plt.subplots()
ax.plot(times, displacements, 'b', label="Displacement")
ax.plot(thetas, cosines, 'r.', label="Cosine")

ax.set_title("Weight on Spring vs. Cosine")
ax.set_xlabel("Time (s)")
ax.set_ylabel("Displacement (m)")
ax.legend()
plt.show()
```

When you run it, you should get a plot of your spring and the cosine graph on the same plot.



The position of the skate is following a cosine curve. Why?

Because a sine or cosine waves happen whenever the acceleration of an object is proportional to -1 times its displacement. Or in symbols:

$$a \propto -p$$

where a is acceleration and p is the displacement from equilibrium.

Remember that if you take the derivative of the displacement, you get the velocity. And if you take the derivative of that, you get acceleration. So, the weight on the spring must follow a function f such that

$$f(t) \propto -f''(t)$$

Remember that the derivative of the $\sin(\theta)$ is $\cos(\theta)$.

And the derivative of the $\cos(\theta)$ is $-\sin(\theta)$

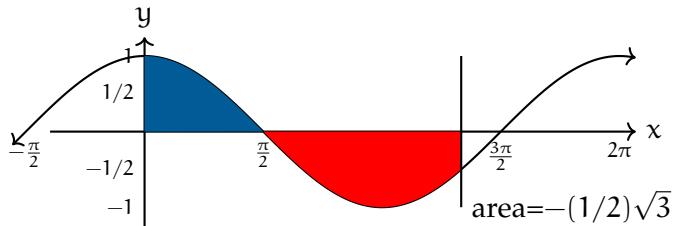
Thus these sorts of waves have an almost-magical power: their acceleration is proportional to -1 times their displacement.

Thus sine waves of various magnitudes and frequencies are ubiquitous in nature and

technology.

43.5 Integral of sine and cosine

If we take the area between the graph and the x axis of the cosine function (and if the function is below the x axis, it counts as negative area), from 0 to $4\pi/3$, we find that it is equal to $-(1/2)\sqrt{3}$



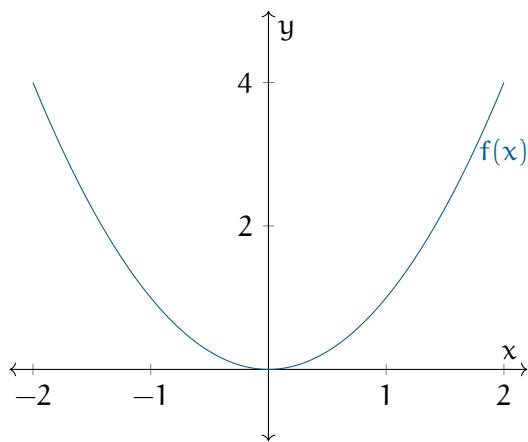
We say “The integral of the cosine function is the sine function.”



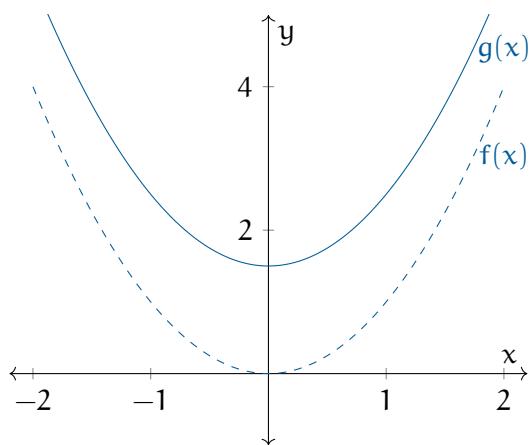
CHAPTER 44

Transforming Functions

Let's say I gave you the graph of a function f , like this:



And then I tell you that the function $g(x) = f(x) + 1.5$. Can you guess what the graph of g would look like? It is the same graph, just translated up 1.5:



There are four kinds of transformations that we do all the time:

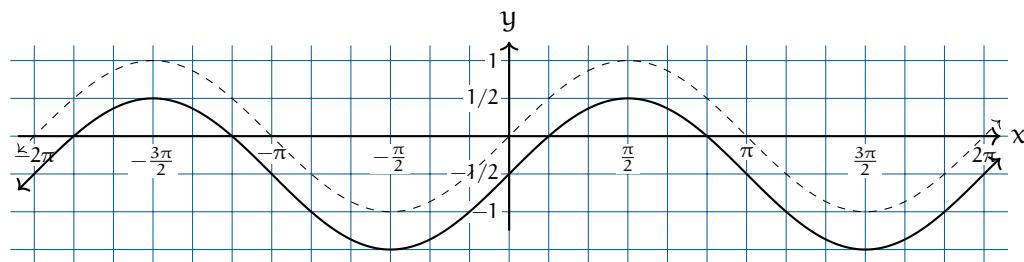
- Translation up and down in the direction of y axis (the one you just saw)
- Translation left and right in the direction of the x axis
- Scaling up and down along the y axis
- Scaling up and down along the x axis

Now I will demonstrate each of the four using the graph of $\sin(x)$.

44.1 Translation up and down

When you add a positive constant to a function, you translate the whole graph up that much. A negative constant translates it down.

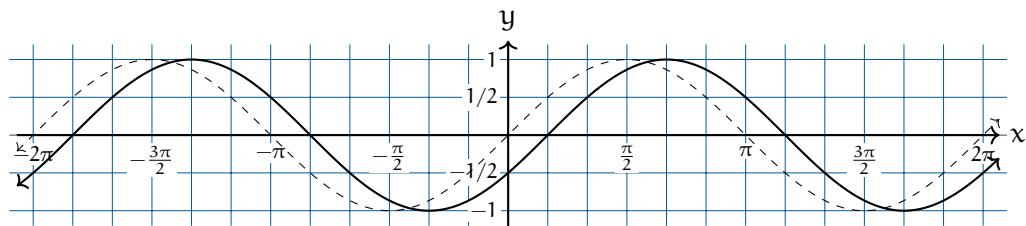
Here is the graph of $\sin(x) - 0.5$:



44.2 Translation left and right

When you add a positive number to x before running it through f , you translate the graph to the left that much. Adding a negative number translates the graph to the right.

Here is the graph of $\sin(x - \pi/6)$:



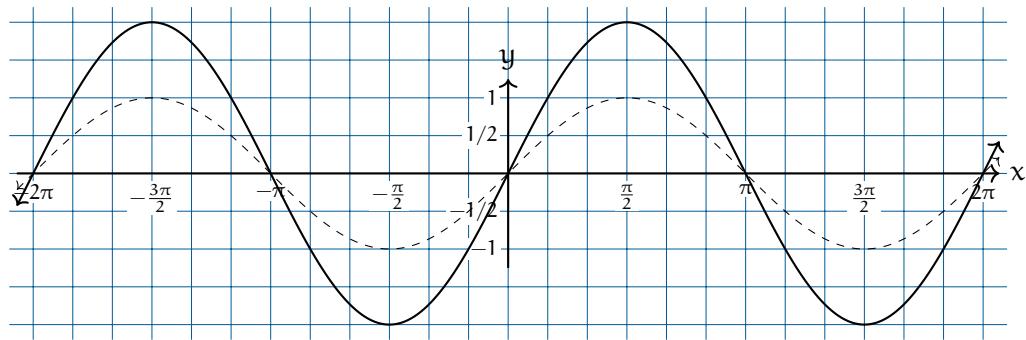
Notice the sign:

- Add to x before processing with the function translates the graph to the *left*.
- Subtract from x before processing with the function translates the graph to the *right*

44.3 Scaling up and down in the y direction

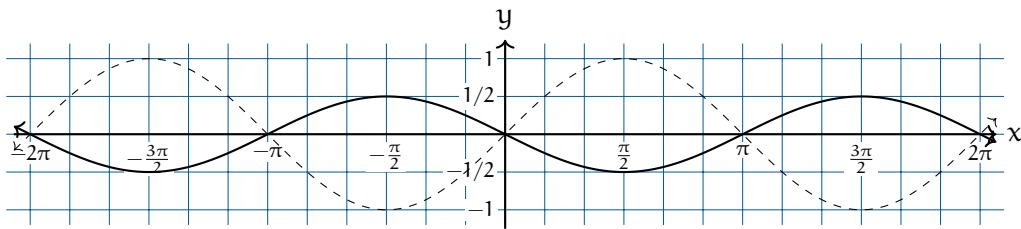
To scale the function up and down, you multiply the result of the function by a constant. If the constant is larger than 1, it stretches the function up and down.

Here is $y = 2 \sin(x)$:



With a wave like this, we speak of its *Amplitude*, which you can think of as its height. The baseline that this wave oscillates around is zero. The maximum distance that it gets from that baseline is its amplitude. Thus, the amplitude here has been increased from 1 to 2.

If you multiply by a negative number, the function gets flipped. Here is $y = -0.5 \sin(x)$

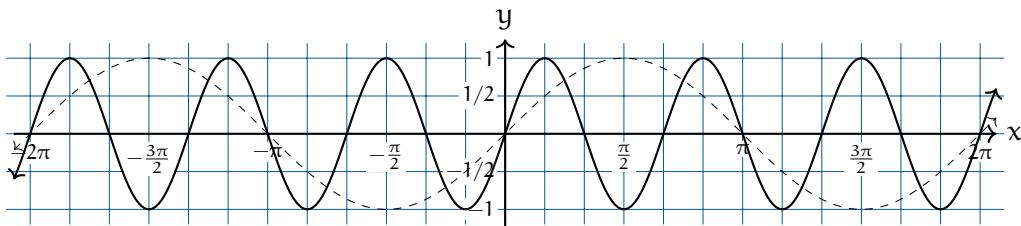


Amplitude is never negative. Thus, the amplitude of this wave is 0.5.

44.4 Scaling up and down in the x direction

If you multiply x by a number larger than 1 before running it through the function, the graph gets compressed toward zero.

Here is $y = \sin(3x)$:

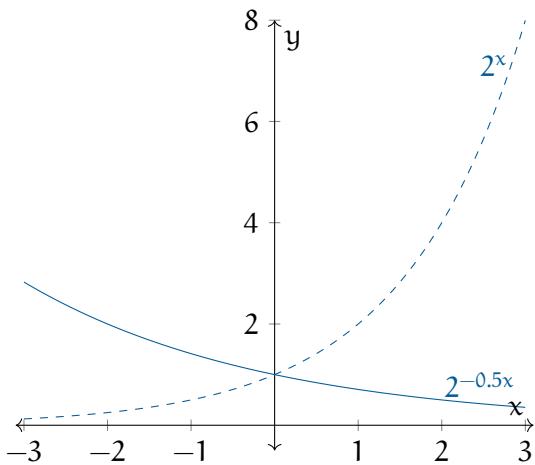


The distance between two peaks of a wave is known as its *wavelength*. The original wave had a wavelength of 2π . The compressed wave has a wavelength of $2\pi/3$.

If you multiply x by a number smaller than 1, it will stretch the function out, away from the y axis.

If you multiply x by a negative number, it will flip the function around the y axis.

Here is $y = 2^{(-0.5x)}$. Notice that it has flipped around the y axis and is stretched out along the x axis.

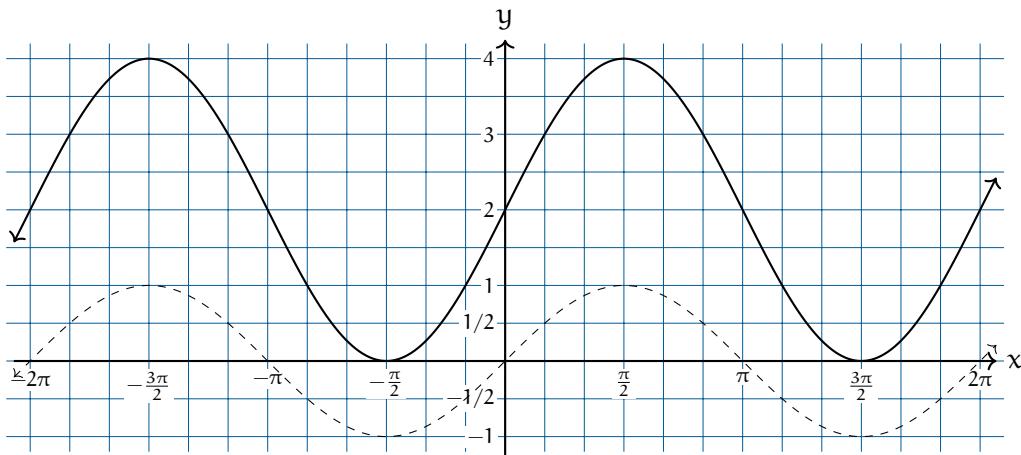


Reflection over x-axis	$(x, y) \rightarrow (x, -y)$
Reflection over y-axis	$(x, y) \rightarrow (-x, y)$
Translation	$(x, y) \rightarrow (x + a, y + b)$
Dilation	$(x, y) \rightarrow (kx, ky)$
Rotation 90° counterclockwise	$(x, y) \rightarrow (-y, x)$
Rotation 180°	$(x, y) \rightarrow (-x, -y)$

44.5 Order is important!

We can combine these transformations. This allows us, for example, to translate a function up 2 and then scale along the y axis by 3.

Here is $y = 2.0(\sin(x) + 1)$:

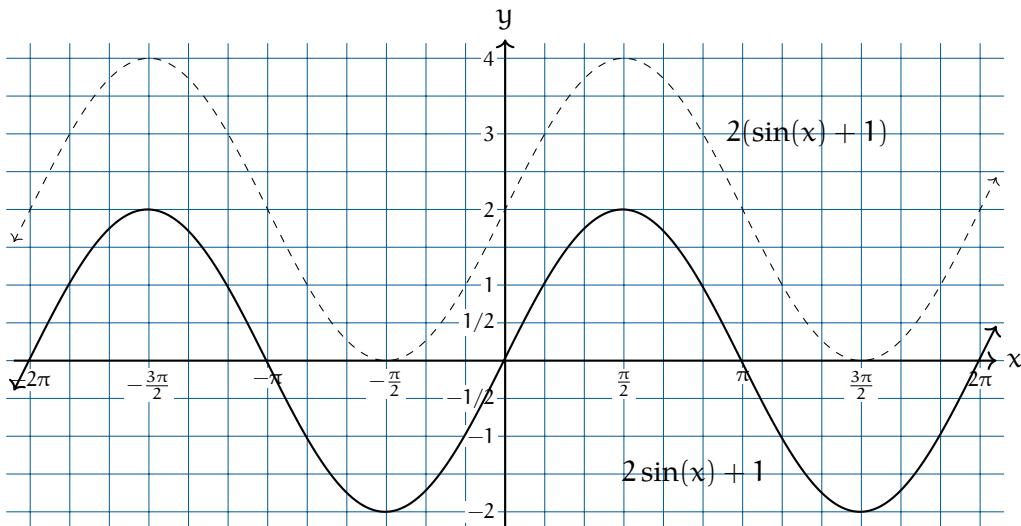


A function is often a series of steps. Here are the steps in $f(x) = 2(\sin(x) + 1)$:

1. Take the sine of x
2. Add 1 to that
3. Multiply that by 2

What if we change the order? Here are the steps in $g(x) = 2 \sin(x) + 1$:

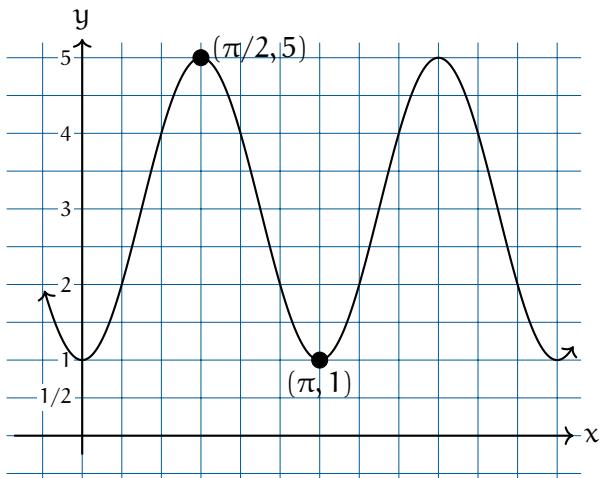
1. Take the sine of x
2. Multiply that by 2
3. Add 1 to that



The moral: You can do multiple transformations of your function, but the order in which you do them is important.

Exercise 54 **Transforms****Working Space**

Find a function that creates a sine wave such that the top of the first crest is at the point $(\frac{\pi}{2}, 5)$ and the bottom of the trough that follows is at $(\pi, 1)$.

**Answer on Page 826**



CHAPTER 45

Sound

When you set off a firecracker, it makes a sound.

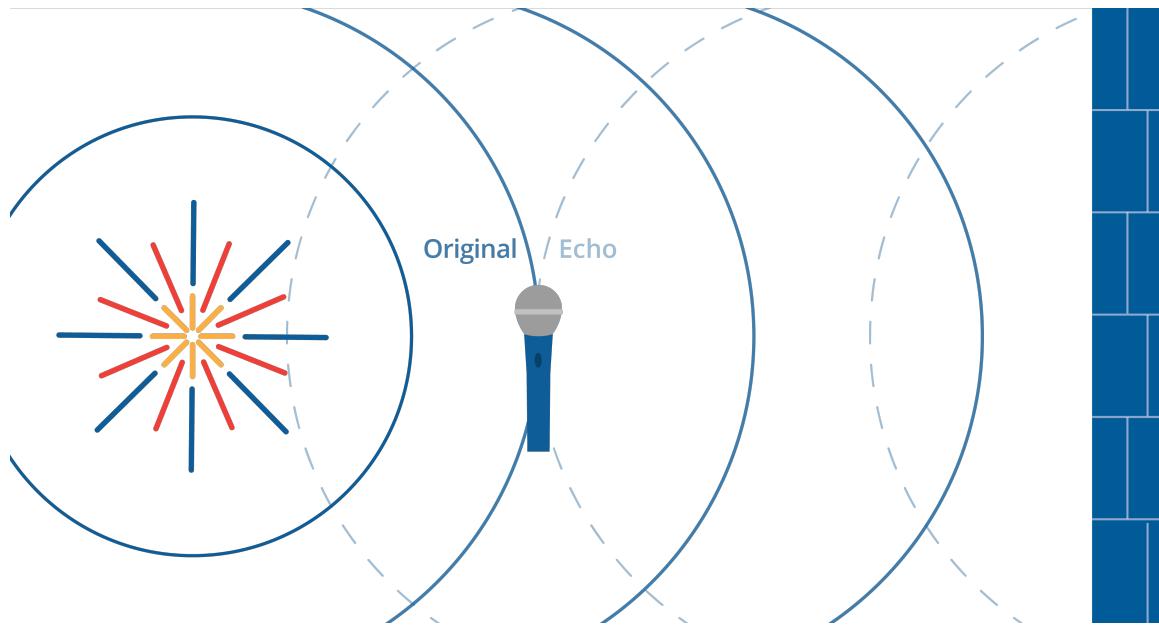
Let's break that down a little more: Inside the cardboard wrapper of the firecracker, there is potassium nitrate (KNO_3), sulfur (S), and carbon(C). These are all solids. When you trigger the chemical reactions with a little heat, these atoms rearrange themselves to be potassium carbonate (K_2CO_3), potassium sulfate (K_2SO_4), carbon dioxide (CO_2), and nitrogen (N_2). Note that the last two are gasses.

The molecules of a solid are much more tightly packed than the molecules of a gas. So after the chemical reaction, the molecules expand to fill a much bigger volume. The air molecules nearby get pushed away from the firecracker. They compress the molecules beyond them, and those compress the molecules beyond them.

This compression wave radiates out as a sphere; its radius growing at about 343 meters per second ("The speed of sound").

The energy of the explosion is distributed around the surface of this sphere. As the radius increases, the energy is spread more and more thinly around. This is why the firecracker seems louder when you are closer to it. (If you set off a firecracker in a sewer pipe, the

sound will travel much, much farther.)



This compression wave will bounce off of hard surfaces. If you set off a firecracker 50 meters from a big wall, you will hear the explosion twice. We call the second one "an echo."

The compression wave will be absorbed by soft surfaces. If you covered that wall with pillows, there would be almost no echo.

The study of how these compression waves move and bounce is called *acoustics*. Before you build a concert hall, you hire an acoustician to look at your plans and tell you how to make it sound better.

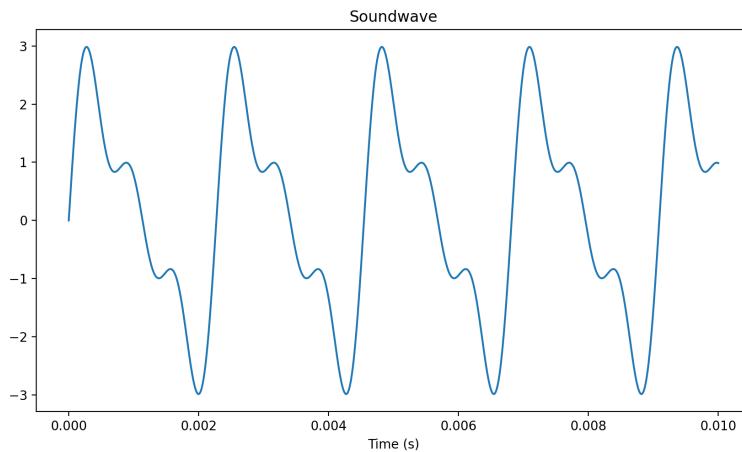
45.1 Pitch and frequency

The string on a guitar is very similar to the weighted spring example. The farther the string is displaced, the more force it feels pushing it back to equilibrium. Thus, it moves back and forth in a sine wave. (OK, it isn't a pure sine wave, but we will get to that later.)

The string is connected to the center of the boxy part of the guitar, which is pushed and pulled by the string. That creates compression waves in the air around it.

If you are in the room with the guitar, those compression waves enter your ear, push and pull your eardrum, which is attached to bones that move a fluid that tickles tiny hairs, called *cilia* in your inner ear. That is how you hear.

We sometimes see plots of sound waveforms. The x-axis represents time. The y-axis represents the amount the air is compressed at the microphone that converted the air pressure into an electrical signal.



If the guitar string is made tighter (by the tuning pegs) or shorter (by the guitarist's fingers on the strings), the string vibrates more times per second. We measure the number of waves per second and we call it the *frequency* of the tone. The unit for frequency is *Hertz*: cycles per second.

Musicians have given the different frequencies names. If the guitarist plucks the lowest note on his guitar, it will vibrate at 82.4 Hertz. The guitarist will say "That pitch is low E." If the string is made half as long (by a finger on the 12th fret), the frequency will be twice as fast (164.8 Hertz), and the guitarist will say "That is E an octave up."

For any note, the note that has twice the frequency is one octave up. The note that has half the frequency is one octave down.

The octave is a very big jump in pitch, so musicians break it up into 12 smaller steps. If the guitarist shortens the E string by one fret, the frequency will be $82.4 \times 1.059463 \approx 87.3$ Hertz.

Shortening the string one fret always increases the frequency by a factor of 1.059463. Why?

Because $1.059463^{12} = 2$. That is, if you take 12 of these hops, you end up an octave higher.

This, the smallest hop in western music, is referred to as *half step*.

Exercise 55 Notes and frequencies**Working Space**

The note A near the middle of the piano, is 440Hz. The note E is 7 half steps above A. What is its frequency?

Answer on Page 827

45.2 Chords and harmonics

Of course, a guitarist seldom plays only one string at a time. Instead, he uses the frets to pick a pitch for each string and strums all six strings.

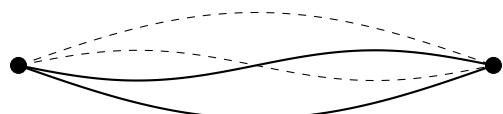
Some combinations of frequencies sound better than others. We have already talked about the octave: if one string vibrates twice for each vibration of another, they sound sweet together.

Musicians speak of “the fifth”. If one string vibrates three times and the other vibrates twice in the same amount of time, they sound sweet together.

If one string vibrates 4 times while the other vibrates 3 times, they sound sweet together. Musicians call this “the third.”

Each of these different frequencies tickle different cilia in the inner ear, so you can hear all six notes at the same time when the guitarist strums his guitar.

When a string vibrates, it doesn’t create a single sine wave. Yes, the string vibrates from end-to-end and this generates a sine wave at what we call *the fundamental frequency*. However, there are also “standing waves” on the string. One of these standing waves is still at the centerpoint of the string, but everything to the left of the centerpoint is going up while everything to the right is going down. This creates *an overtone* that is twice the frequency of the fundamental.



The next overtone has two still points – it divides the string into three parts. The outer parts are up while the inner part is down. Its frequency is three times the fundamental

frequency.



And so on: 4 times the fundamental, 5 times the fundamental, etc.

In general, tones with a lot of overtones tend to sound bright. Tones with just the fundamental sound thin.

Humans can generally hear frequencies from 20Hz to 20,000Hz (or 20kHz). Young people tend to be able to hear very high sounds better than older people.

Dogs can generally hear sounds in the 65Hz to 45kHz range.

45.3 Making waves in Python

Let's make a sine wave and add some overtones to it. Create a file `harmonics.py`

```
import matplotlib.pyplot as plt
import math

# Constants: frequency and amplitude
fundamental_freq = 440.0 # A = 440 Hz
fundamental_amp = 2.0

# Up an octave
first_freq = fundamental_freq * 2.0 # Hz
first_amp = fundamental_amp * 0.5

# Up a fifth more
second_freq = fundamental_freq * 3.0 # Hz
second_amp = fundamental_amp * 0.4

# How much time to show
max_time = 0.0092 # seconds

# Calculate the values 10,000 times per second
time_step = 0.00001 # seconds

# Initialize
time = 0.0
times = []
```

```
totals = []
fundamentals = []
firsts = []
seconds = []

while time <= max_time:
    # Store the time
    times.append(time)

    # Compute value each harmonic
    fundamental = fundamental_amp * math.sin(2.0 * math.pi * fundamental_freq * time)
    first = first_amp * math.sin(2.0 * math.pi * first_freq * time)
    second = second_amp * math.sin(2.0 * math.pi * second_freq * time)

    # Sum them up
    total = fundamental + first + second

    # Store the values
    fundamentals.append(fundamental)
    firsts.append(first)
    seconds.append(second)
    totals.append(total)

    # Increment time
    time += time_step

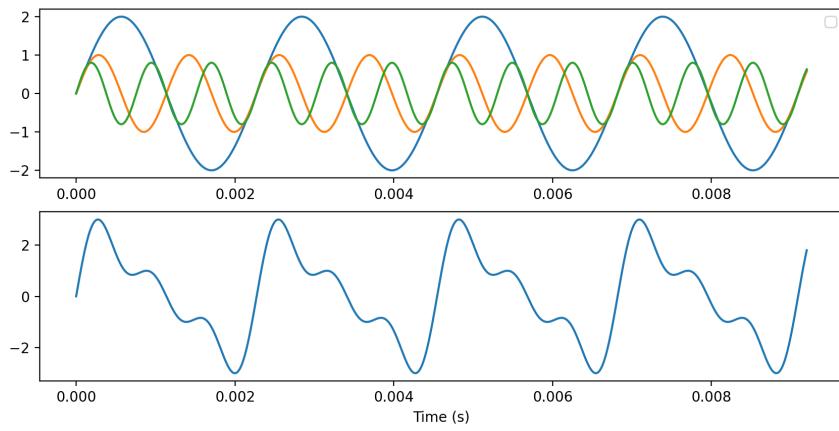
# Plot the data
fig, ax = plt.subplots(2, 1)

# Show each component
ax[0].plot(times, fundamentals)
ax[0].plot(times, firsts)
ax[0].plot(times, seconds)
ax[0].legend()

# Show the totals
ax[1].plot(times, totals)
ax[1].set_xlabel("Time (s)")

plt.show()
```

When you run it, you should see a plot of all three sine waves and another plot of their sum:



45.3.1 Making a sound file

The graph is pretty to look at, but make a file that we can listen to.

The WAV audio file format is supported on pretty much any device, and a library for writing WAV files comes with Python. Let's write some sine waves and some noise into a WAV file.

Create a file called `soundmaker.py`

```
import wave
import math
import random

# Constants
frame_rate = 16000 # samples per second
duration_per = 0.3 # seconds per sound
frequencies = [220, 440, 880, 392] # Hz
amplitudes = [20, 125]
baseline = 127 # Values will be between 0 and 255, so 127 is the baseline
samples_per = int(frame_rate * duration_per) # number of samples per sound

# Open a file
wave_writer = wave.open('sound.wav', 'wb')

# Not stereo, just one channel
wave_writer.setnchannels(1)

# 1 byte audio means everything is in the range 0 to 255
```

```
wave_writer.setsampwidth(1)

# Set the frame rate
wave_writer.setframerate(frame_rate)

# Loop over the amplitudes and frequencies
for amplitude in amplitudes:
    for frequency in frequencies:
        time = 0.0
        # Write a sine wave
        for sample in range(samples_per):
            s = baseline + int(amplitude * math.sin(2.0 * math.pi * frequency * time))
            wave_writer.writeframes(bytes([s]))
            time += 1.0 / frame_rate

        # Write some noise after each sine wave
        for sample in range(samples_per):
            s = baseline + random.randint(0, 15)
            wave_writer.writeframes(bytes([s]))

# Close the file
wave_writer.close()
```

When you run it, it should create a sound file with several tones of different frequencies and volumes. Each tone should be followed by some noise.

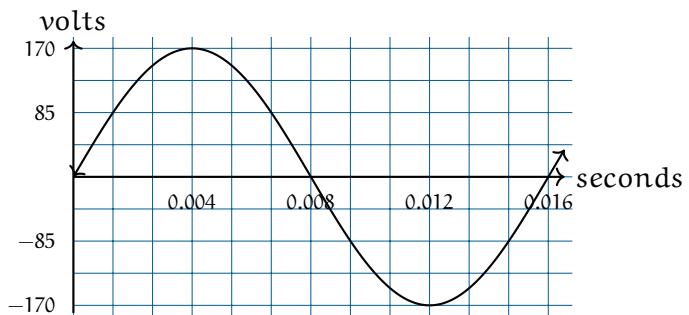


CHAPTER 46

Alternating Current

We have discussed the voltage and current created by a battery. A battery pushes the electrons in one direction at a constant voltage; this is known as *Direct Current* or DC. A battery typically provides between 1.5 and 9 volts.

The electrical power that comes into your home on wires is different. If you plotted the voltage over time, it would look like this:



The x axis here represents ground. When you insert a two-prong plug into an outlet, one is "hot" and the other is "ground". Ground represents 0 volts and should be the same

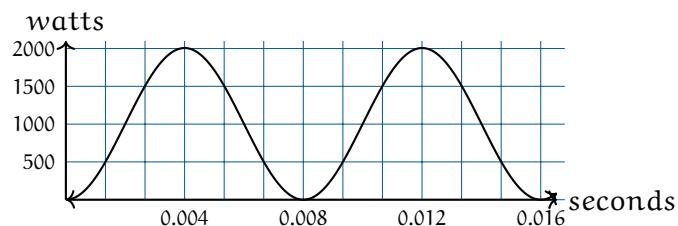
voltage as the dirt under the building.

The voltage is a sine wave at 60Hz. Your voltage fluctuates between -170v and 170v. Think for a second what that means: The power company pushes electrons at 170v and then pulls electrons at 170v. It alternates back and forth this way 60 times per second.

46.1 Power of AC

Let's say you turn on your toaster which has a resistance of 14.4 ohms. How much energy (in watts) does it change from electrical energy to heat? We know that $I = V/R$ and we know that watts of power are IV . So given a voltage of V , the toaster is consuming V^2/R watts.

However, V is fluctuating. Let's plot the power the toaster is consuming:



Another sine wave! Here is a lesser-known trig identity: $(\sin(x))^2 = \frac{1}{2} - \frac{1}{2} \cos(2x)$

So this is actually a cosine wave flipped upside down, scaled down by half the peak power, and translated up so that it is never negative. Note that it is also twice the frequency of the voltage sine wave.

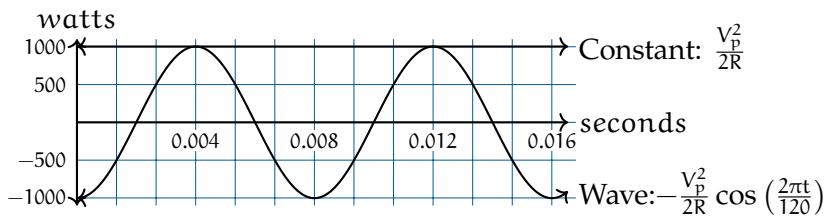
If we say the peak voltage is V_p and the resistance of the toaster is R , the power is given by

$$\frac{V_p^2}{2R} - \frac{V_p^2}{2R} \cos\left(\frac{2\pi t}{120}\right)$$

As a toaster user and as someone who pays a power bill, you are mostly interested in the average power. To get the average power, you take the area under the power graph and divide it by the amount of time.

We can think of the area under the curve as two easy-to-integrate quantities summed:

- A constant function of $y = \frac{V_p^2}{2R}$
- A wave $y = -\frac{V_p^2}{2R} \cos\left(\frac{2\pi t}{120}\right)$



When we integrate that constant function we get $\frac{tV_p^2}{2R}$

When we integrate that wave for a complete cycle we get...zero! The positive side of the wave is canceled out by the negative side.

So, the average power is $\frac{V_p^2}{2R}$ watts.

Someone at some point said "I'm used to power being V^2/R . Can we define a voltage measure for AC power such that this is always true?"

So we started using V_{rms} which is just $\frac{V_p}{\sqrt{2}}$. If you look on the back of anything that plugs into a standard US power outlet, it will say something like "For 120v". What they mean is "For 120v RMS, so we expect the voltage to fluctuate back and forth from 170v to -170v."

Notice that this is the same Root-Mean-Squared that we defined earlier, but now we know that if $y = \sin(x)$, the RMS of y is $1/\sqrt{2} \approx 0.707$.

For current, we do the same thing: If the current is AC, the power consumed by a resistor is $I_{rms}^2 R$, where I_{rms} is the peak current divided by $\sqrt{2}$.

46.2 Power Line Losses

A wire has some resistance. Thinner wires tend to have more resistance than thicker ones. Aluminum wires tend to have more resistance than copper wires.

Let's say that the power that comes to your house has to travel 20 km from the generator in a cable that has about 1Ω of resistance per km. Let's say that your home is consuming 12 kilowatts of power.

If that power is 120v RMS from the generator to your home, what percentage of the power is lost heating the power line? 10 amps RMS flow through your home. When that current goes through the wire, $I^2R = (100)(20) = 2000$ watts is lost to heat.

So the power company would need to supply 14 kilowatts of power, knowing that 2 kilowatts would be lost on the wires.

What if the power company moved the power at 120,000 volts RMS? Now only 0.01

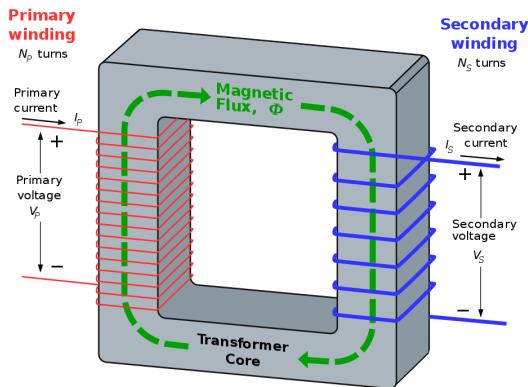
amps RMS flow through your home. When that current goes through the wire $I^2R = (0.0001)(20) = .002$ watts of power are lost on the power lines.

It is much, much more efficient. The only problem is that 120,000 volts would be incredibly dangerous. So the power company moves power long distances at very high voltages, like 765 kV. Before the power is brought into your home, it is converted into a lower voltage using a *transformer*.

46.3 Transformers

A transformer is a device that converts electrical power from one voltage to another. A good transformer is more than 95% efficient. The details of magnetic fields, flux, and inductance are beyond the scope of this chapter, so I am going to give a relatively simple explanation and admit that it is incomplete.

A transformer is a ring with two sets of coils wrapped around it.



(Diagram from Wikipedia)

When alternating current is run through the primary winding, it creates magnetic flux in the ring. The magnetic flux induces current in the secondary winding.

If V_p is the voltage across the primary winding and V_s is the voltage across the secondary winding, they are related by the following equation:

$$\frac{V_p}{V_s} = \frac{N_p}{N_s}$$

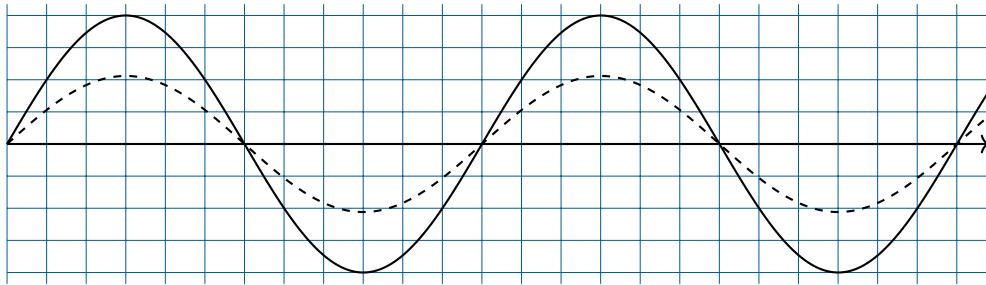
where N_p and N_s are the number of turns in the primary and secondary windings.

There are usually at least two transformers between you and the very high voltage lines.

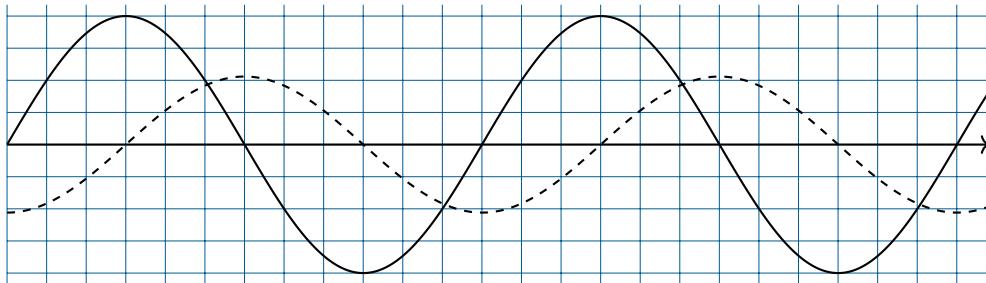
There are transformers at the substation that make the voltage low enough to travel on regular utility poles. On the utility poles, you will see cans that contain smaller transformers. Those step the voltage down to make the power safe to enter your home.

46.4 Phase and 3-phase power

If two waves are “in sync” we say they have the same *phase*.

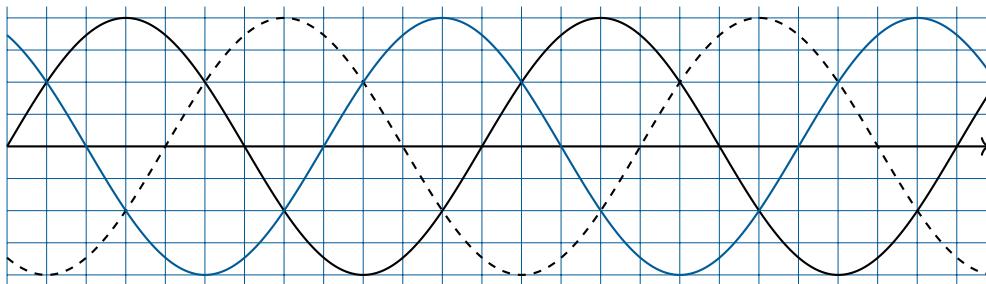


If they are the same frequency, but are not in-sync, we can talk about the difference in their phase.



Here we see that the smaller wave is lagging by $\pi/2$ or 90° .

In most power grids, there are usually 3 wires carrying the power. The voltage on each is $2\pi/3$ out of phase with the other two:



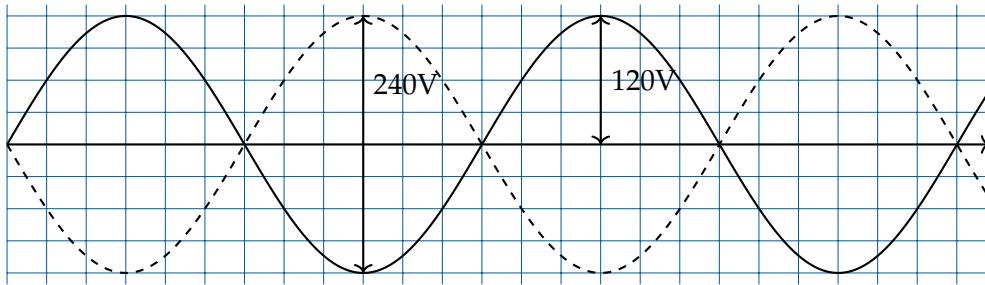
This is nice in two ways:

- While the power in each wire is fluctuating, the total power is not fluctuating at all.
- While the power plant is pushing and pulling electrons on each wire, the total number of electrons leaving the load is zero.

(Both these assume that there each wire is attached to a load with the same constant resistance.)

In big industrial factories, you will see all three wires enter the building. Large amounts of smooth power delivery means a lot to an industrial user.

In residential settings, each home gets its power from one of the three wires. However, two wires typically carry power into the home. Each one carries 120V RMS, but they are out of phase by 180 degrees. Lights and small appliances are connected to one of the wires and ground, so they get 120V RMS. Large appliances, like air conditioners and washing machines, are connected across the two wires so they get 240V RMS.



How do you get two circuits, 180 degrees out of phase, from one circuit? Using a center-tap transformer.

FIXME: Diagram here

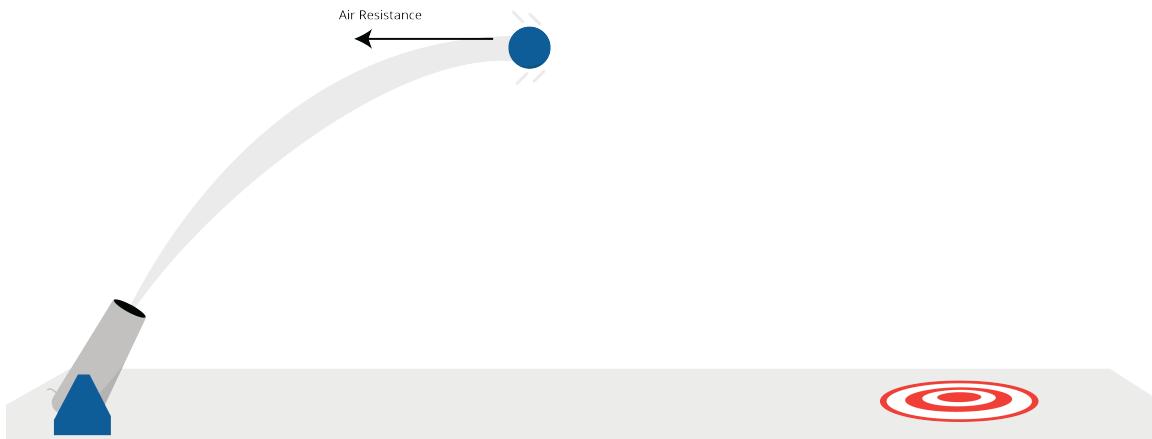


CHAPTER 47

Drag

The very first computers were created to do calculations of how artillery would fly when shot at different angles. The calculations were similar to the ones you just did for the flying hammer with two important differences:

- They were interested in two dimensions: the height and the distance across the ground.
- However, artillery flies a lot faster than a hammer, so they had to worry about drag from the air.



47.1 Wind resistance

The first thing they did was put one of the shells in a wind tunnel. They measured how much force was created when they pushed 1 m/s of wind over the shell. Let's say it was 0.1 newtons.

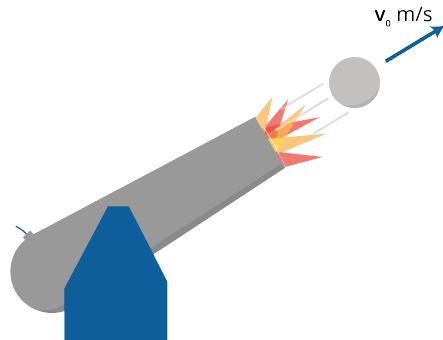
One of the interesting things about the drag from the air (often called *wind resistance*) is that it increases with the *square* of the speed. Thus, if the wind pushing on the shell is 3 m/s, instead of 1 m/s, the resistance is $3^2 \times 0.1 = 0.9$ newtons.

(Why? Intuitively, three times as many air molecules are hitting the shell and each molecule is hitting it three times harder.)

So, if a shell is moving with the velocity vector v , the force vector of the drag points in the exact opposite direction. If μ is the force of wind resistance of the shell at 1 m/s, then the magnitude of the drag vector is $\mu|v|^2$.

47.2 Initial velocity and acceleration due to gravity

Let's say a shell is shot out of a tube at s m/s, and let's say the tube is tilted θ radians above level. Then, the initial velocity will be given by the vector $[s \cos(\theta), s \sin(\theta)]$



(The velocity of the shell is actually a 3-dimensional vector, but we are only going to worry about height and horizontal distance; we are assuming that the operator pointed it in the right direction.)

To figure out the path of the shell, we need to compute its acceleration. We remember that

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

(Note that \mathbf{F} and \mathbf{a} are vectors.) Dividing both sides by m we get:

$$\mathbf{a} = \frac{\mathbf{F}}{m}$$

So let's figure out the net force on the shell so that we can calculate the acceleration vector.

If the shell has a mass of b , the force due to gravity will be in the downward direction with a magnitude of $9.8b$ newtons.

To get the net force, we will need to add the force due to gravity with the force due to wind resistance.

47.3 Simulating artillery in Python

Create a file called `artillery.py`.

```
import numpy as np
import matplotlib.pyplot as plt
```

```
# Constants
mass = 45 # kg
start_speed = 300.0 # m/s
theta = np.pi/5 # radians (36 degrees above level)
time_step = 0.01 # s
wind_resistance = 0.05 # newtons in 1 m/s wind
force_of_gravity = np.array([0.0, -9.8 * mass]) # newtons

# Initial state
position = np.array([0.0, 0.0]) # [distance, height] in meters
velocity = np.array([start_speed * np.cos(theta), start_speed * np.sin(theta)])
time = 0.0 # seconds

# Lists to gather data
distances = []
heights = []
times = []

# While shell is aloft
while position[1] >= 0:
    # Record data
    distances.append(position[0])
    heights.append(position[1])
    times.append(time)

    # Calculate the next state
    time += time_step
    position += time_step * velocity

    # Calculate the net force vector
    force = force_of_gravity - wind_resistance * velocity**2

    # Calculate the current acceleration vector
    acceleration = force / mass

    # Update the velocity vector
    velocity += time_step * acceleration

print(f"Hit the ground {position[0]:.2f} meters away at {time:.2f} seconds.")

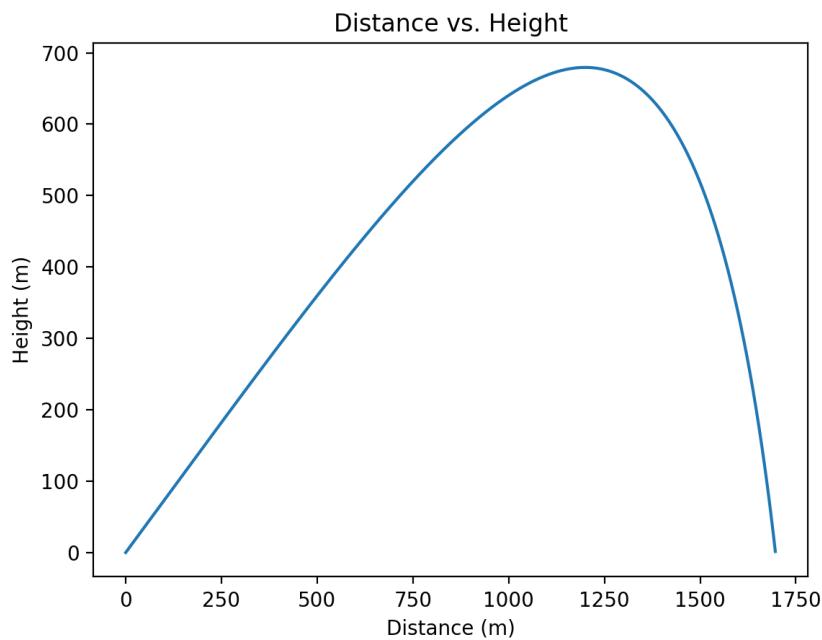
# Plot the data
fig, ax = plt.subplots()
ax.plot(distances, heights)
ax.set_title("Distance vs. Height")
ax.set_xlabel("Distance (m)")
```

```
ax.set_ylabel("Height (m)")
plt.show()
```

When you run it, you should get a message like:

```
Hit the ground 1696.70 meters away at 20.73 seconds.
```

You should also see a plot of the shell's path:



47.4 Terminal velocity

If you shot the shell very, very high in the sky, it would keep accelerating toward the ground until the force of gravity and the force of the wind resistance were equal. The speed at which this happens is called the *terminal velocity*. The terminal velocity of a falling human is about 53 m/s.

Exercise 56 Terminal velocity

What is the terminal velocity of shell described in our example?

Working Space

Answer on Page 827



CHAPTER 48

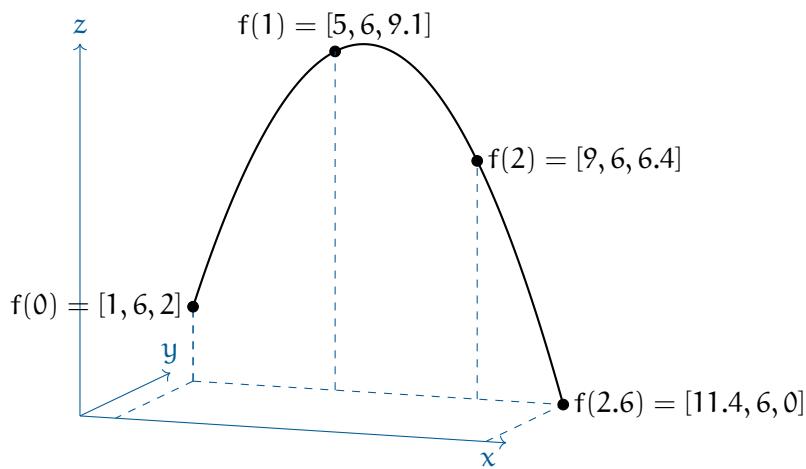
Vector-valued Functions

In the last chapter, you calculated the flight of the shell. For any time t , you could find a vector [distance, height]. This can be thought of as a function f that takes a number and returns a 2-dimensional vector. We call this a *vector-valued* function from $\mathbb{R} \rightarrow \mathbb{R}^2$.

We often make a vector-valued function by defining several real-valued functions. For example, if you threw a hammer with an initial upward speed of 12 m/s and a horizontal speed of 4 m/s along the x axis from the point $(1, 6, 2)$, its position at time t (during its flight) would be given by:

$$f(t) = [4t + 1, 6, -4.8t^2 + 12t + 2]$$

That is, x is increasing with t , y is constant, and z is a parabola.

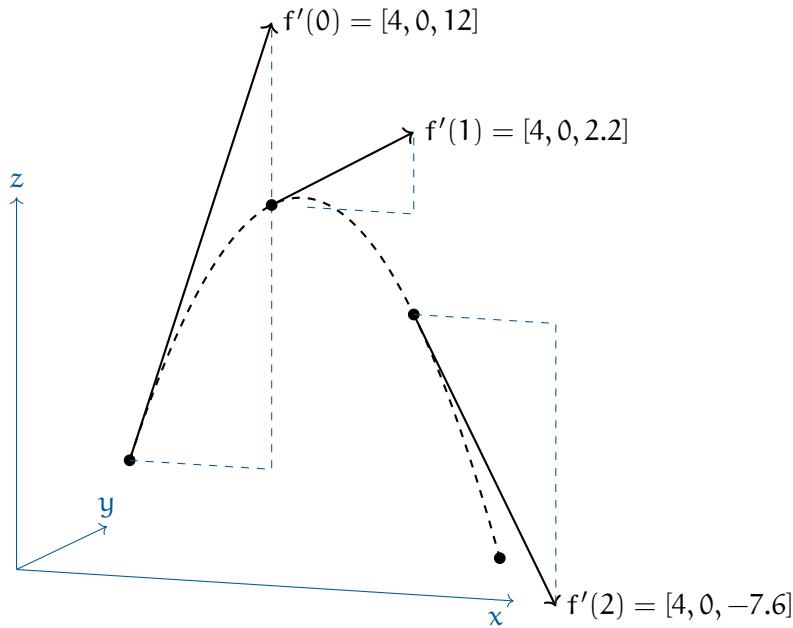


48.1 Finding the velocity vector

Now that we have its position vector, we can differentiate each component separately to get its velocity as a vector-valued function:

$$f'(t) = [4, 0, -9.8t + 12]$$

That is, the velocity is constant along the x -axis, zero along the y -axis, and decreasing with time along the z axis.

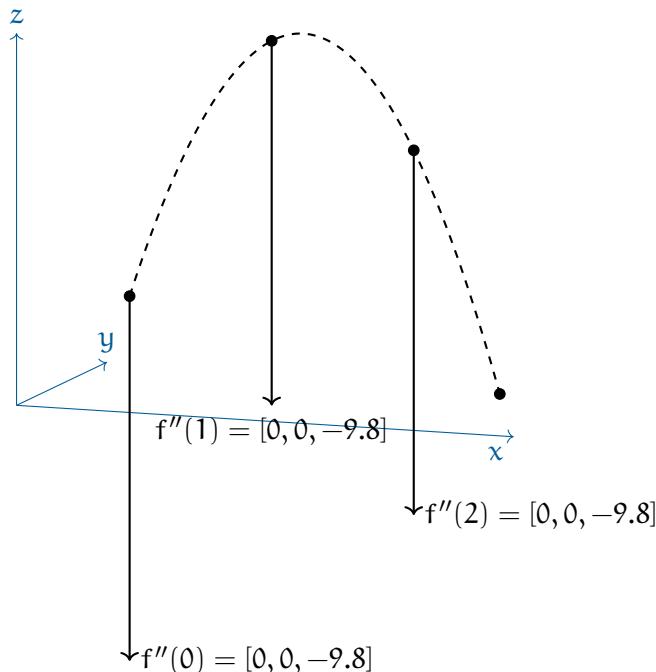


48.2 Finding the acceleration vector

Now that we have its velocity, we can get its acceleration as a vector-valued function:

$$\mathbf{f}''(t) = [0, 0, -9.8]$$

There is no acceleration along the x or y axes. It is accelerating down at a constant 9.8 m/s^2 .





CHAPTER 49

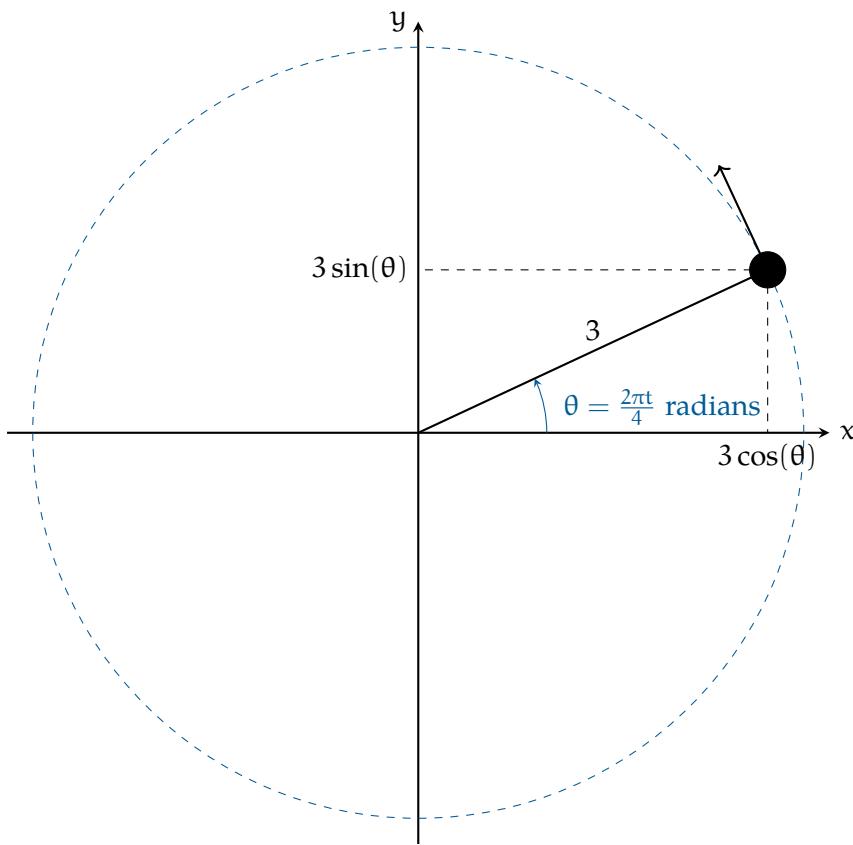
Circular Motion

Let's say you tie a 0.16 kg billiard ball to a long string and begin to swing it around in a circle above your head. Let's say the string is 3 meters long, and the ball returns to where it started every 4 seconds. If you start your stopwatch as the ball crosses the x-axis, the position of the ball at any time t given by:

$$p(t) = [3 \cos\left(\frac{2\pi}{4}t\right), 3 \sin\left(\frac{2\pi}{4}t\right), 2]$$

(This assumes that the ball would be going counter-clockwise if viewed from above. The spot you are standing on is considered the origin $[0, 0, 0]$.)

Notice that the height is a constant – 2 meters in this case. That isn't very interesting, so we will talk just about the first two components. Here is what it would look like from above:



In this case, the radius, r , is 3 meters. The period, T is 4 seconds. In general, we say that circular motion is given by:

$$\mathbf{p}(t) = \left[r \cos \frac{2\pi t}{T}, r \sin \frac{2\pi t}{T} \right]$$

A common question is “How fast is it turning right now?” If you divide the 2π radians of a circle by the 4 seconds it takes, you get the answer “About 1.57 radians per second.” This is known as *angular velocity* and we typically represent it with the lowercase Omega: ω . (Yes, it looks a lot like a “w”.) To be precise, in our example, the angular velocity is $\omega = \frac{\pi}{2}$.

Notice that this is different from the question “How fast is it going?” This ball is traveling the circumference of $6\pi \approx 18.85$ meters every 4 seconds. So the speed of the ball is about 4.71 meters per second.

49.1 Velocity

The velocity of the ball is a vector, and we can find that vector by differentiating each component of the position vector.

For any constants a and b :

Expression	Derivative
$a \sin bt$	$ab \cos bt$
$a \cos bt$	$-ab \sin bt$

Thus, in our example, the velocity of the ball at any time t is given by:

$$\mathbf{v}(t) = \left[-\frac{3(2\pi)}{4} \sin \frac{2\pi t}{4}, \frac{3(2\pi)}{4} \cos \frac{2\pi t}{4}, 0 \right]$$

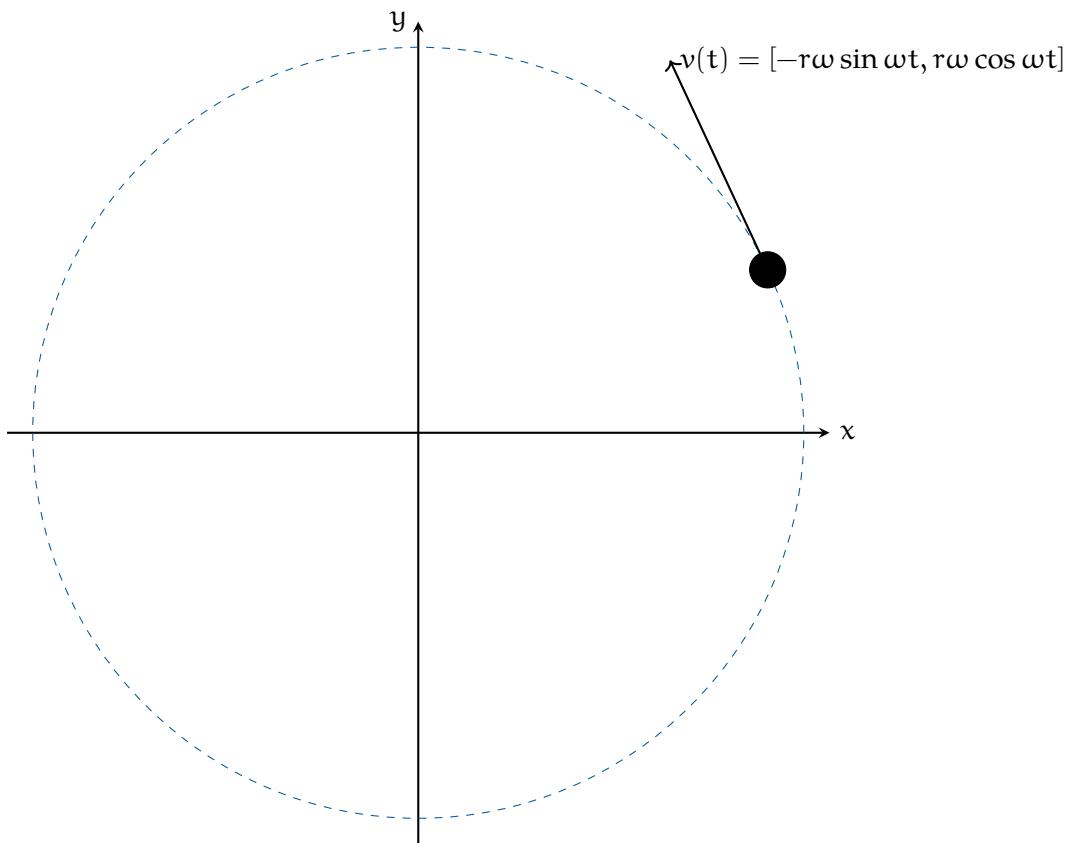
Notice that the velocity vector is perpendicular to the position vector. It has a constant magnitude.

In general, an object traveling in a circle at a constant speed has the velocity vector:

$$\mathbf{v}(t) = [-r\omega \sin \omega t, r\omega \cos \omega t]$$

where $t = 0$ is the time that it crosses the x axis. If ω is negative, that means the motion would be clockwise when viewed from above.

The magnitude of the velocity vector is $r\omega$.

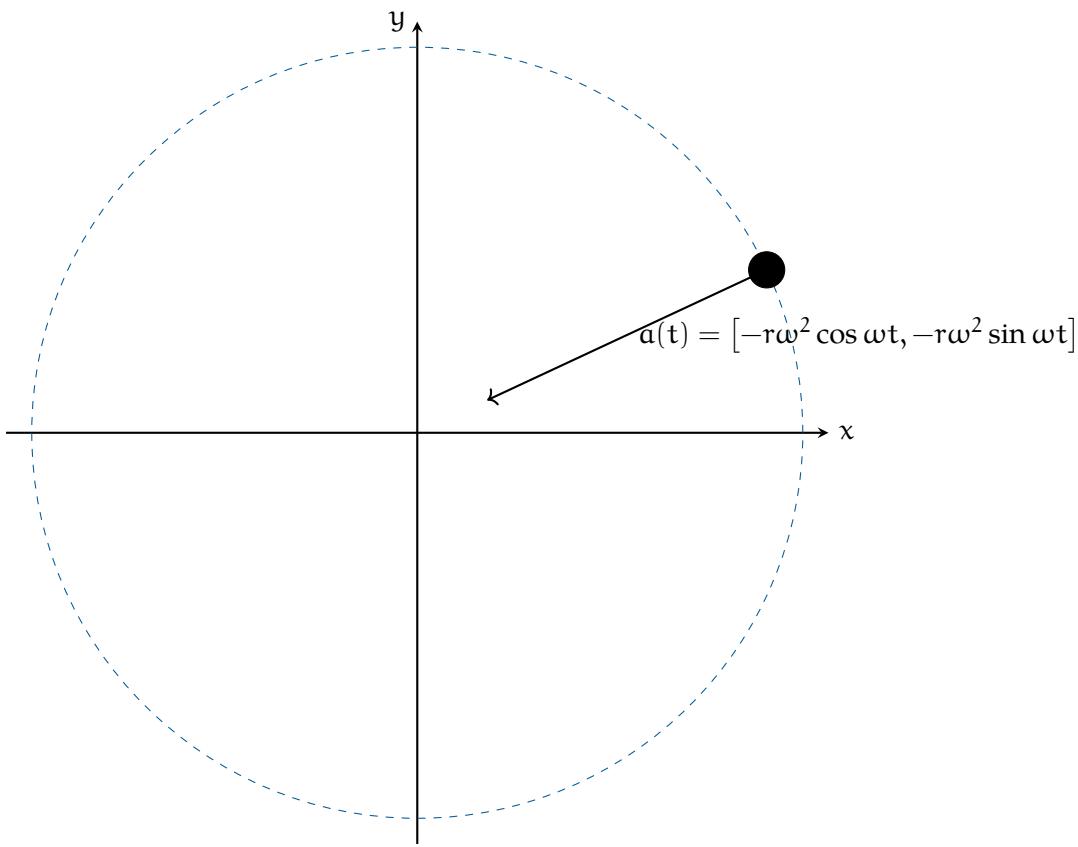


49.2 Acceleration

We can get the acceleration by differentiating the components of the velocity vector.

$$\mathbf{a}(t) = [-r\omega^2 \cos \omega t, -r\omega^2 \sin \omega t]$$

Notice that the acceleration vector points toward the center of the circle it is traveling on. That is, when an object is traveling on a circle at a constant speed, its only acceleration is toward the center of the circle.



The magnitude of the acceleration vector is $r\omega^2$.

49.3 Centripetal force

How hard is the ball pulling against your hand? That is, if you let go, the ball would fly in a straight line. The force you are exerting on the string is what causes it to accelerate toward the center of the circle. We call this the *centripetal force*.

Recall that $F = ma$. The magnitude of the acceleration is $r\omega^2 = 3\left(\frac{2\pi}{4}\right)^2 \approx 7.4$ m/s. The mass of the ball is 0.16 kg. So the force pulling against your hand is about 1.18 newtons.

The general rule is that when something is traveling in a circle at a constant speed, the centripetal force needed to keep it traveling in a circle is:

$$F = mr\omega^2$$

If you know the radius r and the speed v of the object, here is the rule:

$$F = \frac{mv^2}{r}$$

Exercise 57 Circular Motion

Just as your car rolls onto a circular track with a radius of 200 m, you realize your 0.4 kg cup of coffee is on the slippery dashboard of your car. While driving 120 km/hour, you hold the cup to keep it from sliding.

What is the maximum amount of force you would need to use (The friction of the dashboard helps you, but the max is when the friction is zero.)

Working Space

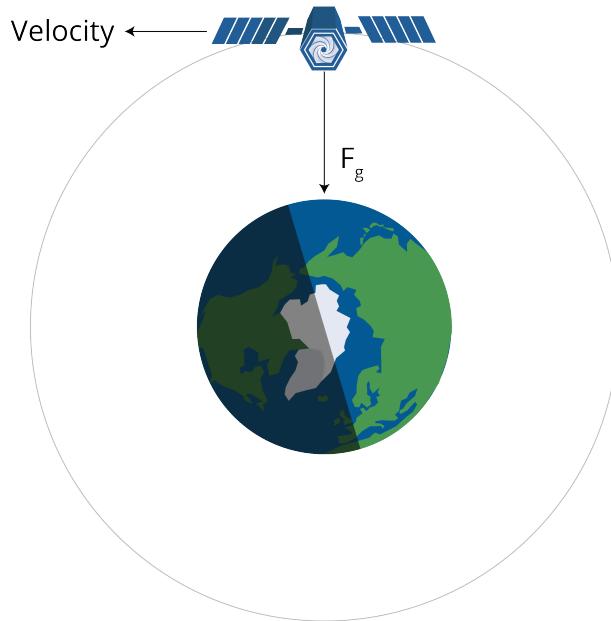
Answer on Page 827



CHAPTER 50

Orbits

A satellite stays in orbit around the planet because the pull of the planet's gravity causes it to accelerate toward the center of the planet.



The satellite must be moving at a very particular speed to keep a constant distance from the planet – to travel in a circular orbit. If it is moving too slowly, it will get closer to the planet. If it is going too fast, it will get farther from the planet.



The radius of the earth is about 6.37 million meters. A satellite that is in a low orbit is typically about 2 million meters above the ground. At that distance, the acceleration due to gravity is more like 6.8m/s^2 , instead of the 9.8m/s^2 that we experience on the surface of the planet.

How fast does the satellite need to be moving in a circle with a radius of 8.37 million meters to have an acceleration of 6.8m/s^2 ? Real fast.

Recall that the acceleration vector is

$$\mathbf{a} = \frac{\mathbf{v}^2}{r}$$

Thus the velocity v needs to be:

$$v = \sqrt{ar} = \sqrt{6.8(8.37 \times 10^6)} = 7,544 \text{ m/s}$$

(That's 16,875 miles per hour.)

When a satellite falls out of orbit, it enters the atmosphere at that 7,544 m/s. The air rushing by generates so much friction that the satellite gets very, very hot and usually disintegrates.

50.1 Astronauts are not weightless

Some people see astronauts floating inside an orbiting spacecraft and think there is no gravity: that the astronauts are so far away that the gravity of the planet doesn't affect them. This is incorrect. The gravity might be slightly less (Maybe 6 newtons per kg instead of 9.8 newtons per kg), but the weightless they experience is because they and the spacecraft is in free fall. They are just moving so fast (in a direction perpendicular to gravity) that they don't collide with the planet.

Exercise 58 Mars Orbit*Working Space*

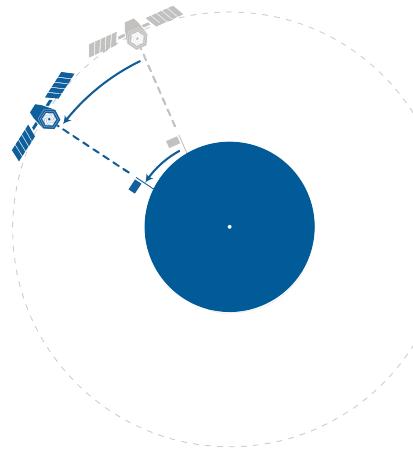
The radius of Mars is 3.39 million meters. The atmosphere goes up another 11 km. Let's say you want to put a satellite in a circular orbit around Mars with a radius of 3.4 million meters.

The acceleration due to gravity on the surface of Mars is 3.721m/s^2 . We can safely assume that it is approximately the same 11 km above the surface.

How fast does the satellite need to be traveling in its orbit? How long will each orbit take?

*Answer on Page 827***50.2 Geosynchronous Orbits**

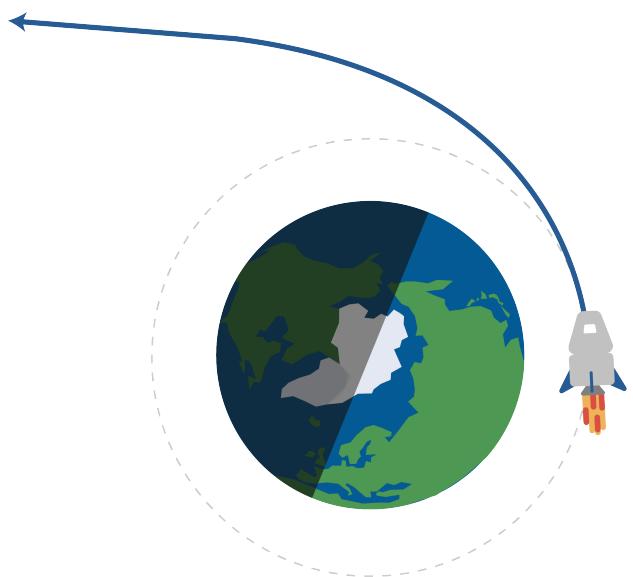
The planet earth rotates once a day. Satellites in low orbits circle the earth many times a day. Satellites in very high orbits circle less than once per day. There is a radius at which a satellite orbits exactly once per day. Satellites at this radius are known as "geosynchronous" or "geostationary" because they are always directly over a place on the planet.



The radius of a circular geosynchronous orbit is 42.164 million meters. (About 36 km above the surface of the earth.)

A geosynchronous satellite travels at a speed of 3,070 m/s.

Geosynchronous satellites are used for the Global Positioning Satellite system, weather monitoring system, and communications system.



FIXME: Add text for escape velocity



CHAPTER 51

Simulation with Vectors

You wrote a python program that simulated the flight of a hammer to predict its altitude. Your simulation dealt only with scalars. Now you are ready to create simulations of positions, velocities, accelerations, and forces as vectors.

In this chapter, you are going to simulate two moons that, as they wandered through the vast universe, get caught in each other's gravity well. We will assume there are no other forces acting upon the moons.

51.1 Force, Acceleration, Velocity, and Position

We talked about the magnitude of a gravitational attraction between two masses:

$$F = G \frac{m_1 m_2}{r^2}$$

Where F is the magnitude of the force in newtons, m_1 and m_2 are the masses in kg, r is the distance between them in meters, and g is the universal gravitational constant:

6.67430×10^{-11} .

What is the direction? For the two moons, the force on moon 1 will pull toward moon 2. And the force on moon 2 will pull toward moon 1.

Of course, if something is big (like the sun), you need to be more specific: The force points directly at the center of mass of the object that is generating the force.

Each of the moons will start off with a velocity vector. That velocity vector will change over time as the moon is accelerated by the force of gravity. If you have a mass m with an initial velocity vector of \vec{v}_0 that is being accelerated with a constant force vector \vec{F} , at time t the new velocity vector will be:

$$\vec{v}_t = \vec{v}_0 + \frac{t}{m} \vec{F}$$

If an object is at an initial position vector of \vec{p}_0 and moves with a constant velocity vector \vec{v} for time t , the new position will be given by

$$\vec{p}_t = \vec{p}_0 + t\vec{v}$$

51.2 Simulations and Step Size

As two moons orbit each other, the force, acceleration, velocity, and position are changing smoothly and continuously. It is difficult to simulate truly continuous things on a digital computer.

However, think about a movie: It shows you many frames each second. Each frame is a still picture of the state of the system. And the more frames per second, the smoother it looks.

We do a similar trick in simulations. We say "We are going to run our simulation in 2 hour steps. We will assume that the acceleration and velocity were constant for those two hours. We will update our position vectors accordingly, and then we will recalculate our acceleration and velocity vectors."

Generally, as you make the step size smaller, your simulation will get more accurate and take longer to execute.

51.3 Make a Text-based Simulation

To start, you are going to write a Python program that simulates the moons and prints out their position for every time step. Later we will add graphs and even animation.

We are going to assume the two moons are traveling the same plane so we can do all the math and graphing in 2 dimensions.

Each moon will be represented by a dictionary containing the state of the moon:

- Its mass in kilograms
- Its position – a 2-dimensional vector represent x and y coordinates of the center of the moon.
- Its velocity – a 2-dimensional vector
- Its radius – Each moon has a radius so we know when the centers of the two moons are so close to each other that they must have collided.
- Its color – We will use that when do the plots and animations. One moon will be red, the other blue.

Then there will be a loop where we will update the positions of the moons and then recalculate the acceleration and velocities.

How much time will be simulated? 100 days or until the moons collide, whichever comes first.

We will use numpy arrays to represent our vectors.

Create a file called `moons.py` and type in this code:

```
import numpy as np

# Constants
G = 6.67430e-11          # Gravitational constant (Nm^2/kg^2)
SEC_PER_DAY = 24 * 60 * 60 # How many seconds in a day?
MAX_TIME = 100 * SEC_PER_DAY # 100 days
TIME_STEP = 2 * 60 * 60     # Update every two hours

# Create the initial state of Moon 1
m1 = {
    "mass": 6.0e22, # kg
    "position": np.array([0.0, 200_000_000]), # m
    "velocity": np.array([100.0, 25.0]), # m/s
    "radius": 1_500_000.0, # m
```

```
        "color": "red" # For plotting
    }

# Create the initial state of Moon 2
m2 = {
    "mass": 11.0e22, # kg
    "position": np.array([0.0, -150_000_000]), # m
    "velocity": np.array([-45.0, 2.0]), # m/s
    "radius": 2_000_000.0, # m
    "color": "blue" # For plotting
}

# Lists to hold positions and time
position1_log = []
position2_log = []
time_log = []

# Start at time zero seconds
current_time = 0.0

# Loop until current time exceed Max Time
while current_time <= MAX_TIME:

    # Add time and positions to log
    time_log.append(current_time)
    position1_log.append(m1["position"])
    position2_log.append(m2["position"])

    # Print the current time and positions
    print(f"Day {current_time/SEC_PER_DAY:.2f}:")
    print(f"\tMoon 1:({m1['position'][0]:,.1f},{m1['position'][1]:,.1f})")
    print(f"\tMoon 2:({m2['position'][0]:,.1f},{m2['position'][1]:,.1f})")

    # Update the positions based on the current velocities
    m1["position"] = m1["position"] + m1["velocity"] * TIME_STEP
    m2["position"] = m2["position"] + m2["velocity"] * TIME_STEP

    # Find the vector from moon1 to moon2
    delta = m2["position"] - m1["position"]

    # What is the distance between the moons?
    distance = np.linalg.norm(delta)

    # Have the moons collided?
    if distance < m1["radius"] + m2["radius"]:
        print(f"*** Collided {current_time:.1f} seconds in!")
        break

    # What is a unit vector that points from moon1 toward moon2?
    direction = delta / distance

    # Calculate the magnitude of the gravitational attraction
```

```
magnitude = G * m1["mass"] * m2["mass"] / (distance**2)

# Acceleration vector of moon1 (a = f/m)
acceleration1 = direction * magnitude / m1["mass"]

# Acceleration vector of moon2
acceleration2 = (-1 * direction) * magnitude / m2["mass"]

# Update the velocity vectors
m1["velocity"] = m1["velocity"] + acceleration1 * TIME_STEP
m2["velocity"] = m2["velocity"] + acceleration2 * TIME_STEP

# Update the clock
current_time += TIME_STEP

print(f"Generated {len(position1_log)} data points.")
```

When you run the simulation, you will see the positions of the moons for 100 days:

```
> python3 moons.py
Day 0.00:
Moon 1:(0.0,200,000,000.0)
Moon 2:(0.0,-150,000,000.0)
Day 0.08:
Moon 1:(720,000.0,200,180,000.0)
Moon 2:(-324,000.0,-149,985,600.0)
Day 0.17:
Moon 1:(1,439,990.7,200,356,896.1)
Moon 2:(-647,995.0,-149,969,507.0)
...
Day 100.00:
Moon 1:(119,312,305.5,283,265,313.5)
Moon 2:(17,393,287.9,-60,319,261.9)
Generated 1201 data points.
```

Look over the code. Make sure you understand what every line does.

51.4 Graph the Paths of the Moons

Now you will use the matplotlib to graph the paths of the moons. Add this line to the beginning of `moons.py`

```
import matplotlib.pyplot as plt
```

Add this code to the end of your `moons.py`:

```
# Convert lists to np.arrays
positions1 = np.array(position1_log)
positions2 = np.array(position2_log)

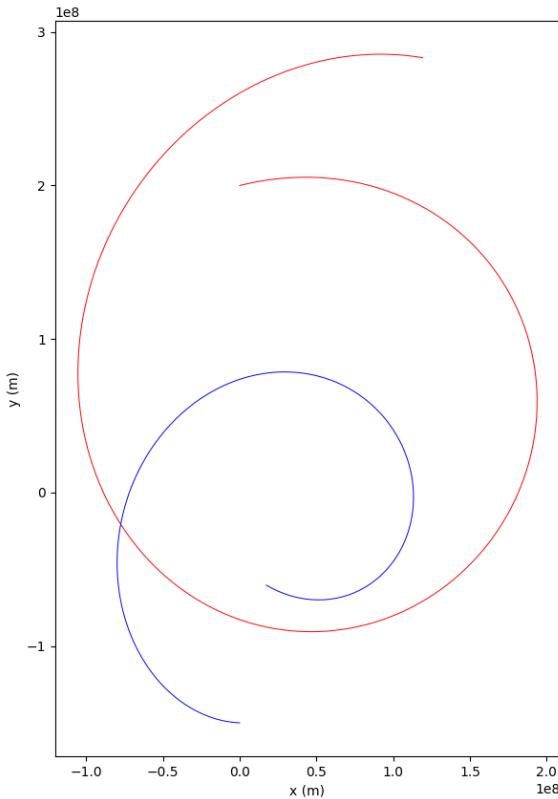
# Create a figure with a set of axes
fig, ax = plt.subplots(1, figsize=(7.2, 10))

# Label the axes
ax.set_xlabel("x (m)")
ax.set_ylabel("y (m)")
ax.set_aspect("equal", adjustable='box')

# Draw the path of the two moons
ax.plot(positions1[:, 0], positions1[:, 1], m1["color"], lw=0.7)
ax.plot(positions2[:, 0], positions2[:, 1], m2["color"], lw=0.7)

# Save out the figure
fig.savefig("plotmoons.png")
```

When you run it, your `plotmoons.png` should look like this:



It is nifty to see the paths, but we don't know where each moon was at a particular time. In fact, it is difficult to figure out which end of each curve was the beginning and which was the ending.

What if we added some lines and labels every 300 steps to put a sense of time into the plot? Add one more constant after the import statements:

```
PAIR_LINE_STEP = 300 # How time steps between pair lines
```

Immediately before you save the figure to the file, add the following code:

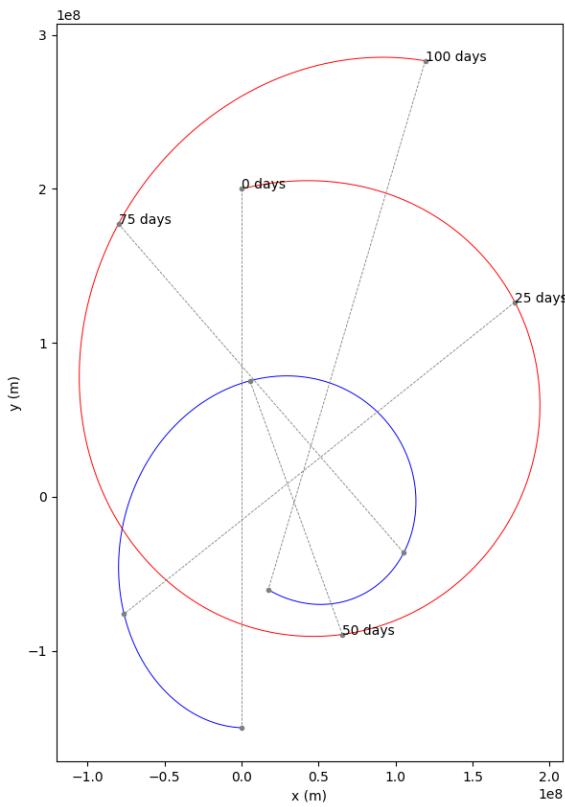
```
# Draw some pair lines that help the
# viewer understand time in the graph
i = 0
while i < len(positions1):

    # Where are the moons at the ith entry?
    a = positions1[i, :]
    b = positions2[i, :]
    ax.plot([a[0], b[0]], [a[1], b[1]], "--", c="gray", lw=0.6, marker=".")

    # What is the time at the ith entry?
    t = time_log[i]

    # Label the location of moon 1 with the day
    ax.text(a[0], a[1], f"{t/SEC_PER_DAY:.0f} days")
    i += PAIR_LINE_STEP
```

When you run it, your plot should look like this:

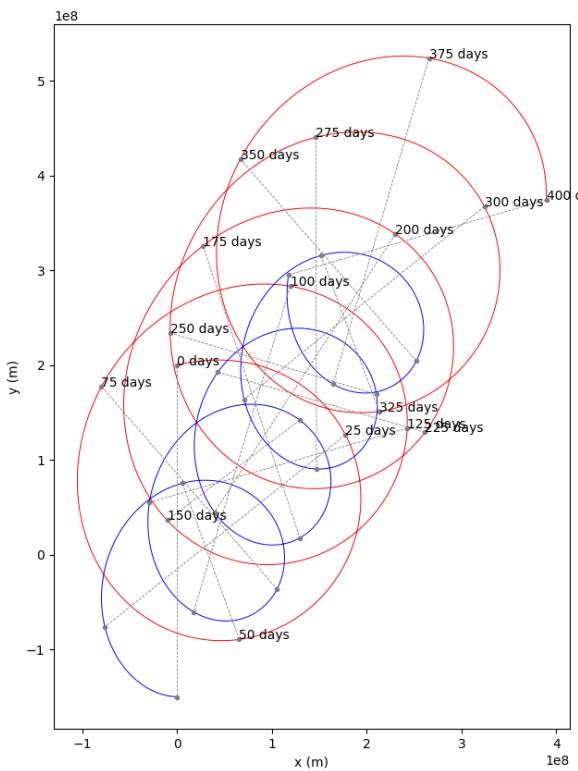


Now you can get a feel for what happened: The moons were attracted to each other by gravity and started to circle each other. The heavier moon accelerates less quickly, thus it makes a smaller loop.

Maybe we will get a better feel for what is happening if we look at more time. Let's increase it to 400 days. Change the relevant constant:

```
MAX_TIME = 400 * SEC_PER_DAY # 100 days
```

Now it should look like this:



Now you can see the pattern: They are rotating around each other and the pair is gradually migrating up and to the right.

51.5 Conservation of Momentum

You are observing a really important idea: the momentum of a system will be conserved. That is, absent forces from outside the system, the velocity of the center of mass will not change.

We can compute the initial center of mass and its velocity. In both cases, we just do a weighted average using the mass of the moon as the weight.

Immediately after you initialize the state of two moons, calculate the initial center of mass and its velocity:

```
# Calculate the initial position and velocity of the center of mass
tm = m1["mass"] + m2["mass"] # Total mass
cm_position = (m1["mass"] * m1["position"] + m2["mass"] * m2["position"]) / tm
cm_velocity = (m1["mass"] * m1["velocity"] + m2["mass"] * m2["velocity"]) / tm
```

Let's record the center of mass for each time. Before the loop starts, create a list to hold them:

```
cm_log = []
```

Inside the loop (before any calculations), append the current center of mass position to the log:

```
cm_log.append(cm_position)
```

Anywhere later in the loop (after you update the positions of the moon), update cm_position:

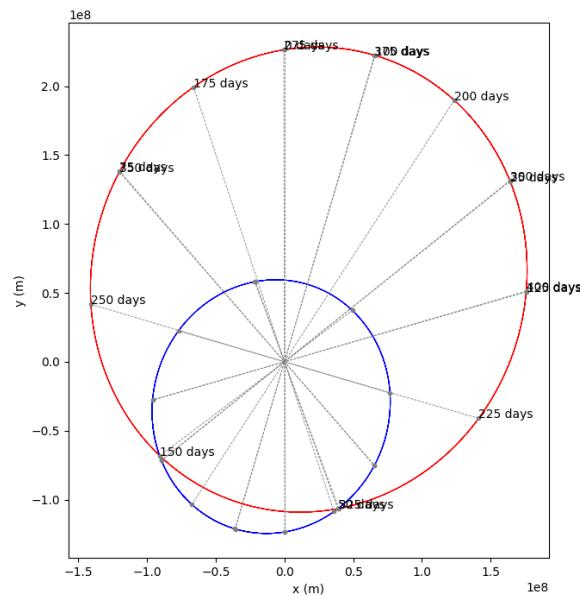
```
# Update the center of mass
cm_position = cm_position + cm_velocity * TIME_STEP
```

Now let's look at the positions of the moons relative to the center of mass. Before you do any plotting, convert the list to a numpy array and subtract it from the positions:

```
cms = np.array(cm_log)

# Make positions relative to the center of mass
positions1 = positions1 - cms
positions2 = positions2 - cms
```

Now when you run it you can really see what is happening:



The moons are tracing elliptical paths. The center of mass is the focus point for both of them.

51.6 Animation

One of the features of matplotlib that not a lot of people understand is how to make animations with it. This seems like a really great opportunity to make an animation showing the position, velocity, acceleration of the moons. We will also show the center of mass.

The trick to animations is that you create a bunch "artist" objects. You create a function that updates the artists. matplotlib will call your functions, tell the artists to draw themselves, and make a movie out of that.

Make a copy of `moons.py` called `animate_moons.py`.

Edit it to look like this:

```
import numpy as np
import matplotlib.pyplot as plt

# Import animation support and artists
from matplotlib.animation import FuncAnimation
from matplotlib.patches import Circle, FancyArrow
from matplotlib.text import Text

# Constants
G = 6.67430e-11 # Gravitational constant (Nm^2/kg^2)
SEC_PER_DAY = 24 * 60 * 60 # How many seconds in a day?
MAX_TIME = 400 * SEC_PER_DAY # 100 days
TIME_STEP = 12 * 60 * 60 # Update every 12 hours
FRAMECOUNT = MAX_TIME / TIME_STEP # How many frames in animation
ANI_INTERVAL = 1000 / 50 # ms for each frame in animation

# The velocity and acceleration vectors are invisible
# unless we scale them up. A lot.
VSCALE = 140000.0
ASCALE = VSCALE * 800000.0

# Create the initial state of Moon 1
m1 = {
    "mass": 6.0e22, # kg
    "position": np.array([0.0, 200_000_000]), # m
    "velocity": np.array([100.0, 25.0]), # m/s
    "radius": 1_500_000.0, # m
    "color": "red", # For plotting
}

# Create the initial state of Moon 2
m2 = {
    "mass": 11.0e22, # kg
    "position": np.array([0.0, -150_000_000]), # m
    "velocity": np.array([-45.0, 2.0]), # m/s
```

```
"radius": 2_000_000.0, # m
"color": "blue", # For plotting
}

# Calculate the initial position and velocity of the center of mass
tm = m1["mass"] + m2["mass"] # Total mass
cm_position = (m1["mass"] * m1["position"] + m2["mass"] * m2["position"]) / tm
cm_velocity = (m1["mass"] * m1["velocity"] + m2["mass"] * m2["velocity"]) / tm

# Start at time zero seconds
current_time = 0.0

# Create the figure and axis
fig, ax = plt.subplots(1, figsize=(7.2, 10))

# Set up the axes
ax.set_xlabel("x (m)")
ax.set_xlim((-1.2e8, 4e8))
ax.set_ylabel("y (m)")
ax.set_ylim((-1.6e8, 5.5e8))
ax.set_aspect("equal", adjustable="box")
fig.tight_layout()

# Create artists that will be edited in animation
time_text = ax.add_artist(Text(0.03, 0.95, "", transform=ax.transAxes))
circle1 = ax.add_artist(Circle((0, 0), radius=m1["radius"], color=m1["color"]))
circle2 = ax.add_artist(Circle((0, 0), radius=m2["radius"], color=m2["color"]))
circle_cm = ax.add_artist(Circle((0, 0), radius=2e8, color="purple"))
varrow1 = ax.add_artist(FancyArrow(0, 0, 0, 0, color="green", head_width=m1["radius"]))
varrow2 = ax.add_artist(FancyArrow(0, 0, 0, 0, color="green", head_width=m2["radius"]))
acc_arrow1 = ax.add_artist(
    FancyArrow(0, 0, 0, 0, color="purple", head_width=m1["radius"]))
)
acc_arrow2 = ax.add_artist(
    FancyArrow(0, 0, 0, 0, color="purple", head_width=m2["radius"]))
)

# This function will get called for every frame
def animate(frame):

    # Global variables needed in scope from the model
    global cm_position, cm_velocity, current_time, m1, m2

    # Global variables needed in scope from the artists
    global time_text, varrow1, varrow2, acc_arrow1, acc_arrow2, circle1, circle2, circle_cm

    print(f"Updating artists for day {current_time/SEC_PER_DAY:.1f}.")

    # Update the positions based on the current velocities
    m1["position"] = m1["position"] + m1["velocity"] * TIME_STEP
    m2["position"] = m2["position"] + m2["velocity"] * TIME_STEP
```

```
# Update day label
time_text.set_text(f"Day {current_time/SEC_PER_DAY:.0f}")

# Update positions of circles
circle1.set_center(m1["position"])
circle2.set_center(m2["position"])

# Update velocity arrows
varrow1.set_data(
    x=m1["position"][0],
    y=m1["position"][1],
    dx=VSCALE * m1["velocity"][0],
    dy=VSCALE * m1["velocity"][1],
)
varrow2.set_data(
    x=m2["position"][0],
    y=m2["position"][1],
    dx=VSCALE * m2["velocity"][0],
    dy=VSCALE * m2["velocity"][1],
)

# Update the center of mass
cm_position = cm_position + cm_velocity * TIME_STEP
circle_cm.set_center(cm_position)

# Find the vector from moon1 to moon2
delta = m2["position"] - m1["position"]

# What is the distance between the moons?
distance = np.linalg.norm(delta)

# Have the moons collided?
if distance < m1["radius"] + m2["radius"]:
    print(f"*** Collided {current_time:.1f} seconds in!")

# What is a unit vector that points from moon1 toward moon2?
direction = delta / distance

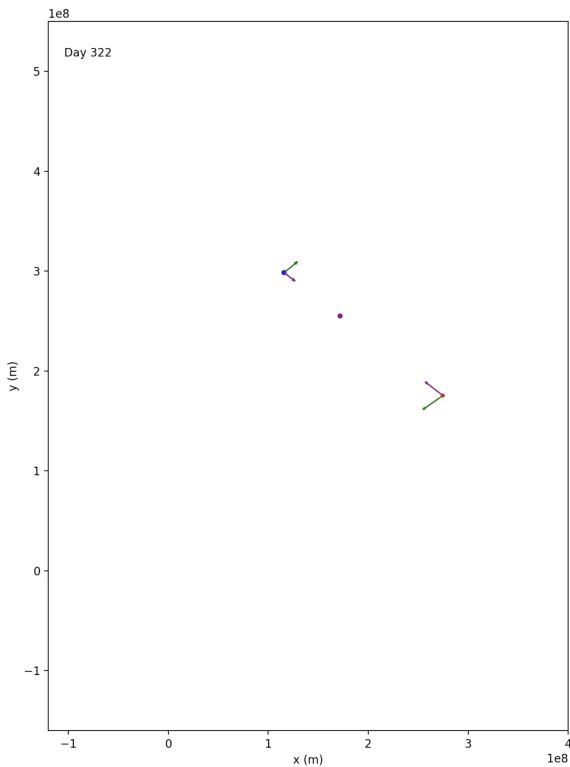
# Calculate the magnitude of the gravitational attraction
magnitude = G * m1["mass"] * m2["mass"] / (distance**2)

# Acceleration vector of moons (a = f/m)
acceleration1 = direction * magnitude / m1["mass"]
acceleration2 = (-1 * direction) * magnitude / m2["mass"]

# Update the acceleration arrows
acc_arrow1.set_data(
    x=m1["position"][0],
    y=m1["position"][1],
    dx=ASCALE * acceleration1[0],
    dy=ASCALE * acceleration1[1],
```

```
)  
acc_arrow2.set_data(  
    x=m2["position"][0],  
    y=m2["position"][1],  
    dx=ASCALE * acceleration2[0],  
    dy=ASCALE * acceleration2[1],  
)  
  
# Update the velocity vectors  
m1["velocity"] = m1["velocity"] + acceleration1 * TIME_STEP  
m2["velocity"] = m2["velocity"] + acceleration2 * TIME_STEP  
  
# Update the clock  
current_time += TIME_STEP  
  
# Return the artists that need to be redrawn  
return (  
    time_text,  
    varrow1,  
    varrow2,  
    acc_arrow1,  
    acc_arrow2,  
    circle1,  
    circle2,  
    circle_cm,  
)  
  
# Make the rendering happen  
animation = FuncAnimation(  
    fig,  
    animate,  
    np.arange(FRAMECOUNT),  
    interval=ANI_INTERVAL  
)  
  
# Save the rendering to a video file  
animation.save("moonmovie.mp4")
```

When you run this, it will take longer than the previous versions. You should have a video file that shows a simulation of the moons tracing their elliptical paths around their center of mass:



51.7 Challenge: The Three-Body Problem

It is time to stretch a little as a physicist and programmer: You are going to make a new version of `moons.py` that handles three moons instead of just two.

This is known as "The Three-Body Problem," and people have tried for centuries to come up with a way to figure out (from the initial conditions) where the three moons would be at time t without doing a simulation. And no one has.

For a lot of problems, the outcome is not very sensitive to the initial conditions. For example, the flight of a cannonball: If it leaves the muzzle of the cannon a little faster, it will go a little farther.

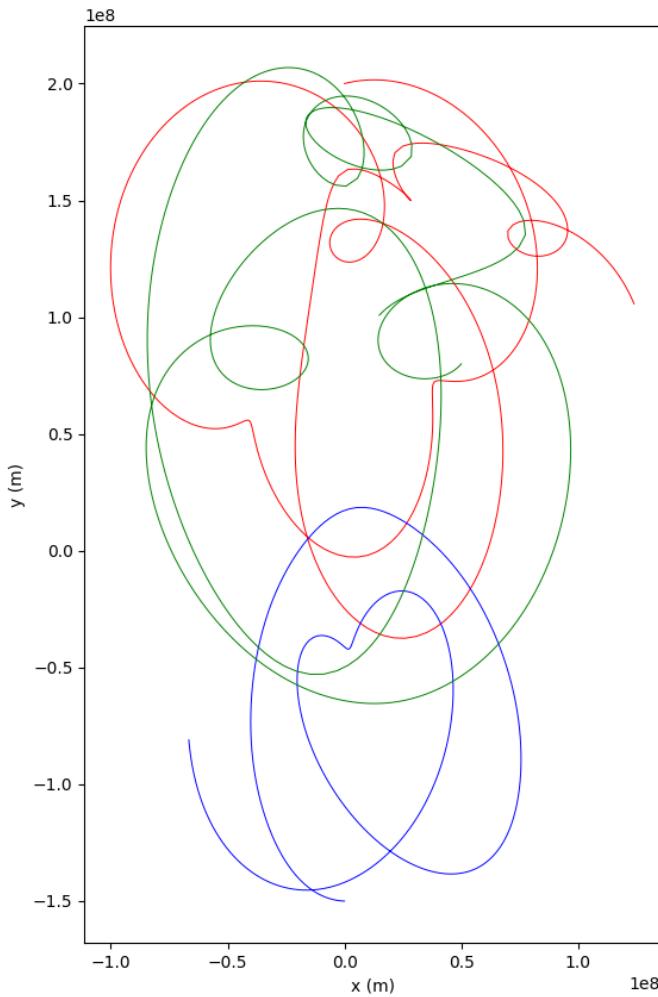
For the three-body problem, the outcome can be radically different even if the initial conditions are very similar.

(There is a whole field of mathematics studying systems that are very sensitive to initial conditions. It is known as *dynamical systems* or *chaos theory*.

Copy `moons.py` to `3moons.py`. Here is a reasonable initial state for your third moon:

```
m3 = {
    "mass": 4.0e22, # kg
    "position": np.array([50_000_000, 80_000_000]), # m
    "velocity": np.array([-30.0, -35.0]), # m/s
    "radius": 1_700_000.0, # m
    "color": "green"
}
```

If I run that simulation for 100 days, I get a plot like this:



Visibly you can see this is very different from the two-body problem that just traced ellipses around the center of mass.



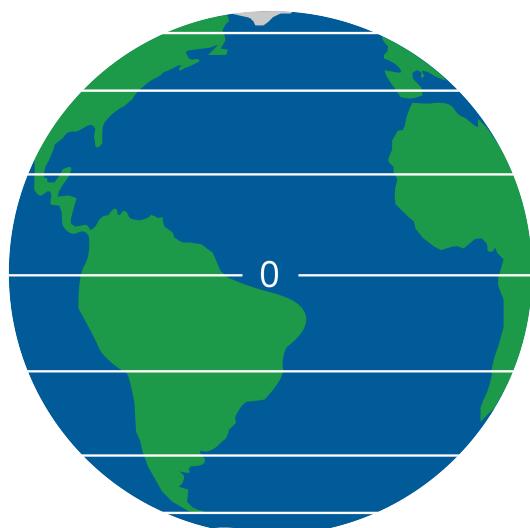
CHAPTER 52

Longitude and Latitude

The Earth can be represented as a sphere, and the position of a point on its surface can be described using two coordinates: latitude and longitude.

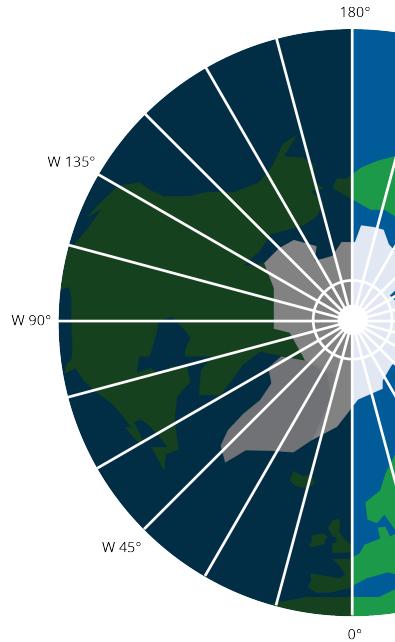
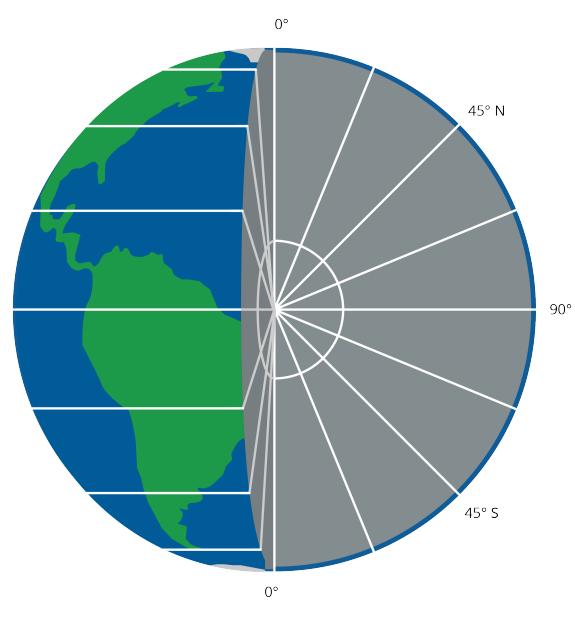
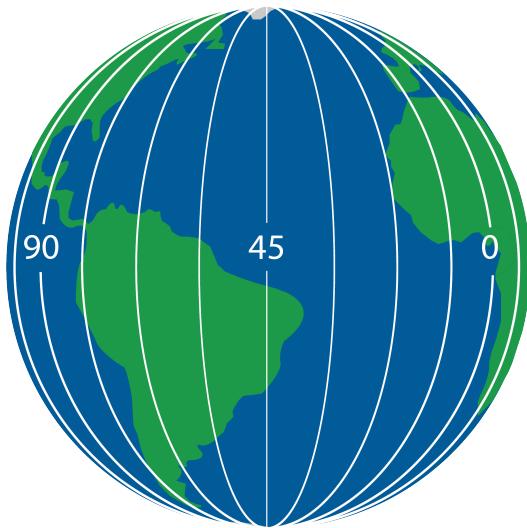


Latitude is a measure of a point's distance north or south of the Equator, expressed in degrees. It ranges from -90° at the South Pole to $+90^\circ$ at the North Pole, with 0° representing the Equator.



Longitude, on the other hand, measures a point's distance east or west of the Prime Merid-

ian (which passes through Greenwich, England). It ranges from -180° to $+180^{\circ}$, with the Prime Meridian represented as 0° .

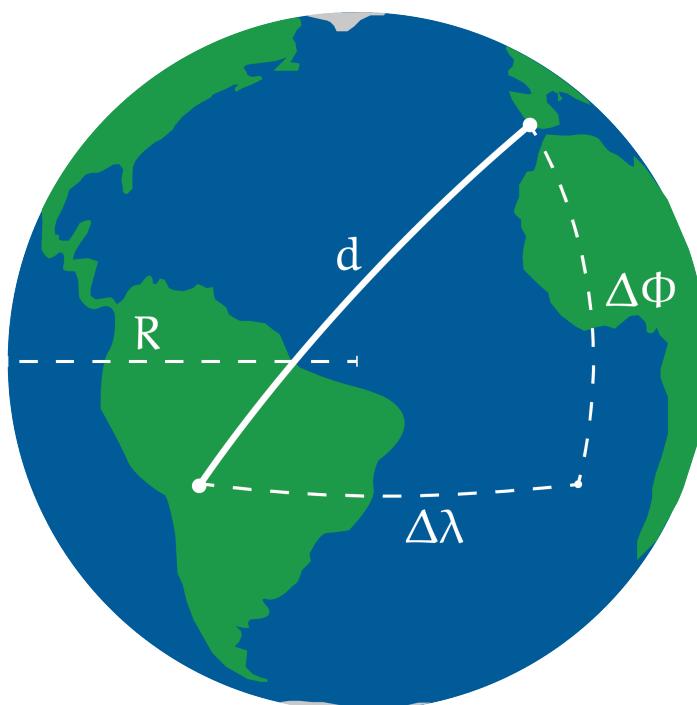


52.1 Nautical Mile

A nautical mile is a unit of measurement used primarily in aviation and maritime contexts. It is based on the circumference of the Earth and is defined as one minute ($1/60^\circ$) of latitude. This makes it directly related to the Earth's geometry, unlike a kilometer or a mile, which are arbitrary in nature. The exact value of a nautical mile can vary slightly depending on which type of latitude you use (e.g., geodetic, geocentric, etc.), but for practical purposes, it's often approximated as 1.852 kilometers or 1.15078 statute miles.

52.2 Haversine Formula

The haversine formula is an equation important in navigation for giving great-circle distances between two points on a sphere from their longitudes and latitudes. It's especially useful when it comes to calculating distances between points on the surface of the Earth, which we represent as a sphere for simplicity.



In its basic form, the haversine formula is as follows:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R \cdot c$$

Here, ϕ represents the latitudes of the two points (in radians), $\Delta\phi$ and $\Delta\lambda$ represent the differences in latitude and longitude (also in radians), and R is the radius of the Earth. The result, d , is the distance between the two points along the surface of the sphere.



CHAPTER 53

Tides and Eclipses

You live with a lot of orbital paths:

- The earth is spinning. If you are standing at the equator, you are traveling at 1,674 km per hour around the center of the planet. We are all spinning east, thus the sun comes up in the east and sets in the west.
- The earth is orbiting the sun. It takes 365.242 days for the earth to go once around the sun. This is why different constellations appear at different times during the year – we only see the stars at night and the direction of night shifts as the earth moves around the sun.
- The moon is orbiting the earth. The moon travels once around the earth once every 27.3 days.

You can see the effects of these orbits on our planet. Let's go over a few.

53.1 Leap Years

Note that it takes 365.242 days for the earth to go around the sun. If we declared "The calendar will *always* be 365 days per year!" then gradually the seasons would shift by 0.242 days every year. After a century, they would have migrated 24 days.

So, we made a rule: "Every fourth year, we will add an extra day to the calendar!" The years 2021, 2022, and 2023 get no February 29th. 2024 gets a February 29th.

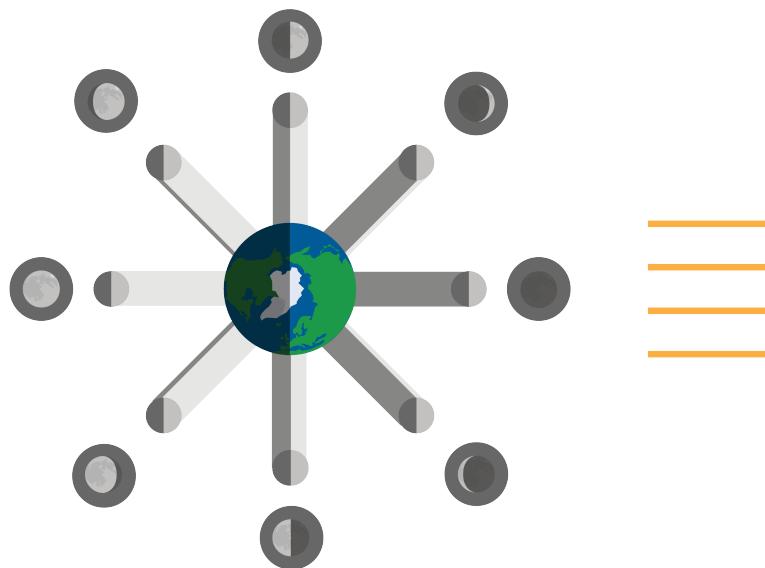
That got us a calendar with an average 365.25 days per year, so the seasons would not have migrated as quickly, but they still would have migrated about three days every four hundred years.

So, we made another rule: "There will be no February 29th in the three century years (multiples of 100) that are not multiples of 400." So the year 1900 had no Feb 29, but the year 2000 had one. Now the average number of days per year is 365.2425.

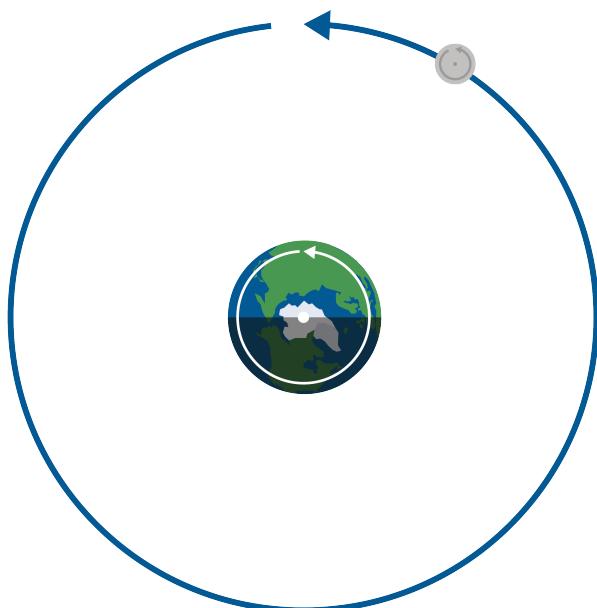
53.2 Phases of the Moon

The earth, the moon, and the sun form a triangle. If you were standing on the moon, you could measure the angle between the light coming from the sun and the the light going to the earth. That angle would fluctuate between 0 degrees and 180 degrees.

- When the angle was close to 0, the people on earth would see a full moon.
- When the angle was close to 90 degrees, the people on earth would see a half moon.
- When the angle was nearing 180 degrees, the people on earth would see a slim crescent.
- When the angle was very close to 180 degrees, the moon would be dark. This is called a "new moon."



Even though it takes 27.3 for the moon to travel around the earth once, it takes 29.5 days to get from one full moon to the next. Why? In the 27.3 days that it took the moon to travel around the earth, the earth has moved about 17 degrees around the sun. To get back into the same triangle configuration takes another 2.2 days.





To explain why we often see a curve in the shadow of the moon, we can look at a ball that has one side painted yellow and the other red.

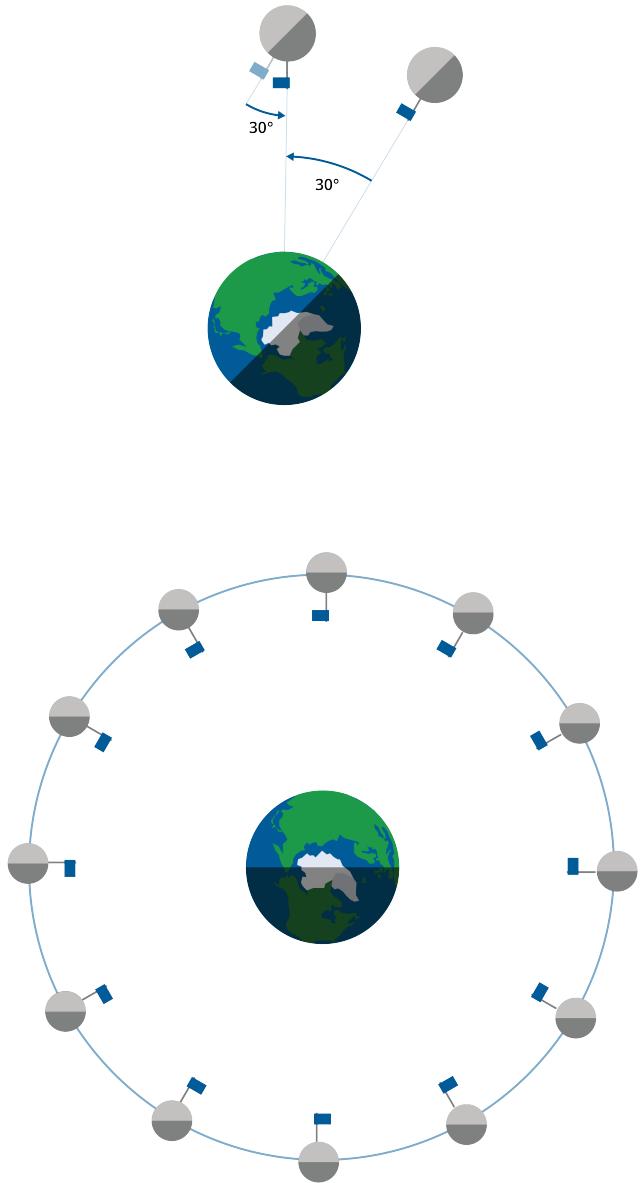


As we rotate the ball, we can see that the straight color boundary between each hemisphere begins to look curved. The curve we see in the moon is due to this same basic principle of how the shading of spheres works.

FIXME: Add text about scale In all of these graphics, we've been using incorrect scale. Here's the true scale of the distance of the earth and the moon with accurate radii.



FIXME: Add text about tidal lock



53.3 Eclipses

While the earth orbits the sun and the moon orbits the earth, the two orbits are *not in the same plane*. We call the plane that the earth orbits the sun in the *ecliptic plane*. The plane of the moon's orbit is about 5 degrees tilted from the ecliptic plane.

Note that the moon passes through the ecliptic plane only twice every 27.3 days. Imagine

that at the instant it passed through the ecliptic plane was also the precise instant of a full moon. The sun, the earth, and the moon would be in a straight line! The earth would cast a shadow upon the moon – it would go from a bright full moon to a dark moon until the moon moved back out of the shadow of the earth. This is known as a *lunar eclipse*.

The diameter of the moon is a little more than a quarter the diameter of the earth, so they don't have to be in perfect alignment for the moon to be darkened. Lunar eclipses actually happen once or twice per year.

Now imagine that at the instant the moon passed through the ecliptic plane was also the precise instant of a new moon. The sun, the moon, and the earth would be in a straight line! The moon would cast a shadow upon some part of the earth. To a person in that shadow, the sun disappear behind the moon. This is known as a *solar eclipse*.

The sun is pretty big, so if the moon blots out just part of it, we call it a *partial solar eclipse*. There are a few partial solar eclipses every year. Note that because the moon's shadow is too small to shade the whole earth, only certain parts of the world will experience any solar eclipses.

Every 18 months or so, there is a total eclipse of the sun. Once again, only certain parts of the world experience it. You can expect to experience a total eclipse of the sun at your home about once every 375 years.

53.4 The Far Side of the Moon

Like the earth, the moon spins on its axis. Due to earth's gravity, the rotation of the moon slowed down until its spin matched the rate it orbits earth. That is: we are always looking at the same side of the moon. Until we orbited the moon, we had no idea what the far side looked like.

Some people call it "The Dark Side of the Moon," but it gets just as much sunshine as the side that faces earth. The name comes from the fact that we lose communication with spacecraft (like the Apollo missions to the moon) when they are on the far side of the moon. When we lose communications with a craft, we often say "It went dark."

53.5 Tides

When we say "The moon orbits the earth," that is a bit of an oversimplification. The force of gravity that pulls the moon toward the earth, also pulls the earth toward the moon. The earth is about 81 times heavier than the moon, so the moon moves more, but the moon definitely moves the earth.

The center of the moon and the center of the rotate around each other. The point they

rotate around is inside the the earth, but it is closer to the surface of the earth than it is to the center of the earth.

Orbits happen, remember, when the centripetal force is equal to gravitational force. So the centripetal force created by the earth being swung by the moon is equal to the gravitational force that the moon exerts on all the mass on the moon.

However.

The parts of the earth that are closer to the moon experience less centripetal force (away from the moon) and more gravitational force (toward the moon).

The parts of the earth that are farther from the moon experience more centripetal force (away from the moon) and less gravitational force (away from the moon).

The effects are not big. For example, you won't notice that you can jump higher when the moon is overhead. You will lose only about 1/200,000 of your weight.

But the ocean is huge.: 1/200,000 of its weight is a lot of force.

The water in the oceans bulges a little both toward the moon and away from it.

The earth is still rotating. If you are at the beach as your longitude slides into one of these bulges, you say "Hey, the tide is rising!" The peak of these bulges is known as "high tide". Because there is a bulge on each side of the planet, high tide comes twice a day.

This is a lunar tide – because it is caused by the moon. There is a similar effect from the sun, but the sun is very, very far away: solar tidal forces are about half as powerful lunar tidal forces. When the sun and the moon work together, the tides are stronger. This is called a *spring tide*. Spring tides don't happen in the spring time; they happen close to full moons and new moons.

When the moon and the sun are working against each other, the tides are weaker. This is called a *neap tide*. Neap tides happen when you see a half moon in the sky.

53.5.1 Computing the Forces

We are enumerating several forces that shape the water on the planet. All these forces are pulling on your body too. In these exercises, you are going to calculate how each force would effect a 1 kg mass on the surface of the earth.

Here are some numbers you will need:

- The mass of the earth: 5.97219×10^{24} kg

- The mass of the sun: 1.9891×10^{30} kg
- The mass of the moon: 7.347673×10^{22} kg
- Radius of the earth at the equator: 6,371 km
- Average distance from the center of the earth to the center of the sun: 149.6×10^6 km
- Average distance from the center of the moon to the center of the earth: 384,467 km.

Exercise 59 Life Among the Orbits 1: Earth Gravity

Working Space

If the earth were still and alone in the universe, there would still be the force of gravity. We have said that that a kilogram on the surface of the earth is pulled toward the center of the earth with a force of 9.8 N.

Confirm that the gravity of the earth pulls a 1kg mass on the surface of the planet with a force of about 9.8 N.

You will need the formula for gravitation:

$$F_g = \frac{gm_1m_2}{r^2}$$

If we measure distance in km and mass in kg, the gravitation constant g is 6.67430×10^{-17} .

Answer on Page 828

Exercise 60 Life Among the Orbits 2: Earth Centripetal Force*Working Space*

What if we add the spinning of the earth? The spinning would try to throw the kg into space. The formula for centripetal force is

$$F_c = \frac{mv^2}{r}$$

Calculate the centripetal force on a 1 kg mass on the surface of the earth. It doesn't fly off into space, so the force due to gravity must be bigger. How many times bigger?

Assume that the mass is on the equator, thus rotating around the earth at 465 m/s.

Does the centripetal force increase, decrease, or stay the same as you get closer to the north pole?

*Answer on Page 828***Exercise 61 Life Among the Orbits 3: The Moon's Gravity***Working Space*

Now we add the moon's gravitational force to our model.

When the moon is directly overhead, how strongly will it pull at the 1 kg mass on the equator?

When the moon is directly underfoot, how strongly will it pull at the 1 kg mass on the equator?

Is that a big difference?

Answer on Page 828

Exercise 62 Life Among the Orbits 4: The Swing of the Moon*Working Space*

Now we add the moon's motion. The moon and the earth swing each other around. This creates a centripetal force. They both travel in nearly a circle centered at their center of mass.

How far is the center of mass of the moon and the earth from the center of the earth? (You can imagine a see-saw with the center of the earth on one end and the center of the moon on the other. Where would the balance point be?)

What point on the surface of the earth is closest to the center of mass? How far is it?

What point on the surface of the earth is farthest from the center of mass? How far is it?

Answer on Page 829

Exercise 63 Life Among the Orbits 5: Lunar Centripetal Force**Working Space**

The moon swings us around that center of mass once every 27.3 days. (Forget about the spinning of the earth for this part.) What is the largest and smallest centripetal forces on the surface of the earth created by this swinging

What is the largest centripetal force on a 1 kg mass with the moon directly underfoot? (You need an answer from the previous question: There is a point on the surface of the earth that is 11,044,000 m from the center of gravity.)

What is the resulting centripetal force on a 1 kg mass with the moon directly overhead? (You will need the other answer from the previous exercise: That point is 1,698,000 m from the center of mass of the moon and the earth.)

For this problem is probably easier to use this formula for centripetal force:

$$F_c = mr\omega^2$$

Where m is mass in kg, r is radius in m, and ω is the angular velocity in radians per second.

Answer on Page 829

Exercise 64 Life Among the Orbits 6: Net Force*Working Space*

Now add together the two forces at both the nearest point to the moon and the farthest.

*Answer on Page 830***53.5.2 Solar Tidal Forces**

The sun has a much larger gravitational effect on the earth than the moon does:

- When the sun is overhead, it will pull on a 1 kg mass with a force of about 0.00593 N.
- When the moon is overhead, it will pull on a 1 kg mass with a force of about 0.0000343 N.

Why are lunar tides about twice as powerful solar tides?

Tides occur because the pull of gravity and the pull of the centripetal force are out of balance somewhere on the planet. The sun is so far away that the effects of gravitational and centripetal forces are very close to equal everywhere on earth.



CHAPTER 54

Electromagnetic Waves

Sound is a compression wave – to travel, it needs a medium to compress: air, water, etc. (Regardless of what you have seen in movies, sound does not travel through a vacuum)

Light is an electromagnetic wave – it causes fluctuations in the electric and magnetic fields that are everywhere. It can cross a vacuum, as it does to reach us from the sun.

Electromagnetic waves travel at about 300 million meters per second in a vacuum. The waves travel slower through things. For example, an electromagnetic wave travels at 225 million meters in water.

Electromagnetic waves come in different frequencies. For example, the light coming out of a red laser pointer is usually about 4.75×10^{14} Hz. The wifi data sent by your computer is carried on an electromagnetic wave too. It is usually close to 2.4×10^6 Hz or 5×10^6 Hz.

Because we know how fast the waves are moving, we sometimes talk about their wavelengths instead of their frequencies. The light coming out of a laser pointer is $300 \times 10^6 / 4.76 \times 10^{14} = 630 \times 10^{-9}$ m, or 630 nm.

Exercise 65 Wavelengths*Working Space*

A green laser pointer emits light at 5.66×10^{14} Hz. What is its wavelength in a vacuum?

Answer on Page 830

We have given names to different ranges of the electromagnetic spectrum:

Name	Hertz	Meters
Gamma rays	$\times 10$	$\times 10$
X-rays	$\times 10$	$\times 10$
Ultraviolet	$\times 10$	$\times 10$
Blue	$\times 10$	$\times 10$
Red	$\times 10$	$\times 10$
Infrared	$\times 10$	$\times 10$
Microwaves	$\times 10$	$\times 10$
Radio waves	$\times 10$	$\times 10$

(You may have heard of “cosmic rays” and wonder why they are not listed in this table. Cosmic rays are actually the nuclei of atoms that have been stripped of their electron cloud. These particles come flying out of the sun at very high speeds. They were originally thought to be electromagnetic waves, and they were mistakenly named “rays”.)

In general, the lower frequency the wave is, the better it passes through a mass. A radio wave, for example, can pass through the walls of your house, but visible light cannot. The people who designed the microwave oven, chose the frequency of 2.45 GHz because the energy from those waves tended to get absorbed in the first few inches of food that it passed through.

54.1 The greenhouse effect

Humans have dug up a bunch of long carbon-based molecules (like oil and coal) and burned them, releasing large amounts of CO₂ into the atmosphere. It is not obvious why that has made the planet warmer. The answer is electromagnetic waves.

A warm object gives off infrared electromagnetic waves. That’s why, for example, motion detectors in security systems are actually infrared detectors: even in a dark room, your

body gives off a lot of infrared radiation.

You may have heard of “heat-seeking missiles.” These are more accurately called “Infrared homing missiles” because they follow objects giving off infrared radiation – hot things like jet engines.

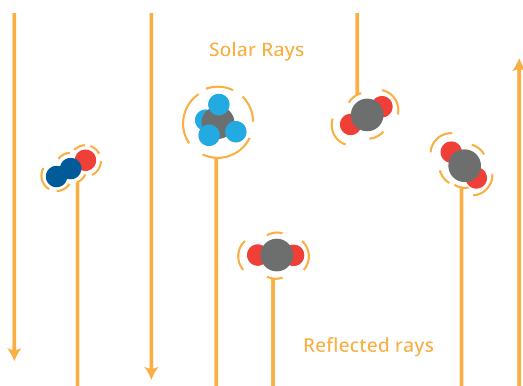
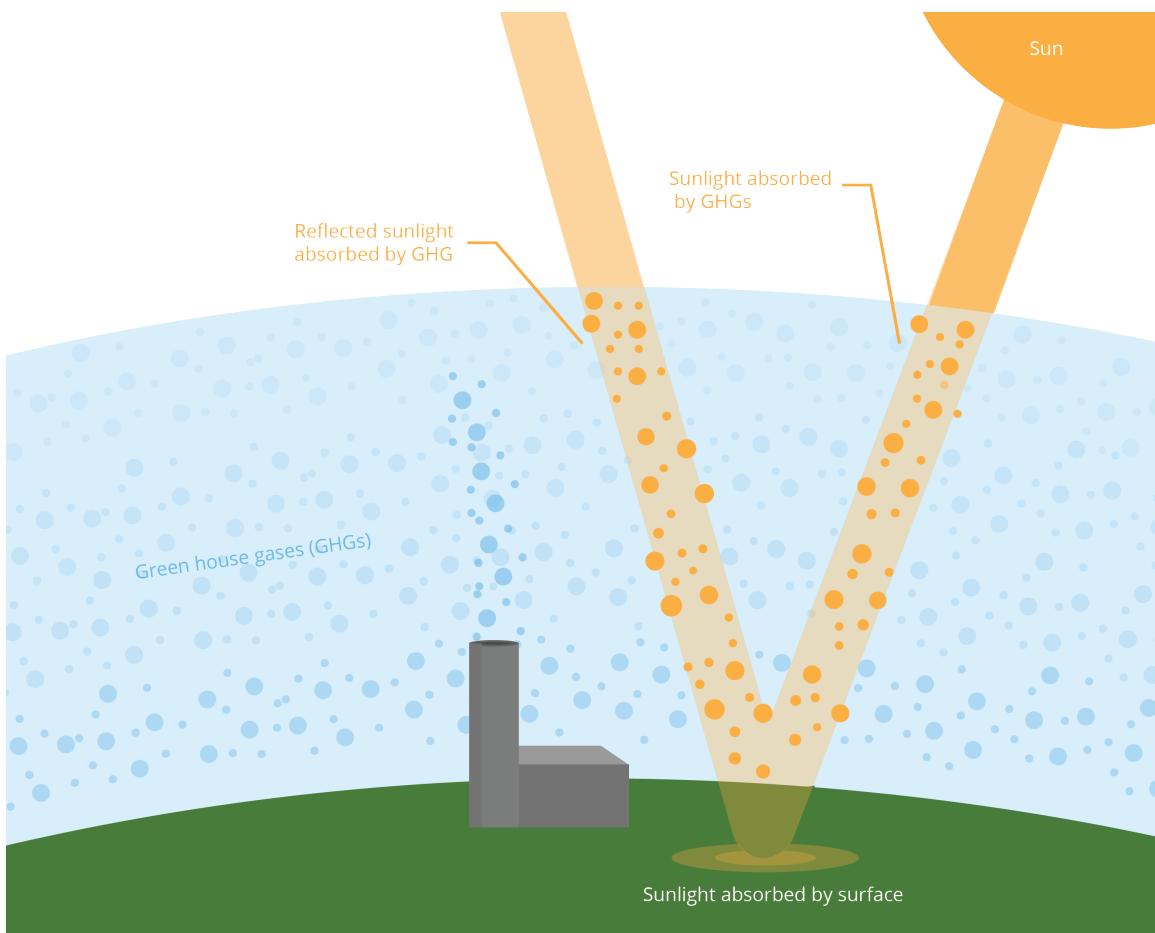
The sun beams a lot of energy to our planet in the form of electromagnetic radiation: visible light, infrared, ultraviolet. (How much? At the top of the atmosphere directly facing the sun, we get 1,360 watts of radiation per square meter. That is a lot of power!)

Some of that radiation just reflects back into space. 23% is reflected by the clouds and the atmosphere, 7% makes it all the way to the surface of the planet and is reflected back into space.

The other 71% is absorbed: 48% is absorbed by the surface and 23% is absorbed by the atmosphere. All of that energy warms the planet and the atmosphere so that it gives off infrared radiation. The planet lives in equilibrium: The infrared radiation leaving our atmosphere is exactly the same amount of energy as that 71% of the radiation that it absorbs.

(If the planet absorbs more energy than it releases, the planet gets hotter. Hotter things release more infrared. When the planet is in equilibrium again, it stops getting hotter.)

So what is the problem with CO₂ and other large molecules in the atmosphere? They absorb the infrared radiation instead of letting it escape into space. Thus the planet must be hotter to maintain equilibrium.



The planet is getting hotter, and it is creating a multitude of problems:

- Weather patterns are changing, which leads to extreme floods and droughts.
- Ice and snow in places like Greenland are melting and flowing into the oceans. This is raising sea-levels.
- Biomes with biodiversity are resilient. Rapidly changing climate is destroying biodiversity everywhere, which is making these ecosystems very fragile.

- In many places, permafrost, which has trapped large amounts of methane in the ground for millenia, is melting.

That last item is particularly scary because methane is a large gas molecule – it absorbs even more infrared radiation than CO₂. As it is escapes the permafrost, the problem will get worse.

Scientists are working on four kinds of solutions:

- **Stop increasing the amount of greenhouse gases in our atmosphere.** It is hoped that non-carbon based energy systems like solar, wind, hydroelectric, and nuclear could let us stop burning carbon. Given the methane already being released, it maybe too late for this solution to work on its own.
- **Take some of greenhouse gases out of our atmosphere and sequester them somewhere.** The trunk of a tree is largely carbon molecules. When you grow a tree where there had not been one before, you are sequestering carbon inside the tree. There are also scientists that are trying to develop process that pull greenhouse gases out of the air and turn them into solids.
- **Decrease the amount of solar radiation that is absorbed by our planet and its atmosphere.** Clouds reflect a lot of radiation back into space. Could we increase the cloudiness of our atmosphere? Or maybe mirrors in orbit around our planet?
- **Adapt to the changing climate.** These scientists are assuming that global warming will continue, and are working to minimize future human suffering. How will we relocate a billion people as the oceans claim their homes? When massive heat waves occur, how will we keep people from dying? As biodiversity decreases, how can we make sure that species that are important to human existence survive?

What are the greenhouse gases and how much does each contribute to keeping the heat from exiting to space? These numbers are still being debated, but this will give you a feel:

Water vapor	H ₂ O	36 - 72 %
Carbon dioxide	CO ₂	9 - 26 %
Methane	CH ₄	4 - 9 %
Ozone	O ₃	3 - 7 %

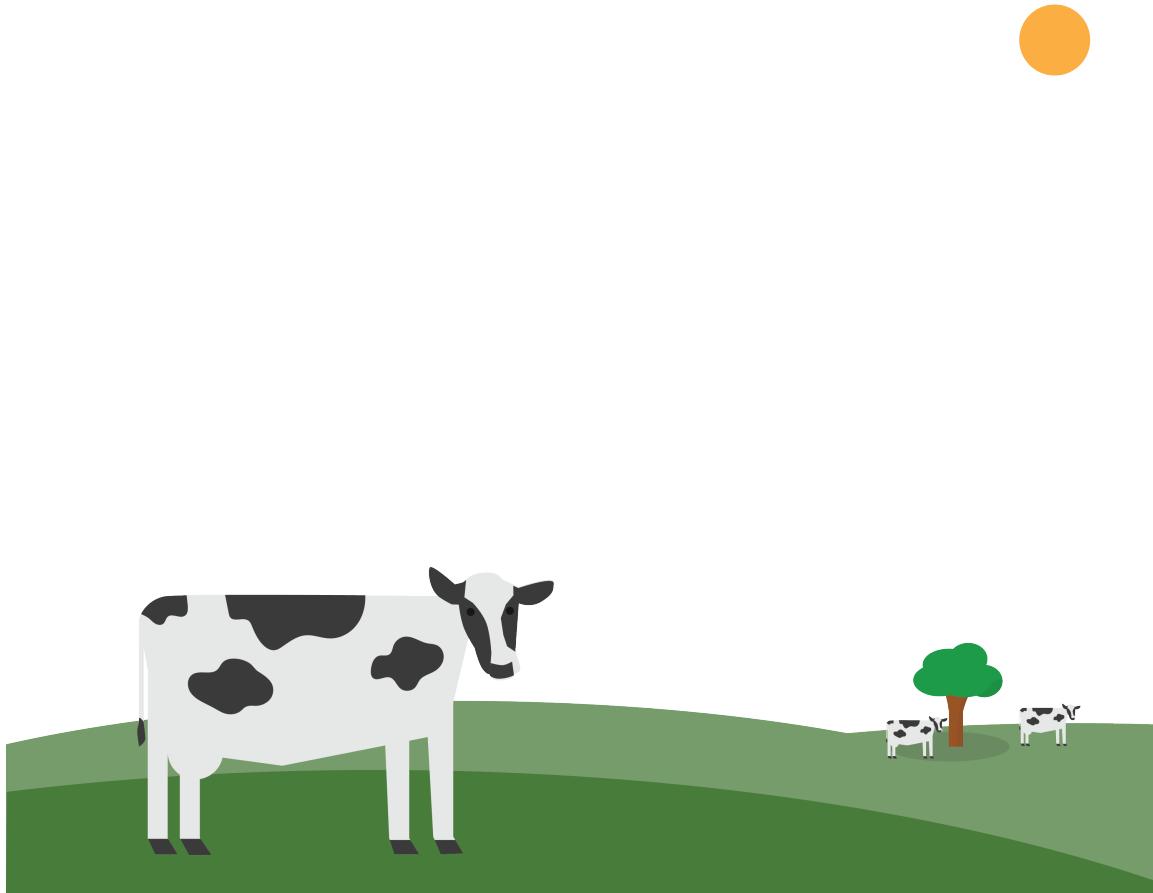
Notice that while we talk a lot about carbon dioxide, the most important greenhouse gas is actually water. Why don't we talk about it? Given the enormous surfaces of the oceans, it is difficult to imagine any way to permanently decrease the amount of water in the air. Also, a lot of water in the air is in the form of clouds that help reflect radiation before it is absorbed.



CHAPTER 55

How Cameras Work

Let's say it is a sunny day and you are standing in a field a few meters from a cow. You use the camera on your phone to take a picture of the cow. How does that whole process work?

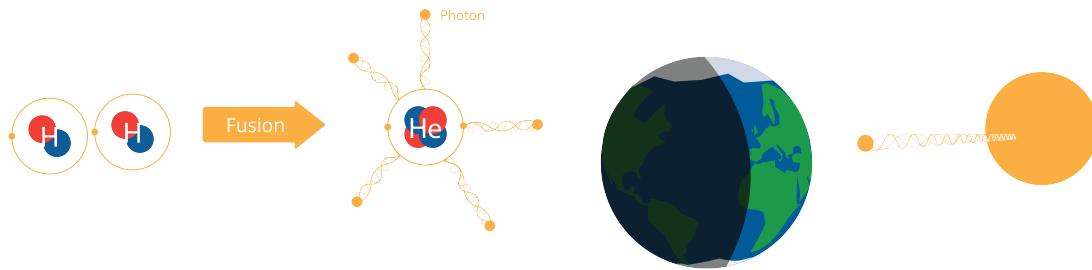


55.1 The Light That Shines On the Cow

The sun is a sphere of hot gas. About 70% of the gas is hydrogen. About 28% is helium. There's also a little carbon, nitrogen, and oxygen.

Gradually, the sun is converting hydrogen into helium through a process known as "nuclear fusion". (We will talk more about nuclear fusion in a later chapter.) A lot of heat is created in this process. The heat makes the gases glow.

How does heat make things glow? The heat pushes the electrons into higher orbitals. When they back down to a lower orbital, they release a photon of energy, which travels away from the atom as an electromagnetic wave.



Heat isn't the only way to push the electrons into a higher orbital. For example, a fluorescent lightbulb is filled with gas. When we pass electricity through the gas, its electrons are moved to a higher orbital. When they fall, light is created.

What is the frequency of the wave that the photon travels on? Depending on what orbital it falls from and how far it falls, the photon created has different amounts of energy. The amount of energy determines the frequency of the electromagnetic wave.

Formula for energy of a photon

If you want to know the amount of energy E in a photon, here is the formula:

$$E = \frac{hc}{\lambda}$$

where c is the speed of light, λ is the wavelength of the electromagnetic wave, and h Planck's constant: $6.63 \times 10^{-34} \text{ m}^2 \text{ kg/s}$

For example, a red laser light has a wavelength of about 630 nm. So the energy in each photon is:

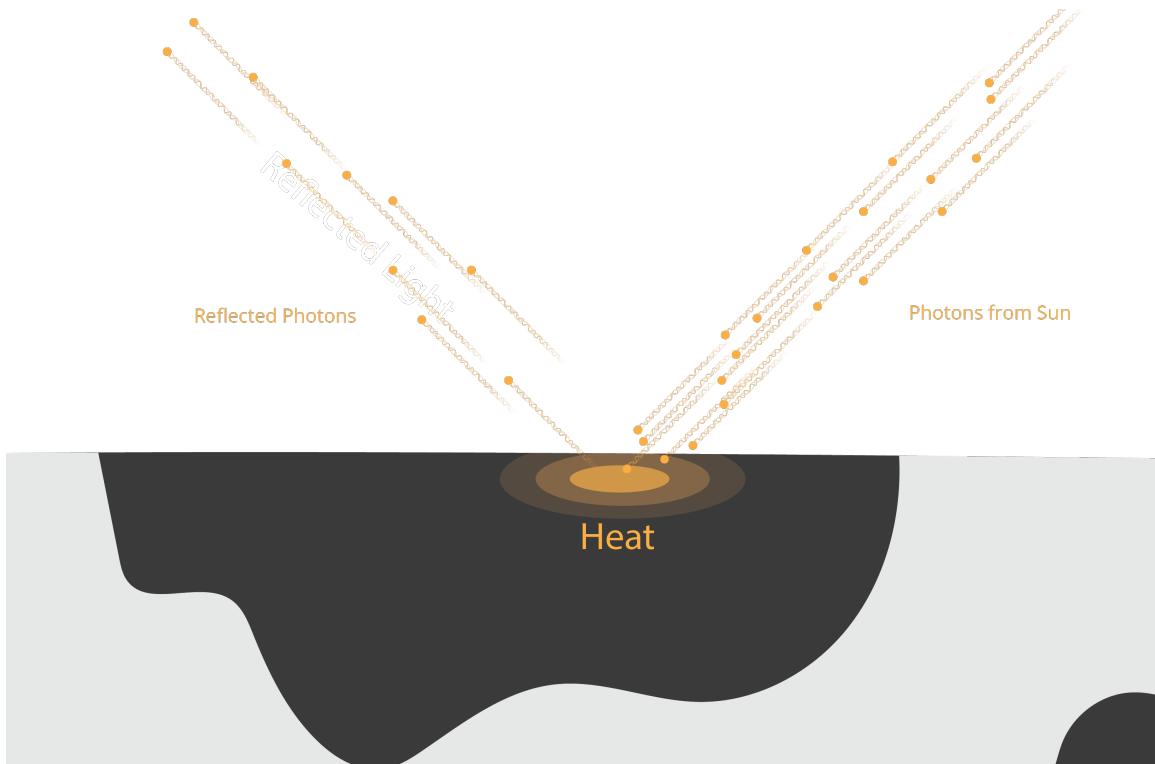
$$\frac{(300 \times 10^6)(6.63 \times 10^{-34})}{630 \times 10^{-9}} = 3.1 \times 10^{-19} \text{ joules}$$

In the sun, there are several kinds of molecules and each has a few different orbitals that the electrons can live in. Thus, the light coming from the sun is made up of electromagnetic waves of many different frequencies.

We can see some of these frequencies as different colors, but some are invisible to humans, for example ultraviolet and infrared.

55.2 Light Hits the Cow

When these photons from the sun hit the cow, the hide and hairs of the cow will absorb some of the photons. These photons will become heat and make the cow feel warm. Some of the photons will not be absorbed – they will leave the cow. When you say “I see the cow,” what you are really saying is “I see some photons that were not absorbed by the cow.”



Different materials absorb different amounts of each wavelength. A plant, for example, absorbs a large percentage of all blue and red photons that hit it, but it absorbs only a small percentage of the green photons that hit it. Thus we say “That plant is green.”

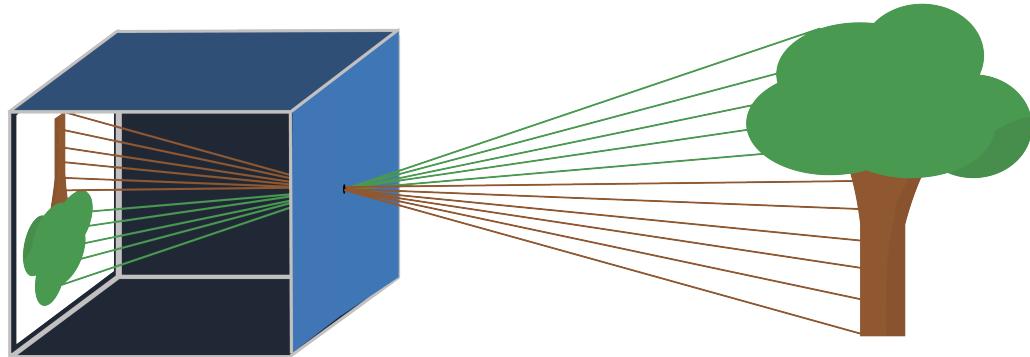
White things absorb very small percentages of photons of any visible wavelength. Black things absorb very *large* percentages of photons of any visible wavelength.

Before we go on, let’s review: The sun creates photons that travel as electromagnetic waves of assorted wavelengths to the cow. Many of those photons are absorbed, but some are not. Some of those photons that are not absorbed go into the lens of our camera.

55.3 Pinhole camera

The simplest cameras have no lenses. They are just a box. The box has a tiny hole that allows photons to enter. The side of the box opposite the hole is flat and covered with film or some other photo-sensitive material.

The photons entering the box continue in the same direction they were going when they passed through the hole. Thus, the photons that entered from high, hit the back wall low. The photons that came from the left, hit the back wall on the right. Thus the image is projected onto the back wall rotated 180 degrees: What was up is down, what was on the left is on the right.



Exercise 66 Height of the image**Working Space**

FIXME: cow swap

Let's say that that the pinhole is exactly the same height as the shoulder of the cow and that the shoulder is directly above one hoof. Then the pinhole, the shoulder, and the hoof form a right triangle.

Now, let's say that the camera is being held perpendicular to the ground. Now, the pinhole, the image of the shoulder, and the image of the hoof on the back wall of the camera also form a right triangle.

These two triangles are similar.

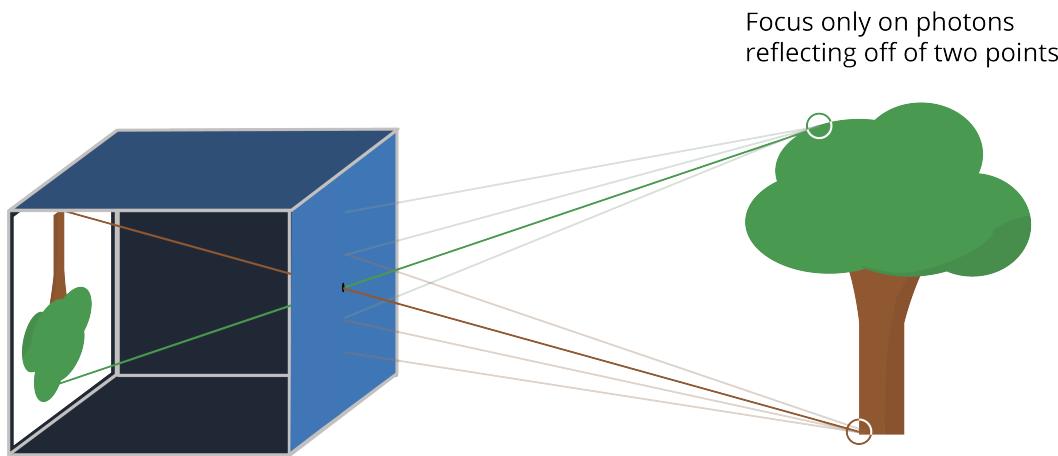
The shoulder is 2 meters from the hoof. The cow is standing 3 meters from the camera. The distance from the pinhole to the back wall of the camera is 3 cm. How tall is the image of the cow on the back wall of the camera?

Answer on Page 830**55.4 Lenses**

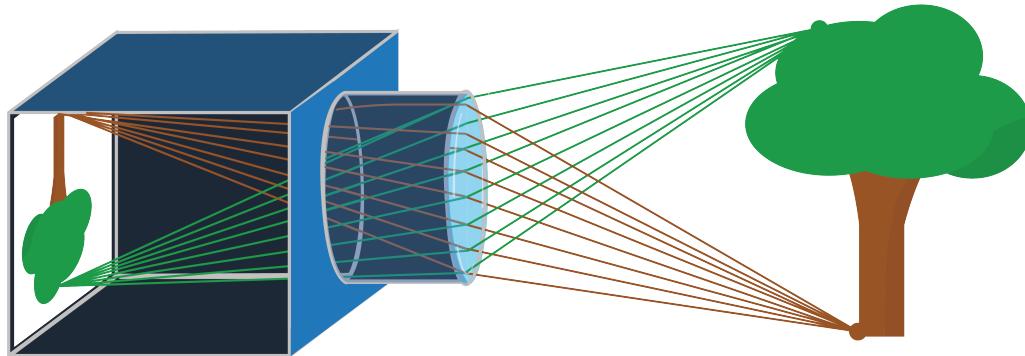
Quick review: A photon leaves the sun in some random direction. It travels 150 million km from the sun and hits a cow. It is not absorbed by the cow, and heads off in a new direction. It passes through the pinhole and hits the back wall of the camera. That seems incredibly improbable, right?

It actually is kind of improbable, especially if there isn't a lot of light – like you are taking the picture at dusk. To increase the odds, we added a *lens* to the camera.

If you focus a lens on a wall, and then you draw a dot on the wall. The lens is designed such that all the photons from the dot that hit the lens get redirected to the same spot on the back wall of the camera – regardless of which path it took to get to the lens.



Note that the image still gets flipped. There is a *focal point* that all the photons pass through.



The distance from the lens to its focal point is called the lens's *focal length*. Telephoto lenses, that let you take big pictures of things that are far away, have long focal lengths. Wide-angle lenses have short focal lengths.

55.5 Sensors

The camera on your phone has a sensor on the back wall of the camera. The sensor is broken up into tiny rectangular regions called pixels. When you say a sensor is 6000 by 4000 pixels, we are saying the sensor is a grid of 24,000,000 pixels: 6000 pixels wide and 4000 pixels tall.

Each pixel has three types of cavities that take in photos. One of the cavities measures the amount of short wavelength light, like blues and violets. One of the cavities measures the long wavelength light, like reds and oranges. One of the cavities measures the intensity of wavelengths in the middle, like greens.

Thus, if your camera has a resolution of 6000×4000 , the image is 72,000,000 numbers: Every one of the 24,000,000 pixels yeilds three numbers: intensity of long wavelength, mid wavelength, and long wavelength light. We call these numbers “RGB” for Red, Green, and Blue.



CHAPTER 56

How Eyes Work

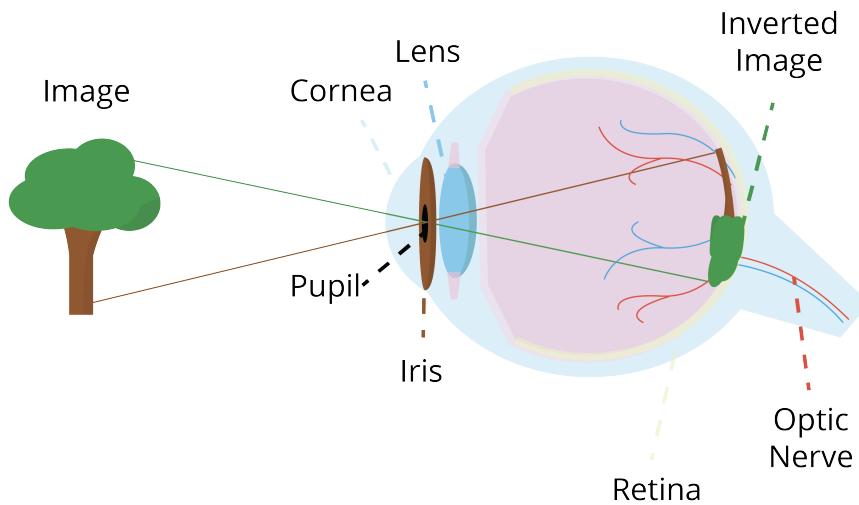
Dr. Craig Blackwell has made a great video on the mechanics of the eye. You should watch it: <https://youtu.be/Z8asc2SfFHM>

Mechanically, your eye works a lot like a camera. The eye is a sphere with two lenses on the front: The outer lens is called the *cornea*, and the second lens is just called “the lens.”

Between the two lenses is an aperture that opens wide when there is very little light, and closes very small when there is bright light. The opening is called the *pupil* and the tissue that forms the pupil is called the *iris*. When people talk of the color of your eyes, they are talking about the color of your iris. The blackness at the center of your iris is your pupil.

There are two types of photoreceptor cells in your retina: rods and cones. The rods are more sensitive; in very dark conditions, most of our vision is provided by the rods. The cones are used when there is plenty of light, and they let us see colors.

The white part around the outside of the eyeball? That is called the *sclera*.



The walls of the eye are lined inside with the *retina*, which has sensors that pick up the light and send impulses down the optic nerve to your brain.

Just like a camera, the images are flipped when they get projected on the back of the eye.

56.1 Eye problems

Now that you know the mechanics of the eye, let's enumerate a few things that commonly go wrong with the eye.

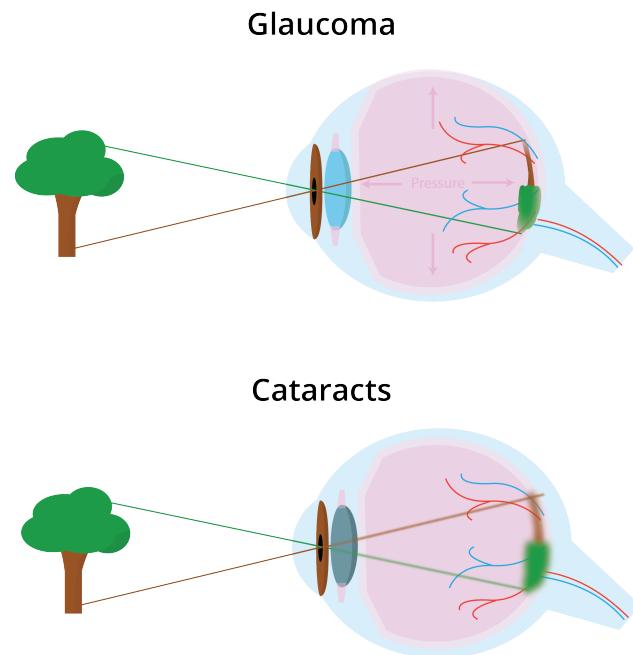
56.1.1 Glaucoma

The space between your cornea and lens is filled with a fluid called *aqueous humor*. To feed the cells of the cornea and lens, the aqueous humor carries oxygen and nutrients like blood would, but it is transparent so you can see. Aqueous humor is constantly being pumped into and out of that chamber. If aqueous humor has trouble exiting, the pressure builds up and can damage the eye. This is known as *glaucoma*.

56.1.2 Cataracts

The lens should be clear. As a person ages (and it can be accelerated by diabetes, too much exposure to sunlight, smoking, obesity, and high blood pressure), the proteins in the lens break down and clump together, becoming opaque. From the outside, the eye will look cloudy. This is called a *cataract*, and it makes it difficult for the person to see.

The problem can be corrected: The person's cloudy lens is removed and replaced with a



clear, manufactured lens.

56.1.3 Nearsightedness, farsightedness, and astigmatism

If you are in a dark room and a tiny LED is turned on, the photons from that LED can pass through your cornea in many different places. If your eye is focusing on that light correctly, all the photons should meet up at the same place on the retina.

FIXME: Diagram here

If the lenses are bending the light too much, the photons meet up before they hit the retina and get smeared a bit across the retina. To the person, the LED would appear blurry. The eye is said to be *nearsighted* or *myopic*.

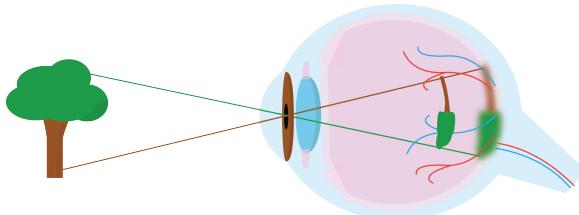
If the lenses are not bending it enough, the photons would meet up behind the retina. Once again, they get smeared a bit across the retina and the LED looks blurry to the

person. The eye is said to be *farsighted* or *hyperopic*.

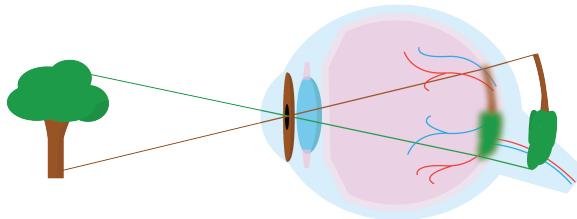
Your lenses are supposed to bend the photons the same amount vertically and horizontally. If one dimension is focused, but the other is myopic or hyperopic, the eye is said to have *astigmatism*.

Myopia, hyperopia, and astigmatism can be corrected with glasses or contact lenses. Doctors can also do surgical corrections, usually by changing the shape of the cornea.

Near Sighted



Far Sighted



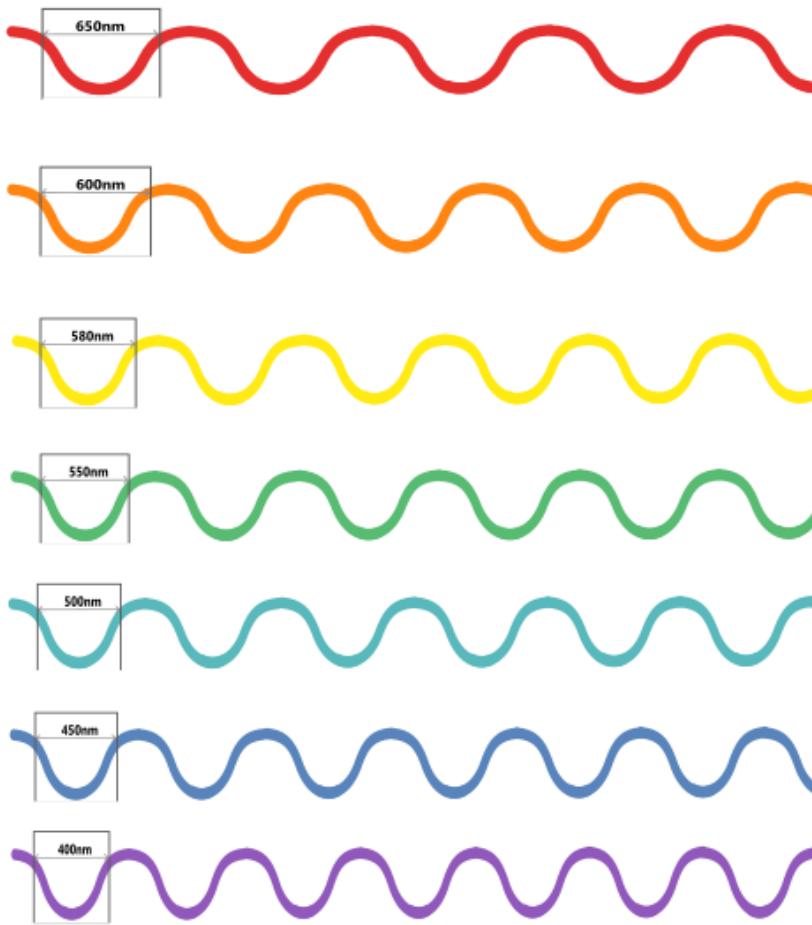
56.2 Seeing colors

TED-Ed has made a good video on how we see color. Watch it here: https://youtu.be/18_fZPHasd0

When a rainbow forms, you are seeing different wavelengths separating from each other. In the rainbow:

- Red is about 650 nm.
- Orange is about 600 nm.
- Yellow is about 580 nm.
- Green is about 550 nm.
- Cyan is about 500 nm.

- Blue is about 450 nm.
- Violet is about 400 nm.



If you shine a light with a wavelength of 580 nm on a white piece of paper, you will see yellow.

However, if you shine two lights with wavelengths of 650 nm (red) and 550 nm (green), you will also see yellow.

Why? Our ears can hear two different frequencies at the same time. Why can't our eyes see two colors in the same place?

As mentioned above, the cone photoreceptors in our eyes let us see colors. There are three kinds of cones:

- Blue: Cones that are most sensitive to frequencies near 450nm.
- Green: Cones that are most sensitive to frequencies near 550nm.
- Red: Cones that let us see the frequencies up to about 700nm.

When a wavelength of 580 nm hits your retina, it excites the red and green receptors, and your brain interprets that mix as yellow.

Similarly, when light that contains both 650 nm and 550 nm waves hits your retina, it excites the red and green receptors, and your brain interprets that mix as yellow.

You can't tell the difference!

Now we know why the sensors on the camera are RGB. The camera is recording the scene as closely as necessary to fool your eye.

A TV or a color computer monitor only has three colors of pixels: red, green, and blue. By controlling the mix of them, it creates the sensation of thousands of colors to your eye.

56.3 Pigments

A color printer works oppositely: Instead of radiating colors, it puts pigments on the paper that absorb certain frequencies. A pigment that absorbs only frequencies near 650 nm (red) will appear to your eye as cyan. This makes sense because the sensation of cyan is created when your blue and green receptors are activated.

Thus, pigment colors come in:

- Cyan: absorbs frequencies around red
- Magenta: absorbs frequencies around green
- Yellow: absorbs frequencies around blue

If you buy ink for a color printer, you know there is typically a fourth ink: black. If you put cyan, magenta, and yellow pigments on paper, the mix won't absorb all the visible spectrum in a consistent manner, and our eyes are pretty sensitive to that, so we would see brown. So we add black ink to get pretty grays and blacks.

We call this approach to color CMYK (as opposed to RGB). If an artist is creating an image to be viewed on a screen, they will typically make an RGB image. If they are creating an image to be printed using pigments, they typically create a CMYK image. (Most of us don't care so much – we just let the computer do conversions between the two color spaces for us.)



CHAPTER 57

Reflection

What happens when light hits a mass?

In a previous chapter, we talked about light as a wave, and we mentioned that each color in the rainbow is a different wavelength. You can also think of light as particles of energy called *photons*. And every photon comes with an amount of energy that determines what color it is.

When we are talking about light interacting with objects, your intuition will be right more often if you think of light as a beam of photons.

When a photon comes from the sun and hits an object, one of several things can happen:

- The energy of the photon is absorbed by the object. It makes the object a little warmer. If a large proportion of photons hitting the mass are absorbed like this, we say the object is “black”.
- The photon bounces off the object. If the surface is very smooth, the photons bounce in a predictable manner, and we call this *reflection* and we say the object is “shiny”.
- If the surface is rough and the photons are not absorbed, the photons are scattered

in random directions. We call this *diffusion*. If most of the photons hitting an object are bounced in random directions, we say that the object is “white”.

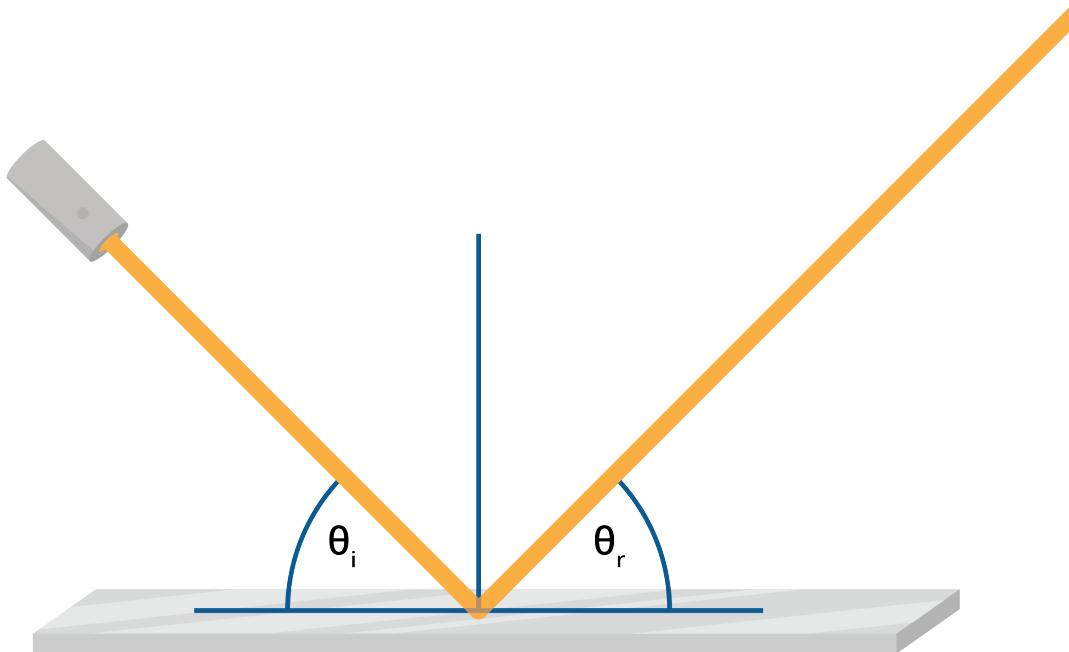
- The photon passes through the mass. If the mass has smooth surfaces and a consistent composition, the photons will pass through the mass in a predictable manner. We say that the mass is “transparent”.
- If the photons pass through, but in an unpredictable, scattering manner, we say the mass is “translucent”.

No object absorbs every photon, but chemists are always coming up with “blacker” materials. Vantablack, for example, is a super-black paint that absorbs 99.965% of all photons in the visible spectrum.

No object reflects every photon, but a mirror is pretty close. Let’s talk about reflections in a mirror.

57.1 Reflection

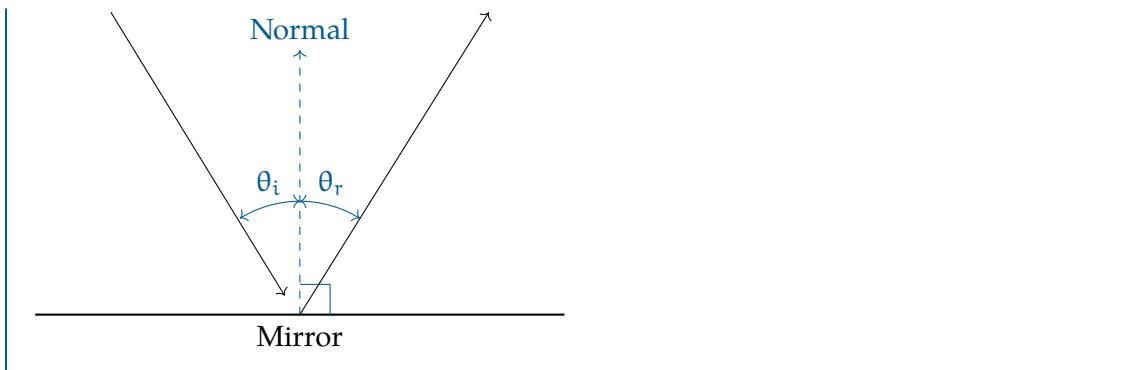
When a beam of light hits mirror, it bounces off the mirror at the same angle it approached from. That is, if it approaches nearly perpendicular to the mirror, it departs nearly perpendicular to the mirror. If it hits the mirror at a glancing angle, it departs at an angle close to the mirrors surface.

**Law of Reflection**

The angle of incidence, denoted as θ_i , is equal to the angle of reflection, denoted as θ_r . This law can be mathematically expressed as:

$$\theta_i = \theta_r$$

where θ_i is the angle between the incident light ray and the normal to the surface, and θ_r is the angle between the reflected light ray and the normal.



Exercise 67 Law of Reflection

Working Space

You are standing 4 meters from a mirror hung on a wall. The bottom of the mirror is the same height as your chin, so you can't see your whole body. You stick a piece of masking tape to your body. You walk forward until you are only 3 meters from the mirror, and put a piece of masking tape on your body at the new cut-off point. Is the new masking tape higher or lower on your body?

Answer on Page 831

Exercise 68 Photons and Color

Working Space

There are red photons.
Are there black photons?
Are there white photons?
Are there yellow photons?

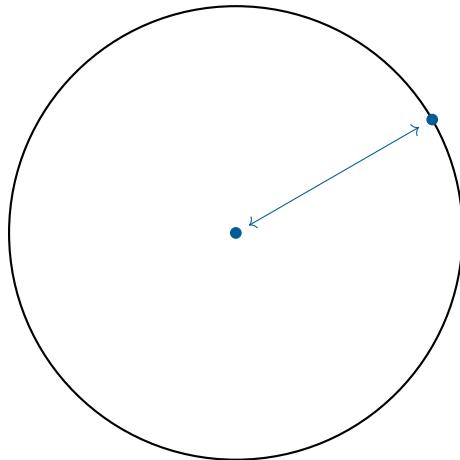
Answer on Page 831

57.2 Curved Mirrors

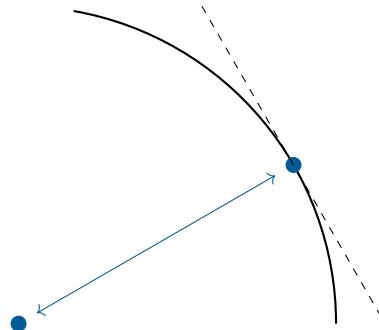
Flat mirrors are common and useful, but things get more interesting once you bend the mirrors. In this section, we are going to talk about a few different kinds of curved mirrors.

57.2.1 The Reflective Properties of Circles and Spheres

For example, if you were inside a circular room (a cylinder, actually), you could imagine standing in the center and pointing a flash light in any horizontal direction. The beam of light would bounce right back to you.



How do you know this? Because the tangent line is always perpendicular to the radius to the point of tangency:



You could create a spherical room with mirror walls. You'd create a platform in the center where you could stand, and if you pointed your flashlight in any direction, its beam of light would shine back at you.

57.2.2 Ellipses and Ellipsoids

Intuitively, you know what an ellipse is: it is an oval. But the ellipse is actually an oval with some special properties. This is a good time to talk about those properties.

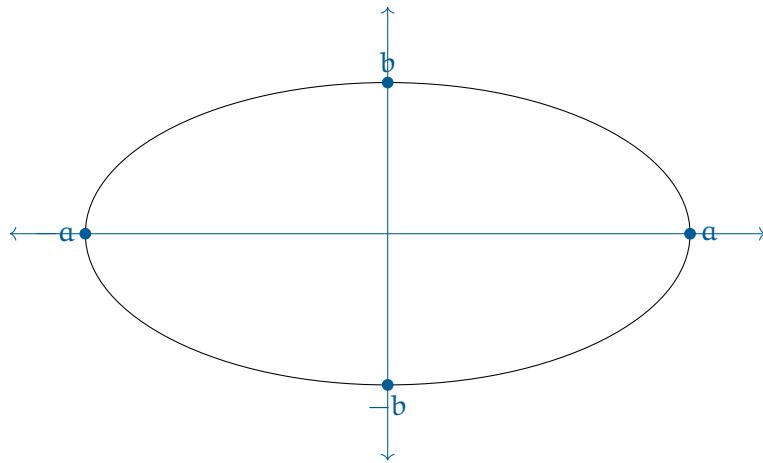
Mathematicians talk about a *standard* ellipse. A standard ellipse is centered on the origin $(0, 0)$ and its long axis is parallel with the x -axis or the y -axis.

Equation for a Standard Ellipse

To be precise, a standard ellipse is the set of points (x, y) that are solutions to the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Note that $(a, 0), (-a, 0), (0, b), (0, -b)$ are all part of the set. The complete set looks like this:

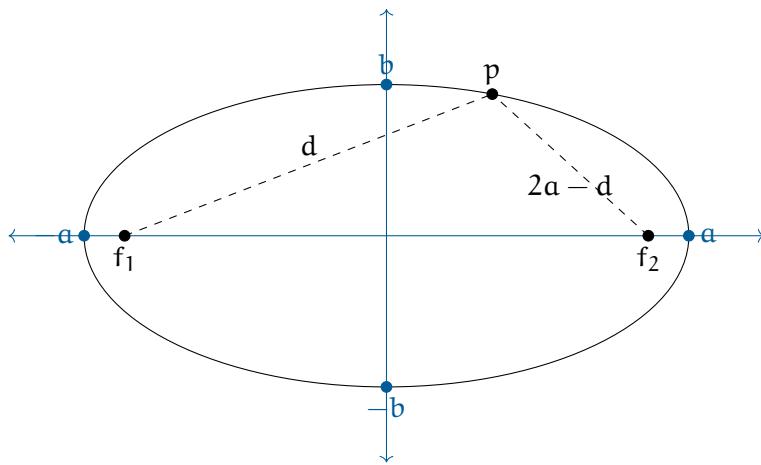


The area contained inside this ellipse is given by

$$A = \pi ab$$

We can now talk about two special points: the *foci*. Each focal point is on the long axis of the ellipse. Let's assume for a second that $a > b$. (Everything works the same if $b > a$, but it gets confusing if we try to deal with both cases simultaneously.)

If p is a point on the ellipse, the distance from p to focal point 1 plus the distance from p to focal point 2 is always $2a$.



How do we find the foci? We know they are on the long axis and that they are symmetrical across the short axis. All we need to know is how far are they from the short axis.

Distance from Center to the Foci

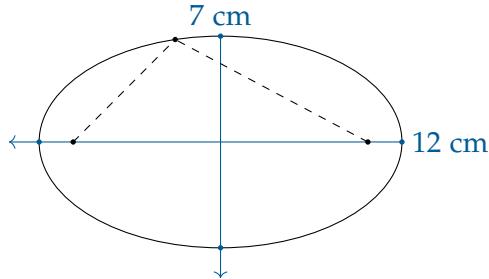
If you have an ellipse with a long axis that extends a from the center and a short axis that extends b from the center. The foci lie on the long axis and are c from the center. Where

$$c = \sqrt{a^2 - b^2}$$

Exercise 69 Foci of an ellipse*Working Space*

You need to draw an ellipse that is 12 cm long and 7 cm wide. You have a string, two pushpins, a ruler, and a pencil. Using the ruler, you draw two perpendicular axes.

You will stick one pin at each focal point. Each end of the string will be tied to a push pin. Using the pencil to keep the string taut, you will draw an ellipse.

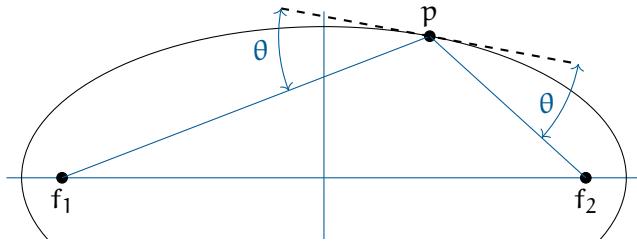


How far from the short axis are the pushpins placed?

How long is the string between them?

*Answer on Page 831***The Reflective Property of Ellipses**

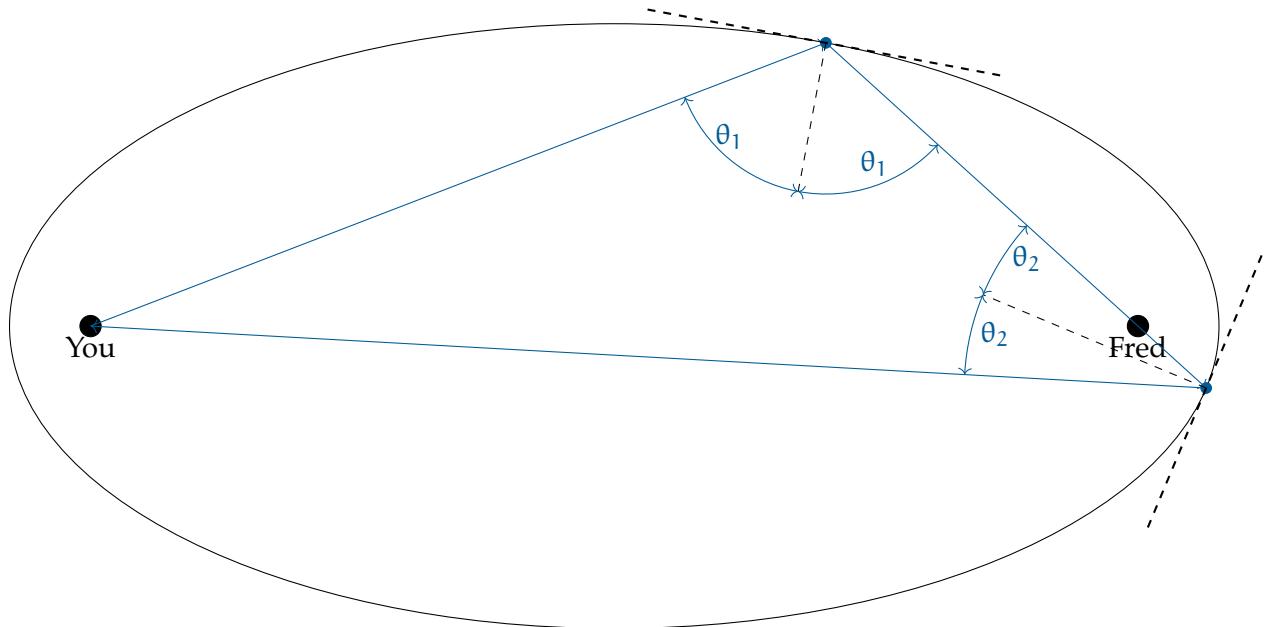
Here is something else that is wonderful about an ellipse: Pick any point p on the ellipse. Draw a line from p to each focal point. Draw the line tangent to the ellipse at p . Fact: The angle between the tangent and the line to focal point 1 is equal to the angle between the tangent and the line to focal point 2.



This is known as “The Reflective Property of Ellipses.”

Imagine you and your friend Fred are at an ellipse-shaped skating rink and edge of the rink are mirrored. You sit at one focal point and your friend sits at the other, if you point a flashlight at the mirror (in any direction!) the beam will bounce off the wall and head directly for Fred.

If Fred ducks out of the way, the beam will bounce again and head back to you.



This will work for sound too. If you whisper on the focal point, Fred (at the other focal point) will hear you surprisingly well because all the soundwaves that hit the wall will bounce (just like the light) straight at Fred.

57.2.3 Elliptical Orbits

One more fun fact about ellipses: We often imagine the planets traveling in circular orbits with the sun at the center – they actually travel in elliptical orbits, with the sun as one of the focal points.

The earth is closest to the sun around January 3rd: 147 million km.

The earth is farthest from the sun around July 3rd: 152 million km.

(Note that these dates are not the same as the solstices: The southern hemisphere is tilted the most toward the sun around December 21 and tilted most away around June 21.)

57.2.4 Ellipsoids

Just as we can pull the ideas of a circle into three dimensions to make a sphere, we can extend the ideas of the ellipse into three dimensions to talk about ellipsoids. Ellipsoids are like blimps.

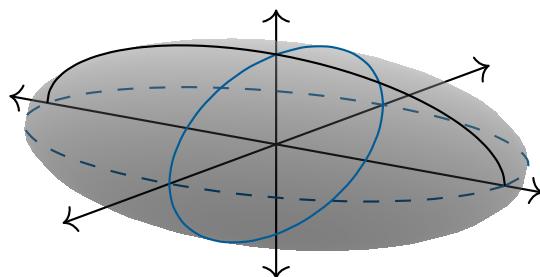
The standard ellipsoids are centered at the origin and aligned with the three axes.

Equation for a Standard Ellipsoid

To be precise, a standard ellipse is the set of points (x, y, z) that satisfy the equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

Note that $(a, 0, 0), (0, b, 0), (0, 0, c)$ are all part of the set. The complete set looks like this:



The volume bounded by this ellipsoid is

$$V = \frac{4}{3}\pi abc$$

Of course, a , b , and c can be any positive number, but in the real world we find ourselves working a lot with ellipsoids where two of the numbers are the same.

Oblate Spheroid

If two axes have the same length and one is shorter, you get something that looks like a sphere compressed in one direction – like a pumpkin. These are called *oblate spheroids*.

The earth is actually an oblate spheroid: the axis that goes through the north and south pole is shorter than the axes that pass through the equator. How much shorter? Just a little: The equator is 6,378 km from the center of the earth. The north pole is 21 km closer.

Prolate Spheroid

If two axes have the same length and one is longer, you get something that looks like a sphere stretched in one direction – like a rugby ball. It is called a *prolate spheroid*.

Like an ellipse, prolate spheroids have two focal points.

Focal Points of a Prolate Spheroid

If the long axis has a radial length of a and the two shorter axes have radial length b , then the focal points are on the long axis. The distance from the center to the focal point is given by

$$c = \sqrt{a^2 - b^2}$$

For any point p on the prolate spheroid, the sum of the distances from p to the focus points will always be $2a$.

It has the reflective property: A photon shot in any direction from one focal point will bounce off the wall and head directly at the other.

Exercise 70 Volume of Ruby Ball*Working Space*

Some jesters once thought it would be fun to make something that looked like a rugby ball, but made out of lead.

A rugby ball is about 30 cm long and has a circumference of 60 cm at its midpoint. A cubic centimeter of lead has a mass of 11.34 grams.

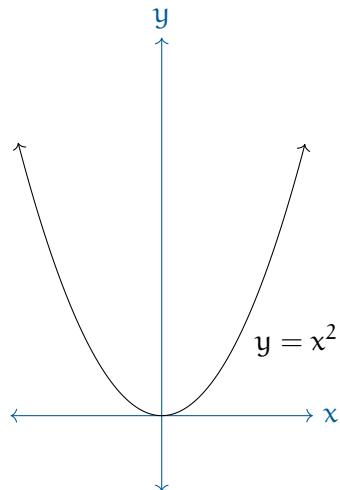
How much would a solid (not hollow) lead ruby ball weigh?

*Answer on Page 832***57.2.5 Parabolas and Parabolic Reflectors**

You are familiar with quadratic functions:

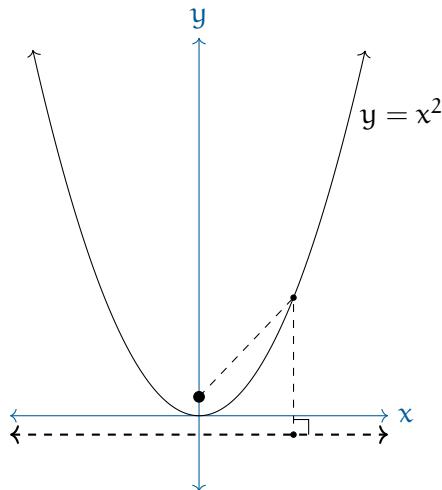
$$y = ax^2 + bx + c$$

If a is not zero, the graph of a quadratic is a curved line called a *parabola*. The first parabola that most mathematicians think of is the graph of $y = x^2$:



Every parabola has a *focus* and a *directrix*. The focus is a point on the parabola's axis of symmetry. The directrix is a line perpendicular to the axis of symmetry. Every point on the parabola is equal distance from the focus and the directrix.

For the graph of $y = x^2$, the focus is the point $(0, \frac{1}{4})$. The directrix is the line $y = -\frac{1}{4}$:



For example, the point $(1, 1)$ is on this parabola. It is $\frac{5}{4}$ from the directorix. How far is it from the focus? 1 horizontally and $\frac{3}{4}$ vertically.

$$\sqrt{1^2 + \left(\frac{3}{4}\right)^2} = \frac{5}{4}$$

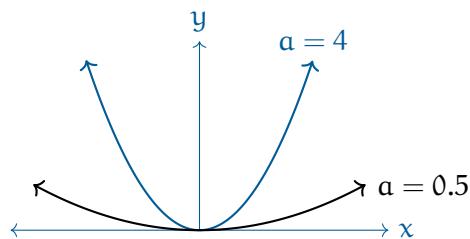
Thus, we have confirmed that $(1, 1)$ is equal distances from the focus and the directrix.

When we think about a parabola and its properties, we usually rotate and translate it to be symmetric around the y-axis, flip it so that it is low in the middle and rising on both sides, and push it up or down until the low point is on the x-axis.

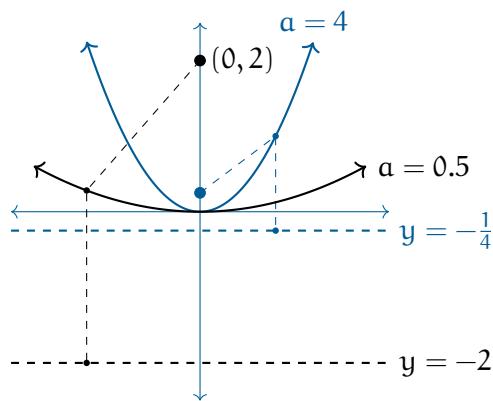
Then, they can all be written:

$$y = \frac{a}{4}x^2$$

where $a > 0$. If a is small, the parabola opens wider.

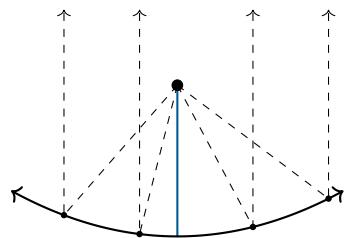


Then the focus is at $(0, \frac{1}{a})$ and the directorix is the line $y = -\frac{1}{a}$.



Reflective Property of a Parabola

Assume you have a parabola-shaped mirror, a beam of light shot from the focus, will bounce off the mirror in the direction of axis of symmetry:



This is why your flashlight has a parabolic mirror: the lightbulb is at the focus so any photons that hit the mirror are redirected straight forward.

(Note that in the real world, we use parabolic dishes: a rotated around its axis of symmetry.)

The reflection works exactly the same in reverse: There are solar cookers that are big parabolic mirrors. They let you put a pot on the focus point. You move the dish until its axis of symmetry is pointed at the sun.

You will also see a lot of antennas have parabolic dishes. Note that photons that come in parallel to the axis of symmetry are redirected to a single point – where the receiver is.



Sometimes in a science museum, you will see two parabolic dishes far apart and pointed at each other. One person speaks with their mouth at the focus of one. The other person listens with their ear at the focus of the other. Even though you are very far apart, it sounds like they are really, really close.



This is because the speaker's parabolic wall focuses the sound energy in a nice beam the size of the wall pointed straight at the listener's parabolic wall. The listener's wall focuses the energy of that beam at the listener's ear.



CHAPTER 58

Refraction

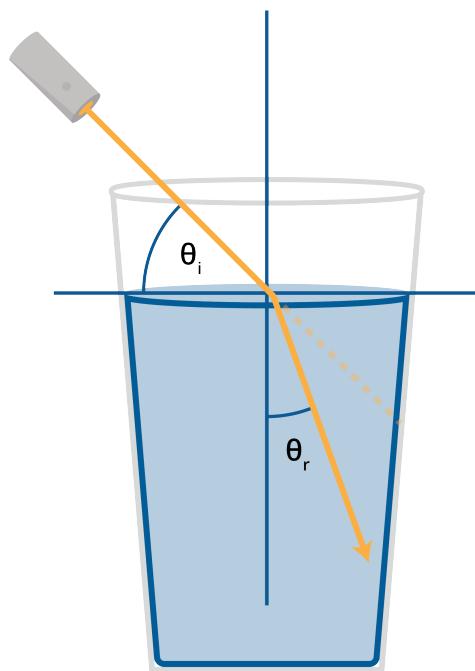
Refraction of light is the phenomenon where light changes its direction when it passes from one medium to another. The change in direction is due to a change in the speed of light as it moves from one medium to another.

This phenomenon is explained by Snell's law, which states:

$$n_1 \cdot \sin(\theta_1) = n_2 \cdot \sin(\theta_2) \quad (58.1)$$

where:

- n_1 and n_2 are the indices of refraction for the first and second media, respectively. The index of refraction is the ratio of the speed of light in a vacuum to the speed of light in the medium. It is a dimensionless quantity.
- θ_1 and θ_2 are the angles of incidence and refraction, respectively. These angles are measured from the normal (perpendicular line) to the surface at the point where light hits the boundary.



The angle of incidence (θ_1) is the angle between the incident ray and the normal to the interface at the point of incidence. Similarly, the angle of refraction (θ_2) is the angle between the refracted ray and the normal.

When light travels from a medium with a lower refractive index to a medium with a higher refractive index, it bends towards the normal. Conversely, when light travels from a medium with a higher refractive index to one with a lower refractive index, it bends away from the normal.

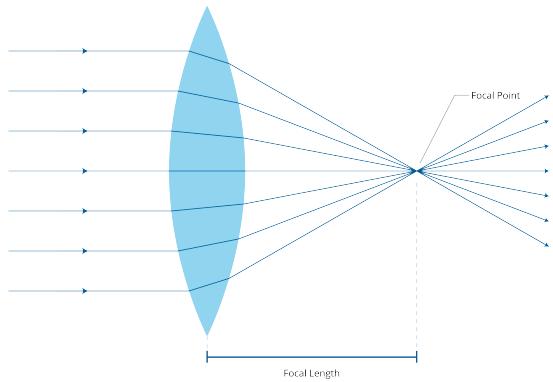


CHAPTER 59

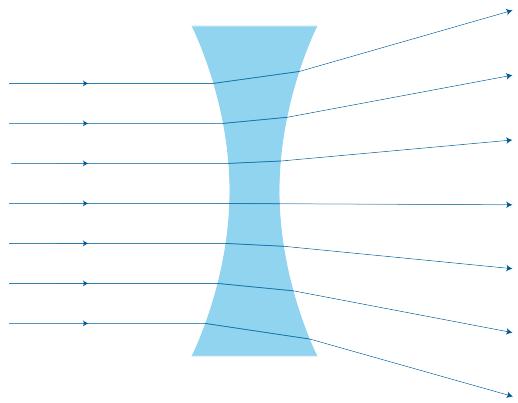
Lenses

Lenses are optical devices with perfect or approximate axial symmetry that transmit and refract light, converging or diverging the beam. There are two main types of lenses, distinguished by their shape and the way they refract light:

- **Converging (or Convex) Lenses:** These are thicker at the center than at the edges. When parallel light rays enter a convex lens, they converge to a point called the focal point. Examples of converging lenses include magnifying glasses and camera lenses.



- **Diverging (or Concave) Lenses:** These are thinner at the center than at the edges. When parallel light rays enter a concave lens, they diverge or spread out. These lenses are often used in glasses to correct farsightedness.



59.1 Focal Length

The focal length of a lens is the distance between the center of the lens and the focal point. It is determined by the lens shape and the refractive index of the lens material. For a converging lens, the focal length is positive, and for a diverging lens, the focal length is negative.

59.2 Refractive Index

The refractive index of a material is a measure of how much the speed of light is reduced inside the material. The refractive index n of a material is given by the ratio of the speed of light in a vacuum c to the speed of light v in the material:

$$n = \frac{c}{v}$$

The refractive index affects how much a light ray changes direction, or refracts, when entering the material at an angle. A higher refractive index indicates that light travels slower in that medium and the light ray will bend more towards the normal.

Lenses work by refracting light at their two surfaces. By choosing the right lens shape and material, lenses can be designed to bring light to a focus, spread it out, or perform more complex transformations.



CHAPTER 60

Images in Python

An image is usually represented as a three-dimensional array of 8-bit integers. NumPy arrays are the most commonly used library for this sort of data structure.

If you have an RGB image that is 480 pixels tall and 640 pixels wide, you will need a $480 \times 640 \times 3$ NumPy array.

There is a separate library (imageio) that:

- Reads an image file (like JPEG files) and creates a NumPy array.
- Writes a NumPy array to a file in standard image formats

Let's create a simply python program that creates a file containing an all-black image that is 640 pixels wide and 480 pixels tall. Create a file called `create_image.py`:

```
import NumPy as np
import imageio
import sys
```

```
# Check command-line arguments
if len(sys.argv) < 2:
    print(f"Usage {sys.argv[0]} <outfile>")
    sys.exit(1)

# Constants
IMAGE_WIDTH = 640
IMAGE_HEIGHT = 480

# Create an array of zeros
image = np.zeros((IMAGE_HEIGHT, IMAGE_WIDTH, 3), dtype=np.uint8)

# Write the array to the file
imageio.imwrite(sys.argv[1], image)
```

To run this, you will need to supply the name of the file you are trying to create. The extension (like .png or .jpeg) will tell imageio what format you want written. Run it now:

```
python3 create_image.py blackness.png
```

Open the image to confirm that it is 640 pixels wide, 480 pixels tall, and completely black.

60.1 Adding color

Now, let's walk through through the image, pixel-by-pixel, adding some red. We will gradually increase the red from 0 on the left to 255 on the right.

```
import NumPy as np
import imageio
import sys

# Check command-line arguments
if len(sys.argv) < 2:
    print(f"Usage sys.argv[0] <outfile>")
    sys.exit(1)

# Constants
IMAGE_WIDTH = 640
IMAGE_HEIGHT = 480

# Create an array of zeros
```

```
image = np.zeros((IMAGE_HEIGHT, IMAGE_WIDTH, 3), dtype=np.uint8)

for col in range(IMAGE_WIDTH):

    # Red goes from 0 to 255 (left to right)
    r = int(col * 255.0 / IMAGE_WIDTH)

    # Update all the pixels in that column
    for row in range(IMAGE_HEIGHT):
        # Set the red pixel
        image[row, col, 0] = r

# Write the array to the file
imageio.imwrite(sys.argv[1], image)
```

When you run the function to create a new image, it will be a fade from black to red as you move from left to right:



Now, inside the inner loop, update the blue channel so that it goes from zero at the top to 255 at the bottom:

```
# Update all the pixels in that column
for row in range(IMAGE_HEIGHT):

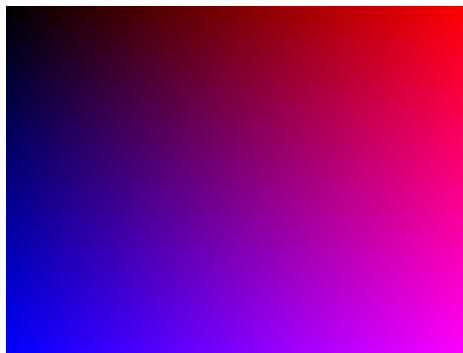
    # Update the red channel
    image[row,col,0] = r

    # Blue goes from 0 to 255 (top to bottom)
    b = int(row * 255.0 / IMAGE_HEIGHT)
    image[row,col,2] = b

imageio.imwrite(sys.argv[1], image)
```

When you run the program again, you will see the color fades from black to blue as you

go down the left side. As you go down the right side, the color fades from red to magenta.



Notice that red and blue with no green looks magenta to your eye.

Now let's add some stripes of green:

```
import NumPy as np
import imageio
import sys

# Check command line arguments
if len(sys.argv) < 2:
    print(f"Usage sys.argv[0] <outfile>")
    sys.exit(1)

# Constants
IMAGE_WIDTH = 640
IMAGE_HEIGHT = 480
STRIPE_WIDTH = 40
pattern_width = STRIPE_WIDTH * 2

# Create an image of all zeros
image = np.zeros((IMAGE_HEIGHT, IMAGE_WIDTH, 3), dtype=np.uint8)

# Step from left to right
for col in range(IMAGE_WIDTH):

    # Red goes from 0 to 255 (left to right)
    r = int(col * 255.0 / IMAGE_WIDTH)

    # Should I add green to this column?
    should_green = col % pattern_width > STRIPE_WIDTH

    # Update all the pixels in that column
    for row in range(IMAGE_HEIGHT):
```

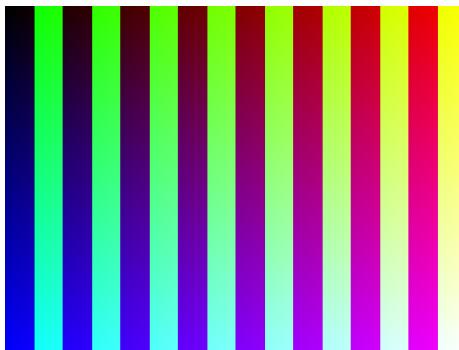
```
# Update the red channel
image[row,col,0] = r

# Should I add green to this pixel?
if should_green:
    image[row,col,1] = 255

# Blue goes from 0 to 255 (top to bottom)
b = int(row * 255.0 / IMAGE_HEIGHT)
image[row,col,2] = b

imageio.imwrite(sys.argv[1], image)
```

When you run this version, you will see the previous image in half the stripes. In the other half, you will see that green fades to cyan down the left side and yellow fades to white down the right side.



60.2 Using an existing image

imageio can also be used to read in any common image file format. Let's read in an image and save each of the red, green, and blue channels out as its own image.

Create a new file called `separate_image.py`:

```
import imageio
import sys
import os

# Check command line arguments
if len(sys.argv) < 2:
    print(f"Usage {sys.argv[0]} <infile>")
```

```
sys.exit(1)

# Read the image
path = sys.argv[1]
image = imageio.imread(path)

# What is the filename?
filename = os.path.basename(path)

# What is the shape of the array?
original_shape = image.shape

# Log it
print(f"Shape of {filename}:{original_shape}")

# Names of the colors for the filenames
colors = ['red','green','blue']

# Step through each of the colors
for i in range(3):

    # Create a new image
    newimage = np.zeros(original_shape, dtype=np.uint8)

    # Copy one channel
    newimage[:, :, i] = image[:, :, i]

    # Save to a file
    new_filename = f"{colors[i]}_{filename}"
    print(f"Writing {new_filename}")
    imageio.imwrite(new_filename, newimage)
```

Now you can run the program with any common RGB image type:

```
python3 separate_image.py dog.jpg
```

This will create three images: `red_dog.jpg`, `green_dog.jpg`, and `blue_dog.jpg`.



CHAPTER 61

Introduction to Polynomials

Watch Khan Academy's **Polynomials intro** video at <https://youtu.be/Vm7H0VT1Ico>

A *monomial* is the product of a number and a variable raised to a non-negative (but possibly zero) integer power. Here are some monomials:

$$3x^2$$

$$\pi x^2$$

$$7x$$

$$-\frac{2}{3}x^{12}$$

$$-2x^{15}$$

$$(3.33)x^{100}$$

$$3$$

$$0$$

The exponent is called the *degree* of the monomial. Examples: $3x^{17}$ has degree 17, $-7x$ has degree 1, and 3.2 has degree 0 (because you can think of it as $(3.2)x^0$).

The number in the product is called the *coefficient*. Example: $3x^{17}$ has a coefficient of 3, $-2x$ has a coefficient of -2, and $(3.4)x^{1000}$ has a coefficient of 3.4.

A *polynomial* is the sum of one or more monomials. Here are some polynomials:

$4x^2 + 9x + 3.9$

$\pi x^2 + \pi x + \pi$

$7x + 2$

$-2x^{10} + (3.4)x - 45x^{900} - 1$

3.3

$3x^{20}$

We say that each monomial is a *term* of the polynomial.

$x^{-5} + 12$ is *not* a polynomial because the first term has a negative exponent.

$x^2 - 32x^{\frac{1}{2}} + x$ is *not* a polynomial because the second term has a non-integer exponent.

$\frac{x+2}{x^2+x+5}$ is *not* a polynomial because it is not just a sum of monomials.

Exercise 71 Identifying Polynomials

Circle only the polynomials.

Working Space

$$-2x^3 + \frac{1}{2}x + 3.9(4.5)x^2 + \pi x$$

7

$$2x^{-10} + 4x - 1 \quad x^{\frac{2}{3}} \quad 3x^{20} + 2x^{19} - 5x^{18}$$

Answer on Page 832

We typically write a polynomial starting at the term with the highest degree and proceed in decreasing order to the term with the lowest degree:

$$2x^9 - 3x^7 + \frac{3}{4}x^3 + x^2 + \pi x - 9.3$$

This is known as *the standard form*. The first term of the standard form is called *the leading term*, and we often call the coefficient of the leading term *the leading coefficient*. We sometimes speak of the degree of the polynomial, which is just the degree of the leading term.

Exercise 72 Standard of a Polynomial

Write $21x^2 - x^3 + \pi - 1000x$ in standard form. What is the degree of this polynomial? What is its leading coefficient?

Working Space

Answer on Page 832

Exercise 73 Evaluate a Polynomial

Let $y = x^3 - 3x^2 + 10x - 12$. What is y when x is 4?

Working Space

Answer on Page 833

I would be remiss in my duties if I didn't mention one more thing about polynomials: mathematicians have defined a polynomial to be a sum of a *finite* number of monomials.

It is certainly possible to have a sum of an infinite number of monomials like this:

$$1 + \frac{1}{2}x + \frac{1}{4}x^2 + \frac{1}{8}x^3 + \frac{1}{16}x^4 + \dots$$

This is an example of an *infinite series*; we don't consider them polynomials. Infinite series are interesting and useful, but I will not discuss them much until later in the course.



CHAPTER 62

Python Lists

Watch CS Dojo's **Introduction to Lists in Python** video at <https://www.youtube.com/watch?v=tw7ror9x32s>

To review, Python list is an indexed collection. The indices start at zero. You can create a list using square brackets.

Now you are going to write a program that makes an array of strings. Type this code into a file called `faves.py`:

```
favorites = ["Raindrops", "Whiskers", "Kettles", "Mittens"]
favorites.append("Packages")
print("Here are all my favorites:", favorites)
print("My most favorite thing is", favorites[0])
print("My second most favorite is", favorites[1])
number_of_faves = len(favorites)
print("Number of things I like:", number_of_faves)

for i in range(number_of_faves):
```

```
print(i, ": I like", favorites[i])
```

Run it:

```
$ python3 faves.py
Here are all my favorites: ['Raindrops', 'Whiskers', 'Kettles', 'Mittens', 'Packages']
My most favorite thing is Raindrops
My second most favorite is Whiskers
Number of things I like: 5
0 : I like Raindrops
1 : I like Whiskers
2 : I like Kettles
3 : I like Mittens
4 : I like Packages
```

After you have run the code, study it until the output makes sense.

Exercise 74 Assign into list

Before you list the items, replace "Mittens" with "Gloves".

Working Space

Answer on Page 833

62.1 Evaluating Polynomials in Python

First, before you go any further, you need to know that raising a number to a power is done with `**` in Python. So for example, to get 5^2 , you would write `5**2`.

Back to polynomials: if you had a polynomial like $2x^3 - 9x + 12$, you could write it like this: $12x^0 + (-9)x^1 + 0x^2 + 2x^3$. We could use this representation to keep a polynomial in a Python list. We would simply store all the coefficients in order:

```
pn1 = [12, -9, 0, 2]
```

In the list, the index of each coefficient would correspond to the degree of that monomial. For example, in the list 2 is at index 3, so that entry represents $2x^3$.

In the last chapter, you evaluated the polynomial $x^3 - 3x^2 + 10x - 12$ at $x = 4$. Now you will write code that does that evalution. Create a file called `polynomials.py` and type in the following:

```
def evaluate_polynomial(pn, x):
    sum = 0.0
    for degree in range(len(pn)):
        coefficient = pn[degree]
        term_value = coefficient * x ** degree
        sum = sum + term_value
    return sum

pn1 = [-12.0, 10.0, -3.0, 1.0]
y = evaluate_polynomial(pn1, 4.0)
print("Polynomial 1: When x is 4.0, y is", y)
```

Run it. It should evaluate to 44.0.

62.2 Walking the list backwards

Now you are going to make a function that makes a pretty string to represent your polynomial. Here is how it will be used:

```
def polynomial_to_string(pn):
    ...Your Code Here...

pn_test = [-12.0, 10.0, 0.0, 1.0]
print(polynomial_to_string(pn1))
```

This would output:

`1.0x**3 + 10.0x + -12.0`

This is not as simple as you might hope. In particular:

- You should skip the terms with a coefficient of zero
- The term of degree 1 has an x , but no exponent
- The term of degree 0 has neither an x nor an exponent

- Standard form demands that you list the terms in the reverse order from that of your coefficients list. You will need to walk the list from last to first.

Add this function to your `polynomials.py` file after your `evaluate_polynomial` function:

```
def polynomial_to_string(pn):  
  
    # Make a list of the monomial strings  
    monomial_strings = []  
  
    # Start at the term with the largest degree  
    degree = len(pn) - 1  
  
    # Go through the list backwards stop after constant term  
    while degree >= 0:  
        coefficient = pn[degree]  
  
        # Skip any term with a zero coefficient  
        if coefficient != 0.0:  
  
            # Describe the monomial  
            if degree == 0:  
                monomial_string = "{}".format(coefficient)  
            elif degree == 1:  
                monomial_string = "{}x".format(coefficient)  
            else:  
                monomial_string = "{}x^{}".format(coefficient, degree)  
  
            # Add it to the list  
            monomial_strings.append(monomial_string)  
  
        # Move to the previous term  
        degree = degree - 1  
  
    # Deal with the zero polynomial  
    if len(monomial_strings) == 0:  
        monomial_strings.append("0.0")  
  
    # Make a string that joins the terms with a plus sign  
    return " + ".join(monomial_strings)
```

Note that in a list n items, the indices go from 0 to $n - 1$. So when we are walking the list backwards, we start at `len(pn) - 1` and stop at zero.

Look over the code and google the functions you aren't familiar with. For example, if you

want to know about the (join) function, google for “python join”.

Now change your code to use the new function:

```
pn1 = [-12.0, 10.0, -3.0, 1.0]
y = evaluate_polynomial(pn1, 4.0)
print("y =", polynomial_to_string(pn1))
print("    When x is 4.0, y is", y)
```

Run the program. Does the function work?

Exercise 75 Evaluate Polynomials

Using the function that you just wrote, add a few lines of code to `polynomials.py` to evaluate the following polynomials:

Working Space

- Find $4x^4 - 7x^3 - 2x^2 + 5x + 2.5$ at $x = 8.5$. It should be 16481.875
- Find $5x^5 - 9$ at $x = 2.0$. It should be 151.0

Answer on Page 833

62.3 Plot the polynomial

We can evaluate a polynomial at many points and plot them on a graph. You are going to write the code to do this. Create a new file called `plot_polynomial.py`. Copy your `evaluate_polynomial` function into the new file.

Add a line at the beginning of the program that imports the plotting library `matplotlib`:

```
import matplotlib.pyplot as plt
```

After the `evaluate_polynomial` function:

- Create a list with polynomial coefficients.

- Create two empty arrays, one for x values and one for y values.
- Fill the x array with values from -3.5 to 3.5. Evaluate the polynomial at each of these points; put those values in the y array.
- Plot them

Like this:

```
# x**3 - 7x + 6
pn = [6.0, -7.0, 0.0, 1.0]

# These lists will hold our x and y values
x_list = []
y_list = []

# Start at x=-3.5
current_x = -3.5

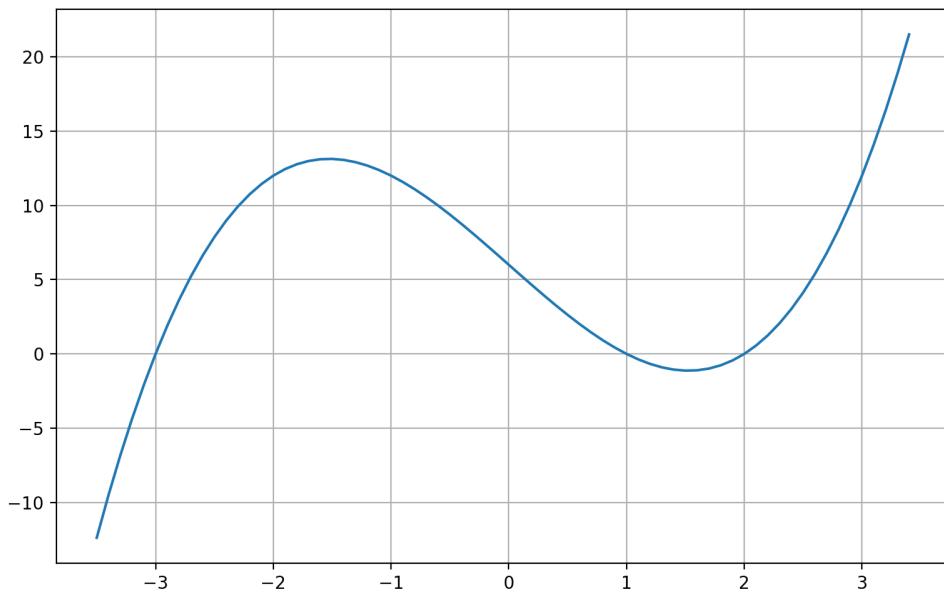
# End at x=3.5
while current_x <= 3.5:
    current_y = evaluate_polynomial(pn, current_x)

    # Add x and y to respective lists
    x_list.append(current_x)
    y_list.append(current_y)

    # Move x forward
    current_x += 0.1

# Plot the curve
plt.plot(x_list, y_list)
plt.grid(True)
plt.show()
```

You should get a beautiful plot like this:



If you received an error that the matplotlib was not found, use pip to install it:

```
$ pip3 install matplotlib
```

Exercise 76 Observations

Where does your polynomial cross the y-axis? Looking at the polynomial $x^3 - 7x + 6$, could you have guessed that value?

Working Space

Where does your polynomial cross the x-axis? The places where a polynomial crosses the x-axis is called *its roots*. Later in the course, you will learn techniques for finding the roots of a polynomial.

Answer on Page 833



CHAPTER 63

Adding and Subtracting Polynomials

Watch Khan Academy's **Adding polynomials** video at <https://youtu.be/ahdKdxsTj8E>

When adding two monomials of the same degree, you sum their coefficients:

$$7x^3 + 4x^3 = 11x^3$$

Using this idea, when adding two polynomials, you convert it into one long polynomial and then simplify by combining terms with the same degree. For example:

$$\begin{aligned}(10x^3 - 2x + 13) + (-5x^2 + 7x - 12) \\&= 10x^3 + (-2)x + 13 + (-5)x^2 + 7x + (-12) \\&= 10x^3 + (-5)x^2 + (-2 + 7)x + (13 - 12) \\&= 10x^3 - 5x^2 + 5x + 1\end{aligned}$$

Exercise 77 Adding Polynomials Practice

Add the following polynomials:

Working Space

1. $2x^3 - 5x^2 + 3x - 9$ and $x^3 - 2x^2 - 2x - 9$

2. $3x^5 - 5x^3 + 3x^2 - x - 3$ and $2x^4 - 2x^3 - 2x^2 + x - 9$

Answer on Page 833

Notice that in the second question, the degree 1 term disappears completely: $(-x) + x = 0$

One more tricky thing that can happen: Sometimes the coefficients don't add nicely. For example:

$$\pi x^2 - 3x^2 = (\pi - 3)x^2$$

That is as far as you can simplify it.

63.1 Subtraction

Now watch Khan Academy's **Subtracting polynomials** at <https://youtu.be/5ZdxnFspyp8>.

When subtracting one polynomial from the other, it is a lot like adding two polynomials. The difference: when make the two polynomials into one long polynomial, we multiply each monomial that is being subtracted by -1. For example:

$$\begin{aligned}
 (2x^2 - 3x + 9) - (5x^2 - 7x + 4) \\
 &= 2x^2 + (-3)x + 9 + (-5)x^2 + 7x + (-4) \\
 &= (2 - 5)x^2 + (-3 + 7)x + (9 - 4) \\
 &= -3x^2 + 4x + 5
 \end{aligned}$$

Exercise 78 Subtracting Polynomials Practice

Add the following polynomials:

Working Space

$$1. (2x^3 - 5x^2 + 3x - 9) - (x^3 - 2x^2 - 2x - 9)$$

$$2. (3x^5 - 5x^3 + 3x^2 - x - 3) - (2x^4 - 2x^3 - 2x^2 + x - 9)$$

Answer on Page 833

63.2 Adding Polynomials in Python

As a reminder, in our Python code, we are representing a polynomial with a list of coefficients. The first coefficient is the constant term. The last coefficient is the leading coefficient. So, we can imagine $-5x^3 + 3x^2 - 4x + 9$ and $2x^3 + 4x^2 - 9$ would look like this: *FIXME: Diagram here*

To add the two polynomials then, we sum the coefficients for each degree. *FIXME: Diagram here*

Create a file called `add_polynomials.py`, and type in the following:

```
def add_polynomials(a, b):
    degree_of_result = len(a)
    result = []
    for i in range(degree_of_result):
        coefficient_a = a[i]
        coefficient_b = b[i]
        result.append(coefficient_a + coefficient_b)
    return result
```

```
polynomial1 = [9.0, -4.0, 3.0, -5.0]
polynomial2 = [-9.0, 0.0, 4.0, 2.0]
polynomial3 = add_polynomials(polynomial1, polynomial2)

print('Sum =', polynomial3)
```

Run the program.

Unfortunately, this code only works if the polynomials are the same length. For example, try making `polynomial1` have a larger degree than `polynomial2`:

```
# x**4 - 5x**3 + 3x**2 - 4x + 9
polynomial1 = [9.0, -4.0, 3.0, -5.0, 1.0]

# 2x**3 + 4x**2 - 9
polynomial2 = [-9.0, 0.0, 4.0, 2.0]
polynomial3 = add_polynomials(polynomial1, polynomial2)
print('Sum =', polynomial3)
```

See the problem?

Exercise 79 Dealing with polynomials of different degrees

Working Space

Can you fix the function `add_polynomials` to handle polynomials of different degrees?

Here is a hint: In Python, there is a `max` function that returns the largest of the numbers it is passed.

```
biggest = max(5,7)
```

Here `biggest` would be set to 7.

Here is another hint: If you have an array `mylist`, `i`, a non-negative integer, is only a legit index if `i < len(mylist)`.

Answer on Page 834

63.3 Scalar multiplication of polynomials

If you multiply a polynomial with a number, the distributive property applies:

$$(3.1)(2x^2 + 3x + 1) = (6.2)x^2 + (9.3)x + 3.1$$

(When we are talking about things that are more complicated than a number, we use the word *scalar* to mean “Just a number”. So this is the product of a scalar and a polynomial.)

In `add_polynomials.py`, add a function to that multiplies a scalar and a polynomial:

```
def scalar_polynomial_multiply(s, pn):
    result = []
    for coefficient in pn:
        result.append(s * coefficient)
    return result
```

Somewhere near the end of the program, test this function:

```
polynomial4 = scalar_polynomial_multiply(5.0, polynomial11)
print('Scalar product =', polynomial_to_string(polynomial4))
```

Exercise 80 Subtract polynomials in Python

Now implement a function that does subtraction using `scalar_polynomial_multiply` and `add_polynomials`.

It should look like this:

```
def subtract_polynomial(a, b):
    ...Your code here...
```

```
polynomial5 = [9.0, -4.0, 3.0, -5.0]
polynomial6 = [-9.0, 0.0, 4.0, 2.0, 1.0]
polynomial7 = subtract_polynomial(polynomial5, polynomial6)
print('Difference =', polynomial_to_string(polynomial7))
```

Working Space

Answer on Page 834



CHAPTER 64

Multiplying Polynomials

Watch Khan Academy's **Multiplying monomials** at <https://youtu.be/Vm7H0VTlIco>.

To review, when you multiply two monomials, you take the product of their coefficients and the sum of their degrees:

$$(2x^6)(5x^3) = (2)(5)(x^6)(x^3) = 10x^9$$

If you have a product of more than two monomials, multiply *all* the coefficients and sum *all* the exponents:

$$(3x^2)(2x^3)(4x) = (3)(2)(4)(x^2)(x^3)(x^1) = 24x^6$$

Exercise 81 **Multiplying monomials**

Multiply these monomials

Working Space

1. $(3x^2)(5x^3)$

2. $(2x)(4x^9)$

3. $(-5.5x^2)(2x^3)$

4. $(\pi)(-2x^5)$

5. $(2x)(3x^2)(5x^7)$

*Answer on Page 834***64.1 Multiplying a monomial and a polynomial**

Watch Khan Academy's **Multiplying monomials by polynomials** at <https://youtu.be/pD2-H15ucNE>.

When multiplying a monomial and a polynomial, you use the the distributive property.

Then it is just multiplying several pairs of monomials:

$$\begin{aligned}
 (3x^2)(4x^3 - 2x^2 + 3x - 7) \\
 &= (3x^2)(4x^3) + (3x^2)(-2x^2) + (3x^2)(3x) + (3x^2)(-7) \\
 &= 12x^5 - 6x^4 + 9x^3 - 21x^2
 \end{aligned}$$

Exercise 82 Multiplying a monomial and a polynomial

Multiply these monomials

Working Space

1. $(3x^2)(5x^3 - 2x + 3)$

2. $(2x)(4x^9 - 1)$

3. $(-5.5x^2)(2x^3 + 4x^2 + 6)$

4. $(\pi)(-2x^5 + 3x^4 + x)$

5. $(2x)(3x^2)(5x^7 + 2x)$

Answer on Page 835

64.2 Multiplying polynomials

Watch Khan Academy's **Multiplying binomials by polynomials** video at https://youtu.be/D6mivA_8L8U

When you are multiplying two polynomials, you will use the distributive property several times to make it one long polynomial. Then you will combine the terms with the same degree. For example,

$$\begin{aligned}
 (2x^2 - 3)(5x^2 + 2x - 7) &= (2x^2)(5x^2 + 2x - 7) + (-3)(5x^2 + 2x - 7) \\
 &= (2x^2)(5x^2) + (2x^2)(2x) + (2x^2)(-7) + (-3)(5x^2) + (-3)(2x) + (-3)(-7) \\
 &= 10^4 + 4x^3 + -14x^2 + -15x^2 + -6x + 21 = 10^4 + 4x^3 + -29x^2 + -6x + 21
 \end{aligned}$$

One common form that you will see is multiplying two binomials together:

$$(2x + 7)(5x + 3) = (2x)(5x + 3) + (7)(5x + 3) = (2x)(5x) + (7)(5x) + (2x)(3) + (7)(3)$$

Notice the product has become the sum of four parts: the firsts, the inners, theouters, and the lasts. People sometimes use the mnemonic FOIL to remember this pattern, but there is a general rule that works for all product of polynomials, not just binomials. Here it is: Every term in the first will be multiplied by every term in the second, and then just add them together.

So, for example, if you have a polynomial s with three terms and you multiply it by a polynomial t with five terms, you will get a sum of 15 terms – each term is a product of two monomials, one from s and one from t . (Of course, several of those terms might have the same degree, so they will be combined together when you simplify. Thus you typically end up with a polynomial with less than 15 terms.)

Using this rule, here is how I would multiply $2x^2 - 3x + 1$ and $5x^2 + 2x - 7$:

$$\begin{aligned}
 (2x^2 - 3x + 1)(5x^2 + 2x - 7) &= (2x^2)(5x^2) + (2x^2)(2x) + (2x^2)(-7) + \\
 &\quad (-3x)(5x^2) + (-3x)(2x) + (-3x)(-7) + \\
 &\quad (1)(5x^2) + (1)(2x) + (1)(-7) \\
 &= 10x^4 + 4x^3 + (-14)x^2 + (-15)x^3 + (-6)x^2 + 21x + 5x^2 + 2x + (-7) \\
 &= 10x^4 + (4 - 15)x^3 + (-14 - 6 + 5)x^2 + (21 + 2)x + (-7) \\
 &= 10x^4 - 11x^3 - 15x^2 + 23x - 7
 \end{aligned}$$

Note that the product (before combining terms with the same degree) has $3 \times 3 = 9$ terms – every possible combination of a term from the first polynomial and a term from the second polynomial.

One common source of error: losing track of the negative signs. You will need to be really careful. I have found that it helps to use + between all terms, and use negative coefficients to express subtraction. For example, if the problem says $4x^2 - 5x - 3$, you should work with that as $4x^2 + (-5)x + (-3)$

Exercise 83 Multiplying polynomials

Multiply the following pairs of polynomials:

Working Space

1. $2x + 1$ and $3x - 2$

2. $-3x^2 + 5$ and $4x - 2$

3. $-2x - 1$ and $-3x - \pi$

4. $-2x^5 + 5x$ and $3x^5 + 2x$

Answer on Page 835

Exercise 84 Observations

Let's say I have two polynomials, p_1 and p_2 . p_1 has degree 23. p_2 has degree 12. What is the degree of their product?

Working Space

Answer on Page 835



CHAPTER 65

Multiplying Polynomials in Python

At this point, you have created a nice toolbox of functions for dealing with lists of coefficients as polynomials. Create a file called `poly.py` and copy the following functions into it:

- `evaluate_polynomial`
- `polynomial_to_string`
- `add_polynomials`
- `scalar_polynomial_multiply`
- `subtract_polynomial`

Now create another file in the same directory called `test.py`. Type this into that file:

```
import poly

polynomial_a = [9.0, -4.0, 3.0, -5.0]
print('Polynomial A =', poly.polynomial_to_string(polynomial_a))

polynomial_b = [-9.0, 0.0, 4.0, 2.0, 1.0]
print('Polynomial B =', poly.polynomial_to_string(polynomial_b))

# Evaluation
value_of_b = poly.evaluate_polynomial(polynomial_b, 3)
print('Polynomial B at 3 =', value_of_b)

# Adding
a_plus_b = poly.add_polynomials(polynomial_a, polynomial_b)
print('A + B =', poly.polynomial_to_string(a_plus_b))

# Scalar multiplication
b_scalar = poly.scalar_polynomial_multiply(-3.2, polynomial_b)
print('-3.2 * Polynomial B =', poly.polynomial_to_string(b_scalar))

# Subtraction
a_minus_b = poly.subtract_polynomial(polynomial_a, polynomial_b)
print('A - B =', poly.polynomial_to_string(a_minus_b))
```

When you run it, you should get the following:

```
Polynomial A = -5.0x^3 + 3.0x^2 + -4.0x + 9.0
Polynomial B = 1.0x^4 + 2.0x^3 + 4.0x^2 + -9.0
Polynomial B at 3 = 162.0
A + B = 1.0x^4 + -3.0x^3 + 7.0x^2 + -4.0x
-3.2 * Polynomial B = -3.2x^4 + -6.4x^3 + -12.8x^2 + 28.8
A - B = -1.0x^4 + -7.0x^3 + -1.0x^2 + -4.0x + 18.0
```

Now you are ready to implement multiplication of polynomials. The function will look like this:

```
def multiply_polynomials(a, b):
    ...Your code here...
```

It will return a list of coefficients.

In an exercise in the last chapter, you were asked “ Let’s say I have two polynomials, p_1 and p_2 . p_1 has degree 23. p_2 has degree 12. What is the degree of their product?” The answer was $23 + 12 = 35$.

In our implementation, a polynomial of degree 23 is held in a list of length 24.

In Python we will be trying to multiply a polynomial *a* and a polynomial *b* represented as lists. What is the degree of that product?

```
result_degree = (len(a) - 1) + (len(b) - 1)
```

Now, we need to create an array of zeros that is one longer than that. Here is a cute Python trick: if you have a list, you can replicate it using the * operator.

```
a = [5,7]
b = a * 4
print(b)
# [5, 7, 5, 7, 5, 7, 5, 7]
```

Here's how you will get a list of zeros:

```
result = [0.0] * (result_degree + 1)
```

We will step through *a* getting the index and value of each entry. You can do this in one line using enumerate:

```
for a_degree, a_coefficient in enumerate(a):
```

For each of those, we will step through the entire *b* polynomial. As you multiply together each term, you will add it to appropriate coefficient of the result.

Here is the whole function:

```
def multiply_polynomials(a, b): # What is the degree of the resulting
polynomial?  result_degree = (len(a) - 1) + (len(b) - 1)

# Make a list of zeros to hold the coefficients result = [0.0] *
(result_degree + 1)

# Iterate over the indices and values of a for a_degree,
a_coefficient in enumerate(a):

    # Iterate over the indices and values of b for b_degree,
    b_coefficient in enumerate(b):
```

```
# Calculate the resulting monomial coefficient =
a_coefficient * b_coefficient degree = a_degree + b_degree

# Add it to the right bucket
result[degree] = result[degree] + coefficient

return result
```

Take a long look at that function. When you understand it, type it into `poly.py`.

In `test.py`, try out the new function:

```
# Multiplication
a_times_b = poly.multiply_polynomials(polynomial_a, polynomial_b)
print('A x B =', poly.polynomial_to_string(a_times_b))
```

This is an example of a *nested loop*. The outer loop steps through the polynomial `a`. For each step it takes, the inner loop steps through the entire polynomial `b`.

65.1 Something surprising about lists

You can imagine that you might want to create two very similar polynomials. Let's say polynomial `c` is $x^2 + 2x + 1$ and polynomial `d` is $x^2 - 2x + 1$. You might think you are very clever to just alter that degree 1 coefficient like this:

```
c = [1.0, 2.0, 1.0]
d = c
d[1] = -2.0
```

If you printed out `c`, you would get `[1.0, -2.0, 1.0]`. Why? You assigned two variables (`c` and `d`) to the *the same list*. So when you use one reference (`d`) to change the list, you see the change if you look at the list from either reference. *FIXME: Diagram of two references to the same list here.*

To create two separate lists, you would need to explicitly make a copy:

```
c = [1.0, 2.0, 1.0]
d = c.copy()
d[1] = -2.0
```



CHAPTER 66

Differentiating Polynomials

If you had a function that gave you the height of an object, it would be handy to be able to figure out a function that gave you the velocity at which it was rising or falling. The process of converting the position function into a velocity function is known as *differentiation* or *finding the derivative*.

There are a bunch of rules for finding a derivative, but differentiating polynomials only requires three:

- The derivative of a sum is equal to the sum of the derivatives.
- The derivative of a constant is zero.
- The derivative of a nonconstant monomial at^b (a and b are constant numbers, t is time) is abt^{b-1}

So, for example, if I tell you that the height in meters of quadcopter at second t is given by $2t^3 - 5t^2 + 9t + 200$. You could tell me that its vertical velocity is $6t^2 - 10t + 9$

Exercise 85 **Differentiation of polynomials**

Differentiate the following polynomials.

Working Space

Answer on Page 835

Notice that the degree of the derivative is one less than the degree of the original polynomial. (Unless, of course, the degree of the original is already zero.)

Now, if you know that a position is given by a polynomial, you can differentiate it to find the object's velocity at any time.

The same trick works for acceleration: Let's say you know a function that gives an object's velocity. To find its acceleration at any time, you take the derivative of the velocity function.

Exercise 86 Differentiation of polynomials in Python

Write a function that returns the derivative of a polynomial in `poly.py`. It should look like this:

Working Space

```
def derivative_of_polynomial(pn):
    ...Your code here...
```

When you test it in `test.py`, it should look like this:

```
# 3x**3 + 2x + 5
p1 = [5.0, 2.0, 0.0, 3.0]
d1 = poly.derivative_of_polynomial(p1)
# d1 should be 9x**2 + 2
print("Derivative of", poly.polynomial_to_string(p1), "is", poly.polynomial_to_string(d1))

# Check constant polynomials
p2 = [-9.0]
d2 = poly.derivative_of_polynomial(p2)
# d2 should be 0.0
print("Derivative of", poly.polynomial_to_string(p2), "is", poly.polynomial_to_string(d2))
```

Answer on Page 835



CHAPTER 67

Python Classes

The built-in types, like strings have functions associated with them. So, for example, if you needed a string converted to uppercase, you would call its `upper()` function: -

```
my_string = "houston, we have a problem!"  
louder_string = my_string.upper()
```

This would set `louder_string` to "HOUSTON, WE HAVE A PROBLEM!" When a function is associated with a datatype like this, it called a *method*. A datatype with methods is known as a *class*. The data of that type is known as *instance* of that class. For example, in the example, we would say "my_string is an instance of the class str. str has a method called upper"

The function `type` will tell you the type of any data:

```
print(type(my_string))
```

This will output

```
<class 'str'>
```

A class can also define operators. `+`, for example, is redefined by `str` to concatenate strings together:

```
long_string = "I saw " + "15 people"
```

67.1 Making a Polynomial class

You have created a bunch of useful python functions for dealing with polynomials. Notice how each one has the word “polynomial” in the function name like `derivative_of_polynomial`. Wouldn’t it be more elegant if you had a `Polynomial` class with a `derivative` method? Then you could use your polynomial like this:

```
a = Polynomial([9.0, 0.0, 2.3])
b = Polynomial([-2.0, 4.5, 0.0, 2.1])

print(a, "plus", b, "is", a+b)
print(a, "times", b, "is", a*b)
print(a, "times", 3, "is", a*3)
print(a, "minus", b, "is", a-b)

c = b.derivative()

print("Derivative of", b, "is", c)
```

And it would output:

```
2.30x^2 + 9.00 plus 2.10x^3 + 4.50x + -2.00 is 2.10x^3 + 2.30x^2 + 4.50x + 7.00
2.30x^2 + 9.00 times 2.10x^3 + 4.50x + -2.00 is 4.83x^5 + 29.25x^3 + -4.60x^2 + 40.50x + -18.00
2.30x^2 + 9.00 times 3 is 6.90x^2 + 27.00
2.30x^2 + 9.00 minus 2.10x^3 + 4.50x + -2.00 is -2.10x^3 + 2.30x^2 + -4.50x + 11.00
Derivative of 2.10x^3 + 4.50x + -2.00 is 6.30x^2 + 4.50
```

Create a file for your class definition called `Polynomial.py`. Enter the following:

```
class Polynomial:
    def __init__(self, coeffs):
        self.coefficients = coeffs.copy()

    def __repr__(self):
```

```
# Make a list of the monomial strings
monomial_strings = []

# For standard form we start at the largest degree
degree = len(self.coefficients) - 1

# Go through the list backwards
while degree >= 0:
    coefficient = self.coefficients[degree]

    if coefficient != 0.0:
        # Describe the monomial
        if degree == 0:
            monomial_string = "{:.2f}".format(coefficient)
        elif degree == 1:
            monomial_string = "{:.2f}x".format(coefficient)
        else:
            monomial_string = "{:.2f}x^{}".format(coefficient, degree)

        # Add it to the list
        monomial_strings.append(monomial_string)

    # Move to the previous term
    degree = degree - 1

# Deal with the zero polynomial
if len(monomial_strings) == 0:
    monomial_strings.append("0.0")

# Separate the terms with a plus sign
return " + ".join(monomial_strings)

def __call__(self, x):
    sum = 0.0
    for degree, coefficient in enumerate(self.coefficients):
        sum = sum + coefficient * x ** degree
    return sum

def __add__(self, b):
    result_length = max(len(self.coefficients), len(b.coefficients))
    result = []
    for i in range(result_length):
        if i < len(self.coefficients):
            coefficient_a = self.coefficients[i]
        else:
            coefficient_a = 0.0
```

```
        if i < len(b.coefficients):
            coefficient_b = b.coefficients[i]
        else:
            coefficient_b = 0.0
        result.append(coefficient_a + coefficient_b)

    return Polynomial(result)

def __mul__(self, other):

    # Not a polynomial?
    if not isinstance(other, Polynomial):
        # Try to make it a constant polynomial
        other = Polynomial([other])

    # What is the degree of the resulting polynomial?
    result_degree = (len(self.coefficients) - 1) + (len(other.coefficients) - 1)

    # Make a list of zeros to hold the coefficients
    result = [0.0] * (result_degree + 1)

    # Iterate over the indices and values of a
    for a_degree, a_coefficient in enumerate(self.coefficients):

        # Iterate over the indices and values of b
        for b_degree, b_coefficient in enumerate(other.coefficients):

            # Calculate the resulting monomial
            coefficient = a_coefficient * b_coefficient
            degree = a_degree + b_degree

            # Add it to the right bucket
            result[degree] = result[degree] + coefficient

    return Polynomial(result)

__rmul__ = __mul__

def __sub__(self, other):
    return self + other * -1.0

def derivative(self):

    # What is the degree of the resulting polynomial?
    original_degree = len(self.coefficients) - 1
```

```
if original_degree > 0:  
    degree_of_derivative = original_degree - 1  
else:  
    degree_of_derivative = 0  
  
# We can ignore the constant term (skip the first coefficient)  
current_degree = 1  
result = []  
  
# Differentiate each monomial  
while current_degree < len(self.coefficients):  
    coefficient = self.coefficients[current_degree]  
    result.append(coefficient * current_degree)  
    current_degree = current_degree + 1  
  
# No terms? Make it the zero polynomial  
if len(result) == 0:  
    result.append(0.0)  
  
return Polynomial(result)
```

Create a second file called `test_polynomial.py` to test it:

```
from Polynomial import Polynomial  
  
a = Polynomial([9.0, 0.0, 2.3])  
b = Polynomial([-2.0, 4.5, 0.0, 2.1])  
  
print(a, "plus", b, "is", a+b)  
print(a, "times", b, "is", a*b)  
print(a, "times", 3, "is", a*3)  
print(a, "minus", b, "is", a-b)  
  
c = b.derivative()  
  
print("Derivative of", b, "is", c)  
  
slope = c(3)  
print("Value of the derivative at 3 is", slope)
```

Run the test code:

```
python3 test_polynomial.py
```




CHAPTER 68

Common Polynomial Products

In math and physics, you will run into certain kinds of polynomials over and over again. In this chapter, I am going to cover some patterns that you will want to start to recognize.

68.1 Difference of squares

Watch **Polynomial special products: difference of squares** from Khan Academy at <https://youtu.be/uNweU6I4Icw>.

If you are asked what is $(3x - 7)(3x + 7)$, you would use the distributive property to expand that to $(3x)(3x) + (3x)(7) + (-7)(3x) + (-7)(7)$. Two of the terms cancel each other, so this is $(3x)^2 - (7)^2$. This would simplify to $9x^2 - 49$.

You will see this pattern a lot. Anytime you see $(a + b)(a - b)$, you should immediately recognize it equals $a^2 - b^2$. (Note that the order doesn't matter: $(a - b)(a + b)$ also $a^2 - b^2$.)

Working the other way is important too: anytime you see $a^2 - b^2$, that you should recognize that you can change that into the product $(a + b)(a - b)$. Making something into a product

like this is known as *factoring*. You probably have done prime factorization of numbers like $42 = 2 \times 3 \times 7$. In the next couple of chapters you will learn to factorize polynomials.

Exercise 87 Difference of Squares

Simply the following products

Working Space

1. $(2x - 3)(2x + 3)$
2. $(7 + 5x^3)(7 - 5x^3)$
3. $(x - a)(x + a)$
4. $(3 - \pi)(3 + \pi)$
5. $(-4x^3 + 10)(-4x^3 - 10)$
6. $(x + \sqrt{7})(x - \sqrt{7})$ Factor the following polynomials:
 7. $x^2 - 9$
 8. $49 - 16x^6$
 9. $\pi^2 - 25x^8$
 10. $x^2 - 5$

Answer on Page 836

We are often interested in the roots of a polynomial, that is we want to know “For what values of x does the polynomial evaluate to zero?” For example, when you deal with falling bodies, the first question you might ask would be “How many seconds before the hammer hits the ground?” Once you have factored a polynomial into binomials, you can easily find the roots.

For example, what are the roots of $x^2 - 5$? You just factored it into $(x + \sqrt{5})(x - \sqrt{5})$. This product is zero if and only if one of the factors is zero. The first factor is only zero when x is $-\sqrt{5}$. The second factor is zero only when x is $\sqrt{5}$. Those are the only two roots of this polynomial.

Let’s check that result. $\sqrt{5}$ is a little more than 2.2. Using your Python code, you can graph the polynomial:

```
import poly.py
import matplotlib.pyplot as plt
```

```
# x**2 - 5
pn = [-5.0, 0.0, 1.0]

# These lists will hold our x and y values
x_list = []
y_list = []

# Start at x=-3
current_x = -3.0

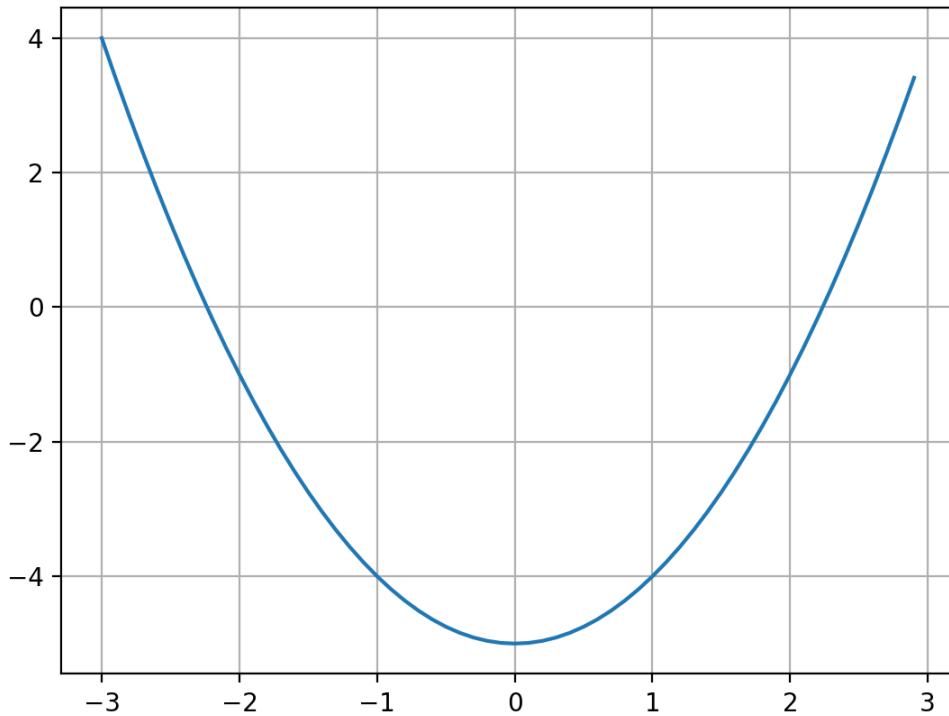
# End at x=3.0
while current_x < 3.0:
    current_y = poly.evaluate_polynomial(pn, current_x)

    # Add x and y to respective lists
    x_list.append(current_x)
    y_list.append(current_y)

    # Move x forward
    current_x += 0.1

# Plot the curve
plt.plot(x_list, y_list)
plt.grid(True)
plt.show()
```

You should get a plot like this:



It does, indeed, seem to cross the x-axis near -2.2 and 2.2.

68.2 Powers of binomials

You can raise whole polynomials to exponents. For example,

$$\begin{aligned}(3x^3 + 5)^2 &= (3x^3 + 5)(3x^3 + 5) \\ &= 9x^6 + 15x^3 + 15x^3 + 25 = 9x^6 + 30x^3 + 25\end{aligned}$$

A polynomial with two terms is called a *binomial*. $5x^9 - 2x^4$, for example, is a binomial. In this section, we are going to develop some handy techniques for raising a binomial to some power.

Looking at the previous example, you can see that for any monomials a and b , $(a + b)^2 = a^2 + 2ab + b^2$. So, for example, $(7x^3 + \pi)^2 = 49x^6 + 14\pi x^3 + \pi^2$

Exercise 88 Squaring binomials

Simply the following

Working Space

1. $(x + 1)^2$
2. $(3x^5 + 5)^2$
3. $(x^3 - 1)^2$
4. $(x - \sqrt{7})^2$

Answer on Page 837

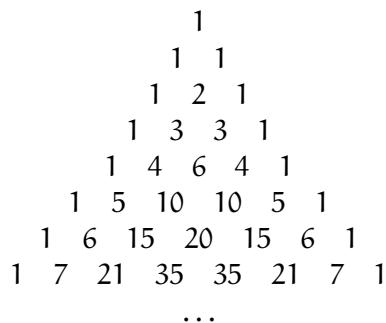
What about $(x + 2)^3$? You can do it as two separate multiplications:

$$\begin{aligned}(x + 2)^3 &= (x + 2)(x + 2)(x + 2) \\ &= (x + 2)(x^2 + 4x + 4) = x^3 + 4x^2 + 4x + 2x^2 + 8x + 8 \\ &= x^3 + 6x^2 + 12x + 8\end{aligned}$$

And, in general, we can say that for any monomials a and b , $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$.

What about higher powers? $(a+b)^4$, for example? You could use the distributive property four times, but it starts to get pretty tedious.

Here is a trick. This is known as *Pascal's triangle*



Each entry is the sum of the two above it.

The coefficients of each term are given by the entries in Pascal's triangle:

$$(a + b)^4 = 1a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + 1b^4$$

Exercise 89 Using Pascal's Triangle

Working Space

1. What is $(x + \pi)^5$?

Answer on Page 837



CHAPTER 69

Factoring Polynomials

We factor a polynomial into two or more polynomials of lower degree. For example, let's say that you wanted to factor $5x^3 - 45x$. You would note that you can factor out $5x$ from every term. Thus,

$$5x^3 - 45x = (5x)(x^2 - 9)$$

And then, you might notice that the second factor looks like the difference of squares, so

$$5x^3 - 45x = (5x)(x + 3)(x - 3)$$

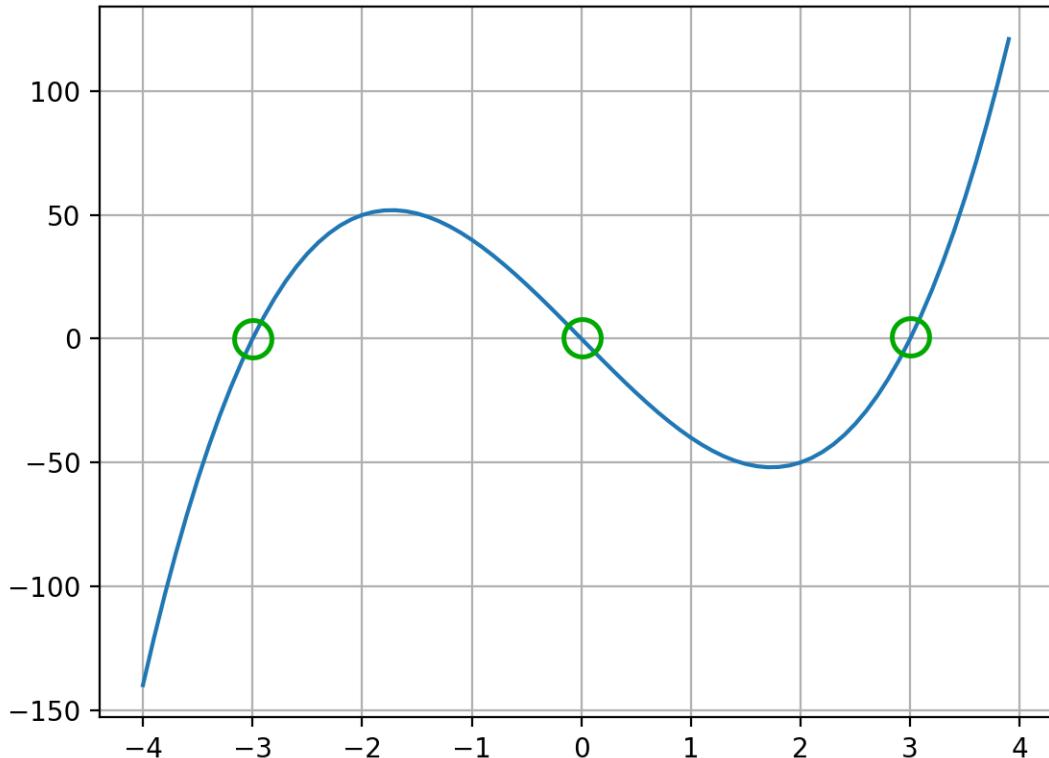
That is as far as we can factorize this polynomial.

Why do we care? The factors make it easy to find the roots of the polynomial. This polynomial evaluates to zero if and only if at least one of the factors is zero. Here we see that

- The factor $(5x)$ is zero when x is zero.
- The factor $(x + 3)$ is zero when x is -3 .
- The factor $(x - 3)$ is zero when x is 3 .

So looking at the factorization, you can see that $5x^3 - 45x$ is zero when x is 0, -3, or 3.

This is a graph of that polynomial with its roots circled:



69.1 How to factor polynomials

The first step when you are trying to factor a polynomial is to find the greatest common divisor for all the terms, and pull that out. In this case, the greatest common divisor will also be a monomial: its degree is the least of the degrees of the terms, its coefficient will be the greatest common divisor of the coefficients of the terms.

For example, what can you pull out of this polynomial?

$$12x^{100} + 30x^{31} + 42x^{17}$$

The greatest common divisor of the coefficients (12, 30, and 42) is 6. The least of the degrees of terms (100, 31, and 17) is 17. So you can pull out $6x^{17}$:

$$12x^{100} + 30x^{31} + 42x^{17} = (6x^{17})(2x^{83} + 5x^{14} + 7)$$

Exercise 90 Factoring out the GCD monomial*Working Space**Answer on Page 837*

So, now you have the product of a monomial and a polynomial. If you are lucky, the polynomial part looks familiar, like the difference of squares or a row from Pascal's triangle.

Often you are trying factor a quadratic like $x^2 + 5x + 6$ in a pair of binomials. In this case, the result would be $(x + 3)(x + 2)$. Let's check that:

$$(x + 3)(x + 2) = (x)(x) + (3)(x) + (2)(x) + (3)(2) = x^2 + 5x + 6$$

Notice that 3 and 2 multiply to 6 and add to 5. If I were trying to factor $x^2 + 5x + 6$, I would ask myself "What are two numbers that when multiplied equal 6 and when added equal 5?" And I would might guess wrong a couple of times. For example, I might say to myself "Well, 6 times 1 is 6. Maybe those work. But 6 and 1 add 7. So those don't work."

Solving these sorts of problems are like solving a Sudoku puzzle: you try things and realize they are wrong, so you backtrack and try something else.

The numbers are sometimes negative. For example, $x^2 + 3x - 10$ factors into $(x + 5)(x - 2)$.

Exercise 91 Factoring quadratics*Working Space**Answer on Page 837*



CHAPTER 70

Practice with Polynomials

At this point, you know all the pieces necessary to solve problems involving polynomials. In this chapter, you are going to practice using all of these ideas together.

Watch Khan Academy's **Polynomial identities introduction here:** <https://youtu.be/EvNKKyhLSpQ> Also watch the follow up here: <https://youtu.be/-6qi049Q180>

FIXME: Lots of practice problems here



CHAPTER 71

Graphing Polynomials

In using polynomials to solve real-world problems, it is often handy to know what the graph of the polynomial looks like. You have many of the tools you need to start to sketch out the graphs:

- To find where the graph crosses the y -axis, you can evaluate the polynomial at $x = 0$.
- To find where the graph crosses the x -axis, you can find the roots of the polynomial.
- To find the level spots on the graph (often the top of a hump or the bottom of a dip), you can take the derivative of the polynomial (which is a polynomial), and find the roots of that.

FIXME: Diagram of those things

For example, if you wanted to graph the polynomial $f(x) = -x^3 - x^2 + 6x$, you might plug in a few values that are easy to compute:

- $f(-2) = -8$

- $f(-1) = -6$
- $f(0) = 0$
- $f(1) = 4$
- $f(2) = 0$

So, right away we know two roots: $x = 0$ and $x = 2$. Are there others? We won't know until we factor the polynomial:

$$\begin{aligned} -x^3 - x^2 + 6x &= (-1x)(x^2 + x - 6) \\ &= (-1x)(x + 3)(x - 2) \end{aligned}$$

So, yes, there is a third root: $x = -3$

What about the level spots? $f'(x) = -3x^2 - 2x + 6$. Where is that zero?

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

We have a formula for quadratics like this:

$$\begin{aligned} x &= -\frac{b}{2} \pm \frac{\sqrt{b^2 - 4c}}{2} \\ &= -\frac{\frac{2}{3}}{2} \pm \frac{\sqrt{\left(\frac{2}{3}\right)^2 - 4(-2)}}{2} \\ &= -\frac{1}{3} \pm \frac{\sqrt{\frac{4}{9} + 8}}{2} \\ &= -\frac{1}{3} \pm \frac{\sqrt{\frac{85}{9}}}{2} \\ &= -\frac{1}{3} \pm \frac{\sqrt{85}}{6} \\ &\approx 1.20 \text{ and } -1.87 \end{aligned}$$

Now you might plug those numbers in:

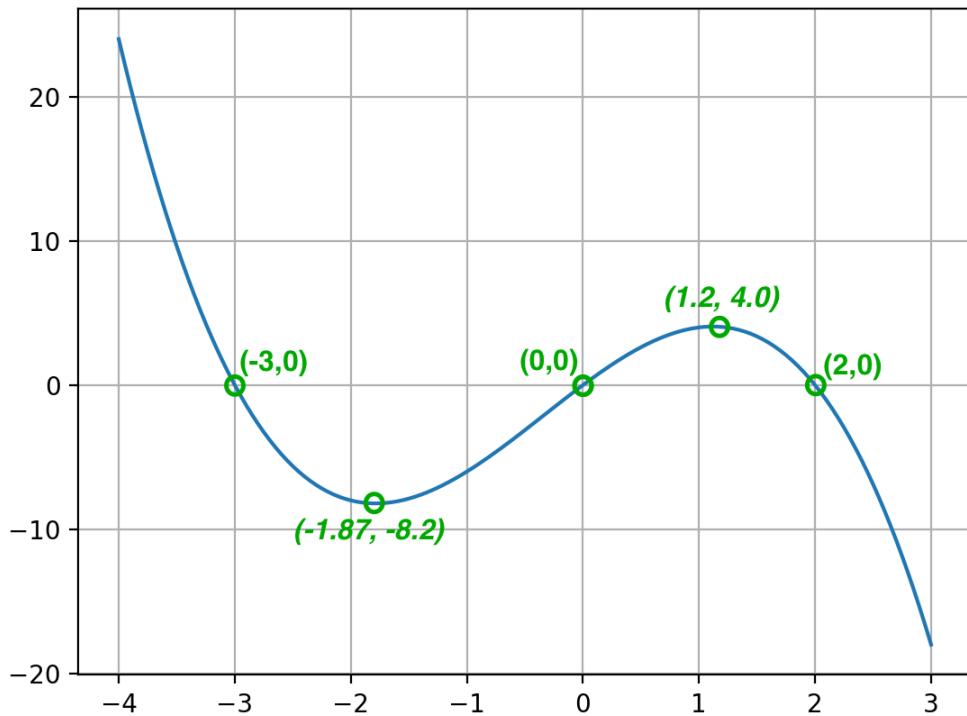
- $f(1.2) \approx 4.0$
- $f(-1.87) \approx -8.2$

71.1 Leading term in graphing

There is one more trick you need before you can draw a good graph of a polynomial. As you go farther and farther to the left and right, where does the function go? That is, does the graph go up on both ends (like a smile)? Or does it go down on both ends (like a frown)? Or does the negative end go down (frowny) while the positive end go up (smiley)? Or does the negative go up (smiley) and the positive end go down (frowny)?

Assuming the polynomial is not constant, there are only those four possibilities. It is determined entirely by the leading term of the polynomial. If the degree of the leading term is even, both ends go in the same direction (both are smiley or both are frowny). If the coefficient of the leading term is positive, the positive end is smiley.

The graph we are working on has a leading term of $-1x^3$. The degree is odd, thus the ends go in different directions. The coefficient is negative, so the positive end points down. Now you can draw the graph, which should look something like this:





CHAPTER 72

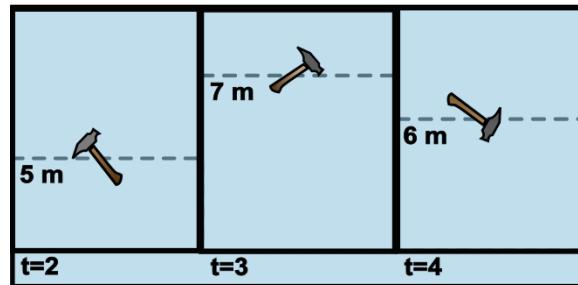
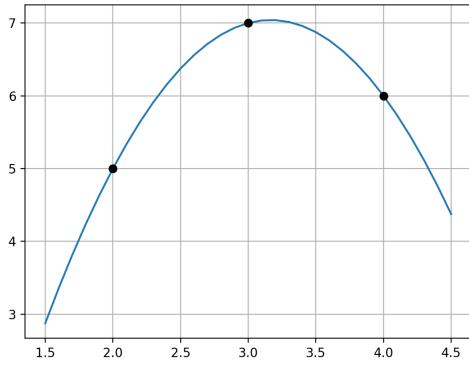
Interpolating with Polynomials

Let's say someone on a distant planet records video of a hammer being throw up into the air. They send you three random frames of the hammer in flight. Each frame has a timestamp and you can clearly see how high the hammer is in each one. Can you create a 2nd degree polynomial that explains the entire flight of the hammer?

That is, you have three points $(t_0, h_0), (t_1, h_1), (t_2, h_2)$. Can you find a, b, c such that the graph of $at^2 + bt + c = t$ passes through all three points?

The answer is yes. In fact, given any n points, there is exactly one $n - 1$ degree polynomial that passes through all the points.

There are a lot of variables floating around. Let's make it concrete: The photos are taken at $t = 2$ seconds, $t = 3$ seconds, and $t = 4$ seconds. In those photos, the height of the hammer is 5m, 7m, and 6m. So, we want our polynomial to pass through these points: $(2, 5), (3, 7), (4, 6)$.



How can you find that polynomial? Let's do it in small steps. Can you create a 2nd degree polynomial that is not zero at $t = 2$, but is zero at $t = 3$ and $t = 4$? Yes, you can: $(x - 3)(x - 4)$ has exactly two roots at $t = 3$ and $t = 4$. The value of this polynomial at $t = 2$ is $(2 - 3)(2 - 4) = 2$. We really want it to be 5m, so we can divide the whole polynomial by 2 and multiply it by 5.

Now we have the polynomial:

$$f_0(x) = \frac{5}{(2-3)(2-4)}(x-3)(x-4) = \frac{5}{2}x^2 - \frac{35}{2}x + 30$$

This is a second degree polynomial that is 5 at $t = 2$ and 0 at $t = 3$ and $t = 4$.

Now we create a polynomial that is 7 at $t = 3$ and 0 at $t = 2$ and $t = 4$:

$$f_1(x) = \frac{7}{(3-2)(3-4)}(x-2)(x-4) = -7x^2 + 42x - 56$$

Finally, we create a polynomial that is 6 at $t = 4$ and zero at $t = 2$ and $t = 3$:

$$f_2(x) = \frac{6}{(4-2)(4-3)}(x-2)(x-3) = 3x^2 - 15x + 18$$

Adding these three polynomials together gives you a new polynomial that touches all three points:

$$f(x) = \frac{5}{2}x^2 - \frac{35}{2}x + 30 - 7x^2 + 42x - 56 + 3x^2 - 15x + 18 = -\frac{3}{2}x^2 + \frac{19}{2}x - 8$$

You can test this with your Polynomial class. Create a file called `test_interpolation.py`. Add this code:

```
from Polynomial import Polynomial
import matplotlib.pyplot as plt
```

```

in_x = [2,3,4]
in_y = [5,7,6]

pn = Polynomial([-8, 19/2, -3/2])
print(pn)

# These lists will hold our x and y values
x_list = []
y_list = []

# Starting x
current_x = 1.5

while current_x <= 4.5:
    # Evaluate pn at current_x
    current_y = pn(current_x)

    # Add x and y to respective lists
    x_list.append(current_x)
    y_list.append(current_y)

    # Move x forward
    current_x += 0.05

# Plot the curve
plt.plot(x_list, y_list)

# Plot black circles on the given points
plt.plot(in_x, in_y, "ko")
plt.grid(True)
plt.show()

```

You should get a nice plot that shows the graph of the polynomial passing through those three points.

In general, then, if you give me any three points $(t_0, h_0), (t_1, h_1), (t_2, h_2)$, here is a second degree polynomial that pass through all three points:

$$\frac{h_0}{(t_0 - t_1)(t_0 - t_2)}(x - t_1)(x - t_2) + \frac{h_1}{(t_1 - t_0)(t_1 - t_2)}(x - t_0)(x - t_2) + \frac{h_2}{(t_2 - t_0)(t_2 - t_1)}(x - t_0)(x - t_1)$$

What if you are given 9 points $((t_0, h_0), (t_1, h_1), \dots, (t_8, h_8))$ and want to find a 8th degree polynomial that passes through all of them? Just what you would expect:

$$\frac{h_0}{(t_0 - t_1)(t_0 - t_2) \dots (t_0 - t_8)}(x - t_1)(x - t_2) \dots (x - t_8) + \dots + \frac{h_8}{(t_8 - t_0)(t_8 - t_1) \dots (t_8 - t_7)}(x - t_0) \dots (x - t_7)$$

FIXME: Do I need to define summation and prod here?

The general solution is, given n points, the $n - 1$ degree polynomial that goes through them is

$$y = \sum_{i=0}^n \left(\prod_{\substack{0 \leq j \leq n \\ j \neq i}} \frac{x - t_j}{t_i - t_j} \right) h_i$$

That would be tedious for a person to compute, but computers love this stuff. Let's create a method that creates instances of Polynomial using interpolation.

72.1 Interpolating polynomials in python

Your method will take two lists of numbers, one contains x -values and the other contains y -values. So comment out the line that creates the polynomial in `test_interpolation.py` and create it from two lists:

```
in_x = [2,3,4]
in_y = [5,7,6]
# pn = Polynomial([-8, 19/2, -3/2])
pn = Polynomial.from_points(in_x, in_y)
print(pn)
```

Add the following method to your `Polynomial` class in `Polynomial.py`

```
@classmethod
def from_points(cls, x_values, y_values):
    coef_count = len(x_values)

    # Sums start with a zero polynomial
    sum_pn = Polynomial([0.0] * coef_count)
    for i in range(coef_count):

        # Products start with the constant 1 polynomial
        product_pn = Polynomial([1.0])
        for j in range(coef_count):

            # Must skip j=i
            if j != i:
                # (1x - x_values[j]) has a root at x_values[j]
                factor_pn = Polynomial([-1 * x_values[j], 1])
                product_pn = product_pn * factor_pn
```

```
# Scale so product_pn(x_values[i]) = y_values[i]
scale_factor = y_values[i] / product_pn(x_values[i])
scaled_pn = scale_factor * product_pn

# Add it to the sum
sum_pn = sum_pn + scaled_pn

return sum_pn
```

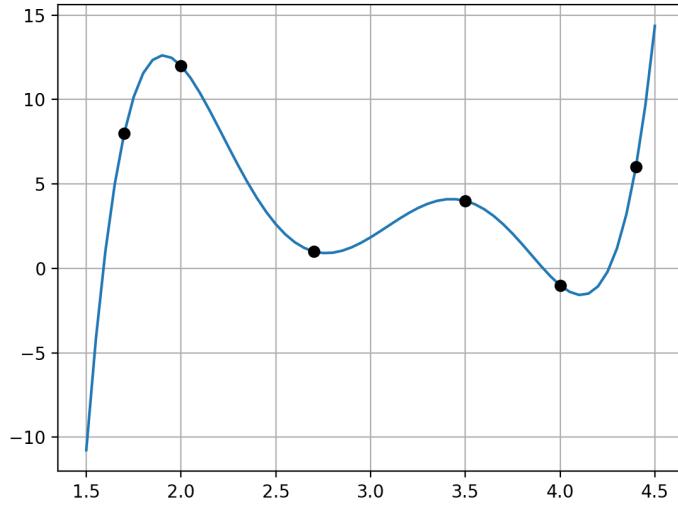
It should work exactly the same as before. You should get the same polynomial printed out as before. You shoud get the same plot of the curve passing through the three points.

How about five points? Change `in_x` and `in_y` at the start of `test_interpolation.py`:

```
in_x = [1.7, 2, 2.7, 3.5, 4, 4.4]
in_y = [8, 12, 1, 4, -1, 6]
```

You should get a polynomial that passes through all five points:

$$11.21x^5 - 171.05x^4 + 1019.44x^3 - 2957.53x^2 + 4161.78x - 2258.75$$



It should look like this:



CHAPTER 73

Limits

The asymptotic behavior we see in rational functions suggests that we need to expand our vocabulary of function characteristics. We examined vertical asymptotes and end behavior through graphs and tables and discussed them in English. The language of limits enables us to discuss these attributes with greater efficiency.

Let us revisit an example from the previous chapter. This function has a hole at $x = 1$, a vertical asymptote at $x = 3$, and a horizontal asymptote of $y = 1$.

$$f(x) = \frac{x^2 - 3x + 2}{x^2 - 4x + 3} = \frac{(x-1)(x-2)}{(x-1)(x-3)}$$

First, consider the vertical asymptote. We see that the graph goes down as it hugs the left side of the vertical asymptote, and goes up as it hugs the right side. We can describe these behaviors as the left- and right-hand limits, respectively. We say that the left-hand limit of f at $x = 3$ is negative infinity. Another way of communicating this is to say that as x approaches 3 from the left, the function approaches negative infinity. Symbolically, we summarize this as $\lim_{x \rightarrow 3^-} f(x) = -\infty$.

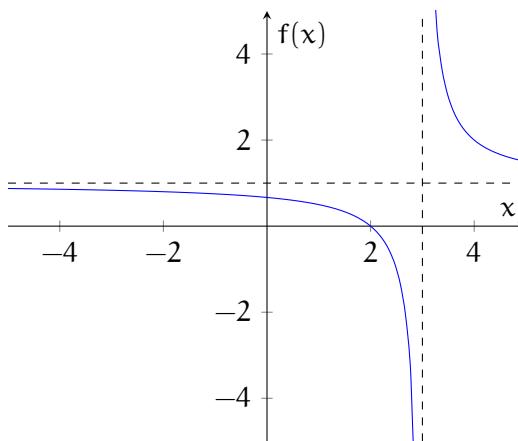


Figure 73.1: Graph of $f(x) = \frac{x^2 - 3x + 2}{x^2 - 4x + 3}$ with asymptotes

Similarly, the right-hand limit of f at $x = 3$ is positive infinity. In other words, as x approaches 3 from the right, the function approaches positive infinity. Symbolically, we write $\lim_{x \rightarrow 3^+} f(x) = \infty$.

The limit of a function at a particular x -value is the y -value that the function approaches as it approaches the given x -value. In the previous example, we could only specify the left- and right-hand limits, because they were different. In cases where the left- and right-hand limits are equal, we can say that the function has a limit there. The hole in our function f is one such value. We see that as we approach the hole from both the left and right, the function takes on values near $\frac{1}{2}$. This is more apparent numerically:

x	0.9	0.99	0.999	1	1.001	1.01	1.1
$f(x)$	0.5238	0.5025	0.5003	undefined	0.4998	0.4975	0.4737

The left-hand and right-hand limits of f at 1 are both $\frac{1}{2}$. Since they are equal, we can also say that the limit of f at 1 is $\frac{1}{2}$. This allows us to efficiently discuss the behavior of f at 1, even though the function is not defined there since substituting 1 into the function gives division by zero.

$$\lim_{x \rightarrow 1^-} f(x) = \lim_{x \rightarrow 1^+} f(x) = \lim_{x \rightarrow 1} f(x) = \frac{1}{2}$$

We can also talk about limits at x -values where nothing weird is happening, that is, no hole or vertical asymptote. For example, as x approaches 4 from the left and right, y approaches 2.

x	3.9	3.99	3.999	4	4.001	4.01	4.1
$f(x)$	2.1111	2.0101	2.0010	2	1.9990	1.9901	1.9091

In this case, since nothing weird is happening, the limit is equal to the function value. This is an example of continuity, which we will discuss in more detail in the next chapter. By contrast, at the vertical asymptote $x = 1$, since the left- and right-hand limits are not equal, we say the function does not have a limit, or the limit does not exist.

Finally, let us consider the horizontal asymptote of f . The graph hugs the line $y = 1$ as x goes far to the left and far to the right. We say that as x approaches negative infinity, f approaches 1, and likewise, that as x approaches positive infinity, f approaches 1. We write these symbolically as $\lim_{x \rightarrow -\infty} f(x) = 1$ and $\lim_{x \rightarrow \infty} f(x) = 1$.

Exercise 92 Limits Practice 1

Determine the left- and right-hand limits of the function as x approaches the given values. At x -values where the limit exists, determine it.

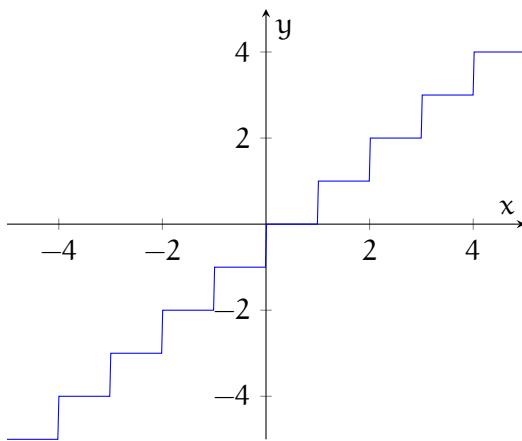
Working Space

- $p(x) = \frac{x+3}{x^2+9x+18}, x = -6, -5, -3, \infty$

Answer on Page 837

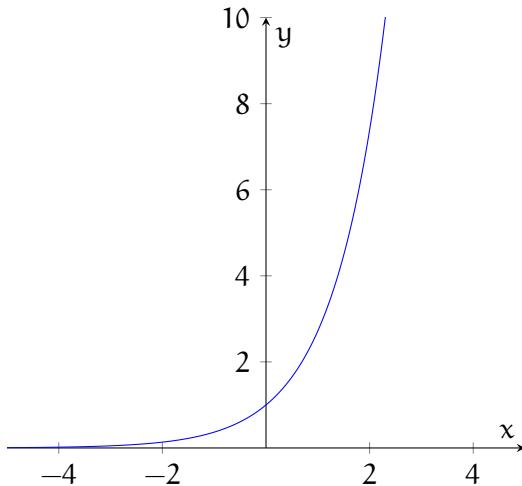
We have seen two weird behaviors of rational functions at certain x -values: holes and vertical asymptotes. Now we will examine another type of weird behavior: jumps. This is a characteristic of some piecewise defined functions. In piecewise defined functions, the domain is divided into two or more pieces, and a different expression is used to give the y -value depending on which piece contains the x -value. One common piecewise defined function is the floor function, sometimes denoted $\lfloor x \rfloor$. The standard floor function rounds any real number down to the nearest integer. So, for a price quoted in dollars and cents, the floor would be just the number of dollars.

When x is exactly 1, the function value is 1: the number of dollars in a price of \$1.00. When x is any number greater than 1 but less than 2, the function value is still 1. Also, $\lfloor 1.01 \rfloor, \lfloor 1.5 \rfloor$, and $\lfloor 1.99999 \rfloor$ are all 1. As we continue to look to the right, once x equals

Figure 73.2: Graph of $y = |x|$

exactly 2, it jumps up to the value 2. So, $\lim_{x \rightarrow 2^-} |x| = 1$, while $\lim_{x \rightarrow 2^+} |x| = 2$.

Besides rational and piecewise defined functions, there are other functions with interesting limits. Consider the standard exponential function, $y = e^x$.

Figure 73.3: Graph of $y = e^x$

As x increases, y increases without bound; that is, $\lim_{x \rightarrow \infty} e^x = \infty$. However, looking far to the left, we see that y hugs the x -axis. This is because raising e to a large negative exponent is the same as 1 divided by e raised to a large positive exponent; that is, 1 divided by a very large number, which yields a very small positive number. In limit notation, $\lim_{x \rightarrow -\infty} e^x = 0$. This example illustrates that horizontal asymptotes need not model end behavior in both directions. Note that this reasoning holds for $y = b^x$ for any $b > 1$, so all such functions have the same horizontal asymptote, $y = 0$.

We know that the natural logarithm function, $y = \ln x$, is the inverse of $y = e^x$. Since inverse functions swap the role of x and y , it stands to reason that a horizontal asymptote

in one function corresponds with a vertical asymptote in the other function, and that is indeed the case.

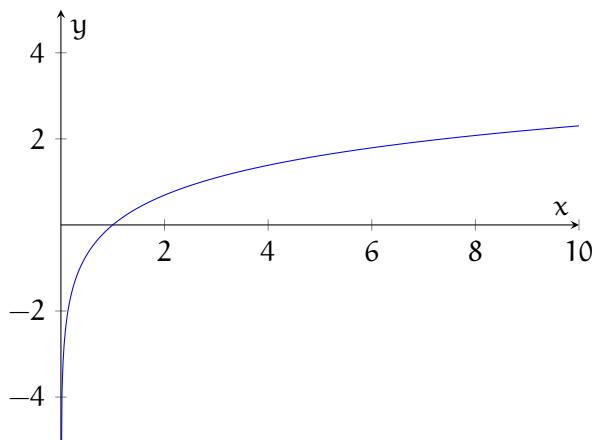


Figure 73.4: Graph of $y = \ln x$

An untransformed logarithm function is defined only for positive inputs. That is because it is not possible to find an exponent of a positive number which will yield a negative or zero result. What type of exponent on a positive number yields a number near zero? That would be a large-magnitude negative number. So, on the logarithm graph, large negative y -values correspond with x -values only slightly greater than zero. So, $\ln x$ (and $\log_2 x$, and indeed $\log_b x$ for any $b > 1$) approaches negative infinity as x approaches 0 from the right. There is no left-hand limit at 0, however. In limit notation, $\lim_{x \rightarrow 0^+} \ln x = -\infty$.

Exercise 93 Limits Practice 2

State the asymptotes of the following transformed exponential and logarithmic functions. Give the limit statement which describes the behavior of the function along the asymptote.

Working Space

1. $y = 3^x + 1, y = \log_2(x - 4), y = 2^{1-x}, y = \log_{10}(-2x)$

Answer on Page 837

We conclude this chapter by considering two functions which each have two horizontal asymptotes. These two seemingly obscure functions are quite important in data science.

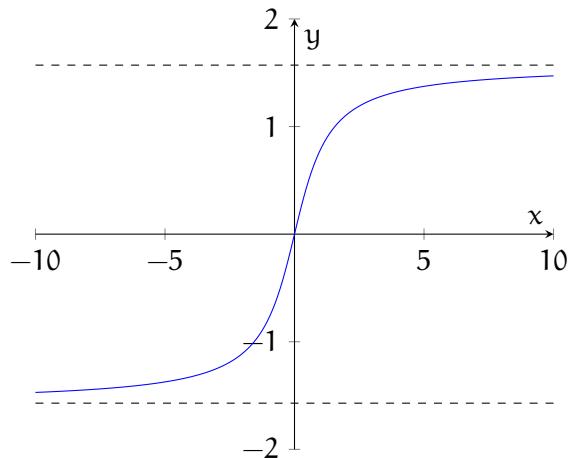


Figure 73.5: Graph of $y = \arctan x$

We know that the arctangent, or inverse tangent, function is the inverse of the piece of the tangent function which passes through the origin. The vertical asymptotes bounding this piece become horizontal asymptotes when the function is inverted.

Here are the equation and graph of the logistic function:

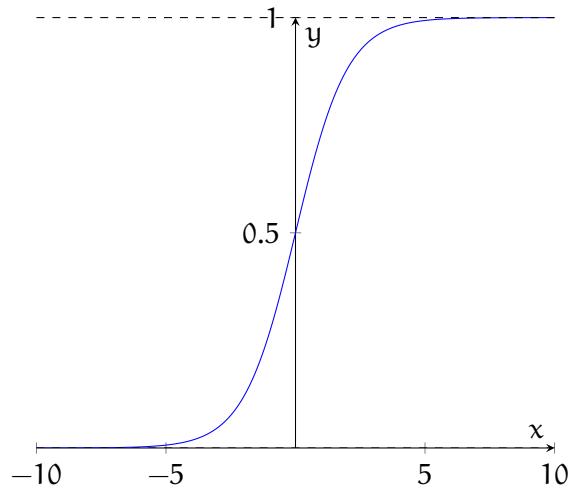


Figure 73.6: Graph of the logistic function, $y = \frac{1}{1+e^{-x}}$

For large magnitude negative values of x , the exponential term in the denominator becomes a very large positive value. The fraction thus becomes a positive number very close to zero. For large magnitude positive values of x , that exponential term becomes a very small positive number. Adding it to 1 yields a denominator just barely greater than 1. Dividing 1 by this number thus yields a function value just barely less than 1. So, the logistic function yields values between 0 and 1, though never equaling either of these values exactly. It is precisely this characteristic which makes the logistic function so useful.

Exercise 94 Limits Practice 3

Using limit notation, state the limits as x approaches negative and positive infinity for the inverse tangent and logistic functions.

Working Space

Answer on Page 838



CHAPTER 74

Rational Functions

We have discussed addition, subtraction, and multiplication of polynomials. What about division?

A quotient of polynomials is called a rational expression. When the polynomials are factored and the stars align, we can simplify the rational expression to a single polynomial, just like we might reduce a fraction to lowest terms.

Example

$$\begin{aligned}\frac{(x+1)(x+5)}{x+5} &= (x+1) * \frac{x+5}{x+5} \\ &= x+1\end{aligned}\tag{74.1}$$

What if the polynomials are not factored? Factor them first.

Example

$$\frac{x^2 + 6x + 5}{x + 5} = \frac{(x+1)(x+5)}{x+5}$$

and simplify as in the previous example.

Now, let us consider a rational expression which can be simplified to a single polynomial - but in the denominator.

Example

$$\begin{aligned}
 \frac{x+5}{x^2+6x+5} &= \frac{x+5}{(x+1)(x+5)} \\
 &= \frac{1}{x+1} * \frac{x+5}{x+5} \\
 &= \frac{1}{x+1}
 \end{aligned} \tag{74.2}$$

Consider this expression as a function: $f(x) = \frac{1}{x+1}$. As you might have guessed, this is called a rational function. We did not bother looking at the result of the previous example as a function, because we already know that function type: it is a line with slope 1 and y-intercept 1. But this rational function is another animal entirely. Let us examine our first rational function with a familiar concept: the y-intercept.

y-intercept: $f(0) = \frac{1}{0+1} = \frac{1}{1} = 1$. The graph contains the point $(0, 1)$.

Does f have an x -intercept? That would be an x -value where $f(x) = 0$. But a fraction equals 0 only when its numerator equals 0; since the numerator of this expression is always 1, f has no x -intercept.

Knowing the y-intercept, and that there is no x -intercept, is a comforting start. But things get weird when we consider a concept that has previously seemed quite simple: domain. Recall that the domain of a function is the set of all values which can be used as inputs. In this case, the domain includes all real numbers, with one exception. The number -1 is not a valid input because $f(-1) = \frac{1}{-1+1} = \frac{1}{0}$, which is undefined. So, we say that the domain is all real numbers except -1 . This means the graph contains a point corresponding to every x -value except -1 .

There is no point at $x = -1$, but there is a point at every other x -value, such as, say, -1.1 , or -0.99999 . So what is happening near $x = -1$?

x	-1.1	-1.01	-1.001	-0.999	-0.99	-0.9
$f(x)$	-10	-100	-1000	1000	100	10

The function is going haywire: as we choose x -values closer and closer to -1 , the resulting function values are larger and larger in magnitude. Also, they are negative on one side, but positive on the other. So how does a graph go from y -values of -10 , to -100 , to -1000 , all in a space of less than 0.1 on the x -axis? And then all of a sudden to big positive numbers on the other side of $x = -1$? All without ever crossing the x -axis (since there is no x -intercept)? Let us look at the graph.

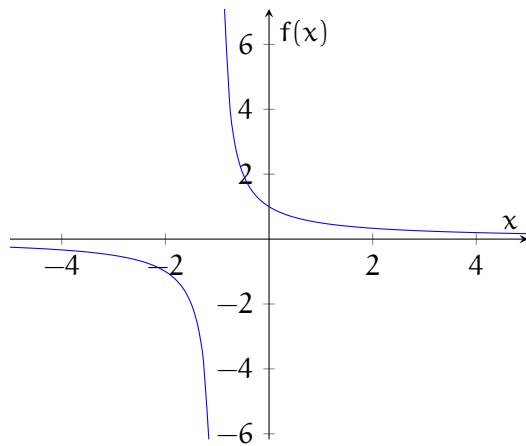


Figure 74.1: Graph of $f(x) = \frac{1}{x+1}$

We can see the y -intercept we found above. We can also see that the graph has no x -intercept, as expected. The phenomenon occurring at $x = -1$ is called a vertical asymptote. One other interesting feature of this graph is how it hugs the x -axis toward the left and right edges of the window. This makes the line $y = 0$ (the x -axis) a horizontal asymptote for this function. We can see why this is happening numerically by considering what happens for x -values far from 0. In this function, the result is a fraction with a numerator of 1 and a denominator that is large in size: a fraction that is close to 0.

x	-1000	-100	-10	10	100	1000
$f(x)$	-0.001	-0.01	-0.1	0.1	0.01	0.001

Let us examine another rational function. Begin by factoring to see if the function can be simplified.

$$g(x) = \frac{x^2 - 3x + 2}{x^2 - 4x + 3} = \frac{(x-1)(x-2)}{(x-1)(x-3)}$$

Consider the domain of g before continuing. Which values of x are valid inputs? Since substituting $x = 1$ or $x = 3$ would result in division by 0, these are not valid inputs. The domain of g is all real numbers except 1 and 3.

Now, for any x -value except 1, $\frac{x-1}{x-1} = 1$. This means that, for all x -values but 1, we can cancel those factors, leaving $g(x) = \frac{x-2}{x-3}$. (We will talk more about what is happening at $x = 1$ in a moment.)

This function has both x - and y -intercepts: y -intercept: $g(0) = \frac{0-2}{0-3} = \frac{2}{3}$. The graph contains the point $(0, \frac{2}{3})$. x -intercept: $g(x) = 0$ where the numerator equals 0 and the denominator does not equal 0. Since $x - 2 = 0$ when $x = 2$, the x -intercept is 2 and the graph contains the point $(2, 0)$.

The graph of g has a vertical asymptote at any x -value where substitution would result in

dividing a nonzero number by zero. Thus, g has a vertical asymptote at $x = 3$.

Does g have a horizontal asymptote? Let us see what happens when we substitute x -values far from 0.

x	-1000	-100	-10	10	100	1000
$g(x)$	0.999	0.990	0.923	1.143	1.010	1.001

As we move further away from the y -axis, the y -values become closer to 1. The horizontal asymptote describes the end behavior of the function, or what the graph looks like far from the y -axis. In this case, if we ignore the portion close to the y -axis, the graph begins to look like the line $y = 1$, making this the horizontal asymptote of g .

So, what is happening at $x = 1$? The value is not in the domain of the function, but there is no vertical asymptote there. That is because substituting any other value for x , even values very close to 1, into $\frac{(x-1)(x-2)}{(x-1)(x-3)}$ gives the exact same number as substituting into $\frac{x-2}{x-3}$. So, there is a hole in the graph at $x = 1$, but nothing strange is happening on either side of 1. (Depending on the graphing software, the hole may not be visible.)

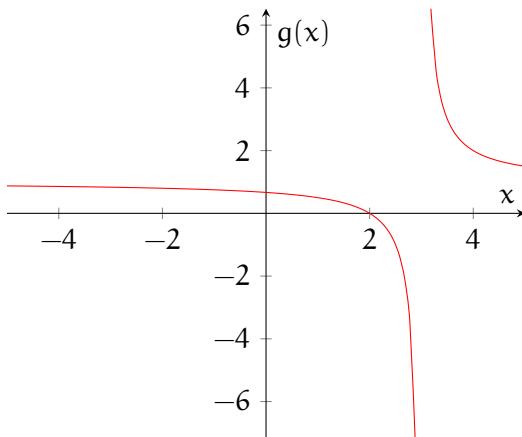


Figure 74.2: Graph of $g(x) = \frac{x^2 - 3x + 2}{x^2 - 4x + 3}$

Exercise 95 Rational Functions Practice 1

Determine the x- and y-intercepts and horizontal and vertical asymptotes of the rational function:

1. $\frac{2x+5}{x+4}$

Working Space

Answer on Page 838

In those examples, common factors cancel, leaving one polynomial. Of course, there is no guarantee that any two polynomials will have common factors, or even be factorable at all. Now, we consider an example which cannot be simplified. We will focus on just the asymptotes here.

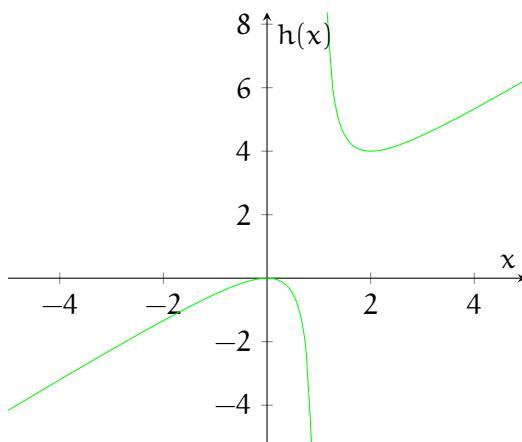
$$h(x) = \frac{x^2}{x - 1}$$

We see that the x-value 1 gives division of a non-zero number by zero, giving a vertical asymptote at $x = 1$. How about a horizontal asymptote? We examine values of h for values of x far from 0.

x	-1000	-100	-10	10	100	1000
$h(x)$	-999	-99	-9	11	101	1001

Rather than seeing function values leveling off as in the previous examples, we see function values that grow in size along with x . The function h has no horizontal asymptote. Let us examine the graph:

This function exhibits a different type of end behavior: that of a line with slope 1. To see that, cover up the portion of the graph near the y-axis and focus on the left and right. The rather dull and time-consuming technique of polynomial long division can be used to rewrite the function as a quotient and a remainder. Feel free to watch the Khan Academy video on the topic, but let us instead use our knowledge of factoring techniques and a clever little trick.

Figure 74.3: Graph of $h(x) = \frac{x^2}{x-1}$

$$\begin{aligned}
 h(x) &= \frac{x^2}{x-1} \\
 &= \frac{x^2 - 1 + 1}{x-1} \\
 &= \frac{x^2 - 1}{x-1} + \frac{1}{x-1} \\
 &= \frac{(x-1)(x+1)}{x-1} + \frac{1}{x-1} \\
 &= x+1 + \frac{1}{x-1}
 \end{aligned} \tag{74.3}$$

We obtain a quotient of $x+1$ and a remainder of 1. It is the quotient which determines the end behavior of the graph. Why? Substituting x -values far from zero makes the remainder term very small, since it becomes a fraction with a large denominator but a numerator of only 1. So for x -values far from zero, the y -value is $x+1$ plus a very small number, so small that we can justifiably ignore it. This means that far from the y -axis, the function acts like the quotient: the line $y = x+1$. We call this line an oblique asymptote. See below how the graph of $h(x)$ hugs that line.

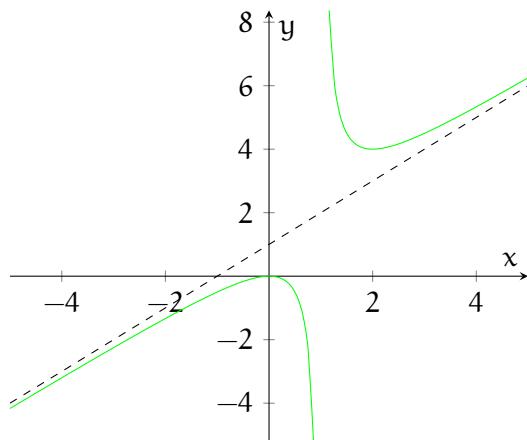


Figure 74.4: Graph of $h(x) = \frac{x^2}{x-1}$ and its oblique asymptote $y = x + 1$

Exercise 96 Rational Functions Practice 2

Factor and simplify the rational function, then determine any holes and vertical and oblique asymptotes of the rational function.

1. $\frac{x^3+2x^2}{x^2+x}$

Working Space

Answer on Page 838

We have seen lines act as end behaviors. Are there other possibilities? Sure! Here is an example with parabolic end behavior.

$$k(x) = \frac{x^3}{x-2}$$

We use our add-subtract trick to reveal the quotient, which describes the end behavior.

$$\begin{aligned}
 h(x) &= \frac{x^3}{x-2} \\
 &= \frac{x^3 - 8 + 8}{x-2} \\
 &= \frac{x^3 - 8}{x-2} + \frac{8}{x-2} \\
 &= \frac{(x-2)(x^2 + 2x + 4)}{x-2} + \frac{8}{x-2} \\
 &= x^2 + 2x + 4 + \frac{8}{x-2}
 \end{aligned} \tag{74.4}$$

The quotient, $x^2 + 2x + 4$, should describe the end behavior. We confirm by graphing both k and the quotient - the parabolic asymptote.

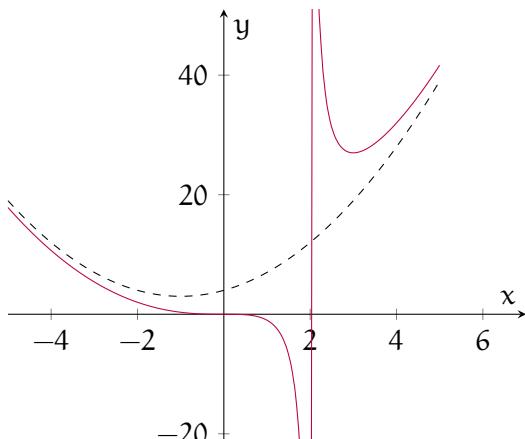


Figure 74.5: Graph of $k(x) = \frac{x^3}{x-2}$ and its parabolic asymptote $y = x^2 + 2x + 4$



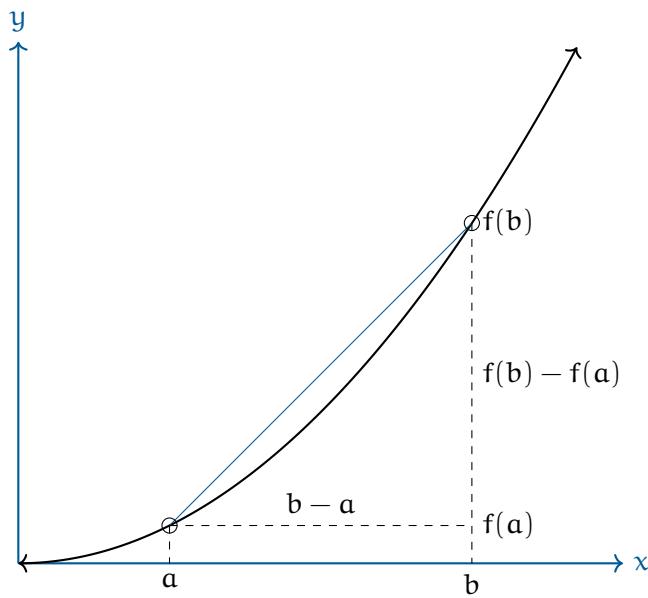
CHAPTER 75

Differentiation

We have done some differentiation, but you haven't been given the real definition because it is based on limits.

The idea is that we can find the slope between two points on the graph a and b like this:

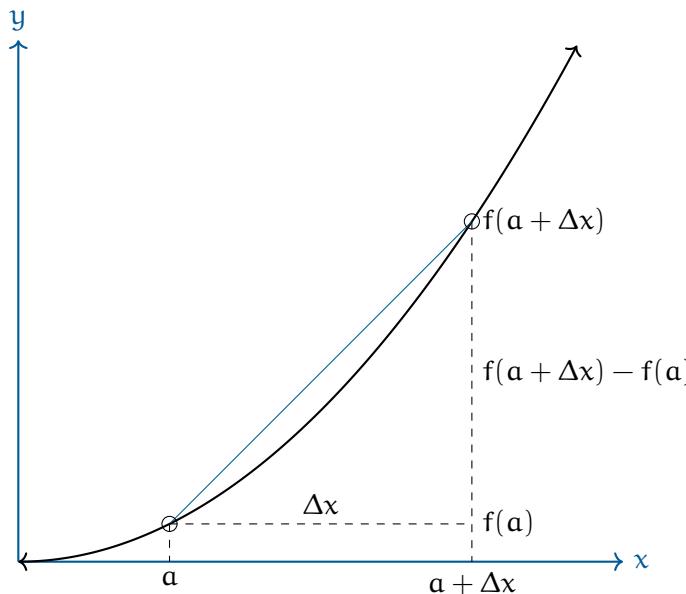
$$m = \frac{f(b) - f(a)}{b - a}$$



If we want to find the slope at a we take the limit of this as the b goes to a :

$$f'(a) = \lim_{b \rightarrow a} \frac{f(b) - f(a)}{b - a}$$

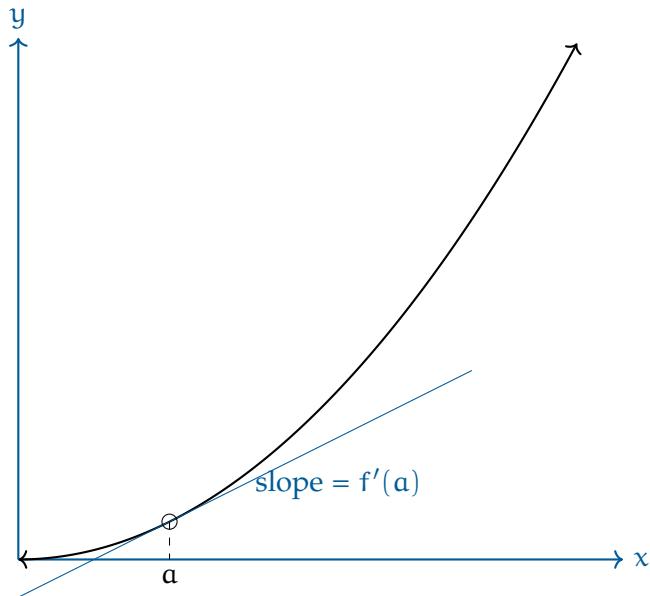
This idea is usually expressed using Δx as the difference between b and a :



Then the formula becomes:

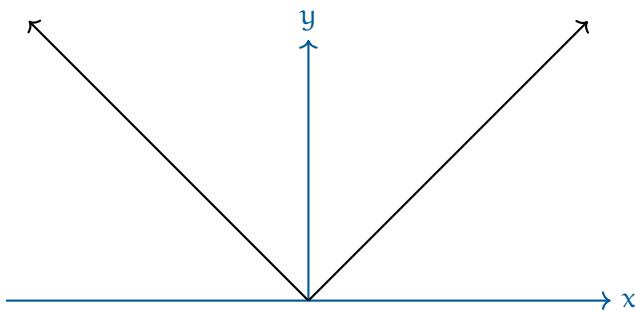
$$f'(a) = \lim_{\Delta x \rightarrow 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}$$

Now, at any point a we can compute the slope of the line tangent to the function at a :



75.1 Differentiability

Warning: Not every function is differentiable everywhere. For example, if $f(x) = |x|$, you get a corner at zero.



To the left of zero, the slope is -1 . To the right of zero, the slope is 1 . At zero? The derivative is not defined.

If a function has a derivative everywhere, it is said to be *differentiable*. Generally, you can think of differentiable functions as smooth – their graphs have no corners.

75.2 Using the definition of derivative

Let's say that you want to know the slope of $f(x) = -3x^2$ at $x = 2$. Using the definition of the derivative, that would be:

$$f'(2) = \lim_{\Delta x \rightarrow 0} \frac{f(2 + \Delta x) - f(2)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{-3(2 + \Delta x)^2 - (-3(2)^2)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{-12 - 12\Delta x + -3(\Delta x)^2 + 12}{\Delta x} = -12$$



CHAPTER 76

Derivatives

In calculus, the derivative of a function represents the rate at which the function is changing at a particular point. It is a fundamental concept that has vast applications in various fields, including physics.

76.1 Definition

The derivative of a function $f(x)$ at a point x is defined as the limit:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h} \quad (76.1)$$

provided this limit exists. In words, the derivative of f at x is the limit of the rate of change of f at x as the change in x approaches zero.

76.2 Applications in Physics

In physics, derivatives play a vital role in describing how quantities change with respect to one another.

76.2.1 Velocity and Acceleration

In kinematics, the derivative of the position function with respect to time gives the velocity function, and further taking the derivative of the velocity function gives the acceleration function. For example, if $s(t)$ represents the position of an object at time t , then the velocity $v(t)$ and acceleration $a(t)$ are given by:

$$v(t) = \frac{ds}{dt} \quad \text{and} \quad a(t) = \frac{dv}{dt} = \frac{d^2s}{dt^2} \quad (76.2)$$

76.2.2 Force and Momentum

In mechanics, the derivative of the momentum of an object with respect to time gives the net force acting on the object, as stated by Newton's second law of motion:

$$F = \frac{dp}{dt} \quad (76.3)$$

where F is the force, p is the momentum, and t is the time.



CHAPTER 77

Rules for Finding Derivatives

Derivatives play a key role in calculus, providing us with a means of calculating rates of change and the slopes of curves. Here, we present some common rules used to calculate derivatives.

77.1 Constant Rule

The derivative of a constant is zero. If c is a constant and x is a variable, then:

$$\frac{d}{dx}c = 0 \tag{77.1}$$

77.2 Power Rule

For any real number n , the derivative of x^n is:

$$\frac{d}{dx}x^n = nx^{n-1} \quad (77.2)$$

77.3 Product Rule

The derivative of the product of two functions is:

$$\frac{d}{dx}(fg) = f'g + fg' \quad (77.3)$$

where f' and g' denote the derivatives of f and g , respectively.

77.4 Quotient Rule

The derivative of the quotient of two functions is:

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{f'g - fg'}{g^2} \quad (77.4)$$

77.5 Chain Rule

The derivative of a composition of functions is:

$$\frac{d}{dx}(f(g(x))) = f'(g(x)) \cdot g'(x) \quad (77.5)$$

77.6 Conclusion

These rules form the basis for calculating derivatives in calculus. Many more complex rules and techniques are built upon these fundamental rules.



CHAPTER 78

Optimization

Optimization is a branch of mathematics that involves finding the best solution from all feasible solutions. In the field of operations research, optimization plays a crucial role. Whether it is minimizing costs, maximizing profits, or reducing the time taken to perform a task, optimization techniques are employed to make decisions effectively and efficiently.

78.1 Optimization Problems

An optimization problem consists of maximizing or minimizing a real function by systematically choosing the values of real or integer variables from within an allowed set. This function is known as the objective function.

A standard form of an optimization problem is:

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad g_i(x) \leq 0, ; i = 1, \dots, m \quad h_j(x) = 0, ; j = 1, \dots, p$$

where

- $f(x)$ is the objective function,
- $g_i(x) \leq 0$ are the inequality constraints,
- $h_j(x) = 0$ are the equality constraints.

78.2 Types of Optimization Problems

There are different types of optimization problems, including but not limited to:

- **Linear Programming:** The objective function and the constraints are all linear.
- **Integer Programming:** The solution space is restricted to integer values.
- **Nonlinear Programming:** The objective function and/or the constraints are nonlinear.
- **Stochastic Programming:** The objective function and/or constraints involve random variables.

These problems are solved using different techniques and algorithms, many of which are a subject of active research.

78.3 Applications

Optimization techniques have a wide variety of applications in many fields such as economics, engineering, transportation, and scheduling problems.



CHAPTER 79

Implicit Differentiation

Implicit differentiation is a technique in calculus for finding the derivative of a relation defined implicitly, that is, a relation between variables x and y that is not explicitly solved for one variable in terms of the other.

79.1 Implicit Differentiation Procedure

Consider an equation that defines a relationship between x and y :

$$F(x, y) = 0$$

To find the derivative of y with respect to x , we differentiate both sides of this equation with respect to x , treating y as an implicit function of x :

$$\frac{d}{dx} F(x, y) = \frac{d}{dx} 0$$

Applying the chain rule during the differentiation on the left side of the equation gives:

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial y} \frac{dy}{dx} = 0$$

Finally, we solve for $\frac{dy}{dx}$ to find the derivative of y with respect to x :

$$\frac{dy}{dx} = -\frac{\frac{\partial F}{\partial x}}{\frac{\partial F}{\partial y}}$$

This result is obtained using the implicit differentiation method.

79.2 Example

Consider the equation of a circle with radius r :

$$x^2 + y^2 = r^2$$

Differentiating both sides with respect to x , we get:

$$2x + 2y \frac{dy}{dx} = 0$$

Solving for $\frac{dy}{dx}$ gives:

$$\frac{dy}{dx} = -\frac{x}{y}$$

which is the slope of the tangent line to the circle at any point (x, y) .



CHAPTER 80

Related Rates

In calculus, related rates problems involve finding a rate at which a quantity changes by relating that quantity to other quantities whose rates of change are known. The technique used to solve these problems is known as "related rates" because one rate is related to another rate.

80.1 Steps to solve related rates problems

80.1.1 Step 1: Understand the problem

First, read the problem carefully. Understand what rates are given and what rate you need to find.

80.1.2 Step 2: Draw a diagram

For most problems, especially geometry problems, drawing a diagram can be very helpful.

80.1.3 Step 3: Write down what you know

Write down the rates that you know and the rate that you need to find.

80.1.4 Step 4: Write an equation

Write an equation that relates the quantities in the problem. This equation will be your main tool to solve the problem.

80.1.5 Step 5: Differentiate both sides of the equation

Now you can use calculus. Differentiate both sides of the equation with respect to time.

80.1.6 Step 6: Substitute the known rates and solve for the unknown

Now that you have an equation that relates the rates, substitute the known rates into the equation and solve for the unknown rate.

80.2 Example

Here is an example of a related rates problem:

A balloon is rising at a constant rate of 5 m/s. A boy is cycling towards the balloon along a straight path at 15 m/s. If the balloon is 100 m above the ground, find the rate at which the distance from the boy to the balloon is changing when the boy is 40 m from the point on the ground directly beneath the balloon.

The problem can be modeled with a right triangle where the vertical side is the height of the balloon, the horizontal side is the distance of the boy from the point on the ground directly beneath the balloon, and the hypotenuse is the distance from the boy to the balloon.

Let x be the distance of the boy from the point on the ground directly beneath the balloon, y the height of the balloon above the ground, and z the distance from the boy to the balloon. From the Pythagorean theorem, we have

$$z^2 = x^2 + y^2 \quad (80.1)$$

Differentiating both sides with respect to time t gives

$$2z \frac{dz}{dt} = 2x \frac{dx}{dt} + 2y \frac{dy}{dt} \quad (80.2)$$

Given that $\frac{dx}{dt} = -15$ m/s (the boy is moving towards the point beneath the balloon), $\frac{dy}{dt} = 5$ m/s (the balloon is rising), $x = 40$ m, $y = 100$ m, we can substitute these into the equation and solve for $\frac{dz}{dt}$.



CHAPTER 81

Multivariate Functions

A real-valued multivariate function is a function that takes multiple real variables as input and produces a single real output.

We generally denote such a function as $f : \mathbb{R}^n \rightarrow \mathbb{R}$, where \mathbb{R}^n is the domain and \mathbb{R} is the co-domain.

For example, consider a function f that takes two variables x and y :

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

Here, $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ takes an ordered pair (x, y) from the 2-dimensional real coordinate space, squares each, and adds them to produce a real number.

In a similar way, a function $g : \mathbb{R}^3 \rightarrow \mathbb{R}$ could take three variables x , y , and z , and might be defined as:

$$g(x, y, z) = x^2 + y^2 + z^2$$

Here, the function squares each of the input variables and then adds them to produce a real number.

These functions are "real-valued" because their outputs are real numbers, and "multivariate" because they take multiple variables as inputs.

The concepts of limits, continuity, differentiability, and integrability can all be extended to multivariate functions, although they become more complex because we now have to consider different directions in which we approach a point, not just from the left or right as in the univariate case. For example, the partial derivative is the derivative of the function with respect to one variable, holding the others constant. It is one of the basic concepts in the calculus of multivariate functions.

For example, given the function $f(x, y) = x^2 + y^2$, the partial derivatives of f are computed as:

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

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



CHAPTER 82

Partial Derivatives and Gradients

This chapter will introduce you to partial derivatives and gradients, equipping you with the tools to study functions of multiple variables. We will explore how these concepts provide valuable insights into optimization, vector calculus, and various fields of science and engineering.

Partial derivatives come into play when dealing with functions that depend on multiple variables. Unlike ordinary derivatives that consider changes along a single variable, partial derivatives focus on how a function changes concerning each individual variable while holding the others constant. In essence, partial derivatives measure the rate of change of a function with respect to one variable while keeping the other variables fixed.

The notation for a partial derivative of a function $f(x, y, \dots)$ with respect to a specific variable, say x , is denoted as $\frac{\partial f}{\partial x}$. Similarly, $\frac{\partial f}{\partial y}$ represents the partial derivative with respect to y , and so on. It is essential to remember that when taking partial derivatives, we treat the other variables as constants during the differentiation process.

The gradient is a vector that combines the partial derivatives of a function. It provides a

concise representation of the direction and magnitude of the steepest ascent or descent of the function. The gradient vector points in the direction of the greatest rate of increase of the function. By understanding the gradient, we gain insights into optimizing functions and finding critical points where the function reaches maximum or minimum values.

Throughout this chapter, we will explore the following key topics related to partial derivatives and gradients:

- Calculating partial derivatives: We will delve into the techniques and rules for computing partial derivatives of various functions, including polynomials, exponential functions, and trigonometric functions. We will also explore higher-order partial derivatives and mixed partial derivatives.
- Interpreting partial derivatives: Understanding the geometric and physical interpretations of partial derivatives is essential. We will discuss the notion of tangent planes, directional derivatives, and the relationship between partial derivatives and local linearity.
- Gradient vectors and their properties: We will introduce the gradient vector and its properties, such as its connection to the direction of steepest ascent, its relationship with partial derivatives, and how it relates to level curves and level surfaces.
- Applications of partial derivatives and gradients: We will explore various applications of these concepts, including optimization problems, constrained optimization, tangent planes, linear approximations, and their relevance in fields like physics, economics, and engineering.

By grasping the concepts of partial derivatives and gradients, you will unlock a powerful mathematical framework for analyzing and optimizing functions of multiple variables. These tools will equip you to tackle advanced calculus problems and gain deeper insights into the behavior of functions in diverse fields.



CHAPTER 83

Vectors and Matrices

Linear algebra is a specialized form of algebra that can represent and manipulate sets of variables that are linearly related to one another. One of the basic operations is the multiplication of a matrix by a vector. As you work through this module, you'll see that you already know the fundamentals (vectors, scalars, dot products) and how to apply these concepts to practical problems. You've also had some experience with matrices in the form of spreadsheets.

Matrices can represent:

- A linear transformation, such as rotation, scaling, and skewing. You apply a transformation to a vector by multiplying the vector by a matrix.
- A system of linear equations. Linear algebra provides various methods that you can use to find the solution vector.

83.1 Applications of Matrix-Vector Multiplication

Many areas in engineering and science rely on matrix-vector multiplication. These are just a few examples. As you encounter more topics in science and engineering, you will find that matrix-vector multiplication is crucial to many other fields.

83.1.1 Computer Graphics

When you play a video game or watch the latest CG animation, matrix operations transform objects in the scene to make them appear as if moving, getting closer, and so on. You can represent the vertices of objects as vectors, and then apply a transformation matrix.

83.1.2 Data Analysis

We live in an era in which it's easy to collect so much data that it's difficult to make sense of the data by just looking at it. You can represent the data in matrix form and then find a solution vector. For example, scientists use this technique to figure out the effectiveness of drug treatments on disease.

83.1.3 Economics

Take a look at financial section of any news organization and you'll see headlines such as "Economic Data Points to Faster Growth" or "Is the Inflation Battle Won?" Economists can use systems of linear equations to represent economic indicators, such as consumer consumption, government spending, investment rate, and gross national product. By using various methods that you'll learn about later, they can get a good idea of the state of the economy.

83.1.4 Engineering

Engineering couldn't do without vector-matrix multiplication. For example, the orbital dynamics of space travel relies on it. Engineers must predict and calculate the motion of planetary bodies, satellites, and spacecraft. By solving systems of linear equations engineers can make sure that a spacecraft travels to its destination without running into a satellite or space rock.

83.1.5 Image Processing

An image is a matrix of pixel values. When you take a selfie and apply a filter, the image app applies a transformation to the image matrix. A simple operation would be to change the tint of the image. A more complex operation would be to skew the image to make it distorted, like a funhouse mirror.

83.2 Vector-Matrix Multiplication

Let's take a look at the general form of vector-matrix multiplication. Given a matrix A of size $m \times n$ and a vector x of size $n \times 1$, the product Ax is a new vector of size $m \times 1$.

You compute the i -th component of the product vector Av by taking the dot product of the i -th row of A and the vector v :

$$(Av)_i = \sum_{j=1}^n a_{ij}v_j$$

where a_{ij} is the element in the i -th row and j -th column of A , and v_j is the j -th element of v .

This is the general form of a matrix and a vector, written to show the specific components of each:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}$$

$$v = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \dots \\ v_m \end{bmatrix}$$

$$Av = \begin{bmatrix} v_1 * a_{11} + v_2 * a_{12} + v_3 * a_{13} + \dots + v_m * a_{1n} \\ v_1 * a_{21} + v_2 * a_{22} + v_3 * a_{23} + \dots + v_m * a_{2n} \\ \dots \\ v_1 * a_{m1} + v_2 * a_{m2} + v_3 * a_{m3} + \dots + v_m * a_{mn} \end{bmatrix}$$

Let's look at a specific example.

$$A = \begin{bmatrix} 2 & 4 & 6 \\ 3 & 5 & 7 \\ 1 & 2 & 3 \\ 8 & 6 & 2 \end{bmatrix}$$

$$v = \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix}$$

Solution:

$$= \begin{bmatrix} -2 * 2 + 1 * 4 + 3 * 6 \\ -2 * 3 + 1 * 5 + 3 * 7 \\ -2 * 1 + 1 * 2 + 3 * 3 \\ -2 * 8 + 1 * 6 + 3 * 2 \end{bmatrix}$$

$$= \begin{bmatrix} 18 \\ 20 \\ 9 \\ -4 \end{bmatrix}$$

$$= (18, 20, 9, -4)$$

Exercise 97 Vector Matrix Multiplication

Multiply the array A with the vector v. Compute this by hand, and make sure to show your computations.

Working Space

$$A = \begin{bmatrix} 1 & -2 & 3 & 5 \\ -4 & 2 & 7 & 1 \\ 3 & 3 & -9 & 1 \end{bmatrix}$$

$$v = \begin{bmatrix} 2 \\ 2 \\ 6 \\ -1 \end{bmatrix}$$

Answer on Page 838

83.2.1 Vector-Matrix Multiplication in Python

Most real-world problems use very large matrices where it becomes impractical to perform calculations by hand. As long as you understand how matrix-vector multiplication is done, you'll be equipped to use a computing language, like Python, to do the calculations for you.

Create a file called `vectors_matrices.py` and enter this code:

```
// import the python module that supports matrices
import numpy as np

// create an array
a = np.array([[ 5,  1 ,3, -2],
              [ 1, -1 ,8,  4],
              [ 6,  2 ,1,  3]])

// create a vector
b = np.array([1,  2,  3,-8])

//calculate the dot product
print(a.dot(b))
```

When you run it, you should see:

```
[16  6  8]
```

83.3 Where to Learn More

Watch this video from Khan Academy about matrix-vector products: <https://rb.gy/frga5>



CHAPTER 84

Linear Combinations

A linear combination of vectors is the addition of two or more scaled vectors. For example, given two vectors, v_1, v_2 and two scalars a_1, a_2 , you'd write their linear combination as:

$$x\mathbf{w} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2$$

The scalars can be any real number. The vectors can be of any dimension.

Let's take a more generalized approach. Given vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n \in \mathbb{R}^m$ and scalars $a_1, a_2, \dots, a_n \in \mathbb{R}$, a linear combination of these vectors is any vector of the form

$$\mathbf{w} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \dots + a_n\mathbf{v}_n$$

Each scalar a_i scales the corresponding vector \mathbf{v}_i , and added together, the results are produce a new vector \mathbf{w} .

Let's look at an example that has 4 vectors and their scalars.

$$a_1 = 1, v_1 = [9, 1, 2]$$

$$a_2 = -1, v_2 = [8, -3, 4]$$

$$a_3 = 3, v_3 = [6, 0, 1]$$

$$a_4 = -4, v_4 = [3, 7, 2]$$

As a linear combination:

$$\mathbf{w} = 1 * [9, 1, 2] + (-1) * [8, -3, 4] + 3 * [6, 0, 1] + (-4) * [3, 7, 2]$$

After multiplying each vector by its associated scalar.

$$\mathbf{w} = [9, 1, 2] + [-8, 3, -4] + [18, 0, 3] + [-12, -28, -8]$$

When combined:

$$\mathbf{w} = [7, -24, -7]$$

Exercise 98 Linear Combination

Calculate the linear combination for vectors v_1, v_2, v_3 and scalars a_1, a_2, a_3 where:

Working Space

$$a_1 = 2, v_1 = [2, 4, 8]$$

$$a_2 = -2, v_2 = [8, -6, 3]$$

$$a_3 = 4, v_3 = [7, 9, 2]$$

Make sure to show all your work.

Answer on Page 838

84.1 Weighted Averages of Vectors

A weighted average of vectors is a specific type of linear combination where the coefficients (or weights) a_i are non-negative and sum to 1:

$$\sum_{i=1}^n a_i = 1, \quad a_i \geq 0$$

A weighted average of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is then defined as

$$\mathbf{w} = a_1 \mathbf{v}_1 + a_2 \mathbf{v}_2 + \dots + a_n \mathbf{v}_n$$

In this case, each a_i not only scales the corresponding vector \mathbf{v}_i , but also represents the proportion of that vector in the final average vector \mathbf{w} .

Weighted averages are useful when you want to attribute the contribution of one feature or item over another. For example, a teacher might figure a student's final grade using exam scores, class participation, and a final project. The exam scores might make up 65% of the final grade, class participation 10%, and a final project 25%. Thus giving the formula for a grade as:

$$\text{Grade} = .65 * \text{ExamScores} + .10 * \text{Participation} + .25 * \text{FinalProject}$$

The teacher defines the weights, making sure they sum to 1.0.

Let's look at an example where the weights don't sum to 1.0. A store that sells umbrellas might have to get the umbrella stock from three different manufacturers. The store owner buys 100 umbrellas at a cost of \$2.10 each, 50 umbrellas cost \$1.85 each, and 200 umbrellas cost \$2.00.

$$\text{TotalCost} = 2.10 * 100 + 1.85 * 50 + 2.00 * 200 = 702.5$$

To calculate the weighted average, divide the total cost by the number of items.

$$\text{WeightedAverage} = 702.5 / 350 = 2.01$$

Exercise 99 **Weighted Average**

A concert sells 300 tickets in the balcony at \$50 each, 100 tickets on the main floor at \$75 each, and 50 tickets in the section closest to the stage at \$150 each. What's the weighted average?

Working Space

Answer on Page 838

84.2 Weighted Averages of Vectors in Python

Create a file called `linearCombos.py` and enter this code:

```
// import the python module that supports matrices
import numpy as np

// an array for number of umbrellas by manufacturer
items = np.array([100, 50, 200])

// weights are the cost of item by manufacturer
weights = np.array([2.10, 1.85, 2.00])

// create an array for total cost for each manufacturer
costPerManufacturer=items * weights

// sum the individuals costs to get the total
totalCost = np.sum(costPerManufacturer)

// get number of items
numItems = np.sum(items)

// you are ready to calculate the weighted average
weightedAverage = totalCost/numItems
print(weightedAverage)
```

When you run this code, you should get a weighted average of \$2.01 when rounded to the nearest cent.



CHAPTER 85

Vector Spans and Independence

A vector span is the collection of vectors obtained by scaling and combining the original set of vectors in all possible proportions. Formally, if the set $S = \{v_1, v_2, \dots, v_n\}$ contains vectors from a vector space V , then the span of S is given by:

$$\text{Span}(S) = \{a_1v_1 + a_2v_2 + \dots + a_nv_n : a_1, a_2, \dots, a_n \in \mathbb{R}\} \quad (85.1)$$

This means that any vector in the $\text{Span}(S)$ can be written as a linear combination of the vectors in S .

Vector spans have practical applications in a number of fields. Computer graphics and physics are two of them. For example, in space travel, knowing the vector span is essential to calculating a slingshot maneuver that will give spacecraft a gravity boost from a planet. For this, you'd need to know the gravity vector of the planet relative to the sun and the velocity vectors that characterize the spacecraft. Engineers would use this information to figure out the trajectory angle that would allow the spacecraft to achieve a particular velocity in the desired direction.

85.1 Vector Independence

A set of vectors $S = \{v_1, v_2, \dots, v_n\}$ is linearly independent if the only solution to the equation $a_1v_1 + a_2v_2 + \dots + a_nv_n = 0$.

is $a_1 = a_2 = \dots = a_n = 0$. This means that no vector in the set can be written as a linear combination of the other vectors.

If there exists a nontrivial solution (i.e., a solution where some $a_i \neq 0$), then the vectors are said to be linearly dependent. This means that at least one vector in the set can be written as a linear combination of the other vectors.

The concept of vector independence is fundamental to the study of vector spaces, bases, and rank. You'll learn more about these concepts in future modules.

85.1.1 Dependent Vectors

Let's start by looking at two vectors.

$$\begin{aligned}v_1 &= \begin{bmatrix} 2 \\ 4 \end{bmatrix} \\v_2 &= \begin{bmatrix} -14 \\ -28 \end{bmatrix}\end{aligned}$$

These two vectors are dependent because $v_2 = -7 * v_1$. This is an obvious example but let's show it mathematically. If linearly independent, the two vectors must satisfy:

$$\begin{aligned}v_1a_1 + v_2a_2 &= 0 \\v_2a_1 + v_2a_2 &= 0\end{aligned}$$

which is:

$$\begin{aligned}2a_1 - 14a_2 &= 0 \\4a_1 - 28a_2 &= 0\end{aligned}$$

To solve, multiply the top equation by -2 and add it to the bottom:

$$\begin{aligned}2a_1 - 14a_2 &= 0 \\0 + 0 &= 0\end{aligned}$$

The bottom equation drops out. Now solve for a_1 in the remaining equation:

$$a_1 = -7a_2$$

As you can see, one vector is a multiple of another.

$$a_1 \neq a_2 \neq 0$$

85.1.2 Independent Vectors

Let's see if these two vectors are independent.

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

To be independent, the two vectors must satisfy:

$$v_1 a_1 + v_2 a_2 = 0$$

$$v_2 a_1 + v_1 a_2 = 0$$

which is:

$$\begin{bmatrix} a_1 + 0 * a_2 \\ 0 * a_1 + a_2 \end{bmatrix}$$

So: $a_1 = a_2 = 0$ These vectors are not only independent, but they are orthogonal (perpendicular) to one another. You'll learn more about orthogonality later.

Here is an example whose solution isn't as obvious. You can solve using Gaussian elimination.

$$v_1 = [2, 1]$$

$$v_2 = [1, -6]$$

Rewrite as a system of equations:

$$a_1 * 2 + a_2 * 1 = 0$$

$$a_1 * 1 + a_2 * (-6) = 0$$

First swap the equations so that the top equation has a coefficient of 1 for a_1 :

$$a_1 - 6a_2 = 0$$

$$2a_1 + a_2 = 0$$

Next multiply row 1 by -2 and add it to row 2:

$$a_1 - 6a_2 = 0$$

$$0 - 11a_2 = 0$$

Multiply row 2 by 1 divided by 11.

$$a_1 - 6a_2 = 0$$

$$0 + a_2 = 0$$

Back substitute a_2 solution into the first equation:

$$a_1 = 0$$

$$a_2 = 0$$

Therefore

$$a_1 = a_2 = 0$$

and the two vectors are linearly independent.

Exercise 100 Vector Independence

Are these vectors independent?

Working Space

[2, 1, 4]

[2, -1, 2]

[0, 1, -2]

Show your work.

Answer on Page 839

85.2 Checking for Linear Independence Using Python

One way to use python to check for linear independence is to use the `linalg.solve()` function to solve the system of equations. You need to create an array that contains the coefficients of the variable and a vector that contains the values on the right-side of each equation. So far, you've either been given equations that equal 0 or you've manipulated each equation to be equal to 0.

Let's first see how to use python to solve the equations in the previous exercise. If the equations are linearly independent, then $a_1 = a_2 = a_3 = 0$

Create a file called linear_independence.Python and enter this code:

```
import numpy as np

A = np.array([[ 2,  2,  0],
              [ 1, -1, 1],
              [4,  2, -2]])
b = np.array([0,0,0])
c = np.linalg.solve(A,b)
print(c)
```

You should get this result, which shows the equations are linearly independent.

```
[ 0. -0.  0.]
```

But what happens if the equations are not independent? Let's make the first two equations dependent by making equation 1 two times equation 2. Enter this code into your file:

```
import numpy as np

D = np.array([[ 2, -2,  2],
              [ 1, -1, 1],
              [4,  2, -2]])
e = np.array([0,0,0])
f = np.linalg.solve(D,e)
print(f)
```

You should get many lines indicating an error. Among the spew, you should see:

```
raise LinAlgError("Singular matrix")
```

So while the `linalg.solve()` function is quite useful for solving a system of independent linear equations, raising an error is not the most elegant way to figure out if the equations are dependent. That's where the concept of a determinant comes in. You'll learn about that in the next section. But for now, let's use the `linalg.solve()` function to find a solution for a set of equations known to be linearly independent.

$$4x_1 + 3x_2 - 5x_3 = 2$$

$$-2x_1 - 4x_2 - 5x_3 = 5$$

$$8x_2 + 8x_3 = -3$$

You will create a matrix that contains all the coefficients and a vector that contains the values on the right-side of the equations.

Enter this code into your file.

```
G = np.array([[4, 3, -5],  
             [-2, -4, 5],  
             [8, 8, 0]])  
h = np.array([2, 5, -3])  
  
j = np.linalg.solve(G, h)  
print(j)
```

You should get this answer:

```
[ 2.20833333 -2.58333333 -0.18333333]
```

85.3 Determinants

The determinant of a matrix is a scalar value that can be calculated for a square matrix. If a matrix has linearly dependent rows or columns, the determinant is 0. One way to figure out if a set of equations are independent is to calculate the determinant.

For a matrix that is 2 by 2, the calculation is easy:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the determinant is:

$$\det(A) = (a * d) - (b * c)$$

For a larger matrix, finding the determinant is more complex and requires breaking down the matrix into smaller matrices until you reach te 2x2 form. The process is called expansion by minors. For our purposes, we simply want to first check to see if a matrix contains linearly independent rows and columns before using our Python code to solve. Modify your code so that it uses the `np.linalg.det()` function. If the determinat is not zero, then you can call the `np.linalg.solve()` function. Your code should look like this:

```
if (np.linalg.det(D) != 0):  
    j = np.linalg.solve(D,e)  
    print(j)  
else:  
    print("Rows and columns are not independent.")
```

85.4 Where to Learn More

Watch this video on *Linear Combinations and Vector Spans* from Khan Academy: <http://rb.gy/g1snk>

If you curious about the *Expansion of Minors*, see: <https://mathworld.wolfram.com/DeterminantExpansionbyMinors.html>



CHAPTER 86

Matrices

You've already had experience with matrices earlier in this module and also when you've used spreadsheets. In this chapter you'll learn the types of matrices and get an introduction to some of the special matrices used for various calculations.

As you know, a matrix is a rectangular array of numbers arranged in rows and columns. The individual numbers in the matrix are called elements or entries. Matrices can be described by their dimensions. For example, a matrix with 2 rows and 3 columns is a 2 by 3 matrix.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

More generally, a matrix with m rows and n columns is referred to as an $m \times n$ matrix or simply an m -by- n matrix, and m and n are its dimensions.

The general form of a 2×3 matrix A is:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

86.1 Types of Matrices

Matrices can be described by their shape:

- **Row Matrix:** has only one row.
- **Column Matrix:** has only one column.
- **Square Matrix:** has the same number of rows and columns.
- **Rectangular Matrix:** has an unequal number of rows and columns.

They can also be described by their unique numerical properties. Special matrices that come in handy for certain types of computations. These are a few of the most common special matrices.

- **Zero Matrix:** contains only entries that are zero.
- **Identity Matrix:** sometimes called the unit matrix, is a square matrix whose diagonal entries are 1 and all other entries are 0.
- **Symmetric Matrix:** a square matrix that equals its transpose.
- **Diagonal Matrix:** has nonzero elements on the main diagonal, but all other elements are zero
- **Triangular Matrix:** This is a special square matrix that can be upper triangular or lower triangular. If upper, the main diagonal and all entries above it are nonzero while the lower entries are all zero. If lower, the main diagonal and all the entries below it are nonzero while the upper entries are all zero.

86.1.1 Symmetric Matrices

If you want to find out if a square matrix is symmetric, you need to transpose it. If the transpose is equal to the original matrix, then the matrix is symmetric.

To transpose a matrix, flip it over its diagonal so that the rows and columns are switched, like this:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

After transposing:

$$A^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

Note that A^T means the transpose of A .

Let's see how this works for the following square matrix, A .

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

Switch the rows and columns to get the transpose:

$$A^T = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

Notice that $A = A^T$, so the matrix is symmetric.

What about matrix B ?

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{bmatrix}$$

Switch the rows and columns to get the transpose:

$$B^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

Note that $B \neq B^T$. So B is not symmetrical.

Exercise 101 Matrix Transposition

Find the transpose of this matrix. Is it
symmetric?

Working Space

$$A = \begin{bmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{bmatrix}$$

Answer on Page 840

86.1.2 Creating Matrices in Python

Create a file called `matrices_creation.py` and enter this code:

```
// import the python module that supports matrices
import numpy as np
// Use the function np.array to define a matrix that
// contains specific values that you supply.
A = np.array([[ 5,  1 ,3],
              [ 1, -1 ,8],
              [ 6,  2 ,1]])
// The transpose function returns
A.transpose()
```

When you run it, you should see:

```
array([[ 5,  1 ,6],
       [ 1, -1 ,2],
       [ 3,  8 ,1]])
```

As you can see, $A \neq A^T$ so A is not symmetric. Try another:

```
// create a matrix, B
B = np.array([[ 5,  1 ,6],
              [ 1, -1 ,2],
              [ 6,  2 ,1]])
B.transpose()
```

When you run it, you should see:

```
array([[ 5,  1,  6],
       [ 1, -1,  2],
       [ 6,  2,  1]])
```

B is symmetric. You can actually transpose any matrix using this function. But a matrix cannot be symmetric unless it is square.

Try this code to see what happens when you transpose a rectangular matrix.

```
// create a matrix, B
J = np.array([[ 5,  1 ,3,  0],
              [ 1, -1 ,8, 11],
              [ 6,  2 ,1,-7]])
J.transpose()
```

Note that transposing a rectangular matrix changes its dimension from 3 by 4 to 4 by 3. You should see a transposed matrix, but it's not symmetric.

```
array([[ 5,  1,  6],  
       [ 1, -1,  2],  
       [ 3,  8,  1],  
       [ 0, 11, -7]])
```

86.1.3 Creating Special Matrices in Python

Use the same file to add this code for creating a zero matrix.

```
// create an 8 by 10 Zero matrix.  
C = np.zeros((8, 10))  
C
```

When you run it, you should see:

```
array([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```

Add the following code to create an 8 by 8 Identity matrix.

```
// create an 8 by 8 Identity matrix  
D = np.eye(8)  
D
```

When you run it, you should see:

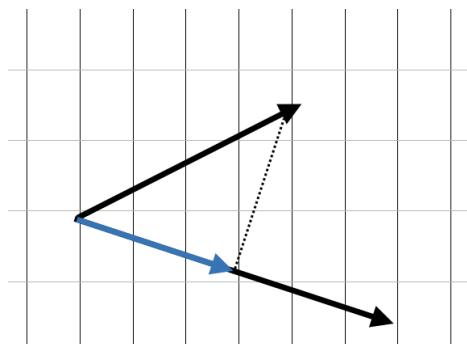
```
array([[1., 0., 0., 0., 0., 0., 0., 0.],  
       [0., 1., 0., 0., 0., 0., 0., 0.],  
       [0., 0., 1., 0., 0., 0., 0., 0.],  
       [0., 0., 0., 1., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 1., 0., 0., 0.],  
       [0., 0., 0., 0., 0., 1., 0., 0.],  
       [0., 0., 0., 0., 0., 0., 1., 0.],  
       [0., 0., 0., 0., 0., 0., 0., 1.]])
```




CHAPTER 87

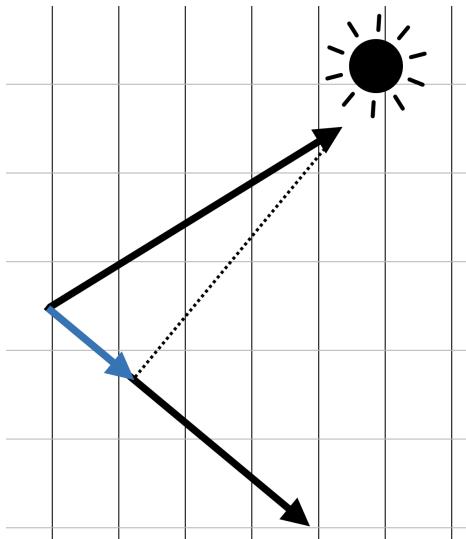
Projections

A projection describes the relationship of one vector to another in terms of direction and orthogonality. Given two vectors, \mathbf{u} and \mathbf{v} , the projection of \mathbf{u} onto \mathbf{v} separates \mathbf{u} into two components. The first component signifies how much \mathbf{u} lies in the direction of \mathbf{v} . The second signifies the component of \mathbf{u} that is orthogonal (perpendicular) to \mathbf{v} . The figure depicts a projection. The perpendicular line dropped from the end of \mathbf{u} is the orthogonal component. The portion of \mathbf{u} that lies in the direction of \mathbf{v} is the blue segment.

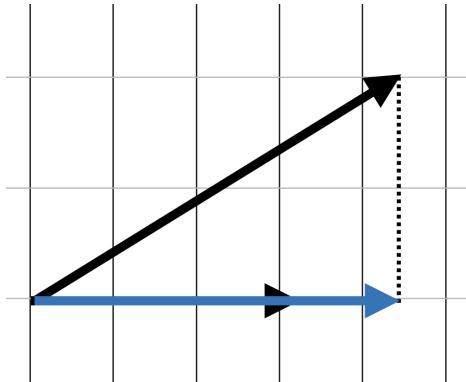


You can also think of a projection as the shadow cast by one vector onto each other by an

overhead light.



The projected vector can be in any direction. The length of the projected vector can extend beyond the vector on which it is projecting.



Projections are useful in many fields. These are a few examples, but there are numerous other applications in science, math, engineering, and finance.

- Investors evaluate risk and return of a portfolio by projecting an asset's return onto a reference portfolio.
- Astronomers analyze the motion of stellar objects by projecting the object's true motion onto the plane of the sky.
- Robotics engineers use projections to prevent robots from running into obstacles by projecting the robot's position onto the optimal path.

As you work your way through this course, you'll have a chance to apply the calculations

you learn in this chapter to a variety of problems. Specifically, the next chapter shows how to transform a set of linearly independent vectors into a set of orthogonal ones. Projections are essential to that transformation.

To calculate the projection of \mathbf{v} onto \mathbf{u} , use this formula:

$$\mathbf{proj}_{\mathbf{v}}(\mathbf{u}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v}$$

Note that the denominator is the magnitude squared of vector \mathbf{v} .

$$(\sqrt{a_1^2 + a_2^2 + \dots + a_n^2})^2$$

You learned previously that this is the same as the dot product of a vector with itself.

$$\mathbf{v} \cdot \mathbf{v}$$

In the examples that follow, we'll simplify to the dot product notation.

Let's look at a specific example:

$$\mathbf{u} = (1, 4, 6)$$

$$\mathbf{v} = (-2, 6, 2)$$

$$\begin{aligned}\mathbf{proj}_{\mathbf{v}}(\mathbf{u}) &= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|^2} \mathbf{v} \\ \mathbf{proj}_{\mathbf{v}}(\mathbf{u}) &= \frac{(1, 4, 6) \cdot (-2, 6, 2)}{(-2, 6, 2) \cdot (-2, 6, 2)} (-2, 6, 2) \\ \mathbf{proj}_{\mathbf{v}}(\mathbf{u}) &= \left(\frac{34}{44} \right) (-2, 6, 2) \\ \mathbf{proj}_{\mathbf{v}}(\mathbf{u}) &= (-1.545, 4.64, 1.545)\end{aligned}$$

Exercise 102 Projections

Find the projection of \mathbf{a} on \mathbf{b} where:

Working Space

$$\mathbf{a} = (1, 3)$$

$$\mathbf{b} = (-4, 6)$$

Answer on Page 840

87.1 Projections in Python

Create a file called `vectors_projections.py` and enter this code:

```
import numpy as np

a = np.array([1, 4, 6])    # vector a
b = np.array([-2, 6, 2])   # vector b

# Use np.dot() to calculate the dot product
projection_a_on_b = (np.dot(a, b)/np.dot(b, b))*b

print("The projection of vector a on vector b is:", projection_a_on_b)
```

87.2 Where to Learn More

Watch this Introduction to Projections from Khan Academy <https://rb.gy/yf0i3>



CHAPTER 88

The Gram-Schmidt Process

The Gram-Schmidt process is a method used to transform a set of linearly independent vectors to a set of orthogonal (perpendicular) vectors. The original vectors and the transformed vectors span the same subspace.

The process was named after two mathematicians: Jørgen Pedersen Gram, a Danish actuary mathematician, and Erhard Schmidt, a German mathematician. The men developed the orthogonalization process independently. Gram introduced the process in 1883 whereas Schmidt did his work in 1907. It wasn't named the Gram-Schmidt process until sometime later, after both mathematicians became well-known in the mathematical community.

This method has many practical applications in science and engineering. These are just two applications of Gram-Schmidt:

1. In signal processing, it can represent an audio signal with fewer components making it easier to isolate and remove noise.
2. In statistics and data analysis, it can reduce the complexity of a dataset so that it is easier to see which aspects or features contribute to the analysis.

The Gram-Schmidt process orthonormalizes a set of vectors in an inner product space, most commonly the Euclidean space \mathbb{R}^n . The process takes a finite, linearly independent set $S = \{v_1, v_2, \dots, v_k\}$ for $k \leq n$, and generates an orthogonal set $S' = \{u_1, u_2, \dots, u_k\}$ that spans the same k -dimensional subspace of \mathbb{R}^n as S .

88.1 The Process

Let's look at how the process works. Given a set of vectors $S = \{v_1, v_2, \dots, v_k\}$, the Gram-Schmidt process is as follows:

1. Let $u_1 = v_1$.
2. For $j = 2, 3, \dots, k$:
 - (a) Let $w_j = v_j - \sum_{i=1}^{j-1} \frac{\langle v_j, u_i \rangle}{\langle u_i, u_i \rangle} u_i$
 - (b) Let $u_j = w_j$

Here, $\langle \cdot, \cdot \rangle$ denotes the inner product.

The set of vectors $S' = \{u_1, u_2, \dots, u_k\}$ obtained from this process is orthogonal, but not necessarily orthonormal. To create an orthonormal set, you simply need to normalize each vector u_i to unit length. That is, $u'_i = \frac{u_i}{\|u_i\|}$, where $\|\cdot\|$ denotes the norm (or length) of a vector.

Among other things, making vectors orthonormal simplifies calculations, makes it easier to define rotations and transformations, and provides a framework for calculations in fields such as quantum mechanics.

88.2 Example Calculation

Given a set of linearly independent vectors, we will use the Gram-Schmidt process to find an orthogonal basis.

Let

$$W = \text{Span}(x_1, x_2, x_3)$$

where

$$x_1 = (1, 2, -2)$$

$$x_2 = (1, 0, -4)$$

$$x_3 = (5, 2, 0)$$

The three orthogonal vectors will define the same subspace as the original vectors.

The first vector of the orthogonal subspace is easy to define. We set it to be the same as x_1 .

$$v_1 = x_1 = (1, 2, -2)$$

The second orthogonal vector is a projection of x_2 onto v_1 . You learned projections in the last chapter, so this should be fairly straightforward.

$$v_2 = x_2 - \frac{x_2 v_1}{v_1 v_1} v_1$$

Substitute the values:

$$v_2 = (1, 0, -4) - \frac{(1, 0, -4)(1, 2, -2)}{(1, 2, -2)(1, 2, -2)} (1, 2, -2)$$

Calculate the coefficient for v_1 :

$$v_2 = (1, 0, -4) - \frac{9}{9}(1, 2, -2)$$

Perform the subtraction:

$$v_2 = (0, -2, -2)$$

The third vector for the orthogonal subspace is a projection onto v_1 and v_2 .

$$v_3 = x_3 - \frac{x_3 v_1}{v_1 v_1} v_1 - \frac{x_3 v_2}{v_2 v_2} v_2$$

Substitute the values:

$$\begin{aligned} v_3 &= (5, 2, 0) - \frac{(5, 2, 0)(1, 2, -2)}{(1, 2, -2)(1, 2, -2)} (1, 2, -2) - \frac{(5, 2, 0)(0, -2, -2)}{(0, -2, -2)(0, -2, -2)} (0, -2, -2) \\ v_3 &= (5, 2, 0) - (9/9)(1, 2, -2) - (-4/8)(0, -2, -2) \\ v_3 &= (5, 2, 0) - (1, 2, -2) + (1/2)(0, -2, -2) \\ v_3 &= (5, 2, 0) - (1, 2, -2) + (0, -1, -1) \\ v_3 &= (4, -1, 1) \end{aligned}$$

This set of vectors is orthogonal, so we need to normalize them so that the vectors are orthonormal. Recall that an orthonormal vector has a length of 1 and is computed using this formula:

$$\text{normalizedVector} = \text{vector}/\sqrt{\sum(\text{vector} * \text{vector})}$$

Thus the normalized set of vectors is:

$$v_1 = (0.33, 0.67, -0.67)$$

$$v_2 = (0.0, -0.71, -0.71)$$

$$v_3 = (0.94, -0.24, 0.24)$$

Exercise 103 Gram-Schmidt Process

Use the Gram-Schmidt process to find an orthogonal basis for the span defined by x_1, x_2 where:

Working Space

$$x_1 = (1, 1, 1)$$

$$x_2 = (0, 1, 1)$$

Answer on Page 840

88.3 The Gram-Schmidt Process in Python

Create a file called `vectors_gram-schmidt.py` and enter this code:

```
# import numpy to perform operations on vector
import numpy as np

# Find an orthogonal basis for the span of these three vectors
x1 = np.array([1, 2, -2])
x2 = np.array([1, 0, -4])
x3 = np.array([5, 2, 0])

# v1 = x1
v1 = x1
print("v1 = ", v1)

# v2 = x2 - (the projection of x2 on v1)
v2 = x2 - (np.dot(x2,v1)/np.dot(v1,v1))*v1
print("v2 = ", v2)

# v3 = x3 - (the projection of x3 on v1) - (the projection of x3 on v2)
v3 = x3 - (np.dot(x3,v1)/np.dot(v1,v1))*v1 - (np.dot(x3,v2)/np.dot(v2,v2))*v2
print("v3 = ", v3)
```

```
v3 = x3 - (np.dot(x3,v1)/np.dot(v1,v1))*v1 - (np.dot(x3,v2)/np.dot(v2,v2))*v2
print("v3 =", v3)

# Next, normalize each vector to get a set of vectors that is both orthogonal and orthonormal
v1_norm = v1 / np.sqrt(np.sum(v1**2))
v2_norm = v2 / np.sqrt(np.sum(v2**2))
v3_norm = v3 / np.sqrt(np.sum(v3**2))
print("v1_norm = ", v1_norm)
print("v2_norm = ", v2_norm)
print("v3_norm = ", v3_norm)
```

88.4 Where to Learn More

Watch this video from Khan Academy about the Gram-Schmidt process: [INSERT LINK](#)



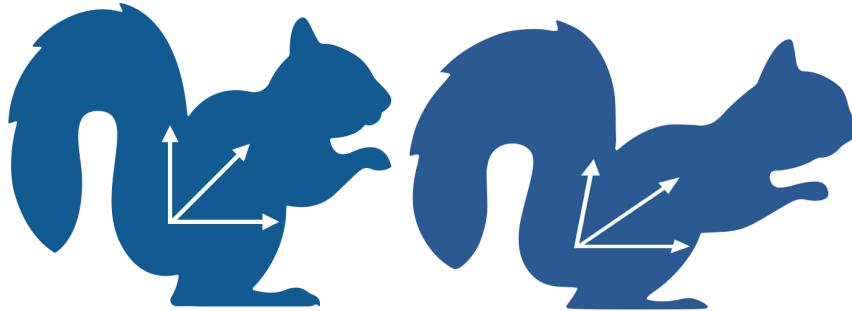
CHAPTER 89

Eigenvectors and Eigenvalues

Like many specialized disciplines, Linear Algebra uses many unfamiliar terms whose origin you might wonder about. Eigenvectors and eigenvalues are two of them. If you know German, you'll recognize that *eigen* means inherent or a characteristic attribute. Named by the German mathematician David Hilbert, an eigenvector mathematically describes a characteristic feature of an object, that remains unchanged after transformation. You can think of an eigenvector as the direction that doesn't change direction.

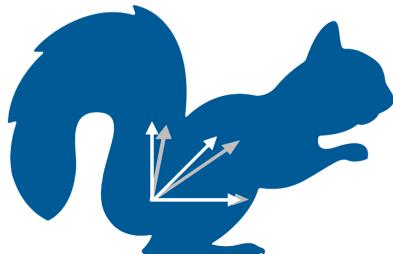
Eigenvalues and eigenvectors are a way to break down matrices that can simplify many calculations and enable us to understand various properties of the matrix. They are widely used in physics and engineering for stability analysis, vibration analysis, and many other applications.

Let's look at a visual example.



You can see that the image on the right is a skewed version of the image on the left. Look closely at the vectors and you'll notice that the one of the vectors is pointing in the same direction in both images, while the direction of the other two vectors has changed. The eigenvector is the one at the bottom that points 0 degrees (or you can think of due east) in both images. Thus the characteristic attribute of both images is their horizontal direction.

When you overlay the vectors from one image over the other, you'll notice that the horizontal vector, while the same direction in both images, is a bit longer in the skewed version. The scale of the stretch is described by an eigenvalue.



89.1 Definition

Given a square matrix A , a non-zero vector v is an eigenvector of A if multiplying A by v results in a scalar multiple of v , i.e.,

$$Av = \lambda v \quad (89.1)$$

where λ is a scalar known as the eigenvalue corresponding to the eigenvector v .

89.2 Finding Eigenvalues and Eigenvectors

You find the eigenvalues of a matrix A by solving the characteristic equation:

$$\det(A - \lambda I) = 0 \quad (89.2)$$

where $\det(\cdot)$ denotes the determinant, I is the identity matrix of the same size as A , and λ is a scalar.

Once you find the eigenvalues, you can find the corresponding eigenvectors by substituting each eigenvalue into the equation $Av = \lambda v$, and solving for v .

89.3 Example

For a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, the characteristic equation is:

$$(a - \lambda)(d - \lambda) - bc = 0 \quad (89.3)$$

Solving this equation gives the eigenvalues. Substituting each eigenvalue back into the equation $Av = \lambda v$ gives the corresponding eigenvectors.

Let matrix $A =$

$$\begin{bmatrix} 5, 4 \\ 1, 2 \end{bmatrix}$$

The characteristic equation is:

$$\begin{aligned} |A - \lambda I| &= 0 \\ \begin{bmatrix} 5 - \lambda, 4 \\ 1, 2 - \lambda \end{bmatrix} &= 0 \\ (5 - \lambda)(2 - \lambda) - (4)(1) &= 0 \\ 10 - 5\lambda - 2\lambda + \lambda^2 - 4 &= 0 \\ \lambda^2 - 7\lambda + 6 &= 0 \\ (\lambda - 6)(\lambda - 1) &= 0 \\ \lambda = 6, \lambda = 1 & \end{aligned}$$

Now that you have the eigen values you can substitute these values into the equation:

$$|A - \lambda I| = 0$$

For $\lambda = 1$:

$$(A - \lambda I)v = 0$$

$$\begin{bmatrix} 5-1, 4 \\ 1, 2-1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 4, 4 \\ 1, 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Next, use elementary row transformation by multiplying row 2 by 4 and then subtracting row 1.

$$\begin{bmatrix} 4, 4 \\ 0, 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now you can expand as an equation:

$$4x + 4y = 0$$

Assume $y = w$

$$4x = -4w$$

$$x = -w$$

The solution is:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -w \\ w \end{bmatrix} = w \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

So the eigenvector is:

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Now we need to substitute the other eigenvalue, 6, into the equation and follow the same procedure for finding the eigenvector.

$$\begin{bmatrix} 5-6, 4 \\ 1, 2-6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -1, 4 \\ 1, -4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Next, use elementary row transformation by adding row 1 to row 2.

$$\begin{bmatrix} -1, 4 \\ 0, 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Expand as an equation:

$$-x + 4y = 0$$

Assume $y = w$

$$-x + 4w = 0$$

$$x = 4w$$

The solution is:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4w \\ w \end{bmatrix} = w \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

So the eigenvector is:

$$\begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

In conclusion, the eigenvectors of the given 2×2 matrix are:

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

89.4 Eigenvalues and Eigenvectors in Python

Create a file called `vectors_eigen.py` and enter this code:

```
# import numpy to perform operations on vector
import numpy as np
from numpy.linalg import eig

a = np.array([[2, 2, 4],
              [1, 3, 5],
              [2, 3, 4]])
eigenvalue,eigenvector = eig(a)

# The values are not in any particular order
print('Eigenvalues:', eigenvalue)

# The eig function returns the normalize vectors
print('Eigenvectors:', eigenvector)
```

89.5 Where to Learn More

Watch this video from Khan Academy, *Introduction to Eigenvectors*: <https://rb.gy/mse7i>



CHAPTER 90

Singular Value Decomposition

In the previous chapter you learned how to calculate eigenvalues and eigenvectors. But not every matrix has them. For those matrices, singular values and singular vectors are analogous features.

Singular Value Decomposition (SVD) is a matrix factorization technique that breaks down a matrix into three matrices that represent the structure and properties of the original matrix. The decomposed matrices make calculations easier and provide insight into the original matrix. Basically, SVD can transform a high dimension, highly variable set of data into a set of uncorrelated data points that reveal subgroupings that you might not have noticed in the original data.

90.1 Definition

For any $m \times n$ matrix A , decomposes the matrix into three matrices.

$$A = U\Sigma V^T \tag{90.1}$$

U and V are orthogonal matrices. Σ is a diagonal matrix that is the same size as A . Its diagonal contain the singular values of A .

- U is an orthogonal matrix whose size is $m \times m$. Its columns are the eigenvectors of AA^T . These are the left singular vectors of A . Because U is orthogonal, $U^TU = I$
- V is an orthogonal matrix whose size is $n \times n$ matrix. Its columns are the eigenvectors of A^TA . These are the right singular vectors of A . Because V is orthogonal, $V^TV = I$.
- Σ is a diagonal matrix that is the same size as A . Its diagonal contains the singular values of A , arranged in descending order. These values are the square roots of the eigenvalues of both A^TA and AA^T .

90.2 Applications of SVD

SVD has numerous applications:

- It's used in machine learning and data science to perform dimensionality reduction, particularly through a technique known as Principal Component Analysis (PCA).
- In numerical linear algebra, SVD is used to solve linear equations and compute matrix inverses in a more numerically stable way.
- It's used in image compression, where low-rank approximations of an image matrix provide a compressed version of the original image.

90.3 Calculating SVD Manually

You might be inclined to skip this example because the computations are lengthy. Why would anyone do this when they can use a computing language, like Python, to calculate the SVD with essentially one command? We show this so you can understand what goes on "under the hood" when you compute SVD programmatically.

After you read through this example, you'll see how to use Python to compute SVD. Then you'll see an example of using SVD for image compression. Finally, you'll have an exercise to compute the SVD. For this, you'll need to write your own Python script.

Let's find the SVD for matrix A . Recall that we want to find U , Σ , and V^T such that:

$$A = U\Sigma V^T \tag{90.2}$$

$$A = \begin{bmatrix} 3, 1, 1 \\ -1, 3, 1 \end{bmatrix}$$

$$U = AA^T$$

Start with A^T :

$$A^T = \begin{bmatrix} 3, -1 \\ 1, 3 \\ 1, 1 \end{bmatrix}$$

$$AA^T = \begin{bmatrix} 3, 1, 1 \\ -1, 3, 1 \end{bmatrix} \begin{bmatrix} 3, -1 \\ 1, 3 \\ 1, 1 \end{bmatrix} = \begin{bmatrix} 11, 1 \\ 1, 11 \end{bmatrix}$$

Next we will find the eigenvalues and eigenvectors of A^T . This is a chance to apply what you learned in the previous chapter. We know that:

$$Av = \lambda v \quad (90.3)$$

So:

$$\begin{bmatrix} 11, 1 \\ 1, 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Rewrite as a set of equations:

$$11x_1 + x_2 = \lambda x_1$$

$$x_1 + 11x_2 = \lambda x_2$$

Then rearrange:

$$(11-\lambda)x_1 + x_2 = 0$$

$$x_1 + (11-\lambda)x_2 = 0$$

Solve for λ :

$$\begin{bmatrix} (11-\lambda), 1 \\ 1, (11-\lambda) \end{bmatrix} = 0$$

And as equations:

$$(11-\lambda)(11-\lambda)-1\cdot 1 = 0$$

$$(\lambda-10)(\lambda-12) = 0$$

These are the eigenvalues.

$$\lambda = 10$$

$$\lambda = 12$$

When substituted into the original equations, you get the eigenvectors. For

$$\lambda = 10$$

:

$$(11-10)x_1 + x_2 = 0$$

$$x_1 = -x_2$$

We'll set

$$x_1$$

to 1 and get this eigenvector:

$$[1, -1]$$

For

$$\lambda = 10$$

:

$$(11 - 12)x_1 + x_2 = 0$$

$$x_1 = x_2$$

We'll set

$$x_1$$

to 1 and get this eigenvector:

$$[1, 1]$$

The matrix is:

$$\begin{bmatrix} 1, 1 \\ 1, -1 \end{bmatrix}$$

Next you need to apply the Gram-Schmidt process to the column vectors. Then you'll have U , the $m \times m$ matrix whose columns are eigenvectors of AA^T . These are the left singular vectors of A . After you apply Gram-Schmidt, you should end up with:

$$U = \begin{bmatrix} 1/\sqrt{2}, 1/\sqrt{2} \\ 1/\sqrt{2}, -1/\sqrt{2} \end{bmatrix}$$

The process for calculating V is the same as the calculation for U , except:

$$V = A^T A$$

$$A^T A = \begin{bmatrix} 3, -1 \\ 1, 3 \\ 1, 1 \end{bmatrix} \begin{bmatrix} 3, 1, 1 \\ -1, 3, 1 \end{bmatrix} = \begin{bmatrix} 10, 0, 2 \\ 0, 10, 4 \\ 2, 4, 2 \end{bmatrix}$$

After applying the process we applied to solve for U , you get:

$$V = \begin{bmatrix} 1/\sqrt{6}, 2/\sqrt{5}, 1/\sqrt{30} \\ 2/\sqrt{6}, -1/\sqrt{5}, 2/\sqrt{30} \\ 1/\sqrt{6}, 0, -5/\sqrt{30} \end{bmatrix}$$

However, you want V_T :

$$V_T = \begin{bmatrix} 1/\sqrt{6}, 2/\sqrt{6}, 1/\sqrt{6} \\ 2/\sqrt{5}, -1/\sqrt{5}, 0 \\ 1/\sqrt{30}, 2/\sqrt{30}, -5/\sqrt{30} \end{bmatrix}$$

You have only to calculate Σ , a diagonal matrix that is the same size as A . The diagonal contains the singular values of A , arranged in descending order. They are the square roots of the eigenvalues of both $A^T A$ and AA^T .

Because the non-zero eigenvalues of U are the same as V , let's use the eigenvalues we calculate for U , 10 and 12. Note that Σ will not be of the correct dimension to reconstruct the original matrix unless we add a column. By adding a zero column you'll be able to multiply between U and V :

$$\Sigma = \begin{bmatrix} \sqrt{12}, & 0, & 0 \\ 0, & \sqrt{12} & 0 \end{bmatrix}$$

You can check your work by multiplying the decomposed matrices. This should return the original matrix.

$$\begin{aligned} A &= U\Sigma V^T \\ &= U = \begin{bmatrix} 1/\sqrt{2}, 1/\sqrt{2} \\ 1/\sqrt{2}, -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{12}, & 0, & 0 \\ 0, & \sqrt{12} & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{6}, 2/\sqrt{6}, 1/\sqrt{6} \\ 2/\sqrt{5}, -1/\sqrt{5}, 0 \\ 1/\sqrt{30}, 2/\sqrt{30}, -5/\sqrt{30} \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{12}/\sqrt{2}, & \sqrt{10}/\sqrt{2}, & 0 \\ \sqrt{12}/\sqrt{2}, & -\sqrt{10}/\sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{6}, & 2/\sqrt{6}, & 1/\sqrt{6} \\ 2/\sqrt{5}, & -1/\sqrt{5} & 0 \\ 1/\sqrt{30}, & 2/\sqrt{30}, & -5/\sqrt{30} \end{bmatrix} \\ &= \begin{bmatrix} 3, 1, 1 \\ -1, 3, 1 \end{bmatrix} \end{aligned}$$

90.4 Singular Value Decomposition with Python

Create a file called `vectors_decomposition.py` and enter this code:

```
# Singular-value decomposition
import numpy as np
from numpy import array
from scipy.linalg import svd
from numpy import diag
from numpy import dot
from numpy import zeros

# Define a matrix
A = array([[1, 2], [3, 4], [5, 6]])

print("Matrix (3x2) to be decomposed: ")
print(A)

# CalculateSVD
U, S, VT = svd(A)
print("Matrix (3x3) that represents the left singular values of A:")
print(U)
print("Singular values:")
print(S)
```

```
print("Matrix (2x2) that represents the right singular values of A:")
print(VT)

# Check if the decomposition by rebuilding the original matrix
# The singular values must be in an m x n matrix
# Create a zero matrix with the same dimension as A
Sigma = zeros((A.shape[0], A.shape[1]))
# Populate Sigma with n x n diagonal matrix
Sigma[:A.shape[1], :A.shape[1]] = diag(S)
# Reconstruct the original matrix
A_Rebuilt = U.dot(Sigma.dot(VT))
print("Original matrix:")
print(A_Rebuilt)
```

90.5 Sign Ambiguity

You might notice that at times the absolute values in the U and V^T matrices are correct but that the signs vary from what you see as the answer. For example, when you compare a manually calculated SVD with one done in Python the signs might not agree. Both decompositions of A are valid. Both decompositions will satisfy:

$$A = U\Sigma V^T$$

Note that the S diagonal values will always be positive.

The sign ambiguity has implications. For example, when using SVD to compress data, if some of the signs are flipped, the data can have artifacts. At this point in your education, you don't need to concern yourself with it except when you are comparing SVD results for the same matrix.

Exercise 104 Single Value Decomposition

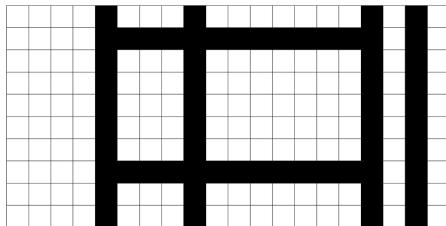
Modify your Python code to calculate SVD for the matrix in the worked out example. Did you arrive at the same answer? Keep in mind that Python will compute square roots and present fractions as decimal. Take a look at the signs for the values in the U and V^T matrices. Are they the same or is this an example of sign ambiguity?

Working Space

Answer on Page 841

90.6 SVD Applied to Image Compression

This image consists of a grid of 20 by 10 pixels, each of which is either black or white.



It's a simple image that has only two types of columns—ideal for data compression. A row is either the first pattern or the second.



We can represent the data as a 20 by 10 matrix whose 200 entries are either 0 for black or 1 for white.

$$\begin{bmatrix} 00001000100000001010 \\ 0000111111111111010 \\ 00001000100000001010 \\ 00001000100000001010 \\ 00001000100000001010 \\ 00001000100000001010 \\ 00001000100000001010 \\ 0000111111111111010 \\ 00001000100000001010 \\ 00001000100000001010 \end{bmatrix}$$

When you perform an SVD on this matrix, there are only two non-zero singular values, 6.79 and 3.72. (You are welcome perform the calculation in Python.) Thus you can represent the matrix as:

$$A = U_1 S_1 V_1 + U_2 S_2 V_2$$

This means there are two U vectors each with 20 entries and two V vectors each with 10 entries, and two singular values. Add those up: $2 \cdot 20 + 2 \cdot 10 + 2 = 62$. This implies that the image can be represented by 62 values instead of 200. If you look back at the image, you can see that there are many dependent columns and very few independent ones.

This is a simple image and a small pixel matrix. But it should give you a sense of how SVD can decompose an image in a way that identifies how much of the image is redundant, and therefore can be compressed.

90.7 Where to Learn More

We Recommend a Singular Value Decomposition. This American Mathematical Society publication focuses the geometry of SVD. What I like about the article is that it shows both graphically and numerically how SVD can be used for data compression on images and for noise reduction. The data compression example in your workbook is based on this article. <https://www.ams.org/publicoutreach/feature-column/fcarc-svd>

Sign Ambiguity in Singular Value Decomposition (SVD). This is a good article for those who want a deeper understanding of sign ambiguity. <https://www.educative.io/blog/sign-ambiguity-in-singular-value-decomposition>

Singular Value Decomposition Tutorial. This PDF starts by defining points, space, and vectors and works through all the concepts you need to tackle SVD. It is one of the few resources that has a completely worked out example of manually calculating SVD. The example in this chapter is from that tutorial. If you read the entire paper, you'll find it is a good review of the concepts you've studied in previous chapters. <https://rb.gy/j6s0w>



CHAPTER 91

Data Tables and pandas

Much of the data that you will encounter in your career will come to you as a table. Some of these tables are spreadsheets, some are in relational databases, some will come to you as CSV files.

Typically each column will represent an attribute (like height or acreage) and each row will represent an entity (like a person or a farm). You might get a table like this:

property_id	bedrooms	square_meters	estimated_value
7927	3	921.4	\$ 294,393
9329	2	829.1	\$ 207,420

Typically, one of the columns is guaranteed to be unique. We call this the *primary key*. In this table, `property_id` is the primary key: every property has one, and no two properties have the same `property_id`.

91.1 Data types

Each column in a table has a type, and these usually correspond pretty nicely with types in Python.

Here are some common datatypes:

Type	Python type	Example
Integer	int	910393
Float	float	-23.19
String	string	'Fred'
Boolean	bool	False
Date	datetime.date	2019-12-04
Timestamps	datetime.datetime	2022-06-10T14:05:22Z

Sometimes it is OK to have values missing. For example, if you had a table of data about employees, maybe one of the columns would be `retirement`, a date that tells you when the person retired. People who had not yet retired would have no value in this column. We would say that they have *null* for `retirement`.

Sometimes there are constraints on what values can appear in the column. For example, if the column were `height`, it would make no sense to have a negative value.

Sometimes a column can only be one of a few values. For example, if you ran a bike rental shop, each bicycle's status would be "available", "rented", or "broken". Any other values in that column would not be allowed. We often call these columns *categorical*.

91.2 pandas

The Python community works with tables of data *a lot*, so it created the pandas library for reading, writing, and manipulating tables of data.

When working with tables, you sometimes need to go through them row-by-row. However, for large datasets, this is very slow. pandas makes it easy (and very fast) to say things like "Delete every row that doesn't have a value for `height`" instead of requiring you to step through the whole table.

In pandas, there are two datatypes that you use a lot:

- a `Series` is a single column of data.
- a `DataFrame` is a table of data: it has a `Series` for each column.

In the digital resources, you will find `bikes.csv`. If you look at it in a text editor, it will

start like this:

```
bike_id,brand,size,purchase_price,purchase_date,status
5636248,GT,57,277.99,1986-09-07,available
4156134,Giant,56,201.52,2005-01-09,rented
7971254,Cannondale,54,292.25,1978-02-28,available
3600023,Canyon,57,197.62,2007-02-15,broken
```

The first line is a header and tells you the name of each column. Then the values are separated by commas. (Thus the name: CSV stands for “Comma Separated Values”.)

91.3 Reading a CSV with pandas

Let’s make a program that reads `bikes.csv` into a pandas dataframe. Create a file called `report.py` in the same folder as `bikes.csv`.

First, we will read in the csv file. pandas has one Series that acts as the primary key; it calls this one the index. When reading in the file, we will tell it to use the `bike_id` as the index series.

If you ask a dataframe for its shape, it returns a tuple containing the number of rows and the number of columns. To confirm that we have actually read the data in, let’s print those numbers. Add these lines to `report.py`:

```
import pandas as pd

# Read the CSV and create a dataframe
df = pd.read_csv('bikes.csv', index_col="bike_id")

# Show the shape of the dataframe
(row_count, col_count) = df.shape
print(f"*** Basics ***")
print(f"Bikes: {row_count},")
print(f"Columns: {col_count}")
```

Build it and run it. You should see something like this:

```
*** Basics ***
Bikes: 998
Columns: 5
```

Note that your table actually had 6 columns. The index series is not included in the shape.

91.4 Looking at a Series

Let's get the lowest, the highest, and the mean purchase price of the bikes. The purchase price is a series, and you can ask the dataframe for it. Add these lines to the end of your program:

```
# Purchase price stats
print("\n*** Purchase Price ***")
series = df["purchase_price"]
print(f"Lowest:{series.min()}")
print(f"Highest:{series.max()}")
print(f"Mean:{series.mean():.2f}")
```

Now when you run it, you will see a few additional lines:

```
*** Purchase Price ***
Lowest:107.37
Highest:377.7
Mean:249.01
```

What are all the brands of the bikes? Add a few more lines to your program that shows how many of each brand:

```
# Brand stats
print("\n*** Brands ***")
series = df["brand"]
series_counts = series.value_counts()
print(f"{series_counts}")
```

Now when you run it, your report will include the number of bikes for each brand from most common to least:

```
*** Brands ***
Canyon      192
BMC         173
Cannondale  170
Trek        166
GT          150
Giant       147
Name: brand, dtype: int64
```

`value_counts` returns a Series. To format this better we need to learn about accessing individual rows in a series.

91.5 Rows and the index

In an array, you ask for data using an the location (as an int) of the item you want. You can do this in pandas using `iloc`. Add this to the end of your program:

```
# First bike
print("\n*** First Bike ***")
row = df.iloc[0]
print(f"{row}")
```

When you run it, you will see the attributes of the first row of data:

```
*** First Bike ***
brand           GT
size            57
purchase_price    277.99
purchase_date     1986-09-07
status          available
Name: 5636248, dtype: object
```

Notice that the data coming back is actually another series.

The last line says that the name (the value for the index column) for this row is 5636248. In pandas, we usually use this to locate particular rows. For example, there is a row with `bike_id` equal to 2969341. Let's ask for one entry from the

```
print("\n*** Some Bike ***")
brand = df.loc[2969341]['brand']
print(f"brand = {brand}")
```

Now you will see the information about that bike:

```
*** Some Bike ***
brand = Cannondale
```

pandas has a few different ways of getting to that value. All of these get you the same thing:

```
brand = df.loc[2969341]['brand'] # Get row, then get value
brand = df['brand'][2969341]      # Get column, then get value
brand = df.loc[2969341, 'brand'] # One call with both row and value
```

91.6 Changing data

One of your attributes needs cleaning up. Every bike should have a status and it should be one of the following strings: "available", "rented", or "broken". Get counts for each unique value in status:

```
print("\n*** Status ***")
series = df["status"]
missing = series.isnull()
print(f"{missing.sum()} bikes have no status.")
series_counts = series.value_counts()
for value in series_counts.index:
    print(f"{series_counts.loc[value]} bikes are \"{value}\"")
```

This will show you:

```
*** Status ***
7 bikes have no status.
389 bikes are "rented"
304 bikes are "broken"
296 bikes are "available"
1 bikes are "Flat tire"
1 bikes are "Available"
```

Right away we can see two easily fixable problems: Someone typed "Available" instead of "available". Right after you read the CSV in, fix this in the data frame:

```
mask = df['status'] == 'Available'
print(f"{mask}")
df.loc[mask, 'status'] = 'available'
```

When you run this, you will see that the mask is a series with `bike_id` as the index and `False` or `True` as the value, depending on whether the row's status was equal to "Available".

When you use `loc` with this sort of mask, you are saying "Give me all the rows for which the mask is True." So, the assignment only happens in the one problematic row.

Let's get rid of the mask variable and do the same for turning `Flat tire` into `Broken`:

```
df.loc[df['status'] == 'Available', 'status'] = 'available'
df.loc[df['status'] == 'Flat tire', 'status'] = 'broken'
```

Now those problems are gone:

```
7 bikes have no status.
389 bikes are "rented"
305 bikes are "broken"
297 bikes are "available"
```

What about the rows with no values for status? We were pretty certain that the bikes were available, we could just set them to 'available':

```
missing_mask = df['status'].isnull()
df.loc[missing_mask, 'status'] = 'available'
```

Or maybe we would print out the IDs of the bikes so that we could go look for them:

```
missing_mask = df['status'].isnull()
missing_ids = list(df[missing_mask].index)
print(f"These bikes have no status:{missing_ids}")
```

But lets just keep the rows where the status is not null:

```
missing_mask = df['status'].isnull()
df = df[~missing_mask]
```

At the end of your program, write out the improved CSV:

```
df.to_csv('bikes2.csv')
```

Run the program and open `bikes2.csv` in a text editor.

91.7 Derived columns

Let's say that you want to add a column with age of the bicycle in days:

```
bike_id,brand,size,purchase_price,purchase_date,status,age_in_days
5636248,GT,57,277.99,1986-09-07,available,13061
4156134,Giant,56,201.52,2005-01-09,rented,6362
7971254,Cannondale,54,292.25,1978-02-28,available,16174
```

Your first problem is that the `purchase_date` column looks like a date, but really it is a string. So you need to convert it to a date. You can do this by applying a function to every item in the series:

```
df['purchase_date'] = df['purchase_date'].apply(lambda s: datetime.date.fromisoformat(s))
```

(With pandas, there is often more than one way to do things. pandas has a `to_datetime` function that converts every entry in a sequence to a `datetime` object. So here is another way to convert the string column in to a date column:

```
df['purchase_date'] = pd.to_datetime(df['purchase_date']).dt.date
```

You can look up `dt` and `date` if you are curious.)

Now, we can use the same trick to create a new column with the age in days:

```
today = datetime.date.today()
df['age_in_days'] = df['purchase_date'].apply(lambda d: (today - d).days)
```

When you run this, the new `bikes.csv` will have an `age_by_date` column.



CHAPTER 92

Data tables in SQL

Most organizations keep their data as tables inside a relational database management system. Developers talk to those systems using a language called SQL (“Structured Query Language”).

Some relational database managers are pricey products you may have heard of before: Oracle, Microsoft SQL Server. Some are free: PostgreSQL or MySQL. These are server software that client programs talk to over the companies network.

There is a library, called `sqlite`, that lets us create files that hold tables. We can use SQL to create, edit, and browse those tables. `sqlite` is free, fast, and very easy to install. So we will use `sqlite` instead of a networked database management system.

If you look in your digital resources, you will find a file called `bikes.db`. I created this file using `sqlite`, and now you will use `sqlite` to access it.

In the terminal, get to the directory where `bikes.db` lives. To open the `sqlite` tool on that file:

```
> textbfsqlite3 bikes.db
```

(If your system complains that there is no sqlite3 tool, you need to install sqlite. See this website: <https://sqlite.org/>)

Please follow along: type each command shown here into the terminal and see what happens.

We mostly run SQL commands in this tool, but there are a few non-SQL commands that all start with a period. To see the tables and their columns, you can run .schema:

```
sqlite> .schema
CREATE TABLE bike (bike_id int PRIMARY KEY, brand text, size int,
    purchase_price real, purchase_date date, status text);
```

That is the SQL command that I used to create the `bike` table. You can see all the columns and their types.

You want to see all the rows of data in that table?

```
sqlite> select * from bike;
4997391|GT|57|269.61|2009-05-03|rented
5429447|Cannondale|50|215.91|2002-02-17|broken
5019171|Trek|58|251.17|1985-07-11|rented
3000288|Cannondale|57|211.08|1993-01-05|broken
880965|GT|52|281.75|1995-08-02|available
...
```

You will see 1000 rows of data!

The SQL language is not case-sensitive, so you can also write it like this:

```
sqlite> SELECT * FROM BIKE;
```

Often you will see SQL with just the SQL keywords in all caps:

```
sqlite> SELECT * FROM bike;
```

The semicolon is not part of SQL, but it tells sqlite that you are done writing a command and that it should be executed.

SQL lets you choose which columns you would like to see:

```
sqlite> SELECT bike_id, brand FROM bike;
4997391|GT
5429447|Cannondale
5019171|Trek
3000288|Cannondale
...
```

Using WHERE, SQL lets you choose which rows you would like to see:

```
sqlite> SELECT * FROM bike WHERE purchase_date > '2009-01-01' AND brand = 'GT';
4997391|GT|57|269.61|2009-05-03|rented
326774|GT|56|165.0|2009-06-27|available
264933|GT|52|302.43|2009-07-09|available
5931243|GT|55|173.56|2009-11-26|rented
4819848|GT|51|221.71|2009-12-11|rented
9347713|GT|52|232.32|2009-06-13|available
3019205|GT|58|262.94|2009-08-22|available
```

Using DISTINCT, SQL lets you get just one copy of each value:

```
sqlite> SELECT DISTINCT status FROM bike;
rented
broken
available

Busted
Flat tire
good
out
Rented
```

You can also edit these rows. For example, if you wanted every status that is Busted to be changed to broken. You can use an UPDATE statement:

```
sqlite> UPDATE bike SET status='broken' WHERE status='Busted';
sqlite> SELECT DISTINCT status FROM bike;
rented
broken
available

Flat tire
good
out
Rented
```

You can insert new rows:

```
sqlite> INSERT INTO bike (bike_id, brand, size, purchase_price, purchase_date, status)
...> VALUES (1, 'GT', 53, 123.45, '2020-11-13', 'available');
sqlite> SELECT * FROM bike WHERE bike_id = 1;
1|GT|53|123.45|2020-11-13|available
```

You can delete rows:

```
sqlite> DELETE FROM bike WHERE bike_id = 1;
sqlite> SELECT * FROM bike WHERE bike_id = 1;
```

To get out of sqlite, type .exit.

Exercise 105 SQL Query

Execute an SQL query that returns the `bike_id` (no other columns) of every Trek bike that cost more than \$300.

Working Space

Answer on Page 841

92.1 Using SQL from Python

The people behind sqlite created a library for Python that lets you execute SQL and fetch the results from inside a python program.

Let's create a simple program that fetches and displays the bike ID and purchase date of every Trek bike that cost more than \$300.

Create a file called `report.py`:

```
import sqlite3 as db

con = db.connect('bikes.db')
cur = con.cursor()
```

```
cur.execute("SELECT bike_id, purchase_date FROM bike WHERE purchase_price > 330 AND bran  
rows = cur.fetchall()  
  
today = datetime.date.today()  
for row in rows:  
    print(f"Bike {row[0]}, purchased {row[1]}")  
  
con.close()
```

When you execute it, you should see:

```
> python3 report.py  
Bike 4128046, purchased 2007-08-06  
Bike 7117808, purchased 1995-03-12  
Bike 7176903, purchased 1986-07-03  
Bike 827899, purchased 2009-03-14  
Bike 363983, purchased 1970-08-16
```




CHAPTER 93

Representing Natural Numbers

The natural numbers are 1, 2, 3, and so on. -5 is not a natural number. π is not a natural number. $\frac{1}{2}$ is not a natural number.

You are used to seeing the natural numbers represented in a base-10 *Hindu-Arabic* numeral system. That is, when you see 2531 you think “2 thousands, 5 hundreds, 3 tens, and 1 one.” Rewritten this is

$$2 \times 10^3 + 5 \times 10^2 + 3 \times 10^1 + 1 \times 10^0$$

In any Hindu-Arabic system, the location of the digits is meaningful: 101 is different from 110. Here are those numbers in Roman numerals: CI and CX. Roman numerals didn’t have a symbol for zero at all.

The Hindu-Arabic system gave us really straightforward algorithms for addition and multiplication. For addition, you memorized the following table:

	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9	10
2	2	3	4	5	6	7	8	9	10	11
3	3	4	5	6	7	8	9	10	11	12
4	4	5	6	7	8	9	10	11	12	13
5	5	6	7	8	9	10	11	12	13	14
6	6	7	8	9	10	11	12	13	14	15
7	7	8	9	10	11	12	13	14	15	16
8	8	9	10	11	12	13	14	15	16	17
9	9	10	11	12	13	14	15	16	17	18

Then when you multiplied two number together, you just multiplied each pair of digits.
 254×26 might look like this:

$$\begin{array}{r}
 2 \quad 5 \quad 4 \\
 \times \quad 2 \quad 6 \\
 \hline
 & 2 & 4 & 6 \times 4 \\
 & 3 & 0 & 6 \times 5 \\
 1 & 2 & & 6 \times 2 \\
 & & 8 & 2 \times 4 \\
 & 1 & 0 & 2 \times 5 \\
 + & 4 & & 2 \times 2 \\
 \hline
 6 & 6 & 0 & 4
 \end{array}$$

For multiplication, you memorized this table:

	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6	7	8	9
2	0	2	4	6	8	10	12	14	16	18
3	0	3	6	9	12	15	18	21	24	27
4	0	4	8	12	16	20	24	28	32	36
5	0	5	10	15	20	25	30	35	40	45
6	0	6	12	18	24	30	36	42	48	54
7	0	7	14	21	28	35	42	49	56	63
9	0	9	18	27	36	45	54	63	72	81



CHAPTER 94

Making Web Requests with HTTP

The Hypertext Transfer Protocol (HTTP) is the protocol used for transmitting hypertext over the World Wide Web. It is the foundation of any data exchange on the web and it is a protocol used for transmitting hypertext requests from clients (like a user's browser) to servers, which respond with the requested resources.

94.1 HTTP Requests

An HTTP request is made up of several components:

- **Method:** The HTTP method, like GET (retrieve data), POST (send data), PUT (update data), DELETE (remove data), etc.
- **URL:** The URL of the resource to retrieve, send data to, update or delete.
- **Headers:** Additional information about the request or response, like the content type of the body.

- **Body:** The body of the request, used when sending data in POST or PUT requests.

94.2 Using HTTP with Web-Based APIs

Software developers often use HTTP to interact with web-based APIs. An Application Programming Interface (API) is a set of rules that allows programs to talk to each other. The developer creates the API on the server and allows the client to talk to it.

When a developer makes a request to an API endpoint, they're asking the server to either send them some data or receive some data from them. The response from the server will often be in a format like JSON or XML, which the developer can then use in their own application.

For instance, a developer might make a GET request to '`https://api.example.com/users`' to retrieve a list of all users. The server would respond with a list of users in a format like JSON.



CHAPTER 95

Using and Creating APIs

As a software engineer, you are likely familiar with building applications that interact with various external services and data sources. One of the most common methods for communication and integration is through HTTP APIs (Application Programming Interfaces). HTTP APIs provide a standardized way for applications to exchange data and functionality over the internet.

This chapter will introduce you to the world of HTTP APIs and explore how you can leverage them in your software development projects. We will cover the fundamental concepts, techniques, and best practices for effectively working with HTTP APIs.

An HTTP API allows two software systems to communicate and exchange information using the Hypertext Transfer Protocol (HTTP). It enables your application to make requests to an API server and receive responses in a structured format, such as JSON (JavaScript Object Notation) or XML (eXtensible Markup Language).

Using HTTP APIs offers a range of benefits for software engineers. It allows you to leverage external services and data sources, enabling your application to access functionality or retrieve valuable information from third-party systems. This opens up opportunities for integration with popular platforms, social media networks, payment gateways, geolo-

cation services, and much more.

Throughout this chapter, we will explore various aspects of working with HTTP APIs, including:

- API endpoints and methods: Understanding how to interact with an API involves identifying the available endpoints and the supported methods, such as GET, POST, PUT, DELETE, etc. We will discuss how to construct API requests and handle different response formats.
- Authentication and authorization: Many APIs require authentication to ensure secure access and protect sensitive data. We will delve into different authentication mechanisms, including API keys, tokens, OAuth, and other authentication protocols commonly used in API integrations.
- Request parameters and payloads: APIs often accept additional parameters or payloads to customize the request or send data for processing. We will explore how to pass query parameters, request headers, and request bodies when interacting with APIs.
- Error handling and status codes: Learning how to handle errors and interpret status codes returned by APIs is crucial for building robust and resilient applications. We will discuss common status codes and best practices for handling various scenarios gracefully.
- Rate limiting and throttling: Many APIs impose restrictions on the number of requests you can make within a given timeframe to prevent abuse and ensure fair usage. We will cover techniques for handling rate limiting and implementing efficient strategies to manage API quotas.
- API documentation and testing: Proper documentation is essential for understanding an API's capabilities, endpoints, and expected behavior. We will explore how to read and interpret API documentation, as well as techniques for testing and validating API integrations.

By mastering the art of using HTTP APIs, you will expand your development toolkit and gain the ability to seamlessly integrate your applications with external services, leverage their functionalities, and build powerful, interconnected systems.

So, let's dive into the world of HTTP APIs and uncover the endless possibilities they offer for enhancing your software engineering projects.



CHAPTER 96

Data Compression and Decompression

Data compression and decompression are fundamental techniques used in modern computing, enabling efficient storage and transmission of data. The concept of entropy, borrowed from the field of information theory, plays a crucial role in determining the compression rate.

96.1 Data Compression and Decompression

Data compression is the process of reducing the amount of data needed to represent a particular set of information. The two main types of data compression are lossless and lossy. Lossless compression ensures that the original data can be perfectly reconstructed from the compressed data, whereas lossy compression allows some loss of data for more significant compression rates.

Decompression is the reverse process of compression, reconstructing the original data from the compressed format.

96.2 Entropy

In information theory, entropy measures the unpredictability or randomness of information content. More specifically, it quantifies the expected value of the information contained in a message. Lower entropy implies less randomness and more repetitiveness, which in turn means the data can be compressed more.

96.3 Entropy and Compression

The role of entropy in data compression is fundamental. The entropy of a source of data is the minimum number of bits required, on average, to encode symbols drawn from the source. It serves as a lower bound on the best possible lossless compression rate.

For a source X with probability distribution $p(x)$, the entropy $H(X)$ is defined as:

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \quad (96.1)$$

If the entropy of the data is high (i.e., the data is random and unpredictable), the potential for compression is low. On the other hand, if the entropy is low (the data is predictable), the data can be compressed to a smaller size.



CHAPTER 97

Dealing with JSON and XML



CHAPTER 98

HTML

HTML, an abbreviation for Hypertext Markup Language, is the standard language for creating web pages and web applications. It's a cornerstone technology of the World Wide Web and forms the structure and layout of web content.

98.1 HTML Elements

An HTML document is composed of a series of elements, which are denoted by tags. Elements have an opening tag and a closing tag with content in between. Some elements, however, are self-closing and do not contain any content. For example, the paragraph tag '`<p>`' is used to denote a paragraph:

```
<p>This is a paragraph.</p>
```

98.2 HTML Document Structure

A typical HTML document has a specific structure, including the following elements:

- **DOCTYPE declaration:** It informs the browser about the version of HTML. For HTML5, it is ‘<!DOCTYPE html>’.
- **html:** This tag encloses the entire HTML document.
- **head:** This contains meta-information about the document, such as its title, meta tags, and links to scripts and stylesheets.
- **body:** This contains the content of the web page that is rendered in the browser.

Here is a basic example of an HTML document:

```
<!DOCTYPE html>
<html>
<head>
    <title>My First HTML Page</title>
</head>
<body>
    <h1>Welcome to My First HTML Page!</h1>
    <p>This is a paragraph.</p>
</body>
</html>
```



CHAPTER 99

Introduction to Text

In computer systems, text is represented in files as a sequence of characters, each of which corresponds to a specific number known as a character code. These character codes are then stored in the file as binary data.

99.1 Newlines and Carriage Returns

Two of the character codes that have special meanings are the newline (often represented as '^n') and the carriage return (often represented as '\r').

The newline character signifies the end of a line of text and the beginning of a new one. The carriage return character moves the cursor to the beginning of the line. The use of these characters can vary between operating systems. Unix-based systems (like Linux and MacOS) use the newline character to indicate the end of a line, while Windows systems use a combination of a carriage return and a newline ('\r\n').

99.2 ASCII

The American Standard Code for Information Interchange (ASCII) is one of the earliest character encodings. It uses 7 bits to represent each character, allowing it to define up to $2^7 = 128$ different characters. These include the English alphabet (in both lower and upper cases), digits, punctuation symbols, control characters (like newline and carriage return), and some other symbols.

99.3 UTF-8

UTF-8 (8-bit Unicode Transformation Format) is a variable-width character encoding that can represent every character in the Unicode standard, yet remains backward-compatible with ASCII. For the ASCII range (0-127), UTF-8 is identical to ASCII. But it can use additional bytes (up to 4 bytes in total) to represent characters that are not included in ASCII, such as characters from other languages, emojis, and many other symbols. This has made UTF-8 a widely used encoding in many modern systems.



CHAPTER 100

Stop Words



CHAPTER 101

Stemming and Lemmatization

Stemming and lemmatization are two fundamental techniques in natural language processing that are used to prepare text data. They help in reducing inflectional forms of a word to a common base form.

101.0.1 Stemming

Stemming is the process of reducing a word to its word stem, i.e., its basic form. For instance, the stem of the word 'jumps' would be 'jump'. A stemming algorithm reduces the words "jumping", "jumped", and "jumps" to the stem "jump".

It's important to note that stemming may not always lead to actual words. For example, the stem of the word "running" could be "runn" depending on the stemming algorithm used.

Stemming is generally simpler and faster than lemmatization, but it is also less precise.

101.0.2 Lemmatization

Lemmatization, on the other hand, reduces words to their base or root form, which is linguistically correct. For example, "running" and "runs" are both changed to "run".

Lemmatization uses a more complex approach to achieve this: it considers the morphological analysis of the words and requires detailed dictionaries which the algorithm can look through to link the form back to its lemma.

To summarise, both stemming and lemmatization help in text normalization and preprocessing, but while stemming can be faster and simpler, lemmatization is more accurate as it uses more informed analysis to create groups of words with similar meanings based on the context.



CHAPTER 102

Alphabets and Accents

In today's interconnected world, software developers often encounter text from diverse languages and cultures. As a developer, it is crucial to have a solid understanding of alphabets and accents to effectively handle and process this multilingual text. Alphabets, the building blocks of written language, vary widely across different nations and regions. Meanwhile, accents, diacritical marks, and other phonetic notations play a crucial role in conveying the correct pronunciation and meaning of words.

This guide aims to provide software developers with a fundamental understanding of alphabets and accents to navigate the complexities of handling text from different nations. By familiarizing yourself with these concepts, you will be better equipped to develop robust applications, support multiple languages, and ensure accurate representation and interpretation of text data.

Alphabets are sets of letters or symbols used to represent the sounds of a language. While the Latin alphabet is widely used in many Western languages, numerous other alphabets exist, such as Cyrillic, Greek, Arabic, Devanagari, and Chinese characters. Each alphabet has its own unique set of letters, often organized in a specific order, and may include uppercase and lowercase variations.

Accents and diacritical marks are additional symbols added to letters to modify their pronunciation or provide additional phonetic information. Accents can appear above, below, or beside a letter, and they can change the sound, stress, or intonation of a word. For example, in French, the acute accent ('e) changes the pronunciation of the letter "e" from / / to /e/.

When working with multilingual text, it is essential to consider various factors:

1. Character encoding: Different alphabets require specific character encodings to represent their letters digitally. Commonly used character encodings include ASCII, Unicode, and UTF-8. Understanding the appropriate encoding for each language is crucial to ensure proper text rendering and avoid data corruption.
2. Text input and validation: Building applications that handle user input requires robust text validation. Account for the diverse set of characters and possible accents that may appear in user-generated content. Implement proper validation and sanitization mechanisms to handle text input securely.
3. Sorting and collation: Sorting text from different languages involves considering the specific rules and conventions of each alphabet. Some languages may have unique sorting orders, while others ignore accents or diacritics when determining the order of words. Take into account the appropriate sorting and collation algorithms to ensure consistent and accurate results.
4. Search and indexing: Efficient search and indexing systems must be capable of handling multilingual text. Consider appropriate text normalization techniques to account for different character representations (e.g., case-insensitive matching, ignoring accents), enabling users to find relevant content across languages and variations in spelling or diacritics.

By grasping the concepts of alphabets and accents, software developers can build robust, inclusive applications that handle multilingual text effectively. Understanding character encodings, implementing proper text validation, considering sorting and collation rules, and enabling efficient search capabilities are crucial steps toward supporting diverse linguistic communities and providing a seamless user experience across different languages.

Now, let's delve deeper into specific alphabets and accents commonly encountered in software development, exploring their unique characteristics and considerations for handling text from different nations.



CHAPTER 103

Making Plots with matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It's highly useful for presenting data in a more intuitive and easy-to-understand manner.

In order to use Matplotlib, you must first import it, typically using the following line of code:

```
import matplotlib.pyplot as plt
```

Let's create a simple line plot. Suppose we have a list of numbers and we want to visualize their distribution:

```
x = [1, 2, 3, 4, 5]
y = [1, 4, 9, 16, 25]

plt.plot(x, y)
plt.show()
```

Here, 'x' and 'y' are the coordinates of the points. The 'plt.plot' function plots y versus x as lines and/or markers. The 'plt.show' function then displays the figure.

Creating a bar plot follows a similar approach:

```
labels = [ 'A' , 'B' , 'C' , 'D' , 'E' ]  
values = [5, 7, 9, 11, 13]
```

```
plt.bar(labels , values)  
plt.show()
```

Here, ‘labels’ are the categories we are plotting, and ‘values’ are the respective sizes of those categories. The ‘plt.bar’ function creates a bar plot.

Matplotlib provides a variety of other plot types and customization options - everything from scatter plots and histograms to custom line styles and colors. Explore the official Matplotlib documentation to learn more about what this powerful library can offer.



CHAPTER 104

Geographical Data



CHAPTER 105

Geocoding and Reverse Geocoding

Geocoding and reverse geocoding are essential processes in geographic information systems (GIS) that are used to convert between addresses and spatial data.

105.1 Geocoding

Geocoding is the process of converting addresses (like "1600 Amphitheatre Parkway, Mountain View, CA") into geographic coordinates (like latitude 37.423021 and longitude -122.083739), which you can use to place markers on a map, or position the map. The resulting latitude and longitude are often used as a key index in merging datasets based on location.

Here is an example of using Google's Geocoding service to get the longitude and latitude of the Dallas County Administration Building:

```
import requests
```

```
import json

# Encode the parameters
parameters = {"address": "411 Elm St, Dallas, TX 75202", "key": "YOUR_API_KEY"}
base_url = "https://maps.googleapis.com/maps/api/geocode/json?"

# Send the GET request
response = requests.get(base_url, params=parameters)

# Convert the response to json
data = response.json()

# Extract the latitude and longitude
if len(data["results"]) > 0:
    latitude = data["results"][0]["geometry"]["location"]["lat"]
    longitude = data["results"][0]["geometry"]["location"]["lng"]
else:
    print(f"Could not find the latitude and longitude.")
```

105.2 Reverse Geocoding

Reverse geocoding, as the name implies, is the opposite process of geocoding. It involves converting geographic coordinates into a human-readable address. This can be useful in applications where you need to display an actual address to a user instead of latitude and longitude coordinates.

Here is an example of using Google's reverse geocoding API to find the address for

```
import requests
import json

api_key = "YOUR_API_KEY"
latitude = 33.9474096
longitude = -118.1179069

# Encode the parameters
parameters = {"latlng": f"{latitude},{longitude}", "key": api_key}
base_url = "https://maps.googleapis.com/maps/api/geocode/json?"

# Send the GET request
response = requests.get(base_url, params=parameters)

# Convert the response to json
data = response.json()
```

```
# Extract the address
if len(data["results"]) > 0:
    address = data["results"][0]["formatted_address"]
else:
    print(f"Could not find the address")
```




CHAPTER 106

Making a Map

Plotly is an open-source data visualization library for Python, R, and JavaScript. It allows for interactive plots, including geographical maps. In this brief example, we will learn how to create a simple annotated map using Plotly in Python.

To start, you need to install Plotly. In Python, you can do this via pip:

```
pip install plotly
```

Once installed, you can create a map with annotations as follows:

```
import plotly.graph_objects as go

fig = go.Figure(data=go.Scattergeo(
    lon = [-75.789],
    lat = [45.4215],
    text = ['Ottawa'],
    mode = 'text',
))
fig.update_layout()
```

```
title_text = 'Annotated Map with Plotly',
showlegend = False,
geo = dict(
    scope = 'world',
    projection_type = 'azimuthal_equal_area',
    showland = True,
    landcolor = 'rgb(243, 243, 243)',
    countrycolor = 'rgb(204, 204, 204)',
),
)

fig.show()
```

This code creates a world map and places a text annotation at the geographic coordinates for Ottawa. The ‘go.Scattergeo’ function is used to define the geographical scatter plot (i.e., the annotation), while the ‘update_layout’ function is used to define the appearance and the properties of the map itself.

In this example, you can replace the latitude, longitude, and text with the values corresponding to your desired location.



CHAPTER 107

Introduction to Discrete Probability

First, let's take care of the word *discrete* vs *discreet*. They sound exactly the same, but “discrete” means “individually separate and distinct” and “discreet” means “careful about what other people know”. So you might say, “You can think of light as a continuous wave or as a blast of discrete particles.” And you might say, “Please go get the box of doughnuts from the kitchen. Oh, and there are a lot of hungry people in the house, so be discreet.”

When we are talking about probabilities, some problems deal with discrete quantities like “What is the probability that I will throw these three dice and the numbers that roll face up sum to 9?”. There are also problems that deal with continuous properties like “What is the probability that the next bird to fly over my house will weigh between 97.2 and 98.1 grams ?” In this module, we are going to focus on the probability problems that deal with discrete quantities.

Watch Khan Academy’s Introduction to Probability at <https://youtu.be/uzkc-qNVoOk>.

Let’s say that I have a cloth sack filled with 100 marbles; 99 are red and 1 is white. If

I ask you to reach in without looking and pull out one marble, you will probably pull out a red one. We say that “There is a 1 in 100 chance that you would pull out a white marble.” Or we can use percentages and say “There is a 1% chance that you will pull out a white marble.” Or we can use decimals and say “There is a 0.01 probability that you will pull out a white marble.” In probability, we often talk about the probability of certain events. “Pulling out a white marble” is an event, and we can give it a symbol like W . Then, in equations we use p to mean “the probability of”. Thus, we can say “There is a 0.01 probability that you will pull out a white marble” which becomes the equation

$$p(W) = 0.01$$

107.1 The Probability of All Possibilities is 1.0

We know that you are either going to pull out a red marble or a white marble, so the probability of a white marble being pulled and the probability of a red marble being pulled must add up to 100%. Therefore, the odds of pulling out a red marble must be 99% or 0.99. If we let the event “Pull out a red marble” be given by the symbol R , we can say:

$$p(R) = 1.0 - P(W) = 1.0 - 0.01 = 0.99$$

Now, let’s say that I make you take a marble from the bag and then toss a coin. What is the probability that you will pull a white marble and then get heads on the coin? It is the product of the two probabilities: $0.01 \times 0.5 = 0.005$, so one-half of a one percent chance. Do the probabilities still sum to 1?

- White and Heads = $0.01 \times 0.5 = 0.005$
- White and Tails = $0.01 \times 0.5 = 0.005$
- Red and Heads = $0.99 \times 0.5 = 0.495$
- Red and Tails = $0.99 \times 0.5 = 0.495$

Yes, the probabilities of all the possibilities still add to 1.

107.2 Independence

In the last section, I told you that the probability of two events (“Pulling a red marble from the bag” and “Getting tails in a coin toss”) is the product of the probability of each event: $0.99 \times 0.5 = 0.495$.

This is true if the two events are *independent*, that is the outcome of one doesn't change the probability of the other. The example I gave is independent: It doesn't matter what ball you pull from the bag, the outcome of the coin toss will always be 50-50.

What are two events that are not independent? The probability that a person is a professional basketball player and the probability that someone wears a shoe that is size 13 or larger is *not* independent. After all, height is an advantage in basketball and most tall people also have large feet. So if you know someone is a basketball player, they likely wear large shoes.

Exercise 106 Rolling Dice

If I give you three dice to roll, what is the probability that you will roll a 5 on all three dice?

Working Space

Answer on Page 841

Exercise 107 Flipping Coins

If I give you five coins to flip, what is the probability that at least one coin will come up heads?

Working Space

Answer on Page 841

107.3 Why 7 is the most likely sum of two dice

If you roll two dice, the sum will be 2 or 12 or any number in between. It is very tempting to assume that the likelihood of any of those numbers is the same. In fact, the probability of a 2 is $\frac{1}{36} \approx 3\%$ and the probability of a 7 is $\frac{1}{6} \approx 17\%$. A 7 is six times more likely than a 12! Why?

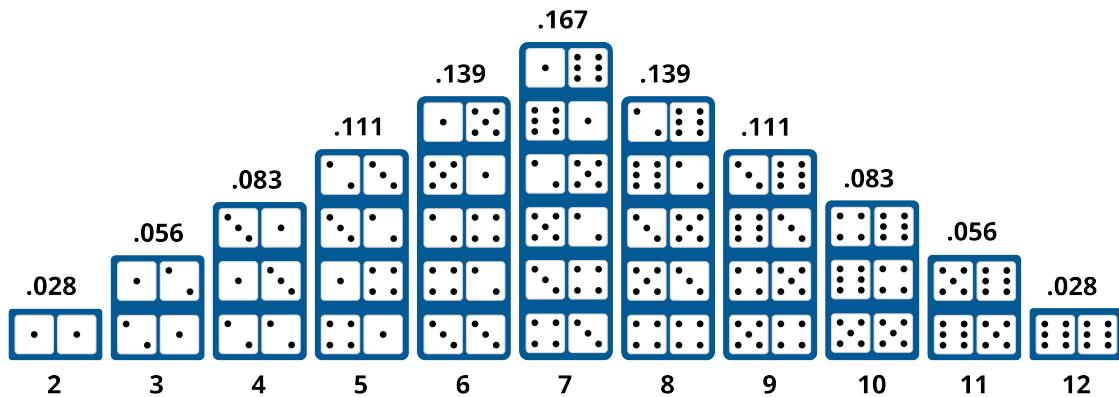
When you roll the first die, there are six possibilities with equal probability. When you roll the second die, there are six possibilities with equal probability. so there are a total of 36 possible events with equal probabilities: 1 then 1, 1 then 2, 2 then 1, 1 then 3, 3 then

1, etc. Only one of these (1 then 1) adds to 2. But six of these sum to 7: 1 then 6, 6 then 1, 2 then 5, 5 then 2, 3 then 4, 4 then 3. So a 7 is six times more likely than a 2.

Here is the complete table:

Sum							Count	Probability
2	1,1						1	1/36
3	1,2	2,1					2	1/18
4	1,3	2,2	3,1				3	1/12
5	1,4	2,3	3,2	4,1			4	1/9
6	1,5	2,4	3,3	4,2	5,1		5	5/36
7	1,6	2,5	3,4	4,3	5,2	6,1	6	1/6
8		2,6	3,5	4,4	5,3	6,2	5	5/36
9			3,6	4,5	5,4	6,3	4	1/9
10				4,6	5,5	6,4	3	1/22
11					5,6	6,5	2	1/18
12						6,6	1	1/36

When I bumped into this, I was skeptical. I decided to test it, so I rolled a pair of dice hundreds of times and made a histogram. It was a tedious and time-consuming task – just the sort of thing that we make computers do for us.



107.4 Random Numbers and Python

You are going to write a simulation of rolling dice in Python. To do this, you will need to generate a random sequence of numbers. The numbers will need to be in the range 1 to 6, and they will need to appear in the sequence with the same frequency. We say the sequence will follow *the uniform distribution*. That is, the probability is uniformly

distributed among the 6 possibilities.

Start python and try a few of the different ways to generate random numbers:

```
> python3
>>> import random
>>> random.random() # Generates a random floating point number between 0 and 1
0.6840892758539989
>>> randrange(5)      # Generates an integer in the range 0 - 4
2
>>> x = ['Rock', 'Paper', 'Scissors']
>>> random.choice(x) # Pick a random entry from the sequence
'Paper'
>>> x
['Rock', 'Paper', 'Scissors']
>>> random.shuffle(x) # Shuffle the order of the sequence
>>> x
['Scissors', 'Paper', 'Rock']
>>> a = list(range(30))
>>> a
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 29]
>>> random.sample(a, 10) # Return 10 randomly chosen items from the sequence
[8, 7, 20, 9, 25, 13, 23, 11, 14, 16]
```

Clearly, Python has a lot of ways to do things that look random. I should be honest with you at this point: they aren't really random. The computer that you are using can't generate random data. Instead, it uses tricks to create data that looks random; we call this *pseudorandom* data. Good pseudorandom algorithms are very important for cryptography and data security.

What if you want real random data? Some companies that are using the decay of radioactive materials to generate real random data. You can pay to download it. For our purposes, Python's pseudorandom numbers are quite sufficient.

If we generate two random numbers in the range 1 through 6 and add them together, we will have simulated rolling a pair of dice. Like this:

```
>>> a = random.randrange(6) + 1
>>> b = random.randrange(6) + 1
>>> a + b
8
```

First, let's write a program that just rolls the dice 100 times and shows the result. Make a file [dice.py](#):

```
import random

roll_count = 100

for i in range(roll_count):
    a = random.randrange(6) + 1
    b = random.randrange(6) + 1
    roll = a + b
    print(f"Toss {i}: {a} + {b} = {roll}")
```

When you run it, you should see something like:

```
> python3 dice.py
Toss 0: 6 + 6 = 12
Toss 1: 4 + 4 = 8
Toss 2: 4 + 2 = 6
Toss 3: 4 + 6 = 10
Toss 4: 4 + 4 = 8
...
Toss 98: 5 + 2 = 7
Toss 99: 5 + 2 = 7
```

Now we want to count occurrences of each possible outcome. Let's use an array of integers. We will start with an array of zeros. And, for example, when we roll a 3, we'll add 1 to item 3 in the array. (We can never roll a zero or a one, so those two entries will always be zero.)

```
import random

roll_count = 100

# Make an array containing 13 zeros
counts = [0] * 13

for i in range(roll_count):
    a = random.randrange(6) + 1
    b = random.randrange(6) + 1
    roll = a + b
    print(f"Toss {i}: {a} + {b} = {roll}")

    # Increment the count for roll
    counts[roll] += 1

print(f"Counts: {counts}")
```

When you run this, at the end you will see a count for each possible outcome :

```
...
Toss 98: 3 + 2 = 5
Toss 99: 6 + 1 = 7
Counts: [0, 0, 2, 6, 16, 11, 13, 14, 11, 11, 6, 9, 1]
```

What was the count that we expected? For example, we expected to see a 2 about once every 36 rolls, right? It might be nice to compare our count to what we expected. Add a few more lines, and we are going to increase the number of rolls. You will probably want to delete the line that prints each roll separately:

```
import random

# Can't ever be 0 or 1
p = [0.0, 0.0, 1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/36, 1/9, 1/12, 1/18, 1/36]
roll_count = 1000

# Make an array containing 13 zeros
counts = [0] * 13

for i in range(roll_count):
    a = random.randrange(6) + 1
    b = random.randrange(6) + 1
    roll = a + b

    # Increment the count for roll
    counts[roll] += 1

for i in range(2,13):
    print(f"{i} appeared {counts[i]} times, expected {p[i] * roll_count:.1f}")
```

Now you should see something like:

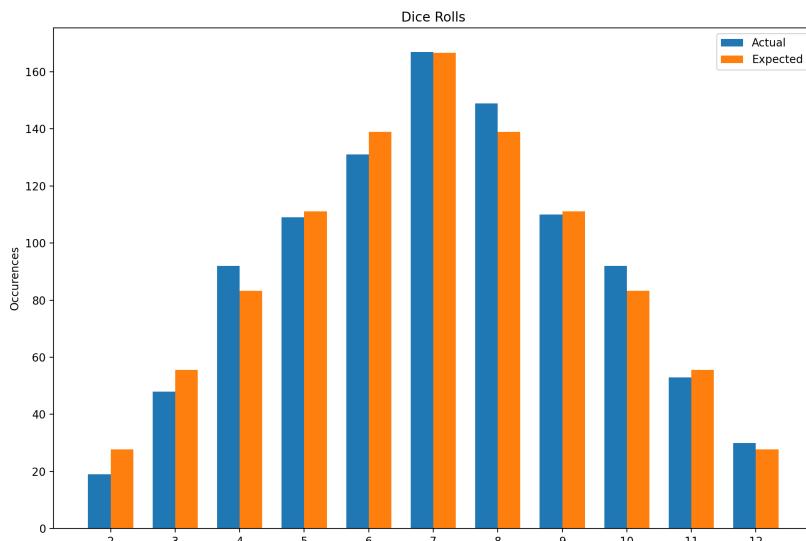
```
2 appeared 39 times, expected 27.8
3 appeared 55 times, expected 55.6
4 appeared 84 times, expected 83.3
5 appeared 110 times, expected 111.1
6 appeared 160 times, expected 138.9
7 appeared 176 times, expected 166.7
8 appeared 124 times, expected 138.9
9 appeared 93 times, expected 111.1
10 appeared 87 times, expected 83.3
```

```
11 appeared 49 times, expected 55.6
12 appeared 23 times, expected 27.8
```

Whenever you are dealing with random numbers, the outcome will seldom be *exactly* what you expected. In this case, however, you should see that your predictions are pretty close.

107.4.1 Making a bar graph

A bar graph is a nice way to look at quantities like this. Let's make a bar graph that shows the actual count and the expected count:



We need to describe the set of rectangles, to do this we will loop through each possible roll (2 - 12) and put data in four lists for each:

```
import random
import matplotlib.pyplot as plt

# Can't ever be 0 or 1
p = [0.0, 0.0, 1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/36, 1/9, 1/12, 1/18, 1/36]
roll_count = 1000

# Make an array containing 13 zeros
counts = [0] * 13
```

```
for i in range(roll_count):
    a = random.randrange(6) + 1
    b = random.randrange(6) + 1
    roll = a + b

    # Increment the count for roll
    counts[roll] += 1

# Gather data for bar chart
bar_width = 0.35
expected = []
actual_starts = []
expected_starts = []
labels = []
actual = []
for i in range(2,13):
    expected.append(p[i] * roll_count)
    actual.append(counts[i])
    actual_starts.append(i - bar_width/2)
    expected_starts.append(i + bar_width/2)
    labels.append(i)

fig, ax = plt.subplots()

# Create the bars
ax.bar(actual_starts, actual, bar_width, label='Actual')
ax.bar(expected_starts, expected, bar_width, label='Expected')
ax.set_xticks(labels)

# Provide labels
ax.set_ylabel('Occurrences')
ax.set_title('Dice Rolls')
ax.legend()
plt.show()
```




CHAPTER 108

Beginning Combinatorics

Discrete probability problems often include some counting. For example, we figured out that there were 36 different ways the two dice, but all of them summed to some number 2 through 12. How many different ways could three 8-sided dice come up? We would need to count them, right? As the numbers get big we will need some tricks so we don't need to write them all down and count them one-by-one.

The branch of mathematics that focuses on tricks for counting is called *combinatorics*.

How can we be sure that there were 36 different configurations for the two 6-sided dice? The first die could have come up as any one of six numbers. For each of those, the second could have come up with any one of six numbers. Thus, the number of possibilities is $6 \times 6 = 36$.

How many different configurations for 3 8-sided dice? $8 \times 8 \times 8 = 8^3 = 512$.

What about seven dice, each with 20 sides? There would be $20^7 = 1,280,000,000$ configurations. See, aren't you glad we don't need to write them all down?

Now, let's say that six people (Anne, Brock, Carl, Dev, Edgar, and Fred) are going to run

a race. You have to make a plaque that says who won first place, who won second place, and who won third. If you want to get all the possible plaques created beforehand, and just pull the right one out as soon as the race ends, how many plaques would you need to get engraved?

In this case, once someone has been given first place, they can't win second or third place. Thus, any of the 6 people can come in first, but once you have engraved that person's name on the plaque, there are only 5 people whose names can appear in second place. Once you have engraved that name, there are only 4 people whose names can appear in third place. Thus, you would get $6 \times 5 \times 4 = 120$ plaques engraved.

What if the plaque includes all 6 places? Then you would need $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ plaques engraved. We use this process often enough that we gave it a name. We say "I need 6 factorial plaques engraved." When we write a factorial, we use an exclamation point:

$$6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$$

We use the word "permutation" to mean a particular ordering. This rule says n items can be ordered in $n!$ ways. Thus mathematicians actually say "If you have a list of n items then we can generate $n!$ different permutations of those items".

In Python, there is a `factorial` function in the `math` library:

```
> python3
>>> import math
>>> math.factorial(6)
720
```

Handy, right? Now you don't need to write a loop to calculate factorials.

Remember when we only wanted the first three names on the plaque? We can do that problem using factorials:

$$6 \times 5 \times 4 = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{6!}{3!}$$

This formulation makes it easy to figure out on any calculator with a "!" button.

The rule on this is to fill m positions from n items, it can be done this many ways:

$$\frac{n!}{(n-m)!}$$

108.0.1 Choose

Let's say that there are 12 kids in a classroom, and you need a team of 4 to wipe down the desks. How many different possible teams are there? You know that if you were giving out four different positions (Like the race gave out 1st, 2nd, and 3rd), the answer would be $12 \times 11 \times 10$ or $12!/(12 - 4)!$.

However, once we pick the 4 people, we don't care what order they are in, right? In this problem, the team "Anne, Brad, Carl, and Don" is the same as the team "Carl, Don, Brad, and Anne".

Thus, the quantity $12!/(12 - 4)!$ is many times too large because it counts each permutation separately. To get the right number, we just divide this by the number of possible permutations for a group of four people: $4!$

That gets us our answer: How many different teams of four can be chosen from 12 people?

$$\frac{12!}{(12 - 4)!4!} = 495$$

In combinatorics, we use this quantity a lot, so we have given it a name: *choose*

We have also given it a notation. "12 choose 4" is written like this:

$$\binom{12}{4}$$

Python has the `math.comb` function:

```
> python3
>>> import math
>>> comb(12, 4)
495
```




CHAPTER 109

Permutations and Sorting

In the previous chapter, we talked about permutations. If you have a list of three letters, like [a, b, c, d], you can rearrange them in $4!$ ways:

a,b,c,d	a,b,d,c	a, d, b, c	a, d, c, b	a, c, b, d	a, c, d, b
b,a,c,d	b,a,d,c	b, d, a, c	b, d, c, a	b, c, a, d	b, c, d, a
c,b,a,d	c,b,d,a	c, d, b, a	c, d, a, b	c, a, b, d	c, a, d, b
d,b,c,a	d,b,a,c	d, a, b, c	d, a, c, b	d, c, b, a	d, c, a, b

You can make Python generate all the permutations for you:

```
from itertools import permutations
all_permutations = permutations(['a', 'b', 'c', 'd'])
for p in all_permutations:
    print(p)
```

109.1 Notation

How do we define or write down a single permutation? You could say something like “Swap the first and second items and swap the third and fourth items.” However, that gets pretty difficult to read. So we usually write a permutation as two lines: the first line is before the permutation and the second line is after. Like this:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

And we can assign permutations to variables. For example, if we wanted the variable A to represent “swapping the first and second item”, we would write this:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix}$$

And if we wanted B to represent “swapping the third and fourth item”, we would write:

$$B = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix}$$

Now, we can *compose* permutations together. For example, we might say:

$$B \circ A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

That is if we have the list [a, b, c, d] and we apply permutation A and then permutation B, we get [b, a, d, c].

Important: Note that permutations are applied from right to left. $B \circ A$ means “Applying A and then B.” Why does this matter? Permutations are not necessarily commutative. That is, if you have two permutations S and T, $S \circ T$ is not always the same as $T \circ S$.

Also, note that “don’t change anything” is a permutation. We call it *the identity permutation*. If you have four items, the identity permutation would be written:

$$I = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

(We use a capital “I” for the identity.)

109.1.1 Challenge

Find an example of two permutations S and T such that $S \circ T$ does not equal $T \circ S$.

109.2 Sorting in Python

One of the common forms of permutation in software is sorting. Sorting is putting data in a particular order. For example, in Python, if you had a list of numbers, you can sort it in ascending order like this:

```
my_grades = [92, 87, 76, 99, 91, 93]
grades_worst_to_best = sorted(my_grades)
```

Do you want to sort backwards?

```
my_grades = [92, 87, 76, 99, 91, 93]
grades_best_to_worst = sorted(my_grades, reverse=True)
```

Note that `sorted` makes a new list with the correct order. If you want to sort the array in place, you can use the `sort` method:

```
my_grades = [92, 87, 76, 99, 91, 93]
my_grades.sort(reverse=True)
```

109.3 Inverses

Think for a second about this permutation:

$$S = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

You could say this permutation shuffles a list a bit. What is its inverse? That is, what is the permutation that unshuffles the items back to where they were originally?

$$S^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}$$

That is, the original moved an item in the first spot to the third spot. The inverse must move whatever was in the third spot back to the first spot.

(Notation note: Because in multiplication, $b \times b^{-1} = 1$, we use “to the negative one” to indicate inverses in lots of places.)

Mechanically, how do you find the inverse? Flip the rows, and then sort the columns using the top number:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix} \text{ flip } \rightarrow \begin{pmatrix} 3 & 4 & 2 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \text{ sort } \rightarrow \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}$$

Let’s say you have two permutations A and B. Permuting by B and then A would look like this:

$$C = A \circ B$$

If you know A^{-1} and B^{-1} , what is C^{-1} ? You would undo-A and then undo-B, so

$$C^{-1} = B^{-1} \circ A^{-1}$$

109.4 Cycles

Here is a permutation:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{pmatrix}$$

When this is applied, whatever is at 1 gets moved to 2, 2 gets moved to 4, and 4 gets moved to 1. That is a *cycle*: $1 \rightarrow 2 \rightarrow 4$ and then it goes back to 1. It involves three locations, so we say it is a *3-cycle*.

There is another cycle in this permutation: $3 \rightarrow 5$ and then it goes back to 3.

Because these cycles share no members, we say the cycles are *disjoint*.

Every permutation can be broken down into a collection of disjoint cycles.

$$T = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{pmatrix} = (1 \rightarrow 2 \rightarrow 4)(3 \rightarrow 5)$$

The first handy thing about this notation is that it makes it easy for us to describe the inverse: we just run the cycles backward:

$$T^{-1} = (4 \rightarrow 2 \rightarrow 1)(5 \rightarrow 3)$$

Starting with the list [a, b, c, d, e], lets repeatedly apply the permutation T

Initial	a, b, c, d, e
T applied	d, a, e, b, c
T \circ T applied	b, d, c, a, e
T \circ T \circ T applied	a, b, e, d, c
T \circ T \circ T \circ T applied	d, a, c, b, e
T \circ T \circ T \circ T \circ T applied	b, d, e, a, c
T \circ T \circ T \circ T \circ T applied	a, b, c, d, e

This permutation, results in six combinations, and then it loops back on itself. The number of combinations is the least common multiple of all the cycles. In this case, there is a 3-cycle and a 2-cycle. The least common multiple of 2 and 3 is 6.



CHAPTER 110

Conditional Probability

Let's say there is a virus going around, and there is a vaccine for it that requires 2 shots. You are working at a school, and you are wondering how effective the vaccines are. Some students are unvaccinated, some have had one shot, and some have had two shots. One day you test all 644 students to see who has the virus. You end up with the following

table:

	V_0	V_1	V_2
T_+	88 students	36 students	96 students
T_-	92 students	76 students	256 students

Here are what the symbols mean:

- V_0 : student has had zero vaccination shots
- V_1 : student has had one vaccination shot
- V_2 : student has had both vaccination shots
- T_+ : student tested positive for the virus
- T_- : student tested negative for the virus

So, for example, your data indicates that 76 students who had only one of the two shots and tested negative for the virus.

Your principal has a few questions. The first is “If I put five randomly chosen students in a study group together, what is the probability that one of them has the virus?”

The first thing you might do is make a new table that shows what is the probability of a randomly chosen student being in any particular group. You just divide each entry by 644 (the total number of students).

	V_0	V_1	V_2
T_+	$p(V_0 \text{ AND } T_+) = 13.7\%$	$p(V_1 \text{ AND } T_+) = 5.6\%$	$p(V_2 \text{ AND } T_+) = 14.9\%$
T_-	$p(V_0 \text{ AND } T_-) = 14.3\%$	$p(V_1 \text{ AND } T_-) = 11.8\%$	$p(V_2 \text{ AND } T_-) = 39.8\%$

(In this table, I expressed the number as a percentage with a decimal point – you had to round off the numbers. If you wanted exact answers, you would have to keep each as a fraction: 36 students represents $\frac{9}{161}$ of the student body.)

110.1 Marginalization

Now we can sum across the columns and rows.

	V_0	V_1	V_2	sum
T_+	0.137	0.056	0.149	$p(T_+) = 0.342$
T_-	0.143	0.118	0.398	$p(T_-) = 0.547$
sum	$p(V_0) = 0.280$	$p(V_1) = 0.174$	$p(V_2) = 0.547$	

If a child is chosen randomly from the entire student body, there is a 34.2% that the student has tested positive for the virus. And there is 17.4% chance that the student has one shot of the vaccine.

This summing of the probabilities across one dimension is known as *marginalizing*. Marginalization is just summing across all the variables that you don’t care about. You don’t care who has the virus, just the probability that a student has not received even one shot of the vaccine? You marginalize all the vaccine statuses.

To answer the principal’s question, the easy thing to do is find the answer of the opposite “if I put five randomly chosen students in a study group together, what is the probability that *none* of them has tested positive for the virus?”

The chance that a randomly chosen student doesn’t have the virus ($p(T_-)$) is 54.7%. Thus the chance that 5 randomly chosen students don’t have the virus is $0.547 \times 0.547 \times 0.547 \times 0.547 \times 0.547 = 0.0489$ Thus the probability of the opposite is $1.0 - 0.0489 = 0.951$

The answer, then, is “If you put 5 kids in a study group together, there is a 95.1 % proba-

bility that at least one of them has the virus."

110.2 Conditional Probability

Now the principal asks you, "What if I make a group of 5 kids who have had both shots of the vaccine? What are the odds that one of them has tested positive for the virus?"

This involves the idea of *Conditional probability*. You want to know the odds that a student doesn't have the virus given that the student has had both shots of the vaccine.

There is a mathematical notation for this:

$$p(T_-|V_2)$$

That is the probability that a student who has had both vaccination shots will test negative for the virus.

How would you calculate this? You would count all the students who had a positive test *and* both vaccination shots, which you would divide by the total number of students who had both vaccination shots.

$$p(T_-|V_2) = \frac{256}{96 + 256} = \frac{8}{11} \approx 72.7\%$$

If we are working from the probabilities, you can get the same result this way: Divide the probability that a randomly chosen student had a positive test *and* both vaccination shots by the probability that a student had both vaccination shots:

$$p(T_-|V_2) = \frac{p(T_- \text{ AND } V_2)}{p(T_-)} = \frac{0.398}{0.547} \approx 72.7\%$$

Notice that this is different from $p(V_2|T_-)$, which is the probability that a student has had both vaccinations, given they tested negative for the virus.

Back to the principal's question: "If you have 5 students who have had both vaccinations, what is the probability that all of them tested negative for the virus?" The probability that one student is virus-free is $\frac{8}{11}$, so the probability that 5 students are virus-free is $\left(\frac{8}{11}\right)^5 \approx 0.203$. So, there is a 79.6% chance that at least one of the five has the virus.

110.3 Chain Rule for Probability

You just used this equality: For any events A and B

$$p(A|B) = \frac{p(A \text{ AND } B)}{p(B)}$$

This is more commonly written like this:

$$p(A \text{ AND } B) = \frac{p(A|B)}{p(B)}$$

This is an abstract way of writing the idea, but the idea itself is pretty intuitive: The probability that I'm going to buy a ticket and win the lottery is equal to the probability that I buy a ticket times the probability that I win, given that I have bought a ticket. (Here A is "win the lottery" and B is "buy a ticket".)

This is known as *The Chain Rule of Probability*. And we can chain together as many events as we want: The probability that you are going to die in the car that you bought with your winnings from the lottery ticket you bought is:

$$p(W \text{ AND } X \text{ AND } Y \text{ AND } Z) = p(W|X \text{ AND } Y \text{ AND } Z)p(X|Y \text{ AND } Z)p(Y|Z)p(Z)$$

where

- W = Dying in car accident
- X = Buying a car with lottery winnings
- Y = Winning the lottery
- Z = Buying a lottery ticket

In English, then, the equation says:

"The probability that you will die in a car accident, buy a car with lottery winnings, win the lottery, and buy a lottery ticket is equal to the probability that you buy a lottery ticket times the probability that you win the lottery (given that you have bought a ticket) times the probability that buy a car with those lottery winnings (given that bought a ticket and won) times the probability that you crash that car (given that you have bought the car, won the lottery, and bought a ticket)."



CHAPTER III

Bayes' Theorem

Let's say that you are holding two bags of marbles. You know that one bag contains 60 white marbles and 40 red marbles. And you know that the other holds 10 white marbles and 90 red marbles. You don't know which is which – and you can't see the marbles.

I say "Guess which bag is mostly red marbles." You pick one.

"What is the probability that this is the bag that is mostly red marbles?" You think "50 percent and there is also a 50 percent probability that it is the mostly-white-marbles bag."

Then you pick one marble from the bag. It is red. Now you must update your beliefs. It is more likely that this is the mostly-red-marbles bag. What is the probability now?

Bayes Theorem gives you the rule for updating your beliefs based on new data.

111.1 Bayes Theorem

Let's say you have two events or conditions C and D. C is "The person has a cough" and D is "The person is waiting to see a doctor."

Using the chain rule of probability, we now have two ways to calculate $p(C \text{ AND } D)$:

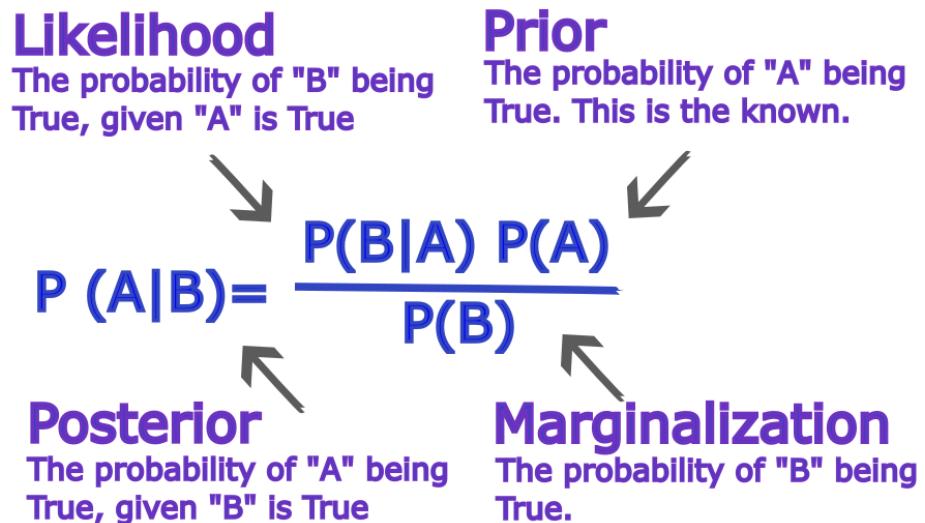
$$p(C \text{ AND } D) = p(C|D)p(D)$$

(The probability the person is at the doctor multiplied by the probability they have a cough if they are at the doctor.)

or

$$p(C \text{ AND } D) = p(C)p(D|C)$$

(The probability the person has a cough multiplied by the probability they are at the doc-



tor if they have a cough.)

Thus:

$$p(D|C) = \frac{p(C|D)p(D)}{P(C)}$$

Now you can calculate $p(D|C)$ (in this case, the probability that you are waiting to see a doctor given that you have a cough.) if you know:

- $p(C|D)$ (The probability that you have a cough given that you are waiting to see a doctor)
- $p(D)$ (The probability that you are waiting for a doctor for any reason.)
- $p(C)$ (The probability that you have a cough anywhere)

Pretty much all modern statistical methods (including most artificial intelligence) are based on this formula, which is known as Bayes' Theorem. It was written down by Thomas Bayes before he died in 1761. It was then found and published after his death.

111.2 Using Bayes' Theorem

Back to the example at the beginning. To review:

- There are two bags that look exactly the same.
- Bag W has 60 white marbles and 40 red marbles.
- Bag R has 10 white marbles and 90 red marbles.
- You pull one marble from the selected bag – it is red.

What is the probability that the selected bag is Bag R? Intuitively, you know that the probability is now more than 0.5. What is the exact number?

In terms of conditional probability, we say we are looking for “the probability that the selected bag is Bag R, given that you drew a red marble?” or $p(B_R|D_R)$, where B_R is “The selected bag is Bag R” and D_R is “You drew a red marble from the selected bag”.

From Bayes' Theorem, we can write:

$$p(B_R|D_R) = \frac{P(D_R|B_R)p(B_R)}{P(D_R)}$$

$P(D_R|B_R)$ is just the probability of drawing a red marble given that the selected bag is Bag R. That is easy to calculate: There are 100 marbles in the bag, and 90 are red. Thus $P(D_R|B_R) = 0.9$.

$P(B_R)$ is just the probability that you chose Bag R before you drew out a marble. Both bags look the same, so $P(B_R) = 0.5$. This is called *the prior* because it represents what you thought the probability was before you got more information.

$P(D_R)$ is the probability of drawing a red marble. There was 0.5 probability that you put your hand into Bag W (in which 40 of the 100 marbles are red) and a 0.5 probability that you put your hand into Bag R (in which 90 of the 100 marbles are red). So

$$P(D_R) = 0.5 \frac{40}{100} + 0.5 \frac{90}{100} = 0.65$$

Putting it together:

$$p(B_R|D_R) = \frac{P(D_R|B_R)P(B_R)}{P(D_R)} = \frac{(0.9)(0.5)}{0.65} = \frac{9}{13} \approx 0.69$$

Thus, given that you have pulled a red marble, there is about a 69% chance that you have selected the bag with 90 red marbles.

111.3 Confidence

Bayes' Theorem, then, is about updating your beliefs based on evidence. Before you drew out the red marble, you selected one bag thinking it might contain 90 red marbles. How certain were you? 0.0 being complete disbelief and 1.0 entirely confidence, you were 0.5. After pulling out the red marble, you were about 0.69 confident that you had chosen the bag with 90 red marbles.

The question "How confident are you in your guess?" is very important in some situations. For example in medicine, diagnoses often lead to risky interventions. Few diagnoses come with 100% confidence. All doctors should know how to use Bayes' Theorem.

And in a trial, a jury is asked to determine if the accused person is guilty of a crime. Few jurors are ever 100% certain. In some trials, Bayes' Theorem is a really important tool.



CHAPTER 112

Definite Integrals

Integrals are a fundamental concept in calculus, which are used to calculate areas, volumes, and many other things. A definite integral calculates the net area between the function and the x-axis over a given interval.

112.1 Definition

The definite integral of a function $f(x)$ over an interval $[a, b]$ is defined as the limit of a Riemann sum:

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*) \Delta x \quad (112.1)$$

where x_i^* is a sample point in the i^{th} subinterval of a partition of $[a, b]$, $\Delta x = \frac{b-a}{n}$ is the width of each subinterval, and the limit is taken as the number of subintervals n approaches infinity.



CHAPTER 113

Antiderivatives

In your study of calculus, you have learned about derivatives, which allow us to find the rate of change of a function at any given point. Derivatives are powerful tools that help us analyze the behavior of functions. Now, we will explore another concept called antiderivatives, which are closely related to derivatives.

An antiderivative, also known as an integral or primitive, is the reverse process of differentiation. It involves finding a function whose derivative is equal to a given function. In simple terms, if you have a function and you want to find another function that, when differentiated, gives you the original function back, you are looking for its antiderivative.

The symbol used to represent an antiderivative is \int . It is called the integral sign. For example, if $f(x)$ is a function, then the antiderivative of $f(x)$ with respect to x is denoted as $\int f(x), dx$. The dx at the end indicates that we are integrating with respect to x .

Finding antiderivatives requires using specific techniques and rules. Some common antiderivative rules include:

- The power rule: If $f(x) = x^n$, where n is any real number except -1 , then the antiderivative of $f(x)$ is given by $\int f(x), dx = \frac{1}{n+1}x^{n+1} + C$, where C is the constant

of integration.

- The constant rule: The antiderivative of a constant function is equal to the constant times x . For example, if $f(x) = 5$, then $\int f(x), dx = 5x + C$.
- The sum and difference rule: If $f(x)$ and $g(x)$ are functions, then $\int(f(x) + g(x)), dx = \int f(x), dx + \int g(x), dx$. Similarly, $\int(f(x) - g(x)), dx = \int f(x), dx - \int g(x), dx$.

Antiderivatives have various applications in mathematics and science. They allow us to calculate the total accumulation of a quantity over a given interval, compute areas under curves, and solve differential equations, among other things.

It is important to note that an antiderivative is not a unique function. Since the derivative of a constant is zero, any constant added to an antiderivative will still be an antiderivative of the original function. This is why we include the constant of integration, denoted by C , in the antiderivative expression.

In summary, antiderivatives are the reverse process of differentiation. They help us find functions whose derivatives match a given function. Understanding antiderivatives is crucial for various advanced calculus concepts and real-world applications.

Now, let's explore different techniques and methods for finding antiderivatives and discover how they can be applied in solving problems.



CHAPTER 114

The Fundamental Theorem of Calculus

The Fundamental Theorem of Calculus is a theorem that connects the concept of differentiating a function with the concept of integrating a function. This theorem is divided into two parts:

114.1 First Part

The first part of the Fundamental Theorem of Calculus states that if f is a continuous real-valued function defined on a closed interval $[a, b]$ and F is the function defined, for all x in $[a, b]$, by:

$$F(x) = \int_a^x f(t) dt \quad (114.1)$$

Then, F is uniformly continuous and differentiable on the open interval (a, b) , and $F'(x) =$

$f(x)$ for all x in (a, b) .

114.2 Second Part

The second part of the Fundamental Theorem of Calculus states that if f is a real-valued function defined on a closed interval $[a, b]$ that admits an antiderivative F on $[a, b]$, and f is integrable on $[a, b]$ (it need not be continuous), then

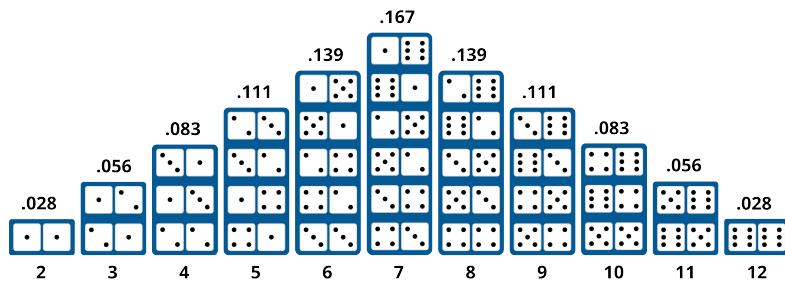
$$\int_a^b f(t) dt = F(b) - F(a). \quad (114.2)$$



CHAPTER 115

Continuous Probability Distributions

When we talked about the probability distribution of the sum of two dice, we assigned a probability to each of the 11 possibilities:



The probabilities all added up to be 1.0. That is a way of saying "100% of the times you throw the dice, the sum will be an integer between 2 and 12."

Now we need to talk about probabilities of properties that are continuous, not discrete. For example, we might want to ask the question "If I randomly pick a cow from all the cows in the world, what is the probability that it will weigh less than 597.34 kg?" What does a probability distribution for a continuous variable look like?

115.1 Cumulative Distribution Function

Imagine that you live in ancient times. You buy, sell, and ship cows. A lot of people come into your office and brag about their cows: "Bessie is heavier than 99% of the cows in world!" So you need to develop some statistics on cow weights.

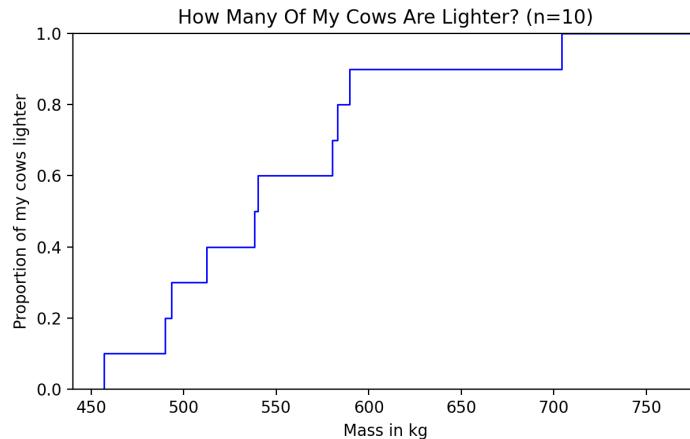
You have ten cows. You weight them:

Cow	Mass in kg
Cow 1	580.22
Cow 2	540.07
Cow 3	538.20
Cow 4	512.39
Cow 5	589.75
Cow 6	456.91
Cow 7	583.09
Cow 8	493.56
Cow 9	489.97
Cow 10	704.15

If someone comes into your shop and says "My Bessie is an astonishing 530 kg!" it would be cool to have a list on the wall that would let you yell back "Half my cows are heavier than that, Silly!" So you sort the cows by weight. For each weight, you say how many of your cows are lighter than that:

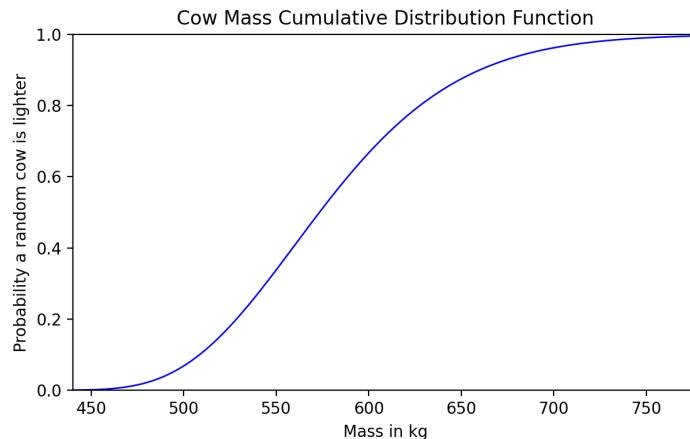
Proportion of cows lighter	Mass in kg
0.00	456.91
0.10	489.97
0.20	493.56
0.30	512.39
0.40	538.20
0.50	540.07
0.60	580.22
0.70	583.09
0.80	589.75
0.90	704.15

In fact, for easy reference, you make a plot:



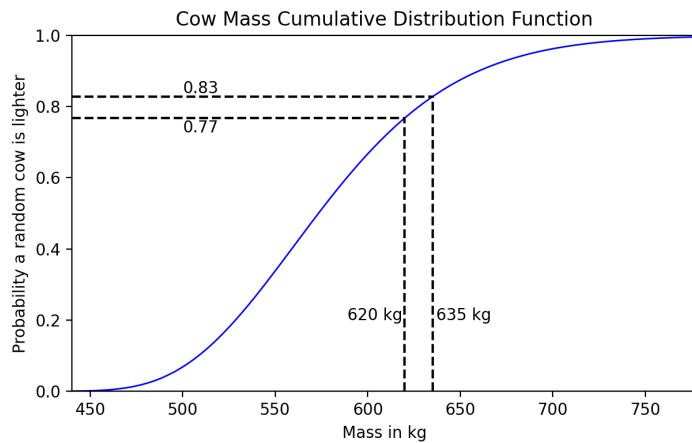
Now for any weight you can quickly look up what proportion of your cows are lighter. (And, if you subtract that from 1, what proportion of your cows are heavier.)

See how jagged that graph is? That is because you only have the data for 10 cows. However, as the years pass and you weight thousands of cows, the plot will become smoother. Because it always accumulates more cows as you move from left to right, this is known as a *cumulative distribution function* or CDF:



A cumulative distribution function always starts at 0 and ends at 1. On that journey, it never decreases.

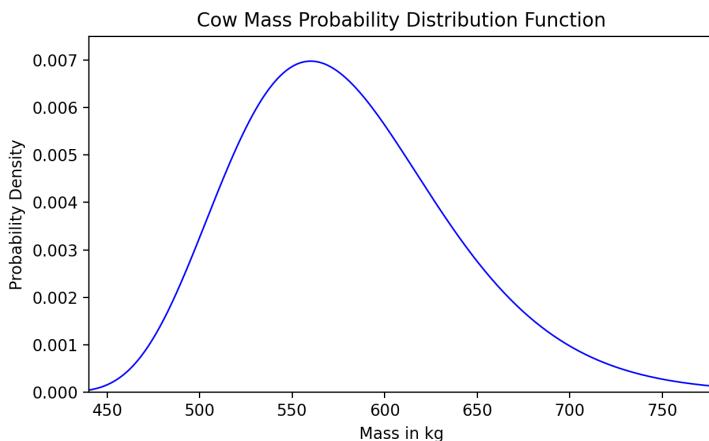
Let's say you want to know what proportion of cows weigh between 620 kg and 635 kg. Using the CDF, you could figure out that 77% of all cows weigh less than 620 kg and 83% of all cows weigh less than 635 kg. Thus 6% of all cows must weigh more than 620 kg and less than 635 kg.



115.2 Probability Density Function

The cumulative density function is handy, but some of its information can be hard to see. For example, how would you answer the question "What is the most common weight of a cow?" You would squint at the CDF and try to determine where it was steepest.

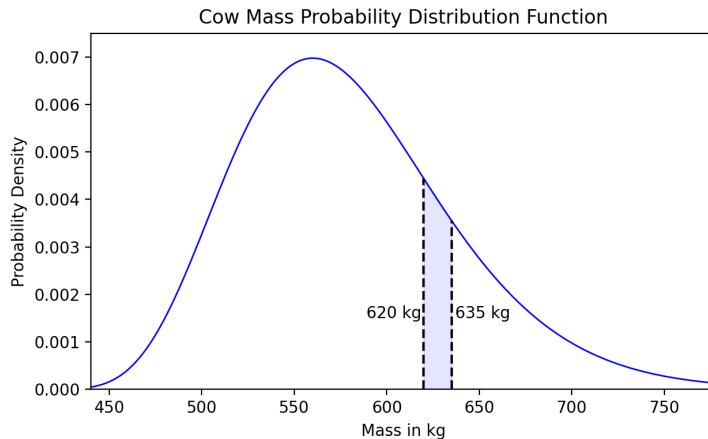
To make these sorts of questions easier to answer, we take the derivative of the CDF to get the *probability density function* (or PDF). For the cows, it would look like this:



Now you can easily see that the CDF was steepest at about 560 kg. We call the highest point on the PDF the *most likely estimator*. For example, you might say "560 kg is the most likely estimator of cow mass." Sometimes we just say "the MLE".

Note that the MLE is often different from the mean or the median. In this case, for example, the distribution is skewed right – there are more cows that are heavier than the MLE than there are cows that are lighter than the MLE. The MLE would be less than the mean or the median.

Once again, let's say you want to know what proportion of cows weight between 620 kg and 635 kg. This is more difficult with a PDF than it is with a CDF. With a PDF, you have to find the area under the curve between $x = 620$ and $x = 635$.

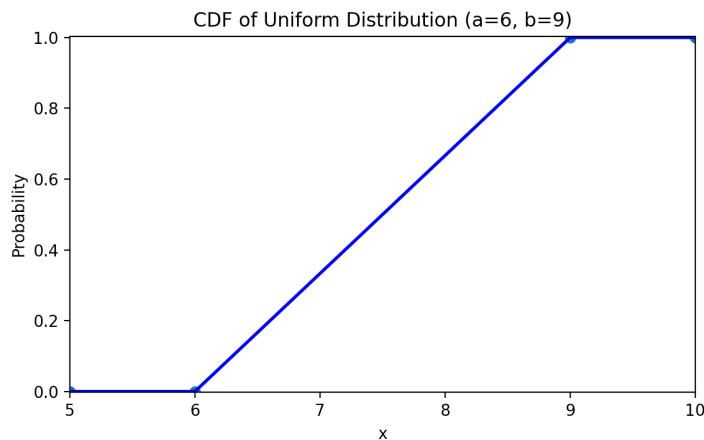


This is why it is called a "probability density" – to get a true probability you need to multiply the density by the width of the region.

What is the area under the entire curve? If you integrated it, you would get the CDF. The CDF goes from 0 to 1.0. The area under a PDF is *always* 1.0.

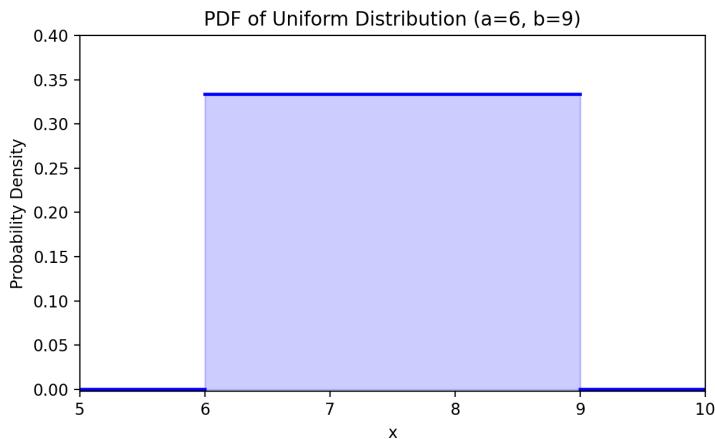
115.3 The Continuous Uniform Distribution

The most simple continuous distribution is the uniform distribution between two numbers a and b . The CDF is a straight line from zero at a to 1 at b . For example, here is the CDF for the uniform distribution between 6 and 9.



That line goes from 0 to 1 over a distance of 3, so its slope is $\frac{1}{3}$ between 6 and 9 and zero

everywhere else. Thus the PDF (its derivative), looks like this:



So we can write the probability distribution of a continuous uniform distribution between a and b as:

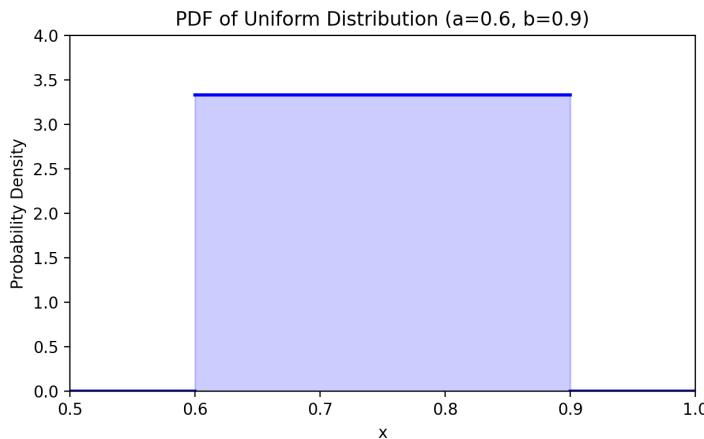
$$p(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b. \end{cases}$$

Notice that if a and b are less than 1 apart, the value of $p(x)$ will be greater than 1. This is a really important difference between a probability and a probability density:

- A probability will always be in the interval $[0, 1]$.
 - A probability density will never be less than 0, but can be much larger than 1.

That said, the probability density will always integrate to 1.

Here is the PDF for a uniform distribution between 0.6 and 0.9:



The mean and median of a uniform distribution between a and b is its midpoint: $\frac{a+b}{2}$.

The variance (σ^2) is $\frac{(b-a)^2}{12}$.

115.4 Continuous Distributions In Python

The SciPy library has functions that let a programmer work with a large collection of different probability distributions.

For example, if you wanted to work with a continuous uniform distribution between 6 and 9, you would import the relevant functions like this:

```
from scipy.stats import uniform
```

Now if you wanted a numpy array containing a sample of 300 numbers generated randomly from that distribution:

```
samples = uniform.rvs(loc=6, scale=3, size=300)
```

The `loc` argument is a . The `scale` argument is $b - a$.

If you wanted to know the value of the probability density function at 8 and 10, you could use the `pdf` function:

```
x_values = np.array([8, 10])
p_values = uniform.pdf(x_values, loc=6, scale=3)
```

Now `p_values` contains 0.33333 and 0.0.

To get the value of the cumulative distribution function at those points, you would use the `cdf` function:

```
cdf_values = uniform.cdf(x_values, loc=6, scale=3)
```

Now `cdf_values` contains 0.666667 and 1.0.

The inverse of the CDF is very useful. It answers questions like "How heavy does a cow have to be to be in top 1%?"

```
bottom_top_percentiles = np.array([0.01, 0.99])
boundaries = uniform.ppf(bottom_top_percentiles, loc=6, scale=3)
```

Now `boundaries` contains 6.03 and 8.97.

The SciPy library supplies these functions (`rvs`, `pdf`, `cdf`, and `ppf`) for over a hundred common continuous probability distributions.

The common "bell curve" shaped distribution is called a Gaussian or Normal distribution. It is described by its mean (the midpoint of the bell) and its standard deviation. For the normal distribution, the standard deviation is the distance you have to go from the mean to reach 68% of the population. We will talk a lot more about the normal distribution in other chapters, but lets take this opportunity to plot the CDF and PDF of a normal distribution with a mean of 32 and a standard deviation of 8.

Create a file called `plot_norm.py` and add the following lines:

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

# Constants
MEAN = 32
STD = 8

# Plotting from the 0.5 percentile to the 99.5 percentile
x_min = norm.ppf(0.005, loc=MEAN, scale=STD)
x_max = norm.ppf(0.995, loc=MEAN, scale=STD)

# Make 200 points between x_min and x_max
x_values = np.linspace(x_min, x_max, 200)

# Get CDF for each x value
```

```
cdf_values = norm.cdf(x_values, loc=MEAN, scale=STD)

# Get PDF for each x value
pdf_values = norm.pdf(x_values, loc=MEAN, scale=STD)

# What is the highest density?
max_density = norm.pdf(MEAN, loc=MEAN, scale=STD)

# Make a figure with two axes
fig, axs = plt.subplots(nrows=2, sharex=True, figsize=(8, 8), dpi=200)
axs[0].set_xlim(left=x_min, right=x_max)

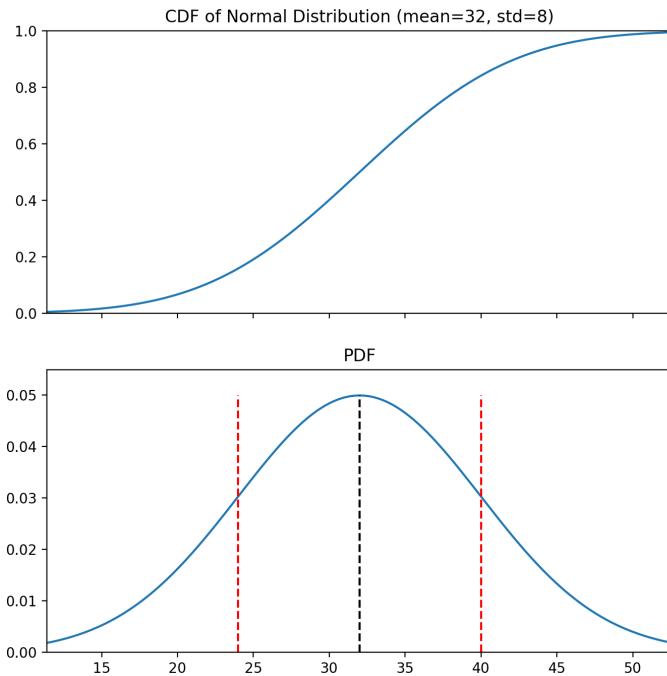
# Draw the CDF on the first axis
axs[0].set_title("CDF of Normal Distribution (mean=32, std=8)")
axs[0].set_ylim(bottom=0.0, top=1.0)
axs[0].plot(x_values, cdf_values)

# Draw the PDF on the second axix
axs[1].set_title("PDF")
axs[1].set_ylim(bottom=0.0, top=max_density * 1.1)
axs[1].plot(x_values, pdf_values)

# Add lines for mean, mean-std, and mean+std
axs[1].vlines(MEAN - STD, 0, max_density, "r", linestyle="dashed")
axs[1].vlines(MEAN + STD, 0, max_density, "r", linestyle="dashed")

# Save out the figure
fig.savefig("norm_32_8.png")
```

The resulting plot should look like this:



What do those vertical lines mean? An ornithologist might tell you "The wingspan of adult robins are normally distributed with a mean of 32 cm and a standard deviation of 8 cm." Then 68% of the population of adult robins would have wingspans between the two red lines.

Exercise 108 **SciPy Stats***Working Space*

Globally, the height of adult women is approximately normally distributed. The mean is 164.7 cm. The standard deviation is 7.1 cm.

Use python and SciPy stats to answer these questions:

- To be in the tallest decile (the top 10%) of adult women, how tall does one need to be?
- What percentage of adult women are between 160 cm and 165 cm?

(In case you are wondering: For men the mean is 178.4 cm and the standard deviation is 7.6 cm.)

Answer on Page 842



CHAPTER 116

The Physics of Gases

Now, let's say you start to heat the helium inside the balloon. As the temperature goes up, the molecules inside will start to move faster.

Remember that the kinetic energy of an object with mass m and velocity v is given by

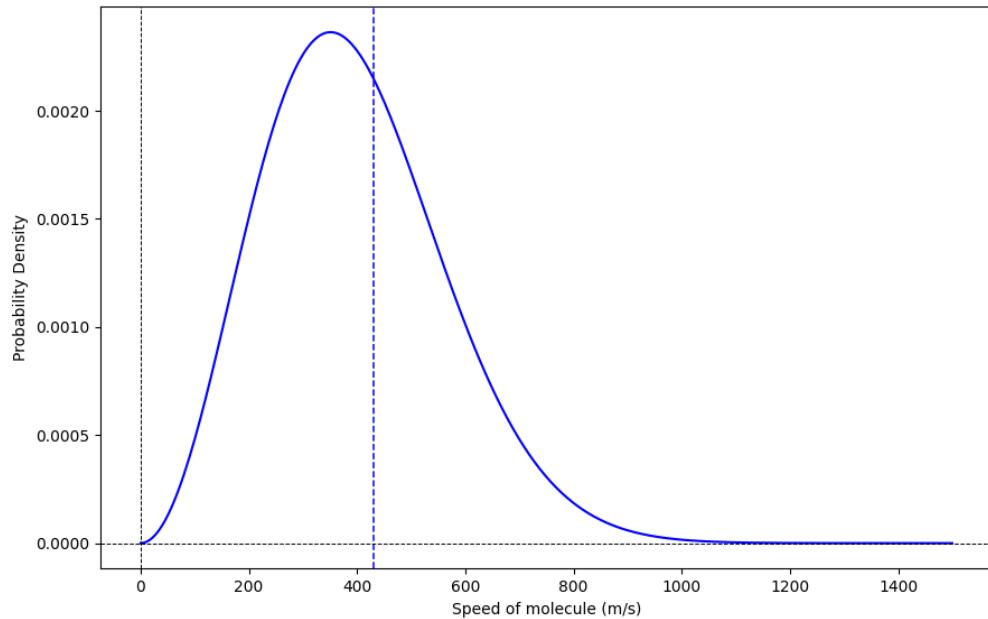
$$k = \frac{1}{2}mv^2$$

So, you could say "As the temperature of the gas increases, the kinetic energy of the molecules increases." But a physicist would say "The temperature of the gas is how we measure its kinetic energy."

116.0.1 A Statistical Look At Temperature

If you say "This jar of argon gas is 25 degrees Celsius," you have told me about the *average* kinetic energy of the molecules in the jar. However, some molecules are moving very slowly. Others are moving really, really fast.

We could plot the probability distribution of the speeds of the molecules. For argon at 25 degrees Celsius, it would look like this:



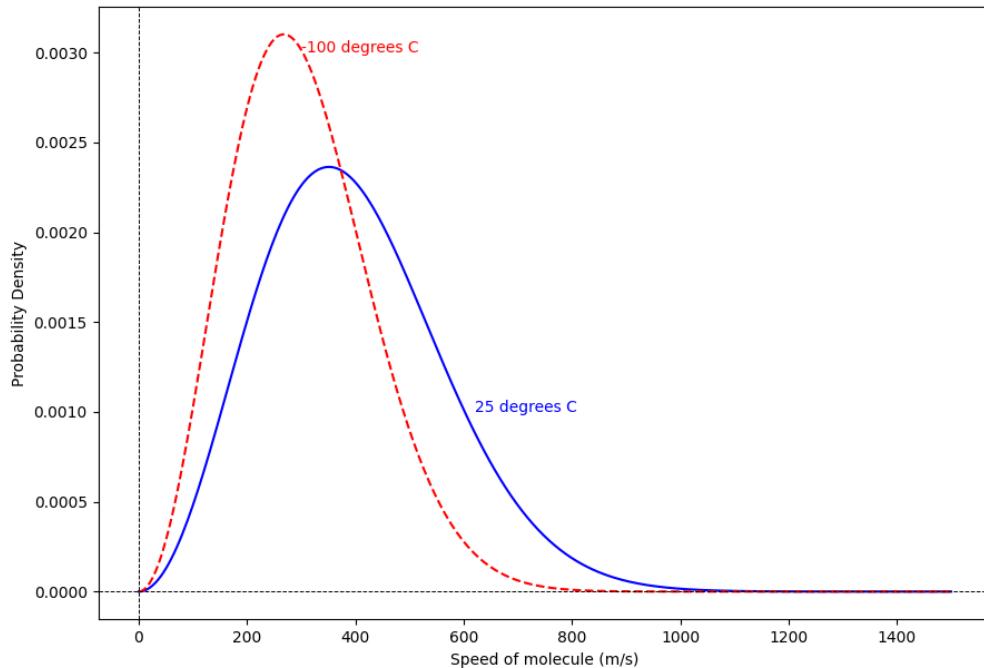
The temperature, remember, is determined by the average kinetic energy of the molecules. Some molecules are moving slowly and have less kinetic energy than the average. Some molecules are moving very quickly and have more kinetic energy. The dotted line is the divider between the two groups: molecules moving at speeds to the left of the line have less kinetic energy than average; those on the right have more kinetic energy than average.

Where is that line? That is the RMS of the speeds of the molecules. That is, if we measured all the speeds of all the molecules $s_1, s_2, s_3, \dots, s_n$, that line would be given by the root of the mean of the squares:

$$v_{\text{rms}} = \sqrt{\frac{1}{n} (s_1^2 + s_2^2 + s_3^2 + \dots + s_n^2)}$$

If you have the same gas at a lower temperature, the distribution shifts toward zero:

Here is probability distribution of molecular speeds for argon gas at 25 degrees and -100 degrees Celsius.



116.0.2 Absolute Zero and Degrees Kelvin

If you keep lowering the temperature, eventually all the molecules stop moving. This is known as *absolute zero* – you can't make anything colder than absolute zero. Absolute zero is -273.15 degrees Celsius.

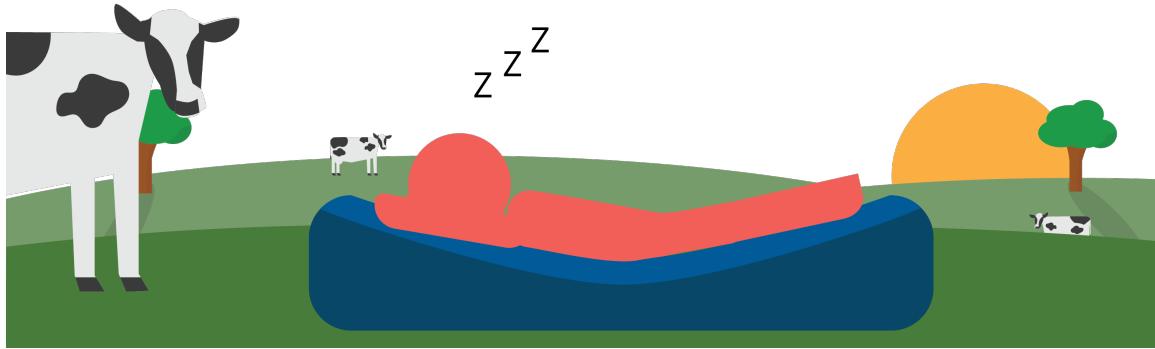
Besides Celsius and Fahrenheit, there is a third temperature system: Kelvin. The Kelvin has the same scale as Celsius, but it starts at absolute zero. So, 0 degrees Celsius is 273.15 degrees Kelvin. And 100 degrees Celsius is 373.15 degrees Kelvin.

Any time you are working with the physics of temperature, you will use Kelvin.

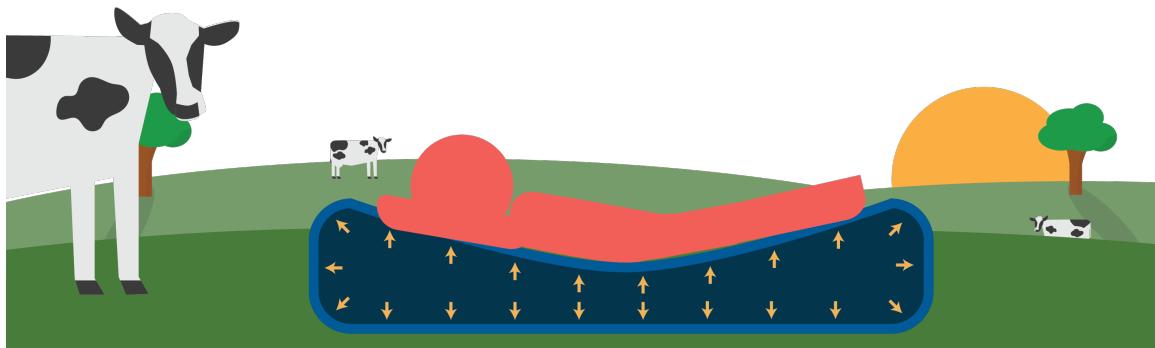
Sometimes, when reading about gases, you will see "STP" which stands for "Standard Temperature and Pressure." STP is defined to be 0° Celsius and 100 kPa.

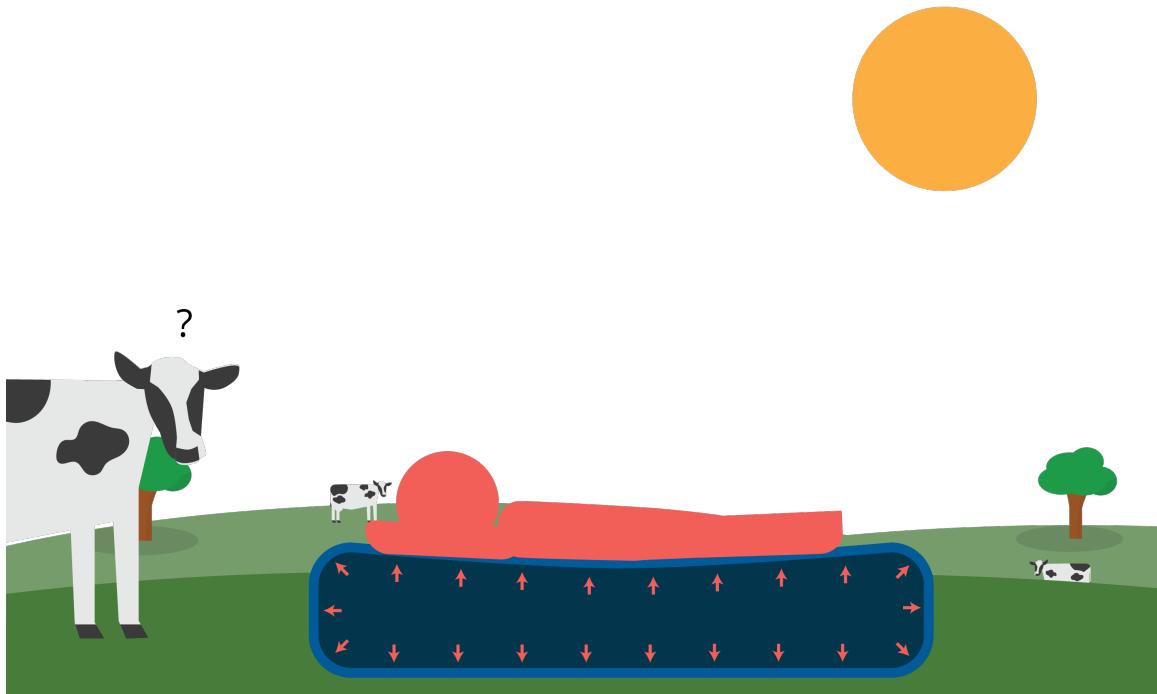
116.1 Temperature and Volume

Let's say you have a half-full air mattress in a field with a person lying on it around dawn. The weight of the person will keep the pressure of the air inside constant (or pretty close).



The molecules in the mattress are not entering or leaving that mattress. However, as the sun rises, the air inside will get warmer and expand. The person will be gently lifted by the expanding air. You might wonder: how much will the air expand?





If you have constant pressure and a constant number of molecules, the volume of the gas is proportional to the temperature in Kelvin:

$$V \propto T$$

Exercise 109 Temperature and Volume

Working Space

At dawn, the air inside mattress at dawn has a volume of 1000 liters and a temperature of 12 degrees Celsius.

At noon, that same air has a temperature of 28 degrees Celsius. The pressure on the gas has not changed at all.

What is the volume of the gas at noon?

Answer on Page 843

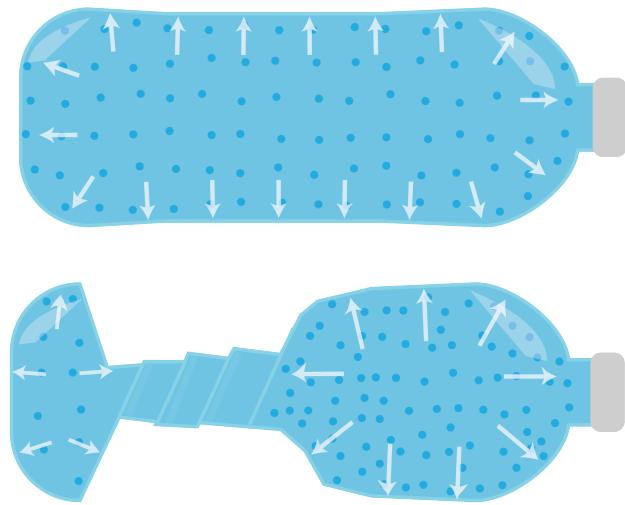
Note: Volume and temperature are only proportional as long as the substance is a gas.

We will talk about liquids and solids soon.

116.2 Pressure and Volume

As you increase the pressure on a gas, the molecules will get pushed closer together, and the volume will decrease.

For example, if you put the cap on an empty plastic bottle and squeeze it. As you put the gas inside the bottle under pressure, its volume will decrease.



If you keep the number of molecules and the temperature constant, the pressure of the gas and its volume are inversely proportional:

$$P \propto \frac{1}{V}$$

"But," you say with disbelief, "if I increase the pressure on my empty water bottle from 5 kPa to 10 kPa, the volume inside won't decrease by half!"

Don't forget that the air inside the bottle is under 101 kPa of atmospheric pressure before you even start to squeeze it.

Exercise 110 Temperature and Volume*Working Space*

At an altitude where the atmospheric pressure is 100 kPa, you seal air in a 1 liter water bottle.

Squeezing the water bottle, you raise the internal pressure by 20 kPa. What is the volume inside the bottle now?

*Answer on Page 843***116.3 The Ideal Gas Law**

You are gradually getting an intuition for the relationship between the number of molecules, the volume, the pressure, and the temperature of a gas. We can actually bring these together in one handy equation.

Ideal Gas Law

$$PV = nRT$$

where:

- P is the pressure in pascals
- V is the volume in cubic meters
- n is the number of molecules in moles
- R is the molar gas constant: 8.31446
- T is the temperature in Kelvin

(You can remember this as the "Pivnert.")

From the name, you might predict the following: The Idea Gas Law is not 100% accurate. But for most purposes, it works remarkably well.

Notice that the ideal gas law says nothing about what kind of gas it is; it works regardless.

Exercise 111 Ideal Gas Law

Working Space

You have a cylinder containing O₂. The chamber inside has a radius of 12 cm and a length of 50 cm. The temperature inside the cylinder is 20 degrees Celsius. The pressure inside the tank is 600 kPa. How many moles of O₂ are inside?

Answer on Page 843

116.4 Molecules Like To Stay Close to Each Other

When two molecules get close to each other a few things can happen:

- They can undergo a chemical reaction: electrons are exchanged or shared and a different molecule or molecules come into existence. This is the realm of chemistry, and we won't go into it in this course.
- One or both of them have so much kinetic energy that they just pass each other or bounce off each other. This is what happens in a gas.
- The two molecules can "stick" together. This is what happens in a liquid or a solid.

Why do they stick together if they aren't combined in a chemical reaction?

First, they don't get *too close*. If they get too close, their electron clouds repel each other with a strong force. This is what happens in a gas when two molecules bounce off of each other.

But if the molecules are quite close to each other, there are forces that will attract them toward each other. These intermolecular forces are beyond the scope of this course, but they called Van der Waals forces and hydrogen bonds. The strength of these forces vary based on the two molecules involved.

Which is why some of the matter around you is in gas form (molecules that don't stick together at the temperature and pressure you are living in because they have weak attrac-

tive forces) and some is non-gas (gangs of molecules with stronger attraction that makes them clump together as a liquid or a solid at that same temperature and pressure).

But. What if we change the temperature and pressure, we can change if and how the molecules clump together. This is known a *phase change*; We will cover it soon.



CHAPTER 117

Kinetic Energy and Temperature of a Gas

As mentioned in the previous chapter, for a particular gas, the temperature (in Kelvin) is proportional to the average kinetic energy of the individual molecules.

Perhaps you want to warm 3 moles of helium gas (trapped in a metal cylinder) from 10 degrees Celsius to 30 degrees Celsius. How would you compute exactly how many Joules of energy this would require?

The amount of energy necessary to raise one mole of a molecule by one degree is known as *molar heat capacity*. (The molar heat capacity of liquid water, for example, is 75.38 J per mole-degree.)

With gases, are actually two different possible situations:

1. Constant volume: As you heat the gas, the pressure and the temperature increase. This molar heat capacity is usually denoted as $C_{V,m}$.

2. Constant pressure: As you heat the gas, the temperature and the volume increase. This molar heat capacity is usually denoted as $C_{P,m}$.

All gases made up of one atom (Helium, for example, is a monoatomic gas.) have the same values for $C_{V,m}$ and $C_{P,m}$:

$$C_{V,m} = \frac{3}{2}R \approx 12.47 \text{ Joules per mole-degree}$$

$$C_{P,m} = \frac{5}{2}R \approx 20.8 \text{ Joules per mole-degree}$$

(Remember from last chapter that R is the ideal gas constant ≈ 8.31446 Joules per mole-degree.)

Exercise 112 Warming Helium

Working Space

You have 3 moles of helium.

- How many Joules would be required to warm 3 moles of helium gas by 20 degrees Celsius at constant volume?
- How many Joules would be required to warm 3 moles of helium gas by 20 degrees Celsius at constant pressure?

Answer on Page 843

117.1 Molecule Shape and Molar Heat Capacity

We told you that gases made up of one atom have the same values for $C_{V,m}$ and $C_{P,m}$:

$$C_{V,m} = \frac{3}{2}R \approx 12.47 \text{ Joules per mole-degree}$$

$$C_{P,m} = \frac{5}{2}R \approx 20.8 \text{ Joules per mole-degree}$$

For any molecule, it is generally true that

$$C_{P,m} \approx C_{V,m} + R$$

It is also true that for any molecule, there is some integer d such that

$$C_{V,m} \approx \frac{d}{2}R$$

For example, for all monoatomic gases, $d = 3$. For diatomic gases (like N_2 and O_2 , d is 5.

d is known as the *degree of freedom* of the molecule. When you study chemistry, they will teach you to predict d based on the shape of the molecule.

Here are the relevant numbers for some gases you are likely to work with:

Gas	type	$C_{V,m}$	$C_{P,m}$	d
He	monoatomic	12.4717	20.7862	3
Ar	monoatomic	12.4717	20.7862	3
O_2	diatomic	21.0	29.38	5
N_2	diatomic	20.8	29.12	5
H_2O (water vapor)	3 atoms	28.03	37.47	7
CO_2	3 atoms	28.46	36.94	7

117.2 Kinetic Energy and Temperature

For a sample of a gas, we can calculate its kinetic energy based on its molar heat capacity, the number of molecules, and the temperature:

$$E_K = C_{V,m}nT$$

where

- E_K is the kinetic energy in Joules
- $C_{V,m}$ is the molar heat capacity of the gas at constant volume
- n is the number of molecules in moles
- T is the temperature in Kelvin

Exercise 113 Warming Helium Revisited**Working Space**

How much kinetic energy does 3 moles of helium have at 10 degrees Celsius?

How much kinetic energy does 3 moles of helium have at 30 degrees Celsius?

What is the difference?

*Answer on Page 844***117.3 Why is $C_{V,m}$ different from $C_{P,m}$?**

What if, instead of keeping the volume constant while we heat the molecules in the helium tank, we keep the pressure constant and let the gas expand? The change in kinetic energy is the same: 748 Joules.

However, we know that the molar heat capacity if we keep pressure constant is $\frac{5}{2}R$, so heating will require $\frac{5}{2}R(3)(20) = 1247$ Joules.

What happened to the 499 missing Joules!? Thermodynamics tells us energy is neither created nor destroyed. So it must have gone somewhere.

That energy was used pushing against the pressure as the gas expanded. For example, maybe the sample was in a balloon in space – the extra energy stretched the surface of the balloon.

The 499 Joules were converted into potential energy.

117.4 Work of Creating Volume Against Constant Pressure

Let's imagine that you had a total vacuum (zero pressure) with a piston. As you pulled the piston out, you would be pulling against the atmospheric pressure. How much energy would that require?

If you increased the volume of the vacuum by V against a pressure of P , you would do VP work.

Let's check to make sure the 499 Joules mentioned above makes sense with this in mind.

No initial pressure was given in the problem, so let's just make one up and see how things work out: 100 kPa. Using the ideal gas law, the initial volume would be:

$$V_1 = \frac{nRT}{P} = \frac{(3)(8.31446)(283.15)}{100,000} = 0.07063 \text{ cubic meters}$$

The volume after we heated the gas and let it expand against 100 kPa would be:

$$V_2 = \frac{nRT}{P} = \frac{(3)(8.31446)(303.15)}{100,000} = 0.07562 \text{ cubic meters}$$

So the volume increased by $0.07562 - 0.07063 = 0.00499$ cubic meters. Multiplying that by 100,000 pa, we get 499 Joules as we expected!

117.5 Why does a gas get hotter when you compress it?

Now imagine that there is gas inside the piston and you push on the piston to compress that air. The work that you do is converted into kinetic energy, and that kinetic energy raises the temperature of the gas.

So, for example, if you had two moles of argon gas in the piston. If you pushed the piston 0.1 meters with an average force of 50 newtons, you will have done 5 Joules of work.

How much would 5 Joules raise the temperature of 2 mole of a monoatomic gas?

$$\Delta T = \frac{5}{(2)(C_{V,m})} = 0.2^\circ \text{ Kelvin}$$

It works both ways: compression makes a gas hotter Decompression makes a gas colder. You can sometimes experience the heat of compression when you pump up a bicycle tire – as you pump the tire will get warmer.

If you compress or decompress a gas without letting any heat enter or depart, we say the compression or decompression was *adiabatic*. In order to solve any interesting problems about heating/cooling due to compression/decompression, you will need to assume the process was adiabatic.

When a spacecraft enters the atmosphere, it has to deal with a lot heat. Some people assume that heat is due to friction of the air rubbing against the spacecraft at over 7,000 meters per second. Actually, most of the heat is due to the compression of the air as it gets pushed out of the way of the spacecraft.

117.6 How much hotter?

Let's say you have a accordion-like container filled with helium at 100 kPa (about 1 atmosphere) and 300 degrees Kelvin. It holds 2 cubic meters. And then you put it in a vice and quickly compress it down to 0.5 cubic meters. Assuming it was adiabatic, how hot would the gas inside be after the compression?

Here is the challenging part: As you crush the container, the temperature and the pressure in the container are both increasing. So as you go, it gets harder and hard to crush. So each milliliter of volume that you eliminate requires a little more work than the milliliter before.

Let's simulate the process in python, and then I'll give you the formula.

In the simulation, you will start with an initial volume of 2 cubic meters and crush it down to 0.5 cubic meters in 40 steps. At each step you will recalculate the temperature and pressure.

Then you will plot the results. Make a file called `gas_crunch.py`:

```
import numpy as np
import matplotlib.pyplot as plt

V_initial = 2.0 # cubic meters
V_final = 0.5 # cubic meters
step_count = 40 # steps

T_initial = 300.0 # kelvin
P_initial = 100000 # pascals

# Constants
R = 8.314462618 # ideal gas constnt
C_v = 3.0 * R / 2.0 # molar heat capacity (constant volume)

# Compute the number of moles
n = P_initial * V_initial/(R * T_initial)
print(f"The container holds {n:.2f} moles of helium")

# How much volume do we need to eliminate in each step?
# (in cubic meters)
step_size = (V_initial - V_final) / step_count

# For recording the state for each step
data_log = np.zeros((step_count, 3))
```

```

# Variables to update in the loop
V_current = V_initial
T_current = T_initial
P_current = P_initial

for i in range(step_count):
    # Record the current state
    data_log[i,:] = [T_current, V_current, P_current/1000.0]

    # Find how much energy to make the step at the current pressure
    E_step = step_size * P_current

    # Find how big the change in temperature will be from that energy
    delta_T = E_step / (n * C_v)

    # Update the current temperature, volume, and pressure
    T_current += delta_T
    V_current -= step_size
    P_current = n * R * T_current / V_current

print(f"Iterative:{T_current:0.3f} K, {V_current:0.3f} m3, {P_current/1000.0:0.3f} kPa")

fig, axs = plt.subplots(3,1,sharex=True, figsize=(8, 6))
axs[0].set_xlim((0,step_count))
axs[0].plot(data_log[:,0], 'k', lw=0.2)
axs[0].plot(data_log[:,0], 'r.')
axs[0].set_ylabel("Temperature (K)")

axs[1].plot(data_log[:,1],'k', lw=0.2)
axs[1].plot(data_log[:,1], 'r.')
axs[1].set_ylabel("Volume (cubic m)")

axs[2].plot(data_log[:,2], 'k', lw=0.2)
axs[2].plot(data_log[:,2], 'r.')
axs[2].set_ylabel("Pressure (kPa)")

axs[2].set_xlabel("Step")

fig.savefig('tvpplot.png')

```

When you run this, you will see how many moles of gas there are and reasonable estimates of the temperature, volume, and pressure:

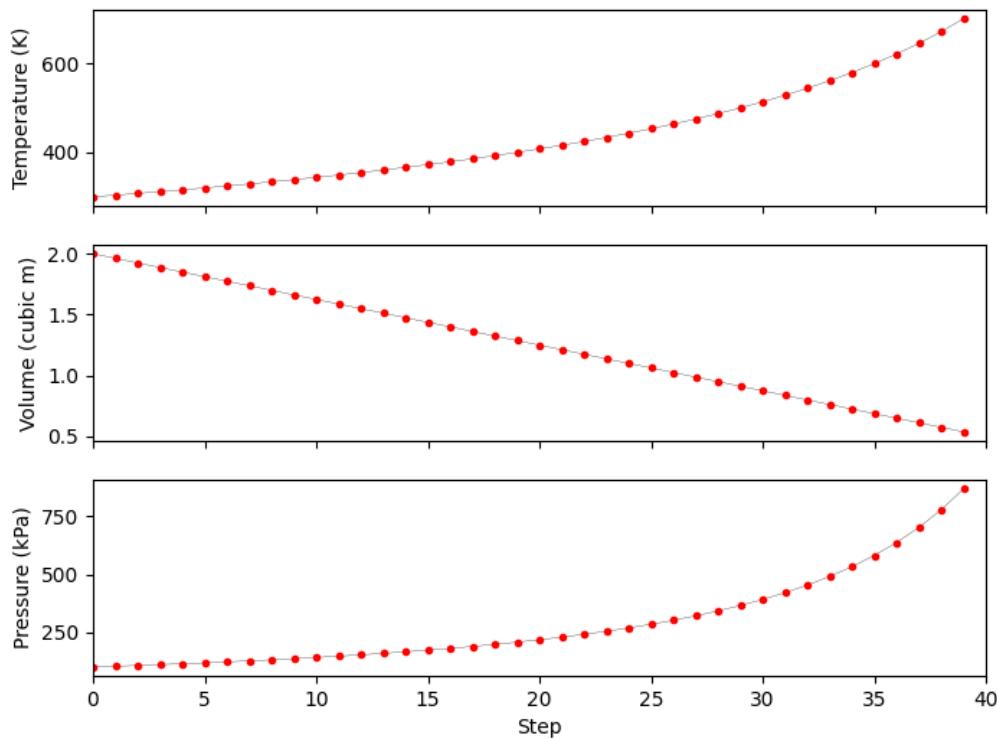
```

> python3 gas_crunch.py
The container holds 80.18 moles of helium

```

Iterative: 733.499 K, 0.500 m³, 977.999 kPa

And a good plot of the intermediate values:

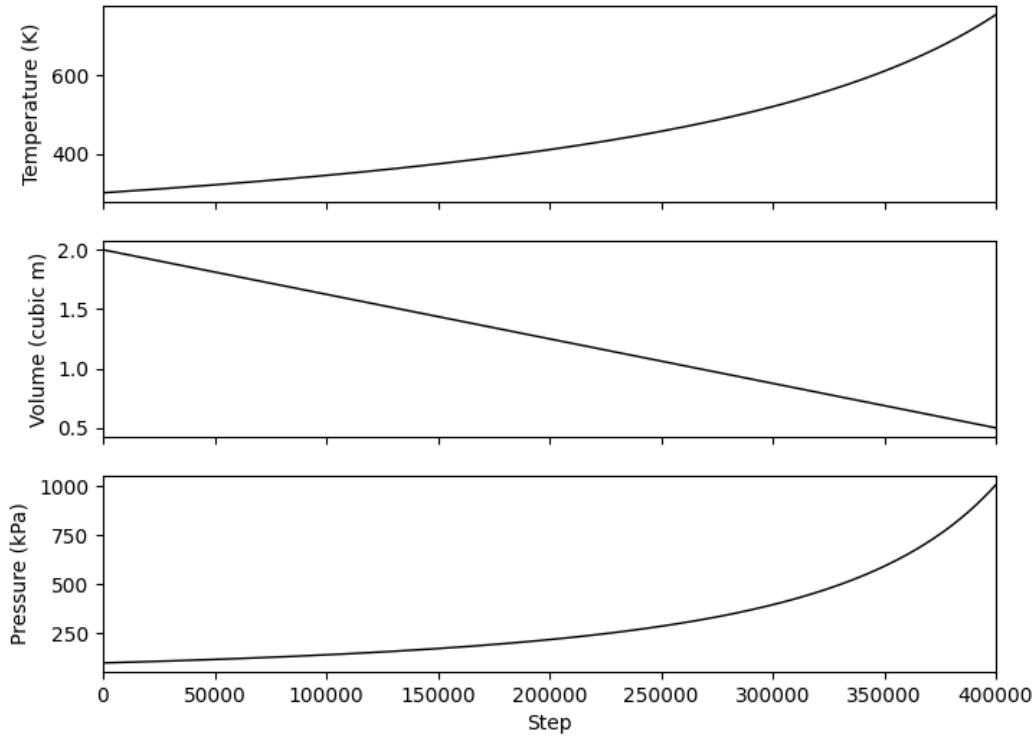


But, we will get better estimates if we break it up into 400 steps instead of 40. Change the line that defines the number of steps:

```
step_count = 400 # steps
```

Now the predicted temperature and pressure should be something like 753.603° K and 1004.803 kPa. (This is much closer to the correct result: 755.953° K and 1007.937 kPa.)

What if you break it into 400,000 steps? Now the result should be really, really close to correct. And the plot is quite accurate:



(You can comment out the lines that make the red dots on the graphs. No one wants to see 400,000 red dots.)

It is inefficient to have to do long simulations to guess the final temperature and pressure. Fortunately, there are two handy rules you can use to skip this:

Adiabatic Compression and Decompression

Let

$$\gamma = \frac{C_{P,m}}{C_{V,m}}$$

In an adiabatic compression or decompression, P and V change, but

$$P(V^\gamma)$$

stays constant.

Also

$$T(V^{(\gamma-1)})$$

stays constant

For a monoatomic gas:

$$\gamma = \frac{C_{p,m}}{C_{V,m}} = \frac{5}{3}$$

$$\gamma - 1 = \frac{2}{3}$$

Before the compression:

$$T(V^{(\gamma-1)}) = 300(2^{0.6667}) = 476.22$$

After the compression it has to be the same:

$$T(V^{(\gamma-1)}) = T(0.5^{0.6667}) = 476.22$$

Thus

$$T = 755.95^\circ \text{ Kelvin}$$

We can then use the ideal gas law to solve for the final pressure:

$$P = \frac{nRT}{V} = \frac{(80.18)(8.31446)(755.95)}{0.5} = 1007937 \text{ pascals}$$

That's hot! As you let it cool back down to 300 degrees Kelvin, how much heat would be released?

$$E = C_{V,m}n\Delta T = (12.47)(80.2)(755.95 - 300) \approx 456 \text{ kJ}$$

117.7 How an Air Conditioner Works

Once again, imagine the accordion-like container filled with helium. Let's say you walked it outside and compressed it from 2 cubic meters to 0.5 cubic meters in a vise. The container would get to 755.95 degrees Kelvin. You keep it compressed, in the vise but let it cool down outside. When it gets back to 300 degrees Kelvin, you walk it back inside.

Now, without letting any molecules in or out of the container, you release the vise. The gas is decompressed and gets very cold – how cold? Cold enough to accept about 456 kJ of kinetic energy from your house. That is, it would absorb heat from your house until the gas inside was the same temperature as your house.

Now you walk outside with your accordion and your vise and repeat:

1. Compress the gas outside.
2. Let the hot gas cool down outside.
3. Walk the room-temperature compressed gas inside.
4. Decompress the gas inside.
5. Let the cold gas warm up inside.

You could keep your house cool on a hot day this way. And this is not unlike how an air conditioner works.

There is a hose filled with refrigerant that does a loop:

- Outside, the refrigerant is compressed and allowed to cool to the outside temperature. (Usually there is a big fan blowing on a coil of refrigerant to speed the process.) Inside, the refrigerant is decompressed and allowed to warm to the inside temperature. (Usually there is a big fan blowing the air of the home past a coil of refrigerant to speed the process.)

In each pass of the loop, the refrigerant absorbs some of the kinetic energy from inside the house, and releases it on the outside.

This same mechanism can be used to heat your house. (Units that both heat and cool are known as *heat pumps*.) The heat pump does the process backwards: The hot compressed refrigerant cools down inside. The cold decompressed refrigerant warms up outside.



CHAPTER 118

Phases of Matter

You have experienced H₂O in three phases of matter:

- Ice is H₂O in the solid phase. At standard pressure, when the temperature of H₂O is below 0° C, it is a solid.
- Water is H₂O in the liquid phase. At standard pressure, when the temperature of H₂O is between 0° C and 100° C, it is a liquid.
- Water vapor (or steam) is the gas phase. At standard pressure, when the temperature of H₂O is above 100° C, it is a gas.

Let's look at some of the properties of the three phases:

Gas	Liquid	Solid
Assumes the volume and shape of its container	Assumes the shape, but not the volume, of its container	Retains its shape and volume
Compressible	Not compressible	Not compressible

118.1 Thinking Microscopically About Phase

As mentioned in an early chapter, there are intermolecular forces that attract molecules to each other. A pair of molecules will have very strong intermolecular forces or very weak intermolecular forces depending on what atoms they are made of.

For example, two helium molecules are very weakly attracted to each other due to weak intermolecular forces. Two molecules of NaCl (table salt) will experience very strong intermolecular attraction.

In a gas, the molecules have lots of room to roam and lots of kinetic energy: The intermolecular attraction has very little effect.

In a liquid, the molecules are sticking close together, but are still moving around, sort of like bees in a hive.

In a solid, the molecules are not changing their configuration, and the kinetic energy they have is just expressed as vibrations within that configuration. You can imagine them like eggs in a carton just vibrating.

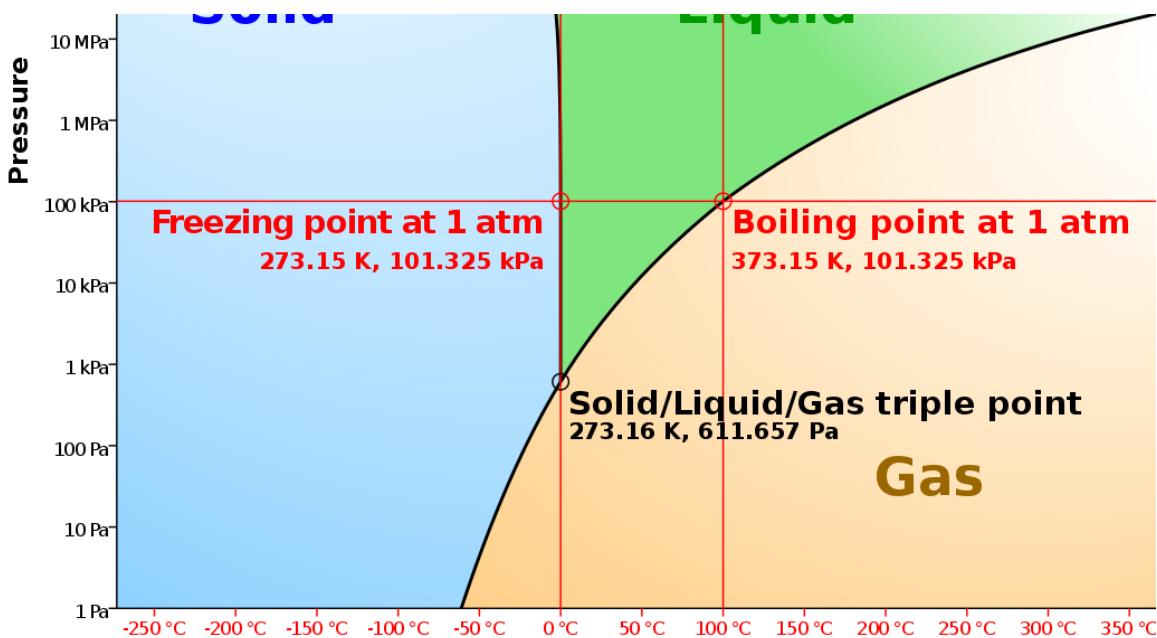
As you would expect, molecules with strong intermolecular attraction require more kinetic energy to change phases. For example, helium is a liquid below -269° C . NaCl, on the other hand, is a liquid between 801° and $1,413^{\circ}\text{ C}$.

The temperatures I just gave you are at standard pressure (100 kPa or 1 atm). Pressure also has a role in phase change: In low pressure environments, it is much easier for the molecules to make the jump to being a gas.

For example, if you climb a mountain until the atmospheric pressure is 70 kPa, your water will boil at about 90° C .

If you rise in a balloon until the atmospheric pressure is 500 Pa, if your water is colder than -2° C , it will be ice. If it is warmer it will vaporize. There is no liquid water at 500 Pa!

For any molecule, we could observe its phase at a wide range of temperatures and pressures. This would let us create a phase diagram. Here is the phase diagram for H₂O:



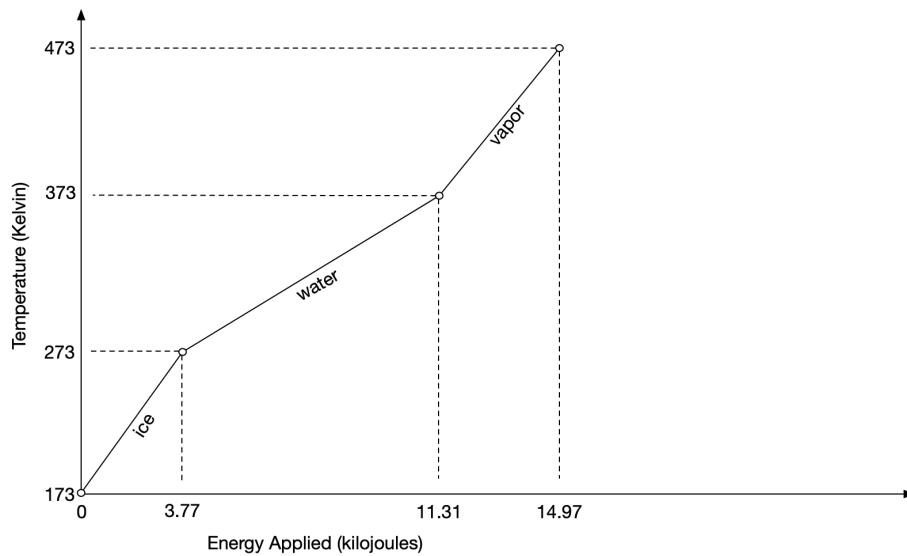
(FIXME: This diagram needs to be recreated prettier.)

118.2 Phase Changes and Energy

The molar heat capacity of ice is about $37.7\text{ J/mol}\cdot\text{K}$. That is it takes about 37.7 Joules of energy to raise the temperature of one mole of ice by one degree kelvin.

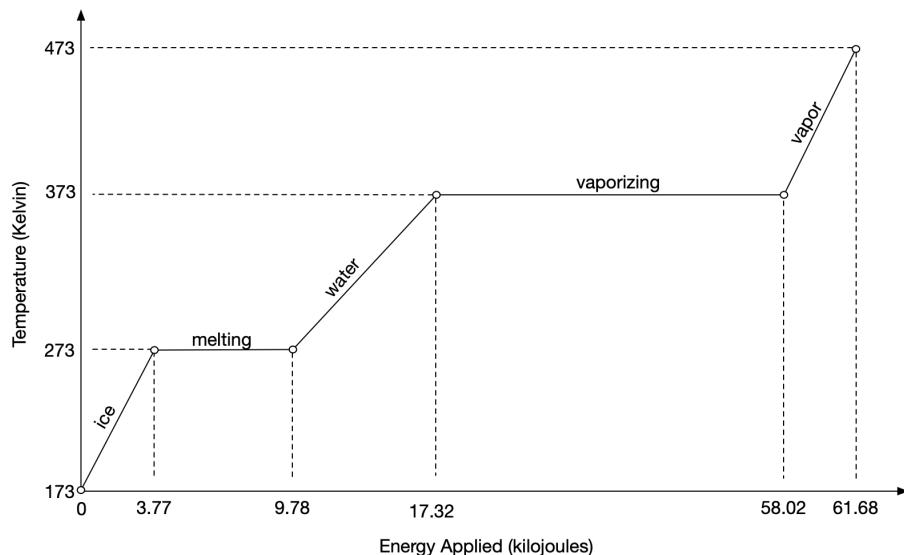
The molar heat capacity of liquid water is about $75.4\text{ J/mol}\cdot\text{K}$. For water vapor, it is about $36.6\text{ J/mol}\cdot\text{K}$.

Imagine you have one mole of ice at 173° K and you are going to gradually add kinetic energy into it until you have steam at 473° K . You might guess (wrongly) that the temperature vs. energy applied would look like this:



However, once molecules are nestled into their solid state (like eggs in cartons), it takes extra energy to make them move like a liquid. How much more energy? For water, it is 6.01 kilojoules per mole. This is known as *the latent heat of melting* or *the heat of fusion*.

Similarly, the transition from liquid to gas takes energy. At standard pressure, converting a mole of liquid water to vapor requires 40.7 kilojoules per mole. This is known as *the latent heat of vaporization*. So the graph would actually look like this:



Note that just as melting and vaporizing require energy. Going the other way (freezing and condensing, respectively) give off energy. Thus, we can store energy using the phase change.

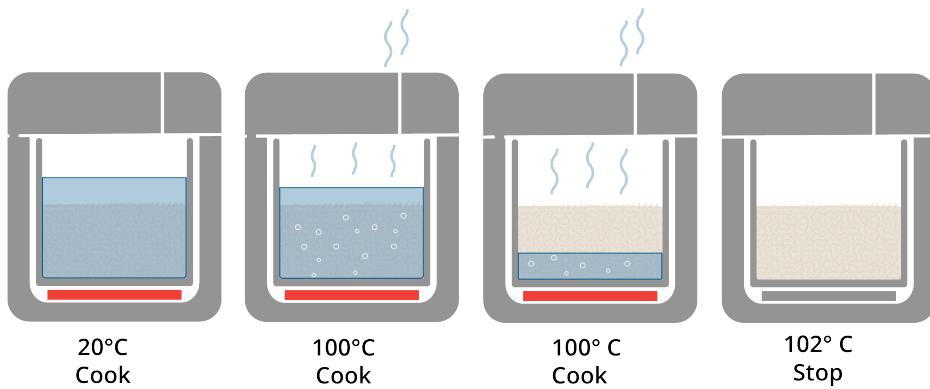
118.3 How a Rice Cooker Works

As you might imagine, a rice cooker is a bowl with a lid and an electric heating element. You put rice and water into the bowl and turn on the heating element. The heating element pushes kinetic energy into the water, which gets warmer and eventually starts to boil.

How does the rice cooker know when to turn off (or at least down to a low-heat "keep warm" mode) before the rice starts to burn?

As long as there is a little liquid water in the bottom of the bowl, the rice won't burn, so the question really is "How does it know when there is no more liquid water in the bottom of the bowl?"

There is a mechanism (and there have been a few different versions of this mechanism) that monitors the temperature of the surface of the bowl. As long as there is liquid water in the bowl, it *cannot* go above 100° C ! When all the water has been absorbed by the rice or turned to steam, the temperature rises above 100° C , and the mechanism cuts off the heat.



Exercise 114 Using Water For Thermal Energy Storage*Working Space*

Tom is building a passive solar house: the front of his house is a greenhouse. He also likes to eat dinner in the greenhouse. He will have barrels (painted black) that hold 159 liters of water. His plan is to let the sun heat the barrels to 33° C by the time the sun goes down. (Any warmer and it would be unpleasant to eat dinner near them.)

At night, he will circulate air past the barrels and through his house. He is OK with the house and the barrels dropping to 17° C .

Looking at the insulation on his house and the expected nighttime temperatures, Tom estimates that he needs to store 300,000 KJ of energy in the barrels.

A mole of water is about 0.018 liters.

The molar heat capacity of water is about 75.38 J/mole-K.

How many barrels of water does Tom need to install in his greenhouse?

Answer on Page 844

Exercise 115 Using Mirabilite For Thermal Energy Storage**Working Space**

Water barrels are going to take up too much of Tom's greenhouse!

There is a substance known as mirabilite, or Glauber's salt, or sodium sulfate decahydrate. It is relatively cheap to produce, and it has a melting point of 32.4°C .

The molar heat capacity of mirabilite is 550 J/mole-K.

The latent heat of melting mirabilite is 82 KJ per mole.

Mirabilite comes in a powder form. Assume that a mole of mirabilite occupies about 0.22 liters.

If Tom fills his barrels with mirabilite, how many barrels will he need?

Answer on Page 845**118.4 Thinking Statistically About Phase Change**

We like to say simple stuff like "At 100°C , water changes to vapor." However, remember what temperature is: Temperature tells you how much average kinetic energy the water molecules have. The key word here is *average*; Some molecules are going faster than average and some are going slower.

A puddle in the street on a warm night will evaporate. It isn't 100°C . Why would the puddle turn to vapor?

While the *average* molecule in the puddle doesn't have enough energy to escape the intermolecular forces, some of the molecules do. When a molecule on the surface has enough velocity (toward the sky!) to escape the intermolecular forces, it becomes vapor and leaves the puddle.

What happens to the temperature of the puddle during this sort of evaporation? Temperature is proportional to the average kinetic energy of the molecules. If a bunch of molecules with a lot of kinetic energy leave, the average kinetic energy (of the molecules

that remain) will decrease.

The most obvious example of this process is sweating: When your body is in danger of getting too hot, sweat comes out of your pores and covers your body. The fastest moving molecules escape your body, taking the excess kinetic energy with them.

118.4.1 Evaporative Cooling Systems

An evaporative cooling system (also called a "Swamp Cooler") uses this idea to cool air. You can imagine a fan drawing warm air through a duct from the outside. Before the air is released inside, it passes very close to a cloth that is soaked with water. The warm air molecules (which has a lot of kinetic energy) slam into the water molecules, some of which get enough kinetic energy to become vapor. Then the cool air and the water vapor enter the room.

"Wait, wait, wait," you say, "The heat hasn't gone away. It is just transferred into the water molecules."

Remember that it takes 40. 7 kilojoules to change a mole of liquid water at 100° C to water vapor at 100°. Escaping those intermolecular bonds takes a lot of energy!

Thus, if a mole of water evaporates, it is because it has absorbed 40.7 kj of heat you can feel (*sensible heat*) and used it to liberate the molecules from their intermolecular bonds. For convenience, physicists call this *latent heat*.

118.4.2 Humidity and Condensation

When a puddle is evaporating on a warm day, there might be some water vapor already in the air. Even as the water molecules in the puddle are evaporating, some water molecules in the air are crashing into the puddle and become liquid again. (We say they *condensed*, thus the word *condensation* to describe the water that accumulates on a cold glass on a warm day.)

When there is a lot of water vapor in the air, the puddle will evaporate more slowly. In fact, if there is enough water vapor in the air, the puddle won't evaporate at all. At this point, we say "The relative humidity is 100%" That is, relative to the amount of water the air will hold, it already has 100% that amount."

Neither sweating nor evaporative cooling systems work well when the relative humidity is high.

As the temperature goes up, the air can hold more water. We usually notice it when it goes the other way: the air cools and has more water vapor than it can hold. Some of the

water vapor condenses into water droplets. If the droplets land on something, we call it "dew". If it is high in the sky, we call the droplets "a cloud". If it is near the ground, we call it "fog."



CHAPTER 119

The Piston Engine

Most cars, airplanes, and chainsaws get their power from burning hydrocarbons in a combustion chamber. We say they have *internal combustion engines*. There are many types of internal combustion engines: jet engines, rotary engines, diesel engines, etc. In this chapter, we are going to explain how one type, piston engines, work. Most cars have piston engines.

Most piston engines burn gasoline, which is a blend of liquid hydrocarbons. Hydrocarbons are molecules made of hydrogen and carbon (and maybe a little oxygen). In the presence of oxygen and heat, hydrocarbons burn – the carbon combines with oxygen to become CO_2 and the hydrogen combines with oxygen to become H_2O . In the process, heat is released, which causes the gases in the cylinder to create a lot of pressure on the piston.

119.1 Parts of the Engine

The engine block is a big hunk of metal. There are cylindrical holes bored into the engine block. A piston can slide up and down the cylinder. There are two valves in the wall of

the cylinder:

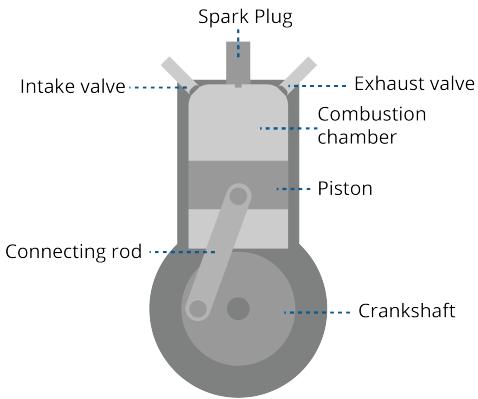
- Before the burn, one valve opens to let ethanol and air into the cylinder.
- After the burn, the other valve opens to let the exhaust out.

There is also a spark plug, which creates the spark that triggers the burn.

As you give the engine more gas, the cylinder does more frequent burns. When the engine is just idling, the cylinder fires about 9 times per second. When you depress the gas pedal all the way down, it is more like 40 times per second.

The cylinder has a rod that connects it to the crank shaft. As the pistons move back and forth, the crank shaft turns around and around. Sometimes a piston is pushing the crankshaft, and sometimes the crank shaft is pushing or pulling the piston. All the cylinders share one crank shaft.

How many cylinders does a car have? Nearly all car models have between 3 and 8 cylinders. The opening of the valve and the firing of the spark plugs are timed so that cylinders all do their burns at different times. This makes the total power delivered to the crank shaft smoother.



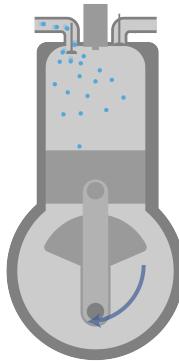
119.2 The Four-Stroke Process

Cars have four-stroke engines – this means for every two rotations of the crank shaft, each cylinder fires once. Smaller engines, like those in chainsaws, are often two-stroke engines – every cylinder fires every time the crank shaft rotates. For now, let's focus on four-stroke engines.

Here is the cycle of a single cylinder:

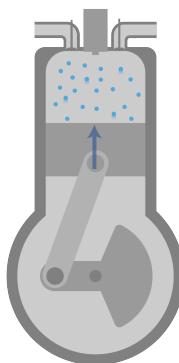
- As the drive shaft turns, it pulls the piston down. The intake valve opens and lets the gas/air mixture into the combustion chamber.

Intake



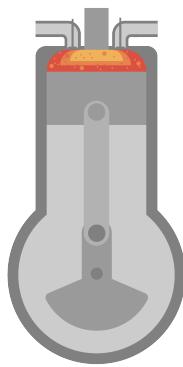
- As the piston reaches the bottom of the stroke, the intake valve closes.
- Now the crank shaft starts to push the piston up, compressing the gas and oxygen.

Compression



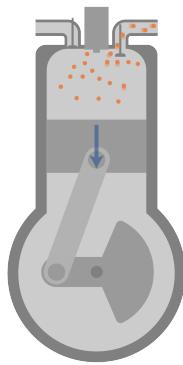
- As the piston reaches the top of its stroke, the spark plug creates a spark. The fuel and oxygen burn quickly. The cool liquid fuel becomes hot carbon dioxide and water vapor.

Combustion



- Now there is very high pressure inside the cylinder. It pushes hard on the piston which pushes the crank shaft.
- When the piston reaches the bottom of this stroke, the exhaust valve opens.
- As the crank shaft pushes the piston up, the carbon dioxide and water vapor is pushed out.

Exhaust



- When the piston reaches the top of this stroke, the exhaust valve is closed.

119.3 Dealing with Heat

Burning fuel inside a block of metal generates a lot of heat. If there is too much heat, parts of the engine will start to melt. So modern car engines are liquid cooled – there are arteries in the engine block carrying a liquid (called “coolant”). The hot coolant is pumped through the radiator (where the air passing through takes away the heat) and then back into the engine.

Note that the heat that is carried away by the coolant is wasted energy. In fact, of the total

energy created in burning the fuel, most car engines only transfer about 30% to turning the crank shaft. About 35% of the heat goes out with the exhaust. About 30% is carried away by the coolant. The remaining waste (usually about 5%) is lost to friction.

119.4 Dealing with Friction

From the description, it is clear that there is a lot of metal sliding against metal, which would grind the engine up quickly if there were no lubrication. In a modern car, the moving parts in the engine are constantly bathed in oil. There is an oil pump that causes it to get sprayed on the crankshaft, the connecting rod, and in the cylinder under the piston. (That is, not on the combustion side.)

The oil eventually falls through the oil into a pan at the bottom of the engine. The oil pump sucks the oil up, pushes it through a filter (so bits of metal are not pumped back into the engine), and then is sprayed on the moving parts again.

119.5 Challenges

With a piston engine, there are a lot of things that can go wrong. Let's enumerate a few:

- *The seal around the piston leaks.* Mechanics say "We aren't getting any compression." The cylinder doesn't get much power to the drive train.
- *The valves open or the spark plug fires at the wrong time.* This is known as a timing problem.
- *The spark plug doesn't make a spark.* The spark plug has two prongs of metal and electrons jump from one to the other. For a good spark, the prongs need to be a very precise distance apart. Sometimes you need to bend one of the prongs to get the right gap. This is known as *gapping*.
- *The mix of fuel and oxygen is wrong.* If there is too much fuel and not enough oxygen (so not all the fuel burns), we say the mix is too rich. If there is not enough fuel (so the pressure created by the burn is as high as possible), we say the mix is too lean.

119.6 How We Measure Engines

If you look up the specs on an engine, you will see the following:

- The number of cylinders
- The cylinder bore, which is the diameter of the cylinder

- The piston stroke, which is the distance the piston travels in the cylinder
- The compression ratio, which is the ratio between the maximum volume of the combustion change and the minimum volume of the combustion chamber.
- What fuel it runs on.

The difference between the minimum and maximum value of the cylinder is known as its *displacement*. The displacement represents the volume of air/fuel sucked into the intake valve before the compression begins.

We often talk about the displacement of the entire engine, which the cylinder's displacement times the number of cylinders. The displacement of an engine can give you a good idea of how much power it can produce.

For motorcycles, the displacement is often part of the name. For example, the Kawasaki Ninja 650 has about 650 cubic centimeters of displacement.

119.7 The Ford Model T and Ethanol

The Ford Model T was the first popular car. It came out in 1908 and remained in production until 1927. It had a four-cylinder engine that would run on ethanol, benzene, or kerosene. For the purposes of this exercise, let's assume you are running yours on ethanol.

A molecule of ethanol has 2 carbon atoms, 6 hydrogen atoms, and 1 oxygen atom. The oxygen in the atmosphere is O_2 . When one molecule of ethanol combines with three molecules of O_2 , 2 molecules of CO_2 and 3 molecules of H_2O are created. Also, a lot of heat is created: 1330 kilojoules for every mole of ethanol burned.

The engine block is usually very hot once the engine has been running. That is important because the ethanol will be completely vaporized at that temperature.

In any sample of air, 21 percent of the molecules will be O_2 .

Exercise 116 Fuel Mix for the Model T**Working Space**

On the Model T, a carburetor mixed the fuel and air before it went into the cylinder. The question to answer in this exercise is: How rich should the mix be at sea level (100 kPa)?

On the Wikipedia page for the Model T, we see the following facts:

- Cylinder bore: 9.525 cm
- Piston stroke: 10.16 cm

You can assume that the air/fuel mixture is 80° C before the pre-burn compression starts. (Thus the ethanol, which boils at 78° C, is in its vapor phase.)

The questions, then, are:

- What is the displacement of a single cylinder?
- How many moles of gas (80° C and 100 kPa) will get sucked through the intake valve?
- How many moles of vaporized ethanol should be part of that?
- How many moles of CO₂ and H₂O are created in each burn?
- How much heat is created in each burn?

(This exercise is a lot of steps, but nothing you don't know. You will need the ideal gas law to figure out how much many moles of air gets dragged into the cylinder.)

Answer on Page 845

119.8 Compression Ratio

Most of the inefficiency of a motor is heat that escapes through the exhaust valve. If your piston stroke were long enough, you could keep increasing the volume (which would cool the gases inside) until the gases inside were the same temperature as the outside world. Then there would be no wasted heat in the exhaust.

For this reason, generally, engines with a higher compression ratios tend to waste less energy through the heat of the exhaust. The Model T had a compression ratio close to 4:1. Modern car engines typically have compression ratios between 8:1 and 12:1.

Cars with really high compression ratios often require fuels with a lot of kilojoules per mole – we say *high octane*. If the fuel doesn't have enough energy, the engine makes loud knocking noises as the pistons don't have enough energy to push through their entire stroke.

It turned out that an easy way to boost the octane of the gasoline was to add a chemical called tetraethyl lead. Gasoline containing tetraethyl lead was known at "Leaded Gasoline" and was intended to prevent the knocking. It is difficult to overstate the damages caused by putting large amounts of lead in the air. Gradually, starting in with Japan in 1986, every country in the world has banned leaded gasoline.

119.9 The Choke and Direct Fuel Injection

Most cars built before 1990 will have a carburetor, which ensures that the ratio between fuel and oxygen is constant regardless of the amount of fuel released by the throttle.

If you go to start an old car on a cold morning, the cold engine will not properly vaporize the fuel and the engine may not have enough power to start. For this reason, most carburetors have a *choke value* that makes the mix richer. (If you pull the choke valve, be sure to push it back after the engine warms up.)

The carburetor was a common source of engine problems and inefficiencies. Starting in the 1990s, car engines started using direct fuel injection: Air still came in through the intake manifold, but fuel was sprayed directly into the cylinder by a fuel injection system.

In modern cars, the fuel injection system is controlled by a computer (an *Engine Control Module* or ECM) which delivers the fuel at the perfect time with the perfect amount based on environmental variables like the temperature of the engine and the barometric pressure (usually related to that altitude at which the engine is operating).



CHAPTER 120

u-Substitution

U-Substitution, also known as the method of substitution, is a technique used to simplify the process of finding antiderivatives and integrals of complicated functions. The method is similar to the chain rule for differentiation in reverse.

Suppose we have an integral of the form:

$$\int f(g(x)) \cdot g'(x) dx \quad (120.1)$$

The u-substitution method suggests letting a new variable u equal to the inside function $g(x)$, i.e.,

$$u = g(x) \quad (120.2)$$

Then, the differential of u , du , is given by:

$$du = g'(x) dx \quad (120.3)$$

Substituting u and du back into the integral gives us a simpler integral:

$$\int f(u) du \quad (120.4)$$

This new integral can often be simpler to evaluate. Once the antiderivative of $f(u)$ is found, we can substitute $u = g(x)$ back into the antiderivative to get the antiderivative of the original function in terms of x .

The method of u-substitution is a powerful tool for evaluating integrals, especially when combined with other techniques like integration by parts, partial fractions, and trigonometric substitutions.



CHAPTER 121

Differential Equations

Differential equations are equations involving an unknown function and its derivatives. They play a crucial role in mathematics, physics, engineering, economics, and other disciplines due to their ability to describe change over time or in response to changing conditions.

121.1 Ordinary Differential Equations

An ordinary differential equation (ODE) involves a function of a single independent variable and its derivatives. The order of an ODE is determined by the order of the highest derivative present in the equation. An example of a first-order ODE is:

$$\frac{dy}{dx} + y = x \quad (121.1)$$

Here, y is the function of the independent variable x , and $\frac{dy}{dx}$ represents its first derivative.

121.2 Partial Differential Equations

Partial differential equations (PDEs), on the other hand, involve a function of multiple independent variables and their partial derivatives. An example of a PDE is the heat equation, a second-order PDE:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} \quad (121.2)$$

In this equation, $u = u(x, t)$ is a function of the two independent variables x and t , $\frac{\partial u}{\partial t}$ is the first partial derivative of u with respect to t , and $\frac{\partial^2 u}{\partial x^2}$ is the second partial derivative of u with respect to x .



CHAPTER 122

Population Proportion Statistics

Let's say that you are trying to get a candidate elected. The candidate asks you "What proportion of the voting population is going to vote for me?" So you go out and ask a random sample of 12 voters. 11 say that they are going to vote for your candidate. What can you tell the candidate?

122.1 Sample Probabilities from Population Proportion

To address these sorts of questions (and there are a lot of them), we start with the opposite question: If we knew what the proportion was in the entire population, what sort of results should we expect in a random sample of just 12?

For example, let's say that 62% of the entire population plan to vote for your candidate. You ask 12 "Will you vote for my candidate?" How many will say "Yes"? You don't know – it depends on the sample. For example, there is some chance that you will just happen to choose all 12 from the 48% of the population that doesn't plan to vote for your candidate.

We can compute the probability of each outcome using the binomial distribution. Let r

be the probability that a random person will say "I plan to vote for your candidate." Let n be the number of people you ask. The probability that exactly k people will say "I plan to vote for your candidate" is given by:

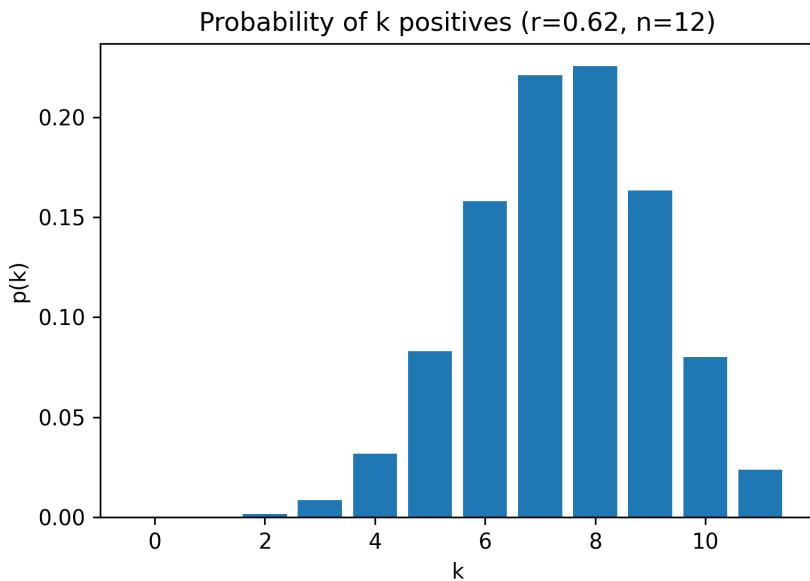
$$p(k) = \binom{n}{k} r^k (1 - r)^{n-k}$$

Note that even though most people support your candidate, there is some chance that no one you ask will say that they will vote for your candidate.

Using $r = 0.62$, we can compute the probability of each outcome:

k	$p(k)$
0	0.000009
1	0.000177
2	0.001593
3	0.008663
4	0.031801
5	0.083017
6	0.158024
7	0.220996
8	0.225358
9	0.163418
10	0.079989
11	0.023729

Looking at this, the most likely outcome is that 8 people will say "Yes." However, there is less than a 1 in 4 chance of that outcome. It is very unlikely that less than 2 people will say "Yes." Here is a bar chart of the data



In this section, we knew the proportion of the population (p) and used that to find the probability of each possible number of positives in a random sample (k). Now we are going to go the other way: You know k , and you are finding the probability of possible values of p .

122.2 Population Proportion from Sample

You ask 12 people if they will vote for your candidate. 9 say "Yes."

Now you do a thought experiment: "If only 10% of the population were going to vote for my candidate, what is the probability that I would see this outcome?"

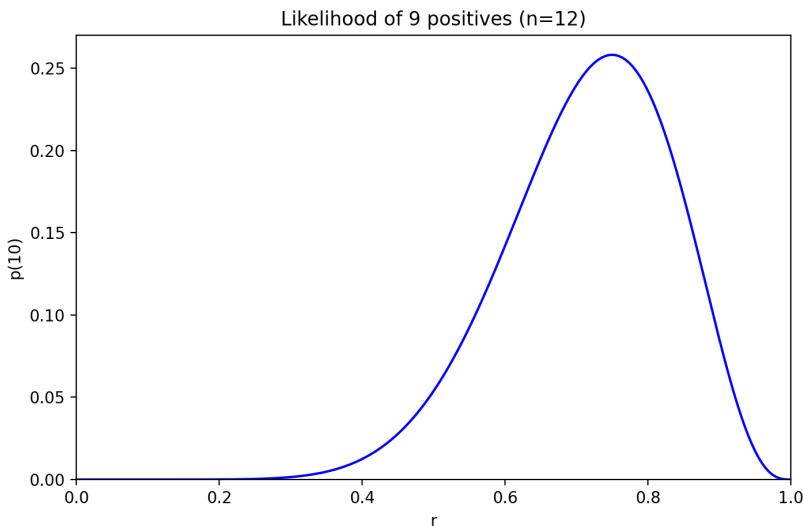
$$p(9) = \binom{n}{k} r^k (1-r)^{n-k} = \binom{12}{9} (0.1)^9 (0.9)^{12-9} = 0.00000016$$

So this outcome would be quite unusual. What if 70% of the population were going to vote for your candidate? What is the probability that you would see this outcome?

$$p(9) = \binom{n}{k} r^k (1-r)^{n-k} = \binom{12}{9} (0.7)^9 (0.3)^{12-9} \approx 0.2397$$

In this case, the observed outcome would be a lot less unusual.

So you decide to plot out the likelihood of this outcome for every possible value of r :



This looks a lot like a probability distribution, but *it is not*. The area under the curve does not integrate to 1.0 – it is significantly less. This is called a *likelihood*.

However, it still tells us something, right? The maximum likelihood estimator is $9/12 = 0.75$.

122.3 From Likelihood to Probability Density Function

How can we make this likelihood into a probability density function? We use Bayes' Law for continuous probability:

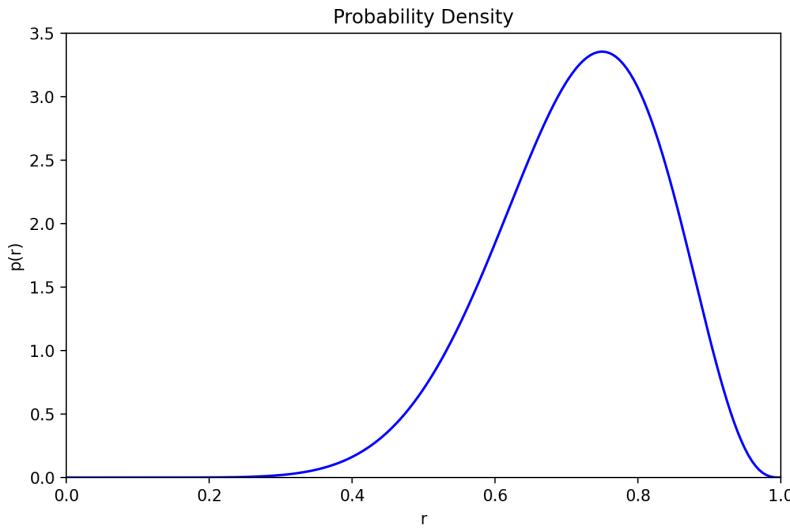
$$p(r|k) = \frac{p(k|r)p(r)}{\int_{r=0}^1 p(k|r)p(r)dr}$$

That is, given that we had k positive responses, what is the probability that the proportion of the population that will vote for your candidate is r ? The numerator of the fraction is the likelihood scaled up or down by our prior belief about the value of r . What is the denominator of the fraction? For this to be a probability distribution, we need it to integrate to 1. This is taken care of by the denominator.

Let's say we have no prior belief about the value of r . That is $p(r)$ is the continuous uniform distribution between 0 and 1, thus $p(r) = 1$ for all possible values of r . Our formula becomes:

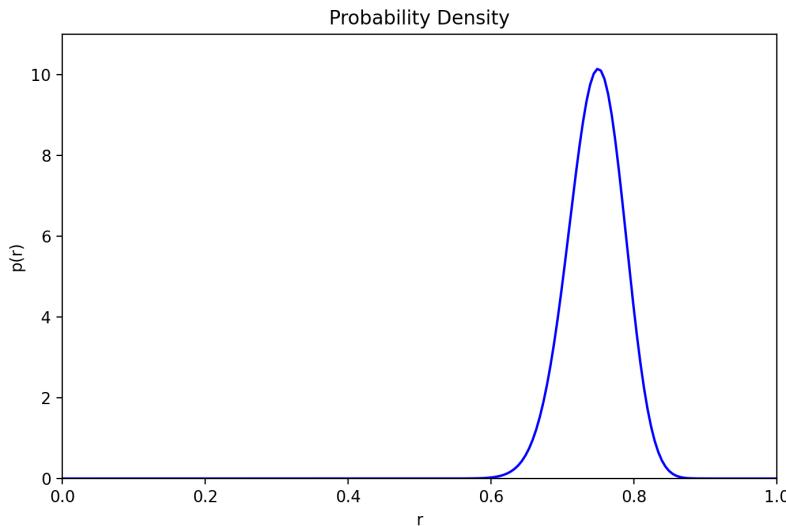
$$p(r|k) = \frac{p(k|r)}{\int_{r=0}^1 p(k|r)dr}$$

That is the likelihood scaled up so that it integrates to 1. If we plot this, we get:



Here then, is your report to your candidate: "I asked 12 voters if they were going to vote for you. 9 said yes. Using a uniform prior, here is what I believe about your support in the general population." And include this graph.

What happens to this graph if you ask 120 voters and 90 say yes?



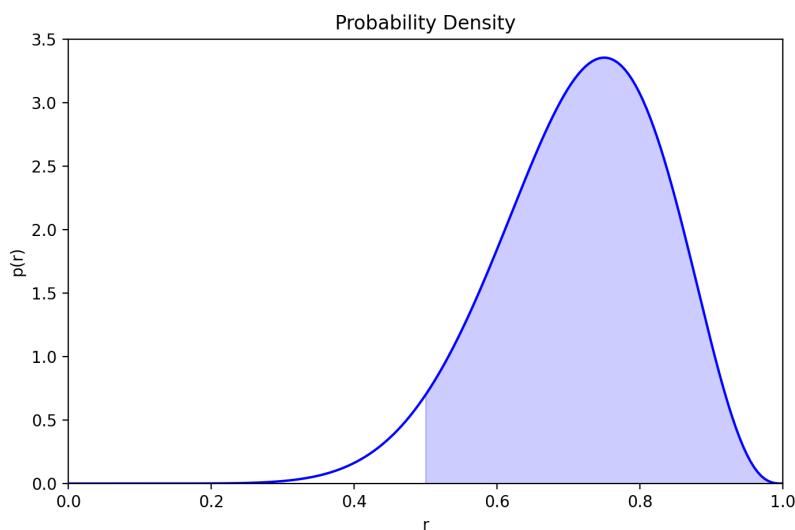
The MLE (0.75) is the same, but because of the much larger sample size, you are more confident when you say "It is probably close to 75%."

122.4 Beta Distribution

This probability distribution that you discovered is actually pretty common. It is known as the *beta distribution*.

The beta distribution has two parameters a and b that determine its shape. If you get k positives out of n , then use $a = k + 1$ and $b = n - k + 1$.

When you make your report to your candidate, they will look at your probability distribution with quiet awe and ask "Based on your sample of 12 people, what is the probability that at least 50% of the population will vote for me?" So, you'd fill in the region and say "This area represents that probability."



Once again, there will be a long silence. And then they will ask "Can you give me a number?" Here is the python code:

```
import numpy as np
from scipy.stats import beta

# Constants
K = 9
N = 12

# What is the probability r <=0.5?
p_less = beta.cdf(0.5, K +1, N - K +1)

# What is the probability r > 0.5?
p_more = 1.0 - p_less
print(f"I'm {p_more * 100.0:.2f}% sure you will win.")
```

This will give you:

I'm 95.39% sure you will win.



CHAPTER 123

The Normal Distribution

The Normal distribution, also known as the Gaussian distribution, is a type of continuous probability distribution for a real-valued random variable. It is one of the most important probability distributions in statistics due to its several unique properties and usefulness in many areas.

123.1 Defining the Normal Distribution

The Normal distribution is defined by its mean (μ) and standard deviation (σ). The probability density function (pdf) of a Normal distribution is given by:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where:

- x is the point up to which the function is integrated,

- μ is the mean or expectation of the distribution,
- σ is the standard deviation,
- σ^2 is the variance.

123.2 Importance of the Normal Distribution

There are several reasons why the Normal distribution is crucial in statistics:

- **Central Limit Theorem:** One of the main reasons for the importance of the Normal distribution is the Central Limit Theorem (CLT). The CLT states that the distribution of the sum (or average) of a large number of independent, identically distributed variables approaches a Normal distribution, regardless of the shape of the original distribution.
- **Symmetry:** The Normal distribution is symmetric, which simplifies both the theoretical analysis and the interpretation of statistical results.
- **Characterized by Two Parameters:** The Normal distribution is fully characterized by its mean and standard deviation. The mean determines the center of the distribution, and the standard deviation determines the spread or girth of the distribution.
- **Common in Nature:** Many natural phenomena follow a Normal distribution. This includes characteristics like people's heights or IQ scores, measurement errors in experiments, and many others.

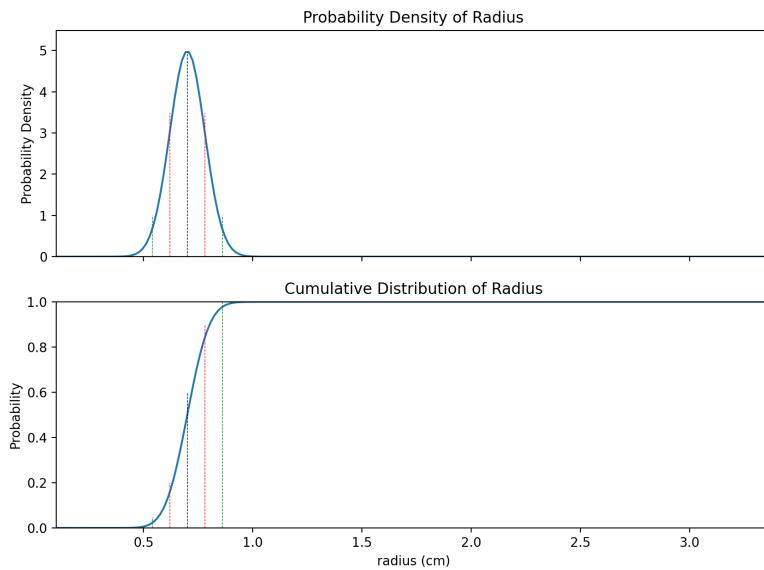
Given its properties, the Normal distribution serves as a foundation for many statistical procedures and concepts, including hypothesis testing, confidence intervals, and linear regression analysis.



CHAPTER 124

Change of Variables

Let's say that I'm making ice spheres, and I tell you that the radius of the ice spheres is normally distributed with a mean of 0.7 cm and a standard deviation of 0.08 cm. Then you can draw the probability distribution and cumulative distribution for that:



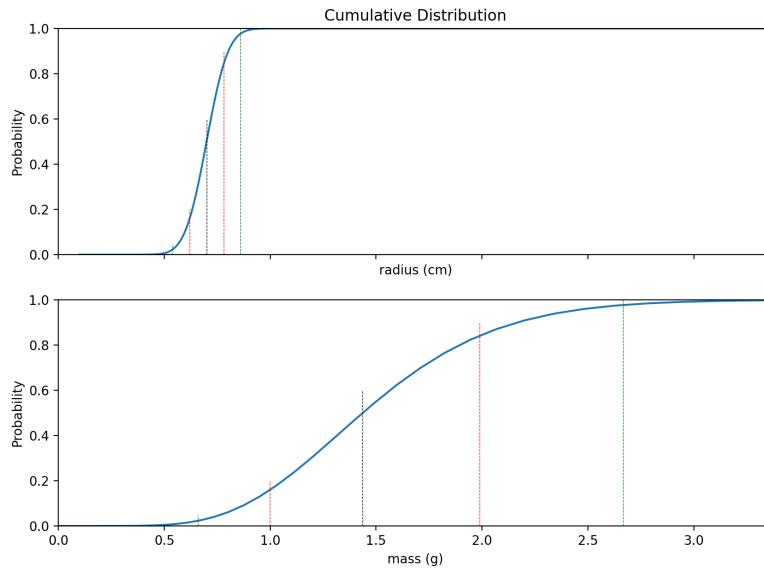
This includes lines indicating the mean and two standard deviations on each side.

Now, let's say I ask you what the cumulative distribution is for the *mass* of the balloons. A cubic centimeter of ice weighs about gram, so if you know the radius of a particular ice sphere, it is easy to compute the mass of it:

$$m = \frac{4}{3}\pi r^3$$

So, for example, if a sphere has a radius of 5cm, its mass in grams is $\frac{4\pi(0.7^3)}{3} \approx 1.44$ g.

Thinking about the graph of the cumulative distribution: if half the balloons have a radius less than 5 cm, than half the balloon have a mass less than 523.6 g. For each point on the cumulative graph, we can use the radius of that point to compute the corresponding mass – the CDF gets stretched out:



If F is the original cumulative distribution function, and g is the function that maps the new variable (mass, in this case) to the old one (radius), then the new cumulative distribution function H is given by

$$H(m) = F(g(m))$$

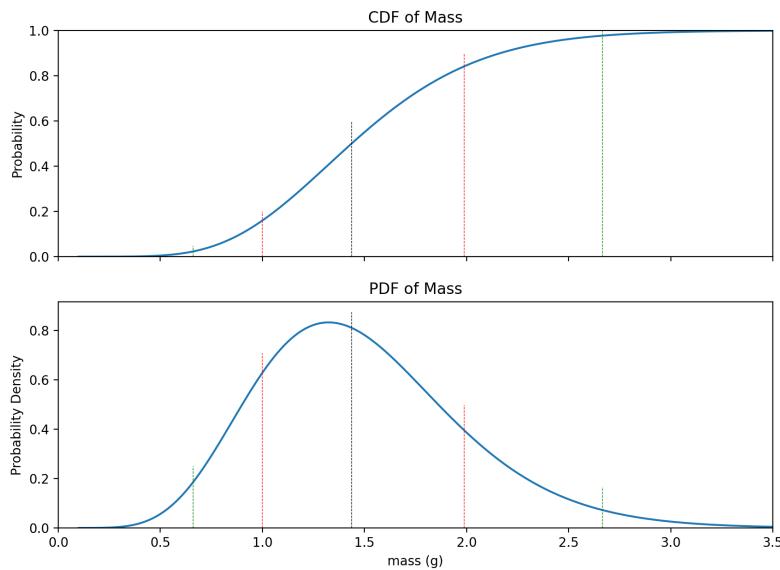
In this case, F is the cumulative function for the normal distribution with mean 0.7 and standard deviation of 0.08. g maps the mass to the radius:

$$g(m) = \left(\sqrt[3]{\frac{3}{4\pi}} \right) m^{\frac{1}{3}}$$

124.1 Making a Probability Density Function

Now we know how to calculate a new cumulative distribution function using the new variable. However, we usually want a probability density.

Here is the CDF and the PDF of the mass of the ice spheres:



Reminder: The probability density function is the derivative of the cumulative distribution function. We know the CDF is

$$H(m) = F(g(m))$$

By the chain rule:

$$H'(m) = F'(g(m))g'(m)$$

The function F is the cumulative distribution for the normal distribution with mean 0.7 and standard deviation of 0.08. So we know its derivative:

$$F'(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where $\mu = 0.7$ and $\sigma = 0.08$.

We've already said that

$$g(m) = \left(\sqrt[3]{\frac{3}{4\pi}}\right)m^{\frac{1}{3}}$$

Which is easy to differentiate:

$$g'(m) = \left(\frac{1}{3}\right) \left(\sqrt[3]{\frac{3}{4\pi}}\right) m^{-\frac{2}{3}}$$

Here, then, is the code to generate that last plot:

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

# Constants
MEAN_RADIUS = 0.7
STD_RADIUS = 0.08

# Range to plot
MIN_MASS = 0.1
MAX_MASS = 3.5

# Number of points to plot
N = 200

# Needed for radius_for_mass and d_radius_for_mass
C = np.power(3 / (4 * np.pi), 1/3)

# In these three functions, x can
# be a number or a numpy array

def mass_for_radius(x):
    return 4 * np.pi * np.power(x, 3) / 3

def radius_for_mass(x):
    return C * np.power(x, 1/3)

# Derivative of radius_for_mass()
def d_radius_for_mass(x):
    return (C/3) * np.power(x,-2/3)

# Compute mean and 2 standard deviations in each direction
m_mean = mass_for_radius(MEAN_RADIUS)
m_minus_std = mass_for_radius(MEAN_RADIUS - STD_RADIUS)
m_plus_std = mass_for_radius(MEAN_RADIUS + STD_RADIUS)
m_minus_2std = mass_for_radius(MEAN_RADIUS - 2 * STD_RADIUS)
m_plus_2std = mass_for_radius(MEAN_RADIUS + 2 * STD_RADIUS)

# Make N possible values for mass
m_values = np.linspace(MIN_MASS, MAX_MASS, N)

# Compute g(m) for each of these masses
```

```
# That is: What is the radius for each of these masses?
r_values = radius_for_mass(m_values)

# Compute F(g(m)) for each of these masses
# That is: What is the cumulative distribution for each those radii?
cdf_values = norm.cdf(r_values, loc=MEAN_RADIUS, scale=STD_RADIUS)

# Compute g'(m) for each of these masses
dg_values = d_radius_for_mass(m_values)

# What is F'(g(m))g'(m)?
pdf_values = norm.pdf(r_values, loc=MEAN_RADIUS, scale=STD_RADIUS) * dg_values

# Sanity check: It should integrate to a little less than 1.0
dx = (MAX_MASS - MIN_MASS)/N
area_under_curve = pdf_values.sum() * dx
print(f"Integral from {MIN_MASS:.2f} to {MAX_MASS:.2f}: {area_under_curve:.3f}")

# Make a figure with two axes
fig, axs = plt.subplots(nrows=2, sharex=True, figsize=(10, 7), dpi=200)

# Draw the CDF on the second axix
axs[0].set_title("CDF of Mass")
axs[0].set_ylim(bottom=0.0, top=1.0)
axs[0].set_xlim(left=0.0, right=MAX_MASS)
axs[0].set_ylabel("Probability")
axs[0].plot(m_values, cdf_values)

# Add lines for mean, mean-std, and mean+std
axs[0].vlines(m_minus_2std, 0, 0.05, "g", linestyle="dashed", lw=0.5)
axs[0].vlines(m_minus_std, 0, 0.2, "r", linestyle="dashed", lw=0.5)
axs[0].vlines(m_mean, 0, 0.6, "k", linestyle="dashed", lw=0.5)
axs[0].vlines(m_plus_std, 0, 0.9, "r", linestyle="dashed", lw=0.5)
axs[0].vlines(m_plus_2std, 0, 1.0, "g", linestyle="dashed", lw=0.5)

# How high does the pdf go?
max_density = pdf_values.max()

# Draw the PDF on the second axix
axs[1].set_title("PDF of Mass")
axs[1].set_ylim(bottom=0.0, top=max_density * 1.1)
axs[1].set_xlim(left=0.0, right=MAX_MASS)
axs[1].set_xlabel("mass (g)")
axs[1].set_ylabel("Probability Density")
axs[1].plot(m_values, pdf_values)

# Add lines for mean, mean-std, and mean+std
axs[1].vlines(m_minus_2std, 0, max_density * .3, "g", linestyle="dashed", lw=0.5)
axs[1].vlines(m_minus_std, 0, max_density * .85, "r", linestyle="dashed", lw=0.5)
axs[1].vlines(m_mean, 0, max_density * 1.05, "k", linestyle="dashed", lw=0.5)
axs[1].vlines(m_plus_std, 0, max_density * .6, "r", linestyle="dashed", lw=0.5)
axs[1].vlines(m_plus_2std, 0, max_density * .2, "g", linestyle="dashed", lw=0.5)
```

```
fig.savefig("pdf.png")
```

124.2 Decreasing Conversions

The last case (mass and radius) is pretty straightforward because the function g is always increasing. What if we have a change of variables where g is decreasing. For example, $V = IR$ so $\frac{V}{R} = I$.

Let's say that you work at a lightbulb factory and you sample the lightbulbs to see what their resistance is. You find the resistances of the lightbulbs are normally distributed with a mean of 24 ohms and a standard deviation of 3 ohms. The voltage will be exactly 12 volts. What is the PDF of the currents that will pass through the lightbulbs?

$$I = \frac{12}{R}$$

so

$$g(x) = \frac{12}{x}$$

is the function that will convert the current to resistance. Taking the derivative, we get:

$$g'(i) = -\frac{12}{x^2}$$



CHAPTER 125

Poisson and Exponential Probability Distributions

In this chapter, we will explore two essential probability distributions: the Poisson distribution and the exponential distribution. These distributions play a crucial role in modeling random events and phenomena, providing insights into the occurrence of events over time or in a discrete set of outcomes.

The Poisson distribution is widely used to describe the number of events that occur within a fixed interval of time or space. It is particularly useful when dealing with rare events or events that occur independently at a constant average rate. For example, the Poisson distribution can model the number of customer arrivals at a store in a given hour, the number of phone calls received by a call center in a day, or the number of defects in a production process.

The Poisson distribution is characterized by a single parameter, often denoted as λ , which represents the average rate of event occurrences in the specified interval. The probability mass function of the Poisson distribution gives the probability of observing a specific number of events within that interval.

On the other hand, the exponential distribution is used to model the time between events occurring at a constant average rate. It is commonly employed in reliability analysis, queuing theory, and survival analysis. For example, the exponential distribution can represent the time between customer arrivals at a service desk, the lifespan of electronic components, or the duration between consecutive earthquakes.

The exponential distribution is characterized by a parameter often denoted as λ , which represents the average rate of event occurrence. The probability density function of the exponential distribution describes the likelihood of observing a specific time interval between events.

In this chapter, we will explore the following key aspects of the Poisson and exponential probability distributions:

- Probability mass function and probability density function: We will dive into the mathematical representation of these distributions and learn how to calculate probabilities and densities for specific events or time intervals.
- Mean and variance: We will discuss how to calculate the mean and variance of the Poisson and exponential distributions, providing measures of central tendency and variability.
- Applications and examples: We will examine real-world scenarios where these distributions find practical applications. From analyzing customer arrival patterns to modeling equipment failure rates, we will explore a range of contexts where the Poisson and exponential distributions prove valuable.
- Relationship between the Poisson and exponential distributions: We will explore the connection between these distributions, as the exponential distribution can emerge as the waiting time between events following a Poisson process.
- Limitations and assumptions: We will also discuss the assumptions and limitations associated with the Poisson and exponential distributions, helping you understand when these models are suitable and when alternative approaches may be necessary.

By developing a solid understanding of the Poisson and exponential probability distributions, you will gain powerful tools for modeling and analyzing random events in various fields. These distributions provide valuable insights into event occurrences, time intervals, and rates, supporting decision-making processes and improving our understanding of stochastic phenomena.



CHAPTER 126

Multiple Integrals

In this chapter, we extend this powerful idea into higher dimensions using the tools of multiple integration. While single integration enables us to calculate the area under a curve or the volume under a surface, multiple integration allows us to calculate volumes in three dimensions, and even hypervolumes in higher dimensions.

We start by discussing double integration, which allows us to find the volume under a surface in three dimensions. This method involves slicing the solid into infinitesimally thin disks, and summing the volumes of these disks.

Next, we'll cover triple integration, a tool that lets us find the volume of more complicated solids in three-dimensional space. The idea is similar to double integration, but instead of slicing the solid into disks, we slice it into infinitesimally small cubes.

To properly implement these techniques, we'll also discuss the different coordinate systems that can be used in multiple integration, such as rectangular, cylindrical, and spherical coordinates, and when it's advantageous to use one system over another.

By the end of this chapter, you will have a deeper understanding of the techniques of multiple integration and how to apply them to find the volumes of various types of solids. The

methods we study here will serve as a foundation for many topics in higher mathematics and physics, including electromagnetism, fluid dynamics, and quantum mechanics.

Exercise 117 Using Polar Coordinates in Multiple Integration

Working Space

1. Use double integration to find the volume of the solid that lies under the surface $z = 4 - x^2 - y^2$ and above the xy -plane.

Answer on Page 846



CHAPTER 127

Multivariate Distributions

The world of probability and statistics doesn't limit itself to the study of single variables. Often, we are interested in the interconnections, relationships, and associations among several variables. In such a scenario, the univariate distributions that we have studied so far become inadequate. To comprehend the joint behavior of these variables and to uncover the underlying patterns of dependency, we must turn to the realm of multivariate distributions.

This chapter aims to introduce the reader to the concept of multivariate probability distributions. These are probability distributions that take into account and describe the behavior of more than one random variable. We will start our exploration with a discussion on the joint probability mass and density functions. These functions extend the concepts of probability mass and density functions for one variable to the situation where we have multiple variables.

Next, we will explore important properties of joint distributions, including the concept of marginal distribution and conditional distribution, which allow us to explore the probability of a subset of variables while conditioning on, or ignoring, other variables. We will also introduce the idea of independence of random variables in the multivariate context.

Subsequently, we will discuss some commonly used multivariate distributions such as the multivariate normal distribution, and the multivariate Bernoulli and binomial distributions. These specific distributions will provide us with practical tools for modelling multivariate data.

Finally, we will delve into covariance and correlation, two key measures that give us a sense of how two variables change together. Understanding these measures is critical for capturing the relationships in multivariate data.



CHAPTER 128

The Multivariate Normal Distribution

The multivariate normal distribution is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. It is used in statistics to describe any set of correlated real-valued random variables.

128.1 Multivariate Normal Distribution

A random vector $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$ follows a multivariate normal distribution if every linear combination of its components has a univariate normal distribution. The distribution is parameterized by a mean vector and a covariance matrix.

The probability density function (pdf) of an n -dimensional multivariate normal distribution is given by:

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

where:

- $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ is the point up to which the function is integrated,
- $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_n]^T$ is the mean vector,
- Σ is the covariance matrix,
- $|\Sigma|$ denotes the determinant of the covariance matrix,
- T denotes the matrix transpose.

128.2 Covariance Matrix

The covariance matrix, Σ , is a symmetric matrix that contains information about the variance of each variable and the covariance between every pair of variables in the distribution.

The element Σ_{ij} is the covariance between the i -th and the j -th random variable, and Σ_{ii} is the variance of the i -th random variable.

The covariance matrix provides a measure of how much each of the dimensions varies from the mean with respect to each other. A positive covariance between two variables indicates that the variables increase or decrease together, whereas a negative covariance indicates that one variable increases when the other decreases.



CHAPTER 129

Sets and Logic

The use of math usually falls into two categories:

- *Developing mathematical tools that let us make better predictions.* This is how engineers and scientists use math. It is usually referred to as *applied math*.
- *Creating interesting statements and proving them to be true or false.* This is known as *pure math*.

A lot of mathematical ideas start out as pure math, and eventually become useful. For example, the field of number theory is devoted to proving things about prime numbers. The mathematicians who created number theory were certain that it could never be used for any practical purpose. After a century or two, number theory was used as the basis for most cryptography systems.

Conversely, some ideas start out as a "rule of thumb" that engineers use and are eventually rigorously defined and proven.

This course tends to emphasize applied math, but you should know something about the tools of pure math.

You can think of all the mathematical proofs as a tree. Each proof proves some statement true. To do this, the proof uses logic and statements that were proven true by other truths. So the tree is built from the bottom up. However, the tree has to have a bottom: At the very bottom of the tree are some statements that we just accept as true without proof. These are known as *axioms*.

All of modern mathematics can be built from:

- A short list of axioms.
- A few rules of logic.

There have been several efforts to codify a small but complete axiomatic system. The most popular one is known as *ZFC*. "Z" is for the Ernst Zermelo, who did most of the work. "F" is for Abraham Fraenkel, who tidied up a couple of things. "C" is for The Axiom of Choice. As a community, mathematicians debate whether the Axiom of Choice should be an axiom; we get a couple of strange results if we include it in the system. If we don't, there are a few obviously useful ideas that we can't prove true.

ZFC has 10 axioms. We simply accept these 10 statements as true, and all the proofs of modern mathematics can be extrapolated from them. The 10 axioms are all stated in terms of sets.

129.1 Sets

A *set* is a collection. For example, you might talk about the set of odd numbers greater than 5. Or the set of all protons in the universe.

We have a notation for sets. For example, here is how define S to be the set containing 1, 2, and 3:

$$S = \{1, 2, 3\}$$

We say that 1, 2, and 3, are *elements* of the set S . (Sometimes we will also use the word "member")

If you want to say "2 is an element of the set S " in mathematical notation, it is done like this:

$$2 \in S$$

If you want to say "5 is *not* an element of the set S " it looks like this:

$$5 \notin S$$

We have notation for a few sets that we use all the time:

Set	Symbol
The empty set	\emptyset
Natural numbers	\mathbb{N}
Integers	\mathbb{Z}
Rational numbers	\mathbb{Q}
Real numbers	\mathbb{R}
Complex numbers	\mathbb{C}

The empty set is the set that contains nothing. It is also sometimes called *the null set*.

Often when we define a set, we start with one of these big sets and say "The set I'm talking about is the members of the big set, but only the one for which this statement is true". For example, if you wanted to talk about all the integers greater than or equal to -5, you could do it like this:

$$A = \{x \in \mathbb{Z} \mid x \geq -5\}$$

When you read this aloud, you say "A is the set of integers x where x is greater than or equal to negative 5."

129.1.1 And and Or

Sometimes you need the members to satisfy two conditions; for this we use "and":

$$A = \{x \in \mathbb{Z} \mid x > -5 \text{ and } x < 100\}$$

This is the set of integers that are greater than -5 *and* less than 100. In this book, we usually just write "and," but if you do a lot of set and logic work, you will use the symbol \wedge :

$$A = \{x \in \mathbb{Z} \mid (x > -5) \wedge (x < 100)\}$$

Sometimes you want a set that satisfies at least one of two conditions. For this you use "or":

$$A = \{x \in \mathbb{Z} \mid x < -5 \text{ or } x > 100\}$$

These are the numbers that are less than -5 or greater than 100. Once again, there is a symbol for this:

$$A = \{x \in \mathbb{Z} \mid (x < -5) \vee (x > 100)\}$$

129.1.2 How simple are sets?

Sets are so simple that some questions just don't make any sense:

- "What is the first item in the set?" makes no sense to a mathematician. Sets have no order.
- "How many times does the number 6 appear in the set?" makes no sense. 6 is a member, or it isn't.

129.1.3 Subsets

If every member of set A is also in set B, we say that "A is a subset of B".

For example, if $A = \{1, 4, 5\}$ and $B = \{1, 2, 3, 4, 5, 6\}$, then A is a subset of B. There is a symbol for this:

$$A \subseteq B$$

Remember the table of commonly used sets? We can arrange them as subsets of each other:

$$\emptyset \subseteq \mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$$

Note that subsets have the transitive property: $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q}$ thus $\mathbb{N} \subseteq \mathbb{Q}$

Note that if A and B have the same elements, $A \subseteq B$ and $B \subseteq A$. In this case, we say that the two sets are equal.

We also have a symbol for "is not a subset of": $A \not\subseteq B$

129.1.4 Union and Intersection of Sets

If you have two sets A and B , you might want to say "Let C be the set containing elements that are in *either* A or B ." We say that C is the *union* of A and B . There is notation for this too:

$$C = A \cup B$$

For example, if $A = \{1, 3, 4, 9\}$ and $B = \{3, 4, 5, 6, 7, 8\}$ then $A \cup B = \{1, 3, 4, 5, 6, 7, 8, 9\}$.

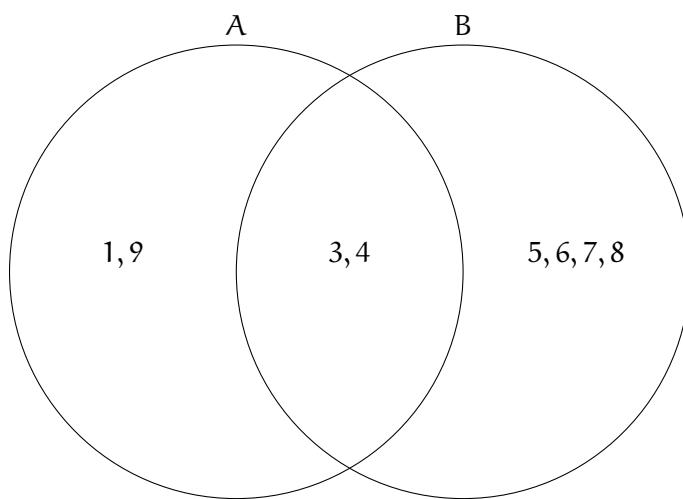
You also want to say "Let C be the set containing elements that are in *both* A and B ." We say that C is the *intersection* of A and B . There is notation for this too:

$$C = A \cap B$$

For example, if $A = \{1, 3, 4, 9\}$ and $B = \{3, 4, 5, 6, 7, 8\}$ then $A \cap B = \{3, 4\}$.

129.1.5 Venn Diagrams

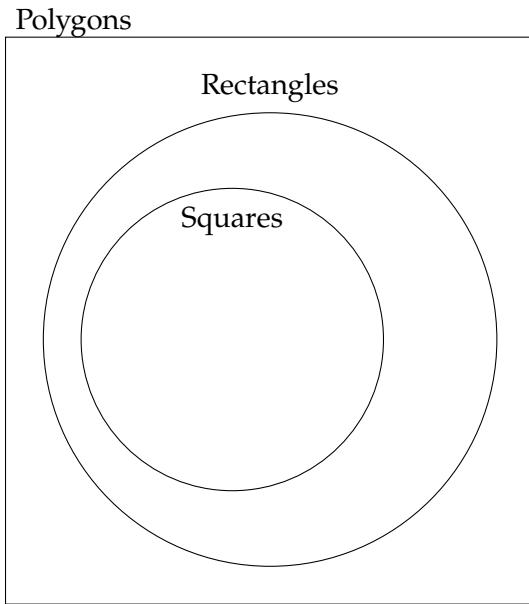
When discussing sets, it is often helpful to have a Venn diagram to look at. Venn diagrams represent sets as circles. For example, the sets A and B above could look like this:



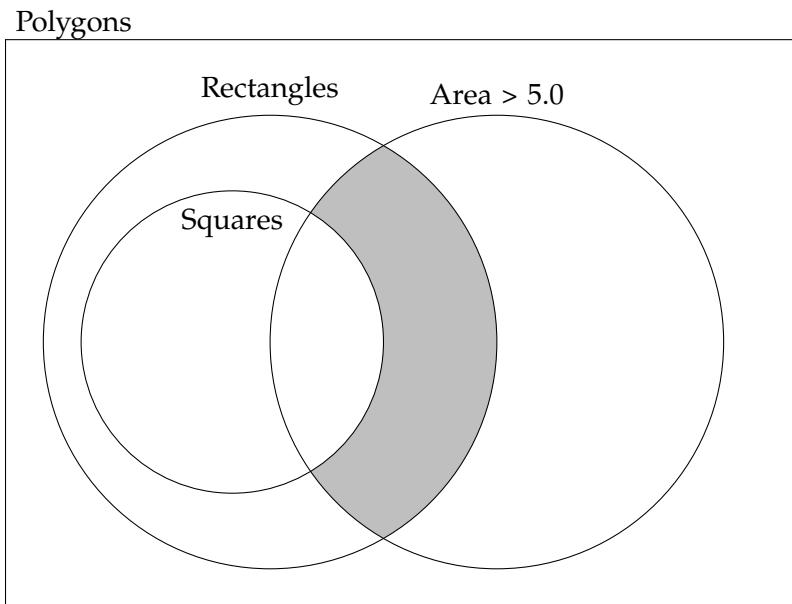
It makes it easy to see that A and B have a non-empty intersection, but they are not subsets of each other.

Often we won't even show the individual elements. For example, in the universe of all

polygons, some rectangles are squares. Here's the Venn diagram:



As the combinations get more complex, we sometimes use shading to indicate what part we are talking about. For example, imagine we wanted all the rectangles with area greater than 5.0 that are not squares. The diagram might look like this:



129.2 Logic

We use a lot of logic in set theory. For example, the shaded region above represents all the polygons for which all the following are true:

- It is a rectangle.
- It is *not* a square.
- It has an area greater than 5.0.

129.3 Implies

In logic, we will often say “ a implies b ”. That means “If the statement a is true, the statement b is also true.” For example: “ p is a square” implies “ p is a rectangle.”

There is notation for this: An arrow in the direction of the implication.

$$p \text{ is a square} \implies p \text{ is a rectangle}$$

Notice that implication has a direction: “ p is a rectangle” does *not* imply “ p is a square.”

Implications can be chained together: If $A \implies B$ and $B \implies C$, then $A \implies C$.

129.4 If and Only If

If the implication goes both ways, we use “if and only if”. This means the two conditions are equivalent. For example: “ n is even if and only if there exists an integer m such that $2m = n$ ”

There is a notation for this too:

$$p \text{ is even} \iff \text{there exists an integer } m \text{ such that } 2m = n$$

There is even notation for “there exists”. It is a backwards capital E:

$$p \text{ is even} \iff \exists m \in \mathbb{Z} \text{ such that } 2m = n$$

129.5 Not

The not operation flips the truth of an expression:

- If a is true, $\text{not}(a)$ is false.

- If a is false, $\text{not}(a)$ is true.

We sometimes talk about “notting” or “negating” a value. We won’t use it much, but there is a symbol for this: \neg .

We might create a *logic table* for negation that shows all the possible values and their negation:

A	$\neg A$
F	T
T	F

This table says “If A is false, $\neg A$ is true. If A is true, $\neg A$ is false.”

Most logic tables are for operations that take more than one input. For example, this logic table shows the values for and-ing and or-ing:

A	B	A and B	A or B
F	F	F	F
F	T	F	T
T	F	F	T
T	T	T	T

Notice that we have to enumerate all possible combinations of the inputs of A and B.

When a variable like A can only take two possible values, we say it is a *boolean* variable. (George Bool did important work in this area.)

Exercise 118 Logic Table**Working Space**

Make a logic table that enumerates all possible combinations of boolean variables A and B and shows the value of the two following expressions:

- $\neg(A \text{ or } B)$
- $(\neg A) \text{ and } (\neg B)$

Answer on Page 847**129.6 Cardinality**

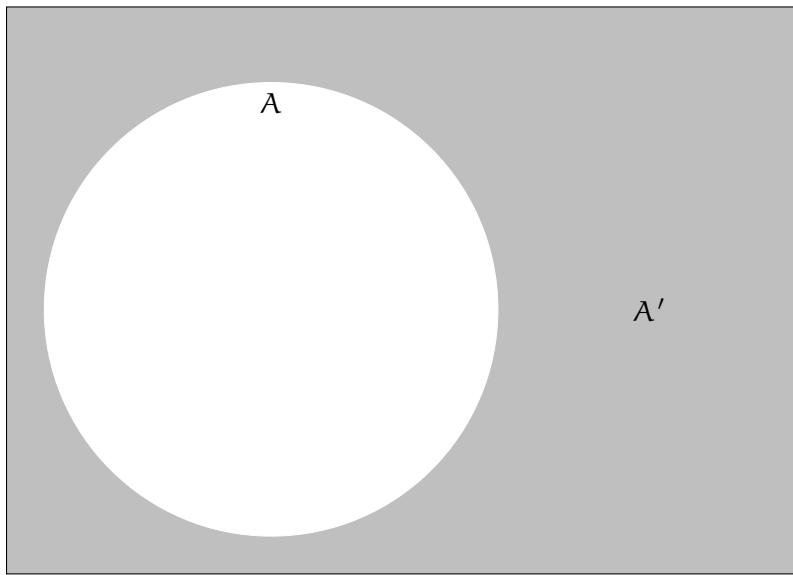
Informally, the *cardinality* of a set is the number of elements it contains. So, $\{1, 3, 5\}$ has a cardinality of 3. The null set has a cardinality of zero.

Things get a little trickier if a set is infinite. We say two infinite sets A and B have the same cardinality if there is some mapping that pairs every member of A with a member of B and mapping that pairs every member of B with a member of A.

129.7 Complement of a Set

Most sets exist in a particular universe, for example you might talk about the even numbers as a set in the integers. Then you can talk about the set's *complement*: the set of everything else. For example, the complement of the even numbers (inside the integers) is the odd numbers.

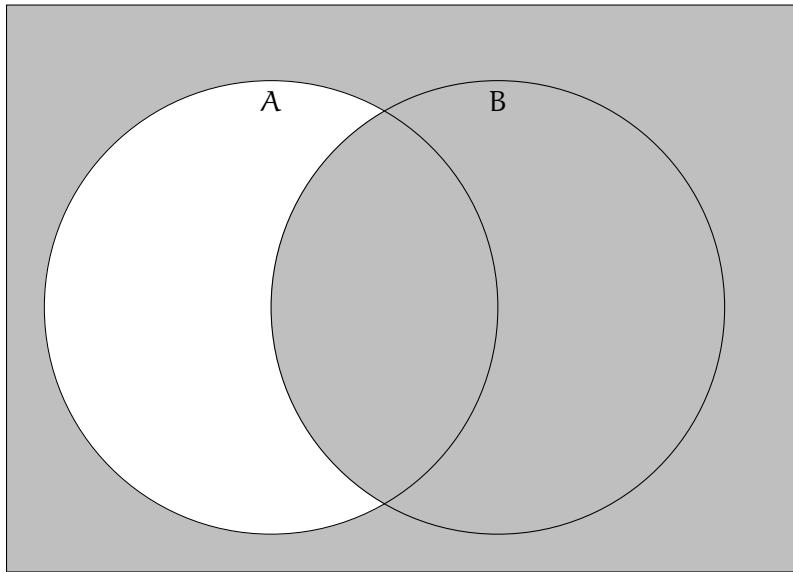
If you have a set A, its complement is usually denoted by A' .



129.8 Subtracting Sets

If you have sets $A = \{1, 2, 3, 4\}$ and $B = \{1, 4\}$, it makes sense to subtract B from A by removing 1 and 4 from A .

If A and B are sets, we define $A - B$ to be $A \cap B'$. Take a second to look at this diagram and convince yourself that the white region represents $A - B$ and that it is the same as $A \cap B'$.



129.9 Power Sets

It is not uncommon to have a set whose elements are also sets. For example, you might have the set that contains the following two sets: $\{1, 2, 3\}$ and $\{2, 3, 4\}$. You might write it like this: $\{\{1, 2, 3\}, \{2, 3, 4\}\}$. (Note that this set has a cardinality of 2 – It has two members that are sets.)

Given any set A, you can construct its *power set*, which is the set of all subsets of A. For example, if you have a set $\{1, 2, 3\}$, its power set is $\{\{1, 2, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1\}, \{2\}, \{3\}, \emptyset\}$.

If a set has n elements, its power set has 2^n elements.

129.10 Booleans in Python

In python, we can have variables hold boolean values: `True` and `False`. We also have operators: `not`, `and`, and `or`.

For example, You could find out what the expression “`a and not b`” is if both variables are false like this:

```
a = False
b = False
result = a or not b
print(f"a={a}, b={b}, a or not b = {result}")
```

This would print out:

```
a=False, b=False, a or not b = True
```

What if you wanted to try all possible values for a and b? You could use `itertools`.

```
import itertools

all_combos = itertools.product([False, True], repeat=2)
for (a, b) in all_combos:
    result = a or not b
    print(f"a={a}, b={b}: a or not b = {result}")
```

Type it in and run it. You should get the whole logic table:

```
a=False, b=False: a or not b = True
```

```
a=False, b=True: a or not b = False
a=True, b=False: a or not b = True
a=True, b=True: a or not b = True
```

If you had three inputs into the expression, your truth table would have eight entries. For example, if you wanted to know the truth table for `aandnot(bandc)`, here is the code:

```
all_combos = itertools.product([False, True], repeat=3)
for (a, b, c) in all_combos:
    result = a and not(b and c)
    print(f"{'a='}{a}, {'b='}{b}, {'c='}{c}: a and not (b and c) = {result}")
```

Type it in and run it. You should get:

```
a=False, b=False, c=False: a and not (b and c) = False
a=False, b=False, c=True: a and not (b and c) = False
a=False, b=True, c=False: a and not (b and c) = False
a=False, b=True, c=True: a and not (b and c) = False
a=True, b=False, c=False: a and not (b and c) = True
a=True, b=False, c=True: a and not (b and c) = True
a=True, b=True, c=False: a and not (b and c) = True
a=True, b=True, c=True: a and not (b and c) = False
```

129.11 The Contrapositive

Here is a statement with an implication: “If it has rained in the last hour, the grass is wet.”

This is *not* equivalent to “If the grass is wet, it has rained in the last hour.” (After all, the sprinkler may be running.)

However, it is exactly equivalent to its *contrapositive*: “If the grass is not wet, it has not rained in the last hour.”

The rule can be written using symbols:

$$(A \implies B) \iff (\neg B \implies \neg A)$$

129.12 The Distributive Property of Logic

Many ideas from integer arithmetic have analogues in boolean arithmetic. For example, there is a distributive property for booleans. These two expressions are equivalent:

- A and (B or C)
- (A and B) or (A and C)

So are these:

- A or (B and C)
- (A or B) and (A or C)

129.13 Exclusive Or

The expression “a or b” is true in any of the following conditions:

- a is True and b is False.
- a is False and b is True.
- Both a and b are True.

Sometimes engineers need a way to say “Either a or b is true, but not both.” For this we use *exclusive OR* (or XOR).

Here, then, is the logic table for XOR

A	B	XOR(a,b)
F	F	F
F	T	T
T	F	T
T	T	F

In python, Logical XOR is done using !=:

```
just_one = (a != b)
```

(Take 10 seconds to confirm that this is the same as the logic table above.)



CHAPTER 130

Linked Lists

A linked list is a linear data structure where each element is a separate object, called a node. Each node holds its own data and the address of the next node, thus forming a chain-like structure.

A simple node in a linked list can be represented in C++ as follows:

```
struct Node {  
    int data;  
    Node* next;  
};
```

In this structure, 'data' is used to store the data and 'next' is a pointer that holds the address of the next Node in the list.

Here is a simple example of creating and linking nodes in a linked list:

```
// Create nodes  
Node* head = new Node();  
Node* second = new Node();  
Node* third = new Node();
```

```
// Assign data
head->data = 1;
second->data = 2;
third->data = 3;

// Link nodes
head->next = second;
second->next = third;
third->next = nullptr; // The last node points to null
```

In this example, we first create three nodes using the ‘new’ keyword, which dynamically allocates memory. We then assign data to the nodes and link them using the ‘next’ pointer.



CHAPTER 131

Trees

Trees are one of the most versatile and widely used data structures in computer science. A tree is a hierarchical data structure consisting of nodes, where each node has a value and a set of references to its child nodes. The node at the top of the hierarchy is called the root, and nodes with the same parent are called siblings.

The power of trees comes from their ability to represent complex relationships between objects, while providing efficient operations for accessing and modifying those objects. Trees can be used to represent hierarchical relationships, to organize data for quick search and insertion, and to manage sorted lists of data, among other uses.

In this chapter, we will delve into the details of the tree data structure. We will start with the definition and properties of trees, including the key concepts of roots, nodes, children, siblings, leaves, and levels. We will then introduce binary trees, a specific type of tree where each node has at most two children, which are referred to as the left child and the right child.

We will explore the various ways to traverse a tree, including depth-first and breadth-first traversals, and discuss the applications and efficiencies of these methods. We will then cover binary search trees, a variant of binary trees that allows for fast lookup, addition,

and removal of items.

Then, we'll take a look at balanced search trees, such as AVL trees and red-black trees, which automatically keep their height small to guarantee logarithmic time complexity in the worst case for search, insert, and delete operations.

Finally, we will explore more advanced topics such as B-trees, tries, and suffix trees, which have applications in databases, file systems, and string algorithms.

By the end of this chapter, you will have a deep understanding of the tree data structure, its variants, and their uses. Armed with this knowledge, you'll be able to choose the right tree structure for your data and implement it effectively in your software.



CHAPTER 132

Searching Trees



CHAPTER 133

Hash Tables

A hash table, also known as a hash map, is a data structure that implements an associative array abstract data type, a structure that can map keys to values. It uses a hash function to compute an index into an array of buckets or slots, from which the desired value can be found.

133.1 Structure of a Hash Table

A hash table is composed of an array (the 'table') and a hash function. The array has a predetermined size, and each location (or 'bucket') in the array can hold an item (or several items if collisions occur, as will be discussed later). The hash function is a function that takes a key as input and returns an integer, which is then used as an index into the array.

133.2 Inserting and Retrieving Data

When inserting a key-value pair into the hash table, the hash function is applied to the key to compute the index for the array. The corresponding value is then stored at that

index.

When retrieving the value associated with a key, the hash function is applied to the key to compute the array index, and the value is retrieved from that index.

133.3 Handling Collisions

A collision occurs when two different keys hash to the same index. There are several methods for handling collisions:

- **Chaining (or Separate Chaining):** In this method, each array element contains a linked list of all the key-value pairs that hash to the same index. When a collision occurs, a new key-value pair is added to the end of the list.
- **Open Addressing (or Linear Probing):** In this method, if a collision occurs, we move to the next available slot in the array and store the key-value pair there. When looking up a key, we keep checking slots until we find the key or reach an empty slot.

133.4 Time Complexity

In an ideal scenario where hash collisions do not occur, hash tables achieve constant time complexity $O(1)$ for search, insert, and delete operations. However, due to hash collisions, the worst-case time complexity can become linear $O(n)$, where n is the number of keys inserted into the table.

Using good hash functions and collision resolution strategies can minimize this issue and allow us to take advantage of the hash table's efficient average-case performance.



CHAPTER 134

Sorting Algorithms

Sorting is a fundamental problem in computer science that has been extensively studied for many years. Sorting is the process of arranging items in ascending or descending order, based on a certain property. In the realm of algorithms, sorting generally refers to the process of rearranging an array of elements according to a specific order. This order could be numerical (ascending or descending) or lexicographical, depending on the nature of the elements.

Sorting algorithms form the backbone of many computer science and software engineering tasks. They are used in a myriad of applications including, but not limited to, data analysis, machine learning, graphics, computational geometry, and optimization algorithms. Thus, understanding these algorithms, their performance characteristics, and their suitability for specific tasks is crucial for anyone venturing into these fields.

This chapter will introduce several sorting algorithms, ranging from elementary methods like bubble sort and insertion sort to more advanced algorithms such as quicksort, mergesort, and heapsort. We will study these algorithms in terms of their time and space complexity, stability, and adaptability, among other characteristics. By the end of this chapter, you should have a solid understanding of how different sorting algorithms work and how to choose the appropriate algorithm for a specific context.

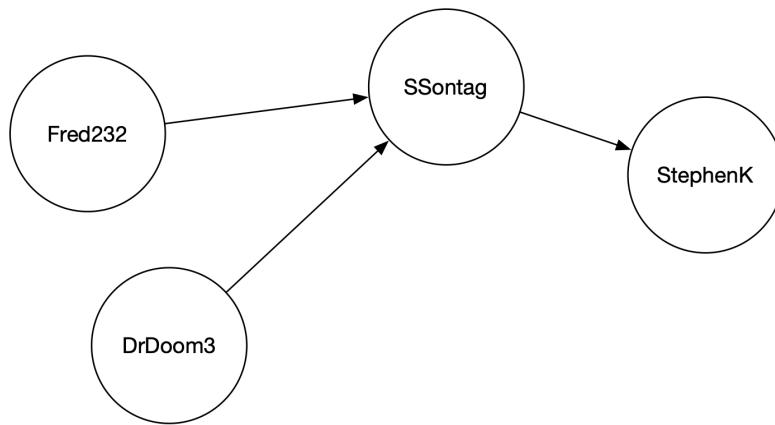
The knowledge of sorting algorithms not only helps in writing efficient code but also strengthens your problem-solving ability and analytical thinking, which are essential skills for succeeding in any technical interview. Let's dive into this fascinating world of sorting algorithms.



CHAPTER 135

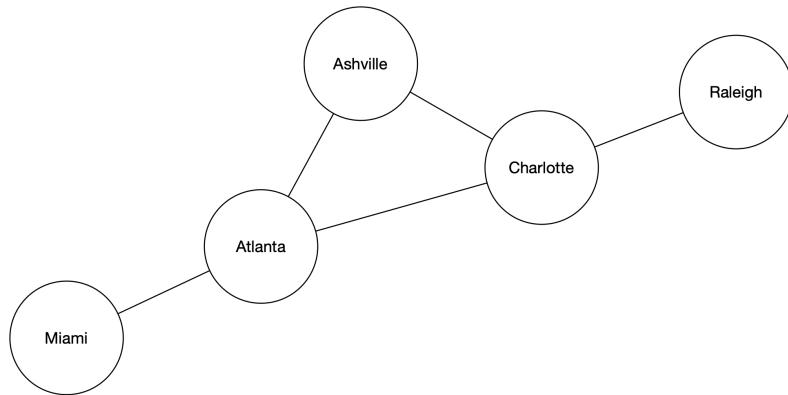
Introduction to Graphs

Some data is easiest to work with if we imagine it as a set of *nodes* connected by *edges*. For example, on some social networks each user can follow any number of other users. We can think of each user as node and the edge points from the user who follows to the user they follow:



This diagram shows four users and three follows. Following is a directed relationship: Fred232 follows SSontag, but SSontag doesn't follow Fred232. So we would say that this is a *directed graph* with four nodes and three edges.

There are also undirected graphs. for example, you can imagine a graph that represents big data lines between cities. All the big data lines allow communications in both directions:



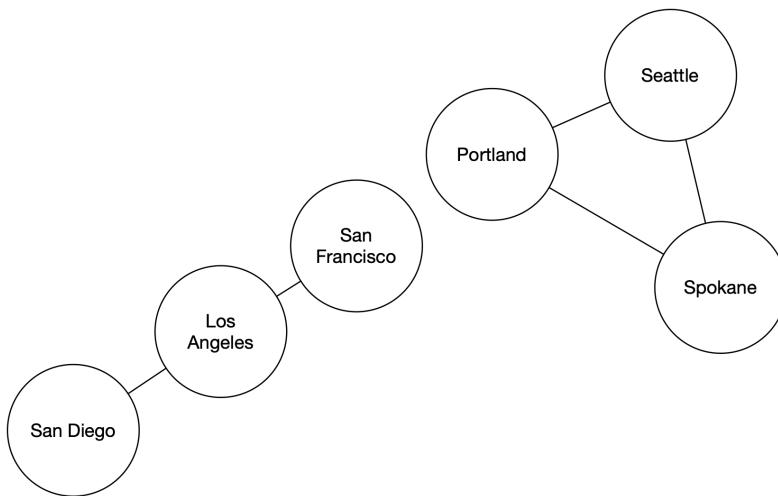
The arrows are gone: if data can flow from Charlotte to Raleigh, then data can flow from Raleigh to Charlotte.

There is a whole branch of mathematics called *Graph Theory* that studies the properties of graphs. Here are two questions that we might ask about this graph:

- What is the shortest number of edges that we would need to follow to get from Miami to Raleigh?
- Does the graph have any paths where you could end up where you started? This is called a *cycle*. This graph has one cycle: Atlanta → Asheville → Charlotte → Atlanta.

There are even database systems that are specifically designed to hold and analyze graph data. Not surprisingly, these are called *Graph Databases*.

Some graphs are *connected*: you can get from one node to any other node by following edges. Is this graph connected?



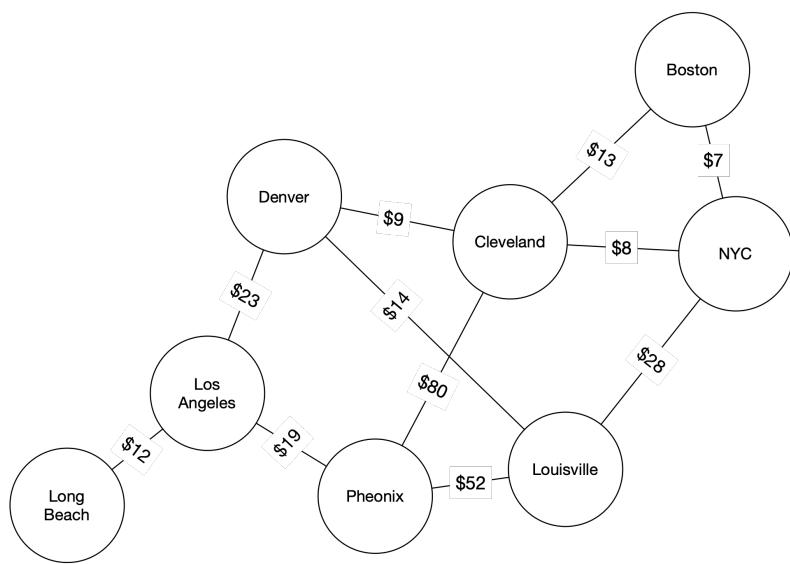
This graph is *not* connected! You can't follow edges from San Diego to Seattle.

In graph data, the nodes and edges often have attributes. For example, a node representing a city might have a name and a population. An edge representing a data line might have a bandwidth (bits per second) and a latency (how many nanoseconds between when you put a bit into the pipe and when it comes out the other end.).

135.1 Finding Good Paths

For a lot of problems, we are trying to find the best path from one node to another. If all the edges are the same, this usually means finding the path that requires walking the fewest edges.

Sometimes the edges have a cost attribute. For example, you might want to find the cheapest way to ship a container from New York City to Long Beach, Calif. In this case the nodes are train depots. Each train line between the depots has a cost. What is the cheapest path?



When edges have costs like this, we call the *weighted edges*.

The graphs that you see here are really small, so finding efficient paths isn't difficult. – you could just try all of them! However, in many computer programs, we are working with millions of nodes and edges. Efficient graph algorithms are *really* important.

135.2 Graphs in Python

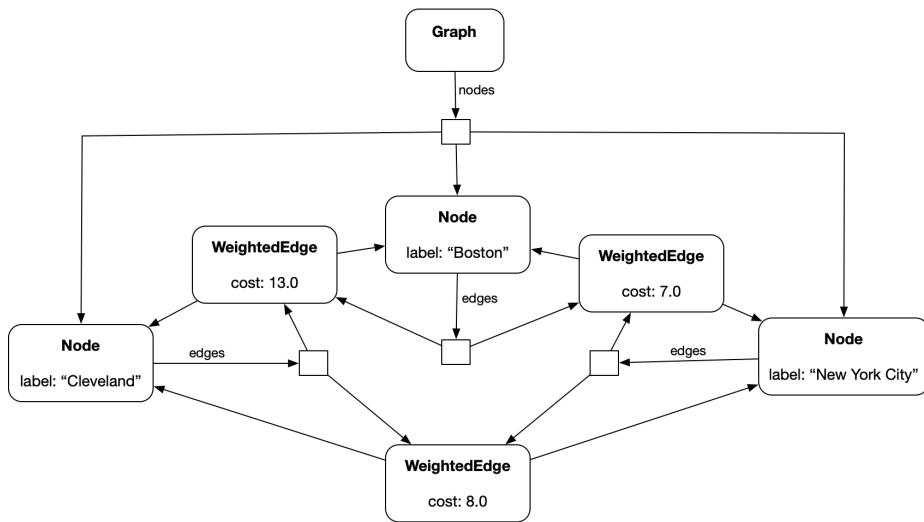
In this section you are going to write Python classes that will let you represent an undirected graph with weighted edges, like the shipping problem above.

(Naturally things would look a little different if the graph were directed or the edges were unweighted, but this is a good starting place.)

Create a file called `graph.py`. This will hold the code for your `Node` and `WeightedEdge` classes. We will also create a `Graph` class that will just hold onto the list of its nodes.

- A `Node` will have a label string and a list of edges that touch it.
- A `Edge` will have a cost and two nodes: `node_a` and `node_b`.
- A `Graph` will have a list of nodes.

Here what the object diagram would look like if you had only three cities:



Put this code into graph.py

```

class Node:
    def __init__(self, label):
        self.label = label
        self.edges = []

    def __repr__(self):
        return f"(node:{self.label}, edges:{len(self.edges)})"

class WeightedEdge:
    def __init__(self, cost, node_a, node_b):
        self.cost = cost
        self.node_a = node_a
        node_a.edges.append(self)
        self.node_b = node_b
        node_b.edges.append(self)

    def other_end(self, node_from):
        if self.node_a == node_from:
            return self.node_b
        else:
            return self.node_a

class Graph:
    def __init__(self):
        self.nodes = []

    def add_node(self, new_node):
        self.nodes.append(new_node)

    def __repr__(self):
        return f"(Graph:{self.nodes})"
  
```

Now lets create some instances of `Node` and `WeightedEdge` and wire them together. Create another file in the same directory called `cities.py`. Put in this code:

```
import graph

# Create an empty graph
network = graph.Graph()

# Create city nodes and add to graph
long_beach = graph.Node("Long Beach")
network.add_node(long_beach)
los_angeles = graph.Node("Los Angeles")
network.add_node(los_angeles)
denver = graph.Node("Denver")
network.add_node(denver)
pheonix = graph.Node("Pheonix")
network.add_node(pheonix)
louisville = graph.Node("Louisville")
network.add_node(louisville)
cleveland = graph.Node("Cleveland")
network.add_node(cleveland)
boston = graph.Node("Boston")
network.add_node(boston)
nyc = graph.Node("New York City")
network.add_node(nyc)

# Create edges
graph.WeightedEdge(12, long_beach, los_angeles)
graph.WeightedEdge(23.0, los_angeles, denver)
graph.WeightedEdge(19, los_angeles, pheonix)
graph.WeightedEdge(52, pheonix, louisville)
graph.WeightedEdge(14, denver, louisville)
graph.WeightedEdge(80, pheonix, cleveland)
graph.WeightedEdge(9, denver, cleveland)
graph.WeightedEdge(8, cleveland, nyc)
graph.WeightedEdge(28, louisville, nyc)
graph.WeightedEdge(7, nyc, boston)
graph.WeightedEdge(13, cleveland, boston)

print(network)
```

Run it:

```
python3 cities.py
```

You should see some rather unexciting output:

```
(Graph:[(node:Long Beach, edges:1), (node:Los Angeles, edges:3), (node:Denver, edges:3),
```

```
(node:Phoenix, edges:3), (node:Louisville, edges:3), (node:Cleveland, edges:4),  
(node:Boston, edges:2), (node>New York City, edges:3)])
```

But we will make it more exciting in the next chapter!



CHAPTER 136

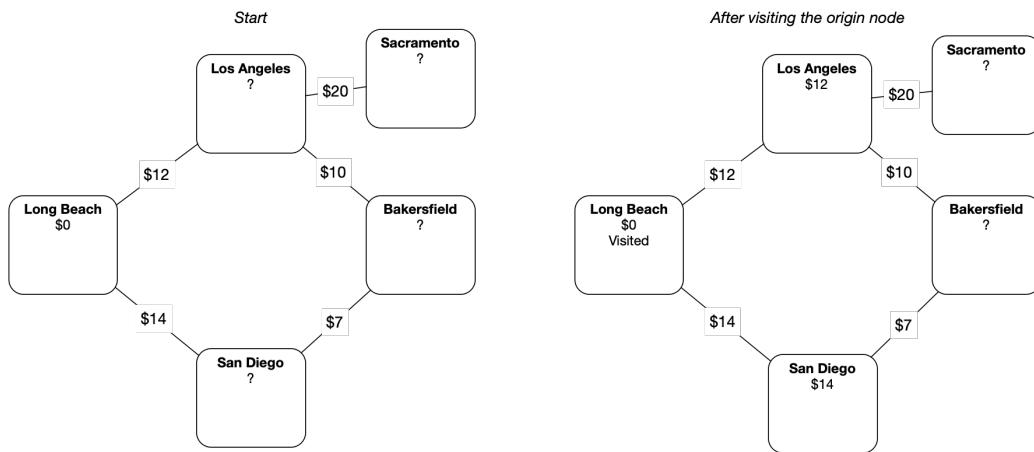
Dijkstra's Algorithm

Edsger W. Dijkstra was a great Dutch computer scientist. He came up with an algorithm for finding the cheapest path through a graph with weighted edges. Today it is known as *Dijkstra's Algorithm*. It is used in a wide variety of common problems. It is also really pretty simple and elegant.

136.1 Algorithm Description

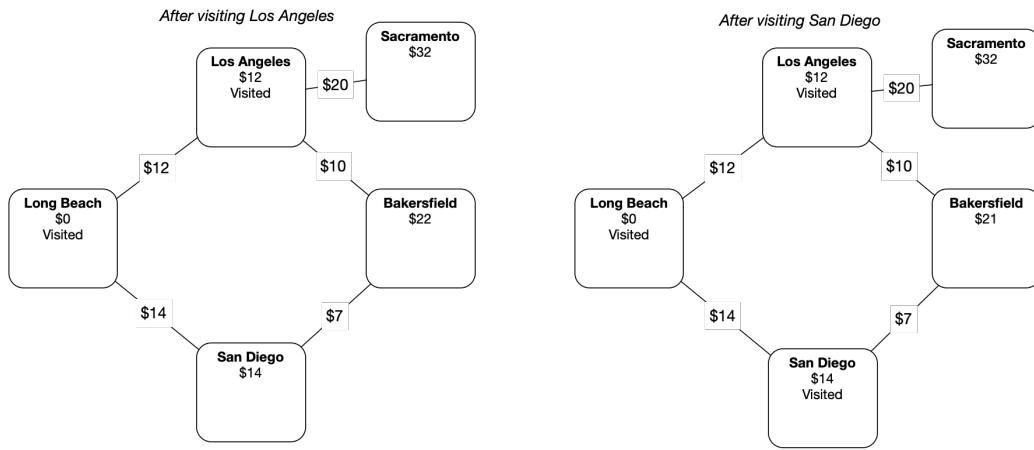
You are going to mark each node with how much it would cost to get there from some origin node. For example, if you are shipping a container from Long Beach, you will mark each city with the cost of getting the container to that city.

You start by marking the price for Long Beach to zero. (The container is already there.) Then, you mark each adjacent city with the cost on the edge. Now you declare Long Beach to be “visited”.



Now, you find the cheapest of the unvisited nodes. In this case, Los Angeles is cheaper than San Diego, so that is the node you will visit next.

You mark all of the unvisited nodes adjacent to Los Angeles, with the price to ship it to Los Angeles plus the cost of shipping the container from Los Angeles to that city. Note that Bakersfield is marked with \$22.



Now the cheapest unvisited node is San Diego. So you mark its neighbors with the cost to ship to San Diego plus the price to ship from San Diego to the neighbor.

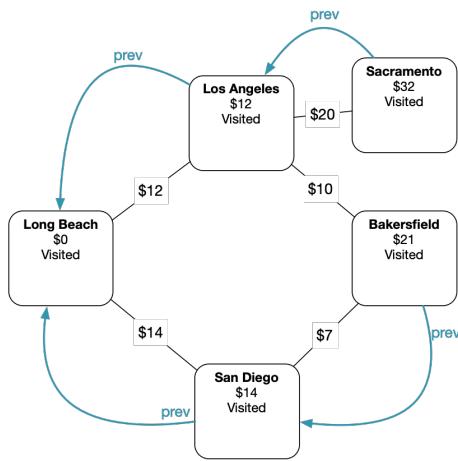
Notice that Bakersfield is already labeled with \$22 from a route through Los Angeles. But the price would be \$21 if you shipped it to Bakersfield via San Diego. Because the new route is cheaper, you change the price to the lower value.

(What does it mean that a node is “visited”? If a node is marked visited, it is marked with a price that won’t get any smaller.)

And you continue visiting the cheapest unvisited node until all the nodes have been visited. Then you know every node has been marked with its lowest price.

In a big graph, each node may be marked several times in this process – each time with a lower price from a cheaper router.

Of course, once you have the price, you will ask “What is the route that gets me that price?” So we will also mark each node with the neighbor from which it would receive the shipment – the previous node. This is easy to do as we execute the algorithm.



Now, to figure out the cheapest route from San Diego to Bakersfield, we start at the destination and follow the `prev` pointer back through San Diego and then to Long Beach.

136.2 Implementation

We don't actually want to sully our graph objects with the three additional pieces of information we need:

- The current minimal cost from the origin node. This is usually called the `dist`, from “distance”.
- The neighbor who gives us the current minimal cost. This is usually called `prev`, from “previous”.
- Whether the city is visited or now.

So we will keep them in collections external to the graph.

For example, to keep track of the `dist`, we will have a `dist` dictionary: Each node will be a key, the current minimal cost will be the value. If the node hasn't received even a first cost, we will put in infinity as the cost.

(After the algorithm is run, if the cost of a node is still infinity, that means that it cannot be reached from the origin node.)

We will also have a `prev` dictionary. The final node will be the key, and its previous neighbor will be the value.

Finally, the graph has a list of all the nodes, so we can just keep a set of the unvisited nodes.

Add a method to the `Graph` class that implements Dijkstra's algorithm:

```
def cost_from_node(self, origin_node):
    # Cost of cheapest path from origin node discovered so far
    # Initially the origin is zero and all the other are infinity
    dist = {k: math.inf for k in self.nodes}
    dist[origin_node] = 0.0

    # The previous city on that cheapest path
    prev = {}

    # All the nodes start as unvisited
    unvisited = set(self.nodes)

    # While there are still unvisited nodes
    while unvisited:

        # Find unvisited node with lowest cost
        min_cost = math.inf
        for u in unvisited:
            if dist[u] < min_cost:
                current_node = u
                min_cost = dist[u]

        # If none are less than inf, we are done
        # This happens in graphs that are not connected
        if min_cost == math.inf:
            return (dist, prev)

        # Remove the lowest cost node from the unvisited list
        unvisited.remove(current_node)

        # Update all the unvisited neighbors
        for edge in current_node.edges:

            # What node is at the other end of this edge?
            v = edge.other_end(current_node)

            # Visited nodes are already minimized, skip them
            if v not in unvisited:
                continue

            # Is this a shorter route?
            alt = dist[current_node] + edge.cost
            if alt < dist[v]:
```

```

# Update the distance and prev dicts
dist[v] = alt
prev[v] = current_node

return (dist, prev)

```

Append some code to your `cities.py` that test this method:

```

(cost_from_long_beach, prev) = network.cost_from_node(long_beach)
print(f"\nMinimum costs from Long Beach = {cost_from_long_beach}")
print(f"\nLast city before = {prev}")

nyc_cost = cost_from_long_beach[nyc]

if nyc_cost < math.inf:
    print(f"\n*** Total cost from Long Beach to NYC: ${nyc_cost:.2f} ***")
else:
    print("You can't get to NYC from Long Beach")

```

When you run it, you should get a list of how much it costs to ship a container to each city from Long Beach:

```

Minimum costs from Long Beach = {(node:Long Beach, edges:1): 0.0,
(node:Los Angeles, edges:3): 12.0, (node:Denver, edges:3): 35.0,
(node:Pheonix, edges:3): 31.0, (node:Louisville, edges:3): 49.0,
(node:Cleveland, edges:4): 44.0, (node:Boston, edges:2): 57.0,
(node>New York City, edges:3): 52.0}

```

You will also get a collection of node pairs. What are these? For each node, you get the node that you would pass through on the cheapest route from Long Beach:

```

Last city before = {(node:Los Angeles, edges:3):(node:Long Beach, edges:1),
(node:Denver, edges:3):(node:Los Angeles, edges:3),
(node:Pheonix, edges:3):(node:Los Angeles, edges:3),
(node:Louisville, edges:3):(node:Denver, edges:3),
(node:Cleveland, edges:4):(node:Denver, edges:3),
(node>New York City, edges:3):(node:Cleveland, edges:4),
(node:Boston, edges:2): (node:Cleveland, edges:4)}

```

Your users won't want to read this; Give them the shortest path as a list. Add a function to `graph.py` that turns the `prev` table into a path of nodes that lead from the origin to the destination:

```

def shortest_path(prev, destination):

```

```
# Include the destination in the path
path = [destination]
current_node = destination

# Keep stepping backward in the path
while current_node in prev:

    # What node should come before the current node?
    previous_node = prev[current_node]

    # Insert it at the start of the list
    path.insert(0, previous_node)
    current_node = previous_node

return path
```

Test that out:

```
if nyc_cost < math.inf:
    print(f"*** Total cost from Long Beach to NYC: ${nyc_cost:.2f} ***")

    path_to_nyc = graph.shortest_path(prev, nyc)
    print(f"*** Cheapest path from Long Beach to NYC: {path_to_nyc} ***")
else:
    print("You can't get to NYC from Long Beach")
```

This should look like this:

```
*** Cheapest path from Long Beach to NYC: [(node:Long Beach, edges:1),
(node:Los Angeles, edges:3), (node:Denver, edges:3), (node:Cleveland, edges:4),
(node>New York City, edges:3)] ***
```

136.3 Making it faster

On really big networks, doing a full Dijkstra's algorithm would take too long. So there are a lot of methods for getting similar results quickly. When you ask for directions from Google Maps, it doesn't do a full Dijkstra's Algorithm for every possible route – it would just take too long.

But there is a way to speed up this implementation. Look at this snippet:

```
# Find unvisited node with lowest cost
```

```
min_cost = math.inf
for u in unvisited:
    if dist[u] < min_cost:
        current_node = u
        min_cost = dist[u]
```

We are scanning through the list of all unvisited nodes, one-by-one, looking for the one with the lowest cost. If we kept this list sorted by cost, then the next one to visit would always be the first one in the list. This is done with a *priority queue* – a list that keeps itself sorted by some priority number – in this case the cost. In python, the standard priority queue is `heapq`.

(So why didn't I implement this using `heapq`? For Dijkstra's Algorithm, the nodes' priority – the current cost – changes as we find cheaper routes. `heapq` doesn't handle the changing priority very gracefully.)

In the next chapter, you will make a priority queue class that will work in this case.



CHAPTER 137

Binary Search

As mentioned in the last chapter, you are going to make a priority queue for use with Dijkstra's Algorithm. Using it will look like this:

```
import kpqueue

myqueue = pqueue.PriorityQueue()
myqueue.add(long_beach, 0) # Inserts first city and its cost
myqueue.add(san_diego, 14) # Puts San Diego after Long Beach
myqueue.add(los_angeles,12) # Inserts LA between Long Beach and San Diego
current_city = myqueue.pop() # Returns first city (Long Beach) and removes it
```

Now if an item gets a new priority, we need to remove it and reinsert it in the new spot.

```
myqueue.add(city_a, 16) # Puts it last in the queue
myqueue.update(city_a, 16, 13) # Moves it to between LA and San Diego
```

137.1 A Naive Implementation of the Priority Queue

Create a file called `kpqueue.py`. Let's do a simple implementation that stores the priority and the data as tuple. And we will keep it sorted by the priority. If two tuples have the same priority, we'll sort by the data.

Type this in to `kpqueue.py`:

```
class PriorityQueue:
    def __init__(self):
        self.list = []

    # Return and remove the first item
    def pop(self):
        if len(self.list) > 0:
            return self.list.pop(0)
        else:
            return None

    def __len__(self):
        return len(self.list)

    def update(self, value, old_priority, new_priority):
        old_pair = (old_priority, value)
        self.list.remove(old_pair)
        self.add(value, new_priority)

    def add(self, value, priority):
        pair = (priority, value)
        # Add it at the end
        self.list.append(pair)
        # Resort the list
        self.list.sort()
```

This will work fine, but it could be much more efficient:

- Every time we add a single element, we resort the whole list.
- The function `remove` is searching the list sequentially for the item to delete.

In a minute, we will revisit these inefficiencies and make the better.

137.2 Using the Priority Queue

We are going to change `graph.py` to use the priority queue. While we are doing, why don't we also shrink the memory footprint of our program a bit.

Notice that as the algorithm is running, each node is in one of three states:

- Unseen: In the earlier implementation, these were the nodes with `math.inf` as their cost.
- Seen, but not finalized: These are “unvisited” but don’t have `math.inf` as their cost.
- Finalized: These are the “visited” nodes – we know that their cost won’t decrease any more.

We can shrink the memory foot print by not putting the unseen into the `dist` dictionary at all. And instead of a separate set for “unvisited” what if we moved finalized nodes and their distances into a separate dictionary?

Rewrite the `cost_from_node` function in `graph.py`:

```
# Visited nodes are already minimized, skip them
if v in finalized_dist:
    continue

# What is the cost to this neighbor?
alt = current_node_cost + edge.cost

# Is this the first time I am seeing the node?
if v not in seen_dist:

    # Insert into the seen_dict, prev, and priority queue
    seen_dist[v] = alt
    prev[v] = current_node
    pqueue.add(v, alt)

else: # v has been seen. Is this a cheaper route?
    old_dist = seen_dist[v]
    if alt < old_dist:
        # Update the seen_dict, prev, and priority queue
        seen_dist[v] = alt
        prev[v] = current_node
        pqueue.update(v, old_dist, alt)

return (finalized_dist, prev)
```

This should be have exactly the same except for the unreachable nodes. If you have a graph that is not connected, there will be nodes that can’t be reached from the origin. In the old version, these had a cost of `math.inf`. Now they just won’t be in the dictionary at all. So, change `cities.py` to deal with this:

```
if nyc in cost_from_long_beach:
```

```
nyc_cost = cost_from_long_beach[nyc]
print(f"\n*** Total cost from Long Beach to NYC: ${nyc_cost:.2f} ***")

path_to_nyc = graph.shortest_path(prev, nyc)
print(f"\n*** Cheapest path from Long Beach to NYC: {path_to_nyc} ***")
else:
    print("You can't get to NYC from Long Beach")
```

If you run `cities.py` now, it should behave exactly like the old version.

But there is a bug. It will rear its head if two cities with the same cost are in the priority queue together. Change `cities.py` so that Denver and Pheonix have the same cost:

```
graph.WeightedEdge(12, long_beach, los_angeles)
graph.WeightedEdge(19, los_angeles, denver)
graph.WeightedEdge(19, los_angeles, pheonix)
```

Now try running it. You should get an error:

```
TypeError: '<' not supported between instances of 'Node' and 'Node'
```

What happened? The `loc_for_pair` method is comparing tuples made up of a `float` and a `Node`. The `float` comes first in the tuple, so that is compared first. However, if the two tuples have the same priority, it then compares nodes.

The error statement say “Nodes don’t have a less-than method; I don’t know how to compare them.”

Each `Node` lives at an address in memory. You can get that address as a number using the `id` function. The ID is unique and constant over the life of the object. It is a rather arbitrary ordering, but it will work for this problem. Add a method to your `Node` class:

```
# Nodes will be ordered by their location in memory
def __lt__(self, other):
    return id(self) < id(other)
```

Fixed.

Now let’s make the priority queue more efficient.

137.3 Binary Search

The phone company in every town used to print a thing called a phone book. The names and phone numbers were arranged alphabetically. As you might imagine, these books often had more than a thousand pages.

If you were looking for “John Jeffers”, you wouldn’t start at the first page and read sequentially until you reached his name. You would open the book in the middle, and see a name like “Mac Miller”, and then think “Jeffers comes before Miller”. Then you would split the pages in your left hand in half and see a name like “Hester Hamburg” and think “Jeffers comes after Hamburg”. Then you would split the pages in your right hand, and so on until you found the page with “John Jeffers” on it.

That is binary search.

Binary Search is a search algorithm that finds the position of a target value within a sorted array. The binary search algorithm works by repeatedly dividing the search interval in half. If the target value is equal to the middle element of the array, the position is returned. If the target value is less or greater than the middle element, the search continues in the lower or upper half of the array respectively.

137.4 Algorithm

The binary search algorithm can be described as follows:

1. If the array is empty, the search is unsuccessful, so return “Not Found”.
2. Otherwise, compare the target value to the middle element of the array.
3. If the target value matches the middle element, return the middle index.
4. If the target value is less than the middle element, repeat the search with the lower half of the array.
5. If the target value is greater than the middle element, repeat the search with the upper half of the array.
6. Repeat steps 2-5 until the target value is found or the array is exhausted.

```
class PriorityQueue:  
    def __init__(self):  
        self.list = []  
  
    # Return and remove the first item  
    def pop(self):  
        if len(self.list) > 0:
```

```
        return self.list.pop(0)
    else:
        return None

def __len__(self):
    return len(self.list)

def add(self, value, priority):
    pair = (priority, value)
    i = self.loc_for_pair(pair)
    self.list.insert(i, pair)

def update(self, value, old_priority, new_priority):
    old_pair = (old_priority, value)
    i = self.loc_for_pair(old_pair)
    del self.list[i]
    self.add(value, new_priority)

def loc_for_pair(self, pair):
    # The range where it could be is [lower, upper)
    # Start with the whole list
    lower = 0
    upper = len(self.list)

    while upper > lower:
        next_split = (upper + lower) // 2
        v = self.list[next_split]
        if pair < v:  # pair is to the left
            upper = next_split
        elif pair > v:  # pair is to the right
            lower = next_split + 1
        else: # Found pair!
            return next_split
    return lower
```

If you try running it now, it should work perfectly.

Now you have a graph class that would find the cheapest path quickly even if it had thousands of nodes with thousands of edges.



CHAPTER 138

Other Graph Algorithms

Now that you are familiar with Dijkstra's algorithm for finding the shortest path in a graph, then you're well-equipped to understand more graph algorithms. This document will discuss two other important algorithms: Depth-First Search (DFS) and the Bellman-Ford algorithm.

138.1 Depth-First Search

Depth-First Search (DFS) is an algorithm for traversing or searching tree or graph data structures. DFS uses a stack (or sometimes recursion which uses the system stack implicitly) to explore the graph in a depthward motion until it hits a node with no unvisited adjacent nodes, then it backtracks.

The procedure is as follows:

1. Push the root node into the stack.
2. Pop a node from the stack, and mark it as visited.

3. Push all unvisited adjacent nodes into the stack.
4. Repeat steps 2 and 3 until the stack is empty.

DFS is particularly useful for solving problems such as connected-component detection in graphs and maze-solving.

138.2 Bellman-Ford Algorithm

The Bellman-Ford algorithm is another shortest path algorithm like Dijkstra's. However, unlike Dijkstra's algorithm, Bellman-Ford can handle graphs with negative weight edges.

The algorithm works as follows:

1. Assign a tentative distance value for every node: set it to zero for our initial node and to infinity for all other nodes.
2. For each edge (u, v) with weight w , if the current distance to v is greater than the distance to u plus w , update the distance to v to be the distance to u plus w .
3. Repeat the previous step $|V| - 1$ times, where $|V|$ is the number of vertices in the graph.
4. After the above steps, if you can still find a shorter path, there exists a negative cycle.

If the graph does not contain a negative cycle reachable from the source, the shortest paths are well-defined, and Bellman-Ford will correctly calculate them. If a negative cycle is reachable, no solution exists, but Bellman-Ford will detect it.



CHAPTER 139

Bayesian Networks

A Bayesian network, also known as a Bayes network, belief network, or decision network, is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).

139.1 Components

A Bayesian Network consists of two main components:

1. A directed acyclic graph (DAG) where each node represents a variable, and the absence or presence of a directed edge between nodes denotes the conditional dependence or independence respectively between the variables.
2. A conditional probability table (CPT) associated with each node which contains the conditional probability distribution of that node given its parents in the DAG.

139.2 Inferences

Bayesian Networks are typically used for reasoning and making inferences under uncertainty. Given observations of a set of variables, we can compute the posterior probabilities of the other variables using Bayes' rule.

There are three main types of inferences that we can make:

- **Causal reasoning (prediction):** Given the causes, what are the effects?
- **Evidential reasoning (diagnosis):** Given the effects, what are the causes?
- **Intercausal reasoning (explaining away):** Given an effect and some of its causes, what can we say about the other causes?

139.3 Learning

Learning a Bayesian Network from data involves two main tasks:

- **Structure learning:** Determining the DAG structure that best fits the data.
- **Parameter learning:** Estimating the parameters (conditional probabilities) of the CPTs given the DAG and data.



CHAPTER 140

Introduction to Classification and Regression

Classification and regression are two types of supervised learning methods in machine learning and statistics. In supervised learning, the goal is to learn a mapping function from inputs x to an output y , given a labeled set of input-output pairs.

140.1 Classification Systems

In classification, the output y is a categorical or discrete value. For example, if we are developing a system to predict whether an email is spam or not, y can take two values: "spam" or "not spam". This is an example of a binary classification problem.

Classification problems that have more than two categories are known as multi-class classification problems. For example, predicting the species of an iris flower from a set of measurements of its petals and sepals is a multi-class classification problem, as there are three species of iris flowers.

140.2 Regression Systems

In regression, the output y is a continuous value. For example, if we are developing a system to predict the price of a house given features like its size, location, number of rooms, etc., the output is a continuous number which represents the price.

140.3 Algorithms

There are many algorithms used to solve classification and regression problems, ranging from simple ones like linear regression for regression problems and logistic regression for binary classification problems, to more complex ones like neural networks, which can be used for both classification and regression problems.

140.4 Performance Metrics

Performance of classification and regression models is evaluated using different metrics. For classification, these include accuracy, precision, recall, and F1 score. For regression, common metrics include mean absolute error, mean squared error, and R-squared.



CHAPTER 141

Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x , is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y , is regarded as the response, outcome, or dependent variable.

Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them in other contexts.

Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, a topic that will be covered later, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

141.0.1 The model behind simple linear regression

Given a scatterplot of the response variable y versus the predictor variable x , we fit the line

$$y = \beta_0 + \beta_1 x + \epsilon \quad (141.1)$$

that minimizes the distances from the observed points to the line!

- y = dependent variable (output/outcome/prediction/estimation)
- β_0 = y -intercept (constant term)
- β_1 = slope of the regression line (the effect that X has on Y)
- x = independent variable (input variable used in the prediction of Y)
- ϵ = error (the difference between the actual and predicted/estimated value)

This line can be used to predict future values of y given new data values of x .



CHAPTER 142

Simple Logistic Regression

While linear regression is used for predicting a continuous response variable, logistic regression is used for predicting a categorical response variable. It's particularly useful when the response variable is binary (i.e., it takes on only two possible outcomes, usually coded as 0 and 1).

The primary idea behind logistic regression is to find the probability of the response variable being true (1) given the values of the predictor variables.

In simple logistic regression, we have only one predictor variable. The form of the logistic regression model is:

$$\ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x \quad (142.1)$$

where:

- p is the probability of the positive class (i.e., the outcome $y = 1$).

- β_0 and β_1 are the parameters of the model.
- x is the predictor variable.

On the left-hand side, we have the natural log of the odds ratio (also called the logit), rather than just p itself. This is done to ensure that the predicted probabilities lie between 0 and 1. The function $\frac{p}{1-p}$ is called the odds, and can take any value between 0 and ∞ .

In contrast to linear regression, where the parameters are estimated using least squares, the parameters in logistic regression are usually estimated using maximum likelihood estimation.

Maximum likelihood estimation finds the parameter values that make the observed data most likely under the model.

In a simple logistic regression model, the probability that $Y = 1$ given x is:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (142.2)$$

And the probability that $Y = 0$ given x is:

$$1 - p(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \quad (142.3)$$



CHAPTER 143

Standardizing Data

Data standardization is a preprocessing step in many machine learning algorithms. Standardization transforms the variables in the dataset to have a mean of zero and a standard deviation of one.

The standardization of a variable X is calculated as follows:

$$Z = \frac{X - \mu}{\sigma} \quad (143.1)$$

where:

- Z is the standardized variable.
- X is the original variable.
- μ is the mean of X .
- σ is the standard deviation of X .

143.1 Why Do We Standardize Data?

There are several reasons why standardization is essential:

143.1.1 Homogeneity of Variances

Some statistical techniques assume that all variables have the same variance. Standardizing the data ensures this assumption.

143.1.2 Interpreting Coefficients

In regression analysis, standardizing allows us to interpret the coefficients of the predictors as the change in the response variable associated with a one-standard-deviation increase in the predictor.

143.1.3 Algorithm Convergence

For many machine learning algorithms (like gradient descent), standardization can help the algorithm converge more quickly to the optimum.

143.1.4 Comparing Variables

Standardization puts different variables on the same scale, allowing for meaningful comparisons. For example, it would be challenging to compare a variable measured in kilograms with another measured in kilometers without standardization.

143.1.5 Preventing Numerical Instabilities

Standardizing can help prevent numerical instabilities in computations, particularly when dealing with high-dimensional data.

Remember, though standardization is useful and necessary in many situations, it's not always required. For instance, tree-based models are scale-invariant and don't require standardization.



CHAPTER 144

One-Hot Encoding

In machine learning and data analysis, it is common to encounter categorical variables. Categorical data refers to variables that contain label values rather than numeric values. Examples include color ("red", "blue", "green"), size ("small", "medium", "large"), or geographic designations (city names, country names, etc.). Most machine learning algorithms require numerical input and output variables. One-hot encoding is a process of converting categorical data into a format that could be provided to machine learning algorithms to improve prediction.

144.1 Why One-Hot Encoding?

While some machine learning algorithms can work with categorical variables directly, many machine learning algorithms cannot operate on label data. They require all input variables and output variables to be numeric. Hence, categorical data needs to be converted to a numerical form. One-hot encoding is a popular method to transform categorical variables into a format that works better with classification and regression algorithms.

144.2 How does One-Hot Encoding Work?

In one-hot encoding, for each unique value in the categorical variable, we create a new binary feature that takes a value of 1 if the original feature value matches the unique value and 0 otherwise. If a categorical variable has n unique values, we would create n new features.

For example, consider the categorical variable "color" with three categories: "red", "blue", and "green". The one-hot encoding process will result in three new features, namely "is_red", "is_blue", and "is_green".

Color	is_red	is_blue	is_green
red	1	0	0
blue	0	1	0
green	0	0	1
red	1	0	0

This encoding helps to convey the information in the categorical variable to the learning algorithm effectively.

However, it's worth noting that one-hot encoding can significantly increase the dimensionality of the data, which can be problematic for some models. Therefore, it is not always the best choice, and other encoding methods might be more suitable depending on the situation.



CHAPTER 145

Multiple Logistic Regression

The simple logistic regression model, discussed in the last chapter, uses only one predictor variable, while multiple logistic regression, as the name implies, allows for more than one predictor variable.

145.1 Multiple Logistic Regression

In multiple logistic regression, we want to model the relationship between a binary response variable and multiple predictor variables. Let y be the binary response variable and x_1, x_2, \dots, x_p be p predictor variables. The multiple logistic regression model has the form:

$$\ln \left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

where $P(Y = 1|X)$ is the probability of the event $Y = 1$ given the predictor variables, and $\beta_0, \beta_1, \dots, \beta_p$ are the parameters of the model. This equation can also be rewritten in terms

of the probability $P(Y = 1|X)$:

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}$$

In this model, each one-unit increase in X_i multiplies the odds of $Y = 1$ by e^{β_i} , holding all other predictors constant.

145.2 Divide by 4 Rule

The "Divide by 4" rule is a rule of thumb for interpreting the coefficients in logistic regression. It says that for small values of β_i , a one-unit increase in X_i will change the probability $P(Y = 1|X)$ by approximately $\beta_i/4$ at the average value of X_i .

The rule arises from the derivative of the logistic function at its midpoint, and provides a useful and simple way to get an approximate sense of the effect size when interpreting the coefficients.



CHAPTER 146

The Training/Validation/Testing Process

In machine learning, it's essential to assess the performance of a model accurately. This assessment helps us choose the best model and tune its parameters. The data used to develop machine learning models is typically divided into three sets: training, validation, and testing.

146.0.1 Training Set

The training set is used to train the model, i.e., to adjust the model's weights and biases in the case of neural networks, or to determine the best split in decision trees, among other things. The model learns from this data, which is why it's called the "training" set.

146.0.2 Validation Set

The validation set is used to tune model parameters (hyperparameters), to choose the model architecture (for example, the number of hidden layers in a neural network), or to determine the degree of the polynomial in polynomial regression, among other uses. This set provides an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

146.0.3 Testing Set

The testing set is used to provide an unbiased evaluation of the final model fit on the training dataset. The test set serves as a proxy for real-world data that the model has not seen before. It's important to only use the test set once, after all training and validation is complete, to avoid "leaking" information from the test set into the model.

The key to this process is that each set of data is separate and independent. Mixing data between the sets can lead to overly optimistic or pessimistic assessments of a model's performance.

In practice, the division of data into these three sets can be done randomly (often with 70% for training, 15% for validation, and 15% for testing), or using more structured methods like cross-validation, depending on the amount and nature of the available data.



CHAPTER 147

Evaluating Classification Systems

The confusion matrix is a tabular method used in machine learning to evaluate the performance of a classification model. It allows for the visualization of the model's performance and to compute various performance metrics.

147.1 Definition of a Confusion Matrix

A confusion matrix is a specific table layout that presents the performance of a classification model. For a binary classification problem, it is a 2x2 matrix that compares the actual and the predicted classifications.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

147.2 Performance Metrics

Using the confusion matrix, we can compute several performance metrics:

- **Accuracy:** The proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. It is calculated as $(TP + TN) / (TP + TN + FP + FN)$.
- **Precision:** The proportion of positive identifications that were actually correct. It is calculated as $TP / (TP + FP)$.
- **Recall (Sensitivity):** The proportion of actual positives that were identified correctly. It is calculated as $TP / (TP + FN)$.
- **Specificity:** The proportion of actual negatives that were identified correctly. It is calculated as $TN / (TN + FP)$.
- **F1 Score:** The harmonic mean of precision and recall. It tries to find the balance between precision and recall. $F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$.



CHAPTER 148

Evaluation Binary Classifiers

Accuracy, recall, precision, and the F1 score are widely used metrics to measure and compare the performance of binary classifiers. This chapter will delve into these evaluation measures, providing insights into their interpretation and practical applications.

148.0.1 Binary Classification

Before diving into the evaluation metrics, let's clarify the concept of binary classification. In binary classification, we aim to assign each instance in a dataset to one of two mutually exclusive classes. For example, classifying emails as spam or not spam, identifying whether a patient has a specific medical condition or not, or predicting whether a credit card transaction is fraudulent or legitimate are common binary classification tasks.

To evaluate the performance of a binary classifier, we need metrics that can provide insights into how well the classifier performs in distinguishing between the two classes.

148.0.2 Accuracy

Accuracy is a widely used metric for evaluating binary classifiers. It measures the overall correctness of the classifier's predictions by calculating the ratio of correctly classified instances to the total number of instances in the dataset. Mathematically, accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

While accuracy provides a general overview of the classifier's performance, it may not be sufficient in certain scenarios. This is especially true when the dataset is imbalanced, meaning that one class significantly outweighs the other in terms of the number of instances.

148.0.3 Precision and Recall

Precision and recall are evaluation metrics that provide insights into the classifier's performance on specific classes, allowing us to identify potential trade-offs between false positives and false negatives.

Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). It can be expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). Mathematically, recall can be represented as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision and recall are complementary metrics. Precision focuses on the quality of positive predictions, while recall emphasizes the classifier's ability to identify positive instances. The choice between precision and recall depends on the specific requirements of the problem at hand.

148.0.4 F1 Score

The F1 score combines precision and recall into a single metric, providing a balanced evaluation measure that considers both false positives and false negatives. It is the harmonic mean of precision and recall, and it can be calculated as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score ranges between 0 and 1, where a value of 1 represents perfect precision and recall. It is particularly useful when we want to strike a balance between precision and recall, considering both the false positives and false negatives in the classifier's predictions.



CHAPTER 149

The k-Nearest Neighbor Classifier

The k-nearest neighbors (k-NN) algorithm is a type of instance-based learning algorithm used for classification and regression. Given a new, unknown observation, k-NN algorithm searches through the entire dataset to find the 'k' training examples that are closest to the new instance, and predicts the label based on these 'k' nearest neighbors.

149.1 The k-NN Algorithm

The algorithm can be summarized as follows:

1. Given a new observation x , compute the distance between x and all points in the training set.
2. Identify the 'k' points in the training data that are closest to x .
3. If k-NN is used for classification, output the most common class among these 'k'

points as the prediction. If k-NN is used for regression, output the average of the values of these 'k' points as the prediction.

The distance between points can be calculated using various metrics, the most common one being the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where n is the number of features, and x_i and y_i are the corresponding features of x and y .

149.2 Choosing the Right 'k'

The choice of 'k' has a significant impact on the k-NN algorithm. A small 'k' (like 1) can capture a lot of noise and lead to overfitting, while a large 'k' can smooth over many details and potentially lead to underfitting. Cross-validation is typically used to select an optimal 'k'.

149.3 Considerations

Although the k-NN algorithm is simple to understand and implement, it can be computationally intensive for large datasets, as it requires computing the distance between every pair of points. Additionally, it's sensitive to the choice of the distance metric and the scale of the features.



CHAPTER 150

Naive Bayes Classifier

The Naive Bayes classifier is a simple yet effective algorithm for classification tasks. It is a probabilistic classifier based on Bayes' theorem and some additional simplifying assumptions, which make it particularly suitable for high-dimensional datasets.

150.1 Bayes' Theorem

Bayes' theorem describes the relationship of conditional probabilities of statistical quantities. In the context of classification, it can be written as:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

where:

- $P(C|X)$ is the posterior probability of class C given predictor (features) X.
- $P(C)$ is the prior probability of class.

- $P(X|C)$ is the likelihood which is the probability of predictor given class.
- $P(X)$ is the prior probability of predictor.

150.2 The Naivety of Naive Bayes

The "Naive" in Naive Bayes comes from the assumption that each feature in the dataset is independent of all other features, given the class. This is a strong (and often unrealistic) assumption, hence the name "naive". Despite this unrealistic assumption, the Naive Bayes classifier often performs well in practice.

In the context of text classification, this naivety translates into assuming that every word in a document is independent of all other words, given the document's class.

150.3 Working of Naive Bayes Classifier

When given an instance to classify, the Naive Bayes classifier calculates the posterior probability of that instance belonging to each possible class. The classifier then outputs the class with the highest posterior probability.

For computational reasons, and because the denominator $P(X)$ is constant given the input, we typically use the following simplification in practice:

$$P(C|X) \propto P(X|C) \cdot P(C)$$

which means that we can focus on maximizing $P(X|C) \cdot P(C)$.

Naive Bayes classifiers are highly scalable and are known for their simplicity, speed, and suitability for high-dimensional datasets.



CHAPTER 151

Evaluating the Fit of a Linear Regression Model

The fit of a linear regression model can be evaluated using several statistical metrics. Three common ones include the residuals, the coefficient of determination (R-squared or R^2), and the root mean squared error (RMSE).

151.1 Residuals

Residuals are the differences between the observed and predicted values. For an observation i , the residual e_i is calculated as

$$e_i = y_i - \hat{y}_i$$

where y_i is the observed value and \hat{y}_i is the predicted value. By plotting these residuals against the predicted values, we can visually inspect the model's fit. Ideally, the residuals

should be randomly scattered around zero, and there should be no clear pattern in the residual plot.

151.2 R-Squared (R^2)

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be predicted from the independent variables. It provides a measure of how well the model's predictions fit the data. R^2 is calculated as

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares. An R^2 value of 1 indicates a perfect fit, while an R^2 of 0 indicates that the model does not explain any of the variability of the response data around its mean.

151.3 Root Mean Squared Error (RMSE)

RMSE is a frequently used measure of the differences between the values predicted by a model and the values actually observed. It's the square root of the average of squared differences between prediction and actual observation. RMSE is calculated as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

A lower RMSE indicates a better fit to the data.

It's important to note that these metrics should not be used in isolation to evaluate the model's fit. They should be used in combination along with the understanding of the underlying problem and domain knowledge.



CHAPTER 152

Linear Regression and Gradient Descent

Linear regression models can be fitted using an optimization algorithm known as gradient descent. This is especially useful when the number of features is large, making the normal equation computationally expensive.

In gradient descent, we start with an initial guess for the model parameters and iteratively update these parameters to minimize the cost function, which is usually the mean squared error (MSE) for linear regression. For a linear regression model with parameters θ , the update rule is given by

$$\theta := \theta - \alpha \nabla J(\theta)$$

where α is the learning rate and $\nabla J(\theta)$ is the gradient of the cost function evaluated at θ . For MSE, the gradient is given by

$$\nabla J(\theta) = \frac{2}{n} \mathbf{X}^T (\mathbf{X}\theta - \mathbf{y})$$

where \mathbf{X} is the feature matrix, \mathbf{y} is the vector of target values, and n is the number of observations.

152.1 Standardizing Inputs

Standardizing inputs can improve the performance of gradient descent. By ensuring all features have a similar scale, we can avoid a situation where the cost function has a very elongated shape, causing gradient descent to take a long time to converge.

More specifically, standardization transforms the features so they have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation for each feature:

$$x'_i = \frac{x_i - \mu_i}{\sigma_i}$$

where x_i is a feature vector, and μ_i and σ_i are its mean and standard deviation, respectively.

By standardizing the inputs, each feature contributes approximately proportionately to the final distance, helping the gradient descent algorithm converge more quickly and efficiently.



CHAPTER 153

Generalized Linear Models

In statistics, generalized linear models (GLMs) are a flexible generalization of ordinary linear regression models for response variables that are not normally distributed. If you're already familiar with multiple linear regression, you're well on your way to understanding GLMs.

153.1 Components of a Generalized Linear Model

A GLM consists of three components:

1. A random component: This is a specification of the probability distribution of the response variable (e.g., normal, binomial, Poisson distributions, etc.). This differs from ordinary linear regression, which assumes that the response variable follows a normal distribution.
2. A systematic component: This is the linear predictor, a linear combination of the explanatory variables, just as in ordinary linear regression.
3. A link function: This is a function that connects the mean of the response variable

to the linear predictor. The choice of link function depends on the nature of the response variable and the range of its possible values.

153.2 Formulation of a Generalized Linear Model

The GLM can be formulated as follows:

$$g(E(Y)) = \eta = X\beta \quad (153.1)$$

Here, Y is the response variable, X represents the matrix of explanatory variables, β is the vector of parameters to be estimated, η is the linear predictor, $E(Y)$ represents the expected value of Y , and $g(\cdot)$ is the link function.

153.3 Fitting a Generalized Linear Model

The parameters β in a GLM are typically estimated using maximum likelihood estimation (MLE). The specifics of this process depend on the probability distribution of the response variable and the link function.

153.4 Examples of Generalized Linear Models

Examples of GLMs include:

- Logistic regression: This is a GLM with a binomial response variable and a logit link function.
- Poisson regression: This is a GLM with a Poisson response variable and a log link function.



CHAPTER 154

Link Functions

In generalized linear models, the link function provides the relationship between the linear predictor and the mean of the distribution function. Different choices of link function can be used to model different types of relationships. Here are a few commonly used link functions:

1. **Identity link:** The identity link function is the simplest form of link function, where the response variable is expected to be the linear combination of the predictors. This is the default link function for Gaussian family distributions.

$$g(\mu) = \mu$$

2. **Log link:** The log link function is used when modeling positive data and count data. This link function is the default for Poisson and exponential family distributions.

$$g(\mu) = \log(\mu)$$

3. **Logit link:** The logit link function is often used when modeling binary response data, and is the default link function for binomial family distributions. It gives the log-odds, or the logarithm of the odds $p/(1 - p)$.

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

4. **Probit link:** The probit link function is another common choice for binary response data. It is based on the cumulative distribution function of the standard normal distribution.

$$g(\mu) = \Phi^{-1}(\mu)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution.

5. **Inverse link:** The inverse link function is often used in modeling rates or times. It is the canonical (or default) link function for the Gamma family distributions.

$$g(\mu) = \mu^{-1}$$

Different link functions can substantially impact the model's interpretation, so it's crucial to choose a link function that aligns with the nature of the data and the scientific question at hand.



CHAPTER 155

Decision Trees for Classification

A decision tree is a popular machine learning algorithm used for both regression and classification problems. In this discussion, we will focus on its application in classification tasks.

155.1 Decision Trees for Classification

A decision tree for classification uses a tree structure to predict the class of an object based on its features. The tree is made up of nodes that split the data based on a feature value, and leaves that represent a class label. The idea is to create a tree that has minimum impurity, i.e., at the end of the tree, we would like each leaf to contain data points that belong to a single class.

155.2 Gini Impurity

Gini impurity is a measure of misclassification, which applies in a multiclass classifier context. It gives an idea of how often a randomly chosen element from the set would be

incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

The Gini impurity for a node of the tree is calculated as:

$$\text{Gini}(p) = 1 - \sum_{i=1}^J (p_i)^2 \quad (155.1)$$

where p_i is the fraction of items classified to label i at a node and J is the total number of classes.

A Gini impurity of 0 is the best score, where all elements in a partition fall into a single category.

155.3 How Gini Impurity is Used

During the construction of a decision tree, the best feature to split on at each node is chosen by minimizing the Gini impurity of the child nodes. The algorithm will consider all features and all possible split points for each feature to find the split that yields the lowest weighted average Gini impurity.



CHAPTER 156

Bagging and Random Forests

Bagging (Bootstrap Aggregating) and Random Forests are ensemble machine learning methods that are primarily used to improve the stability and accuracy of prediction models.

156.1 Bagging

Bagging, an abbreviation for Bootstrap Aggregating, is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages the output (for regression) or performs a vote (for classification).

Given a standard training set D of size n , bagging generates m new training sets D_i , each of size n' , by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i . If $n' = n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e) \approx 63.2\%$ of the unique examples of D , the rest being duplicates.

156.2 Random Forests

Random Forests is a substantial modification of Bagging that builds a large collection of de-correlated trees, and then averages them. When building these decision trees, each time a split in a tree is considered, a random sample of k features is chosen as split candidates from the full set of features. The split is allowed to use only one of those k features. A fresh sample of k features is taken at each node, and the best feature/split-point among the k is chosen.

For classification problems, $k = \sqrt{p}$ is typically taken, where p is the number of features in the model. For regression problems, the inventors recommend $k = p/3$, with a minimum node size of 5 as the default.

In Random Forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

1. Each tree is constructed using a different bootstrap sample from the original data.
2. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k -th tree.
3. Let $y_{\text{tree } k}(x)$ be the class prediction of the k -th Random Forest tree for x . Then the Random Forest classifier does a majority vote over all trees:

$$y_{\text{RF}}(x) = \text{majority}\{y_{\text{tree } k}(x), k = 1, \dots\}$$



CHAPTER 157

Boosting

Boosting is a machine learning ensemble meta-algorithm primarily used to reduce bias, and to a lesser extent variance, in supervised learning. It works by iteratively learning weak classifiers and adding them to a final strong classifier in a way that the subsequent weak learners try to correct the mistakes of the previous ones.

157.1 AdaBoost

AdaBoost, short for Adaptive Boosting, is one of the first and simplest boosting algorithms. Given a set of n training examples $(x_1, y_1), \dots, (x_n, y_n)$ where y_i are binary outputs, the algorithm works as follows:

1. Initialize weights $w_i = 1/n$ for $i = 1, \dots, n$.
2. For $t = 1$ to T :
 - Train a weak learner h_t using the weighted examples.
 - Compute the weighted error $\epsilon_t = \sum_{i:h_t(x_i) \neq y_i} w_i$.

- Set $\alpha_t = \frac{1}{2} \log \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$.
 - Update the weights: $w_i = w_i \exp(-\alpha_t y_i h_t(x_i))$ for $i = 1, \dots, n$, and normalize them so that they sum to one.
3. The final model is $H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$.

157.2 Gradient Boosted Trees

Gradient Boosted Trees is a generalization of boosting to arbitrary differentiable loss functions. It works by sequentially adding predictors to an ensemble, each one correcting its predecessor by fitting the new predictor to the residual errors.



CHAPTER 158

Clustering using k-Means

K-means is a popular unsupervised learning algorithm used for data clustering. The goal of k-means is to group data points into distinct non-overlapping subgroups, or clusters, based on their features.

158.1 The K-Means Algorithm

Given a dataset $X = \{x_1, x_2, \dots, x_N\}$, where each x_i is a d -dimensional vector, and an integer k , the k-means clustering algorithm seeks to find k cluster centroids $C = \{c_1, c_2, \dots, c_k\}$ such that the distance from each data point to its nearest centroid is minimized.

The k-means algorithm works as follows:

1. Initialize k centroids randomly.
2. Assign each data point to the nearest centroid. This forms k clusters.
3. For each cluster, update its centroid by computing the mean of all points in the cluster.

4. Repeat steps 2 and 3 until the centroids do not change significantly or a maximum number of iterations is reached.

The measure of distance typically used in k-means is the Euclidean distance. For two d -dimensional vectors $x = (x_1, \dots, x_d)$ and $y = (y_1, \dots, y_d)$, the Euclidean distance is defined as:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_d - y_d)^2} \quad (158.1)$$

158.2 Choosing K

Choosing an appropriate value for k is a significant aspect of the k-means algorithm. One common method is the Elbow Method, which involves plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.



CHAPTER 159

Neural Nets for Regression

Neural Networks are powerful computational models that are used for a variety of tasks in machine learning, including regression. Unlike linear regression, which is a linear approach to modelling the relationship between a dependent variable and one or more independent variables, a Neural Network can model complex non-linear relationships.

159.1 Neural Networks

A neural network is composed of nodes (neurons) grouped into layers. The input layer takes in the features of the dataset, hidden layers perform computations on these inputs, and the output layer produces the final prediction. Each node in a layer is connected to each node in the next layer through "edges". These edges carry weights, which are the parameters of the model that are learned during training.

A key part of each node is an activation function. It transforms the weighted sum of the node's inputs into an output value that is passed onto the next layer. Common activation functions include the ReLU (Rectified Linear Unit), sigmoid and hyperbolic tangent functions. For regression tasks, usually a linear activation function is used in the output layer,

as the output can be any real number.

159.2 Neural Networks for Regression

To perform regression using a neural network, you would train the network to map the input features to a continuous target variable.

Here is the basic process:

- **Feedforward:** Compute the output of the network given the input features. This involves calculating the weighted sum of inputs for each node and applying the activation function.
- **Loss Calculation:** Calculate the loss (difference between the network's prediction and the actual value). For regression tasks, common loss functions include Mean Squared Error (MSE) or Mean Absolute Error (MAE).
- **Backpropagation:** Update the network's weights to minimize the loss. This is done by computing the gradient of the loss function with respect to each weight in the network, and then adjusting the weights in the direction that decreases the loss.
- **Iteration:** Repeat the feedforward, loss calculation, and backpropagation steps for a number of epochs (complete passes through the dataset) or until the loss converges to a minimum.

This way, the neural network learns to approximate the function that best maps input features to the target variable, thus performing regression. The advantage of using neural networks over linear regression is that they can capture complex non-linear relationships between variables.



CHAPTER 160

Neural Networks for Classification

Neural Networks can also be used for classification tasks, which involve predicting a discrete class label output for an instance. The process of using a Neural Network for classification is similar to using it for regression, but there are key differences in the output layer and loss function.

160.1 Neural Networks for Classification

For a binary classification problem, where the output can be either of two classes, the output layer of the neural network typically consists of a single neuron with a sigmoid activation function, which squashes the output between 0 and 1. This output can be interpreted as the probability that the instance belongs to a particular class.

For a multi-class classification problem, where the output can be one of more than two classes, the output layer typically has as many neurons as there are classes, and a softmax activation function is used, which gives the probability distribution over the classes.

Here is the basic process:

- **Feedforward:** Compute the output of the network given the input features, just as in regression.
- **Loss Calculation:** Calculate the loss (difference between the network's prediction and the actual class). For classification tasks, common loss functions include Cross-Entropy Loss.
- **Backpropagation:** Update the network's weights to minimize the loss, the same way as in regression.
- **Iteration:** Repeat the feedforward, loss calculation, and backpropagation steps for a number of epochs (complete passes through the dataset) or until the loss converges to a minimum.

Once trained, the neural network can classify a new instance by performing a feedforward pass and predicting the class with the highest probability in the output layer.



CHAPTER 161

Deep Learning

Deep learning is a subfield of machine learning that focuses on algorithms inspired by the structure and function of the brain called artificial neural networks. While you may have encountered simple, shallow neural networks, deep learning involves neural networks with many layers, hence they are often referred to as "deep" neural networks.

161.1 Deep Learning

Deep learning models learn to represent data by training on a large number of examples. Unlike shallow neural networks that have one or two layers of hidden nodes, deep networks can have tens or even hundreds of layers of hidden nodes. Each layer in these networks performs a nonlinear transformation of its inputs and is trained to extract increasingly abstract features with each additional layer.

161.2 Chain Rule

In order to understand how these networks are trained, we need to revisit a fundamental concept from calculus, the chain rule. The chain rule is used for differentiating compositions of functions. It essentially says that the derivative of a composed function is the product of the derivatives of the composed functions.

Suppose we have a function $y = f(g(x))$, then the derivative of y with respect to x is:

$$\frac{dy}{dx} = f'(g(x)) \cdot g'(x) \quad (161.1)$$

This rule becomes indispensable when calculating the gradient of the loss function in a deep learning model with respect to the model parameters.

161.3 Backpropagation

Backpropagation is the method used to train deep learning models by calculating the gradient of the loss function with respect to each weight in the network. The name "backpropagation" comes from the fact that the calculation of the gradient proceeds backwards through the network, with the gradient of the final layer of weights being calculated first and the gradient of the first layer of weights being calculated last.

Mathematically, backpropagation uses the chain rule to efficiently compute these gradients. Starting from the final layer, the chain rule is repeatedly applied to propagate the gradient backwards through the network, storing intermediate results as it goes along. Once the gradient has been calculated, the weights are updated using a gradient descent step.



APPENDIX A

Answers to Exercises

Answer to Exercise 1 (on page 27)

To get the train to 20 meters per second in 120 seconds, you must accelerate it with a constant rate of $\frac{1}{6}\text{m/s}^2$. You remember that $F = ma$, so $F = 2400 \times \frac{1}{6}$. Thus, you will push the train with a force of 400 newtons for the 120 seconds before the bomb goes off.

Answer to Exercise 2 (on page 28)

$$F = G \frac{m_1 m_2}{r^2} = (6.674 \times 10^{-11}) \frac{(6.8^3)(6 \times 10^{24})}{(10^5)^2} = 6.1 \times 10^6$$

About 6 million newtons.

Answer to Exercise 3 (on page 34)

The average hydrogen atom has a mass of 1.00794 atomic mass units.

The average oxygen atom has a mass of 15.9994.

$$2 \times 1.00794 + 15.9994 = 18.01528 \text{ atomic mass units.}$$

Answer to Exercise 4 (on page 35)

From the last exercise, you know that 1 mole of water weighs 18.01528 grams. So 200 grams of water is about 11.1 moles. So you need to burn 11.1 moles of methane.

What does one mole of methane weigh? Using the periodic table: $12.0107 + 4 \times 1.00794 = 16.04246$ grams.

$$16.0424 \times 11.10 = 178.1 \text{ grams of methane.}$$

Answer to Exercise 5 (on page 41)

At the top of the ladder, the cannonball has $(9.8)(5)(3) = 147$ joules of potential energy.

At the bottom, the kinetic energy $\frac{1}{2}(5)v^2$ must be equal to 147 joules. So $v^2 = \frac{294}{5}$. Thus it is going about 7.7 meters per second.

(Yes, a tiny amount of energy is lost to air resistance. For a dense object moving at these relatively slow speeds, this energy is negligible.)

Answer to Exercise 6 (on page 46)

$$4.5 \text{ kWh} \left(\frac{3.6 \times 10^6 \text{ joules}}{1 \text{ kWh}} \right) \left(\frac{1 \text{ calories}}{4.184 \text{ joules}} \right) = \frac{(4.5)(3.6 \times 10^6)}{4.184} = 1.08 \times 10^6 \text{ calories}$$

Answer to Exercise 7 (on page 46)

$$\frac{0.1 \text{ gallons}}{2 \text{ minutes}} \left(\frac{3.7854 \text{ liters}}{1 \text{ gallons}} \right) \left(\frac{1000 \text{ milliliters}}{1 \text{ liters}} \right) \left(\frac{1 \text{ minutes}}{60 \text{ seconds}} \right) = \\ \frac{(0.1)(3.7854)(1000)}{(2)(60)} \text{ ml/second} = 3.1545 \text{ ml/second}$$

Answer to Exercise 8 (on page 51)

Paul is exerting $(70)(9.8)$ newtons of force at 4 meters from the fulcrum, so he is creating a torque of $2,744$ newton-meters of torque on the see-saw. Jan is creating $(50)(9.5) = 490$ newtons of force.

If r is the distance from the fulcrum to Jan's seat, to balance $490r = 2744$, so $r = 5.6$ meters.

Answer to Exercise 9 (on page 53)

To lift the barrel would require $136 \times 9.8 = 1,332.8$ newtons of force.

Letting L be the length of the ramp:

$$300 = \frac{2}{L} 1332.8$$

So $L = 8.885$ meters.

Answer to Exercise 9 (on page 54)

$$583 = (70)(2.2) \frac{53}{n}$$

Thus $n = 14$ teeth.

Answer to Exercise 11 (on page 55)

We are looking for r , the radius of the piston head in meters. The area of the piston head is πr^2 .

The pressure in pascals of the brake fluid is given by $12/(\pi r^2)$.

$$2,500,000 = \frac{12}{\pi r^2}$$

So $r = \sqrt{\frac{12}{\pi \times 2.5 \times 10^6}} = 0.001236077446474$ meters.

Answer to Exercise 12 (on page 59)

Equilibrium will be achieved when the box has displaced 10 kg of water. That is, when it has displaced 0.01 cubic meters.

The area of the base of the box is 0.12 square meters. So if the box sinks x meters into the water it will displace $0.12x$ cubic meters.

Thus at equilibrium $x = \frac{0.01}{0.12} \approx 0.083$ m. So, the box will sink 8.3 cm into the water before reaching equilibrium.

Answer to Exercise 13 (on page 60)

$$p = dgh = (900)(3.721)(5) = 16,744.5 \text{ Pa}$$

Answer to Exercise 14 (on page 65)

$$E_C = (1200)(0.4)(T - 10) = 480T - 4800$$

Total energy stays constant:

$$0 = (12600T - 252000) + (900T - 72000) + (480T - 4800)$$

Solving for T gets you $T = 23.52^\circ C$.

Answer to Exercise 15 (on page 66)

During the 3 minutes, you want the coffee to give off as much of its heat as possible, so you want to maximize the difference between the temperature of the coffee and the temperature of the room around it.

You wait until the last moment to put the milk in.

Answer to Exercise 16 (on page 78)

Kinetic energy? $E = mv^2 = (55)(11^2) = 6,655 \frac{\text{kg}\text{m}^2}{\text{s}^2} = 6,655$ joules.

Frictional force? $F = \mu N = (0.7)(55)(9.8) = 377.3$ newtons.

Distance? $D = \frac{6,655}{377.3} = 17.6$ seconds.

Answer to Exercise 17 (on page 80)

The empty sled is pushing directly down on the floor with a force of $(50)(9.8) = 490$ N.

The force to overcome the static friction is:

$$270 = 490\mu_s$$

Thus $\mu_s = 0.551$

The force to match kinetic friction is:

$$220 = 490\mu_k$$

Thus $\mu_k = 0.449$

Once you are on the sled, it is pressing directly down on the floor with a force of $(50 + 55)(9.8) = 1,029$ N.

The force to overcome the static friction is:

$$F = (1,029)(0.551) = 567 \text{ N}$$

Once the sled is moving, friction is counteracting some of your force. How much?

$$F_f = (1,029)(0.449) = 462 \text{ N}$$

So all of your acceleration is due to the remaining $567 - 462 = 75 \text{ N}$.

We know that $F = ma$. In this case $F = 75 \text{ N}$ and $m = 105 \text{ kg}$. So

$$a = \frac{75}{105} = 0.714 \text{ meters per second per second}$$

Answer to Exercise 18 (on page 86)

$$\mu = \frac{1}{6} (87 + 91 + 98 + 65 + 87 + 100) = 88$$

Answer to Exercise 19 (on page 88)

The mean of your grades is 88.

The variance, then is

$$\sigma^2 = \frac{1}{6} \left((87 - 88)^2 + (91 - 88)^2 + (98 - 88)^2 + (61 - 88)^2 + (87 - 88)^2 + (100 - 88)^2 \right) = \frac{784}{6} = 65\frac{1}{3}$$

The standard deviation is the square root of that: $\sigma = 8.083$ points.

Answer to Exercise 20 (on page 89)

In order the grades are 65, 87, 87, 91, 98, 100. The middle two are 87 and 91. The mean of those is 89. (Speed trick: The mean of two numbers is the number that is half-way between.)

Answer to Exercise 21 (on page 99)

The formula for the RMS is “=SQRT(SUMSQ(A2:A1001)/COUNT(A2:A1001))”.

Answer to Exercise 22 (on page 110)

$$V = IR \text{ so } I = \frac{V}{R} = \frac{24V}{6\Omega} = 4A.$$

Answer to Exercise ?? (on page 112)

There is a total resistance of 8Ω , so your 16V will push 2A of current around the circuit.

2A going through a 5Ω resistor represents a 10V drop.

2A going through a 3Ω resistor represents a 6V drop.

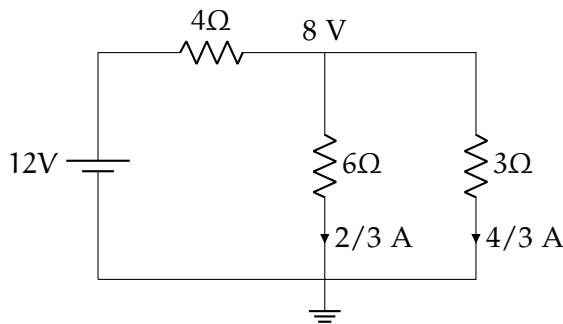
Answer to Exercise 24 (on page 114)

The effective resistance of the 6Ω and the 3Ω is 2Ω because

$$\frac{1}{R_T} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

So the battery experiences a resistance of $4\Omega + 2\Omega = 6\Omega$. A 12V will push 2A through a resistance of 6Ω .

The voltage drop across the 4Ω resistor is $2A \times 4\Omega = 8V$. Thus there will be a 4V drop across the two resistors in parallel. So $2/3$ A will flow through the 6Ω resistor. $4/3$ A will flow through the 3Ω resistor.



Answer to Exercise 25 (on page 116)

$$F = K \frac{|q_1 q_2|}{r^2} = (8.988 \times 10^9) \frac{(-5 \times 10^{-9})(-5 \times 10^{-9})}{0.12^2} = \frac{224.7 \times 10^{-9}}{0.0144} = 15.6 \times 10^{-6}$$

15.6 micronewtons.

Answer to Exercise 26 (on page 126)

On earth, holding a 100 kg man aloft requires 980 Newtons of force.

$980/700 = 1.4$, so you need a cable with a cross-section area of 1.4 square millimeters.

$$\pi r^2 = 1.4$$

So $r = \sqrt{1.4/\pi} \approx .67$ millimeters. So the cable would have to have a diameter of at least 1.34 millimeters.

Answer to Exercise 27 (on page 136)

$$180^\circ - (92^\circ + 42^\circ) = 46^\circ$$

Answer to Exercise 28 (on page 137)

$$360^\circ$$

Answer to Exercise 29 (on page 140)

$$10 \text{ because } 6^2 + 8^2 = 10^2$$

$$12 \text{ because } 5^2 + 12^2 = 13^2$$

$$8 \text{ because } 8^2 + 15^2 = 17^2$$

$$3\sqrt{2} \approx 4.24 \text{ because } 3^2 + 3^2 = (3\sqrt{2})^2$$

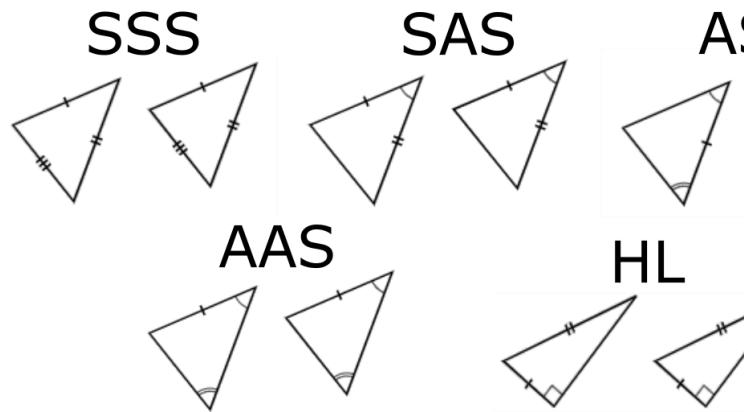
Answer to Exercise 30 (on page 147)

Congruent by the Side-Side-Right Congruency Test.

Congruent by the Side-Side-Side Congruency Test.

Congruent by the Side-Angle-Angle Congruency Test.

We don't know if they are congruent. The measured angle is not between the measured sides.



Answer to Exercise 31 (on page 152)

The table has a radius of 3 meters.

So the area of its top is $3^2\pi \approx 28.27$.

$$28.27 \text{ square meters} \left(\frac{1 \text{ liter}}{6 \text{ square meters}} \right) = 4.72 \text{ liters}$$

Answer to Exercise 32 (on page 153)

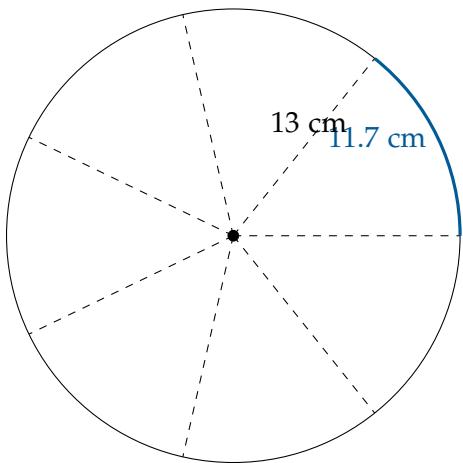
The diameter is

$$\frac{c}{\pi} = \frac{64}{\pi} \approx 20.37 \text{ centimeters}$$

Answer to Exercise 33 (on page 153)

The circumference of the pie is $26\pi \approx 81.7$ centimeters.

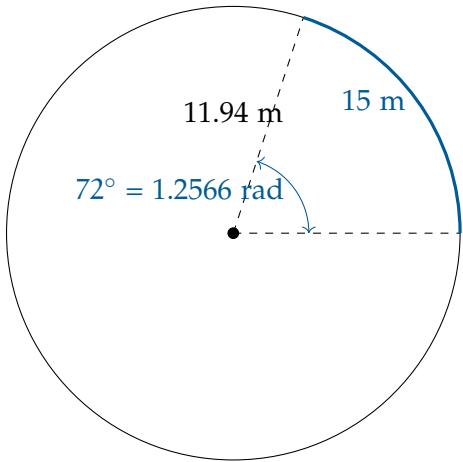
The length of the crust for each piece would be about $\frac{81.7}{7} = 11.7$ cm.

**Answer to Exercise 34 (on page 154)**

$$72 \text{ degrees} \left(\frac{2\pi \text{ radians}}{360 \text{ degrees}} \right) \approx 1.2566 \text{ radians}$$

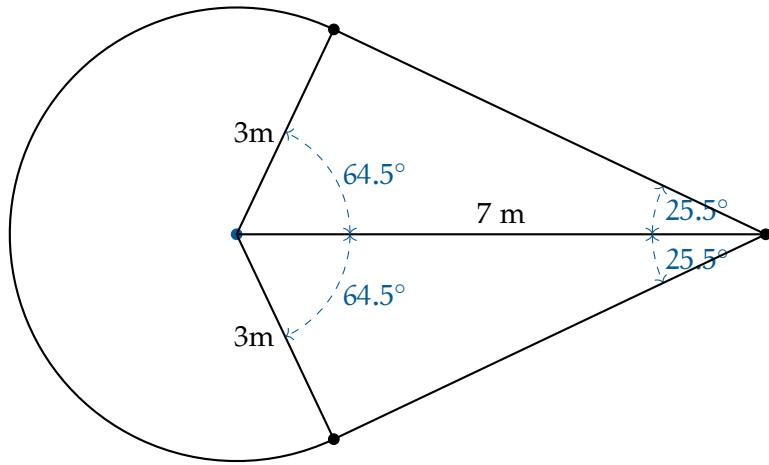
$$15 = 1.2566r$$

$$r = 11.94 \text{ meters}$$

**Answer to Exercise 35 (on page 156)**

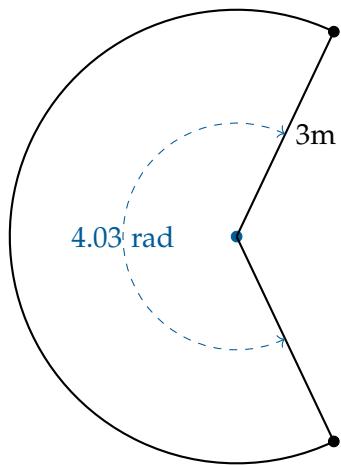
The trick here is to take advantage of the fact that the tangent is perpendicular to the

radius to make right triangles:



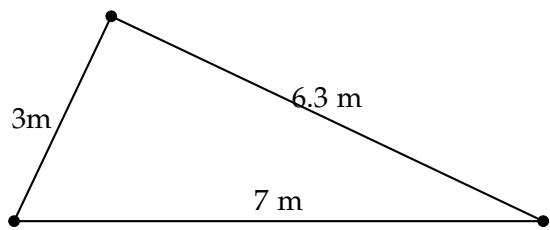
The wedge has radius 3 and represents $360 - 2(64.5) = 231^\circ \approx 4.03$ radians.

We are finding the area of this piece:



The area of this piece is $(4.03)(3^2) = 36.27$ square meters.

If a right triangle has a hypotenuse of 7m and one leg is 3m, the other leg is $\sqrt{7^2 - 3^2} = 2\sqrt{10} \approx 6.3$ m.



A right triangle with legs of 3m and 6.3m has an area of 9.45 square meters.

There are two of them, so the total area is $36.27 + 2(18.9) = 74.07$ square meters.

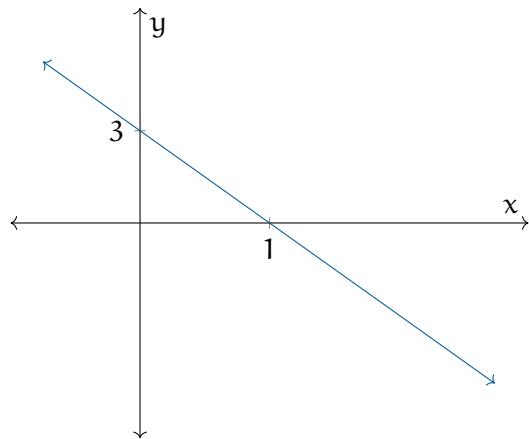
Six square meters per liter, so you need $\frac{74.07}{6} = 12.35$ liters of paint.

Answer to Exercise 36 (on page 158)

You can only take the square root of nonnegative numbers, so the function is only defined when $x - 3 \geq 0$. Thus the domain is all real numbers greater than or equal to 3.

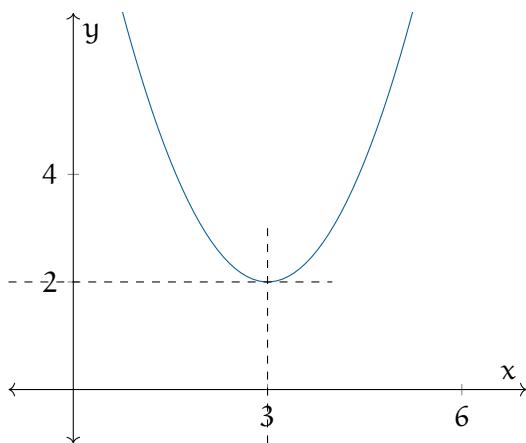
Answer to Exercise 37 (on page 159)

The graph of this function is a line. Its slope is -3. It intersects the y axis at (0, 3)



Answer to Exercise 38 (on page 162)

This graph is the graph of $y = x^2$ that has been moved to the right by three units and up two units:



To prevent any horizontal line from containing more than one point of the graph, you would need to use the left or the right side: Either $\{x \in \mathbb{R} | x \leq 3\}$ or $\{x \in \mathbb{R} | x \geq 3\}$. Most people will choose the right side; the rest of the solution will assume that you did too.

To find the inverse we swap x and y : $x = (y - 3)^2 + 2$

Then we solve for y to get the inverse: $y = \sqrt{x - 2} + 3$

You can take the square root of nonnegative numbers. So the function $f^{-1}(x) = \sqrt{x - 2} + 3$ is defined whenever x is greater than or equal to 2.

Answer to Exercise 39 (on page 166)

The volume of the sphere (in cubic meters) is

$$\frac{4}{3}\pi(1.5)^3 = 4.5\pi \approx 14.14$$

The mass (in kg) is $14.14 \times 7800 \approx 110,269$

The kinetic energy (in joules) is

$$k = \frac{110269 \times 5^2}{2} = 1,378,373$$

About 1.4 million joules.

Answer to Exercise 40 (on page 167)

In your mind, you can dissemble the tablet into a sphere (made up of the two ends) and a cylinder (between the two ends)

The volume of the sphere (in cubic millimeters) is

$$\frac{4}{3}\pi(2)^3 = \frac{32}{3}\pi \approx 33.5$$

Thus the cylinder part has to be $90 - 33.5 = 56.5$ cubic mm. The cylinder part has a radius of 2 mm. If the length of the cylinder part is x , then

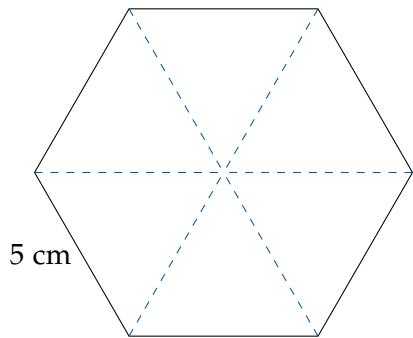
$$\pi 2^2 x = 56.5$$

$$\text{Thus } x = \frac{56.5}{4\pi} \approx 4.5 \text{ mm.}$$

The cylinder part of the table needs to be 4.5mm. Thus the entire tablet is 8.5mm long.

Answer to Exercise 41 (on page 170)

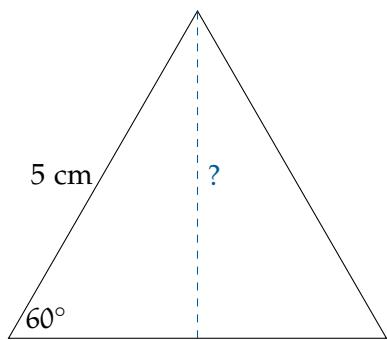
First, you need to find the area of the base, which is a regular hexagon:



All the angles in this picture are 60° or $\frac{\pi}{3}$ radians. Thus, each line is 5 cm long.

Thus, we need to find the area of one of these triangles and multiply that by six.

Every triangle has a base of 5cm. How tall are they?



$$5 \sin 60^\circ = 5 \frac{\sqrt{3}}{2}$$

Which is about 4.33 cm.

Thus, the area of single triangle is

$$\frac{1}{2}(5) \left(5 \frac{\sqrt{3}}{2}\right) = 25 \frac{\sqrt{3}}{4}$$

And the area of the whole hexagon is six times that:

$$75 \frac{\sqrt{3}}{2}$$

Thus, the volume of the pyramid is:

$$\frac{1}{3}hb = \frac{1}{3}13 \left(75 \frac{\sqrt{3}}{2}\right)$$

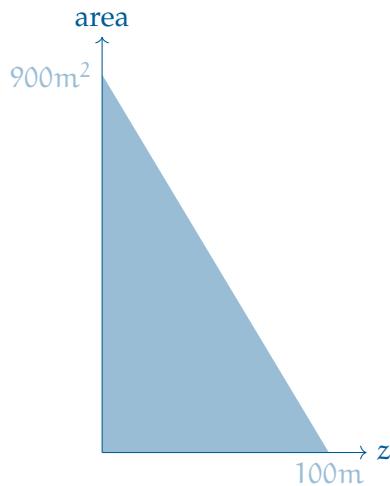
About 281.46 cubic centimeters.

Answer to Exercise 42 (on page 172)

The area at height z is given by:

$$a = \frac{1}{2}w^2 = \frac{1}{2} \left(30 \sqrt{1 - \frac{z}{100}}\right)^2 = \frac{1}{2}900 \left(1 - \frac{z}{100}\right)$$

If we plot that, it looks like this:



What is the area of the blue region? $\frac{1}{2}(900)(100) = 45,000$

The building will be 45 thousand cubic meters.

Answer to Exercise 43 (on page 182)

Solve for when the velocity is zero.

$$t = \frac{12}{9.8} = 1.22 \text{ seconds after release.}$$

Answer to Exercise 44 (on page 194)

For what t is $-4.9t^2 + 12t + 2 = 0$? Start by dividing both sides of the equation by -4.9.

$$t^2 - 2.45t - 0.408 = 0$$

The roots of this are at

$$x = -\frac{b}{2} \pm \frac{\sqrt{b^2 - 4c}}{2} = -\frac{-2.45}{2} \pm \frac{\sqrt{(-2.45)^2 - 4(-0.408)}}{2} = 1.22 \pm 1.36$$

We only care about the root after we release the hammer ($t > 0$).

$1.22 + 1.36 = 2.58$ seconds after releasing the hammer, it will hit the ground.

Answer to Exercise 45 (on page 202)

- $[1, 2, 3] + [4, 5, 6] = [5, 7, 9]$
- $[-1, -2, -3, -4] + [4, 5, 6, 7] = [3, 3, 3, 3]$
- $[\pi, 0, 0] + [0, \pi, 0] + [0, 0, \pi] = [\pi, \pi, \pi]$

Answer to Exercise 46 (on page 202)

To get the net force, you add the two forces:

$$\mathbf{F} = [4.2, 5.6, 9.0] + [-100.2, 30.2, -9.0] = [-96, 35.8, 0.0] \text{ newtons}$$

Answer to Exercise 47 (on page 203)

- $2 \times [1, 2, 3] = [2, 4, 6]$
- $[-1, -2, -3, -4] \times -3 = [3, 6, 9, 12]$
- $\pi[\pi, 2\pi, 3\pi] = \pi^2, 2\pi^2, 3\pi^2$

Answer to Exercise 48 (on page 205)

- $|[1, 1, 1]| = \sqrt{3} \approx 1.73$
- $|[-5, -5, -5]| = |-5 \times [1, 1, 1]| = 5\sqrt{3} \approx 8.66$
- $|[3, 4, 5] + [-2, -3, -4]| = |[1, 1, 1]| = \sqrt{3} \approx 1.73$

Answer to Exercise 49 (on page 210)

The momentum of the first car is 12,000 kg m/s in the north direction.

The momentum of the second car is 24,000 kg m/s in the east direction.

The new object will be moving northeast. What angle is the angle compared with the east?

$$\theta = \arctan \frac{12,000}{24,000} \approx 0.4636 \text{ radians} \approx 26.565 \text{ degrees north of east}$$

The magnitude of the momentum of the new object is $\sqrt{12,000^2 + 24,000^2} \approx 26,833 \text{ kg m/s}$

Its new mass is 2,500 kg. So the speed will be $26,833/2,500 = 10.73 \text{ m/s}$.

Answer to Exercise 50 (on page 212)

The original forward momentum was 1.2 kg m/s. The original kinetic energy is $(1/2)(0.4)(3^2) = 1.8 \text{ joules}$.

Let s be the post-collision speed of the ball that had been at rest. Let x and y be the forward and sideways speeds (post-collision) of the other ball. Conservation of kinetic energy says

$$(1/2)(0.4)(s^2) + (1/2)(0.4)(x^2 + y^2) = 1.8$$

Forward momentum is conserved:

$$0.4 \frac{s}{\sqrt{2}} + 0.4x = 1.2$$

Which can be rewritten:

$$x = 3 - \frac{s}{\sqrt{2}}$$

Sideways momentum stays zero:

$$(0.4) \frac{s}{\sqrt{2}} - 0.4y = 0.0$$

Which can be rewritten:

$$y = \frac{s}{\sqrt{2}}$$

Substituting into the conservation of kinetic energy equation above:

$$(1/2)(0.4)(s^2) + (1/2)(0.4)\left(3 - \frac{s}{\sqrt{2}}\right)^2 + \left(\frac{s}{\sqrt{2}}\right)^2 = 1.8$$

Which can be rewritten:

$$s^2 - \frac{3}{\sqrt{2}}s + 0 = 0$$

There are two solutions to this quadratic: $s = 0$ (before collision) and $s = \frac{3}{\sqrt{2}}$. Thus,

$$y = \frac{3}{2}$$

and

$$x = 3 - \frac{3}{2} = \frac{3}{2}$$

So both balls careen off at 45° angles at the exact same speed.

Answer to Exercise 51 (on page 214)

- $[1, 2, 3] \cdot [4, 5, -6] = 4 + 10 - 18 = -4$
- $[\pi, 2\pi] \cdot [2, -1] = 2\pi - 2\pi = 0$
- $[0, 0, 0, 0] \cdot [10, 10, 10, 10] = 0 + 0 + 0 + 0 = 0$

Answer to Exercise 52 (on page 215)

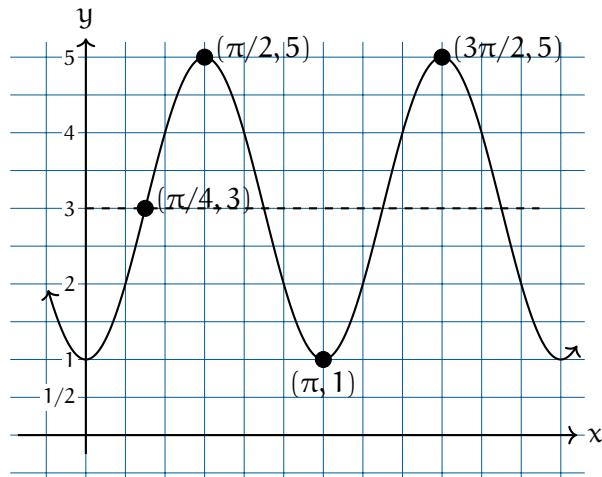
- $[1, 0] \cdot [0, 1] = 0$. The angle must be $\pi/2$.
- $[3, 4] \cdot [4, 3] = 24$. $\|[3, 4]\| \|[4, 3]\| \cos(\theta) = 24$. $\cos(\theta) = \frac{24}{(5)(5)}$. $\theta = \arccos(\frac{24}{25}) \approx 0.284$ radians.

Answer to Exercise 53 (on page 252)

$$p = 101,332 \times \left(1 - 2.25577 \times 10^{-5} \times h\right)^{5.25588}$$

and $h = 9,144$. Thus,

$$p \approx 30.1 \text{kPa}$$

Answer to Exercise 54 (on page 287)

This wave has an amplitude of 2. Its baseline has been translated up to 3.

This wave has wavelength of π . A sine wave usually has a wavelength of 2π , so we need to compress the x axis by a factor of 2.

The wave first crosses its baseline at $\pi/4$. The sine wave starts by crossing its baseline, so we need to translate the curve right by $\pi/4$.

$$f(x) = 2 \sin\left(2x - \frac{\pi}{4}\right) + 3$$

Answer to Exercise 55 (on page 292)

A is 440 Hz. Each half-step is a multiplication by $\sqrt[12]{2} = 1.059463094359295$ So the frequency of E is $(440)(2^{7/12}) = 659.255113825739859$

Answer to Exercise 56 (on page 308)

The force of gravity is $9.8 \times 45 = 441$ newtons.

At any speed s , the force of wind resistance is $0.05 \times s^2 = 0.05s^2$ newtons.

At terminal velocity, $0.05s^2 = 441$.

Solving for s , we get $s = \sqrt{\frac{441}{0.05}}$

Thus, terminal velocity should be about 94 m/s.

Answer to Exercise 57 (on page 318)

$$\frac{120 \text{ km}}{1 \text{ hour}} = \frac{1000 \text{ m}}{1 \text{ km}} \frac{120 \text{ km}}{1 \text{ hour}} \frac{1 \text{ hour}}{3600 \text{ seconds}} = 33.3 \text{ m/s}$$

$$F = \frac{mv^2}{r} = \frac{0.4(33.3)^2}{200} = 2.2 \text{ newtons}$$

Answer to Exercise 57 (on page 322)

$$v = \sqrt{3.721(3.4 \times 10^6)} = 3,557 \text{ m/s}$$

The circular orbit is $2\pi(3.4 \times 10^6) = 21.4 \times 10^6$ meters in circumference.

The period of the orbit is $(21.4 \times 10^6)/3,557 \approx 6,000$ seconds.

Answer to Exercise 59 (on page 354)

The earth and 1 kg on the surface would attract each other with a force of:

$$F_g = \frac{(6.67430 \times 10^{-17})(5.97219 \times 10^{24})(1)}{6,371^2} = \frac{3.98583 \times 10^8}{4.0590 \times 10^7} = 9.7987 \text{ N}$$

Thus, if the earth were still and alone in the universe, the oceans would form a perfect sphere.

Answer to Exercise 60 (on page 355)

$$F_c = \frac{(1)(465)^2}{6,371,000} = 0.03373 \text{ N}$$

So the spinning of the earth is trying to throw you into space, but the force of gravity is about 289 times more powerful.

This centripetal force decreases as you move from the equator to the north pole. In fact, at the north pole, there is no centripetal force. Thus, the spinning of the earth makes the oceans an oblate ellipsoid instead of a perfect sphere: the diameter going from pole-to-pole is shorter than a diameter measured at the equator.

You should feel a teensy-tiny bit lighter on your feet at the equator than you do at the north pole: 0.34% lighter.

Answer to Exercise 61 (on page 355)

Overhead, the moon is $384,467 - 6,371 = 378,096$ km from your 1 kg mass.

$$F_g = \frac{gm_1m_2}{r^2} = \frac{(6.67430 \times 10^{-17})(7.347673 \times 10^{22})(1)}{378,096^2} = \frac{4.9040574 \times 10^6}{1.42956585216 \times 10^{11}} = 3.43058 \times 10^{-5} \text{ N}$$

This is a very small force: The force due to earth's gravity is nearly three hundred thousand times stronger.

Underfoot, the moon is $384,467 + 6,371 = 390,838$

$$F_g = \frac{gm_1m_2}{r^2} = \frac{(6.67430 \times 10^{-17})(7.347673 \times 10^{22})(1)}{390,838^2} = \frac{4.9040574 \times 10^6}{1.52754 \times 10^{11}} = 3.2103 \times 10^{-5} \text{ N}$$

The force due to the moon's gravity is about 6% stronger when the the moon is overhead than when it is underfoot.

Answer to Exercise 62 (on page 356)

If we let r be the distance (in km) from the center of the earth to the center of mass, the distance from the center of the mass to the center of the moon is $384,467 - r$.

To find the balance point, multiply each mass by how far it is from the center of mass:

$$(5.97219 \times 10^{24})r = (7.347673 \times 10^{22})(384,467 - r)$$

Solving for r :

$$r = \frac{4,730.15}{1 + 0.0123} = 4,673 \text{ km}$$

The point on the earth closest to this? It is where the moon is directly overhead. The it is $6,371 - 4,673 = 1,698$ km from the center of mass.

The point on the earth farthest from this? It is where the moon is directly underfoot. The it is $6,371 + 4,673 = 11,044$ km from the center of mass.

Answer to Exercise 63 (on page 357)

First, lets figure out ω . It travels through 2π radians in 27.3 days. $27.3 \text{ days} = 2,358,720 \text{ seconds}$. $\omega = \frac{2\pi}{2,358,720} = 2.663811435 \times 10^{-6}$

$$F_c = (1)(11,044,000)(2.663811435 \times 10^{-6})^2 = 7.8365 \times 10^{-5}$$

Now the weakest:

$$F_c = (1)(1,698,000)(2.663811435 \times 10^{-6})^2 = 1.20512 \times 10^{-5}$$

Answer to Exercise 64 (on page 358)

Closest to the moon, the gravitational force of the moon and the centripetal forces are in the same direction: toward the moon.

$$F_{\text{total}} = 1.20488 \times 10^{-5} + 3.43045 \times 10^{-5} = 4.6356 \times 10^{-5} \text{ N}$$

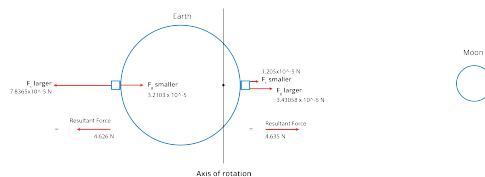
Farthest from the moon, the gravitational force of the moon and the centripetal forces are in opposite directions:

$$F_{\text{total}} = 7.8367 \times 10^{-5} - 3.2104 \times 10^{-5} = 4.62604 \times 10^{-5} \text{ N}$$

This is great conclusion: The two forces are basically equal: one pulls the water closest to the moon toward the moon, the other pulls water farthest from the moon away from the moon.

Both forces are pretty small: The force due to earth's gravity is about 211,000 times more than either.

And that is why there are two basically equally large high tides every day.



Answer to Exercise 65 (on page 360)

$$\frac{300 \times 10^6}{5.66 \times 10^{14}} = 530 \times 10^{-9} = 530 \text{ nm}$$

Answer to Exercise 66 (on page 370)

The two triangles are similar, one is 2 m and 3m. The other is x cm and 3 cm.

The image of the cow is 2 cm tall.

Answer to Exercise 67 (on page 382)

Assuming the mirror is truly vertical and the floor is truly horizontal, the new cut off should be exactly the same as the old one: It should be below your chin the same amount that your eyes are above your chin.

Illustration Here

Answer to Exercise 68 (on page 382)

Are there white photons? No. What we call “white” is a blend of photons that are several different colors.

Some people like to say white light is the combination of all visible colors of photons in equal amounts. That seems oddly specific and unusual.

Maybe it is better to imagine it from the human experience of white light. In our eyes, we have three different types of color-sensing cones, which generally correspond to the red, the green, and the blue regions of the spectrum. When all three are excited about equal amounts, humans experience that as white. On your computer screen, for example, what you see as white is just a blend of three colors of photons: a red, a green, and a blue.

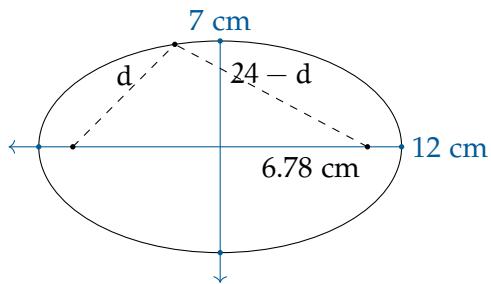
Are there black photons? No. What we call “black” is an absence of photons in the visible range.

Are there yellow photons? Yes! There is a region of the color spectrum that is yellow. It has a wavelength of about 527 nm. Photons at this energy level excite both our green-sensitive and red-sensitive cones. Your computer monitor does not actually create light with a 527 nm wavelength. Instead, it creates red light and green light, which our eyes interpret as yellow.

Answer to Exercise 69 (on page 386)

The length of the string is easy: $2 \times 12 = 24$ cm.

The distance from the center to the focal point is $\sqrt{12^2 - 7^2}$ approx 6.78 cm.



Answer to Exercise 70 (on page 390)

We need the distance from the center out to each of the three axes. We know that $a = \left(\frac{1}{2}\right) 30 = 15$ cm.

We can calculate the b and c (which are equal) using the circumference given: $2b\pi = 60$, so $c = b \approx 9.55$ cm.

The volume, then is

$$V = \frac{4}{3}\pi(15)(9.55)(9.55) \approx 5,730 \text{ cubic centimeters}$$

The mass would be $5,730 \times 11.34 = 64,973$ grams or about 65 kg.

Answer to Exercise 71 (on page 408)

$$-2x^3 + \frac{1}{2}x + 3.9$$

$$(4.5)x^2 + \pi x$$

$$\boxed{7}$$

$$2x^{-10} + 4x - 1$$

$$x^{\frac{2}{3}}$$

$$\boxed{3x^{20} + 2x^{19} - 5x^{18}}$$

Answer to Exercise 72 (on page 409)

Standard form would be $-x^3 + 21x^2 - 1000x + \pi$. The degree is 3. The leading coefficient is -1 .

Answer to Exercise 73 (on page 409)

$4^3 - (3)(4^2) + (10)(4) - 12 = 64 - 48 + 40 - 12$. So $y = 44$

Answer to Exercise 74 (on page 412)

```
favorites[3] = "Gloves"
```

Answer to Exercise 75 (on page 415)

```
pn2 = [2.5, 5.0, -2.0, -7.0, 4.0]
y = evaluate_polynomial(pn2, 8.5)
print("Polynomia 2: When x is 8.5, y is", y)

pn3 = [-9.0, 0.0, 0.0, 0.0, 0.0, 5.0]
y = evaluate_polynomial(pn3, 2.0)
print("Polynomial 3: When x is 2.0, y is", y)
```

Answer to Exercise 76 (on page 417)

The polynomial crosses the y-axis at 6. When x is zero, all the terms are zero except the last one. Thus you can easily tell that $x^3 - 7x + 6$ will cross the y-axis at $y = 6$.

Looking at the graph, you tell that the curve crosses the y-axes near -3, 1 and 2. If you plug those numbers into the polynomial, you would find that it evaluates to zero at each one. Thus, $x = -3$, $x = 1$, and $x = 2$ are roots.

Answer to Exercise 77 (on page 420)

$3x^3 - 7x^2 + x - 18$ and $3x^5 - 7x^3 + x^2 - 12$

Answer to Exercise 78 (on page 421)

$x^3 - 3x^2 + 5x$ and $x^5 - 3x^3 + 5x^2 - 2x + 6$

Answer to Exercise 79 (on page 422)

```
def add_polynomials(a, b):
    degree_of_result = max(len(a), len(b))
    result = []
    for i in range(degree_of_result):
        if i < len(a):
            coefficient_a = a[i]
        else:
            coefficient_a = 0.0

        if i < len(b):
            coefficient_b = b[i]
        else:
            coefficient_b = 0.0

        result.append(coefficient_a + coefficient_b)
    return result
```

Answer to Exercise 80 (on page 423)

```
def subtract_polynomial(a, b):
    neg_b = scalar_polynomial_multiply(-1.0, b)
    return add_polynomials(a, neg_b)
```

Answer to Exercise 81 (on page 426)

$$(3x^2)(5x^3) = 15x^5$$

$$(2x)(4x^9) = 8x^{10}$$

$$(-5.5x^2)(2x^3) = -11x^5$$

$$(\pi)(-2x^5) = -2\pi x^5$$

$$(2x)(3x^2)(5x^7) = 30x^{10}$$

Answer to Exercise 82 (on page 427)

$$(3x^2)(5x^3 - 2x + 3) = 15x^6 - 6x^3 + 6x^2$$

$$(2x)(4x^9 - 1) = 8x^{10} - 2x$$

$$(-5.5x^2)(2x^3 + 4x^2 + 6) = 11x^5 - 22x^4 + 33x^2$$

$$(\pi)(-2x^5 + 3x^4 + x) = -2\pi x^5 + 3\pi x^4 + \pi x$$

$$(2x)(3x^2)(5x^7 + 2x) = 30x^{10} + 12x^4$$

Answer to Exercise 83 (on page 429)

$$(2x + 1)(3x - 2) = 6x^2 - x - 2$$

$$(-3x^2 + 5)(4x - 2) = -12x^3 + 6x^2 + 20x - 10$$

$$(-2x - 1)(-3x - \pi) = 6x^2 + (4 + 2\pi)x + \pi$$

$$(-2x^5 + 5x)(3x^5 + 2x) = -6x^{10} + 12x^6 + 10x^2$$

Answer to Exercise 84 (on page 429)

The degree of the product is determined by the term that is the product of the highest degree term in p_1 and the highest degree term in p_2 . Thus, the product of a degree 23 polynomial and a degree 12 polynomial has degree 35.

Answer to Exercise 85 (on page 436)

Answer to Exercise 86 (on page 437)

```
def derivative_of_polynomial(pn):  
  
    # What is the degree of the resulting polynomial?  
    original_degree = len(pn) - 1  
    if original_degree > 0:
```

```
degree_of_derivative = original_degree - 1
else:
    degree_of_derivative = 0

# We can ignore the constant term (skip the first coefficient)
current_degree = 1
result = []

# Differentiate each monomial
while current_degree < len(pn):
    coefficient = pn[current_degree]
    result.append(coefficient * current_degree)
    current_degree = current_degree + 1

# No terms? Make it the zero polynomial
if len(result) == 0:
    result.append(0.0)

return result
```

Answer to Exercise 87 (on page 446)

$$(2x - 3)(2x + 3) = 4x^2 - 9$$

$$(7 + 5x^3)(7 - 5x^3) = 49 - 25x^6$$

$$(x - a)(x + a) = x^2 - a^2$$

$$(3 - \pi)(3 + \pi) = 9 - \pi^2$$

$$(-4x^3 + 10)(-4x^3 - 10) = 16x^6 - 100$$

$$(x + \sqrt{7})(x - \sqrt{7}) = x^2 - 7$$

$$x^2 - 9 = (x + 3)(x - 3)$$

$$49 - 16x^6 = (7 + 4x^3)(7 + 4^3)$$

$$\pi^2 - 25x^8 = (\pi + 5x^4)(\pi - 5x^4)$$

$$x^2 - 5 = (x + \sqrt{5})(x - \sqrt{5})$$

Answer to Exercise 88 (on page 449)

$$(x + 1)^2 = x^2 + 2x + 1$$

$$(3x^5 + 5)^2 = 9x^{10} + 30x^5 + 25$$

$$(x^3 - 1)^2 = x^6 - 2x^3 + 1$$

$$(x - \sqrt{7})^2 = x^2 - 2x\sqrt{7} + 7$$

Answer to Exercise 89 (on page 450)

$$(x + \pi)^5 = x^5 + 5\pi x^4 + 10\pi^2 x^3 + 10\pi^3 x^2 + 5\pi^2 x + \pi^5$$

Answer to Exercise 90 (on page 453)

Answer to Exercise 91 (on page 453)

Answer to Exercise 92 (on page 469)

$$\lim_{x \rightarrow -6^-} p(x) = -\infty, \quad \lim_{x \rightarrow -6^+} p(x) = \infty$$

$$\lim_{x \rightarrow -5^-} p(x) = \lim_{x \rightarrow -5^+} p(x) = \lim_{x \rightarrow -5} p(x) = 1$$

$$\lim_{x \rightarrow -3^-} p(x) = \lim_{x \rightarrow -3^+} p(x) = \lim_{x \rightarrow -3} p(x) = \frac{1}{3}$$

$\lim_{x \rightarrow \infty} p(x) = 0$ called simply a limit, although it is a left-hand limit

Answer to Exercise 93 (on page 472)

$$\lim_{x \rightarrow -\infty} 3^x + 1 = 1; \lim_{x \rightarrow 4^+} \log_2(x - 4) = -\infty; \lim_{x \rightarrow \infty} 2^{1-x} = 0; \lim_{x \rightarrow 0^-} \log_{10}(-2x) = -\infty$$

Answer to Exercise 94 (on page 473)

$$\lim_{x \rightarrow -\infty} \tan^{-1} x = -\frac{\pi}{2}, \lim_{x \rightarrow \infty} \tan^{-1} x = \frac{\pi}{2}; \lim_{x \rightarrow -\infty} \frac{1}{1+e^{-x}} = 0, \lim_{x \rightarrow \infty} \frac{1}{1+e^{-x}} = 1$$

Answer to Exercise 95 (on page 479)

x-intercept: $(-5/2, 0)$; y-intercept: $(0, 5/4)$; horizontal asymptote: $y = 2$; vertical asymptote: $x = -4$

Answer to Exercise 96 (on page 481)

Factored form: $\frac{x^2(x+2)}{x(x+1)}$; hole: $(0, 0)$; vertical asymptote: $x = -1$; oblique asymptote: $y = x + 1$

Answer to Exercise 97 (on page 506)

$$Av = (11, 37, -43)$$

Answer to Exercise 98 (on page 510)

$$\mathbf{w} = 2 * [2, 4, 8] + (-2) * [8, -6, 3] + 4 * [7, 9, 2]$$

$$\mathbf{w} = [4, 8, 16] + [-16, 12, -6] + [28, 36, 8]$$

$$\mathbf{w} = [16, 56, 18]$$

Answer to Exercise 99 (on page 512)

$$\text{TotalSales} = 50 * 300 + 75 * 100 + 150 * 50 = 30,000$$

$$\text{NumberTickets} = 300 + 100 + 50 = 450$$

$$\text{WeightedAverage} = 30,000/450 = 66.67$$

Answer to Exercise 100 (on page 516)

Rewrite as a system of equations:

$$\begin{aligned} 2 * a_1 + 2 * a_2 + 0 * a_3 &= 0 \\ 1 * a_1 - 1 * a_2 + 1 * a_3 &= 0 \\ 4 * a_1 + 2 * a_2 - 2 * a_3 &= 0 \end{aligned}$$

Simplify

$$\begin{aligned} 2a_1 + 2 * a_2 &= 0 \\ a_1 - a_2 + a_3 &= 0 \\ 4a_1 + 2a_2 - 2a_3 &= 0 \end{aligned}$$

Swap row 2 and 1:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 2a_1 + 2 * a_2 &= 0 \\ 4a_1 + 2a_2 - 2a_3 &= 0 \end{aligned}$$

Multiply row 1 by -2 and add to row 2:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 0 + 3 * a_2 - 2a_3 &= 0 \\ 4a_1 + 2a_2 - 2a_3 &= 0 \end{aligned}$$

Multiply row 1 by -4 and add to row 3:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 0 + 3 * a_2 - 2a_3 &= 0 \\ 0 + 6a_2 - 6a_3 &= 0 \end{aligned}$$

Multiply row 2 by -4 and add to row 3:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 0 + 3 * a_2 - 2a_3 &= 0 \\ 0 + 0 - 2a_3 &= 0 \end{aligned}$$

Multiply row 3 by -1 and add to row 2:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 0 + 3 * a_2 + 0 &= 0 \\ 0 + 0 - 2a_3 &= 0 \end{aligned}$$

Divide row 3 by -2 and row 2 by $\frac{1}{3}$:

$$\begin{aligned} a_1 - a_2 + a_3 &= 0 \\ 0 + a_2 + 0 &= 0 \\ 0 + 0 + a_3 &= 0 \end{aligned}$$

Backsubstitute a_2 and a_3 into row 1:

$$\begin{aligned} a_1 + 0 + 0 &= 0 \\ 0 + a_2 + 0 &= 0 \\ 0 + 0 + a_3 &= 0 \end{aligned}$$

Therefore

$$a_1 = a_2 = a_3 = 0$$

Answer to Exercise 101 (on page 523)

$$A = A^t = \begin{bmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{bmatrix}$$

Answer to Exercise ?? (on page 529)

Compute dot product of **a** and **b**:

$$1 * -4 + 3 * 6 = -4 + 18 = 14$$

Compute the dot product of **b** and **b**

$$16 + 36 = 52$$

$$14/52 * (-4, 6) = (-1.076, 1.61)$$

Answer to Exercise 103 (on page 534)

The first vector of the orthogonal subspace is:

$$v_1 = x_1 = (1, 1, 1)$$

The second vector of the subspace is a projection of x_2 onto v_1 .

$$v_2 = x_2 - \frac{x_2 v_1}{v_1 v_1} v_1$$

Substitute the values:

$$v_2 = (0, 1, 1) - \frac{(0, 1, 1)(1, 1, 1)}{(1, 1, 1)(1, 1, -1)} (1, 1, 1)$$

$$v_2 = (0, 1, 1) - (2/3)(1, 1, 1)$$

$$v_2 = (-2/3, 1/3, 1/3)$$

Normalize:

$$v_1 = v_1 / \sqrt{|v_1|}$$

$$\begin{aligned}
 v_1 &= (1, 1, 1) / \sqrt{|v_1|} \\
 v_1 &= (0.577, 0.577, 0.577) \\
 v_2 &= v_2 / \sqrt{|v_2|} \\
 v_2 &= (0, 1, 1) \sqrt{|v_2|} \\
 v_2 &= (-0.816, 0.408, 0.408)
 \end{aligned}$$

Answer to Exercise 104 (on page 548)

$$\begin{aligned}
 U &= \begin{bmatrix} -0.70710678 & -0.70710678 \\ -0.70710678 & 0.70710678 \end{bmatrix} \\
 \text{Singularvalues} &= [3.464101623, 1.16227766] \\
 V^T &= \begin{bmatrix} -0.408 & -0.816 & -0.408 \\ -0.894 & 0.447 & 0.0 \\ -0.183 & -0.365 & 0.9129 \end{bmatrix}
 \end{aligned}$$

Answer to Exercise 105 (on page 562)

```
SELECT bike_id FROM bike WHERE purchase_price > 330 AND brand='Trek'
```

Answer to Exercise 106 (on page 597)

probability of all 5's = $\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \left(\frac{1}{6}\right)^3 = \frac{1}{216} \approx 0.0046$

Answer to Exercise 106 (on page 597)

probability of at least one heads = $1.0 - \text{probability of all tails} = 1.0 - \left(\frac{1}{2}\right)^5 = 1.0 - \frac{1}{32} = \frac{31}{32} \approx 0.97$



Answer to Exercise 108 (on page 639)

```
from scipy.stats import norm

# Constants
MEAN = 164.7
STD = 7.1

# What is the cutoff for the top decile?
cutoff = norm.ppf(0.9, loc=MEAN, scale=STD);
print(f"To be in the top 10 percent, you must be at least {cutoff:.2f} cm")

# What proportion of women are between 160cm and 165cm?
shorter_than_160 = norm.cdf(160, loc=MEAN, scale=STD)
shorter_than_165 = norm.cdf(165, loc=MEAN, scale=STD)
between = shorter_than_165 - shorter_than_160
print(f"{between * 100.0:.2f}% of adult women are between 160 and 165 cm.")
```

When run, this will give you:

```
> python3 women.py
To be in the top 10 percent, you must be at least 173.80 cm
26.29% of adult women are between 160 and 165 cm.
```

Answer to Exercise 110 (on page 647)

First, we convert the temperatures into Kelvin:

- Dawn: $12 + 273.15 = 285.15$
- Noon: $28 + 273.15 = 301.15$

So, the temperature T has increased by a factor of $\frac{301.15}{285.15} \approx 1.056$

Thus the volume of the air mattress has also increased by a factor of 1.056.

So the air mattress that had a volume of 1000 liters at dawn, will have a volume 1056 liters at noon.

Answer to Exercise 110 (on page 647)

What is the pressure in kPa?

- Before squeezing: 100 kPa
- While squeezing: 120 kPa

So, the pressure P has increased by a factor of $\frac{120}{100} = 1.2$

$$1/1.2 \approx 0.833$$

The air in the bottle had a volume of 1 liter before squeezing, so it has a volume of 833 milliliters while being squeezed.

Answer to Exercise 111 (on page 648)

First, let's convert the known values to the right unit:

- Radius = 0.12 m
- Length = 0.5 m
- $T = 20 + 273.15 = 293.15$ degrees Kelvin
- $P = 600 \text{ kPa} = 600,000 \text{ Pa}$

The volume of the cylindrical chamber is $V = \pi r^2 h = \pi(0.12)^2 0(0.5) \approx 0.0226$.

The Ideal Gas Law tell us that $PV = nRT$. We are solving for n.

$$n = \frac{PV}{RT} = \frac{(600,000)(0.0226)}{(8.31446)(293.15)} \approx 5.68 \text{ moles of O}_2$$

Answer to Exercise 112 (on page 652)

$$E = C_{V,m}(3 \text{ moles})(20 \text{ degrees Celsius}) = (12.47)(3)(20) = 748 \text{ Joules}$$

$$E = C_{V,m}(3 \text{ moles})(20 \text{ degrees Celsius}) = (20.8)(3)(20) = 1247 \text{ Joules}$$

Answer to Exercise 113 (on page 654)

10 degrees Celsius is 283.15 degrees Kelvin. 30 degrees Celsius is 303.15.

For any gas:

$$E_K = C_{V,m} n T$$

And $C_{V,m} = 12.47$ for all monoatomic gases.

So the energy at 10 degrees Celsius:

$$E_1 = (12.47)(3)(283.15) = 10,594 \text{ Joules}$$

The energy at 30 degrees Celsius:

$$E_2 = (12.47)(3)(303.15) = 11,342 \text{ Joules}$$

The difference?

$$E_2 - E_1 = 11,342 - 10,594 = 748 \text{ Joules}$$

Which is consistent with your earlier exercise.

Answer to Exercise 114 (on page 668)

When one mole of water goes from 33° to 17° , it will give off $(75.38)(33 - 17) = 1,206$ Joules or 1.206 kJ.

Tom needs 300,000 kJ, so he needs $300,000 / 1.206 = 248,739.72$ moles of water.

How many liters is that? $248,739.72 * 0.018 = 4,477.31$ liters.

How many barrels is that? $4,477.31 / 159 = 28.16$ barrels. He will need 29 barrels.

Answer to Exercise 115 (on page 669)

When one mole of mirabilite goes from 33° to 17° , it will give off $(550)(33 - 17) + 82,000 = 90,800$ Joules or 90.8 kJ.

Tom needs 300,000 kJ, so he needs $300,000 / 90.8 = 3,304$ moles of mirabilite.

How many liters is that? $3,304 * 0.22 = 726.9$ liters.

How many barrels is that? $726.9 / 159 = 4.57$. He will need 5 barrels.

Answer to Exercise 116 (on page 679)

The pre-compression temperature is $80^\circ\text{C} + 273.15 = 353.15^\circ\text{K}$.

The radius of the cylinder is $9.535 / 2 = 4.7625$ cm.

The area of a cross section of the cylinder is $\pi r^2 = \pi(4.7625)^2 \approx 71.26$ ml.

So the change in volume between the minimum and maximum volume is $(71.26)(10.16) = 724$ ml, or 0.724 liters.

(With four cylinders, the total displacement of a Model T is thus $(4)(724) = 2,896$ cc.)

Now we use the ideal gas to figure out how many moles of gas will fit into 0.724 liters at 100 kPa and 353.15°K .

$$n = \frac{PV}{RT} = \frac{(100)(0.724)}{(8.314)(353.15)} = 0.02466 \text{ moles of air+fuel}$$

So, if we suck n_a moles of air and n_e moles of vaporized ethanol in to the cylinder:

$$n_a + n_e = 0.02466$$

So

$$n_a = 0.02466 - n_e$$

21% of n_a is O_2 :

$$n_{O_2} = 0.21n_a = 0.21(0.02466 - n_e) = 0.005178 - 0.21n_e$$

For a clean burn, we need 3 times as many O_2 molecules as ethanol molecules. Thus:

$$3n_e = n_{O_2} = 0.005178 - 0.21n_e$$

Solving for n_e :

$$n_e = \frac{0.005178}{3.21} = 0.001613 \text{ moles of ethanol}$$

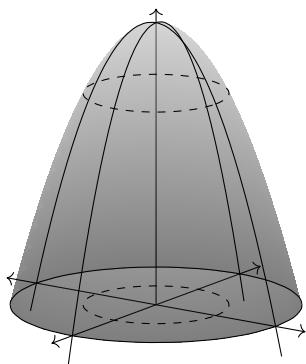
For every molecule of ethanol that burns we get 2 molecules of CO_2 : 0.003226 moles.

For every molecule of ethanol that burns we get 3 molecules of H_2O : 0.004839 moles.

How much heat? $(0.001613)(1330) = 2.145$ kilojoules from each burn.

Answer to Exercise 117 (on page 706)

We are finding the volume of the solid that lies under the surface $z = 4 - x^2 - y^2$ and above the xy -plane.



We can use polar coordinates to simplify the double integral. In polar coordinates, $x = r\cos(\theta)$ and $y = r\sin(\theta)$, so $x^2 + y^2 = r^2$. The volume under the surface and above the xy -plane is given by

$$V = \int \int (4 - r^2)r \, dr \, d\theta, \quad (126.1)$$

where r ranges from 0 to 2 (since $4 - r^2 \geq 0$ if $0 \leq r \leq 2$) and θ ranges from 0 to 2π .

Hence,

$$\begin{aligned} V &= \int_0^{2\pi} \int_0^2 (4r - r^3) dr d\theta \\ &= \int_0^{2\pi} \left[2r^2 - \frac{1}{4}r^4 \right]_0^2 d\theta \\ &= \int_0^{2\pi} (8 - 4) d\theta \\ &= \int_0^{2\pi} 4 d\theta \\ &= [4\theta]_0^{2\pi} \\ &= 8\pi. \end{aligned}$$

So the volume of the solid is 8π cubic units.

Answer to Exercise 118 (on page 719)

A	B	not (A or B)	(not A) and (not B)
F	F	T	T
F	T	F	F
T	F	F	F
T	T	F	F

Notice that the two expressions are equivalent!

DeMorgan's Rule says "not (A or B)" is equivalent to "(not A) and (not B)".

It also says "not (A and B)" is equivalent to "(not A) or (not B)"



INDEX

- \mathbb{R}^2 , 786
- \emptyset , 713
- \mathbb{C} , 713
- \mathbb{N} , 713
- \mathbb{Q} , 713
- \mathbb{R} , 713
- \mathbb{Z} , 713
- absorption
 - photon, 379
- acceleration, 180
- Accents, 583
- accuracy, 778
- acute triangle, 135
- AdaBoost, 797
- adding
 - monomials, 419
 - polynomials, 419
- Alphabets, 583
- amp or ampere, 103
- and, 713
- Antiderivatives, 625
- atmospheric pressure, 249
- atom, 22
- atomic mass, 34
- atomic mass unit, 32
- Avogadro's number, 35
- Bagging, 795
- bar chart, 242
- bar graph
- in python, 602
- barometric pressure, 249
- Bayesian Network, 759
- Bellman-Ford algorithm, 758
- Binary Search, 755
- boolean variables, 721
- Boosting, 797
- BTU, 39
- calories, 39
- cardinality, 719
- career, 21
- cement, 123
- chain rule, 490
- chemical energy, 39
- chemical reaction, 24
- choose function, 607
- circle
 - area of, 152
 - reflections in, 383
- circumference, 152
- class in python, 440
- Classification, 761
- coefficient
 - polynomial, 407
- collisions in hash tables, 732
- combinatorics, 605
- Complex Numbers, 195
- compound interest, 237
- concrete, 123
- confusion matrix, 775

- conic sections, 175
constant rule, 489
contrapositive, 722
conversion factors, 45
Coulomb's law, 115
coulombs, 103
covariance matrix, 710
- data compression, 571
decision trees, 793
deep learning, 805
degree
 polynomial, 407
depth-first search, 757
derivative, 487
DFS, 757
Diacritical Marks, 583
differential equations, 683
differentiation
 polynomials, 435
diffusion, 380
Dijkstra's Algorithm, 743
Dijkstra, Edsger, 743
discrete vs. discreet, 595
distance
 in 3 dimensions, 142
distance using Pythagorean theorem, 141
dot product, 213
- e, 271
earth
 shape of, 389
edge, 735
efficiency, 41
eigenvalue, 537
eigenvector, 537
electricity, 39
electrons, 22
elements, 24
ellipse, 384
 focus points, 384
ellipsoid, 388
end
 behavior, 478
endothermic, 25
energy
 conservation of, 41
- Forms of, 38
entropy, 572
equilateral triangle, 133
equilibrium, 59
exothermic, 25
exponential decay, 268
exponential growth, 239
exponents, 261
 fractions, 263
 negative, 262
 zero, 262
- f1 score, 779
factorial, 606
factoring polynomials, 451
fertilizer, 120
floats
 formatting, 207
focal length, 398
fundamental theorem of calculus, 627
- Gaussian distribution, 709
generalized linear models, 789
geocoding, 589
Gini impurity, 794
GLMs, 789
gradient, 502
Gradient Boosted Trees, 798
gradient descent, 787
Gram-Schmidt process, 532
graph, 735
 connected, 736
 database, 736
 directed, 736
 undirected, 736
graph theory, 736
- Haber-Bosch process, 121
half-life, 267
hash function, 731
hash table, 731
Haversine formula, 344
heat, 39
histograms, 89
hole, 478
horizontal
 asymptote, 477

-
- HTML, 575
 HTTP, 567, 569
 identity link function, 791
 if and only if, 717
 implicit differentiation, 493
 implies, 717
 independent, 597
 integration, 183
 intersection, 715
 isosceles triangle, 134
 isotopes, 31
 Joule, 37
 Joule's law, 105
 k-Means Clustering, 799
 k-nearest neighbor, 781
 kinetic energy, 40
 Kirchhoff's voltage law, 111
 latitude, 341
 lenses, 397
 linalg, 206
 line graph, 243
 linear combinations, 509
 linear regression, 787
 link functions, 791
 linked list, 725
 lists, python, 186
 ln, 271
 log, 269
 in python, 270
 log link function, 791
 logarithm, 269
 change of base, 271
 identities, 270
 natural, 271
 logic, 717
 logit link function, 791
 longitude, 341
 matplotlib, 186, 585
 subplots, 188
 mean, 86
 median, 88
 metric system
 prefixes, 44
 millimeters mercury, 255
 mirror, 380
 mole, 35
 molecules, 24
 monomial, 407
 coefficient, 407
 degree, 407
 multiplication
 polynomials, 425
 multivariate normal, 709
 naive Bayes classifier, 783
 nautical mile, 344
 neural net classifiers, 803
 neural net regression, 801
 neutron, 31
 neutrons, 22
 nitrogen, 120
 nitrogen cycle, 120
 nitrogen fixation, 120
 node, 735
 normal distribution, 693
 not, 718
 np, 206
 null, 552
 NumPy, 206
 oblate spheroid, 389
 oblique
 asymptote, 480
 obtuse triangle, 135
 ODEs, 683
 Ohm's law, 105
 ohms, 105
 one-hot encoding, 769
 optimization, 491
 or, 713
 orbit
 elliptical, 388
 ordinary differential equation, 683
 parallel, 149
 partial derivative, 501
 partial differential equations, 684
 PDEs, 684
 periodic table of elements, 33
 permutations, 609

- composing, 610
- cycles, 612
- identity permutation, 610
- inverses, 611
- perpendicular, 149
- phosphorus, 120
- pie chart, 244
- plotly, 593
- polynomial, 407
 - definition of, 407
 - graphing, 457
- potash, 120
- potassium, 120
- potential energy
 - gravitational, 40
- power, 216
- power rule, 489
- power set, 721
- precision, 778
- pressure, 249
- Priority Queue, 751
- priority queue, 749
- probability, 596
- probit link function, 792
- product rule, 490
- projection, 527
- prolate spheroid, 389
- proton, 31
- protons, 22
- Pythagorean theorem, 140
- quadratic functions, 184
- quadratic mean, 91
- quitting, 21
- quotient rule, 490
- radioactive decay, 266
- Random Forest, 796
- random number generation, 599
- rational
 - expression, 475
 - function, 476
- rebar, 124
- recall, 778
- reflection, 379
 - law of, 381
- refractive index, 398
- Regression, 762
- residual, 785
- resistance, 104
 - in parallel, 113
- reverse geocoding, 590
- right triangle, 135
- RMS, 91
- RMSE, 786
- Root Mean Squared Error, 786
- root-mean-squared, 91
- samples, 86
- scatter plot, 245
- set, 712
 - sets of sets, 721
- singular value decomposition, 543
- sorting, 611
- specific heat capacity, 63
- Spreadsheet, 229
 - Entering formula, 94
- spreadsheet, 93
 - graphing multiple series, 247
 - graphs, 234
- standard form
 - polynomial, 408
- standardizing data, 788
- steel reinforced concrete, 124
- straw
 - drinking, 252
- subset, 714
- summation symbol, 86
- svd, 543
- symbolic vs. numeric solutions, 230
- text, 577
- thermal equilibrium, 63
- translucent, 380
- transparent, 380
- triangle, 133
 - triangle inequality, 134
- union, 715
- units table, 44
- urea, 121
- urine, 121
- Vantablack, 380

variance, 87
vectors, 199
 adding, 200
 angle between, 216
 in python, 205
 magnitude of, 204
 multiplying by a scalar, 202
 subtraction, 204
Venn diagrams, 715
vertical
 asymptote, 477
voltage, 104
volts, 105
volume
 oblique cylinder, 168
 pyramid, 169
 rectangular solid, 166
 right cylinder, 167
 sphere, 166
watts, 106
Web APIs, 569
weighted averages, 511
work, 38, 216

XOR, 723