

Class Notes
Introduction to Modern Physics
Physics 321 – Plan II
A Different Approach
Under Construction

Austin M. Gleeson¹
Department of Physics
University of Texas at Austin
Austin, TX 78712

January 23, 2021

¹gleeson@physics.utexas.edu

Preface

These notes were prepared as the primary text for the course **Phy 321** *Introduction to Modern Physics* which is required for almost all students majoring in the Plan II degree program at the University of Texas at Austin.¹ These notes provide a comprehensive review of the material covered by that course. The students taking the course have a broad range of backgrounds, many with second majors in the usual degree plans offered by the University. The course is offered in one semester and in its original form was intended to cover all the fundamental issues at the base of modern physics. After an initial attempt offering it at the sophomore level, it was decided that the material was more appropriate to higher level students and, in its current offering, the students are dominantly juniors and seniors. Regardless, because of the diversity of backgrounds of these students ranging from engineering to religious studies with all backgrounds in between there is a significant problem with setting the technical level of the course and thus the text. In order to assist the reader in working with the text, sections that do not materially effect the conceptual understanding of the material which is the primary intent of the course are marked with an asterisk. These asterisked sections are intended to supplement the regular sections by going further and showing how more technically developed approaches make the concepts more appropriate and specific to the complex world in which we find ourselves. Students in this course will not be expected to manipulate material with the tools used in the asterisked sections. They should review these sections for content independent of the details of the manipulations. This is an important ability to develop when reading or studying ideas that have evolved very elaborate technical descriptions. As is always the case, you must see the trees to see the forest but the essence of the forest is more than a collection of the botanical descriptions of the trees.

¹The Plan II Program is the original and most prestigious honors program offered by the University of Texas.

Another point of issue is that these notes are “Under Construction”. They are a living document and is constantly under revision. Because of the nature of the real world where distractions and interruptions are unavoidable, please put up with the paragraphs the seem to end without conclusion or even sentences that are not completed. The nature and purpose of this document requires that it never be completed and finished. It is more fun that way.

Contents

Preface	i
1 Introduction	1
1.1 Purpose of This Course	1
1.2 Physics that you should know	3
1.2.1 Introduction	3
1.2.2 Kinematics	4
1.2.3 Dynamics	5
1.2.4 An Important Aside on Causal Agents:	6
1.3 The Role of Mathematics	10
1.3.1 Mathematics and Symbols That You Should Know . .	11
1.4 First Day Handout	16
1.4.1 Fermi Problems	16
1.4.2 Things Everyone Should Know	16
1.4.3 Order of Magnitude Estimates	19
1.4.4 Home Experiments	20
1.4.5 Review Syllabus	21
1.4.6 Text	21
1.5 What is Physics?	22
1.5.1 Range of Phenomena	24
1.5.2 Reductionism and General Principles	28
1.6 A Joke	29
1.7 Problems	30
2 Measurement	33
2.1 The Role of Measurement	33
2.2 Measurability	34
2.3 Role of Standards	38
2.3.1 The Story of Length	40

2.3.2	Accuracy and Precision of Standards	42
2.4	Quantities of Physics	43
2.5	Dimensional Analysis	44
2.5.1	Uses of Dimensional Analysis	45
2.5.2	Scaling Laws	47
2.6	Fundamental Dimensional Constants	48
2.6.1	Sizes	48
2.6.2	Modern Standards	50
2.6.3	Useful Dimensions	52
3	Pre 19th Century Physics	55
3.1	Introduction	55
3.2	Least Time Formulation of Light Propagation	57
3.2.1	Speculation on the form of Fermat's Theory	62
3.3	Applications of Fermat's Principle	63
3.3.1	Light Travels in Straight Lines	63
3.3.2	Refraction & Snell's Law	64
3.3.3	Total Internal Reflection	68
3.3.4	Lenses	69
3.3.5	Rays in a General Inhomogeneous Space and Mirages.	71
3.3.6	Reflection and Mirrors	72
3.3.7	Mathematical Digression	76
3.4	Newton and Color	79
4	Fresnel/Young/Huygens Theory	83
4.1	Recapitulation of Fermat's Least time principal	83
4.2	Problems with Fermat's Least Time	85
4.3	Huygens	86
4.4	Thomas Young and Interference	88
4.4.1	Detail Analysis of Interference for the Double Slit	92
4.5	Phasors	98
4.5.1	Example of Three Slits and More	102
4.6	The Theory of Light Propagation	106
4.6.1	Mirror Reflection à la Fresnel	107
4.6.2	Diffraction Grating	113
4.6.3	Diffraction	115
4.6.4	Lens and Spherical Mirror Revisited	117
4.6.5	How do we get least time from Fresnel's Theory?	117
4.6.6	Recovery of Fermat's Least Time	119
4.7	Polarization	120

4.8	Introduction of the Field	121
5	19th Century Physics	123
5.1	Introduction	123
5.2	Action at a Distance and Field Dynamics	123
5.2.1	Action at a Distance	125
5.2.2	Local Field Theory	126
5.3	The Stretched String*	128
5.4	Maxwell's Theory of Electromagnetism	140
5.5	Other Classical Field Systems	151
5.5.1	Temperature	151
5.5.2	Sound	151
5.5.3	The "Wave"	151
5.5.4	Surface Disturbances on Water	151
6	Dynamics and Action	153
6.1	Dynamics	153
6.1.1	Background on Formulation of Action	153
6.1.2	Introduction to Action	155
6.1.3	Definition of Action	157
6.1.4	Trajectory of a Free Particle	160
6.1.5	Proof that the Least Action Reproduces Newtonian Physics	162
6.2	Examples of Actions	166
6.2.1	Gravitation near a flat earth	166
6.2.2	Same Example done another way*	169
6.2.3	Simple Scattering of Two Particles	171
6.2.4	A Small Sphere Hanging from a Fixed Point	176
6.2.5	Action for the Stretched String	177
6.2.6	Digression on averages and slicing*	178
7	Basic Principles of Physics	183
7.1	Symmetry	183
7.2	The Nature of Symmetry in Physics	188
7.2.1	Discrete Transformations	189
7.2.2	Continuous Transformations	189
7.2.3	Identity Transformation	190
7.2.4	Examples of symmetry in situations like physics	190
7.3	Examples of Symmetry in physics	191
7.3.1	Physics transformations:	192

7.4	Symmetry and Action	194
7.4.1	Introduction	194
7.4.2	Galilean Invariance	197
7.4.3	More on Symmetry and Action	198
7.4.4	Noether's Theorem	199
8	Special Classical Physical Systems	205
8.1	Introduction	205
8.2	The Harmonic Oscillator	205
8.2.1	Importance	205
8.2.2	Dynamics	207
8.2.3	Examples of harmonic oscillator systems	209
8.3	The Stretched String Revisited	210
8.3.1	Distributed Systems	210
8.3.2	Concluding Remarks	212
9	The Special Theory of Relativity	213
9.1	Pre-History of concepts about light	213
9.2	Galilean Invariance	215
9.3	Implications of and for Maxwell's Equations	219
9.4	Pursuit of a special frame	223
9.5	Michelson-Morley Experiment	223
10	Kinematics of Special Relativity	229
10.1	Special Relativity	229
10.1.1	Principles of Relativity	229
10.2	Harry and Sally and Space Time Diagrams	232
10.2.1	Introduction	232
10.2.2	The Paradox of Harry and Sally	232
10.3	The Relativity of Simultaneity	233
10.3.1	Harry and Sally's Movements in a Diagram	234
11	The Nature of Space-Time	241
11.1	The Problem of Coordinates	241
11.2	The Lorentz Transformations	246
11.2.1	The Relatively Moving Clock	252
11.2.2	Derivation of the Lorentz Transformation	254
11.2.3	Details of the Derivation of the Lorentz Transformations	255
11.3	Using Lorentz Transformations	257
11.3.1	Time Dilation	257

11.3.2	Length contraction	259
11.3.3	The Doppler Effect	260
11.3.4	Addition of velocities	262
11.3.5	Time for Different Travelers	265
11.3.6	Visual Appearance of Rapidly Moving Objects	267
12	Events, Worldlines, Intervals	269
12.1	Introduction	269
12.2	Place and Path in the Two Dimensional Plane	269
12.3	Minkowski Space-time	280
12.3.1	Future, Past, and Elsewhere	281
12.4	Causality and Trajectories	287
12.5	The Hyperbolic Hangle*	289
12.5.1	The same result directly using calculus*	294
12.6	Four Vectors and Invariants*	295
12.7	Harry, Dorothy, and Sally Revisited	299
13	Paradoxes of Relativity	303
13.1	The Twin Paradox	303
13.1.1	The Problem	303
13.1.2	The Solution	303
13.2	The Boy in the Barn	304
13.2.1	The Problem	304
13.2.2	The Solution	304
13.3	The Bandits and the Bullet Train	306
13.3.1	The Problem	306
13.3.2	The Solution	307
13.3.3	Consequences*	308
14	Uniform Acceleration	309
14.1	Events at the same proper distance from some event	309
14.2	Uniformly accelerated motion	310
14.2.1	Details of the calculation of the acceleration	311
14.3	The proper time along the trajectory	314
14.3.1	Timelike Trajectories and Accelerated Motion	314
14.4	Examples using accelerated motion	315
14.4.1	Deceleration	315
14.4.2	Accelerated Rocket	317
14.4.3	John Bell's Problem	319
14.5	The Accelerated Reference Frame	322

14.6 Events at the Same Proper Time	330
15 Relativistic Dynamics	331
15.1 Relativistic Action	331
15.1.1 The Action for a Free Particle	331
15.2 Energy and momentum of a single free particle	334
15.3 Mass	336
15.4 Kinetic Energy of a Single Particle	337
15.5 Transformations of Momentum and Energy	338
15.6 The Energy, Momentum, and Mass of Light	339
15.7 Interactions	340
15.8 Multi-particle Systems	340
15.9 Rest energy of composite and elementary systems	340
15.10 Applications of Energy Momentum	340
15.10.1 Nuclear Reactions	340
15.10.2 A Relativistic Rocket	342
16 Introduction to General Relativity	345
16.1 The Problem	345
16.2 Free Fall Observers and the Equivalence Principle	346
16.3 The Equivalence Principle	348
16.3.1 The Monkey and the Hunter	349
16.4 Direct Effects from the Equivalence Principle	350
16.4.1 Universality and Eötvös–Dicke	350
16.4.2 Bending of Light Rays	352
16.4.3 Clocks and Accelerations in Towers	353
16.5 Intrinsic Effects of Gravity	355
16.5.1 Distortion of Elastic Bodies	355
16.5.2 Gravitation and Tidal Forces	357
17 Geometry and Gravitation	363
17.1 Introduction to Geometry	363
17.2 Gaussian Curvature*	365
17.3 Example of negative curvature: the Pringle*	366
17.4 Curvature and Geodesics	368
17.5 The Theorema Egregium and the Line Element	370
17.6 Geometry in Four or More Dimensions*	371
17.7 Coordinate Labels in General Relativity	372
17.8 Einstein Equations	372

18 Effects of Gravitation	373
18.1 Curvature around a Massive Body	373
18.2 The Universe	373
18.2.1 Background Ideas	373
18.2.2 Copernican Principle	376
18.2.3 Olber's Paradox	378
18.2.4 Hubble Expansion	380
18.2.5 The Age of the Universe	385
18.2.6 Models of Expanding Universes*	385
18.2.7 The Space Time Structure	400
18.2.8 Black Body Background	401
18.2.9 Problems with the Expanding Universe	401
18.2.10 The Cosmological Constant	401
18.2.11 The Standard Model of the Universe	401
19 Introduction to Quantum Theory	403
19.1 Introduction	403
19.2 Blackbody Radiation	404
19.2.1 Background on Thermal Processes	404
19.2.2 Basic Thermodynamics	408
19.2.3 -1^{st} Law: Equilibrium & Variables	409
19.2.4 0^{th} Law: Temperature Exists	411
19.2.5 Radiation in a Cavity	417
19.2.6 Attempts to describe the spectrum	421
19.2.7 Planck's Explanation of the Spectrum	422
19.3 Photo-Electric Effect	423
19.4 Young's Double Slit Experiment Revisited	426
19.5 Electrons and Young	430
19.6 Action and Quantum Mechanics	431
19.7 Constructing the Amplitude.	433
19.7.1 A Mathematical Aside – The Population Equation The Exponential Function	435
19.8 New Clocks or Even More on Phasors*	439
19.9 The Uncertainty Relations	442
19.9.1 The Uncertainty Principle and the Quantum Mechan- ical Harmonic Oscillator	443
19.9.2 Oscillator Ground State Wave Function	444
19.10 Localized States	445
19.10.1 An Aside on the Particle in the Box	445
19.10.2 Returning to the Oscillator*	447

19.11	Time Development of Quantum Systems*	449
19.11.1	Motion in Quantum Mechanics	449
19.11.2	Relation between the Quantum and the Classical Oscillator	451
19.11.3	Classical Motion of the Quantum Oscillator*	452
19.11.4	An Aside on the Binomial & Poisson Distribution	452
19.11.5	A Return to Classical Motion of the Quantum Oscillator*	457
20	Single Particle Quantum Mechanics	463
20.1	Single Particle Quantum Mechanics for Massive Particles	463
20.1.1	Introduction	463
20.1.2	Implications of Young's Double Slit with Electrons	464
20.1.3	The Schrödinger Probability Amplitude Field*	465
20.1.4	The Field as a Causal Agent	467
20.1.5	The Magic of Separable Field Configurations	467
20.1.6	Dimensional Considerations in Single Electron Quantum Mechanics	467
20.1.7	More than One Low Energy Electron	468
20.2	Examples	468
20.2.1	Free Particle	468
20.2.2	Particle in the box	468
21	EPR, Quantum Measurement and Bell's Theorem	471
21.1	A Two Level System	471
21.2	More on polarized light as a two level system	474
21.3	EPR and Bell's Theorem	477
21.3.1	What is a particle and what is the field ?	481
22	Quantum Field Theory	483
22.1	Introduction	483
22.2	The Many Photon State	484
22.3	The Quantum Stretched String	485
22.4	The field	486
22.5	Elementary Particles	486
22.6	Fundamental Processes	487
22.7	Feynman Diagrams	490
22.7.1	Introduction	490

23 Interface of Large Scale and Micro-physics	491
23.1 Structure in the Universe	491
23.2 The Inflationary Universe	491
23.3 String Theory	493
23.4 Quantum Gravity	493

Chapter 1

Introduction

1.1 Purpose of This Course

Today, it is apparent that you cannot function in society without contact with science. Not only are the things we use in our daily lives based on the discoveries of science, but also our attitudes and beliefs are derived from conceptual aspects of science. More importantly, the methods used for the acquisition of scientific knowledge are extraordinary and successful; a better system has not been developed. Every student at The University of Texas, in particular, every student in the Plan II program, must be exposed to the basic concepts and methodology of modern science. In addition, it is important that as a part of that exposure, you learn about the concepts of modern physics. Physics has been the most successful of the sciences, and its fundamental methods, based on experimental verification, reduction, and synthesis, has become a paradigm for all the other sciences.

This course is a part of a sequence of science courses that is required for all Plan II students. It concentrates on the conceptual foundations of modern physics. This course is different from any other that is taught at The University of Texas at Austin or anywhere else that I am aware of for two reasons. These concepts are very important but difficult to understand. The junior Plan II student has had enough other preparatory material and shown a maturity that makes it possible to discuss these issues. In addition, the students in the Plan II sequence have diverse majors and many will take or have taken a physics course at the university level. For these students, it is important that this course offer new ideas. Fortunately, all these other courses deal with the material at a more applied level and do not treat modern physics in the detail required to understand the basic conceptual ideas.

Most other physics courses spend almost all of their time developing the concepts of classical physics. This is because so much of our lives is effected by these ideas and concepts. We live in a world that is dominated by the objects that are dealt with in classical physics. Without the foundation of classical physics, it is impossible to understand the ideas of modern physics. In the case of the Plan II students, we are fortunate to be able to assume a reasonable level of understanding of the ideas of classical physics. It is anticipated that all students taking this course will have had an introductory physics course in high school or at The University of Texas at Austin.

Another feature of most university physics courses is that they serve as a foundation course for subsequent studies of a more specialized nature; therefore, these courses have to cover a certain content. That content is also predominantly based on classical physics. These courses also tend to be dominated by problem solving techniques and preparation for subsequent standardized tests such as the LSAT or MCAT. In this course, in contrast, the primary emphasis will be conceptual. Many of the concepts that will be discussed will not be used in your future work. One of the purposes of this course is to provide an opportunity to understand the basis for our current descriptions of matter and the universe. These ideas are often contrary to every day experience and thus require a level of understanding that is generally more abstract and subtle. The skills required for this type of reasoning are valuable in almost any context and are another reason that this course is offered.

We will treat only modern physics. Well, almost. There are some aspects of classical physics that are not treated adequately in most physics courses, but they are essential to the understanding of modern physics (optics, field theory, action, thermodynamics and symmetry). In addition, because this course will emphasize conceptual foundations, it will spend little time on the “things” of modern physics. These “things”, such as models of the atom, transistors, or lasers, are covered by most courses with a modern physics component and we will only deal with them as they provide examples for the development of the basic concepts of our interpretation of what we see in nature.

The goal of this course is to develop a sense of the processes that are inherent in the articulation of the discoveries of modern physics. Skepticism is inherent in an honest and objective search for truth; in the use of reason to establish a successful model of a subtle and difficult to understand phenomena; and in an analytic approach that reduces phenomena so as to discover its essential minimum. In some sense, the detailed calculations that are required in the homework and the tests are not important; instead, what is

important is the *process* by which they are derived. Rather than just having students rewrite or restate the principles, in this course, an understanding of the basic concepts is developed in the application of theoretical principles to diverse examples. I hope that the student who successfully completes this course will appreciate the value of scientific reasoning. The fact that the universe is knowable by these techniques, and that the successful enterprise of physics leads to a knowledge system that is more powerful than any other because it deals with the objective reality that we all share.

This is intended to be a terminal course in physics. There is no follow up course and the issues that are discussed here are at the limit of our current knowledge. This course will show you how our modern theories are developed and what is the basis for our belief in these theories. It will also show you how to deal with ideas that are outside your usual approach to understanding. I hope that you are willing to undertake the task and will enjoy the exercise.

1.2 Physics that you should know

1.2.1 Introduction

This course is intended for people with minimal formal exposure to physics. Basic ideas and relevant definitions will all be introduced as they are required. This does not mean that some background information or experience is not helpful. Prior exposure to physical reasoning and to physics vocabulary should make the material more accessible to you. Some concepts that we all use and discuss in our daily lives, such as energy, become more refined in the context of physics, and they will be treated this way in our course. More important than a physics background will be experience with consistent logical reasoning usually based on quantitative but simple techniques and curiosity about the world.

It is important that you have a quantitative understanding of the phenomena about us. When you discuss things that you see, you should use specific terms, such as density and speed, and you must understand what they mean. Whenever possible, you should discuss them in a quantitative manner. Basic mathematical concepts, such as area and volume, are essential. An extended discussion of the mathematical requirements is in Section 1.3.1 on page 11. You should familiarize yourself with all of the items on the “Things that Everyone Should Know” list, Section 1.4.2 on page 16. Simple exercises that we all do, like computing a route for a trip or estimating the cost of a vacation trip, are important skills; they are quanti-

tative and require complex reasoning. Many of these skills will be essential for Fermi problems, Section 1.4.1 on page 16, and are an important part of this course.

The following is a brief outline some of the important ideas from basic physics and skills that any student in this course should know. It is rather terse and, in some places, abstract. You may have to read it carefully to recognize that it is something that you already know but may know in a different way of expressing it. If you do not know them, and you feel that you will have difficulty, you should work with someone to develop a basic understanding at the level described here.

1.2.2 Kinematics

Kinematics is the study of the relationships between the quantities that are involved in the study of motion. You should realize that to describe a place you need to select an origin or reference point from which to measure displacements. Displacements are the separation between two places. Displacements are measured in lengths. We will discuss the issue of length, Section 2.2 on Page 34 and place, Section 10.1 on Page 229 in great detail later. Again, these are actually subtle issues and our attitude toward them has changed in modern physics.

You should know how to identify a place and suitably articulate it in a one, two or three dimensional space. You should realize that this descriptor of place is a vector quantity of suitable dimension and that as such it has both a magnitude and a direction. In general, the displacement vector can be thought of as the most accessible of the larger class of objects called vectors and the rules of vector algebra are those of common sense applied to displacements. The displacement in the three dimensional space that we find ourselves in can be stated as the triplet of numbers that are the magnitude of the displacements in the three basic coordinate directions or as a magnitude and a direction and notationally have the usual designation $\vec{r}$.

Vectors can be added to produce new vectors, and they have simple addition rules. There are two general rules: To find the sum of two vectors place the tail of the second vector on tip of first. The sum is the displacement produced by going from the tail of the first to the tip of the relocated second. Said another way, two displacements can be combined, and their result is also a displacement. This is a general property of all objects called vectors; there is a rule for addition and the addition of two vectors is also a vector.

The magnitude of a vector is its length. The magnitude of the displace-

ment is the distance (actually the shortest distance of the many possible distances that depend on the path) between the initial and final places. Note that distance is always a positive quantity whereas a displacement can be either positive or negative.

Velocity is the time rate of change of displacement. In this sense it is a difference of two displacements and thus is also a vector. The length of the velocity vector is the speed. Note that speeds are always positive. Velocities can be added using the same "tip to tail" addition rule that was used for displacement. Note that if you change the displacement in any way, you have a non-zero velocity. Even if you do not change the distance, but change just the direction of the displacement, you have a velocity.

Acceleration is the time rate of change of velocity. It is basically a difference in velocities and thus is also a vector. Accelerations can be added using the same "tip to tail" rule. If you change the velocity in any way, you have a non-zero acceleration. Even if you do not change the speed, but change just the direction, you have an acceleration. You should understand situations in which acceleration stays constant but velocity changes.

It should be obvious that if you know the position of an object for all times, you know the velocity and the acceleration for all times. You should also realize that if you know the acceleration for all times, the initial position, and the initial velocity, then you know the position for all subsequent times.

Any description of motion depends on a choice of reference frame from which all displacement, velocity, and acceleration measurements are made.

If any of the above descriptions make no sense to you, take the time to ask the Professor in office hour or one of the tutors in a tutoring session or even another student until you are comfortable with the descriptions above.

1.2.3 Dynamics

Dynamics is the study of the causes of motion. The motion is the temporal evolution of systems in space. Newtonian physics is based on the idea that space and time are absolute. In other words, time and space are unaffected by what is in it and how it moves. This idea will be expanded in the considerably in the course but it is from this basis from which we must start.

A primary notion is that there are forces. These forces represent the effect of other bodies on the body whose motion is under study. Your third-grade definition of a force—a push or a pull—is as good as any for a start. In this sense, forces are contact actions of one body on another. To do physics,

we need to expand this idea beyond contact forces to action at a distance influences, see Section 5.2 on Page 123.

To get a better understanding of forces, consider the world made up of several parts. This system of parts is isolated and thus all influences are from the parts on each other. This is the essence of reductionism, see Section 1.5.2 on Page 28: you can reduce the whole to its parts and the action of any part on a given part does not depend on the remaining other parts. The important point is that a force is the effect of one body on another and is only considered when you replace the body by its force, see Figure 1.1 on Page 7. We are interested in the motion of body one. We talk about the force of body two on body one and the force of body three on body one and so forth. Once we know the forces and use the fact that force in simple cases is a vector quantity¹ and obeys the usual rules for vector addition, we can get the total force by addition. In a real sense, bodies two and three etc. are replaced by their forces. Later in the semester, we will have to broaden our idea of force so that it becomes separated from the body that is its source, see Section 5.2.1 on page 125, and just talk about it as a thing unto itself. For now, all forces are due to other bodies and they have meaning only in the sense that they are there when we want to discuss the effect that one body has on the other.

1.2.4 An Important Aside on Causal Agents:

Force will be the first of what will be called ‘causal agents’. This is the construct that attributes the effect being manifest through agency and that these effects can be traced to some articulated source or agent and these effects are articulated for that agent as if the other agents were not present and that in some form these effects can be added and that that is also all that can be operative, i. e. $\vec{F}_{1,2}$ brings about changes in the state of body 1 and is due to presence of body 2 and this effect is independent of the other possible things that could effect 1 that if we enumerate all the agents and through some process ‘add’ their effects we will know the net effect on body 1.² The idea

¹More generally, the influence of one body on another may manifest itself as more than a simple directed quantity, a vector. In particular, gravitational interactions result in a shear, a stretch in one direction and a compression in the orthogonal direction, see Section 16.5 on page 355.

²You probably should reread this sentence several times to get its full import. The fact that this process seems to work for all systems and effects that we have discovered is the basis for our ability to articulate what is happening in all circumstances.

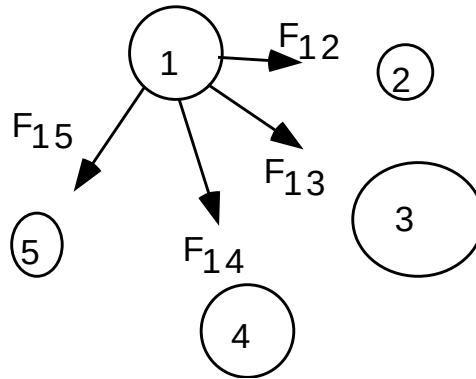


Figure 1.1: **Adding Forces** A system composed of 5 parts. The forces are there in the sense that F_{12} is the push or pull on body 1 due to body 2. F_{12} can depend only on the relationship between bodies 1 and 2 and F_{12} does not depend on the presence of the other bodies. Similarly F_{1i} is the effect of body i on body 1. Note also there is a set of forces that act on body 2 and so forth.

of causal agent and the free agent independence and addition description of their operation will recur in properties of light, Section 4 on page 83; field theory, Section 5.2.2 on page 126; action, Section 6.2 on page 166; gravity, Section 16.5 on page 355 and quantum phenomena, Chapter 19 on page 403. The general term for causal agents that add separately to produce a net effect is called superposition.

The rest of basic dynamics is contained in what are generally called Newton's three laws of motion. The first law states that if a body has no net forces acting on it, it will continue in its present state of motion. This means that the velocity of an unforced body is unchanged; there is no acceleration. Newton took this idea from Galileo. We will look at this law from a different perspective and, in fact, closer to the original spirit of Galileo. An object at rest and subjected to no forces remains at rest. An object with a velocity, $\vec{v}$, subjected to no force will continue to move at a velocity $\vec{v}$. In a sense, there is no difference for an object at rest and an object with a uniform velocity. This is called Galilean invariance and will play a very important role in what we do in this course. This law can be stated in many forms and each way provides new insight into its meaning. One of the more intuitive is that, for any body that is subject to no net forces, there exists a reference frame in which the body is and remains at

rest, see Sections 7.4.2 on page 197, and Section 9.2 on page 215. Since by reference frame, we mean an unforced observer, an observer that also notes no forces, there may appear to be some circularity in this definition. The important observation is the observers that detect no forces are those that are in uniform motion. Another way of interpreting this result is to say that all force-free motions have constant velocity and that uniform motion, motion with constant velocity, is the same experimentally as no motion. This was a long way around to the statement that all uniformly moving coordinate systems are equivalent and that it is meaningless to say how fast you are going in any absolute sense. You can measure accelerations absolutely but you cannot measure velocities except as relative concepts.

In order to present the second law, we need the concept of mass. For our present purposes, we can take the simple definition of mass: it represents the amount of matter in an object. We will spend considerable time in this course clarifying the idea of mass; it was a difficult concept for Newton and the modern interpretations are also subtle. In its simplest form, Newton's second law states that a body responds to the presence of an unbalanced force by accelerating. The acceleration is the net force divided by the mass of the body, the famous $\vec{F} = m\vec{a}$. It is important to note that acceleration is a kinematic quantity and is defined once we have a length and a time. It is important to note that this is not the definition of force. In all applications, $\text{vec}F$ is a relationship descriptor and $m\vec{a}$ is a response descriptor of the object under consideration.

Newton's third law states that if two bodies exert forces on one another, these forces are equal and opposite. The force of body two on body one is equal to the negative of the force of body one on body two, $\vec{F}_{21} = -\vec{F}_{12}$. This law is also known as the law of action reaction. The intuitive rationale for the truth of this law is that the two interacting bodies are combined into a single composite, that body should have no residual force; the internal forces must cancel or the composite would accelerate. As we develop more complex systems and more complex descriptors, we will come to circumstances in which there will be a need to abandon this relationship. In particular, there are things other than discrete objects, in particular fields, see Section 4.8 on page 121.

It is very important to realize that if you know the forces acting on a body, either as a function of position or time, and you know the initial position and velocity, then you know the subsequent motion, i. e. the position as a function of time, see Section 1.2.2 on page 4. This is the essence of causality. Given the initial position and velocity, and knowing the forces between all the bodies determines all the subsequent behavior of the bodies.

We will find that there is more to the world than just localizable point objects and that our requirements for causality have to increase to account for all the phenomena observed in the universe, see Section 5.2 on page 123.

You should know several simple examples of forces. There are two types: basic and phenomenological. Basic forces are those that we attribute to the fundamental aspects of matter, such as electric force between charged particles and gravitation between massive bodies. Phenomenological forces are due to very complex involvements of many things but, despite the complications, are simple to describe.

An example of a phenomenological force is the normal force that stops my hand from moving through the table when I lean on it. In this case, the atoms of my hand and the atoms of the table act to produce whatever force is necessary so that my body is supported. Another example is the Hook's Law spring. Here, a complicated structure of coiled metal, when exposed to a force is deformed. If the force is proportional to the stretch of the spring, $\vec{F} = -k\vec{x}$, this is a Hook's Law spring. Many coil springs and lots of other things act like a Hook's Law spring and this is a very useful concept. You should understand the motion of a system that is well described by a Hook's Law spring.

There are four basic forces: strong, weak, electromagnetic, and gravitational. You should know about these forces, along with the simplest forms of the two classical forces: the electrical force between two charges, Q_1 , and Q_2 at locations $\vec{r}_1$, and $\vec{r}_2$; $\vec{F}_{12} = \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{|\vec{r}_2 - \vec{r}_1|^3} \times (\vec{r}_2 - \vec{r}_1)$ and the gravitational force between two masses, m_1 , and m_2 at locations $\vec{r}_1$, and $\vec{r}_2$; $\vec{F}_{12} = -G \frac{m_1 m_2}{|\vec{r}_2 - \vec{r}_1|^3} \times (\vec{r}_2 - \vec{r}_1)$. Here $\frac{1}{4\pi\epsilon_0}$ or more succinctly, ϵ_0 ³ and G are fundamental constants of nature. That means that we have no explanation for why they take on the values that they have and assume, particularly in the case of G , that we probably never will. As we will see shortly, Section 2.6 on page 48, the values of the fundamental constants determine the size of things.

From forces and kinematics almost all of physics can be developed. Certain derived concepts are so important that they take on a fundamental nature. For example: work done by a force which is the force times the distance through which the force acts and kinetic energy which is the en-

³At one time ϵ_0 and the corresponding constant from the magnetic force law called μ_0 were thought to be fundamental but, with the development of the Maxwell's Theory of Electromagnetism, see Section 5.4 on page 140 especially the material around Equation 5.34 on page 150 and the Special Theory of Relativity, see Chapter 9 on page 213, they are combined to form a velocity, $c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}$, the speed of light, which is now the fundamental constant and the values of ϵ_0 and μ_0 are defined.

ergy of motion, and which, for slow moving particles, is $\frac{m\vec{v}^2}{2}$ where $\vec{v}$ is the velocity. For special cases, there is also an energy of position called the potential energy. For instance, for places not too high above the earth, the potential energy for an object of mass, m , at a height h is mgh where g is the acceleration of objects released from places not too high above the earth.

There are two types of momentum: linear which is usually $m\vec{v}$ and rotational or angular which is usually $mr\omega$ where r is the distance from the axis of rotation and ω is the angular speed around that axis. The direction of the angular momentum is along the axis of rotation.

You should be aware of the famous conservation laws, such as conservation of energy and momentum. There are two forms for the law of conservation of energy—the equivalence of work and the total energy (both kinetic and potential energy). There is also a related energy conservation law that comes from thermodynamics, the study of heat processes. In this law, energy is not only mechanical energy, it is also thermal energy and involves concepts like temperature. In this course, we will find a more general definition of energy and momentum, see Section 7.1 on page 183.

1.3 The Role of Mathematics

To most people, there is a big difference between mathematics and physics. This is not the case and, until very recently, all the mathematics that existed had been developed in response to a need for a language that could describe a physical phenomena. The mathematics was not developed and at hand for use. In most cases, the physicist or physicist/mathematician had a problem and invented new mathematics that was needed to provide an appropriate description of the physical system under study; Newton invented the calculus to have a language to describe objects that changed their position; Dirac invented the delta function to describe phenomena in quantum mechanics and this led to distribution theory.

One of the most important points in this course is to clarify the relationship between mathematics and physics. Mathematics is a carefully articulated set of rules for the manipulation of carefully defined objects. The objects and the manipulations are constructed to have the aspects of interest to the problem at hand. Almost all the mathematics taught at the university level was invented to analyze a physics problem. It is only in the past century that the elements of the mathematics have become rich enough that mathematicians have been able to develop systems that do not have a

counterpart in physics. Even in these cases, it is possible that these “mathematical” systems may find a surprising and new application in physics. In this regard, mathematics is a tool for the analysis of phenomena. A very powerful tool since it has had all of its logical elements carefully vetted so that all the manipulations are consistent. It is also an intuitive tool that you could develop if you think about it. For us, mathematics is a language that, because its algorithms are precise and logically consistent, enables anyone to completely understand what is said. It is the objectification of the thinking process. Mathematics is the process for reducing our thought processes to an algorithm. Mathematics is not a substitute for thinking but it provides a framework in which details of the thinking process are codified.

1.3.1 Mathematics and Symbols That You Should Know

Mathematics is both a language for the description of phenomena and a tool for analysis. Both of these aspects are important. Mathematical terms, such as “radian”, “linearity”, “variable”, and “sum” should all be well understood. Techniques of analysis, such as analytic geometry and algebra, are invaluable in the analysis of complex situations. In addition to understanding and using the vocabulary of mathematics, you must also understand concise notation. In the following sections I will detail the essential mathematical skills that will be required for this course.

Number Skills

. It may seem trivial but many people do not have a sense of quantity. This is often traced to an inability to appreciate the order of magnitude of a number, see Section 1.4.3 on page 19 for further comments. A great way to assess magnitude is by using scientific notation; three million is 3×10^6 . Of course, to use scientific notation, you need to understand the use of exponents, $x^a \times x^b = x^{a+b}$. Using these rules, you can perform algebraic and numeric manipulations with large and small quantities.

Regardless of the ease of manipulation, it is important to realize that an increase by a factor of 10 appears in the exponent as an addition of 1. This is a big change. What if you were suddenly ten times taller? Are there people that are ten times taller than you? It is in this sense that people are the order of 10^0 meters tall, see Section 1.4.3 on page 19.

Scientific notation also allows you to discuss the precision of a quantity. In other words, the notation allows you to report the size and the units but also how precisely determined the value is. In Section 1.4.2 on page 16, there

is a list of things that people should know. These are expressed in scientific notation. Most of the items on that list are measured quantities and as such have a certain precision, see Sections 1.4.3 on page 19, and 2.3.2 on page 42. The general rule used in scientific notation is that the precision is the range of values obtained by increasing and decreasing the last digit by one unit. In addition, the number has a certain accuracy. The number is accurate if the “real” value is within the precision. For example, we say the the radius of the earth is 6.4×10^3 km. The precision of this value is at the level of the second digit. By writing it this way we are indicating that we expect that the “real” value is between 6.3×10^3 km and 6.5×10^3 km. The value 6.4×10^3 km is accurate if the “real” value is somewhere in the range of the expressed precision. An interesting example in understanding scientific notation and precision and accuracy is the value of g , the gravitational acceleration, on the list. There is some ambiguity. Does the 10 indicate the power of ten or the front digits. If it is the front digits, it is accurate in the sense that the precision implies that the real value of g is between $11 \frac{\text{m}}{\text{sec}^2}$ and $9 \frac{\text{m}}{\text{sec}^2}$. This is what is meant by the way that it is written. It is not $1 \times 10 \frac{\text{m}}{\text{sec}^2}$. Maybe it would have been better to write it as $10 \times 10^0 \frac{\text{m}}{\text{sec}^2}$ but that would be overkill. It is also important to note that it is $10 \frac{\text{m}}{\text{sec}^2}$. It is also $32 \frac{\text{ft}}{\text{sec}^2}$. Physics quantities come with dimensions and the numeric value depends on the units used. The statement that g is 10 is just plain wrong.

There may be several reasons why you do not give the “exact” real value. A very important one may be that it, like most physics quantities, is a measured quantity and thus there is an intrinsic limit to how well it can be known. When you look up the values of quantities in a good text book they will generally give the value with much more precision than I have shown. In our example of the earth’s radius, you will find numbers like 6.371×10^3 km which is the value given in the text for Phy 302K [Giancoli]. I am not sure what motivates the author to select that level of precision. It is too much to remember and is more than is necessary for most purposes. If you look it up in a tables book, it will be measured to a much higher precision. In the CRC, a popular table book for physical constants, it is given as 6.378245×10^3 km. Note though that this is defined as the “mean equatorial radius” of the earth because at this level of precision we have to be very precise about what we are discussing. The distance from the center of the earth to the edge varies by more than this at different places. This is another reason that you may limit the precision of a value: variations in the thing you are measuring. The earth’s radius is a good example. The earth is not a perfect sphere. The earth is an oblate spheroid with small

lumps and the north south radius and the mean equatorial radius differ by approximately 21 km. Even if you discuss the equatorial radius there is a 20 km variation due to mountains and valleys. That is why the table book calls it the “mean” equatorial radius.

How much precision is appropriate? I take a very pragmatic view on this subject. You should only use the precision that you need for the problem at hand, and usually you do not need much. In this age of hand calculators, there is a tendency to use the precision of the calculator. I do not own a calculator and feel strongly that the precision should be set by the problem and not the calculating instrument. Going back to the value of the gravitational acceleration on the list of things that everyone should know, Sec 1.4.2 on page 16, you will notice that I list the acceleration of gravity as $10 \frac{\text{m}}{\text{s}^2}$ and not as the famous $9.8 \frac{\text{m}}{\text{s}^2}$. These values differ by 2 parts in 100. In our day to day observations, we are not measuring lengths and times to that precision. So why insist that the acceleration of gravity be to such a high precision? In fact, for the purposes of this course, you will be able (in most cases) to work to a precision of one significant figure. Sometimes less.

Among your number sense skills you should also have some feel for how probabilities operate. If A and B are independent events and A has a probability of occurrence of $P(A)$ and B has the probability of occurrence of $P(B)$, then the probability of occurrence of both A **and** B occurring is $P(A \cup B) = P(A) \times P(B)$. The probability of A **or** B occurring is $P(A \cap B) = P(A) + P(B)$. The notation of $\cup$ for **and** and $\cap$ for **or** as $\cap$ for and and **or** for or is from basic set theory. Consider the set of all independent events. Consider the toss of a die, any toss can be considered an independent event and every toss has an outcome that is a 1 or a 2, or a 3, or a 4, or a 5, or a 6. For a single toss of an unloaded die, these are all equally likely, $\frac{1}{6}$. There is a subset that is the events that are in A. For example, our set A can be the events that are the two highest values, a 5 or a 6. The $P(A) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. We can also construct spaces that are multiple tosses. For example, two independent tosses. This is a much larger space of events, 36 independent events to be exact. The set of all events with each of the two tosses that have the highest values are for members; 55, 56, 65, 66 or 4 events. Thus the probability of this happening is $\frac{4}{36} = \frac{1}{9}$. It is also useful to develop the concepts of conditional probabilities; $P(A|B)$ the probability of A occurring given that B has occurred. These are handled with Bayes’ Theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}. \quad (1.1)$$

Later in this course, we will encounter quantum mechanical concepts which

are very dependent on probabilistic interpretations and we will have to develop some of these ideas much more carefully, see Chapter 19 ‘Introduction to Quantum Theory’ on page 403.

Number sense manifests itself most significantly in our Fermi problems, see Section 1.4.1 on page 16. In most of these you will be working at a very low level of precision: for example at the 30% level. When that is the case, you can forget about small effects below that level. For example, suppose you want to estimate the total biomass on the earth. You do not have the information that would allow you to make a better estimate than possibly 50% precision. At that level of precision, you do not need to worry about the mass tied up in mammals; it is negligible. A useful assistance for Fermi problems is the book “Innumeracy” by John Allan Paulos [Paulos 1988]. I recommend it very highly.

Algebra, Trigonometry, and Analytic Geometry

These are all subjects that you should have studied in high school. You should be able to use ideas such as linearity, the relationship between solvability and number of equations and unknowns, and the role of redundant solutions. In this course, you should expect to encounter situations with simple polynomial equations and linear systems of up to three unknowns. I will spend some time developing the properties of the exponential function, see Section 19.7.1 on page 435, and its inverse (the logarithm), but these should not be totally new to you.

The trigonometric definitions and relations will all be required. You should be prepared to encounter situations that deal with the simpler identities, simple angle addition, and very simple trigonometric equations. I will always go slowly in these places.

From analytic geometry, you should be able to analyze problems graphically and recognize the shapes of the conic sections: parabola, hyperbola, and ellipse. You should also be able to do the opposite – identify the shape from the equation. You should be able to translate and rotate the simple forms discussed above and solve systems of simultaneous equations.

Calculus

All of you have had some introduction to the basic ideas of calculus. You will not be expected to use any calculus, but you must understand the concept of the derivative and its inverse, the integral. Although you will not have to perform significant manipulations using calculus, you must be

able to recognize its importance in some of the manipulations that I will perform. In addition, you will be asked to approximate calculus procedures, such as the computation of a slope or an integral, and you should realize what the approximation means. I will use the symbols of calculus, $\sum_{paths}$ for sum over paths, or $\int_a^b$ to summarize an argument. In addition, I will use the shorthand of Δ for difference or change. You will see the symbol for differentiation as $\frac{d}{dx}$ or $\frac{\delta}{\delta x}$ and should interpret it the change in something that is produced by a small change in x . Also I will introduce some very sophisticated symbols for some of the manipulations of fields. This is due to the fact that fields generally depend on several variables, see Section 5.2.2 on page 126. In these cases, the symbols $\frac{\partial}{\partial x}$ means the change in something for a change in x with the other variables held constant. Similarly, $\frac{\partial}{\partial t}$ is the change in something for a change in time with the other variables held constant. These terms will be carefully described in words. Again, these are a shorthand for a numeric computational process, and you should not allow terror to replace reason.

In this course, I will use the language of mathematics where it is appropriate to state relationships. In some cases, this will be a rather sophisticated use of the concepts of Mathematics. This is the most concise and careful way to say things. There will be many algebraic manipulations and a great deal of quantitative manipulation – sorry. I cannot cover the material in any other way.

Spirit of the Mathematics

One of the primary goals of this course is to convince you that, regardless of your previous mathematical background, you can do some quantitative analysis of a problem, any problem. Often, this will mean using rather crude analytic tools, such as rectified paths what in ‘mathematical’ language are called line integrals, but some analysis is better than none. By the end of this course, I hope that you will feel comfortable working problems until you find a satisfactory answer. I will push you until you overcome the stage in which you feel that you cannot get an answer because you lack some analytic skill. Don’t say that you cannot understand something until you learn some esoteric mathematical skill. All things are understandable without advanced mathematics or, at least, I will try to convince you that is the case.

1.4 First Day Handout

The first class day handout lists the general policies and grading procedures for the course. The following items are special aspects of this course that merit special comment:

1.4.1 Fermi Problems

These are all simple reasoning problems that are generally solved by making some very basic and plausible assumptions based on your experience or by using simple facts that you already know, see Section 1.4.2 on page 16. You actually know more than you think you do, and you can apply this information in many interesting circumstances. These problems also point out the value of having a quantitative perspective and often deal only with order of magnitude estimates, see Section 1.4.3 on page 19.

Fermi problems are named after the famous Italian-American physicist, Enrico Fermi, who was well-known for setting and solving them. For example, he taught at the University of Chicago and he would ask his class to estimate the number of piano tuners in Chicago. Fermi was associated with the Manhattan project, and there are several stories about him and order of magnitude estimates. In one, as he is being escorted about the laboratory in a jeep on the dusty roads of New Mexico, he asks the driver how thick a layer of dust could accumulate on a car window before falling off. Knowing the strength of chemical bonds and the size of atoms etc., he could quickly calculate the adhesion and check his result with the amount of the dust on the windshield. The most famous story is how he estimated the strength of the blast at the first atom bomb test, the Trinity Test, by releasing a sheet of paper as the shock wave passed by him and noting its deflection of the paper.

1.4.2 Things Everyone Should Know

In order to develop your quantitative perspective you have to know some things. Many of these things you can know just by looking around; some have to be put together from other facts. In any case, the world is a knowable place, and you already have many of the instruments you need to know it. The sights that you see, the sounds that you hear, and tactile feel of the world around us, supplemented with simple devices, can all be understood and fit into a pattern that allows for all of us to lead a fuller and more meaningful life. Achieving this requires the willingness to approach the

world in a quantitative fashion, along with the willingness to probe the world with simple experimental questions. Although in many cases we can reason out the magnitude of some of these facts, such as the radius of the earth, others, like Avogadro's number, just have to be remembered. The following is a list of things that I think you should know, and you will be expected to know them.

**Some Things That Everyone Should Know
Order of Magnitude**

gravitational acceleration	10 m/s^2
densities of solids and liquids	$10^3 \frac{\text{kg}}{\text{m}^3}$
density of air at sea level	1 kg/m^3
length of day	10^5 s/day
length of the year	$\pi \times 10^7 \text{ s/year}$
earth's radius	$6 \times 10^3 \text{ km}$
angle of width of finger at arm's length	1^0 or $\frac{\pi}{180} \approx \frac{1}{60}$
thickness of paper	0.1 mm
mass of a paper clip	0.5 gm
heat output per person	70 W
speed of walking	$2 \frac{\text{foot}}{\text{sec}} \approx 2 \frac{\text{km}}{\text{hr}}$
highest mountain, deepest ocean	10 km
earth moon separation	$4 \times 10^5 \text{ km}$
earth sun separation	$\frac{3}{2} \times 10^8 \text{ km}$
atmospheric pressure	weight of 1 kg/cm^2 or a 10 m column of water

Avogadro's number	$6 \times 10^{23} \frac{\text{atoms}}{\text{gm mol}}$
$\hbar$ or $\frac{\text{Planck's constant}}{2\pi}$	$1 \times 10^{-34} \text{ J s}$ or $\frac{2}{3} \times 10^{-21} \text{ MeV s}$
atomic diameter	10^{-10} m
nuclear diameter	10^{-15} m
atomic masses	$\frac{3}{2} \times 10^{-27} - 4 \times 10^{-25} \text{ kg}$
energy conversion . $1 \text{ eV} \approx \frac{3}{2} \times 10^{-19} \text{ J}$ (Not an accident. See charge of the electron.)	
energy content of a chemical bond	$2 - 5 \text{ eV}$
energy content of temperature, Boltzman's Constant .	$10^{-4} \frac{\text{eV}}{\text{°K}} \approx 10^{-23} \frac{\text{J}}{\text{°K}}$
energy content of food	$1 \text{ Cal} = 10^3 \text{ cal}$ and $1 \text{ cal} \approx 4 \text{ J}$
charge of the electron	$\frac{3}{2} \times 10^{-19} \text{ C}$
electron mass	10^{-30} kg
ratio of the electron and proton masses	$1/2000$
speed of light	$3 \times 10^8 \frac{\text{m}}{\text{s}}$
speed of sound	$\frac{10^3}{3} \frac{\text{m}}{\text{s}}$
frequency of sound	$2 \times 10^3 \text{ Hz}$
wavelength of light	$6 \times 10^{-7} \text{ m}$
population of the US	$\pi \times 10^8 \text{ people}$
population of Austin	$1 \times 10^6 \text{ people}$
π^2	10

Law of Seventy	$\ln 2 \approx 0.7 \approx \frac{2}{3}$
e^x	$\approx 1 + x + \frac{x^2}{2}$ for $x \ll 1$
e	$\frac{11}{4}$
e^{-1}	$\frac{1}{3}$
$(1+x)^n$	$\approx 1 + n x$ for $x \ll 1$
$\sin x$	$\approx x - \frac{x^3}{3!}$ for $x \ll 1$
$\cos x$	$\approx 1 - \frac{x^2}{2!}$ for $x \ll 1$
$\sin(\alpha + \beta)$	$= \sin \alpha \cos \beta + \cos \alpha \sin \beta$
$\cos(\alpha + \beta)$	$= \cos \alpha \cos \beta - \sin \alpha \sin \beta$
$\ln(1+x)$	$\approx x - \frac{1}{2}x^2$ for $x \ll 1$

1.4.3 Order of Magnitude Estimates

As you can see from the above list, for most purposes it is only important to know “about” how big something is. Often because to the crudeness of the measurement, you can only know something to within an order of magnitude. In most cases, this means to within a power of ten. In some cases, you can know it a little better say to within a percentage, say 10%. The range into which you can know a certain value is the precision. It is important to realize that all measurements and estimates have both an accuracy and a precision.

In a reasonable estimate or measurement, the true value should be within the range set by the precision. In other words, when you state a value, you are really giving its value and range of correctness for that value. By consensus, an easy way to express the range of precision is the range allowed by letting the last non-zero digit in the number increase and decrease by one unit. For example, you think that the population of the US is 330,000 people. Absurd, but this is an example. By saying 330,000 you are implying that the population is between 320,000 and 340,00. The actual value falls

well outside this range and, thus, this is an accuracy problem. These issues, along with the use of precision, are discussed in Section 1.3.1 on page 11.

The issue here is that often it is valuable to know something within a very large range called order of magnitude. In this case there are no digits. This is usually appropriate because the range of phenomena is so large. For instance in the “Things” list above the energy content of temperature is given as $10^{-4} \frac{\text{eV}}{\text{°K}} \approx 10^{-23} \frac{\text{J}}{\text{°K}}$. In fact, these numbers are known to a very high precision but that is not relevant to most usual quotidian applications. For instance, is something hot enough for the chemical bonds to break?

It is just a simple fact that our useful knowledge of quantitative value is generally order of magnitude. Better precision is often not available. For our analytic purposes, we will generally work to the level of one digit and a power of ten, a little better than straight order of magnitude. My own operating level of precision is a bit looser than this since I restrict myself to the numbers 1, 2, π , 5 and seven. By the way, I also do not own a calculator, see Equation 2.5 on page 46.

1.4.4 Home Experiments

In each homework assignment, there will be a home experiment. These are simple exercises that require basic materials that will be easily available or provided to you. There are two reasons why we perform home experiments. One: physics is an experimental science. The only route to knowledge is through experiment and, no matter how wonderful your reasoning, if it disagrees with experiment, then the reasoning has to be replaced. In a course like this one, there is a tendency to see all progress as theoretical when in fact it is the other way around. The problem is that most of the experiments that treat the basic concepts of modern physics are not accessible using a simple home apparatus. Therefore, many of the home experiments are not directly related to the course content. We still set them because they keep us aware of the experimental basis of our knowledge. The second reason is that the world is a knowable place, and it is only by manipulation that we can really test our understanding. This is an important point that differentiates physics from other subjects. In physics, you do not just accept the picture of the world that you are given, you test it, and then you change the conditions to see if it behaves as predicted. This same idea is important to the whole study of physics. You must always seek other ways to test and verify each idea. This is the essence of scientific skepticism. “But if that is the case, then shouldn’t this happen?”

1.4.5 Review Syllabus

The syllabus is a general outline of what I would like to cover in this class. There are two different syllabi, one for the fall semester and one for the spring. In both first, I will cover the background of classical physics. This will take a few weeks. My approach to classical physics will be based on principles that will be new to all of you, but which are the techniques used by modern physicists. Next in the fall, we will cover the modern physics of large scale phenomena. This is the theory of space-time called “general relativity”. Before we can cover general relativity, you must have a solid understanding of the “special theory” of relativity. Finally, at the conclusion of general relativity, we will discuss some aspects of cosmology. Then, we will develop the modern theory of light. This is our introduction to quantum phenomena. We will use our study of quantum phenomena to develop our understanding of what things really are. In the spring semester, we will reverse the order of the two general modern physics topics.

As stated above, this syllabus outlines what I would like to cover. It is very aggressive, and in all likelihood we will settle for less. In the fall, it will be mostly large scale phenomena and, in the spring, microscopic physics. A great deal depends on how the class proceeds.

1.4.6 Text

These notes are the primary text for the course. You should read them carefully before lecture. Another text is “QED” by Richard Feynman [Feynman 1985]. This book is an discussion of microscopic physics, and in our case, it is an introduction to the study of light. Another text that treats these issues is Rae’s book, “Quantum Physics: Illusion or Reality” [Rae 1994]. This is the basis for our discussion of the nature of the material world on the small scale. The auxiliary material on relativity comes from the Space Time Traveler by Moore [Moore 1998].

In addition, the book “Innumeracy” by John Allan Paulos [Paulos 1988] is a great foundation for quantitative reasoning. You should read it immediately. It is an easy and enjoyable read. Readings from the other books will be set at the appropriate times. There will also be some specialized handouts during the semester.

1.5 What is Physics?

Physics is an incredible accomplishment. It sets the tone for all our understanding of all the phenomena of the world around us. Other bodies of knowledge generally aspire to the level of prediction that is required for a physics theory to be accepted. Very few, if any, achieve it. The development of our understanding of the physical world is the greatest accomplishment of mankind. Two thousand years from now, when they write of the significant accomplishments of the 20th century, they will record as the most significant events the discovery of the quantum mechanics and relativity. That is, of course, if they were not discovered earlier on another planet. The great wars of this century will only be noted in passing. Richard Nixon and the collapse of the Soviet Union will hardly merit a footnote.

It is important to realize how the process of physics works. The basic operating procedure of physics is easy to state. It was hard to discover and is often harder still to adhere to in many circumstances but it is the most successful approach to knowledge that has been developed. It is the careful observation of the world followed by the development of idealized objects that reproduce this behavior followed by the extension of these ideas until they either fail of their own accord (because of some intrinsic property) or they are found to no longer agree with experiment. When that happens, you search for a new construction that includes both the successful results of the first theory *and* extends to include the new range of phenomena all of which agrees with this new construction. The process is always under continuous development. In this sense, it is hard to envision an end to physics. The phenomena may become more remote from our day to day experience and use but there will always be new questions that emerge as our understanding grows. Also as our understanding of newly accessible phenomena clarifies, new technologies develop that provide new observations that test the existing paradigms that force us to extend the applications of physics or force us to question our current understanding.

Since this approach to the basis of physics may be new to you, it will be worthwhile to give some examples. One set of examples is obvious. The material of this course was selected to illustrate this point. It follows the development of the several modern theories of light. We do not go back to the ancient Arabic and Greek theories of light but begin with Fermat's Least Time Theory since it was the first to be based on the experimental methods articulated by Galileo. For an interesting readable account of the ideas that preceded Fermat and even the controversy that surrounded Fermat's ideas, see the book by Park [Park 1997]. Although developed in about 1660, this

idea of the least time of travel for rays of light is still the principle that governs modern lens design. The subsequent development of Fresnel was a consequence of new families of experiments, interference and diffraction, that could not be understood with Fermat's theory. Note that Fresnel not only had to develop an algorithm for computing optical phenomena associated with interference and diffraction, he had to develop a method that contained the results of Fermat in the correct limit. It is in this sense that a new theory supplements an old theory. It does not make it incorrect. As stated above, Fermat's least time is still used in modern complex lens design. You could also use all the machinery of Fresnel to do lens design but most of what you learn is more than you need to know to make a good lens. Similar statements can be made about how quantum mechanics extends classical mechanics and general relativity extends Newton's theory of gravitation. Actually in all these cases, since the new theory invariably encompasses a greater range of phenomena, it is better to view the older theory as a special case of the new theory.

In another example, we study the General Theory of Relativity which is the name for the modern theory of gravitation. We all know that Newton developed a very successful theory of gravity. As Einstein tried to develop a new theory of gravity consistent with his ideas from special relativity, it was among the most difficult problems of the new theory to replicate all the successes of Newton's theory. He also had to find new phenomena that Newton's theory did not get correct and again these were difficult to come up with. It was for this reason that for many years it was very hard to find confirming experiments for the General Theory. Thus, Einstein's theory does not replace Newton's but includes it as a limiting case. A similar argument can be made for the development of quantum field theory as a replacement for classical mechanics.

This is such an important subject that it bears repeating. Physicists invent idealized forms and endow them with properties that are seen in nature. The manipulation of the idealized forms leads to behavior that mimics those seen experimentally. Once a set of forms provide a complete set of descriptors for some class of phenomena further manipulation of the forms may lead to behaviors that have not yet been manifested experimentally. This is the essence of prediction from theory. If the forms fail to behave like the phenomena being modeled, the theory fails, [Popper 2002], and the idealized form is extended or in some extreme cases replaced until it produces behavior that is observed including all the behavior that had been described successfully before. In this sense, the great theoretical achievements such as Maxwell's equations, see Section 5.4 on page 140, are a catalogue of the

results of thousands of experimental observations described through a concise organization of idealized forms.

This same perspective on theory construction is also helpful in understanding the role of mathematics in physics. Historically, mathematicians have studied and extended the properties of the idealized forms that have emerged from the observations of nature. In more recent times, because of the richness of the ever growing set of forms, they have been able develop new forms and codify a greater range of phenomena even those needed for descriptions of natural phenomena. In some cases, physicists have found that a set of forms that was developed by mathematicians only for their abstract beauty have application in nature and, in some cases, physicists have had to ignore the constraints that the mathematics have imposed only to then open a new realm of mathematical investigation, a healthy give and take.

1.5.1 Range of Phenomena

A particular aspect of our understanding of the world around us is the huge range of phenomena encompassed. No other body of knowledge covers such a broad range. Current understanding of physical length scales ranging over forty orders of magnitude are generally understood. Similarly, time scales ranging through similar orders of magnitude have been successfully described.

The following are interesting accessible materials that indicate this aspect of modern physics descriptive success.

“Powers of Ten”

This is a film that starts by looking at a man on a blanket at a beach along Lake Michigan in Chicago. It then expands the point of view until it covers the known universe. Then it focuses in until it is looking at a scale so small that you can see the quarks in a proton.

When students view this film there are many reactions. There are 40 powers of ten between the largest scale phenomena and the small scale observations. This is truly a fantastic range. No other field of knowledge can come close to a basis of explanation with that range. Generally, we feel that we have some handle on all the observed physical phenomena that occur in this interval. The real ends of our knowledge come at the peripheries. On the large scale, gravity dominates and we require the use of the General Theory of Relativity. Currently, we have difficulty with the origins of

space-time or, in the vernacular, the origin of the universe and the union of gravity with microphysics. On the small scale, we have troubles with the basic constituents of matter. Both of these subjects are important and are the two fundamental themes of this course. The Theory of Relativity is our look into space-time and quantum mechanics has given us framework for describing the fundamental constituents of matter. The problem of combining quantum mechanics with the General Theory is among the most pressing problems of current physics.

In the film “Powers of Ten,” people are always impressed by the contrasting periods of activity and inactivity as the scale of length is changed. This pattern indicates the separation of phenomena with differing length scales. Atoms come in only one range of sizes, and the same holds true for galaxies. Stars and biological systems occur in a range of sizes. Sizes are set by the same laws of physics that govern the behaviors of the matter. In the case of atoms, it is the mass of the electron and Planck’s constant that determine size and behavior. We will have a great deal more to say about Planck’s constant in the course of this semester.

Plot of Masses and Lengths

See Insert

The attached insert is a scatter plot of lengths versus masses. In a scatter plot, you pick two independent variables, in this case length and mass, and for each element of the group under study put at point on the plot at the coordinates associated with that element. For instance, we could be studying GPA versus height. Then on the graph with the height as the ordinate and the GPA as the abscissa, each student with their GPA and height is represented as a point on the plot. These are called scatter plots because, if the variables are unrelated, they will not fall in a pattern on the plot. You would expect that the scatter plot of points of height versus GPA would be all over the allowed ranges of the variable. If it were the case that smaller people are smarter as measured by GPA, the correlation would be manifest in this kind of scatter plot. In the insert, we are scatter plotting the length of a thing against the mass.

The first feature of this scatter plot is that for our case the range of values for length and mass is extraordinary. This was also pointed out in the movie “Powers of Ten,” see Section 1.5.1 on page 24. For this reason, to fit all the phenomena on a single piece of paper, we use as the ordinate and abscissa the log of the length and the log of the mass. This is thus a

log log scatter plot. As stated earlier, no discipline can claim a quantitative understanding of phenomena in such a range, approximately forty powers of ten in each variable. The next issue is that we want a scatter plot of things of interest. We want a point on the plot for people. What is the mass for a people? There is a range of masses for adult males that can vary by about 40%. This is a small range of variation and is within any size point that can be drawn on the plot. Thus for mass we can chose the generic value of 60 kg and not worry about men or women or any other variation. What do you do for length? There are several candidate lengths for a person – height, ear lob length,... What do you pick? Generally, once you are talking about people all the other length choices scale as the height, i. e. the ratio of the index finger length to the height is a universal constant that is the same for all people. Thus we can chose the height or largest dimension to set a standard. These are examples of scaling laws which are discussed more in Section 2.5.2 on page 47. In this general sense, we can now ascribe a length scale to most objects and be reasonably consistent. On the huge range of scales that we are dealing with the variations within the category that we are interested in is negligible.

Expanding out from humans, we want to put on other biological systems. Again, it turns out that, on the range of scales we have here, the different biological systems such as bacteria and ducks have a definite mass and length range and can be represented reasonably well by points. Note that all biological systems fall within a small region in the center of the diagram. It is no accident that this stuff is central but notice that it is also a small part of the total plot. It is central because we got to pick the scale of length and mass, see Section 2.3 on page 38, and we are biological systems. It is a rather small region of the plot because biological systems are complicated and cannot be too small and still have all the parts that it takes to operate. They cannot be too large for it to be a coherent whole. There was a report some years ago of a mold in Wisconsin that was several kilometers across. Even here, we could debate whether this constituted one living system.

Also note, the patterns of phenomena on the plot. The points are not scattered on the plot but fall on a straight line. What is the implication of the straight line? A straight line on a log-log plot implies a power law

relationship between the variables⁴:

$$\log M = a \log L + b \Rightarrow M = 10^b L^a = c L^a \quad (1.2)$$

where a is the slope of the line and b is the intercept and $c \equiv 10^b$. In our case, a is three and this linear relationship is just another reflection on the item in ‘Things That Everyone Should Know’, Section 1.4.2 on page 16 that most solids and liquids have the same density of about $10^3 \frac{\text{kg}}{\text{m}^3}$. This is really no surprise. All biological systems have about the same density as water.

This feature of the straight line carries on as objects like battleships and pyramids are added except that we note that it is not the same line but one displaced a bit. Again we are seeing that all the heavy things also have about the same density. It is just slightly higher and really justifies our statement that all things have nearly the same density including the heavy things. What is the meaning of this near equality of all densities? The first thing to note is that although we had said that we would make a scatter plot of all things we really have not. We have put on this plot only things that are composed of atoms touching each other or what are called condensed matter systems as opposed to gases for example. We did not put on things like the atmosphere. You notice this deviation as you look to the top of the diagram. As we add the planets, they are still pretty much on the line but objects like galaxies and pulsars are far from the line and not even a point on the plot. We can roughly conclude that all objects made out of atoms that are touching are of comparable density. Well, we know that this is not strictly true. Solid lead has a density that is 10 times that of water. Since the atomic weight of lead and water are also in the ratio of about 10 to one, we conclude that the size of lead atom is about the same as oxygen, the dominant mass in water. In other words, all atoms are about the same size. This is a striking fact. A lead atom has 82 electrons and is still about the size of a hydrogen atom which has only one. In other words, the scaling law for the size of an atom with mass is not the usual one but is instead approximately M^0 . This is a consequence of the competition between the attraction of the Coulomb force and the Pauli exclusion principle. There is certainly a tremendous amount of information on this diagram. This will hopefully make more sense as the semester develops.

⁴Fortunately, the creator of this plot used the log base 10 instead of the usual $\ln$ base e which is the more usual in scientific computation, see Section 19.7.1 on page 435 since we tend to think in powers of 10. Maybe if we were more sophisticated in a daily parlance, we would think in powers of e . It would be hard to make that change. Another base that is commonly used is base 2; the basis for all modern machine computations. Another use of the base 2 system is manifest in the law of seventy on ‘Things’.

There are several other features of the plot that will make sense as the semester develops. Black holes are a consequence of the theory of general relativity and these forbid mass and length relations in a large part of the diagram. Protons and neutrons also obey the Pauli principle but since they have a different mass than the electron, objects made of them – nuclei and pulsars – have a different density line. Again, this will all become clear as the semester develops.

It is worthwhile to point out that the same clustering of phenomena that you saw in the film “Powers of Ten” is present here. This separation of phenomena into groupings is one of the great accidents of physics and a very fortunate one. When you are dealing with atoms, you do not need to worry about gravitation. The masses in atoms are small enough and the gravitational force small enough that you only have to consider the electromagnetic force. Also the velocities are small enough that you can neglect the effects of special relativity. In the nucleus again you can neglect gravity. In a galaxy you have to worry about gravitation but, since most of the matter is electrical neutral, you can neglect the electromagnetic effects. It seems unlikely that the great success of modern physics could have been achieved without the ability to categorize and separate the phenomena that we deal with.

1.5.2 Reductionism and General Principles

Much of the success of physics stems from the ability to reduce the whole into smaller parts, to understand the small parts, and to reconstruct the whole from our understanding of these smaller pieces. The recent developments of many other sciences, such as biology, can be attributed to the ability to use these reductionist techniques. Although there has been some speculation that, with our current theories of matter, we may be approaching the limit of this technique and that a final theory may be on the horizon, this is, at best, speculation. For now there is no reason to believe that the successes of reductionism have all been identified.

Another aspect of the reductionist argument that most people fail to understand is that the elements that constitute the whole do not have to act like the whole. The only requirement is that when the whole is reconstituted, it must behave as observed. For example, a building is made of bricks. A building is not like the brick it is made of, and we all seem to be able to accept that. Similarly, a brick is not like a building, even though bricks are the primary constituents of the building, and again we accept that. When discussing the elementary constituents of matter, the tendency is to require

that the constituents have properties like the whole. This is at the heart of the problem of understanding “wave particle duality” problems of modern quantum mechanics. There is no requirement that the elements that are at the basis of the reductionist process must be like the objects that they constitute. This prejudice is not only apparently not required of the theory, but it does not hold for the elementary constituents of matter.

Although most of the successes of physics are attributed to reductionism, the great global principles of physics have also played an important role. In fact, the idea of understanding is usually based on some underlying global principal such as the concept of a mechanical system. We understand something when we can reduce its operation to that of a mechanical system. This was the accepted approach during the later part of the 19th century. As you will learn in this course, we now have a different criteria. Today, we understand something when we can describe its operations from an action principle. Another example is the idea of force used in Section 6.1 on page 153. In that case, you realize that you can separate the effects from individual sources, see Figure 1.1 on page 7, and that these operate the same way independently of the presence of other sources. This separability is at the heart of reductionism. In addition, the level and the object that you add as independent sources are generally derived from a global principal. The identification of an objective reality and the articulation of causality are important global concepts that do not exist in all cultures. The assumption of their role in the material world is at the heart of modern physics.

1.6 A Joke

The National Association of Dairy Farmers wish to support the invention of an automatic milking machine. They decide to do this by commissioning three lines of investigation: one lead by a mathematician, one lead by an engineer and one lead by a physicist. Each of these will have one year to study the problem and then report their results.

After the year has passed they call in their experts and ask for the reports. The first respondent is the mathematician. He fills the screen with equations and at the end of his/her presentation states that “thus we have proven that a solution exists.” The second presenter is the engineer. He/she roles in a rather large kludge of a machine and states that “not only does it produce milk but it produces hamburger automatically.” The final



Figure 1.2: **Spherical Cow** When studying any complex system, it is always useful to first eliminate many of the peripheral issues. In other words, do the spherical cow problem first.

presenter is the physicist and after a discussion of the groups work he/she concludes that “they have a complete and operational machine that is appropriate for the spherically symmetric cow.”

1.7 Problems

1. How many people die in Austin each year? Indicate your reasoning and pick one of the following: 1×10^2 , $\pi \times 10^2$, 8×10^2 , 1×10^3 , $\pi \times 10^3$, 8×10^3 , 1×10^4 , $\pi \times 10^4$, 8×10^4 , 1×10^5 , $\pi \times 10^5$, 8×10^5 .
2. Estimate the order of magnitude of the mass of a speck of dust, a grain of salt, a mouse, an elephant, the water that is equivalent to 1 inch of water over 1 mi^2 of rainfall, a small hill, and Mount Everest.
3. How tall is Jester Dormitory? (Say, JesterWest) Find some way to measure the height to within an accuracy of 5%. If it is to 5%, do we

need to specify the tower of Jester?

4. Pick a tree. (Any tree! Well, any live deciduous tree with more than 10 leaves!) How many leaves are on it? Try to get a fairly accurate count by counting the number of leaves in some small volume that you can measure accurately and then roughly measuring the leaf-bearing volume of the tree.
5.
 - (a) What is the height of the National Debt in pennies stacked on top of each other.
 - (b) Suppose these pennies were distributed uniformly across the land area of the contiguous 48 states. What distance would separate each penny from its nearest neighbor?
 - (c) How many tons of copper would be required to make these pennies?
 - (d) Suppose they were distributed by dropping them from the sky. If you were standing outside, how many pennies would hit your head, on average?
 - (e) If you stuck your finger straight up, what is the probability that a penny would land on it? and stick?

Home Experiment 1. You were given two pieces of paper and a string. Using the string as your unit of length, measure the perimeter of the two pieces of paper and the distance between the two corners indicated as A and B in the figure. Do this with a precision of at least 10%. For each sheet of paper, take the ratio of the length of the perimeter to the distance between A and B. Discuss these results in terms of dimensional variables and scaling. Discuss the use of your string as a candidate for the international standard of length. Consider the area of the two pieces of paper. How does the area scale with perimeter?

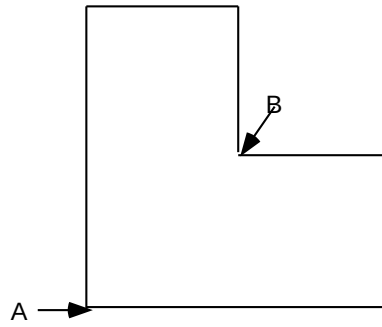


Figure 1.3: Figure of Home Experiment # 1

Chapter 2

Measurement

2.1 The Role of Measurement

At the very center of physics is the essential role of experiment. Even the most carefully crafted theoretical system can only be valid if it agrees with experiment. Experiment is the process of careful observation of the world around us. In the process of performing experiments, in some cases, it is possible to control some parts of the activity of observation but that is not the important part of experiment. Experiment is the drawing of coherent information from a situation. In a sense, the idea of theory construction is to develop a method which can consistently bring into a concise set of statements the results of all possible experiments on a given system.

In order to make consistent observations you have to make measurements. Measurement can be both qualitative and quantitative. Often times qualitative measurements can differentiate different ideas of how some process occurs. We will see this in the discussion of the foundations of quantum mechanics, see Chapter 19 on page 403. Most of the time, however, to differentiate competing ideas based on what we see, our observations have to be quantitative. Actually when you think about it, even most qualitative observations are really just very rough quantitative assessments. The reddening of the sky at sunset says a great deal about how the atmosphere works. What do we mean by red in this case. A definite range of values in the wavelength of the light and thus a quantitative assessment. In this sense, all of physics is based on the process of measurement—quantitative observation.

In its simplest form, measurement is basically the comparison of two related things. Whenever a certain circumstance is seen, the ‘caused’ sit-

uation emerges. On both ends of this observation, measurements must be made to know what was set up and what was the result. Therefore, we must understand measurement if we want to understand physics.

A process of measurement is basically a comparison of situations. This process is then formalized by using standards and comparing with these standards. This is best understood when we talk about length. The objects under consideration are separated. After some time, there appears to be a different separation. To quantify this set of events, we can find another separation that does not appear to be changing. An example of separated objects is two scratches on a rigid bar. Comparing the separations under consideration with the separation of the two scratches on the bar allows us to communicate the nature of the new separations. Of course, what is being measured is the length before and after. What most people miss in a discussion of length measurements is the fact that the process of measuring, the identification of a standard and a process is essentially the definition of length. The case of separation measurements is central to our study and available to our experience and that is the example that I will elaborate on throughout this course, but it is also true for all the other cases of measurement. A few other examples of things you can measure are temperature, hardness, intensity of earthquakes, and time. All measurements are of some attribute of a thing that satisfies some common general criteria and the rules for comparing the attribute to be measured and a standard of comparison with the same attribute.

Before going into more detail about these processes, it should be clear that there is a great deal of arbitrariness in this process of establishing the measurement protocol. Not only are there choices of comparison systems, called standards, but there is even an arbitrariness in establishing the processes. It should also be clear that the phenomena under study does not depend on these choices. This arbitrariness will have important ramifications, see Section 2.5 on page 44.

2.2 Measurability

As a first step in developing any system of measurement, we have to agree that the attribute in question is measurable. To be measurable, the attribute must satisfy an objective equivalence or reflexive relationship: if $A \geq B$ and if $B \geq C$ then $A \geq C$. The $\geq$ is an example of areflexive relationship. In other words, a reflexive relationship allows you to establish an ordered set of configurations for the attribute. Once you have an ordered set, you can

then map that ordering onto the real line. This is all an abstract way to say that then you can assign numeric values. You will often see us using this trick of mapping an ordered set onto the real line. This action of ordering and assigning a numeric scale is what is meant by setting a standard.

For example, again consider the case of length. If a place A is farther from some selected origin than a place B and another place C is closer than B , then A is farther than C . This is the reflexive part of the act of measuring. This kind of ordering does not work for things like beauty. There are actually two problems with measuring beauty. Firstly, the ordering of objects of art in terms of their beauty is generally not objective and, secondly, several different measures have to be brought together for an assessment. The different measures can lead to different orders. In other words, it is not clear that an objective ordering is possible. In our sense then, beauty cannot be measured.

Once you have established an ordering, you can place values of the measured thing on a numeric scale. That's where the values actually come from. You also have to remember that this mapping onto the real line assumes an underlying continuity that is often there but in some cases may not be. On the other hand, since the points on the real line are dense, if in your ordering, you left something out there is always room to stick something into a gap. All this amounts to saying that the ordering is important and the specific mapping onto the line is not. For many things any other mapping that preserves the ordering is as valid as the one that you are using. There are some ease of use criteria that make some choices better than others. An important one of these is additivity. For instance a distance that is twice as far is assigned a numeric value that is twice as large. In all the attributes quantified to date there has been some sense of combining systems to produce a larger measures. Measures such as distance that can be put on a scale that adds are called extrinsic. Length is extrinsic. Time is extrinsic. Density on the other hand is not. It is said to be intrinsic. If you take twice as much stuff at the same density you do not have twice the density.

The next step in establishing a measuring system is choosing the standards, see Section 2.3 on page 38. Before I do that though I have to emphasize that regardless of how you establish your standards there is an attribute with a property called measurability. It exists. For our example, length is the measured thing and there are many possible standards and systems but all of them are merely different articulations of the attribute that is length. In this case, we say that there is a dimensional content that is length.

On the other hand, It is also important to realize that all things with the dimensional content of a length are not a "length" in the sense of separation.

In some special circumstances, these quantities can turn out to be a separation but that does not have to be the case. An obvious example is an area. The square root of the area has the dimensional content of a length. This is in the sense that if the area was that of a square, the separation, a length in the fundamental sense, of the corners along an edge is the square root of the area. Another example, you can have the dimensional content of a length when you have a separation which is our prototypical “length”, constituted from a speed times a time, or a force times a time squared divided by a mass. In all these cases, there are circumstances in which although they may or may not represent a separation, they are a length. For example in the case of a velocity times a time which has the dimension of a length, this is a length when the velocity is that of an object and the time is a time of flight. For the rules governing the manipulation of dimensional quantities see Section 2.4 on page 43. Another important example of this type is idea of the gravitational acceleration g . g is the gravitational force per unit mass at a place in the vicinity of a massive body. It is not an acceleration. But the dimensional content is $\frac{\text{force}}{\text{mass}}$ which is the same as a $\frac{\text{length}}{\text{time}^2}$ which the dimensional content of an acceleration in the usual definition as the time rate of change of velocity. In certain circumstances, g is the value that the acceleration takes. For instance, when gravity is the only force acting on the body, g is the acceleration that the body will have. Certainly, if there are circumstances in which g is an acceleration, it must have the dimensional content of an acceleration.

This leads to the next issue. How many dimensional quantities are there? For historical reasons, length, time, and mass are taken to be the primary quantities and things like velocity, a $\frac{\text{length}}{\text{time}}$, are considered derivative. Are there more? As many as you like. To see that let’s look at an obvious example. Volume is a length^3 in the sense of the discussion above for area. You find it by multiplying three lengths. At the same time you could have an independent system for the measurement of volumes. For instance the gallon is a measure of volume. You could have a standard gallon and a protocol for measuring volumes based on this standard gallon. In this case, if you can find empirically that a certain number of cubic inches are contained in the standard gallon, $1 \text{ gallon} = 231 \text{ in}^3$. This would appear as a law of nature and could be called the Law of Volumes. Instead, we use ordinary geometry to conclude that this law is actually a result of our understanding of geometry. This example may seem a little forced but consider a slightly more subtle situation. Consider the case of the inertial and gravitational mass. This will be discussed in great detail later when we look at the problem of General Relativity in Chapter 16 on page 345 but for now we

need only know that there are two rather independent properties of mass. We all know that $\vec{F} = m\vec{a}$ and that the mass in this expression indicates how difficult it is to change the velocity of an object. This is called the inertial mass. You measure inertial mass in situations in which objects are accelerated. An alternative concept of mass is the mass that acts to generate the gravitational force. The attractive gravitational force, $\vec{F}_{1,2}$ of one mass, m_1 , on a second mass, m_2 , is $\vec{F}_{1,2} = G \frac{m_1 m_2}{r_{12}^2} \vec{r}_{12}$, where G is Newton's Gravitational Constant which has the value $6.7 \times 10^{-11} \frac{\text{m}^3}{\text{s}^2 \text{kg}}$ in the MKS system and $\vec{r}_{12}$ is the separation vector from body one to body two. This mass would be measured by placing two bodies at a known separation and measuring the force between them. Since these two ideas of mass are so completely different, it is difficult to conceive of why they are given the same name and treated identically. In a very real sense, there are two kinds of mass. We might want to differentiate by calling them by different names which for our discussion will be inertial and attractant. Stuff has so much attractant, a_t or inertial, i_n . You could measure i_n in a situation with a standard force and an acceleration according to $\vec{F} = i_n \vec{a}$. You would also likely define a measurement system for a_t based on the gravitational force but in the form $\vec{F}_{1,2} = \frac{a_{t1} a_{t2}}{r_{12}^2} \vec{r}_{12}$ without the use of an empirical constant such as G . In other words, you would say that two bodies of attractant one generated a force of one newton between themselves when placed one meter apart. Then by examining the motion of bodies under the influence of each others gravitational forces discover the empirical law that inertia and attractant are related by $a_t = \sqrt{G} i_n$. Of course, this is not how the subject was developed. Newton realized immediately that objects move under the influence of gravity in a fashion that is independent of their mass and that therefore gravitational and inertial mass are related by $m_{\text{inertial}} = m_{\text{gravitational}}$ and he never really discriminated between them. The lesson for us is that you can have an independent unit system for anything that can be measured.

On the other hand, it is the practice to consider mass, length, and time as special or primary. In this sense all the other measures are derivative of these three. What would have been empirical relations between measured quantities become definitions such as velocity is the change in separation for a change in time, or in more complex cases become expressed as a law of physics such as $\vec{F} = m\vec{a}$. Why only three and how did we get here? It is a result of the effort of physics to unify all phenomena into as few categories as possible. To classical physicists, these three, mass, length, and time, were the irreducible set from which all others could be constructed. We now

take a different perspective. There are two ways to look at the modern situation. We have found that as our understanding of nature has improved certain intrinsic quantities have been discovered. For instance, the Special Theory of Relativity has provided a special significance for c , the speed of light. Although it is the speed of light in a vacuum, it is more significant as a measure linking intervals of space and time, see Chapter 9 on page 213. This type of quantity, c , can be used to set a scale of units and these in turn can be used to set scales for length, mass, and time, see Section 2.6 on page 48. In one view you can say that these fundamental dimensional constants provide a basis for a system of measurement as discussed in Section 2.6 on page 48 or they can be viewed as the discovery of new physical law to reduce the number of primary dimensions. In this second view, you could now say that we are down to two and shortly may be reduced to one. Before we are in a position to look at this question closely, we will need to develop some of our technical skills for the manipulation of measured quantities in Section 2.5 on page 44.

2.3 Role of Standards

Once you decide that something is measurable, you have to pick a standard for comparison. A standard is something that has the property that you wish to measure. You arbitrarily select the standard and a protocol for using it. For example, for years the meter was the length between two scratches on a bar in Paris. You will obtain different values for the measured quantity depending on the standard of comparison. The distance between Austin and College Station is the same *regardless* of the standard, it is a length, but the *numeric* value depends on the standard or unit system used, miles or feet. There are several criteria for the choice of the standard. It should be convenient, stable, and accessible. Beyond these criteria, the choice can be rather arbitrary.

It is very important to again emphasize that **the standard along with the algorithm for comparison, is the definition of the thing that is being measured.** For example, the definition of “hardness” is determined as the quantity you get according to the algorithm stated for finding “hardness”. Algorithm “A” is established as the prescription to measure the quantity that will be called “hardness,” a specifically shaped diamond needle under a certain pressure moved across the surface of interest. The application of the algorithm to a certain material sets a standard reference that, in partnership with the algorithm, becomes the definition of hardness.

This process can be applied similarly to length and temperature, etc. The “unit” is the name of the particular standard being used. Lengths are in meters or feet; earthquakes are measured in Richters¹.

Since the choice of standard is arbitrary, nothing important can depend on it. The quantities can change but not what happens. This is our first case of a symmetry, a subject that we will discuss at length, see Section 7.1 on page 183. The symmetry under changes of standards, like all symmetries, leads to important consequences. The most important of these is the useful tool of analysis called “Dimensional Analysis,” Section 2.5 on page 44.

It is also important to reemphasize that although we can change the standard, there is still an intrinsic measured quantity; the distance between Austin and College Station is a length; that is its dimensional content of the measured quantity. When we measure the distance we use a specific unit, the mile. We can have lots of units and they are arbitrarily chosen, but we always have a distance whose dimensional content is length. It is useless to state the value of physical quantities without stating the standard that is used to measure them. Conversely, depending on the choice of standard, you can get any value for a quantity and so our sense of big or small. The distance between Austin and College Station is about 100 miles, a nominal distance on our scale. In a distance measure based on atomic diameters, the distance between the cities is huge.

In any measurement, there is also always an accompanying algorithm that establishes a method of comparison. An algorithm is a rule in which all the steps are defined and can be carried out by any person. For length, there is a standard length: the distance between two scratches on a platinum bar stored at the International Bureau of Standards in Paris. The method of comparison for length is to lay a length to be measured next to the standard to see if it is longer or shorter or what multiple or fraction the measured length is. Actually, what we really do is to establish a set of secondary standards that are even subdivisions, or multiples, of the original. This secondary standard is the same in the places where it can be compared directly and then applied in the other domains where the new standard

¹There is a rather annoying habit among physicists to name things after dead physicists. This practice is most apparent with units of measurement. For example the standard system of units, the MKS or **M**eter, **K**ilogram **S**econd system or **SI** is used throughout the world except the United States, Liberia and Myanmar. Its unit of energy is the Joule named for the physicist who lead in the development of our understanding of processes involving energy. It is more helpful to call the unit of energy in the MKS system the $\frac{\text{Kilogram} \cdot \text{Meter}^2}{\text{Second}^2}$ or $\frac{\text{KM}^2}{\text{S}^2}$ its dimensional content in the MKS system. Better yet, is to call energy a $\frac{\text{Mass} \cdot \text{Length}^2}{\text{Time}^2}$ or $\frac{\text{M} \cdot \text{L}^2}{\text{T}^2}$. See Section 2.5 on page 44

works. This algorithm is useful for medium lengths such as measured on the earth but for astronomical distances and extremely small distances we need an alternatives; you cannot lay a rod down and compare. For situations like this, we find alternative measures which are found to replicate the laying down next to the standard or secondary standard in contexts that we can test and then extend the protocol to those cases for which the laying down process cannot be implemented. For astronomical lengths, we use what is called the standard candle. Basically for certain objects we know how bright some object is close up and use the fact that the brightness decreases with distance in a known pattern and thus from the perceived brightness infer the distance. This approach is used for intergalactic distance measurements.

Actually now, for lengths we don't have a standard. The modern construction for length measurements uses the speed of light and a time for the algorithm. It is simply better to use the definition of the meter as a product of the defined speed for light in a vacuum and a defined time. This has become the standard for all cases, see Section 2.3.1 on page 40.

2.3.1 The Story of Length

The story of length is interesting and pertinent. Length is probably the most basic of measured quantities and its history shows many of the characteristics of all measure systems. The need to measure lengths clearly goes back to antiquity. In particular, measurement of segments of the earth's surface was an important activity even at the time that man was still a hunter gatherer. At best, the distances were measured in crude and qualitative ways. With the advent of agriculture, length measurements took on an even greater significance. Not only was there a need to measure plots of land but there was also a need to standardize the units of measure. In all likelihood, the early measures were a crop yield. The tendency to measure land by yield persisted well into the nineteenth century. This measure was ultimately displaced by the more objective measure based on a predetermined length. As societies became more organized, standards were introduced and managed by the those in control of those societies and the control of the instruments of measure became one of the primary duties of government.

Earlier standards such as the length of the king's foot were a reasonable standard. They could at least be required universally but still they were not stable or convenient when you wanted to use them. At some point, a secondary standard, two marks on a rod, that was made from the primary, the king's foot, became the standard and was kept in a special place. A part of the problem was that there were different kings and different municipali-

ties had different standards. It was so chaotic that in some cases merchants used one length standard to purchase materials and a shorter one by the same name for selling them. It was in this context, that the metric system and the idea of the meter was developed. A solution to the universality and consistency problem.

In 1791, The French Academy of Sciences decided to make a standard of length that was “natural.” The hope being that if it was natural it would be universal and stable. The need for a better system of measurement was acknowledged by everyone. The Academy was encouraged by the soon to be replaced regime of Louis the XVI and despite the turmoil of the French Revolution was continued by the several new regimes that followed. The Academy choose as the unit of length the meter which which was defined as $\frac{1}{10,000,000}$ of the quadrant of the Earth’s circumference running from the North Pole through Paris. This was an interesting choice because it was difficult to measure accurately and hardly accessible. In some sense it is not even “natural.” Because it is not the length of the quadrant but the length of a quadrant of the smooth surface that is at sea level, a quadrant of the geoid, an idealized model of the shape of the earth. At the time of this selection as the meter, a competing idea was to make the standard of length the length of a pendulum whose period was one second. The second at the time being defined as $\frac{1}{60} \times \frac{1}{60} \times \frac{1}{24}$ of the day. This idea was dismissed because of the known variation of g , the acceleration of gravity, and the reluctance to base one fundamental unit on another. The variation of g would require that the meter be defined at one specific location on the earth and the hope was that this standard would be universal and accepted by all nations. The meridian through Paris was chosen not because it was in France but because it provided the longest land mass along a meridian that was in a major country. The problem of the dependence on time for length is interesting in light of our current definition, see later.

It was not long before people realized that the original choice was not a reasonable one. Not only was it hard to measure and access, it changes over time. The struggle to measure the meter as defined by the French Academy of Science is an interesting story as told by Alder [Alder 2002]. Also when it was measured later and more carefully, it was wrong. The current best measurement of the quadrant of the geoid is 10,002,290 meters. Although this is better precision than we need for this class, it is not sufficient for a modern industrial society. A new more precise measure is needed. The secondary, the bar in Paris, became the standard.

By 1960 advances in the techniques of measuring the wavelength of the emission lines of atomic radiation had made it possible to establish a more

accurate and easily reproducible standard not dependent on any artifact. In 1960, the meter was thus defined in the International System of Units as equal to 1,650,763.73 wavelengths of the orange-red line in the spectrum of the krypton-86 atom in a vacuum. It should also be obvious that these new standards were becoming more precise in order to accommodate the needs of a modern technological society for exacting metrology.

By the 1980s, advances in laser measurement techniques had yielded values for the speed of light of unprecedented accuracy. With the success of the Special Theory of Relativity, see Chapter 11 on page 241, it was realized that the speed at which light in vacuum traveled was a universal constant. It was decided in 1983 by the General Conference on Weights and Measures that the accepted value for this constant, the speed of light, would be exactly 299,792,458 meters per second. The meter is now thus defined as the distance traveled by light in a vacuum in $\frac{1}{299,792,458}$ of a second. This is a subtle but dramatic change in our understanding of length. We no longer use a fundamental distance as the basis for our measure of length. Instead, we use a velocity and a time. Now length is the secondary quantity and length is derivative. This idea can be extended to create a system of units that is based on the Fundamental Constants of Nature, see Section 2.6 on page 48.

2.3.2 Accuracy and Precision of Standards

In the past few years, there have been many changes to the choice of standards. The principle reasons for change has been the need for increased accuracy in measurement. In a modern industrial society, it is essential for successful commerce to be able communicate size in a confident precise manner. In a sense, you can never measure better than your standard can be interpreted. In the section on the “The Story of Length,” Section 2.3.1 on page 40, you can ask what is wrong with always using as the definition of the meter the distance between the scratches on a bar at the International Bureau of Standards. This is the definition of the meter and how can another definition be more accurate? When technology advances, and people need to make measurements in the micron and sub-micron range or at astronomical distances, a standard based on scratches on bars cannot be reproducible on these size scales. On the the microscopic scales, where in the scratch is the end of the meter? In a sense, the standard is always accurate and is the definition. But if there is an intrinsic error in the process of reading the standard or if the definition is ambiguous, the definition has only a range of usefulness.

By producing a standard that can be compared with greater precision, all measurements have an improved accuracy. Please note the contrast between the use of the words precision and accuracy in the preceding sentence, see Section 1.3.1 on page 11. In the astronomical case, the comparison algorithm cannot be implemented. There is no way to lay out rods between galaxies. Why not work with an algorithm that can be used? One of the beauties of the use of the speed of light to define length is that the primary standard can be used directly in the measuring process.

2.4 Quantities of Physics

As stated above, most of the quantities of physics are measured. I would go so far as to state that all the important quantities are measured. Since all measurements are comparisons, all quantities have a unit. The lessons of the previous sections are that when you talk about a quantity in physics you always keep track of its dimensional content and when you state a numeric value for a physical quantity, you must also state the unit to which it is compared (i.e., length in meters, mass in kilograms). There are also some non-measured quantities that come from the manipulation of measured quantities. These quantities are dimensionless. There are two sources of dimensionless quantities, mathematical manipulations and cancellation of dimensional content. An example of the first is the “ $\frac{1}{2}$ ” in the formula for the distance moved by an object with constant acceleration, a , in time, t :

$$d = \frac{1}{2} a t^2. \quad (2.1)$$

You can also say the same thing about the “2” in the exponent. These quantities are not measured quantities and there is no sense in discussing their precision and they are dimensionless. They come from the processes of mathematics (the algorithms) that we develop to help us understand important concepts.

Another way that we derive dimensionless quantities is by canceling dimensions. The dimensional content of a compounded physical quantity is algebraic reduction of the dimensional content of the elements of the quantity. In equation 2.1 on page 43, the combination of variables on the right side of the equation, $\frac{1}{2} a t^2$, has the dimensional content of the factors composing it, $a t^2 \stackrel{\text{dim}}{=} \frac{L}{T^2} \times T^2 \stackrel{\text{dim}}{=} L$. Note that I have used the fact that the $\frac{1}{2}$ is dimensionless. In this case, the time dimension dropped out of the term of interest.

Another example in the category of a dimensionless measured quantity is angle.

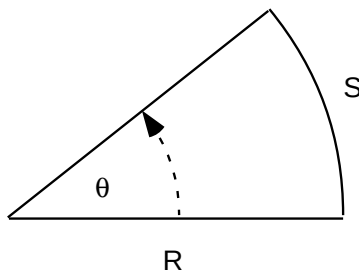


Figure 2.1: **The Radian** The definition of the angle measure called the radian is the ratio of the arc length S to the radius R . The dimensional content of angle is thus $\theta = \frac{S}{R} \stackrel{\text{dim}}{=} \frac{L}{L} \stackrel{\text{dim}}{=} L^0$.

An angle is the ratio of two lengths, see Figure 2.1. It is measured in radians using the ratio of the arc length to the radius for a given opening. In this example, S is a length and R is also a length and, for the angle defined as the ratio, the two lengths cancel out. Angle is a dimensionless quantity.

2.5 Dimensional Analysis

Because you must always maintain the dimensional content of a physical quantity and yet you can measure it in any unit, you obtain a powerful analytic tool called dimensional analysis. The physics behind this is that, since the unit choice is arbitrary, nothing important can depend on the unit used. This is an example of a symmetry which will be discussed in great detail later, see Section 7.1 on page 183.

Another way to say that you are maintaining the dimensional content is to say that in all relationships involving physics quantities all the terms must be homogeneous in their dimensional content. This is because all the relevant terms of physics are measured quantities and, as stated in Section 2.3 on page 38, all measurements are comparison processes. This is really based on the fact that size is a relative concept; we are large compared to atoms, but atoms are large compared to nuclei. All determinations of measured quantities are a relational operation and large or small is a matter of choice of unit.

We already took advantage of this idea in our discussion of the dimensional content of g in Section 2.2 on page 34. g is the gravitational force

per unit mass and has the dimensional content $\frac{\text{force}}{\text{mass}}$. If gravity is the only force acting on a body of mass m , then the force on that body is $f = mg$ and Newton's Law says that the body with total force f has an acceleration equal to the force divided by the mass, $a = \frac{f}{m}$, or $a = g$ in that case. Thus although from the definition, g has the dimensional content of $\frac{\text{force}}{\text{mass}}$, if this equation, $a = g$, is true g must also have the dimensional content of a which is $\stackrel{\text{dim}}{=} \frac{L}{T^2}$. In other words, since the dimensional content can be manipulated algebraically, both of these quantities must have the same dimensional content. For example, a length divided by a time squared has the same dimensional content as an acceleration. An acceleration times a time squared has the same dimensional content as a length. Is it a length? In some cases, it will be, i. e. it is twice the length displaced under constant acceleration, but it is not a length it merely has the dimensional content of a length and only in certain circumstances is it a length.

2.5.1 Uses of Dimensional Analysis

The simplest and most useful application of Dimensional Analysis is the recognition that, since the dimensional content is manipulated algebraically, that you can use it to make sure that your algebraic manipulations are correct. If you have done a problem asking you to find the time of oscillation of the pendulum of length, l , in the earth's gravitational field, g , and you have obtained $T \stackrel{?}{=} \frac{1}{2\pi} \sqrt{\frac{g}{l}}$ you can be sure that you made an error because the dimensional content of both sides of the equation are inconsistent. Note that you cannot tell a thing about the correctness of the dimensionless $\frac{1}{2\pi}$ part.

The requirement that the dimensional content of all equations be homogeneous is a lot like the idea that you must only add like things. You can only add apples to apples. You cannot add apples to bananas.

Take, for example, this equation:

$$s = \frac{g}{2}t^2 + v_0t + s_0 \quad (2.2)$$

Now look at it dimensionally²:

$$L \stackrel{\text{dim}}{=} \frac{L}{T^2} \times T^2 + \frac{L}{T} \times T + L \quad (2.3)$$

²Throughout these notes, in cases where we are only analyzing the results for dimensional content, we will indicate the equation with the symbol $\stackrel{\text{dim}}{=}$

Using algebraic calculations, we see that each term on the right side of the equation is a length, i. e. $L \stackrel{\text{dim}}{=} L + L + L \stackrel{\text{dim}}{=} L$. This is what is meant by saying that the equation is dimensionally homogeneous, every term has the same dimensional content.

You should get into the habit of checking for the dimensions in an equation. It is a great algebra checker. If you had a formula that said $s = \frac{g}{2}t$, the dimensional content is not homogeneous. Therefore, it is wrong.

Check that the dimensional content of any equation that you write is consistent. It is a good habit to get into.

Probably another place that you have used Dimensional Analysis is in the changing of units. When you are using a given standard as the dimension, then you are using a specific unit. Again, let's examine lengths. Length is the dimension. Several length units are the meter, the foot, and the light year. They are all lengths (L). A neat unit is the "lightnanosecond". It is a length that is about equal to the foot. It is defined as the distance that light travels in one nanosecond. How long is it in inches? In some sense, this is a silly question. It is always the same length. It has different numeric values depending on the unit used. The calculation is simple:

$$1 \text{ lightnanosecond} = 3 \times 10^8 \frac{\cancel{\text{meters}}}{\cancel{\text{sec}}} \times 10^{-9} \cancel{\text{sec}} \times \frac{39 \text{ inches}}{\cancel{\text{meter}}} = 11.7 \text{ inches} \quad (2.4)$$

You always maintain the dimensional content of all quantities by multiplying by a dimensionless ratio equal in value to one. For example. one foot is 12 inches. Therefore you can multiply any quantity by $\frac{12 \text{ inches}}{1 \text{ foot}}$. An example is the problem of finding how many seconds there are in a year. I will also use this example to show how to do numeric calculations to within a reasonable precision without a calculator. These tricks use the algebraic relations on 'Things'.

Seconds per year:

$$\begin{aligned} 1 \cancel{\text{yr}} \times \frac{365 \cancel{\text{days}}}{\cancel{\text{yr}}} \times \frac{24 \cancel{\text{hrs}}}{\cancel{\text{day}}} &\times \frac{60 \cancel{\text{min}}}{\cancel{\text{hr}}} \times \frac{60 \text{ sec}}{\cancel{\text{min}}} = 365 \cdot 25 \left(1 - \frac{1}{25}\right) \cdot 3600 \text{ sec} \\ &= 365 \cdot \frac{10^2}{4} \cdot 3600 \cdot \left(1 - \frac{4}{10^2}\right) \text{ sec} \\ &= 365 \cdot 10^2 \cdot 10^3 \cdot \left(1 - \frac{1}{10}\right) \cdot \left(1 - \frac{4}{10^2}\right) \text{ sec} \\ &\approx (365 - 36 - 15) \cdot 10^5 \text{ sec} \\ &\approx \pi \times 10^7 \text{ sec}. \end{aligned} \quad (2.5)$$

2.5.2 Scaling Laws

The opposite case of using the dimensional content to check algebra is to use the dimensional content of variables to determine the relationships between the sevariables to produce the relationships. In other words, once you identify the important variables, you must find what combination has the correct dimension. If this combination is unique, then to within dimensionless factors you know the relationships between the variables. These are called scaling laws. Kepler's laws are the direct result of the dimensions of G , the constant from the universal law of gravitation, $f = G \frac{m^2}{r^2}$.

$$G \stackrel{\text{dim}}{=} \text{force} \frac{\text{distance}^2}{\text{mass}^2} \stackrel{\text{dim}}{=} \frac{L^3}{M \times T^2} \quad (2.6)$$

Suppose you want to know the time (T) that it takes a planet with orbit radius (R) to complete an orbit around a body of mass (M). Since these three are the only variables that can matter, the only combination of these variables with the correct dimension is:

$$T = \sqrt{\frac{R^3}{GM}} \quad (2.7)$$

This argument is based on the dimensional content of the variables in the problem. Note that it cannot identify any dimensionless variables such as 2π .

Let's take another example: With one motion of my arm I can throw a ball so high. How much higher will it go if I move my arm through the same motion in half the time? First, break the problem into two parts. (1) To move my arm through the same distance in half the time is to say that I have *doubled* the speed of my throw. (2) How does the height scale with the initial speed? To find the answer, we consider that the only combination of speed and acceleration of gravity (g) that gives a distance is $\frac{v^2}{g}$. So, if I halve the time of the motion (i.e., if I double the velocity), the height will increase by a factor of four. Analysis of this kind is called a scaling relationship.

Another example of simple scaling problem. You are walking with a small child that is $\frac{1}{2}$ your height. Assuming that you are walking in the same fashion with an unforced gait, what is the ratio of your speeds? Answer $-\sqrt{2}$.

The basic idea of this analysis is to identify the relevant variables, and then determine which ones can be combined to form something with the correct dimensions for an answer. In this case we need a speed. This is dimensionally $s \stackrel{\text{dim}}{=} \frac{L}{T}$. The relevant variables are the length scale (L) and the acceleration of gravity (g) (this is what is meant by an "unforced gait").

The unique combination of (L) and (g) that is a speed is $\sqrt{Lg}$. Since the length dimension is the height, the ratio of the heights is $\sqrt{2}$. The value of g is unchanged in this example. The situation of the comparing the speed of astronauts on the moon to their gate on earth where the value of g is different from on earth would be a different story.

2.6 Fundamental Dimensional Constants

2.6.1 Sizes

The scale of all things is not arbitrary. In the film “Powers of Ten,” Section 1.5.1 on page 24 and in the “Plot of Masses and Lengths,” Section 1.5.1 on page 25, we saw that things come in certain sizes. There are no atoms the size of the sun! From our discussion of dimensions, we realize that from the freedom of choice of standards or units that all numeric values of size are possible. What is it then that sets the sizes of things? We also realize that, If the fundamental laws were only expressed by purely mathematical symbols, there would be no factors that could lead to sizes or periods of time. If you want large departures of size using similar rules of the game, you will need to have factors in the rules that reflect the different sizes. These are the dimensional parameters that appear in the equations. These are the determinants of size. Said another way, sizes have to come from somewhere; mathematics cannot provide them.

Let’s discuss a concrete example. As discussed earlier in Section 1.5.1 on page 25, all atoms are about the same size. We will discuss this case in detail in Chapter 19 on page 403 but for now all we have to realize is that the size of an atom has to come from the dimensional variables that govern the system. The size of an atom, in particular the hydrogen atom, is set by the fact that it is a system that is composed of an electron held close to a proton by the electric force and using the dynamics associated with quantum mechanics. This says that the size must be determined from combinations of the mass of the electron, $m_e \stackrel{\text{dim}}{=} M$, and the constants associated with the electric forces, $\frac{e^2}{4\pi\epsilon_0} \stackrel{\text{dim}}{=} \frac{ML^3}{T^2}$. You can work this out similar to the analysis of the dimensional content of Newton’s Gravitational Constant, G . The use of quantum dynamics brings in Planck’s constant. Looking up the units of Planck’s Constant in the table of “Things that Everyone Should Know”, Section 1.4.2 on page 16, shows that it is an energy times a time, a Joule Sec, and thus has dimensional content of $\frac{ML^2}{T}$. This is a particularly important combination of dimensions and has its own name, Action, which we will

discuss in great length later, see Section ?? on page ??. All three of these parameters have dimensional content and there is a unique combination that leads to a length. Work it out. Thus, we see that the size of atoms is set by the parameters that describe the system to within a dimensionless factor which we always assume is of order unity.

Thus we have a rather general result. Although we use mathematics to express our laws, the variables are physical variables and therefore they have dimension. Similarly, in the articulation of any law, there may be and, in general, there will be constant parameters that are themselves dimensional. In a world with no fundamental dimensional constants there would be no scales of size or time. Since we know that phenomena come in specific sizes, fundamental dimensional constants must exist. In any problem, sizes are set by the dimensional parameters of the problem. This means that something in nature is restricting the sizes that we see. Look at the chart of all the things in the universe, Section 1.5.1 on page 25. The things on this chart are concentrated in specific places. This is because of the dimensional constants of the laws governing their behavior: Plank's constant $\hbar$, the gravitational constant G , the speed of light c , the mass of the electron m_e , and the mass of the proton m_p . These constants set the scales of the phenomena that we observe.

Of the family of dimensional constants of nature, some of these are thought to be more fundamental than others. This is in the sense that in some complete theory of everything, all phenomena would be derived from these. The “Fundamental” dimensional constants are $\hbar$, G , and c .

$$\hbar \stackrel{\text{dim}}{=} M \frac{L^2}{T}, \quad G \stackrel{\text{dim}}{=} \frac{L^3}{T^2 M}, \quad c \stackrel{\text{dim}}{=} \frac{L}{T} \quad (2.8)$$

This choice is based on the fact that there are indications in our current understanding of nature that all the others will be computed from them or a set that is closely related to them. For instance, in string theory, the latest candidate for a “Theory of Everything,” we seem to have a successful approach to a quantum mechanical theory of gravity. The theory has, in addition to $\hbar$ and c , one dimensional parameter, the string tension. The value of the string tension is set once you require that the theory reproduce classical gravity. In this way the tension is set by G . If string theory is to be a “Theory of Everything,” then all the masses and strengths of interactions would follow.

2.6.2 Modern Standards

To describe the motion of material objects, there are three independent types of measurements that must be made. All others are combinations of these three. Historically, we used length (L), mass (M), and time (T). Using the laws of physics, all other quantities are then derived from these three fundamental ones. For instance, you may think that there is a measurable thing called force, a push or pull between two bodies. Yes, you could even develop a standard and an algorithm for comparing forces. You might then think that this is an independent unit. But, you also have $\vec{F} = m\vec{a}$. For this equation to be valid for all systems, a force $\stackrel{\text{dim}}{=} \frac{ML}{T^2}$, and thus be reduced to the length, mass, and time dimensions. Thus, force can be viewed as just some special combination of our basic units, a derived unit.

Actually though, why couldn't force be fundamental and one of the other units derived? When you think about it you realize that these are obviously a choice. Units are chosen for several reasons: convenience, utility, and reproducibility. If you define length by using scratches on a special bar, it is convenient and reproducible; but, it will not work for extreme cases, so you need to find another method for those cases.

In a discussion of "The Story of Length," Section 2.3.1 on page 40, we have seen that length has now become the derivative concept and that a certain velocity, the speed of light, times a time which clearly has the dimensional content of a length is the defining concept for length. In that sense, now instead of time, mass, and length, we use time, mass, and a special velocity as fundamental standards. In fact, if you think about it you realize that if we use a unit like the "lightnanosecond" mentioned earlier for a length, we are actually working in a system that uses as it fundamental quantities a time and a velocity, the speed of light, instead of a time and a length. If you look at a good table book for the value of the speed of light, you will be told 2.99792458×10^8 m/s (defined). Here we have chosen a speed as our basic unit instead of a length. You define the speed of light to be the appropriate value to reproduce your old standard of the meter as the distance between two scratches on a bar. The bar is now a secondary standard with a finite precision. The primary motivation for the change in definition was the need for increased precision in length measurements. It turns out that it is easier to make very precise measurements of time intervals. Thus defining or better said using as a standard the speed of light and a time you produce a very precise standard of length.

So – What is so special about length, time and mass? The answer is nothing.

Once it is realized that there is nothing special about length, time, and mass there are many options available that may be more useful. What we need is three dimensional entities from which all the others can be found. We know that we need three because the classical system had at least three; length, mass, and time. Our new choice has to be able to reproduce these three. Someday, we might use a time, the speed of light, a velocity, and Planck's Constant, an action. Is this possible? Will it work? In fact, it is likely. Measurements of the Josephson effect involving superconductors would allow a direct definitions of Plancks Constant and thus its use as a defining unit.

Someday, we will may be able to use a velocity, the speed of light c , an action, Planck's Constant h , and Newton's Gravitational Constant, G , the other fundamental dimensionisinal constants of physics? Probably not. The drive to use fundamental constants comes not from a desire for "naturalness" that so drove the metric choice but from the need for high precision for commercial and scientific applications. The trouble is that it is diddicult to measure G with any precision.

When you think about it you realize that the attempt to base standards on "Fundamental" dimensional parameters is an old one. In the "Story of Length," the French encyclopedists wanted to use the size of the earth. They thought that it would be fundamental. The original definition of mass was based on the density of water—the mass of a given volume of water. The trouble in both these cases is that although these are useful and in principle will work, they are not fundamental and thus always have an intrinsic limit to relevance. We currently reserve the designation "fundamental dimensional constants" only for the three constants $\hbar$, c , and G with the idea that all length mass and time scales will be derived from them. It is our hope that the laws of physics are complete enough that we will ultimately derive all the others from these three. They enter the laws of physics at the most basic level, and we do not expect that we will find a more basic source for them in the future. I am convinced that. if we could produce precise secondary standards from them, we would use a standard system based entirely on them. Our current measurements of $\hbar$ are becoming very precise and, some time soon, we will use it as one of our standards. The problem is that G since gravity is so weak that it may never be measured at the precision required.

Systems of units:

old old	length, density, and time
old	length, mass, and time
new	speed of light, mass, and time
Post modern	speed of light, gravitational constant, and action

Actually the ambiguity in the choice of unit systems is used to simplify calculations. By setting some chosen unit to take a special value, usually one, calculations can take on an especially simple form. The most common place where this is seen is in the Special Theory of Relativity. Many cumbersome c 's are eliminated if $c \equiv 1$. This is effectively what is happening when you are using the usual time units, say years, and distance in lightyears. In the end to recover the usual units, you just have to realize whether you are speaking of length or a time. This is carried to an extreme in many computations in which three entities are set to one and all units disappear.

That there are three independent fundamental dimensional constants is not an accident. We expect that they will give us all the structure that we see in the universe. But in the post modern view, almost any three basic independent dimensions will do. In olden times, we had length, time, and mass. If you think about it, you realize that you could as well choose a time, speed, and an energy. The other quantities like length are related to the fundamental constants by the laws of physics. For instance, with a standard force and mass you can derive an acceleration. The modern system uses the speed of light from which time and length are derived dimensions.

Are some choices of unit standards better than others?

We should select those that best fit the criteria—reproducible, available, stable, and precise in the sense that secondary standards are precise. If it turns out that some unit standards are themselves basic laws of physics, then what could be more reproducible? We should use these if they are precise enough for the uses we need.

2.6.3 Useful Dimensions

The following table is a collection of physical quantities, **Quantity**; the vernacular name, **Name**; the usual symbol, **Sym.**; the associated dimensional content, **Dim.** and the SI unit, **SI**. The three dimensional entities are mass, M, length, L, and time, T. The symbols for the SI units for these are m, meter, s, second, and kg, kilogram. Certain SI units are not primary but are honorific. These are the Newton, N, for force which is a $\frac{\text{kg m}}{\text{s}^2}$, the Joule, J, which is a $\frac{\text{kg m}^2}{\text{s}^2}$ for energy, the Watt, W, for power which is a $\frac{\text{kg m}^2}{\text{s}^3}$, the Pascal, Pa, for pressure which is a $\frac{\text{kg}}{\text{m s}^2} \dots$

Table of Useful Dimensions

Quantity	Name	Sym.	Dim.	SI
distance	spatial interval	x	L	m
time	temporal interval	t	T	s
area	area	A	L^2	m^2
volume	volume	V	L^3	m^3
speed	rate of change of distance	v	$\frac{L}{T}$	$\frac{m}{s}$
acceleration	rate of change of speed	a	$\frac{L}{T^2}$	$\frac{m}{s^2}$
angle	opening between two lines	θ	$\frac{L}{L} = L^0$	radian
angular speed	rate of change of angle	ω	$\frac{1}{T}$	$\frac{\text{radian}}{s}$
mass	amount of stuff	m	M	kg
mass density	stuff per unit volume	ρ	$\frac{M}{L^3}$	$\frac{kg}{m^3}$
angular inertia	angular mass	I	ML^2	$kg\ m^2$
force	mass $\times$ acceleration	F	$\frac{ML}{T^2}$	N
force moment	torque	τ	$\frac{ML^2}{T^2}$	N m
pressure	force per unit area	P	$\frac{M}{LT^2}$	Pa
energy	force through a distance	E	$\frac{ML^2}{T^2}$	J
energy density	energy per unit volume	u	$\frac{M}{LT^2}$	$\frac{J}{m^3}$
power	energy per unit time	P	$\frac{ML^2}{T^3}$	W
momentum	mass $\times$ velocity	p	$\frac{ML}{T}$	$\frac{kg\ m}{s}$
angular momentum	angular inertia $\times$ angular speed	L	$\frac{mL^2}{T}$	$\frac{kg\ m^2}{s}$
action	energy times time	S	$\frac{ML^2}{T}$	J s
gravitational constant	strength of gravity	G	$\frac{L^3}{MT^2}$	$\frac{m^3}{kg\ s^2}$
Planck's constant	unit of action	h	$\frac{ML^2}{T}$	Js
elastic modulus	compressibility	E	$\frac{M}{LT^2}$	Pa

Chapter 3

Pre 19th Century Physics

3.1 Introduction

We begin our study of modern physics by examining the phenomena associated with light. Although the phenomena of light is among the oldest examined and there are theories of light that must go back to the first humans, light is particularly interesting from the modern perspective because of the central role that it has played in the development of our ideas, particularly quantum mechanics and relativity. To trace the development of our ideas about light from the ancients to today would take the entire semester and not allow any room for modern physics. For a concise coverage of these ancient ideas, the book by Park [Park 1997] or [Ronchi 1970] are excellent. Instead, we will start with one of two threads of development that emerged in the 17th century and as we will see are relevant to our modern approach to light and related phenomena.

In the 1660's, Fermat¹ proposed that light travels between two points over the path, called a light ray, that has the least travel time of all the possible paths connecting those two points, Section 3.2 on page 57. At the time of its formulation, there was a competing theory: a particle theory often identified with Newton. The particulate theory was the generally accepted description, because it successfully accounted for all the phenomena known at that time to be related to light, basically reflection and refraction. Fermat's approach equally well described the reflection and refraction experiments of the day. In a sense, there was a stalemate with the great prestige of Newton providing the edge to the particulate approach. A significant

¹ Pierre de Fermat, born in 1607 and died in 1662, was responsible for significant contributions to mathematics and optics.

difference between the two theories was that Fermat's Theory required that light traveled slower in a dense media whereas the particulate approach as given by Newton required that light travel faster. At the time of formulation of these competing theories, it was impossible to measure the speed of light in dense media. Once the confirming experiment supported Fermat's theory, it became the accepted approach. We now know that, in some sense, some of the aspects of the particulate theory are correct for a certain range of observation but we will get to this later in Chapter 19 on page 403. With the measurement of the speed of light in dense matter, the particulate theory was then superseded by a Fermat's theory. Fermat's formulation was very successful in describing all the phenomena associated with light that was known at his time and, in fact, most of the common phenomena that we associate with light, see Section 3.3 on page 63. Newton's lasting contribution was the description of the relationship of color to light, see Section 3.4 on page 79. Interestingly, the modern interpretation of the behavior of light comes very close to what Newton developed and some argue that Newton was close to the discovery of quantum mechanics at least as it is applied to light.

As new phenomena, interference and diffraction, associated with light were observed, it became clear that a new construction was needed. Extending and clarifying a construction associated with Huygens, a contemporary of Newton, and Thomas Young, Fresnel formulated a new approach that in the appropriate limits reproduced to all the success of Fermat but incorporated the new phenomena of interference and diffraction, see Section 4 on page 83. Integral to the success of this approach is the idea of an underlying continuous system, the ether, that was the basis for the the phenomena associated with light. Much of the intuition of the new construction was based on the understanding of fluid flows and, in particular, sound that had been developed earlier. These also depended on the properties of an underlying mechanical system, air for sound and water, for surface waves², for their interpretation. There was similar material basis for light. Later, in what at the time appeared as an independent investigation, Maxwell was attempting to construct a mechanical model of electric and magnetic forces. In his mechanical model of electric and magnetic forces, Maxwell realized that disturbances traveled at the speed of light and he immediately identi-

²It is common practice to call all disturbances that travel in a field system as waves. This is not completely clear since in any field situation that supports traveling solutions any disturbance will travel but is called a wave and the disturbance does not have to take the shape of a sine or cosine. Where appropriate I will try to be more formal and call them travelers.

fied these as light. It also provided a material basis for the propagation of light called the aether. The methods of analysis that emerged are a special case of a local field theory, Section 5.2.2 on page 126, but this carries us well into the 19th century and Chapter 5 on page 123. In the last century, the classical theories of light were superseded by two successful modifications of our understanding of light, the theory of Special Relativity which lead to a new interpretation of the nature of gravity and a full fledged quantum field theory of light, Quantum Electrodynamics or Q. E. D. The exposition of these precursor theories will take most of our time but is necessary to appreciate the nature of our final theories. Although the theories we describe here are not the modern theory, they are interesting predecessors to it, and they provide us with valuable insights into the fundamental concepts of the modern theory.

We will cover the thread of development from Fermat's and Fresnel's approaches to light in some detail here because they will allow us to develop both an intuition about these phenomena but also develop technical tools that are necessary to articulate the modern approaches. These theories also provide a wonderful example of how transition occurs in physical theory and we will see a similar transition to the modern theory. Hopefully, you will also see how the Fresnel approach was required to produce in the appropriate limit the Fermat theory. This is the usual case. The older approach had to have some successes or it would not be accepted. The identification of new phenomena that the older theory could not accommodate are the stimulus for the new approach. Despite its ability to accommodate the newly realized phenomena, the new theory must also fit the old successes. This later issue is often the hardest part in the development of a new theory.

3.2 Least Time Formulation of Light Propagation

. Fermat's 'Least Time Principle' describing how light travels between two points is an excellent example of a theory that agrees with the data and appears to be computationally simple. An interesting feature of this theory is the interaction of the development of the theory with the concomitant development of new mathematical tools. This theory appears on the surface to have a simple and straight forward computational basis but on careful examination, reveals deep and subtle mathematical complications. This is also typical of all theory development – new mathematical understanding will generally be required for the successful implementation of the theory, see Section 3.3.7 on page 76.

The rule is stated very simply. If you want to find the path that light travels when it moves between any two points. You find all possible paths. On each path, find the time that it takes the light to travel over the path. The light travels between two points in space over the path that has the least travel time. This statement of the rule is so intuitive that two things tend to happen; you think that it is obvious or you tend to say that this is what *light* does.

This process of selecting a path on the basis of some extremum property is very common. You have often selected least time paths or, at least, a least something path. Maybe you want to conserve gasoline, or go the shortest distance, or avoid speed traps. But you have an extremum rule. This is a satisfying way for choosing an action. Similarly, you then feel that it makes sense that light would do this also. There are several problems with this idea. There are lots of choices about what to extremize. Not only that but it implies an anthropomorphic basis for the behavior of inanimate phenomena. But realize that Fermat is not saying that the light calculates the travel time on each path and then selects the least time path. He says instead that, if **you** want to find out the path, you must identify all paths, a prodigious undertaking, see Figure 3.1 on page 59, and actually very subtle issue, know the speed with which the light travels at all points in space, calculate the time for each path, and finally pick the path with the least time. Clearly, *light* does not do all these things. These are activities of people. It is interesting to point out that for light to do this calculation of time on all paths and choosing the right one, we need a natural argument for how light does this. Interestingly, this was accomplished by Fresnel, see section 4.6.5 on page 117. This is an valuable example which we will discuss of how a new theory recovers and clarifies the older theory. Regardless, these least time paths that light travel over are called the rays of the light and they are where the light goes in the Fermat picture. The experimental verification of the predicted path is to place a barrier at a point on the path and see if the light no longer connects the two points that were the end points.

This formulation of the rule raises many interesting conceptual questions beside the anthropic one of how the “light” does it. Note that it is formulated in such a way as to specify where the light goes between two points. This algorithm does not start at one point and in a direction and decide point by point how the light progresses; it does not propagate the light. This rule is not a local rule which is the usual way that we look at how systems develop. This makes it what is called a global rule. You need to determine the time of travel for the total path. You start with two points that are well separated in space. Of course, once you know the path, you can come

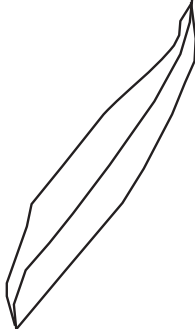


Figure 3.1: **Least Time Path** When light travels between two points, it travels over the path that requires the least time of travel. In order to find that path, simply find the travel time over all paths and choose the minimum of the set.

back and apply it to any pair of points along that path and, in all cases, the segment of path that was obtained as the least time path is also a minimum in family of paths between those points. This holds no matter how closely placed the new points are so that the rule can take on a local character. The only problem is that this local information can be obtained only after the path between the two separated points has been found.

Just how simple, algorithmic, is this rule. This seems algorithmic since it is a prescription that anyone can follow. You have followed it many times when you pick a travel route between two cities. What's the best way to go between Austin and Houston? You take a map with all the roads indicated on it. You classify all routes. On any route, you divide the trip into segments and then estimate your speed in each segment. From the speed and the length of the segment, you can calculate the time for that segment and then you add up the time for each segment to get a total.

$$T(route) = \sum_{\text{segments}} \Delta t_i = \sum_{\text{segments}} \frac{\Delta s_i}{v_i}. \quad (3.1)$$

where Δt_i is the time in each segment labeled i and Δs_i is the length of that segment i and v_i is the speed in that segment. You somehow make an ordered list of routes and repeat this process for all routes. Once you have $T(route)$ for all routes, you look down the list of travel times and select the one with the least time. That is the route that you take if you want the least time. In a similar fashion, if the light goes between two points, with this algorithm, if you know the speed of light at every place, you can find

the routes in space through which the light travels.

Is it really this simple? First, let's take a closer look at the algorithmic nature of this process. Despite its apparent simplicity, is it really well defined? It requires that we look at all paths. How many paths are there? A lot. In contrast to our highway problem, there are an infinity of paths. The problem of making sure that you have all paths is a complex one, and we will reserve a detail discussion for later, Section 3.3.7 on page 76. Just be assured that the requirement for an examination of all paths is not simply met and, in fact, this is one of those cases in which new mathematics had to be developed to meet the needs of the physics. Related is the fact that need to make a table of paths so that we can scan down it to make the choice of the least time, i. e. time as a function of path. This requires that we are able to make an ordered set of paths. Is this always possible or even ever possible? Again new mathematics will be required. It is worth noting that, generally when you do the highway problem, you have so few paths that you can keep track of them in your mind and maintain an order in that fashion.

Another problem is that, for each path in order to calculate the time that it takes the light to travel from end to end, you must know the speed of light at each point on the path and, since you must do this for all paths and since in the family of all paths all the points in the space will be touched. This implies that you will need to know the speed for all points in the space. A great deal of information. Also in a manner similar to the highway problem, you need a speed for each segment. This implies that you must also sensibly rectify the curved parts of the path, see Figure 3.2 on page 61. This is because of both the variation in the speed of light at different points and the curvature of the path. This is what you do when you calculate the time of travel between two cities. You add up the segments with comparable speeds and you make the curved parts out of straight segments that approximate the path.

How do you decide how big to make the straight pieces? You should pick the size of the straight intervals so that they follow that path with close precision and so that the speed of light is reasonably constant throughout the segment. Depending on the precision that you need when you calculate the time you may use a more coarse or a more fine grid.

With this done, you can calculate the time Δt_i in any straight line segment where Δs_i is the length of the straight line segment and v_i is the speed of light in that segment:

$$\Delta t_i = \frac{\Delta s_i}{v_i} = \frac{\sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2}}{v_i} \quad (3.2)$$

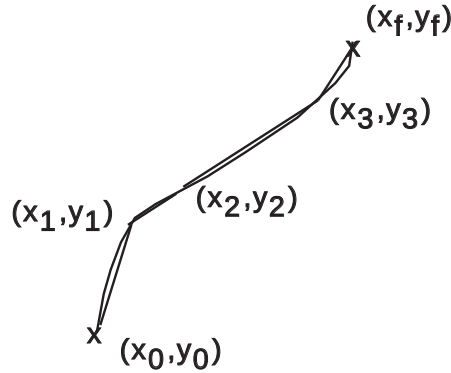


Figure 3.2: **Least Time Path in Inhomogeneous Medium** In order to calculate the time over a curved path in an inhomogeneous space, a space in which the speed of light varies from place to place, you must sensibly rectify the path, i. e. reduce the path to a set of straight line segments. The length of the segments depend on the precision of the calculation and how it is impacted by the variation of the speed and the amount of curvature of the path.

and then add the time for all the segments to get the total travel time, $T(path)$,

$$T(path) = \sum_{\text{segments of path}} \Delta t_i = \sum_{\text{segments of path}} \frac{\Delta s_i}{v_i}. \quad (3.3)$$

where I have now added the phrase “of path” to the right side of the equation to emphasis that throughout the computation we must somehow keep track of the specific path from a large family of paths with which we are dealing. Then, by means of some protocol, we select the path with the least travel time and call it the “least time path” and, according to the theory, this will be the path that the light will travel. In other words, placing a obstacle at a point in this particular path will extinguish the light between the two points.

This hypothesis “explains” reflection, refraction, and many other optical phenomena. Although, in the years since the 17th century, we have developed several layers of superseding theories of light phenomena, the Fermat rules are still the basis for much of the design of optical instruments and the basic explanation for many atmospheric optical phenomena. In the following sections, I will illustrate this. This is not to say that only Fermat’s

Theory will explain these phenomena. In fact, a particle based theory would do as well. As stated earlier, the clinching evidence for the Fermat theory was the requirement that to get refraction required that light in the more dense media traveled slower. This is often the case. People will cite examples of phenomena “explained” by a theory that works but a satisfactory explanation is not unique to the theory being cited. There are usually small but compelling differences when different theories are in competition. It has to be this way or there would be no competition.

3.2.1 Speculation on the form of Fermat’s Theory

Since this is our first attempt at theory construction, it may be appropriate to speculate on the nature of this construction. Firstly, this is not a Newtonian approach which is local at each place. The path that nature chooses for the light is based on a global measure – the total time of travel of the path. Newtonians would have had the light move from place to place by means of some rule that held at each place at each instant. In Section 6.1.5 on page 162, we show how another global extremum rule, a rule about least action, similar to this one, can recover a local statement about how the system develops. Generally, the idea is that, if we can assume that the extremum is reached smoothly in the very rich path space, then paths which differ slightly have the about the same value. In particular, the requirement that two paths that are the same everywhere except at an isolated point and that the deviation of the path at this point is small then the global measure has almost the same value implies a condition that constrains the effects at that point³. This constraint is a local statement on the path development. This result is intuitive from our experience in finding least time paths for travel. The least time path for a trip is always made up of segments that are themselves the least time path between the points at the ends of that segment. In other words, the least time path is always made up of locally, between nearby points, paths that are the least time between those points.

A similar observation is that, although the word time is an important part of the formulation of this rule, there is no real time involved. By this I mean that there is no real evolution of the system. The path is what it is. The time in this approach is just some global measure on path space. This observation is especially relevant when we realize that, at the time

³Another way to look at the same result, is to realize that, once you have identified the natural path, the segment of the path between any two points on the natural path is a natural path between those points. If the two points are arbitrarily close you have the above result.

of Fermat's formulation, the speed of light had not been measured. The situation was worse than that. At the time, it was not clear whether or not light even had a velocity. On some occasions, Descartes who was the pre-eminent natural philosopher of the time argued the light was instantaneous and at other times he argued that light had a finite velocity. I have to assume that Fermat choose time as the measure because he knew that there were circumstances in which length did not work and what else could it be. Thus he formulated a global measure which weighted each path segment with the inverse of velocity, $\frac{\text{time}}{\text{length}}$, and then predicted that, if it could be measured, you would find that light traveled slower, a higher inverse velocity, in dense media. Of course, this was his great success. It still leaves the question of what other measures are there. We know from our experience planning travel that all kinds of measures are possible. Instead of least time, there is the most scenic route. In that case, we would develop a measure of scenic, for example hilly, and apply the measure $\frac{\text{hilly}}{\text{length}}$ to each segment and add the contributions. We could even count unpleasant scenery as negative hillyness and develop a measure that can have either sign. It will turn out that, when we expand our study to include dynamics, we will need a new global measure in a path space in space-time called the "action" and we will find that the naturally occurring path in space-time, called the trajectory, is the one that is the least action.

3.3 Applications of Fermat's Principle

3.3.1 Light Travels in Straight Lines

Let's start with the simplest observation. What are the paths of light in a homogeneous medium? A homogeneous medium is one in which every point is the same. In particular, the speed of light must be the same at every point. Thus, in this type of medium, the least time path is the same as the shortest path. By definition, the shortest path is a straight line.

Proof:

$$T(path) = \sum_{\text{segments}}^{\text{path}} \frac{\Delta s_i}{v_i} \quad (3.4)$$

I have added the path designation to remind you that you must do this for each path. Based on the fact that all the v_i are the same at every point in a homogeneous medium, the $v_i = v$, the common speed for light for all points in the medium, and can be factored from the terms in the sum and you will

have:

$$T(path) = \frac{1}{v} \sum_{\text{segments}}^{\text{path}} \Delta s_i \quad (3.5)$$

Since $\sum_{\text{segments}}^{\text{path}} \Delta s_i$ is the definition of the length of the path, we see that the time for any path is proportional to the length of the path. Thus the least time path is the shortest-length path which, by definition, is the straight line path.

3.3.2 Refraction & Snell's Law

Refraction is the phenomena that occurs when light passes through a medium that has a varying speed for light. In this case, the ray bends. As the simplest case, chose a system of two media that are themselves homogeneous, separated by a planar interface, and place the two end-points in the different media. Both media are homogeneous, but they have a different speed for light called v_1 in media 1 and v_2 in media 2.

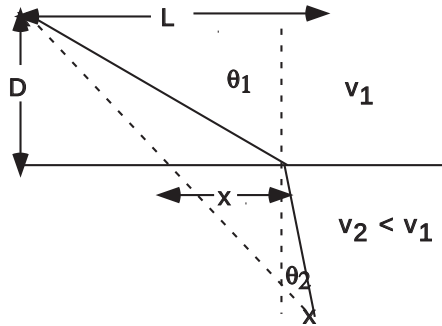


Figure 3.3: **Light Path in Two Homogeneous Media** The light path in two homogeneous media is straight in each part but kinked at the interface. In this example, the starting point of the ray is a distance D from the interface on each side and separated by a distance L measured along the interface direction. The distance along the interface from the point at which a straight light path would strike the interface plane and where the path strikes the interface is x . The angle between the normal to the interface and the path segments in each media are θ . The media are labeled by the speed for light in each media, v_1 and v_2 .

Our first problem is to determine how to discuss the paths that connect the two points. There are an infinity of them, see Section 3.3.7 on page 76. Physical intuition tells us though that the least time paths in a homogeneous

medium must be straight lines and thus the path with the least time overall must be among the paths that are straight within either of the two media and kinked at the interface, see Figure 3.3 on page 64. A path that is curved in one of the media would clearly be a longer time path than the one with the same start point and hitting the other media at the same point and then traveling in the second media. This is an example of how a global rule does have some local content. This ability to reduce the path space to kinked straight line segments is an important reduction in the nature of the problem. With this reduction in the size of the path space, we can label the paths with the distance of the kink position from the place at which the path would meet the interface if the two media were the same, i. e. the straight line path between the two points, see Figure 3.3 on page 64. Two things have been accomplished. We now have an ordering for the family of paths that we wish to investigate. Even more significantly, we have reduced the path space to one that can be mapped onto the real line. In this case, we are labeling the paths with the parameter x . Remember that functions are mappings of the real line onto the real line. This then gives us access to all the usual tools of mathematics.

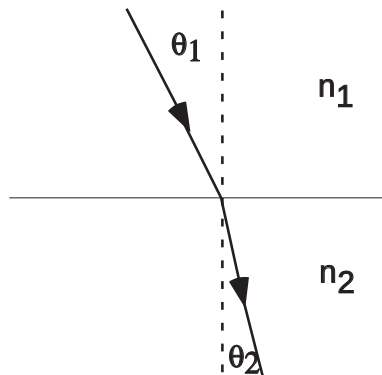


Figure 3.4: **Snell's Law** Snell's Law states that, when light passes from one optical medium to another, the ray of light bends at the interface according to $n_1 \sin(\theta_1) = n_2 \sin(\theta_2)$, where θ_i is the angle of the ray to the normal at the interface and n_i is the index of refraction for the material.

Once the path has been reduced to two straight line segments, it is easy to find the least time path. In this example for simplicity of analysis, I will pick two points that are equidistant from the interface as measured along the normal to the interface and that distance is D . The two points are a distance L apart as measured along the interface, see Figure 3.3 on page 64.

The time for the path with intercept x is

$$T(\text{paths}) = T(x) = \frac{\sqrt{D^2 + (\frac{L}{2} + x)^2}}{v_1} + \frac{\sqrt{D^2 + (\frac{L}{2} - x)^2}}{v_2} \quad (3.6)$$

The least time path is the one that has the minimum value for $T(x)$ for all x . This is the x value at which the slope of the T versus x curve is zero. The easiest way to find the slope means taking the derivative of T with respect to x . This is a small bit of calculus which I do not expect you to carry out. You can check my calculus if you like. I just want you to accept that it can be done and agree that the derivative is the slope and that the minimum occurs when the slope is equal to zero.

Taking the derivative, you get

$$\frac{dT}{dx} = \frac{(\frac{L}{2} + x)}{v_1 \sqrt{D^2 + (\frac{L}{2} + x)^2}} - \frac{(\frac{L}{2} - x)}{v_2 \sqrt{D^2 + (\frac{L}{2} - x)^2}}. \quad (3.7)$$

Setting the derivative equal to zero, and solving for the path with the least time yields

$$\frac{(\frac{L}{2} + x_0)}{v_1 \sqrt{D^2 + (\frac{L}{2} + x_0)^2}} = \frac{(\frac{L}{2} - x_0)}{v_2 \sqrt{D^2 + (\frac{L}{2} - x_0)^2}} \quad (3.8)$$

where x_0 is the label of the least time path. Using some simple trigonometry, we can relate the angle of the least time path with the normal at the interface to x_0 , see Figure 3.3 on page 64. From the figure, we have that $\sin(\theta_1) = \frac{(\frac{L}{2} + x_0)}{\sqrt{D^2 + (\frac{L}{2} + x_0)^2}}$ and $\sin(\theta_2) = \frac{(\frac{L}{2} - x_0)}{\sqrt{D^2 + (\frac{L}{2} - x_0)^2}}$ so that

$$\frac{\sin(\theta_1)}{v_1} = \frac{\sin(\theta_2)}{v_2} \quad (3.9)$$

or

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2) \quad (3.10)$$

where $n_i \equiv \frac{c}{v_i}$ and is called the index of refraction. c is the speed of light in a vacuum. Since $v_i \leq c$, $n_i \geq 1$. This is known as Snell's Law, see Figure 3.4 on page 65.

Following any derivation, it is useful to see if this agrees with our intuition. The light wants to spend the least time traveling between the two points. It is better to have more distance in the faster medium. Think of the lifeguard at the beach. She sees someone off to the side drowning. Although

she is a good swimmer, she can run faster on the beach than she can swim. Therefore, instead of going directly to the person drowning, she runs a little further up the beach past the point on the direct line to get to the victim in the shortest possible time.

It is worthwhile to note that, in the particulate theory of light, the path of the particles is bent toward the normal by the fact that the particles travel faster in the dense medium. Once it was found that light travels slower in the dense medium, the particle theory was not tenable. This is an often cited example of the Popper hypothesis of the use of falsifiability to prove or disprove theory in physics, [Popper 1973].

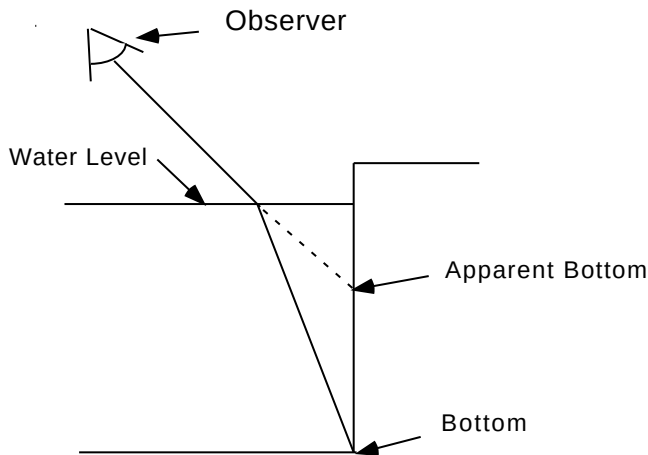


Figure 3.5: **Apparent Depth in Water** To an observer outside and above viewing a pool of water sees it as much shallower than it is. This is because the brain reconstructs the light ray that comes to the eye as a straight line path. Since the density of water is greater than air, the ray from the intersection of the bottom and the side of the pool and the eye of the observer is bent toward the normal in the water.

A direct application of Snell's Law is the observation that when viewed from outside a pool does not appear as deep as it actually is. The ray from the edge at the bottom of the pool to the eye is refracted, see Figure 3.5 on page 67. Since the speed of light is lower in the water than in the air, the ray bends away from the normal in the air. The observer's brain assumes that light travels in straight lines and thus places the intersection of the side and bottom of the pool at a much shallower depth.

The discerning reader may protest the interpretation of the apparent depth above does not make sense. A single ray cannot determine a point, in

this case the intersection of the bottom and edge of the pool. To find this point at the bottom of the pool, you need the intersection of at least two rays. As we all know, for humans the trick of depth perception is binocular vision. Even with one eye, the seeing involves the collection of several rays around the one shown and then the lens in the eye provides a measure of distance. Thus there are other rays that are not shown in Figure 3.5 on page 67 and these in combination with the ray shown ultimately determines the depth. This is a general truth. To find an image you require a collection of rays. Again, the discerning reader may note that, even without the binocular vision argument above, there really are many rays from the bottom edge point coming out the pool. These range in angle from the ray we have shown to rays that do not hit the eye but are from the bottom edge. A particularly simple ray from the edge is the one that runs along the side of the pool from the bottom straight up. Since this is parallel to the normal it is not refracted. It is also not detected by the eye. Regardless, there is a whole collection of rays fanning between this ray and the one depicted in Figure 3.5. The eye collects the rays from this fan that are near the ray depicted. The lens in the eye focuses these rays to a point on the retina. Its intersection with the refracted ray determines the position of the point on the bottom.

3.3.3 Total Internal Reflection

There is an interesting case of refraction that can occur when the light exits a dense or slow medium into a less dense or faster medium. Rearranging Snell's Law, Equation 3.9 on page 66, $\sin \theta_2 = \frac{v_2}{v_1} \sin \theta_1$, we can see that there can be cases in which θ_2 cannot be found.

If θ_1 is large and, thus, close to $\frac{\pi}{2}$, $\sin \theta_1$ will be close to one. Since for this case $v_2 > v_1$, $\frac{v_2}{v_1} > 1$. In this case, the product of the two terms in the rearranged Snell's Law could be greater than one and since the sin function is always less than or equal to one, there is no angle θ_2 that can satisfy the law.

In this case, the light does not penetrate the less dense surface but instead reflects from the surface with the surface acting as a very good mirror. As we see in the a later subsection, Section 3.3.6 on page 72, in mirrors the angle of incidence and the angle of reflection are the equal. The angle of incidence in the dense medium that just skims the interface is called the critical angle, θ_c .

This effect is illustrated in the pool problem of the previous section, Section 3.3.2 on page 64. In Figure 3.6 on page 3.6, rays rising from the

bottom/edge, fill all the space above the water. An observer in this space looking down will see the bottom/edge but not at the actual depth. On the other hand an observer in the water sees a more complex situation. Looking down the observer anywhere in the water will see the bottom/edge directly. An observer in the water far enough away from the edge sees the bottom/edge and the sky, maybe more likely the local vegetation, in the same direction. Note that closer to the edge of pool there is no internal reflection and he/she will see only sky or vegetation. In fact, for an observer near the surface, he/she will see almost all the way up to the edge. The situation in an actual pool is even more complex since the surface is rough and changing. The swimmer whose eye is close to the surface sees a rather chaotic mix of sky and bottom.

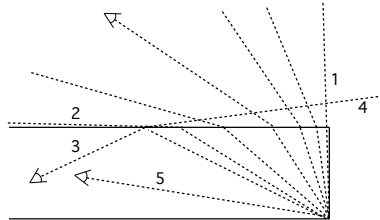


Figure 3.6: **Total Internal Reflection in a Pool** In the pool edge problem of Figure 3.5 on page 67, light leaving the bottom/edge point spreads its fan of rays. Ray 1 is normal to the water air interface and starts the fan. For ray 2, the angle of incidence is just below the critical angle, θ_c and the emergent ray skims the surface. Ray 3 is a ray that has been totally internally reflected. Ray 4 is a ray from the sky that is along the same direction to the underwater observer as ray 3. Ray 5 is from the bottom/edge directly to the observer. An observer above the water looking down can see the bottom/edge from anywhere. An observer in the water looking down can see the bottom/edge from anywhere. An observer in the water, if he/she is far enough away, looking up can see the both the bottom/edge and the sky and coming from the same direction. Which or these is perceived will depend on their relative brightness.

3.3.4 Lenses

We are all familiar with lenses. We have them in our bodies, our eyes. Many of us have them hung on our faces or added to the surface of our eyes to help the working of the biological lens. In all these cases, they are used to bring

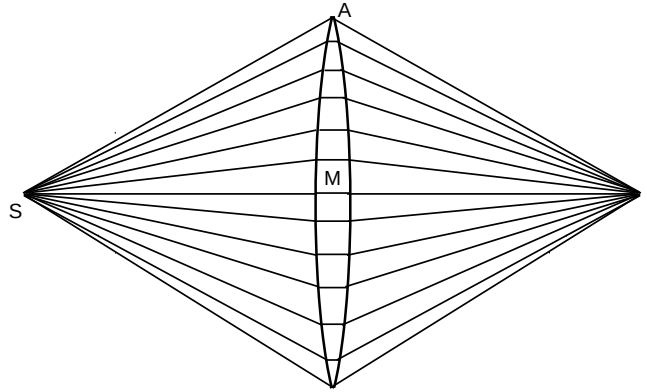


Figure 3.7: **Light in a Lens** The path that goes from S to A to I and the path that goes from S to M to I have the same time. All the rays shown in this figure are between focal points, S and I, and the thickness of the lens is adjusted so that the time for each path is the same.

a spreading beam of light from a source, the object, into a more narrow region and concentrate it at a different place, the image. In the case of our eyes, the deposition of the fan of light starting at a leaf of a tree intercepted by lens is focused on to the retina as an image. Generally all lens are for for imaging or for the concentration of the light energy, focusing, but some lens designs are for the opposite reason, the spreading the light. A lens is another example of a system consisting of two different homogeneous media interacting. The difference with the discussion of the previous section is that here the interface between the two media is curved. The Fermat explanation for focusing or concentrating the light energy is that the glass of the lens is shaped so that all the rays between two points of interest, the object place and image place, that are intercepted by the lens have the same travel travel time and that are on the axis of the lens have the same time of travel, see Feynman [Feynman 1985].

For the configuration shown in Figure 3.7 on page 70, without the lens, the axial ray would be the least time ray and the only one between the points. By placing glass in the path, the time is increased for this ray. In a similar fashion, glass but with a smaller thickness is placed in the way of each of the other paths between the two points in precisely the manner that each path has the same travel time. We will carry out the details of this

computation in a homework assignment. In this case, all the rays that pass through the lens are least time rays. Note how this explains why, when you block a portion of a lens, you do not block a portion of the image but only decrease its brightness. It also explains the concentration of the energy, rays which without the lens would have gone to other points also act at the same point. When we get to the Fresnel/Young/Huygens construction, Section 4 on page 83, we will discover an even more compelling interpretation of the operation of the lens.

3.3.5 Rays in a General Inhomogeneous Space and Mirages.

An inhomogeneous space is one in which, at different places, the light travels with different speeds. In the previous example, we discussed the most trivial example of an inhomogeneous space, two homogeneous media with an interface. In a general inhomogeneous space, the speed of light can vary at each point in the space and you have to calculate the time for the path carefully.

After you select the path, to calculate the time over the path, you must rectify the path and note the different speeds in each segment before adding the times of all the parts, see Figure 3.2 on page 61 and Equation 3.3 on page 61. You select the segments on the basis of the curvature of the path and the rate at which the speed of light is changing. You are working with a certain precision and the length of the segment of path must be the same as that of the straight segment and the speed of light can only vary over the segment within the desired precision.

mirages



Figure 3.8: **Mirages** Due to the bending of light caused by the variation of the density of air and thus variation of the speed of light, to an observer looking down on a hot surface, the ray of light that comes to his/her eyes is not from the road surface but actually comes from the sky.

Mirages are a common experience for Texans. In the summer, the road surface gets extremely hot. A mirage is an example of a phenomena using

the two previous situations, an inhomogeneous space and total internal reflection. When the road is heated to a high temperature from the sun above it, it heats the air immediately over it and that air is thus less dense than air further up. The speed of light in air increases as density of the air decreases. A light ray moving down toward the road surface is moving from a more dense to a less dense medium and is refracted away from the normal. This bends it to larger angles to the normal as it goes closer to the ground and finally reflecting and turning upward. Therefore, for points over a hot road, the least time path is bent upward. This means that when you look down you are actually seeing the sky, and your brain thinks the shimmering blue of the sky is water on the ground, see Figure 3.8 on page 71. The blue spot that you see is shimmering because the less dense warmer air at the bottom is unstable under the dense cool air and the are rising air currents which cause the shimmer.

The opposite effect is associated with looking over a cool surface. An example is with a phenomena known as “ghost ships.” In this case, the cool surface is the ocean in the early hours of the morning when the sun has come up to heat the air over the surface. In this case the temperature profile drops as you get closer to the surface and this bends the light ray downwards and the images of a nearby ship seems to hover in the air.

3.3.6 Reflection and Mirrors

Plane Mirror

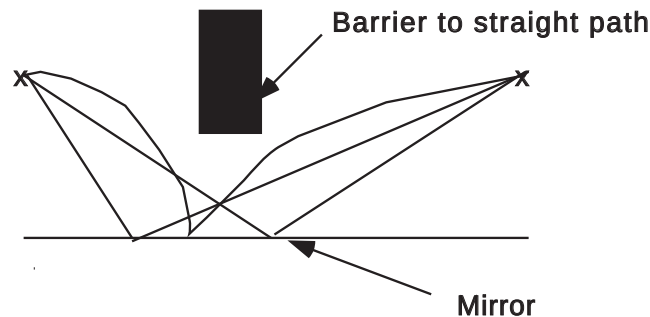


Figure 3.9: **Reflection** Light paths around a reflecting surface. Paths directly connecting the two end points of the paths are blocked by a barrier.

An optical phenomena that appears to be simpler than refraction is reflection. This phenomena is also easily seen to be consistent with the

Fermat's Least Time Principle but, since it was also consistent with the competing particle theory of light, we chose to cover the more complex case first. In the case of reflection, we want to find the light path between two points above a mirrored surface. The trick, in this case, is to realize that we must consider only paths that touch the mirror once. For example, we place a barrier between the points so that the direct path is blocked. The observant student might comment that even with the barrier in place, there are shorter time paths than those obtained by using the mirror. For example those that just graze the edge of the barrier. Why not select these? Later and for different reasons, i. e. diffraction in Section 4.6.3 on page 115, we will. For now though, we will just take it as our definition of reflection that the family of paths under consideration are those that touch the mirror once. Maybe reflection is not that simple after all?

Again, the path is in a region that is homogeneous and, thus, we anticipate that the least time path is the shortest distance path. In this case, you must use the mirror to get past the barrier. What is then the shortest path? Or better said, what is the shortest of all the paths between two points that touch the mirror at only one point? For the simple case of two initial points equidistant from the mirror surface and with a piece of string, it is easy to convince yourself that the path that touches the mirror at the mid point of the interval between the points is the shortest and it, therefore, makes equal angles of incidence and reflection. You should describe how you can use a piece of string to show this.

. Thus:

$$\theta_1 = \theta_2 \quad (3.11)$$

This apparently very simple rule can be used to interpret many interesting situations.

An image is formed when several rays from the same point are brought together by the eye. In addition, the brain extends each set of rays, so that it places the image where the set of rays converge treating them as if they were straight lines. This is similar to the cases that we had above with viewing the bottom of a pool and mirages. Consider the case of two plane mirrors that are perpendicular, using Fermat's Least Time and extending the rays as straight lines, you find three images in addition to the original object, see Figure 3.11 on page 75.

Curved Mirror

Now let's examine the case of a curved mirror. For example, look into a spoon—the bigger and more polished the better. You see yourself shrunk

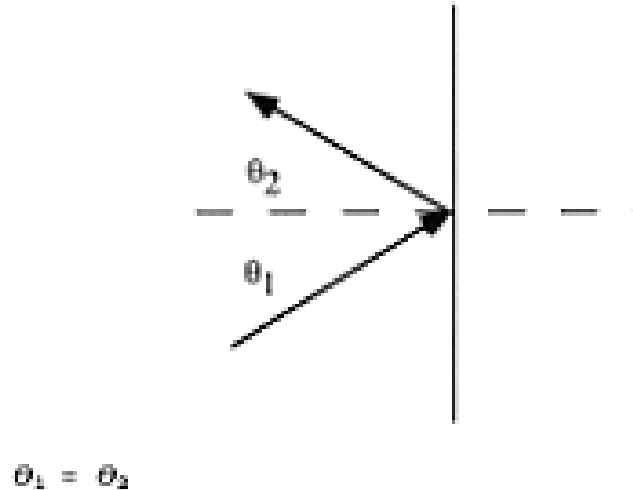


Figure 3.10: **Law of Reflection** For light reflecting from a mirrored surface, the least time ray is the one in which the angle of incidence with the normal is equal to the angle of reflection with the normal.

and upside down. The situation here is the reflection correspondent to the lens. The surface is curved in such a fashion that for the selected points, all rays have the same travel time. Why upside down and shrunk. Look at the light that comes from the tip of the larger arrow in Figure 3.12 on page 76. For this discussion to be strictly correct the arrow though large should be small compared to the mirror radius. The big arrow is you, the object. The rule is simple. At the mirror the angle of incidence must equal the angle of reflection but, since the direction of the normal to the mirror is different at the different points on the surface, different rays reflect in different directions. The three rays that are shown are all least time paths. These three rays are representative and any ray from the large arrow will pass through the point of convergence, the tip of the small arrow. These three are shown because they are particularly simple to describe: using simple trigonometry and the bending of the mirror, it is possible to show that the ray that starts parallel to the axis always reflects so that it passes through the point at $\frac{R}{2}$, where R is the radius of the spherical mirror, from the axis; the ray from the object that goes to the vertex of the mirror reflects so that it is symmetrically located below the axis; and the ray that passes through the point on the axis at $\frac{R}{2}$ from the vertex will emerge on reflection

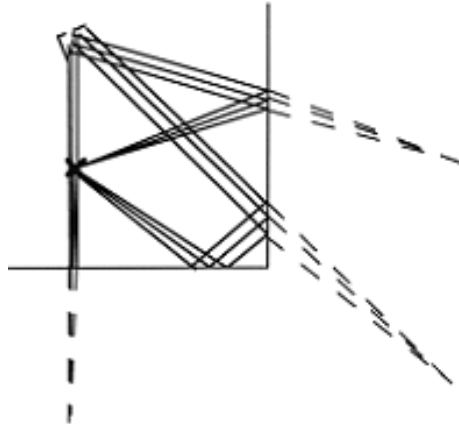


Figure 3.11: **Perpendicular Mirrors** An object viewed from the front of two perpendicular mirrors produces three images. There is the image in each of the two plane mirrors and the image produced by both mirrors.

parallel to the axis. The eye that receives these reflected rays reconstructs the image as the tip of the small arrow, upside down and smaller. Again, the rays that reach your eyes or are all least time rays. This is why when you cover part of the mirror or, in the case of the spoon, really only have a fraction of the mirror, you still see the entire image not a part of it. This is in contrast to the case when you remove part of a plane mirror. In this case, you lose part of the image. We will make more of this later, see Section 4.6.4 on page 117.

The idea of the brain reconstructing the image as the crossing point of the reflected rays reaches its extreme when you move the spoon closer. What happens? Using the same three rays, when the object arrow moves closer to the vertex of the mirror than $\frac{R}{2}$, the image is larger than the object and is upright. You have to get pretty close to the spoon for the image to make much sense and the image is usually too big to interpret as your face. In fact you need a really big spoon for this to work. The interesting thing that emerges from the diagram is the the reflected rays really never cross. Only the extrapolation of the rays beyond the back of the mirror cross. Thus this image is behind the spoon, region where they do not actually go, and is called a virtual image in this case. The name comes from the fact that in the case of the real image there is a point in space where the rays cross. In this case, you can put your finger there and destroy the image. For the virtual image, there is no point at which the rays cross; it was extrapolated

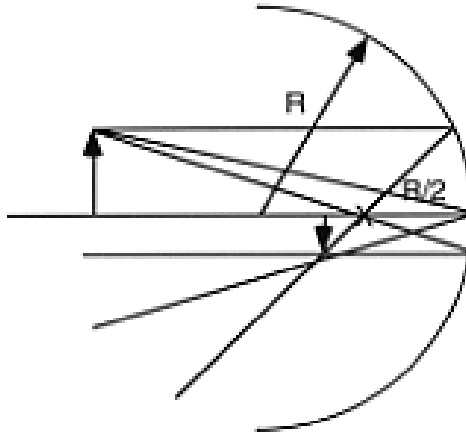


Figure 3.12: **Spherical Mirror** Light rays focusing an image near a spherical mirror. Three rays for which the angle of incidence is the same as the angle of reflection are shown. The axial ray passes through $\frac{R}{2}$ after reflection. A ray through $\frac{R}{2}$ produces an axial ray after reflection. A ray to the vertex produces a reflected ray that is symmetric below the axis. If the object is as shown further than $\frac{R}{2}$ from the vertex the image is smaller and upside down.

in your mind and placing your finger there does nothing to the image.

3.3.7 Mathematical Digression

In our articulation of Fermat's Principle, we casually assumed that it made sense to use the phrase "all possible paths" between two points. In a normal space, that's a lot of paths. To start with does it even make sense to identify "all paths". If you think about it, it means that somehow you produce an ordering so that you can go through the lists to examine all possible cases. An ordering is mapping of the paths onto an ordered set. Without much thought, it should be clear that there are a lot of paths – an infinity, a very large infinity. Are there too many paths to order them like the integers? Two common examples of large sets are the integers for a discrete but infinite set or the points on a line for an infinite but continuous set. The counting of infinite sets is a subtle issue. There are as many integers as there are odd numbers. That's because they can be ordered together – put into a one to one correspondence.

How do you determine the number of paths? You count them or order

them. Counting is a process of matching the elements of two sets, one the set in question, in our case paths, and a given set whose properties are better understood. The smallest of the standard sets of choice are the discrete infinite set that is the number of integers. Sets that have the same number of elements as this are relatively nice to deal with and once an identification with the numbers is established the elements can be manipulated like numbers. Sets of this size are said to be in the class $\aleph_0$. Anytime that you make a table, you are making a mapping between the set of integers and your set of objects that enter the table. If you have an ever larger set of objects, you have a set the size of $\aleph_0$ and you have ordered it with the integers.

In order to use the tools of analysis you need to deal with a system that has the right number of members. Functions are mappings of the real line onto the real line. The real line is, in fact, an example of the next larger infinite set, $\aleph_1$. It is bigger than the number of integers which also happens to be the same size as the number of rationals. The set made of all the points on the real line is the same size as the number of irrationals. A simple ordering argument, we can show that the number of points on a line and the the number of points in a plane are the same. Again, an example of a property of these infinite sets that is not intuitive. In other words, there are as many points on one line as on any countable number of lines.

It is relatively straight forward to convince yourself that the number of paths is larger than the number of points on a line. This makes for a problem. Most of what we can do in analysis is dealt with through functions. By definition, functions are mappings of the real line on to the real line. Thus, our manipulations with paths cannot be considered functions and all the things that we learned about the manipulation of functions does not hold. Mappings of path space onto the real line are called functionals and thus our ambition of finding the least time as a function of path is a functional.

In our first example, refraction, we used our intuition to label the paths as the same as the point of intersection of the path with the interface of the media. This is clearly only a small sample of all the paths. The important point about our selection of the point of intersection was not only for convenience, it was a reduction in the size of the path space to one that allowed it to only the same number of paths as the points on the real line. This choice allows us to write the time T as a function of x , a point on the real line, $T(x)$. Thus although it is nice to think of x as the distance along the interface, its real role is as a label in path space and one that is only $\aleph_1$. This makes T , a real number, into a function in the sense that it provides a mapping of x onto the associated time for the path. We are then free to use usual mathematics to find the minimum.

Reiterating, in general, time over the path is it not a function of the path, because the number of paths is greater than the number of elements in $\aleph_1$; there is no rule for matching paths with points on the real line. The number of paths can be quite large and, using some other information such as our intuition, depending on the restrictions that you put on the family of paths that you consider, the family will be in some class, $\aleph_i$ where $i \geq 2$. In these situations, you cannot call T a function. It is called a “functional” instead. That is to remind us that the ordinary procedures of mathematics are not adequate. Thus this very simple algorithmic looking rule,

$$T = \sum_{path, (x_0, y_0)}^{(x_f, y_f)} \frac{\Delta s_i}{v_i}, \quad (3.12)$$

is actually a complicated mathematic structure. For us, being straight forward people with a simple outlook on life, we will ignore most of these complications and go ahead and, in all our cases find a family that is $\aleph_1$, when we are operating in path space. In other words, we will select some small class of paths and label them one or more intervals on the line. In this way, we reduce the functional to a function.

The other interesting mathematical feature of this supposedly simple algorithm is the need to evaluate a complicated object. These are the problems of sensibly rectifying path either because of curvature or the variation in the speed as the path moves through points in space. These issues were discussed earlier in Section 3.2 on page 57 and Figure 3.2 on page 61. The point is that, although it is often the case that a rule for interpreting a phenomena can be stated simply, there are often subtle issues that require a great deal of mathematical development to disentangle. Much of modern mathematics is devoted to the untangling of what appears to be on the surface very simple physics problems.

Finally, it is important to make clear what we are doing drawing the path the light travels over, the ray. Firstly, the attitude that this is the light gives us the ability to make the theory falsifiable in the Popper sense of falsifiability⁴. Put an obstacle on the path and the light no longer goes to the other point. It is stopped by the obstacle. The path, the ray, is the light. It is the ‘reality’ of the light. It is what light is made of. If block any point not on the ray we have not effected the light. It is a wonderfully simple thing that is the light. The way we measure light is it brightness.

⁴In his famous book *The Logic of Scientific Discovery*[Popper 2002] Karl Popper argues that science should adopt a methodology based on falsifiability, because no number of positive experiments can prove the truth of a hypothesis.

Brightness is a positive definite quantity and so is the number of rays. The concept of light as rays leads easily to the idea of geometric fall off⁵ As we go further into our discussion of light, we will find that although, as a convenience, we draw a line connecting places and call it light. The line will have to become a much richer construct until we no longer call it a line but will call it a propagator which is a thing with much richer embellishments required to account for the observations. A simple example is that a line is either there or not, a one or a zero. We will persist in drawing lines when we need richer constructs.

3.4 Newton and Color

It is a common experience to use a piece of shaped glass, a triangular cut of glass called a prism, to produce a rainbow of color from sun light. This is commonly described in the following way: this is basically a refractive phenomena and a simple extension of Fermat's Least Time Principle can be used to describe it. A narrow beam of white light incident at a non-normal angle on one surface of the glass is refracted; the beam changes direction. The spread of color appears because the different colors in the light have different speeds in the glass with the blue being faster than the red and all colors slower than for light in air, see Section 3.3.2 on page 64. Thus the blue is bent less than the red. The separated rays then emerge from the other interface of the glass spread in this familiar rainbow pattern. This spread of color can be seen by placing a piece of paper after the second interface, see the first part of Figure 3.13 on page 80.

Actually Fermat did not describe the phenomena of color. Among the early studies of color and the best were carried out by Newton in a series of experiments over many decades that he brings together in his treatise called "Opticks", [Newton 1730]. Although his most famous presentation of physics was in his monumental work "Principia" his "Opticks" could be considered his best or even the best physics book ever, Newton developed an interpretation of the nature of light and its relationship to color phenomena. The beauty of the book is that as opposed to the "Principia" it deals with the observed phenomena directly and does not develop an underlying structural basis "explaining" light's color behavior. Interestingly, it was Newton, the advocate of a particulate theory, who first articulated the ideas about the white light being composed of the colored components. Prior to Newton's

⁵Light from a local source, a point source, gives less brightness at greater distances. The rays spread out.

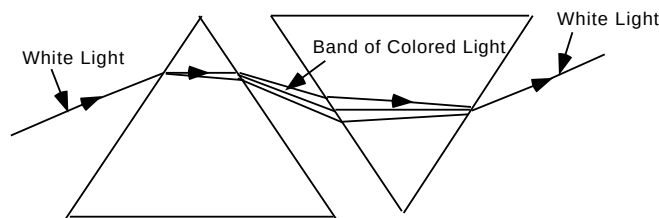


Figure 3.13: **Newton's Experiment with Light and Color** Newton's experiment showing that light is composed of colored components. A narrow beam of light is incident on a prism and produces a broadened and colored band which can be reconstituted back into a narrow white beam of light with a second prism.

interpretation, the idea was that the color in the prism came from the glass and was not an intrinsic property of the light. To show otherwise, Newton placed two prisms, in the path of a narrow beam of sunlight. The beam emerging from the first prism was traveling in a different direction from the original beam, as expected from refraction theory, either by particulate or least time principles. As usual, the beam was spread over a band of angles and, when a piece of paper is located in the beam, after the first prism, a broad smear of light appears and the different parts are a different color, the rainbow alluded to above. Newton then went one step further and inserted the second prism and allowed the spread beam to enter it. When arranged carefully, he found that this reconstituted the original beam in the original direction, see Figure 3.13 on page 80. Newton's interpretation was that the color was intrinsic to the light and; in other words, white light has constituents which we perceive as the colors; and the bending in the glass spread the constituent parts differentially to spread the beam. The same process reversed was then able to reconstitute the beam of white light.

In the Fermat least time approach, the blues travel in the glass at a faster speed than the reds and thus the blue colors are bent less by the prism,. In the particulate theory, the blues would travel in the glass slower than the reds. This is an example of a phenomena the is consistent with both interpretations but for different reasons. This difficulty could not be resolved until it was possible to measure the speed of light in materials.

Regardless, it was Newton who realized that white light was a complex phenomena and that white light was composed of an internal structure – the colors. This realization had immediate and important impact in the interpretation of visual phenomena. You saw different colors because, by

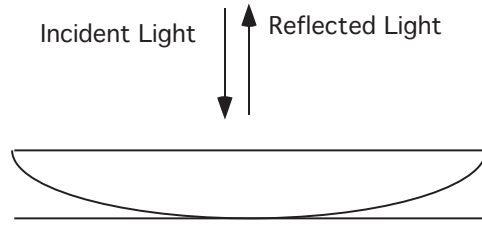


Figure 3.14: **Newton's Rings** Newton's set a curved lens on a plane glass surface and illuminated it from above. When viewing the reflected light from above, there is a series of rainbow colored rings surrounding a central dark spot.

some mechanism, you removed from the white light the other colors or you created color by combining various components such as red and blue to make purple.

It is important to realize that all of this science of color is independent of our modern interpretation of color as the frequency of the oscillations of the light. That came later although the seed had been planted by another observation of Newton. In another experiment, Newton placed a small lens on top of a plane glass surface and illuminated the combination from above, see Figure 3.14 on page 81. Viewing the reflected light from above, there are a series of concentric rainbow colored rings around the central dark spot. This is a direct indication that the color label of the components of the light can be associated with another feature of the light – length. The idea is that because of the curved surface of the lens at different distances from the center, the light, reacts differently depending on the color. Moreover, this phenomena is periodic with the reds repeating as multiples of the gap between the lens and the plate varied. This implied that there was associated with color some sense of length, i. e. the different colors fit the varying places better. We, of course, realize that having a length and speed, the speed of light, is equivalent dimensionally to having a time. The light has components and these are labeled with a time. With the full development of the wave approach of Thomas Young and Christian Huygens, see Section 4 on page 83, color became identified with a very specific interpretation of the time, the time label was the period repeat of the wave at any place, the frequency.

Thus Newton described the basic phenomena of color: White light from the sun or most other luminous bodies is a composite system. There is an internal constituent that is recognized as the color. The different colors can be labeled with a continuously varying parameter, an element of an $\mathbf{R}^1$, that

represents a periodic structure, repeats, as manifest in the Newton's Rings experiment shown in Figure 3.14 on page 81 . The color identification can be characterized as a length peculiar to that color called the wavelength λ or, since light is identified with a particular speed, the characteristic length can also be identified with a time, T , which is now identified more often with its inverse, a frequency, $f = \frac{1}{T}$. Regardless of the interpretation, the length label, λ , and time labels, T or f^6 , that designate the color are connected by the speed of light,

$$\frac{\lambda}{T} = v_{light}, \quad (3.13)$$

which could be slightly different for the different colors, labeled by either λ or T , and for different media. We cannot understand this variation in the velocity until we understand better how the light interacts with media through which the light moves. In a vacuum⁷, there is no medium and thus no interaction with and the speed of light is the same for all colors. Air is such a tenuous media that for all but the most precise measurements light shows no effect on the speed making it difficult for people of Newton's time to observe these phenomena in air.

⁶There is an entire vocabulary associated with labeling periodic entities and even idioms in usage. The frequency, f , as defined above is the cycle frequency with units $\frac{cycles}{sec}$ or in SI Hertz. This is the common engineering unit. For physicists, the preferred frequency is the radian frequency, ω or $\frac{2\pi}{T}$, which has units of $\frac{radians}{sec}$. This is preferred since the argument of a harmonic function should be a radian.

⁷In modern interpretation, the vacuum is a rather richly structured entity whose properties effect the behavior of the things passing through the vacuum. Chapter 22 on page 483 and Chapter 23 on page 491 discuss these issues.

Chapter 4

Fresnel/Young/Huygens Theory

4.1 Recapitulation of Fermat's Least time principal

It really works. It tells you how the light moves between two points. It says that the light moves over a certain path. If you block that path there is no light, see Figure 4.1 on page 83.

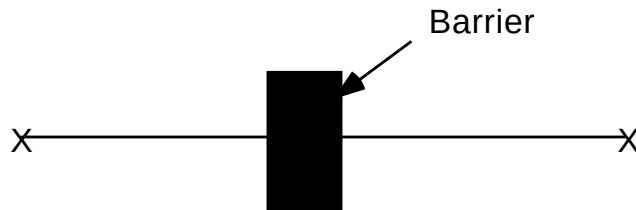


Figure 4.1: Blocking the Least Time Ray If you have light traveling between two points in a homogeneous medium, Fermat's Principle predicts that the light travels over the straight line connecting those points. If you now place a barrier between them you get no light at the second point

So, was it a satisfying theory? This is really a bad question. The only true test of a theory is its agreement with observation. In this case, the predictions were in concordance with experiment, as known at that time. There were competing theories— Newton's particle theory for example — and these were also in concordance with experiment at the time. Later experi-

ments measuring the speed of light in dense matter showed that the particle theory was untenable or, at least, Newton's particle theory¹. For example, for a light beam moving from air to glass at some angle, the light bends toward the normal but, for that to be the case, the particle theory requires that the speed be greater in the glass than in the air. This is inconsistent with the measured speed of light in the dense medium. As stated earlier, this is an often cited example of falsifiability, see Section 3.3.2 on page 64 and reference [Popper 1973]. You should realize that Newton could have inferred from his corpuscular theory that the more dense the medium the slower the particles move. We will return to this issue later, Section 4.4 on page 88.

Fermat's least time was also a "satisfying" principal. It was based on an extremum statement. Of all possible paths, the light takes the *least* time. People seem to like theories based on extremum statements – least effort, most money, greenest grass². . . Yet, we know that light, whatever it is, does not itself measure time over all possible paths and then select the shortest path as the one to take. This would be anthropomorphizing the process. You can argue, and we will, that there must be an underlying theory will provide a description of light propagation that leads to this least time path. This will be the Fresnel/Young/Huygens Theory. It is important to emphasize that, regardless of how successful we are at finding a new theory that "explains" the old theory, no matter what we do, the rule is to find how the light travels not how it decides to travel. You give me a physical situation and I will tell you what I do to tell you what will happen; our algorithms are not the same as "light's."

Another feature of Fermat's theory is that it is a consistent theory. It is a global theory in the sense that to find the least time path, you need to compute the time over the entire path. It is also the case though that, once you have the least time path between two points, if you pick two points within the path and find the least time path, then for those two points on that path, the least time path between them is the segment of the previous path between the points. In this fashion, Fermat's theory is also local in that all subsegments, no matter how small.

There is another fashion in which Fermat's Theory is local. In Fermat's theory, light travels only over the least time paths and, therefore, only samples the places that are on these paths. If you block a place that is not on

¹Our current theory of light has as its basis the idea that light is composed of things called photons which are particles but with curious properties. See Section ?? on page ??

²There is also a technical reason why physicists prefer extremum rules. If the space of labels is smooth in label values, you can show that the extremum value is unique.

the least time path, it does not change the light. In this theory, the light is a local phenomena; it does not sample the whole space. It is where the least time ray goes. You may ask how the light knows which path is the least time path without going everywhere to find out but, as stated above, that is not the issue. Granted that when we find the least time path, we calculate the time for all the possible paths; therefore, we have to know the state of all the points in the space. Thus we have to have global information. I emphasize that is how we find where the light travels. We still say that the light is on the least time ray.

4.2 Problems with Fermat's Least Time

Fermat's Least Time is a wonderful theory for many applications. It is still used today to do lens design and in acoustics design. The problem with Fermat's theory is that there are situations that it does not describe well. There are cases in which you actually do find light between points that, according to Fermat's Theorem, should be dark. Let's look at one simple example. In a homogeneous space, the light travels in a straight line between the two points. Now place a barrier between two points. Fermat's Theory would argue that the light should be blocked and no light should be seen at the second point. The problem is that in a very sensitive experiment there *is* some light that goes between these points. It is important to note that this light is not as bright at the point as the original unblocked case. For this reason, it was some time before it was actually discovered! In other words, given the technology of the time, Fermat's Theory worked for what was known. This light, seen where it was not supposed to be, is a phenomena called diffraction. It was a phenomena that was not predicted by Fermat's Theory and, even after its observation, was not accommodated by it.

The fact is that we now know that light does not strictly follow Fermat's rules and restrict itself to least time paths. Light can travel 'around' an obstacle, a phenomena called diffraction, In some sense, as we will find, the light is all over the place. It propagates according to a rule that was first stated (although not too clearly!) by Huygens. Huygens' approach was made more concrete by Thomas Young and later applied to diffraction by Fresnel. The Fresnel/Young/Huygens's theory not only explained effects such as diffraction, it also explained why Fermat's Theory worked. Why for so many cases the light, although light sampled the entire space, in the appropriate circumstances, it was concentrated around least time paths.

This is a very typical pattern in the development of a theory. A new set

of phenomena are observed that are not predicted by the standing theory; consequently, a new theory emerges that encompasses the old theory along with all of its successful predictions *and* contains the new phenomena. In the case of light, there was an indication of the nature of the new theory available but the statement of its rules were not articulated well enough to provide a true test of the theory.

Diffraction is a rather complex phenomena. Although the first discussions of the new approach by Huygens was a attempt to describe the situation in diffraction, it was Thomas Young's description of the interference, a closely related phenomena, that is easier to understand and we will treat it first and then extend the discussion to diffraction.

4.3 Huygens

The Huygens theory is based on a new physical construct called the "field". The field concept developed over a long period of time, reaching its clearest articulation in the period following Maxwell in the later part of the 19th century. It will be discussed in more detail in the Chapter 5 on page 123 and in particular Section 5.2.2 on page 126. In reality, the field concept has continued to develop well into this century as the ideas of quantum mechanics and relativity are combined with field theory. We will look at all these issues, see Chapters 19 on page 403 through Chapter 22 on page 483. For now though, all we need to say is that a field is a distributed entity; it is defined at every point in space and has a dynamic, a rule for its time development. In other words, it is quantity, call it ϕ that depends on the spatial coordinate and time, $\phi(\vec{x}, t)$ and there is an accompanying set of rules for the spatial and temporal development. For details see Section 5.2.2 on page 126.

Huygens's ideas about the way that light behaved was based on his interpretation of how sound and disturbances on the surface of water, waves, behaved. As we all know, water does not pile. Therefore, when a water surface is disturbed and you get a small pile of water somewhere, the surface of the water wants to return to the level of the water around it. The interesting thing is that when you make this disturbance in the surface, a copy of it moves away from the source location and spreads over the surface for a fair distance without changing shape appreciably. Sound had a similar interpretation. A small area of compressed air, high density, tries to return to the density of the surrounding air. This compresses the nearby air and this then compresses the air nearby. Periodic disturbances of air propagate

long distances. In fact, the idea of propagation in a distributed medium is ubiquitous.

Huygens used the intuitive idea that each part of the disturbance acts like a new source of disturbance and that subsequent configurations can be found by allowing the disturbance to spread by ‘adding’ the many sources, see Figure 4.2 on page 87. If you know the configuration of the disturbance at any time, subsequent patterns can be found by allowing the current disturbance to act as a series of new sources and then allow the disturbance from these sources to spread and the consequent form of the disturbance is given by adding the effects of each of the new sources. This pattern is repeated until the disturbance spreads throughout the space.

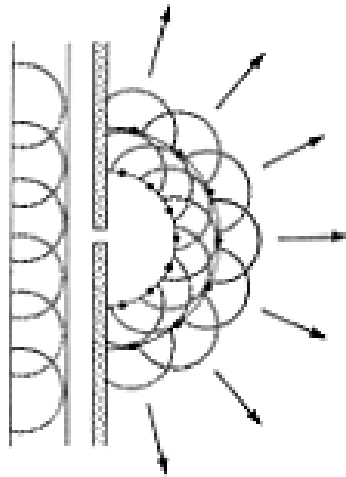


Figure 4.2: The Huygens Construction The form of a disturbance at some later time is determined by using the present configuration and using each part of the present disturbance as a source for future disturbances. From each of the sources that are the present disturbance, a new wave spreads. The new disturbance is generated by ‘adding’ the these new disturbances. To Huygens, ‘adding’ was the process of forming the envelope of the new sources. This process is repeated as necessary to find the final configuration of the field. The figure shows a disturbance prior to reaching the slotted screen and two levels of construction subsequent to the slot.

The field is manifest here in the sense that at all points in the space there is the potential for disturbance. What is this potential for disturbance? Following Maxwell, see Section 5.4 on page 140, we now know that for light

the disturbance that is ‘added’ is the electric field. At the time of Huygens, the idea of the field had not yet been well articulated and all forces were action at a distance, see Section 5.2.1 on page 125. The question at the time of Huygens was what was the disturbance. The only observable consequence of light was its brightness. As it relates to our particular application to light then is this potential for a light related to the brightness? This would be the obvious solution and is probably what Huygens thought. It took another step to get to a working algorithm that worked for light and, in fact, all other wave phenomena. Thomas Young realized the brightness was only a positive quantity and that there had to be some more elemental entity that could take on both signs but still itself not be the observable quantity, the brightness or intensity of the light. He postulated that there was an amplitude for light at every point in space. This is what it is that is being added. The brightness of the light which we see is the square of this amplitude. If you have an amplitude value at some point, the amplitude around it is the superposition of all the earlier amplitudes. In other words, it is the unmeasurable quantity, the amplitude for light, which is the causal element and has its effects added. Then you find what you see by squaring this amplitude to get the brightness at a point.

This is what lead to the idea of light as a wave and holds true for all kinds of waves. How Young got here is the subject of the next section.

4.4 Thomas Young and Interference

As we try to understand exactly what is going on, let’s look at a phenomena called interference. The best and simplest situation in which to observe this phenomena is with an apparatus called the Young’s Double Slit, named for Thomas Young. Monochromatic light from a single concentrated source falls onto an opaque screen with two narrow but long slits and the image is then projected onto a distant viewing screen, see Figure 4.3 on page 89. The long direction of the slit is perpendicular to the plane of the figure.

On the distant viewing screen, you see a series of bright and dark bands in the color of the light, see Figure 4.4 on page 89. The variation in brightness is very striking. It varies very rapidly as you move across the screen.

It is even more striking when compared with the pattern that you get when you cover one of the slits. With only one of the slits open, the screen is uniformly illuminated. There is no sharp shadow. This is because we are using a very narrow slit and this pattern is explained by diffraction which we will deal with later, see Section 4.6 on page 106.

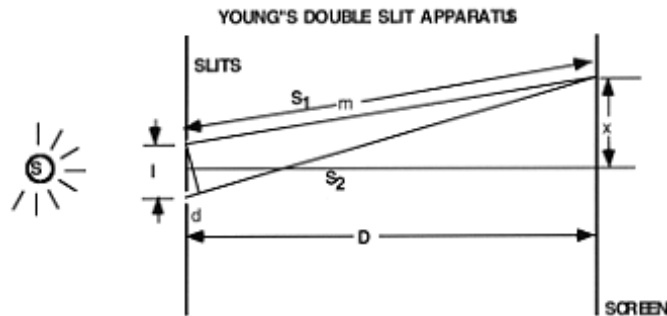


Figure 4.3: **Young's Double Slit Apparatus** Light of a single color, called monochromatic light, shines on an opaque screen with two narrow slits. The light subsequently passes through to a distant screen on which the brightness can be observed. s_1 and s_2 are the distances from the slits to a general point on the screen and $s_2 - s_1 = d(x)$. In the approximation that D is large compared to all the other lengths, the two triangles shown are similar and thus $\frac{x}{D} \approx \frac{d}{l}$.

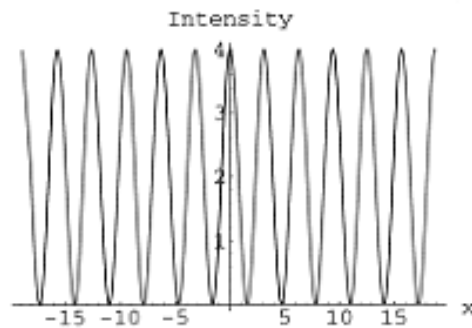


Figure 4.4: **Brightness of Light on the Screen in a Young's Double Slit Apparatus** The variable x is distance measured as you move across the screen from the point midway between the projection point of the slits, see Figure 4.3 on page 89. The light varies from very bright in the center to dark and bright very rapidly.

For now, we can understand what is happening qualitatively from the Huygens construction. For a very narrow slit, the light at the slit acts as a new source of light which spreads out a spherical wave. We know enough dimensional analysis to know that we have to say what “narrow” means.

We need a length in the problem. Obviously, this comes from the light and is the wavelength, the speed of light divided by the frequency of the light. The slit must be comparable to the wavelength to be considered narrow. Regardless, the illumination in the single slit case is shown graphically in Figure 4.5 on page 90.

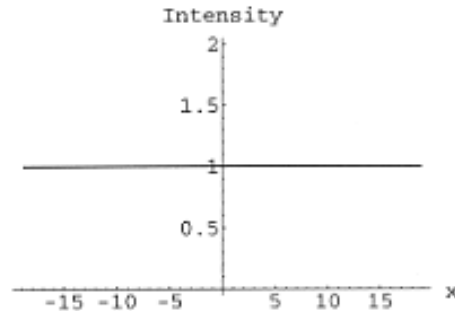


Figure 4.5: **Brightness of light on the screen in a Young's Double Slit apparatus when one of the slits is covered** The variable x is measured as you move up the screen from the point midway between the projection point of the slits, see Figure 4.3 on page 89. In this case, the illumination is uniform over the screen. The intensity scale in both this figure and in Figure 4.4 on page 89 are the same.

Note that, besides the rapid variation in brightness of the light in Figure 4.4 on page 89, the bright places in the two slit case are much brighter than the single slit case – four times as bright.

How do we describe what is happening? First, let's look at the point on the screen that is located at the mid-point between the slits, $x = 0$. By symmetry, it receives the same "thing" from each slit. Thus at this point, your expectation is that the brightness will be twice what it is from a single slit; there are two slits and the light from each add. Instead it is four times as bright. What adds at the screen from each slit is not the brightness. There must be a new entity that adds and brightness is constructed from it. In a sense, what we need is a new physical entity that is causal in its agency. The effects from each source, the slits, are added but the measured "thing", the brightness is constructed from it. The method for constructing the brightness is obvious, square what is added. The evidence is that what comes from the slits which, when two slits are open, must be twice what comes from either slit independently and must be squared to give the observed brightness, 2 squared is 4.

Besides the action of this “added” entity in producing the bright spot at the central point, it must account for the dark spots. Since brightness is purely positive, there is no way to add two brightnesses and get zero. In other words, the dark spots emerge because the thing being added can be positive or negative and the addition of these from the two slits can add to zero at special places. The idea squaring to get the extra brightness at the symmetric place is also consistent with positive nature of brightness.

How do we attach a mechanism to this observation? Young who was English was still and adherent to the corpuscular basis for light. His explanation required the corpuscles to vibrate and interfere with each other. It took quite an interesting describe what is happening? The light leaves each slit and travels over the path designated by Fermat to the screen. One difference between Young’s idea and Fermat’s theory is in what we have traveling to the screen. In Fermat’s theory the brightness, or intensity, is what is traveling. In Fermat’s theory, you only have brightness and with the addition of the observations of Newton about color which, of course, is associated with the frequency by the non-Newtonians.

In the Young/Huygens theory, light has an additional attribute, called the amplitude. We now associate it with a combination of the electric and magnetic fields, see Section 5.4 on page 140. This is a rather complex construction and we do not have to deal with most of these complexities now and certainly Thomas Young did not. To Young, the amplitude which in contrast to the brightness could be positive or negative was the light. It was the causal agent of the light and effects could be added but it was not directly observable, the brightness which was its square was. The brightness at any place, the measurable quantity, is determined in a two-step process, find the amplitude, the causal agent, the thing that added in its effects, and then square this to find the brightness.

The detail description of what is happening in the experiment goes like this, see Figure 4.3 on page 89 . At each slit, there is an amplitude created there by the bright source of light behind the opaque screen with the slits. Since the light at the slit was monochromatic, the amplitude varies harmonically with a definite frequency and is the same at each slit, see Section 3.4 on page 79. This amplitude is the thing that then travels over the path. The amplitude along the path is the same as the amplitude at the start but the time for the variations associated with the color is effectively delayed by how long it took the light to get from the slit to the screen. This is the Huygens part of the construction. At any point, the amplitudes from all the sources, in this case the two slits, add to make a net amplitude, and the intensity or brightness is then found from this net amplitude by squaring.

This clearly works for the spot on the screen at the projection of the middle of the slits, the symmetric point on the screen. Because the slits are identical and the opaque screen is illuminated uniformly, the amplitudes at the slits are the same. At the middle point on the screen the amplitudes from each slit is the same since both paths from the slit to the screen are the same length. The the amplitudes from each slit at the screen are the same and the net amplitude is twice what the amplitude of either slit alone. The brightness is thus four times what you get from one of the slits alone.

What about the oscillating pattern of bright and dark at other points as you move up the screen? We saw from our analysis of the central point that the bright spots were where the amplitudes from the two slits were the same and adding produced a net amplitude that was twice either one alone. The dark places must be places at which the two amplitudes had the same magnitude but opposite signs. How does the amplitude vary in space and time in order to make this work? Here, Young uses the combination of the time variation associated with color and the shift in time associated with a finite speed combined with distance. At points on the screen other than the central point, the travel distances from each of the two slits are different, s_1 and s_2 in Figure 4.3 on page 89 and thus at a generic point on the screen, the amplitude from the each of the two slits are related to the amplitude at the slits at different times. This is basis for the little clocks that Feynman discusses in QED, [Feynman 1985].

In summary, the idea was that light travels over Fermat Least Time paths but the important point is that it is the amplitude that travels and that is added. Not only that but the amplitude is intrinsically oscillatory and the frequency of the oscillation is identified with the colors. This oscillatory amplitude is the causation agent in light. It is not directly observable but its square is. This constitutes $\frac{2}{3}$ of the Fresnel/Young/Huygens construction.

Wait a minute. If light is intrinsically oscillatory, why don't we see it turning on and off? The trick is that the observable entity, the intensity, is the square of the amplitude and all or our sensors of light cannot detect the brightness instantaneously but only averaged over a time period that is long with respect to the period of the light. We will show how this works in the next section, Section 4.4.1 on page 92.

4.4.1 Detail Analysis of Interference for the Double Slit

The previous discussion of the operation of the Young's Double Slit was qualitative and to make progress, we will have to become more quantitative. In this section, I will do a rather tedious analysis of the Young experiment.

This will show the true nature of what is going on but also motivate the introduction of phasors, a technical tool that makes the analysis simpler. These phasors are the little clocks in Feynman's QED, [Feynman 1985].

Light of a given color is an intrinsically oscillating system. The different components of light identified as colors by Newton are the different frequencies, f . f is what is called the revolution frequency and is used by engineers and the units are cycles per second or Hertz. In this sense, each color of light is identified with a certain time period, $T = \frac{1}{f}$ or a length, the wavelength, $\lambda = cT$ where c is the speed of light. Physicists prefer the radian frequency which is $\omega = 2\pi f$. In addition, light is represented by an amplitude which is the element 'added' by Young. What you see and can measure is the square of that amplitude. In other words, the amplitude associated with light travels over the least time path between two points and, as it travels, it carries a little clock whose angle of advance of its arm is the travel time for that segment of path divided by the period of the oscillation for that color of light; $\Delta\theta_i = 2\pi \frac{\Delta t_i}{T}$, where t_i is the travel time for light in segment i and T is the period of the light.

Let's apply this analysis to the double slit experiment described above and shown in Figure 4.3 on page 89. There is a constant level of brightness at each of the slits. If we put a screen there we will see a brightness $I_{observed}$. Yet, we know that we want the light to be oscillatory. We accomplish this by saying that the light has an amplitude at slit i where i is 1 or 2 and that $A_i(t)$ that varies harmonically with time, $A_i(t) = A_{i0} \cos(\omega t)$, where A_{i0} is the maximum value of the amplitude, $\omega \equiv \frac{2\pi}{T}$, and T is the period for light of that color. Unfortunately, the coefficient " A_0 " of the harmonic factor is also often called the "amplitude" of the harmonic signal. It should be clear from the context which amplitude is which.

What we measure at any place is the brightness: $I = A^2(t) = A_0^2 \cos^2(\omega t)$. But this oscillating brightness is not what we see. We see a steady brightness. The resolution of this difficulty is to realize that on the time scale that we view the light, the light passes through many periods. From the list of "Things that Everyone Should Know", Section 1.4.2 on page 16, we see that the wavelength of visible light is 6×10^{-7} m or that the frequency is 5×10^{14} sec⁻¹ or a period of 2×10^{-15} s. Thus if the time resolution of the eye is millisecs, we will see the average of tens of millions of cycles. Thus, the brightness we see is the long time average of many periods of

$$I_{observed} = \langle I(t) \rangle_t = \langle A^2 \cos^2(\omega t) \rangle_t,$$

where $\langle \rangle_t$ means to take the time average³ of the quantity inside the brackets.

This average is easy to compute if we remember a little bit of high school trigonometry⁴. Realizing that the time average of $\cos(2\omega t)$ for several periods is clearly zero, we get that

$$I_{observed} \equiv \langle A^2(t) \rangle_t = \frac{A_0^2}{2}.$$

Another way to see this is by plotting $\cos^2(\omega t)$ for many periods to see that the average is $\frac{1}{2}$. Thus when we see steady brightness of $I_{observed}$, it comes from an amplitude of $A(t) = \sqrt{2 I_{observed}} \cos(\omega t)$. If you could struggle your way through that discussion and keep track of the two amplitudes, $A(t)$ and A_0 , you are on solid ground.

Going back to our apparatus, if we arrange it so that the amplitude at slit one is

$$A_1(t) = A_{10} \cos(\omega t) \quad (4.1)$$

where, as explained above, A_{10} is determined by the brightness of the light at slit one.

Let's consider the case that slit one is the only one open. The amplitude at the screen at a given time t is the original amplitude at slit one delayed by the time it takes the light to go from the slit to the screen. In other words, the amplitude of the light at the screen at time t is the same as the amplitude of the light at the slit at time $t - \frac{s_1}{c}$ where s_1 is the distance between the slit and the screen and c is the speed of light. This part of the calculation is a combination of Fermat's Least time, a straight path, and Huygens' construction. The addition of Young is the use of the amplitudes for the Huygens' construction and not the brightness.

This means that the amplitude on the screen from slit one alone is

$$A_{1 \text{ screen}} = A_{10} \cos \left(\omega \left(t - \frac{s_1}{c} \right) \right). \quad (4.2)$$

This result is not as trivial as it seems. Let's cast it in slightly different form.

$$A_1(t) = A_0 \cos \left(\omega t - \frac{2\pi \frac{s_1}{c}}{T} \right) = A_0 \cos \left(\omega t - 2\pi \frac{s_1}{\lambda} \right) \quad (4.3)$$

where I have used the fact that $\omega \equiv \frac{2\pi}{T}$ and that $cT = \lambda$. From the ωt term, we see that this is an amplitude that oscillates with a period T so that

³The time average over the interval 0 to T for a function of time is defined as $\langle f \rangle_t = \frac{1}{T} \sum_0^T f(t) \Delta t$ or $\langle f \rangle_t = \frac{1}{T} \int_0^T f(t) dt$

⁴The required relationship is $\cos^2(\omega t) = \frac{1}{2}(1 + \cos(2\omega t))$.

this is, in fact, still light and that the color of the light at the screen is the same as the color of the light at slit one. The only difference is that there is an extra time independent term in the argument of the cosine function. All this does is shift the the argument that goes in at the start. Again, since this signal varies so rapidly that our sensors can only see the time average over many many periods, this starting angle is not detectable. Since this shift is the only factor that changes as you move to different parts of the screen, the brightness at the screen is uniform. These extra time independent terms will become important later on in this analysis. The argument that goes into the cosine function is important and given a name. It is called the phase. This same terminology holds for sine functions and, to get all the terminology out, the pair of functions, cosine and sine, are called the harmonic functions.

If you choose to have only slit two open, you would have a similar situation at the screen from slit two. Since the two slits are located symmetrically relative to the source, the amplitude at slit two is the same as that of slit one and thus the amplitude at the screen from slit two alone would be

$$A_{2 \text{ screen}}(t) = A_0 \cos \left(\omega \left(t - \frac{s_2}{c} \right) \right) \quad (4.4)$$

where I have used the fact that, at a general point on the screen, the two distances, s_1 and s_2 , will be different. Again, this by itself produces an illumination that is uniform and the same color as the original light. Note that if you have just one of the slits open, say slit one, the intensity is $I_{\text{observed screen}} = \langle A_{1 \text{ screen}}^2(t) \rangle_t = \frac{A_0^2}{2}$. As before, the brightness is the time average of the amplitude squared. We can use the observed I to find the appropriate A_0 .

What happens when both slits are open? The net amplitude at the screen is sum of the two amplitudes from the slits as if they operated independently. This is the point of the fact that the amplitudes of independent sources ‘add;’ the amplitudes are the causal agents. The amplitudes are the fundamental causal agents. They carry the information about the slits to the screen. This process of adding independent sources as if the other was not present is called superposition. This is not the first time that we have used superposition. Our discussion in Section 6.1 on page 153 and shown in Figure 1.1 on page 7 about forces, treated each force as if the other bodies were not present, the force F_{12} is the force on body one from body two independent of the presence of the other bodies.

In terms of our situation here, the amplitude at the screen is the superposition of the amplitudes from the two slits. You do not add the brightness;

they do not superpose. Thus we have:

$$\begin{aligned}
 A_{T_s} &= A_{1 \text{ screen}} + A_{2 \text{ screen}} \\
 &= A_0 \left(\cos \left(\omega \left(t - \frac{s_1}{c} \right) \right) + \cos \left(\omega \left(t - \frac{s_2}{c} \right) \right) \right) \\
 &= 2A_0 \left(\cos \left(\omega \left(\frac{s_1 - s_2}{2c} \right) \right) \cos \left(\omega \left(t - \frac{s_1 + s_2}{2c} \right) \right) \right) \quad (4.5)
 \end{aligned}$$

where I have again used a trig identity⁵ from high school to add the two cosine functions and A_{T_s} is the total amplitude at the screen. Note that s_1 and s_2 depend on the position on the screen.

In this case, there is an oscillating signal at the screen. This is the because of the term $\cos \left(\omega \left(t - \frac{s_1 + s_2}{2c} \right) \right)$. Again, this is light of the same color with a position dependent phase that is not observable for fast frequencies with slow detectors like our eyes, see discussion of time averaging above. The important feature of this superposed amplitude is that the amplitude at the screen now has a position dependent amplitude $2A_0 \cos \left(\omega \left(\frac{s_1 - s_2}{2c} \right) \right)$. As you move to different positions on the screen, there will be different brightnesses and even zero brightness at places where $\cos \left(\omega \left(\frac{s_1 - s_2}{2c} \right) \right) = 0$ or $\omega \left(\frac{s_1 - s_2}{2c} \right)$ is an odd multiple of $\frac{\pi}{2}$.

The total intensity at the screen is this amplitude squared.

$$\begin{aligned}
 I_{T_s} &= |A_1 + A_2|^2 \\
 &= \left(2A \cos \left(\omega \left(\frac{s_1 - s_2}{2c} \right) \right) \cos \left(\omega \left(t - \frac{s_1 + s_2}{2c} \right) \right) \right)^2 \quad (4.6)
 \end{aligned}$$

Using the fact that the intensity that we measure is the long time average, we replace the $\cos^2 \left(\omega \left(t - \frac{s_1 + s_2}{2c} \right) \right)$ term by $\frac{1}{2}$. Using $s_1 - s_2 \equiv d(x)$, the difference in the distances for rays from slit one and slit two in the Figure 4.3 on page 89.

For the geometry of Figure 4.3 on page 89, in which the slit separation is very small compared to the distance from the slits to the screen then the triangles for the inclination of the rays and the difference of distances are similar and thus $\frac{d(x)}{l} \approx \frac{x}{D}$. Putting this all together, we can finally write

$$\begin{aligned}
 \langle I_{T_s} \rangle_t &= 4 \langle I_{1_s} \rangle_t \left(\cos \left(\frac{\omega d(x)}{2c} \right) \right)^2 \\
 &= 4 \langle I_{1_s} \rangle_t \left(\cos \left(\frac{x \omega l}{2c D} \right) \right)^2 \quad (4.7)
 \end{aligned}$$

⁵The identity is $\cos(\alpha) + \cos(\beta) = 2 \cos \left(\frac{\alpha + \beta}{2} \right) \cos \left(\frac{\alpha - \beta}{2} \right)$

where $\langle I_{1_s} \rangle_t$ is the intensity at the screen if you only have slit one open. Equation 4.7 on page 96 describes the brightness pattern that is observed as you move up or down a distance x measured from the central position on the screen. It predicts a rapidly changing pattern of bright and dark spots.

It is important to once again emphasize that it is the amplitudes that add. The amplitude carries the causal information. The brightness, or intensity, at a point is derived as the square of the amplitude. What you see is the intensity. In physics language, what you see is the “energy per unit area per unit time”. This is an interesting place to comment about knowing. Originally, in Fermat Theory, we dealt only with the intensity and in a very restricted sense – the light was either there or not there. What was manipulated in the construction of the theory was what was measured; the light being there or not being there. We could have created a measurement system by defining so much being there by using standard sources and adding independent paths connecting the sources to some place. This would be the concept of ‘brightness’.

With Fresnel/Young, there is a new concept closely related to the old ‘being there’, the brightness as a measured quantity in units of energy per unit area per unit time. In some sense, it is ‘being there’ on a measured scale; how much energy is at that place. Now there is also a new idea, one that is the basis of the brightness, the amplitude, which itself cannot be detected but only its square. The amplitude, the thing that cannot be measured directly, is the manipulated quantity, additive causal agent, and the measured quantity, the brightness is found from it. Later with Maxwell, it was discovered that the amplitude for light was a measurable entity, a special combination of the electric field and this can in principle be measured or, at least, we thought so. In fact, it is only recently that direct measurements of the field strength has become possible. At the time of Young and Fresnel, the amplitude could not be measured. We will return to this issue later in quantum mechanics, see Section 19.6 on page 431, where the wave function, the causal entity, is not measurable but the probabilities, the square of the wave function, is measured. In the case of quantum mechanics in contrast to that of light, we do not anticipate that someone will discover a new interpretation of the wave function that will then make it directly measurable. In fact, when we learn more about light, see Chapter 22 on page 483, we will discover that we really cannot know the electric field but we have a ways to go before we can discuss that. It is because of this complex interplay between the particles of light and the amplitude of the light that we now think that the amplitude of quantum mechanics will not be directly measured.

We now know how to construct the amplitude for light with a given

frequency. What do you do if you do not have monotonic light. For any form of the light, you can treat it as a superposition of several frequencies or different colors? Evaluate what happens for each frequency, add the amplitudes, and then square. If we know what happens to harmonically oscillating signals, then we know what happens to anything. Interestingly, if, as always happens for visible light, you take the long time average, the mixed frequency terms in the square drop out, $\langle A_{\omega_i} A_{\omega_j} \rangle_t = 0$ for all $\omega_i \neq \omega_j$ and thus

$$\begin{aligned}
 \langle I_{\text{Tot}} \rangle_t &= \langle (A_{\omega_1} + A_{\omega_2} + \cdots + A_{\omega_n})^2 \rangle_t \\
 &= \langle A_{\omega_1}^2 \rangle_t + \langle A_{\omega_2}^2 \rangle_t + \cdots + \langle A_{\omega_n}^2 \rangle_t \\
 &= \langle I_{\omega_1} \rangle_t + \langle I_{\omega_2} \rangle_t + \cdots + \langle I_{\omega_n} \rangle_t. \quad (4.8)
 \end{aligned}$$

This translates into the statement that you've heard since childhood that light is made up of individual colors.

This has been a rather difficult and long analysis of the Young's Double Slit Experiment. It was important to slog through it so that we could appreciate the simplicity of the approach developed in the next section. The other advantage of doing this is that it emphasizes the elements of the analysis that is often overlooked but essential for understanding the phenomena. It also makes clear what the assumptions of the subsequent analysis entail.

4.5 Phasors

In the previous section, we derived the intensity pattern observed in the Young's Double Slit Experiment directly using harmonic functions. In order to complete the analysis, we had to remember and deal with rather difficult properties of these functions. In Feynman's book, QED, [Feynman 1985], he uses an alternative approach based on clocks to keep track of how light propagates. There is a clock that is carried by the light as it propagates. The clocks have one hand and the length of the hand represents the magnitude of the amplitude of the light. The rotation rate of the hand of the clock is the frequency. This is a descriptive way to introduce the idea of a mathematical entity called the phasor.

The complications of the analysis of the double slit with harmonic functions is associated with the difficulty of adding the two harmonic functions and the complications of their time dependence. Shortly we will have to deal with more than one slit. In that case, the direct manipulation with

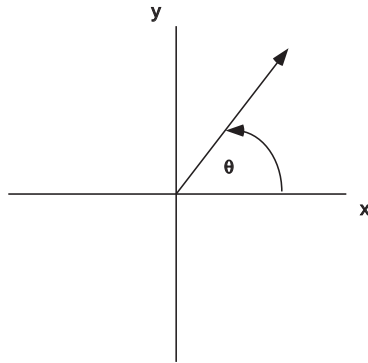


Figure 4.6: **Phasors:** A phasor is a two dimensional vector. It can be added and subtracted, see Figure 4.7 on page 100. These are the clocks in Feynman's QED, [Feynman 1985].

harmonic functions is almost impossible. Thus, the invention and use of phasors is essential to our understanding of how light operates.

Let's do the problem of the double slit with phasors or the clocks in the Feynman's language. In a sense, the introduction of the phasor seems to be an added complication. The use of phasors is connected to our problem because it is a two dimensional vector and for a two-dimensional vector, the x component is $A \cos \theta$, where A is the length of the vector and θ is the direction as measured from the x axis. The direction θ can now be varied with time as $\omega t + \theta_0$ where θ_0 is some initial angle. Adding the two dimensional vectors by the usual process of tip to tail addition and taking the x component you get what you would have if you had added two harmonic functions, see Figure 4.7 on page 100.

You could legitimately ask at this point what could be the possible advantage of using phasors. There are two, one a general one and one that is special to the circumstances at hand. The general one is that for cases with more than two amplitudes or with different amplitudes, it is just easier to add the phasors. The second is that for our case, the different phasors that are to be added all have the same frequency. This means that the several phasors to be added all move together and that the net amplitude also moves with them as if they were rigidly connected. Said another way, since they all move at the same frequency, you can remove the common rotation rate from them all and treat them as in a fixed orientation with respect to each other and a fixed orientation with respect to the directions of the x and y axis. The fact that we take the long time average means the we do not

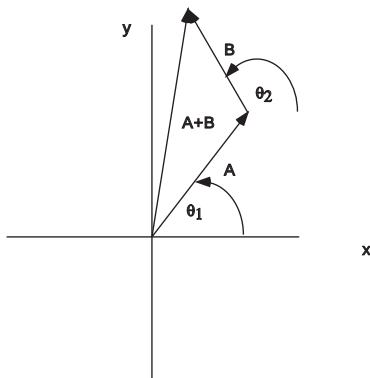


Figure 4.7: **Adding Phasors** Two phasors, $\mathbf{A}$ and $\mathbf{B}$ are added to produce a new phasor, $\mathbf{A} + \mathbf{B}$, by placing the tail of the second phasor, $\mathbf{B}$, on the tip of the first phasor $\mathbf{A}$. The resultant phasor, $\mathbf{A} + \mathbf{B}$, is the phasor connecting the tail of $\mathbf{A}$ to the tip of the relocated $\mathbf{B}$.

care about the orientation of the net phasor; we are concerned with only its length whose square is related to the brightness.

Please note that when you interpret the systems as a clock the angle advances oppositely to the phasor convention that I have chosen. My convention is the usual one and Feynman's is the clock convention. These are obviously conventions chosen for convenience and do not effect the physics. I will start with the usual one.

Let's do the double slit with phasors. For each slit there is a phasor. The two phasors associated with the light at the slits are the same in length and angle since the light there is identical. We are free to pick the direction of the phasors arbitrarily so pick them as straight up. The phasors at the screen are related to the original phasors by the delay in traveling from the slit to the screen, i. e. the phasors are rotated through an angle $\theta_1 = -\omega \frac{s_1}{c}$ and $\theta_2 = -\omega \frac{s_2}{c}$ respectively. Remember that the actual angle is $\theta_1 = \omega t - \omega \frac{s_1}{c}$ but we have removed the rapid time variation since all the phasors have the same frequency of rotation.

In fact, we now realize that, since the orientation of the axis system was arbitrary and our concern is with the phasors for the light at the screen, it would be more convenient to orient the system in the direction that is convenient for the phasors at the screen. In other words, pick one of the two final phasors, say the one associated with slit one, and chose the axis system so that it is oriented straight up and the other one is oriented in a direction $\theta_2 = -\omega \frac{s_2 - s_1}{c}$. Again, defining $d(x) \equiv s_2 - s_1$, the angle for the phasor for

the light from the second slit is oriented at an angle $\theta_2 = -\omega \frac{d(x)}{c}$. With this set of conventions, we have now recovered Feynman's clock conventions, see Figure 4.8 on page 101.

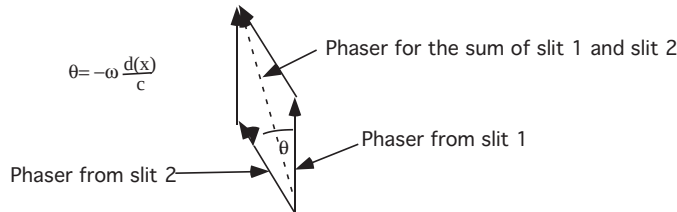


Figure 4.8: Adding phasors in Young's Experiment The addition of the two phasors in the Young's double slit experiment leads to a final phasor which when squared yields the brightness. The angle between the phasors is determined by the difference in the distance traveled by the two rays which varies with the position in the screen. When $\theta = 0$ or a multiple of 2π the resultant phasor is twice that of either slit alone. When $\theta = \pi$ or an odd multiple of π the resultant phasor is zero.

The resultant phasor is the sum of these two phasors from each slit alone and is thus the phasor found by placing the tail of the second phasor on the tip of the first phasor. For the point on the screen at the center, the situation has both phasors pointing straight up and the net phasor thus has the length of twice what one of them was. This phasor represents the amplitude of the light in the sense that its length is the factor which multiplies a harmonic function that varies with time with a radian frequency ω . The angle of the resultant phasor plays no role in our considerations. This configuration produces a brightness that is four times that of either one slit alone. This situation repeats itself whenever the angle θ is a multiple of π . On the other hand, whenever $d(x)$ is such that it is an odd multiple of $\frac{c\pi}{\omega}$, the brightness is zero. For the Young's double slit, when the slit width is very narrow and slit separation is small, and the screen far away, the angle between the phasors is

$$\theta = \omega \frac{d(x)}{v} = \frac{\omega}{c} \frac{l}{D} x \quad (4.9)$$

where ω is 2π times the frequency of the light, $d(x)$ is the difference in the distance that each of the two rays travel from the slits to the screen, and c is the speed of light. In arriving at the last part of Equation 4.9 on page 101, I have used the fact that with these conditions, the right triangle made with d as the side and l as the hypotenuse is similar to the right triangle with x

and D as sides and that since $D \gg x$, $\frac{d}{l} \approx \frac{x}{D}$, see Figure 4.3 on page 89.

You get the interference pattern from varying $d(x)$. As you move up the screen $d(x)$, the difference in distance from the two slits to the common point on the screen, increases linearly with the position up the screen. At the mid-point, the two phasors are together. As you move up the screen, the two phasors separate. The intensity goes from $I_0 = 4I_1$ to 0 as the angle θ starts at zero and opens to π . As you move further up the screen, the angle continues to open so that the intensity returns to its original value when θ is 2π . As you continue further up the screen, the angle continues to increase and the pattern repeats.

These phasors are those little clocks in Feynman’s book “QED” [Feynman 1985].

Also It is important to remember that the clock is not in our physical space time. It is carried over the path on which the light travels. The only way to measure it is in the net result of the comparison with the other clocks on other paths. This is like the amplitude. We like to think of the amplitude as a extension into some dimension. Again, for now, it is not directly measured and does not extend into any of the dimensions of space time that we can measure.

4.5.1 Example of Three Slits and More

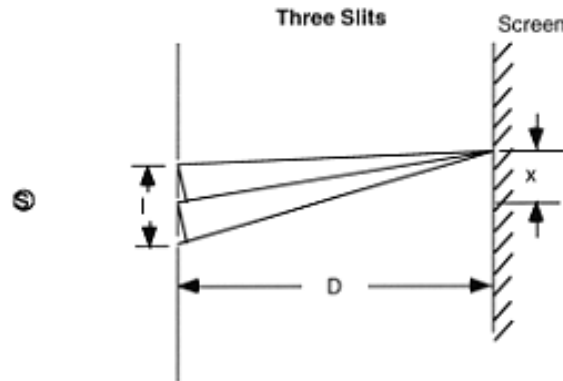


Figure 4.9: **Three Slits** Light from a monochromatic source illuminates and opaque screen with three slits. The light from the slits is projected on a distant screen.

To test our understanding of the Young/Huygens construction, let's look at the case of three slits. Again, light from a single source shines on an opaque screen which, in this case, has three very narrow closely spaced slits and the light emerging from the slits is projected onto a distant screen, see Figure 4.9 on page 102. To keep matters simple the interval between the slits is the same and the light is monochromatic.

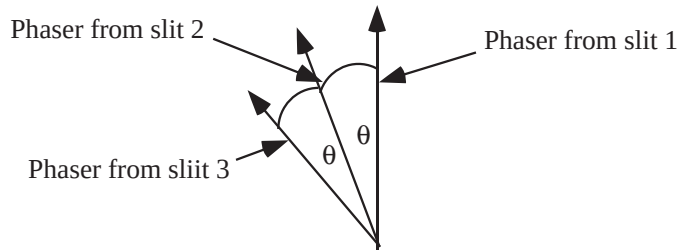


Figure 4.10: **phasors for Three Slits** The three phasors for the light on the screen in the apparatus shown in Figure 4.9 on page 102. The angle between the phasors, θ , for slit 2 and slit 1 is computed in the same way as for the double slit case. It is the radian frequency times the difference in distance to the point on the screen from the second slit minus the distance traveled by the ray from the first slit divided by the speed of light. The angle between the second and third phasor is determined similarly. Since the slits are separated by the same distance these angles are the same.

Instead of two phasors, there are now three. Since the arrangement is symmetric, all three phasors have the same magnitude. As in the case of the double slit, the angle between the phasors is determined by the difference between the distances traveled over the paths from slit to the point on the screen. Because we picked the intervals between the slits to be equal, the angle between the phasor representing the top or first ray and second or middle ray and the angle of the phasor between the middle ray and the third or bottom ray are always equal. As before, as the point of interest moves up the screen, the relative angle between the phasors opens out. The relative angle between the phasors is still 2π times the frequency of the light times the difference in distance of travel of the first and second or second and third rays divided by the speed of light. Again using the notation similar to the section on the double slit and using Figure 4.9 on page 102, we can calculate that ϕ is

$$\phi = \omega \frac{s_2 - s_1}{v}$$

$$\begin{aligned}
&= \omega \frac{d(x)}{v} \\
&\approx \frac{\omega}{v} \frac{l}{2} \frac{x}{D}
\end{aligned} \tag{4.10}$$

where l is the total distance between the slits, v is the speed of light, D is the distance to the screen from the opaque slit screen, and x is the position up the screen from the midpoint.

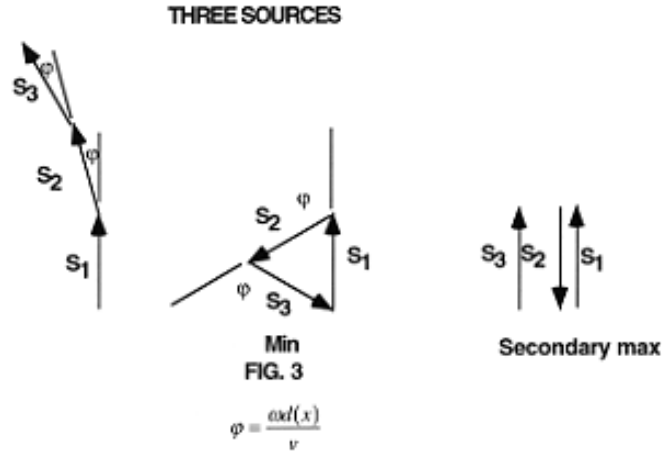


Figure 4.11: **Phasor Configurations in Three Slit Case** As you move up the screen in the case of three slit illumination different configurations for the three phasors emerge. At the midpoint, the three phasors are aligned producing a bright spot that is nine times as bright as that of a single slit. When the three phasors close to form a triangle there is a dark spot. Advancing further, there is a secondary bright spot that is the same as that of a single slit. Here v is the speed of light.

The striking feature of this result is the difference in the pattern of bright and dark that emerges as the point of interest moves up the screen. The three phasors form a phasor fan that opens out uniformly as you move up the screen. This generates an interesting pattern of phasors, see Figure 4.11 on page 104. At the midpoint, the three phasors are aligned and the amplitude is three times that of a single slit. The brightness is thus nine times that of a single slit. As you move up the screen, the fan opens out. Remember that the angle between the phasors is linear in the distance above the midpoint, see Equation 4.10 on page 104. Also remember that the phasor from the third slit is always advanced by twice the amount of the second phasor.

When the angle $\phi = \frac{2\pi}{3}$, the angle between the first and second phasor, the fan has opened to the point that, when the vectors are added, the vectors form a closed triangle. Here the third phasor is at $2\phi = \frac{4\pi}{3}$. In this case, the amplitude and thus the brightness is zero. The next special case is when $\phi = \pi$ and the third phasor is at 2π . Here the phasors from slit 1 and slit 2 cancel and the resultant amplitude is the same as from one of the slits. The brightness is the same as that of one of the slits. Moving further up the screen, the next interesting place when $\phi = \frac{4\pi}{3}$. Again, the triangle closes and the amplitude and brightness is zero. Moving further, we get $\phi = 2\pi$ and the third phasor is at 4π . Here the three phasors are again aligned and the amplitude is three times and the brightness is nine times that of a single slit. From here on, the pattern will repeat. The intensity pattern is shown in Figure 4.12 on page 105.

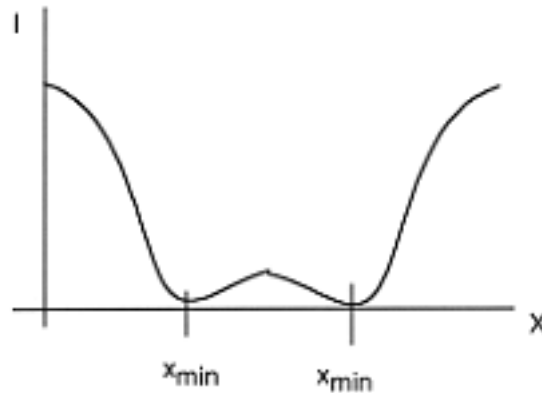


Figure 4.12: Intensity Pattern for Three Slits The intensity pattern for the three slit experiment of Figure 4.9 on page 102. At the central maximum, the brightness is nine times that of a single slit. As you move up the screen, the brightness drops to zero. It then recovers to a brightness the same as that of a single slit. From here the pattern is symmetric about this point having a minimum of zero and then a place with a brightness nine times that of a single slit. Moving up from here the pattern repeats.

Work your way through the four and five slit case. What happens when there are lots of slits?

4.6 The Theory of Light Propagation

Thomas Young first articulated the operations of the amplitude for the limited circumstances of the double slit apparatus. The the complete articulation of the Huygens construction was clarified by Fresnel. The idea is an extension of Young's construction for other circumstances. The basic idea is that there is a two step process for the development of the brightness of light at a point from some source of light. The light propagates between two points in space by having its amplitude travel over all available paths. The amplitude is the quantity that is additive in its sources. The brightness at any point is the light amplitude at that point squared. For instance, in the double slit experiment, the light can only come from one or the other of the two slits. We were able to generalize that to any number of discrete slits. The net light amplitude at the screen is the sum of the amplitudes from each of the slits. What we need to do is generalize this even further to the case of a continuum of slits. This was the contribution of Fresnel.

What we found in the previous section is that the easiest way to keep track of the amplitude is to think of it as a clock hand or a phasor with the rules for addition being those of two dimensional vectors. At each point, there is a clock with both a hand length, magnitude, and an angle. For a given ray, the magnitude of the clock hand is the square root of the brightness at that point. If you know the brightness at any point, or set of points, the rule for calculating the value of the light amplitude is to calculate how the clock hand changes as you drag it over all possible paths from the starting point to the ending point. When you had only slits as sources, this meant adding the phasors from each slit transported over the straight line path. In some sense, these are all possible paths although we know this not the case. We will have to proceed in two steps. We will work through an example with many but not all paths and then use the results to justify what we did. For now though stated as a principle, we say that we are using all possible paths. For a path that is not a straight line and in a medium in which the speed of light varies, you appropriately segment the path. You then need to know how the angle varies as you move the phasor over the path. The rule is simple. For an increment of path of length Δs , you advance the clock hand by an amount $-\frac{2\pi}{vT}\Delta s = -\omega\frac{\Delta s}{v}$, where v is the speed of light at that point on the path and T is the period of the light. Ultimately, you add all the little clock hands from each of the paths to get the net light amplitude at the final point.

We will do this for a specific case but one that is rich enough to indicate the generalization to all cases.

4.6.1 Mirror Reflection à la Fresnel

Let's consider the case of the reflection. Here we will follow and expand considerably the discussion of Feynman [Feynman 1985].

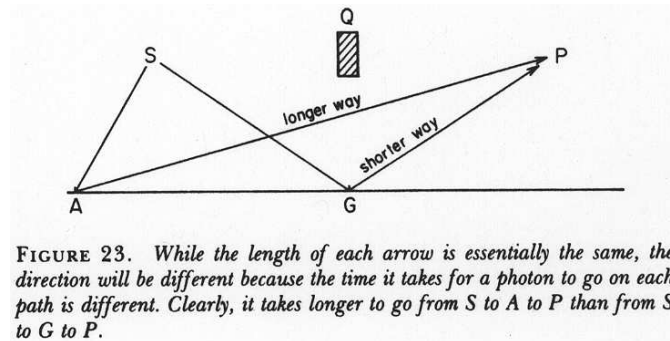


Figure 4.13: **Least Time for Reflection** Figure 23 from Feynman's QED, [Feynman 1985], which shows the least time path for a mirror.

For a mirror, we know that Fermat's Least Time says that the light follows the path that requires the least time, Section 3.3.6 on page 72. Thus in Figure 4.13 on page 107, we would say that the light travelled over the path labeled SGP and not SAP. In other words, if we place an obstacle in path SAP, we do not block the light but, if we place an obstacle in the path SGP, we do block the light from the mirror.

Yet from our analysis of the multiple slit case, light in some sense travels over all the paths available to it and the brightness at P is due to a constructive interference at P. It will be a major point of our efforts to understanding light to reconcile these seemingly divergent concepts, light in a ray in path SGP and light all over the mirror.

To start, we need to construct a situation like the multi-slit case. We will want the light to use the entire mirror and then find out why the regions away from the center have no change in the light at P when a barrier is placed near them. Fresnel developed the algorithms that enabled us to do this by extending the techniques used for slits to a continuous surface such as in this case the mirror.

The important part to remember is that, in Fresnel's Theory, light now travels over the *entire* mirror. The first problem is to decide how you carry out the algorithm? For barriers with slits, we just used all the available paths and that worked. Therefore, it seems natural in this case to divide the mirror into parts and see how each part's phasor contributes to the

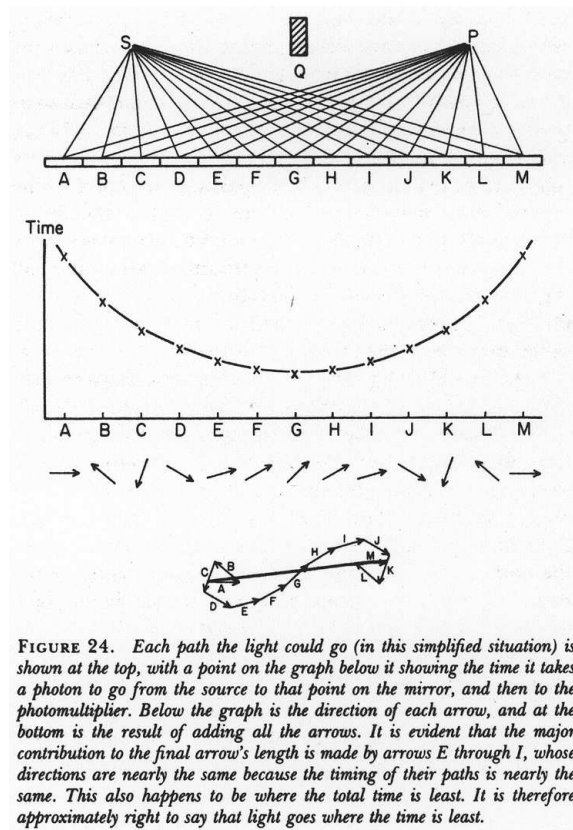


Figure 4.14: **Mirror with phasors** Figure 24 from Feynman's QED, [Feynman 1985], which shows the phasor paths for the mirror and the time for each phasor.

brightness of the light at the end point. This exercise is shown in Figure 4.15 on page phasor108.

With the mirror divided into segments, we can follow the pattern of Section 3.3.6 on page 72 and find the time for light to travel from S to P for a path that touches the center of that segment. Now plot the time of arrival for each segment versus the segment label, see Figure 4.15 on page 110. As you expect, the segments reach a minimum at the center. This is the idea of Fermat's least time; the least time path is the one over which the light travels. Use these times to orient a phasor that is associated with each segment. Remember the the direction of any one of the phasors is arbitrary and he choses the first segment, the one in Figure 4.15 on page 110 labeled

A, to be horizontal. Looking at the pattern of the phasors, note that the phasors for the light paths near the central path all have an orientation that is similar to that of the middle one. In contrast, for a group selected from around another area more toward the edges of the mirror, you find that the angles of the phasors for the group members differ from each other. In terms of the phasors, the phasors in the group are not aligned but instead point in all the directions around the clock. If you add the phasors for a group of segments away from the center you see that a group of phasors these wrap around and the net effect is to produce no net phase. Whereas, the phasors from a group selected around the center when added will produce a net phasor with finite magnitude. An even more impressive display is to add the phasors in a tip to tail fashion each of the phasors as you move from the segment A to the other end. Here you see how the segments in the middle add and are the major contributions to the net phasor. The square of the net phasor is the brightness at the point P.

Another way of stating this situation, is to say that not only do you get a minimum time for the path from the center of the mirror, the least time path, but that the minimum is a soft soft one, i. e. the slope of the curve is zero. Another expression for this situation is that the time for paths around the minimum time path is stationary. Paths around any other point on the mirror are not stationary.

Relating this situation to the phasors, this means that the variation in the angle of the phasors as you move through the central region changes very slowly whereas the variation in the angle of the phasors for a family of segments elsewhere is large. This implies that, as you move through the family of paths that are centered around the minimum, there is little change in the angle of the phasors for the members of the family. This is not the case for a small family of paths that come from points on the mirror that are not at the center. The idea is that the phasors from the ends wrap and add to zero whereas the phasors from the middle region reinforce. This is shown in lower part of Figure 4.15 on page 110 where the mirror has been divided into 13 segments and the phasor from each segment is added incrementally added tip to tail as we move across the mirror surface. The primary contribution to the resultant phasor is from the segments at the middle of the mirror. There is some indication of the wrapping pattern at both ends.

Lets look at the end regions in some detail. The situation in Figure 4.15 on page 110 is not detailed enough to show the power of the technique.

Making many small divisions of the mirror, we find that the paths around the point B cancel themselves out. This is why you can cut off the ends of

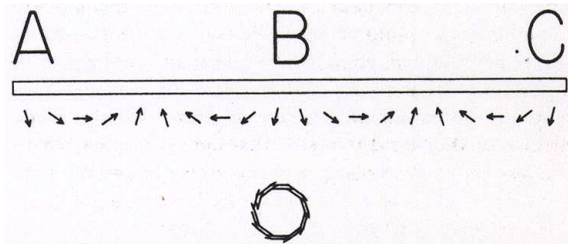


FIGURE 25. To test the idea that there is really reflection happening at the ends of the mirror (but it is just cancelling out), we do an experiment with a large piece of mirror that is located in the wrong place for reflection from S to P . This piece of mirror is divided into much smaller sections, so that the timing from one path to the next is not very different. When all the arrows are added, they get nowhere: they go in a circle and add up to nearly nothing.

Figure 4.15: **Phasors from Ends of Mirror** Figure 25 from Feynman's QED, [Feynman 1985], which shows the phasors for the different paths touching parts of the mirror near the end.

the mirror or block it and not lose any light; the phasors from that part of the mirror all cancel each other out so that the net contribution is zero. In other words, the light uses paths from the entire mirror but the segments on the ends do not contribute to the net phasor. That is why, in Fermat's theory, you did not include these long time paths from the ends of the mirror.

Is there light on the ends? Not in the sense that you can block the light by placing an obstacle there. The segments around G , on the other hand, all have nearly the same travel time which implies that the phasors are all pointing in the same direction and the sum is the principle contribution to the net amplitude. Thus, the light does not travel over the single path touching at G . To block the light completely you need an obstacle that blocks not only the path at G but also the segments around G . In other words, the least time path is at the center of a cluster of paths whose phasors all reinforce each other and thus produce a large amplitude. The paths that the light travels are not necessarily all least time paths. They are the cluster of paths that have a small variation in the travel time.

We can do a much more detail calculation of the situation that Feynman develops in QED, [Feynman 1985]. We will use lots more paths by dividing the mirror more finely. I have added another detail to the Feynman example that he omits. It is to take into account for the drop in the amplitude with distance in a three dimensional space. Choose the two points of interest at a distance of two unit distances apart. The points are one unit above in the same unit. The wavelength of the light is $\frac{1}{10}$ of a unit. Please note that there

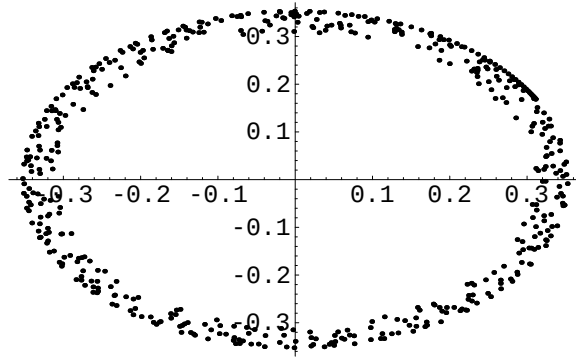


Figure 4.16: **Phasors from Mirror with 1000 Parts** Because there are so many paths, each phasor is represented only by its endpoint. There is no apparent overall direction to the family of phasors.

is no length scale in this problem so that we can choose this unit arbitrarily. For the first analysis, there are 1000 paths for points on the mirror between the points. Figure 4.16 on page 111 shows the ends of all 1000 phasors. It does not appear that there is a non zero resultant phase, i e. there is no concentrated sets of points.

This can be seen not to be the case if we just add them all up. Figure 4.17 on page 111, shows the result of just adding all the phasors. There is a non-zero result which is, of course, the resultant amplitude.

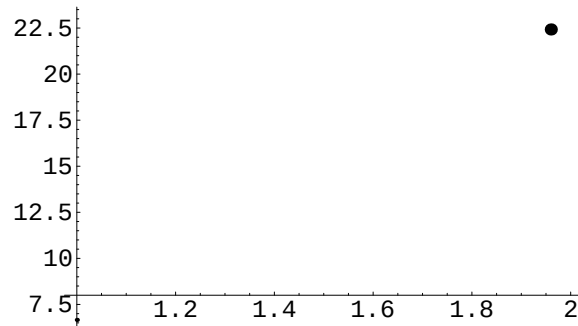


Figure 4.17: **Sum of the phasors from Mirror with 1000 Parts** Although the phasors seem to be in all directions a direct sum yields a non-zero result. Again, the end point of the phasor is a dot on this plot,

The square of this amplitude is the brightness of the light at the second

point. Of course, this is consistent with the earlier result in Feynman's QED.

It is interesting to add the individual phasors incrementally tip to tail as you move from one end the mirror to the other. Again, since we have so many elements, the curve appears to be smooth instead of the kinked curve that you see in Feynman. Figure ?? on page 112 shows in detail the spiral that is characteristic of problems of this type. In this representation, it is clear that the regions located away from the central region smoothly wrap to cancel each other out and the non-zero part is coming only from the middle sections.

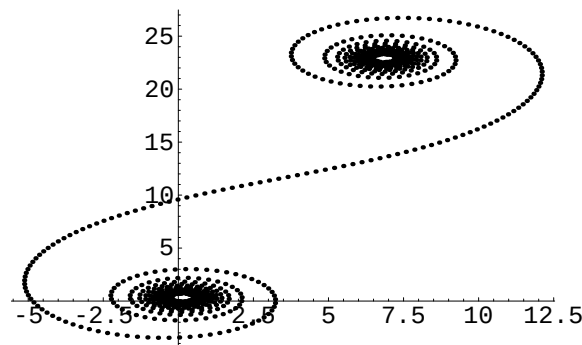


Figure 4.18: **Tip to Tail Sum of the Phasors from Mirror with 1000 Parts** Here we see the emergence of the spiral phasor that is characteristic of these phasor sums.

Let's return to the question of whether or not the light pattern used the ends of the mirror. There are paths that use the ends. There are phasors for those paths. It is just that the net effect of a collection of paths from the ends wrap to add to zero. This is very clear in the high resolution version of the analysis. The question of whether the light uses the end of the mirror is like the question: For a body at the center of the earth, is there gravity from the earth acting on it? The answer is yes but, for a body located at that point, the gravitational attraction from the parts of the surrounding earth all add to zero and the net strength of the gravitational force and thus the weight is zero. For the body at the center of the earth, we do an interesting thought experiment. What would happen if we had an antigravity shield the could eliminate the gravitational interaction from the matter in the earth on one side? The answer is obvious. There would now be a net gravitational force to the other side.

We do not have an antigravity shield and cannot do the experiment

described above but, for the case of the mirror and light, we could selectively eliminate the contribution of segments of the mirror whose phasors point in a certain direction. For instance, we could cover with a light absorbing material segments whose phasors point in a direction opposite to that of the middle section of the mirror. From Figure 4.18 on page 112, we see that these are regions of finite size on the surface. In other words, in any one of the coils at the ends of the spiral take each phasor that has any component that points opposite to the resultant direction and relate them back to where the path contacts the mirror. Since these are all from a given part of the mirror, darken those regions. This will make a stripped mirror with somewhat less than half the regions darkened. This would make a very bright spot at P, significantly brighter than with the segments of the mirror uncovered. This is also the basis for a diffraction grating. In this variation, only the ends of the mirror are used. Because the loops from the paths using segments at the ends of the mirror are all about the same size the regions of darkening on the mirror are also the same size. Note how in Figure 4.18 on page 112, even with the added term for the fall off with distance, the loops after the first few wraps quickly become about the same size. This allows for the easy manufacture of the mask on the mirror, the darkened regions are the same size. As we will see shortly, the rate of looping is strongly dependent on the wavelength of the light. This is the basis for a device called a diffraction grating.

4.6.2 Diffraction Grating

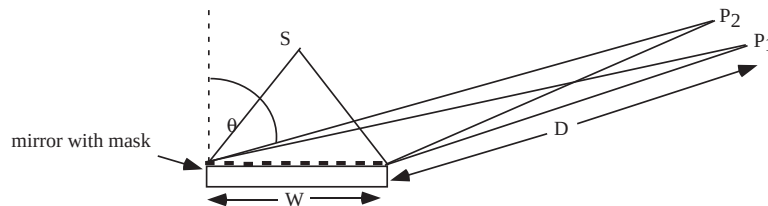


Figure 4.19: **Diffraction Grating** A situation in which you only use the ends of the mirror. By masking intervals on the mirror, you can generate a pattern of reinforcing phasors at particular positions for a given wavelength.

Following the geometry of Figure 4.19 on page 113, light from small source S uniformly illuminates a ruled mirrored surface with half the area covered. Thus if we can arrange that the point P is located such that the

phasors from the ruled mirror surface repeat after each rule, the rules or mask will block out half the loop so that we only get reinforcing phasors. There are N rules on the surface and thus we can treat this as an N slit problem. With the geometry shown, all the phase differences are due to the paths from the mirror to P . Setting the distance from the mirror to P as D , which is large on the scale of all other lengths in the problem, and designating the position on an arc by the angle, θ , measured from the vertical, for positions above the mirror, $\theta = 0$ all the phasors from the mirror are aligned. As we move the mirror to larger angles, the N phasors start to fan out with uniform angular spacing between them. The difference in distance traveled from the mirror to P by any pair of adjacent paths is $d \approx \theta \frac{W}{N}$. The angle between the phasors, $\phi \approx \frac{\omega}{c} d = \frac{\omega W}{cN}$. All the phasors are equally spaced through a full circle and there is no brightness when $\phi_{zero} \equiv \frac{2\pi}{N-1}$ or

$$\theta_{zero} = \frac{N}{N-1} \frac{c}{\omega W} 2\pi \quad (4.11)$$

Since $\frac{c}{\omega} = \frac{\lambda}{2\pi}$ and N is very large,

$$\theta_{zero} \approx \frac{\lambda}{W}. \quad (4.12)$$

For all practical cases, $\lambda \ll W$, which implies that θ_{zero} is very small and, thus, there is a very narrow very bright light beam being reflected up from the mirror. It is more interesting to look for other directions in which the phasors reinforce. Remembering what we learned from the multiple slits in Section 4.5.1 on page 102, we realize that we can find the next bright place by looking where all the phasors realign again. This requires that $\phi_{max} = 2\pi$ or

$$\theta_{max} = N \frac{\lambda}{W}. \quad (4.13)$$

If N is comparable to $\frac{\lambda}{W}$, we can get a second maximum within the first quadrant. Remember that the brightness of this maximum is N^2 times the brightness of a single slit and that in the several slit case the intermediate brights were of the order of a single slit in brightness. Thus there is a very narrow very bright beam at the angle θ_{max} with hardly any light at any other angle. Not only that but, if the source of light is mixed, say half red and blue and, since the wavelength of the red is twice the wavelength of the blue, the separate colors will separate into narrow beams where the red is at twice the angle as the blue. This device is a very effective tool for the analysis of the structure of a light beam. It has found numerous

applications in Astronomy, Physics, and Chemistry. It is the reason that you see a rainbow when you look from the side onto a CD illuminated from above.

4.6.3 Diffraction

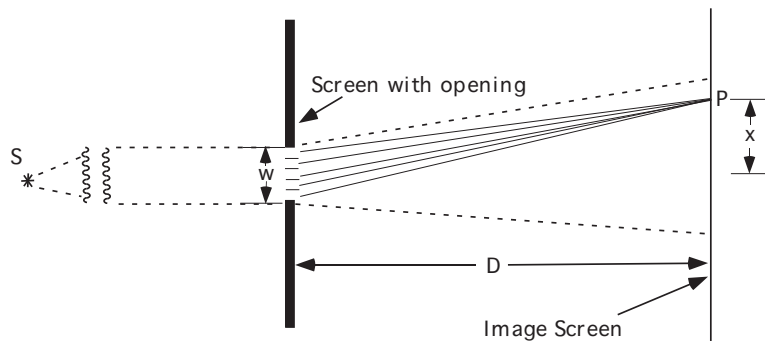


Figure 4.20: **Diffraction** Monochromatic light from a distant source illuminates an opaque screen with an opening and the light is projected onto an image screen. The image is larger than the geometric image of the opening.

Diffraction is the name of the process that occurs when light passes through an opening; think of light from a distant source uniformly illuminating an opaque screen with a circular hole in it, see Figure 4.20 on page 115. In Fermat's Least Time approach, you would expect a sharp shadow of the opening, i. e. on a silvered screen used for imaging, you expect to see a circular spot like the original opening. In the Fresnel construction, there is illumination outside the geometric image of the opening. This is what is observed.

We can understand the Fresnel result by considering the opening as made of several slits and allowing the number of slits to get very large. In this case, we again have a multi-slit case. The analysis follows that of the diffraction grating and although we expect that the image of a very distant source to have zero opening angle, there is a small opening angle given by Equation 4.12 on page 114. Thus in Figure 4.20 on page 115, we see that for points on the screen at a distance from the center of less than $x_{zero} = \frac{\lambda}{W}D$, there is image illumination on the screen. This formula was only derived using the simplest geometry, a very long slit. For a circular aperture, this is

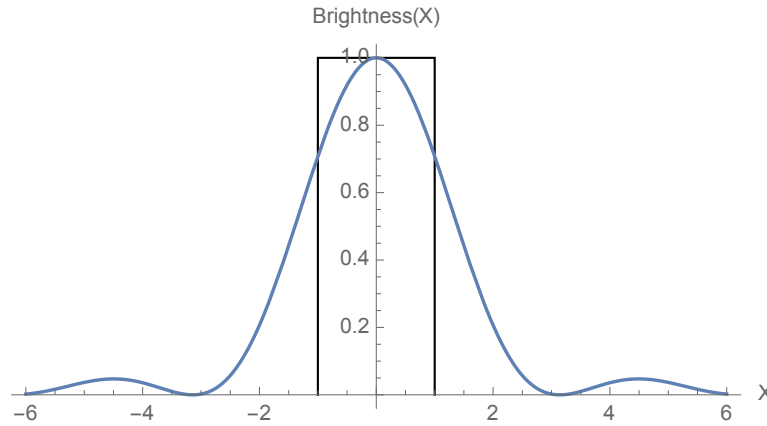


Figure 4.21: **Diffraction Brightness Pattern** The brightness pattern, shown in blue, of monochromatic light passing through a circular hole and hitting a distant screen as shown in Figure 4.20 on page 115. The solid black line is the expected brightness if a shadow was formed.

corrected to

$$x_{zero} = 1.2 \frac{\lambda}{W} D, \quad (4.14)$$

where W is the diameter of the aperture.

There are several ways to see a large effect. The small W case is for openings that have widths close to the wavelength of the light. When viewing the Young's double slit, the pattern of brights and darks that our simple theory predicts do not extend forever as expected. Instead there is an envelope that modulates the pattern. This is the diffraction pattern set by the slit width of the individual slits. Even for a modest W , if D is large enough there is a discernible effect. There were reflectors placed on the moon in several of the Apollo missions. If the reflectors are a fraction of a meter in size and using estimates from "Things Everyone Should Know," Section 1.4.2 on page 16, the image size at the earth is about 5×10^2 meters.

An interesting application of this result is the understanding of the limits of resolution of imaged objects. A point source of light is focused by passing the light through a lens generally with a circular aperture. The image then of this point source of light is now a smear of radius given by Equation 4.14 on page 116. In some sense, this is the pixel size for this imaging system and, unless the image points are separated by an amount greater than that, the two points cannot be resolved. For instance, two headlights, separated by a distance of 2 meters, are imaged by the eye. Again using "Things," Sec-

tion 1.4.2 on page 16, for an eye that is 5 cm in diameter, D in Equation 4.14 on page 116, and aperture of 0.5 cm, the “pixel” size of the eye is about 10^{-5} . At any distance greater than about 10 km, you could not discriminate the two lights. This assumes that you have a perfect lens. Since most of us do not, we cannot even do this good. The point is that, no matter how much they correct your vision, you cannot do better than this.

4.6.4 Lens and Spherical Mirror Revisited

The Fresnel construction gives us new insight into the operation of simple optical devices like the lens and spherical mirror. In both of these, the trick to finding the relation between the object and image point was to construct rays that had the same travel time, see Section 3.3.4 on page 69. In this new context, we realize that the light going between the object and image uses the entire lens. This is different than what happens for the mirror. In the mirror, the light concentrates on the rays around single least time ray that has the angle of incidence equal to the angle of reflection. In other words, if we block half the mirror we remove a portion of the image. If we block the lens, we still have the same image. It is just not as bright.

4.6.5 How do we get least time from Fresnel’s Theory?

Given that, how do you recover all the successes of Fermat’s Least Time Theory?

The ideas of the Fresnel/Young/Huygens approach to light seems to be very different from the rules developed by Fermat. In the Fresnel construction, light is understood to fill all the space available and use all possible paths. In the Fermat case, the light was completely localized. Fermat’s approach was successful in many applications and, as is always the case when developing a superseding theory, it is incumbent on the new theory to recover the working results of the earlier theory. Let’s see how Fresnel recovers Fermat for the simple case of light in a homogeneous medium.

Another point to notice is how these examples combines Fermat’s Least Time and Huygens’s Construction. We actually did not use all possible paths. In the Young’s double slit case, between the slits and the screen, we used the least time, or “straight line” path. In the mirror, we used the straight line path from S to the mirror and and the mirror to P . Why were we allowed to do that? From the earlier analysis, we now have some idea. Not only are the omitted paths longer, they are all parts of families of paths that, like the ends of the mirror, each have rapidly varying phasors and the

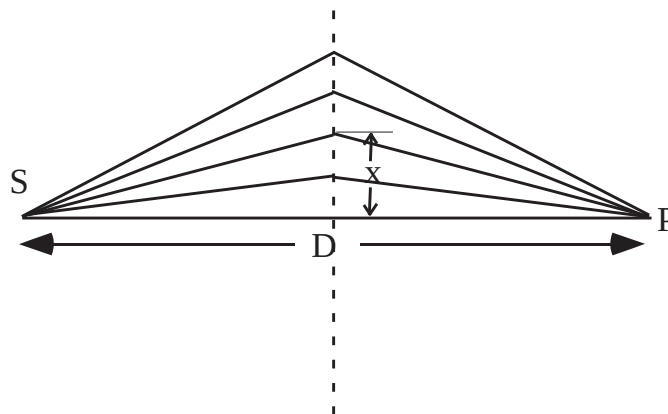


Figure 4.22: **Fresnel Construction in a Homogeneous Medium** From among all paths connecting points S and P separated by a distance D, consider only the paths that are represented by these once kinked paths labeled by the kink distance from the straight line path. As in the case of the mirror, paths at a distance from the straight path, those with large x , will have phasors that vary rapidly with the label x and thus like the ends of the mirror not contribute significantly to the light going between S and P.

net effect of the family of related paths is that they do not contribute to the light at P. Again as a simple example consider the case of light going between two points S and P, separated by a distance D in a homogeneous medium, see Figure 4.22 on page 118. This especially simple example will shed light on these two questions.

In order to reduce the path space to manageable size, we restrict ourselves to the once kinked straight line paths. We see that the lengths of these paths are

$$l(x) = 2\sqrt{x^2 + \left(\frac{D}{2}\right)^2}. \quad (4.15)$$

Once these paths are assigned a phasor, we can see without going through the details that a situation like that with the mirror is obtained. A family of paths around $x = 0$ will have related phases and reinforce whereas families around other values of x will have rapidly varying phasors and thus add to zero. Again, the minimum at $x = 0$ is soft. From this analysis, not only do we see that our dealing with the simplest paths made sense – non-simple paths have families whose phasors cancel – but also that Fermat's Least Time Hypothesis should be replaced by a statement that says the light is in

regions in which the family of rays are stationary. This implies that light is concentrated in the regions of paths that have either a minimum and a maximum time path. The only real criteria is that the family of paths be slowly varying in phase as you move through the members of the family of paths.

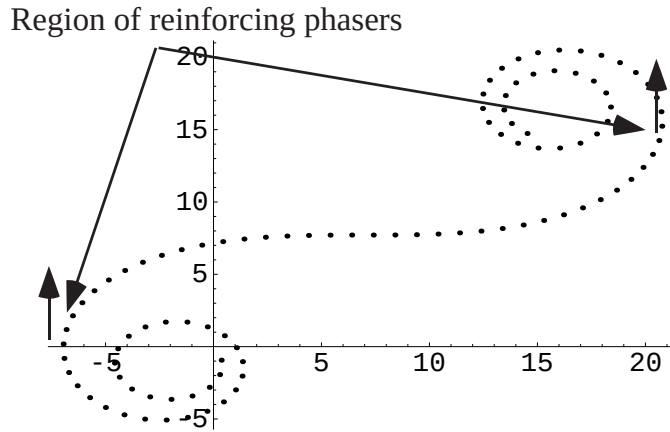


Figure 4.23: **Fresnel Spiral in a Homogeneous Medium** By choosing an orientation of the phasors such that the phasor for the central path, $x = 0$ in Figure 4.22 on page 118, is horizontal, we can define the family of paths that reinforce the central path as those paths moving out from the central path that have no part of their phasor opposite the central path's phasor. The last paths to do this are those with their phasors oriented upward.

4.6.6 Recovery of Fermat's Least Time

In the example in Figure 4.22 on page 118, we can also get an estimate of how thick the region is that can be called the ray of light, i. e. how big must the barrier be that blocks the light from going between S and P? In order to simplify the analysis, we can take advantage of the arbitrariness of the direction of phases to use a length measure that guarantees that the middle, $x = 0$ path has a phasor that is horizontal, see Figure 4.23 on page 119. In this way, we can define the region of interest as those paths between the places where, moving out from the central region, the phasors point up, at an angle of $\frac{\pi}{2}$. The appropriate length measure is one that vanishes for the

central path,

$$l_0(x) = 2\sqrt{x^2 + \left(\frac{D}{2}\right)^2} - D. \quad (4.16)$$

With this length measure, the phases of the kinked path labeled by x are

$$\begin{aligned} \theta(x) &= 2\pi \frac{l_0(x)}{\lambda} \\ &= 2\pi \left(\frac{2\sqrt{x^2 + \left(\frac{D}{2}\right)^2} - D}{\lambda} \right). \end{aligned} \quad (4.17)$$

Setting this equal to $\frac{\pi}{2}$, and rearranging, the paths at the end of the reinforcing region satisfy

$$2\sqrt{x_{edge}^2 + \left(\frac{D}{2}\right)^2} - D = \frac{\lambda}{4}. \quad (4.18)$$

Using the fact that $\frac{\lambda}{D} \ll 1$ and consequently $\frac{x_{edge}}{D} \ll 1$, we get for the two solutions of bounding paths

$$x_{edge} = \pm \frac{\sqrt{\lambda D}}{2\sqrt{2}}.$$

Thus for small, optical, wavelengths and meter separations the band of light is very thin, $\approx 2 \times 10^{-4}$ m. This is truly a ray as in the sense of Fermat. Thus in this case and plausibly for other cases, we get that for optical wavelengths and reasonable separations, the Fresnel construction reproduces the results of Fermat Least Time and, in fact, enhances it by replacing “Least” with “Least or Most.” Note also that the band of light becomes thicker if the separations become large enough. This is consistent with our results for diffraction, see Section 4.6.3 on page 115.

4.7 Polarization

Another important feature of the phenomenology of light was discovered in the 17th century but not studied carefully until the early years of the 19th century. Certain crystals, calcite and Iceland spar in particular, produce rays with two refraction angles. Text when viewed through one of these crystals appear doubled, see Figure 4.24 on page 121.

Besides the constituent nature of light that manifests itself as color, there was an intrinsic doubling of the number of constituents.

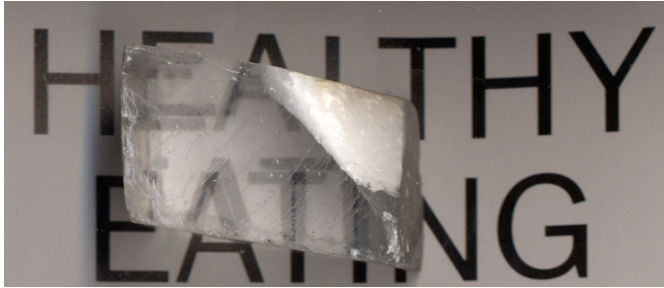


Figure 4.24: **Birefringent Crystal** A calcite of island spar crystal placed over illuminated text produces two copies of the text transmitted through the crystal.

4.8 Introduction of the Field

The Fresnel/Young/Huygens construction brings with it the need for a new physical construct, the light amplitude defined by Young is everywhere, $A(\vec{x}, t)$ is defined everywhere for any time. It brings us to a need to develop techniques for handling objects like this. A physical object that is defined at all points in space is given the general title of a field. In this case, the field is the amplitude for light, but the Huygens/Fresnel construction applies to all propagating signals such as sound, surface waves on water, etc. In the more general case, not only do we want to know the value of the field at each place, but we will want to understand how it changes in time. We will not have the advantage of taking time averages to make what is an evolving system look static. Thus in the general theory of fields we need to know about the development in both space and time. The development of the ideas and techniques of field theory took place in the later half of the 19th century and were applied to optical phenomena by Maxwell. Although this is not modern physics, it is so basic to our understanding of modern physics that we will now spend some time developing it, see Chapter 5 on page 123. Because of Maxwell, we now understand what the amplitude for light is and, in a sense, is no longer thought to be unmeasurable. It is a special field combination whose spatial and temporal variation produces the electric and magnetic fields. These fields can be and are regularly measured although to do so at optical frequencies is still too difficult but we are getting close.

Chapter 5

19th Century Physics

5.1 Introduction

During the 19th century our understanding of the nature of light as developed from James Clerk Maxwell's unification of the electromagnetic forces and the articulation of new rules of mechanics that opened the way for the study of more complex systems than allowed by Newton's laws progressed significantly. These two developments started independently but eventually merged to produce the current basis for our description of phenomena in the universe. The first of these came about by the realization that the combined electric and magnetic forces when supplemented with an extra term called the electric displacement current allow traveler solutions for the combined fields, see Section 5.4 on page 140. His repudiation of action at a distance formalism as the basis for electric and magnetic forces lead to the realization that a consistent discussion of these effects had to be based on a local in space and time dynamic. Subsequent work by many physicists clarified the meaning of his discovery and it is the basis for our modern theory of light.

5.2 Action at a Distance and Field Dynamics

The previous construction of Fresnel/Young/Huygens tell us how to construct an amplitude for light at any point in space given the amplitude at some other point in space. This is the first part of the construction of a field. A field is something, generally a measured quantity, that is defined at every point in space. At each point in space you can measure the entity. In addition, as you move from one point to a nearby point the value of the something changes smoothly; it varies as you change places. There will even

be a rule on how the change as you move from point to point is manifest. To appreciate these rather abstract comments let us look at several examples.

There are numerous examples of fields. The temperature in a room is a field. Temperature is measured for instance by a mercury bulb thermometer. As you move the thermometer from point to point, you will get different values for the temperature. If the room is not too drafty, the temperature at nearby points will be similar; the temperature varies smoothly as you move to nearby points. You can even intuit certain rules for how the temperature changes as you move from point to point. For instance, you can guess that a point at the center of a surrounding group of points, the temperature will be the average of the temperatures of the surrounding points. It is because of rules like this that you expect that the temperature varies smoothly as you go among nearby points. Other obvious examples of fields are air pressure in a room, height above or below the normal height of water in a pool, or the transverse displacement of a stretched string. With some amount of smoothing you can make a field from such things as population density on the earth. Any system that is defined over a manifold, continuous or not, is a field.

The discussion of the previous examples generally did not deal with the time variation. It is not until we endow something with a time dependence that the something becomes interesting. In fact, as we will see, Section 7.4.4 on page 199, we cannot really talk about energy until we have temporal evolution. In the Fresnel/Young/Huygens construction of the amplitude for light, we eliminated the effect of the time variation by “seeing” only the brightness, the amplitude squared, and averaging for long times so that the short time oscillations of the phasors cancelled out, see Section 4.4.1 on page 92. Thus although the brightness as a field can be interpreted as slowly varying there is an intrinsic time variation that makes light especially interesting.

In other words, a field is something that is defined over some manifold, usually space, that has a temporal evolution. The rules for the behavior of the field are usually local in the sense that its variation in space and time is determined by what is going on at those points of space at those times. This is the meaning of local causality. It is one of the bedrock principles of modern physics. It ranks with reductionism as one of our formulating rules. The basic idea is that what happens to an entity happens because of what is going on at the place at which the entity is or the immediate neighborhood. This is in sharp contrast to the situation in theories that are based on action at a distance dynamics. Newton’s Laws of gravitation are an example of an action at a distance theories. To a large extent, it was

the attempt to remove these action at a distance formulation and replace them with locally causal theories that motivated the development of field theories.

5.2.1 Action at a Distance

My former colleague, Johnny Wheeler called it ‘spooky’ action at a distance. Newton, its inventor, was not comfortable with the concept but could not come up with something better. In a letter[Cohen,1958] to the theologian Robert Bentley, Newton wrote:

that gravity should be innate, inherent and essential to Matter, so that one body may act upon another at a Distance thro’ a Vacuum, without the Mediation of any thing else, by and through which their Action and Force may be conveyed from one to another, is to me so great an Absurdity that I believe no Man who has in philosophical Matters a competent Faculty of thinking, can ever fall into it. Gravity must be caused by an Agent acting constantly according to certain laws; but whether this Agent be material or immaterial, I have left the consideration of my Readers.

Regardless of his own reservations and because of the success of the Newtonian approach, physicists became accepting of the anomalous nature of action at a distance and the early formulations of most laws were all in the pattern of action at a distance. Fortunately, Maxwell could not believe these and, for the case of electricity and magnetism, this lead him to the development of the first first-principle unified field theory. Prior to Maxwell’s work there were field theories but these were derivative of an underlying structure. For example, the rules of fluid flow were formulated in a field theory vocabulary. But this was understood to be a consequence of the underlying structure of the fluid. Maxwell’s formulation of the nature of the electric and magnetic systems was actually a statement on the intrinsic properties of these entities. In order to understand this important idea let’s review the situation with action at a distance theories and the contrast to field theories.

All the satisfactory theories prior to the 19th century were not what we now call locally causal theories but instead were based on action at a distance theories, actions resulted from situations that were at a distance from the object of interest. Newton’s theory of the gravitational force is a perfect example. In Newton’s approach to gravitation, a bodies motion

is determined by the separation from a remote other body at the instant under consideration. The moon determined its acceleration from knowledge of the earth's position which is at a distance at that instant. It is hard to accept that, if the earth suddenly ceased to exist that, at instant, the moon would instantaneously react by traveling off in a straight line, no longer in orbit. There are two issues here. First the idea that somehow that moon is influenced not by things going on where it is and the fact that the earth's disappearance should be realized by the moon instantaneously; it should take some time. Consider the case that I am standing in the front of the lecture hall and announce that I am going to make the clock at the back of the room run differently. If I could do that, you would infer that I had a wire or used sound waves or some other mechanism to communicate the change to the immediate vicinity of the clock. Whatever ultimately changed the clocks running was at the place of the clock not at a distance.

Coulomb's law and all the other laws of electromagnetism that were formulated before the 19th Century were action at distance laws. A charge here effected a charge there.

The solution to this basic philosophical conundrum is in the idea of strict locality for all phenomena and the vehicle is the concept called the field. Of course, in physics, a philosophic problem is not a good reason for doing something. The idea must be tested experimentally. The proof of the construction is in the testing. Through his treatment of electromagnetic phenomena as a field theory, Maxwell was lead to predict that light was a disturbance of the electromagnetic field. When this prediction was verified by Heinrich Hertz in 1887, there was a general acceptance of Maxwell's approach. Since that time, we have found that all fundamental theories are field theories; the ultimate modern expression of the nature of matter and energy being through the machinery of quantum field theory. For this reason, it is important to understand the idea of the field. For now we will develop the classical field, we will add the complications of quantum mechanics, see Chapter 19 on page 403.

5.2.2 Local Field Theory

Maxwell developed a local field theory to describe the phenomena associated with what is called electricity and magnetism. He reduced all the known laws of electricity and magnetism into four reasonably simple equations. In so doing, he unified the electric and magnetic forces and predicted the fundamental nature of light. These are considerable accomplishments in their own right but also he somewhat inadvertently clarified the idea of the

field and the idea of causality. His was not the first field theory; it was the first field theory of a fundamental force system. The first local field theory and the easiest to appreciate was the description of fluid flow. It was the success of a field theory of fluid flow that motivated him to attempt to write the rules of the electricity and magnetism in this field theory form.

How fluids move through space is very complex. At any point in the fluid there are several variables that are necessary to describe the state of the fluid. These variables such as density, velocity, and temperature are all fields, defined at each point in space and subject to change by some set of rules that are determined by the values of these variables at that point and nearby points and by the nature of the fluid. For example, if the temperature at a point is higher than its neighbors, that temperature will tend to decrease because of heat flow from the neighbors. Also depending on the nature of the fluid, the density may increase and this will cause flow away from the point. How much effect each variable has on the magnitude of the the other variables and how fast these variables respond will depend on the fluid. The parameters such as the thermal conductivity and compressibility of the fluid which will control the rates at which these effects can take place are measured phenomenologically for each fluid. It is not hard to understand that the properties of a fluid in motion are controlled by local effects; flow at a point depends on the temperature and pressure and flow at the point and neighboring points not on what is going on some distance away. The rules for the fluid flow are thus local. The difference with the results of Maxwell is that we know there is an underlying structure, the atoms. In the case of the electromagnetic field, it is not made of anything but itself. The inability to associate a reality to the field independent of an underlying structure is the basis for the famous search for an ether, see Section 9.4 on page 223.

In fact, Maxwell suffered from that same problem. He discovered his equations by trying to fill space with a hypothetical something that exhibited reasonable mechanical properties and attributing the electric and magnetic forces to whirling vortices in the pervasive medium. The idea was that charges produced vortices in this medium and that the whirling of the vortices close to the charge then produced other vortices etc. until space was filled with whirling vortices and the amount of whirling at any place was the electric force. In other words, in order to understand his own equations, he needed an ether, the famous ether that Einstein disposed of later. He also needed to have the vortices properties be determined by the charge or the whirlyness locally. To the modern physicist, the idea of an underlying mechanical system seems out of place and a little weird. In fact, several years back, there was a collection of articles published that were “lighthearted”

musings by well known scientists, [Weber 1973]. These articles were written as joke. Among the collected articles was the original paper by Maxwell justifying his vortices in the ether as a mechanism for the electromagnetic field. At the time of the writing, there was nothing lighthearted about it.

5.3 The Stretched String*

Since the concept of the field and its dynamical rules are rather hard to grasp in the abstract, let's look at a particularly simple mechanical field system – the transverse displacement of a stretched string. I have to emphasize that this is a field with an obvious underlying mechanical structure – the string, a system with mass and an internal force, the tension. This is in contrast with the fields that we will deal with later. These fields are themselves the fundamental entities. The other thing to realize is that the string that we deal with is an idealized element. It has zero thickness and bends with no resistance. Its only possible displacement is transverse to its alignment.

The displacement of the string in a direction transverse to its direction is a field defined on all the points along the string. This field is much simpler than the electromagnetic field which is a field composed of two vector quantities, the electric and magnetic forces. The string field also obeys a simple mechanical rule for its dynamics.

Like most mechanically based systems the dynamics of the string has two simple sources, energy of the motion of its masses and a potential energy that is due to its configuration. For the case of a string held tightly with a tension T_e and with only transverse displacements, the potential energy is the work associated with making the string longer. The displacement of the string in the transverse direction is the field that we will consider and any non-zero displacement causes the string to be longer and thus changes the potential energy. These are global approaches to the behavior of the string and will be useful to us later when we use a more universal approach to dynamics based on a concept called action, see Section ?? on page ?? . For now because our goal will be an understanding of the electromagnetic field, we will use a more local approach and find that the electromagnetic field has many of the same properties as this the simplest of fields. In this approach the electromagnetic field is just a more complex field and the complications do not add any to the understanding of the field nature of the system. For example, the stretched string is a one dimensional field defined on a one dimensional manifold, the distance along the direction of the string. The field variable, $y(x, t)$, is also simple in that it is the transverse displacement

of the string from its equilibrium position where x is the position along string. Both y and x range over a one dimensional range of values. The electromagnetic field is a pair of vectors in its field variable and it ranges over a three dimensional manifold, space.

You may also be perplexed by the idea of a stretched string under tension. Our experience is that a string has to be fastened to be under tension. If that is the case, think of the string as tightly stretched between fixed walls. The problem with this is that the walls add complications of their own and for the first pass are not necessary. Here we deal with an infinite string under tension. Later, we will deal with the walls, see Section 8.3 on page 210.

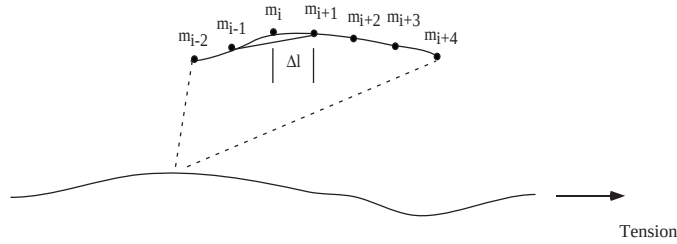


Figure 5.1: **The Stretched String** A string that can move in the transverse direction under tension is a simple example of a local field. In the figure, a section is magnified. In this section, the string is divided into small segments of length Δl and the mass of each segment is concentrated at a point. The dynamic of the string is that the mass at segment at location i has a force on it if its transverse displacement is different from the average of its two neighbors. Thus in the case shown, by drawing a straight line between masses at $i - 1$ and $i + 1$, we can see that at the place of segment i , the neighbors' average is below i 's current position. Thus i has a downward force on it.

The local statement of the dynamics of the string are easy to understand; the rule is very simple and intuitive: The force on a segment of the string caused by the transverse displacement of that piece of the string is proportional to the negative of the average of the displacement of that segment of the string from the displacement of its neighbors.

In order to implement this algorithm, divide the string into small segments of length Δl and concentrate the mass in the segment at a point, see Figure 5.1 on page 129. In the example shown, the segment of string labeled i is above the position of the average of its two neighbors. Thus there is a force to bring it to the position of the average. The proportionality constant for this force has the dimensions of a force per unit length and is thus the

twice the tension in the string divided by the length of the segment of string; twice since both neighbors pull. ρ is the mass per unit length of the string and thus the mass of each segment is $\rho\Delta l$. Using $\vec{F} = m_i\vec{a}_i$ and using the position along the string x as label for the piece of string, the transverse displacement of the string at x is $y(x, t)$, the average of the two neighbors of x is $\frac{y(x+\Delta l, t) + y(x-\Delta l, t)}{2}$, the force equation for the segment at x is

$$\rho\Delta l a_y(x, t) = -\frac{2T_e}{\Delta l} \left[y(x, t) - \frac{y(x + \Delta l, t) + y(x - \Delta l, t)}{2} \right], \quad (5.1)$$

where T_e is the tension in the string. This equation can be reorganized to

$$\rho\Delta l a_y(x, t) = T_e \left[\frac{y(x + \Delta l, t) - y(x, t)}{\Delta l} - \frac{y(x, t) - y(x - \Delta l, t)}{\Delta l} \right], \quad (5.2)$$

but this is the slope¹ of the transverse displacement curve at $(x + \frac{\Delta l}{2}, t)$,

$$\frac{y(x + \Delta l, t) - y(x, t)}{\Delta l} = \text{Slope}[y](x + \frac{\Delta l}{2}, t) = \frac{\Delta y}{\Delta x}(x + \frac{\Delta l}{2}, t). \quad (5.3)$$

Similarly,

$$\frac{y(x, t) - y(x - \Delta l, t)}{\Delta l} = \text{Slope}[y](x - \frac{\Delta l}{2}, t) = \frac{\Delta y}{\Delta x}(x - \frac{\Delta l}{2}, t). \quad (5.4)$$

Thus,

$$\begin{aligned} \rho a_y(x, t) &= \frac{T_e}{\Delta l} \left[\text{Slope}[y](x + \frac{\Delta l}{2}) - \text{Slope}[y](x - \frac{\Delta l}{2}) \right] \\ &= T_e [\text{Slope}[\text{Slope}[y]](x, t)], \\ &= T_e \left[\frac{\Delta}{\Delta x} \left(\frac{\Delta}{\Delta x} (y)(x, t) \right) \right]. \end{aligned} \quad (5.5)$$

Since

$$a_y(x, t) \equiv \frac{\Delta v}{\Delta t} = \frac{\Delta}{\Delta t} \left(\frac{\Delta y}{\Delta t} \right) (x, t), \quad (5.6)$$

we have

$$\rho a_y(x, t) = \rho \frac{\Delta}{\Delta t} \left(\frac{\Delta y}{\Delta t} \right) (x, t) = T_e \left[\frac{\Delta}{\Delta x} \left(\frac{\Delta y}{\Delta x} (x, t) \right) \right]. \quad (5.7)$$

¹In the limit $\Delta l \rightarrow 0$, this is the derivative of $y(x, t)$ with respect to x which is one of the two variables of y and thus the partial derivative of $y(x, t)$ with respect to x , usually designated $\frac{\partial y}{\partial x}(x, t)$. We actually stick with the idea of *Slope* and designate it as $\frac{\Delta y}{\Delta x}(x, t)$.

In this last equation, the term on the right is the definition of the second partial derivative of $y(x, t)$ with respect to x . This operation of taking the *Slope* of the *Slope* of the field at a place is very common in field dynamics and it is worthwhile giving it a separate identification. Let's define the operation $\frac{\Delta}{\Delta x} \frac{\Delta}{\Delta x}$ as the *Bendy* operation². In this convention we can rewrite the string dynamic, Equation 5.8, as

$$\rho \frac{\Delta}{\Delta t} \left(\frac{\Delta y}{\Delta t} \right) (x, t) = T_e [Bendy [y] (x, t)]. \quad (5.8)$$

Note also that the acceleration of the string in the transverse direction is the second derivative with respect to time. In the limit that Δl is zero and using partial derivatives because we have both x and t dependence, this force equation becomes

$$\rho \frac{\partial^2 y}{\partial t^2} (x, t) = T_e \frac{\partial^2 y}{\partial x^2} (x, t). \quad (5.9)$$

$$= T_e Bendy_x [y] (x, t). \quad (5.10)$$

This is an excellent example of the general form in which the dynamics of fields are expressed. They are generally partial differential equations because we are interested in how the field changes for changes in position and time. Equation 5.10 on page 131 is second order in the time derivatives because that is how the dynamic operates; it emerged from a mechanical force law. Other orders of time derivatives are possible and it is not uncommon to have laws that are first order in time. In fact, it is preferable because the interpretation of the evolution is simpler. Maxwell's Equations for electromagnetism, Equations 5.26 through 5.29 on page 148, are an example. The stretched sting or any higher order temporal evolution can be reduced to a first order temporal evolution by defining new fields. Defining a new field, $v(x, t) \equiv \frac{\partial y}{\partial t}$, we can get an evolution that has only first time derivatives.

$$\begin{aligned} \frac{\partial y}{\partial t} (x, t) &= v(x, t) \\ \rho \frac{\partial v}{\partial t} (x, t) &= T_e \{Bendy_x (y(x, t))\} = T_e \frac{\partial^2 y}{\partial x^2} (x, t). \end{aligned} \quad (5.11)$$

²There is a formal mathematical construction that is closely related to the Bendy operator called the curvature operator, $\kappa \equiv \frac{\Delta}{\Delta s} \left(\frac{\Delta}{\Delta s} \right)$, where s is the arc length along the curve. When increments in $\Delta s \equiv \sqrt{(\Delta y)^2 + (\Delta x)^2}$, and Δx are about the same, Bendy and curvature are about the same. This occurs when the slope, $\frac{\Delta y}{\Delta x}$, is small compared to one. Also note that $Bendy_x \equiv \frac{\partial^2}{\partial x^2} (x, t)$ in the continuum limit where you read $Bendy_x(x, t)$ as the bendy of the field at the place x at time t .

In a very real sense, you could say the the magnetic part of the electromagnetic system is a manifestation of this kind of substitution.

The fact that there are only values of the field and spatial derivatives of the field on the right side of the Equation 5.10 on page 131 is the expression of the locality of the dynamic. How the field evolves at a place depends only on what is going on at that point. Also note that the only parameters in the field equation are ρ and T_e . These express the intrinsic properties of the medium in which the field operates. By dividing Equation 5.10 on page 131 by ρ , we can reduce the effective number of parameters to one, $\frac{T_e}{\rho} \stackrel{\text{dim}}{=} \frac{L^2}{T^2}$. This has the dimensions of a velocity squared. The fact that there is only this parameter in the dynamic says a great deal about the nature of the evolution of the fields. There are not enough parameters to construct a length or a time. Thus for this field there is no intrinsic size except as it is put in by the starting conditions or put into the problem by boundaries like walls. Thus this particular field system, the stretched rope, is characterized by movement of field configurations. Since the parameter of the medium is a velocity squared, the movement is in both directions with a characteristic speed, $\pm\sqrt{\frac{T_e}{\rho}}$. It is important to remember that the movement of a piece of string is only in the transverse direction whereas the movement of the field configurations is along the direction in which the string is aligned. This is a difficult situation to describe. If you attribute all reality to the hunk of string the only motion is up and down in the transverse direction. Yet the configuration of the string moves along the string. We will find that there is energy and momentum associated with the configuration of the string and that this thus moves with the configuration along the string. Thus we have the problem of the ‘string’ only moving up and down but energy and momentum flowing along the string.

The converse of the above result that the parameters of the system are not sufficient to determine a size or time scale is that the medium, in the case of the stretched string are ρ and T_e , implies that the disturbances in the string travel with a speed set by the medium, $\sqrt{\frac{T_e}{\rho}}$ and that this speed is independent of the form of the disturbance. In other words disturbances travel with speed $\pm\sqrt{\frac{T_e}{\rho}}$ without distortion. For this reason, systems with this field dynamic are called wavelike. This is the definition of a wavelike medium. Although many systems are wavelike such as sound and light, other field systems may not be. For instance the dynamic for temperature

flow in one spatial dimension is

$$\frac{\partial T_{emp}}{\partial t}(x, t) = a^2 \{Bendy_x(T_{emp}(x, t))\} = a^2 \frac{\partial^2 T_{emp}}{\partial x^2}(x, t). \quad (5.12)$$

where a^2 is called the diffusion constant and is the ratio of the heat conductivity to the heat capacity of the material. Notice that $a^2 \stackrel{\text{dim}}{=} \frac{L^2}{T}$ and thus there is no special speed or length or time that is characteristic of the field.

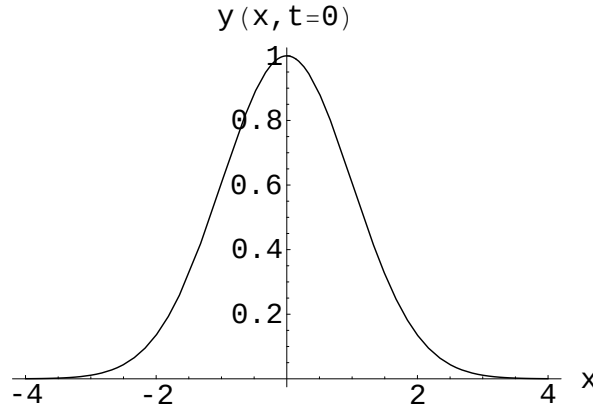


Figure 5.2: **A Simple Displacement Pulse in a String** A simple pulse in a stretched string under tension. At $t = 0$, the string is distorted but no part of the string moving.

In order to better understand the operation of field dynamics let's work through the example of the string under tension. Consider our case of a stretched string with mass per unit length ρ and tension T_e . At $t = 0$, we put a distortion in the string as shown in Figure 5.2 on page 133. Note that at $t = 0$, the string is displaced but no part of the string is moving. It is simplest to interpret the operation of the dynamic in the first order time derivative form, Equation 5.11 on page 131. In this form, it is clear that a complete description of the initial configuration of the string involves the specification of two fields, the initial velocity field and the initial displacement field. In other words for the case in Figure 5.2 on page 133 at $t = 0$, the velocity of all parts of the string is zero and there is a simple pulse of displacement in the string. Other starting configurations are possible. You could have the situation in which the string has no displacement and the string has a distribution of transverse velocity. The difference in the operation of a harpsichord and piano is the the strings are plucked or distorted in the

harpsichord and hammered in a piano. You can also have situations with both an initial displacement and velocity.

The dynamic of the string requires that all points on the string be at the average of its neighbors. An easy way to compute the average is to pick two neighbors, points on the string close to the point of interest and equidistant from it, and connect the points by a straight line. At the point of interest, x , the point on the line is the average of the two neighbors. Thus from Figure 5.3 on page 134, we see that the center of the string is pulled strongly down and the edges are pulled up. The points of steepest drop are not pulled at all. This last point is interesting to note. The string is not pulled to the neutral position. Each segment is pulled only by its neighbors. If the string were pulled to the neutral position there would be a force for the entire time of descent and then the string would still have a velocity when it reached the neutral position and thus would overshoot and there would be oscillation at each disturbed point on the string. As we know, the disturbance in the stretched string is removed by the dynamic with the string returning gently to its neutral position.

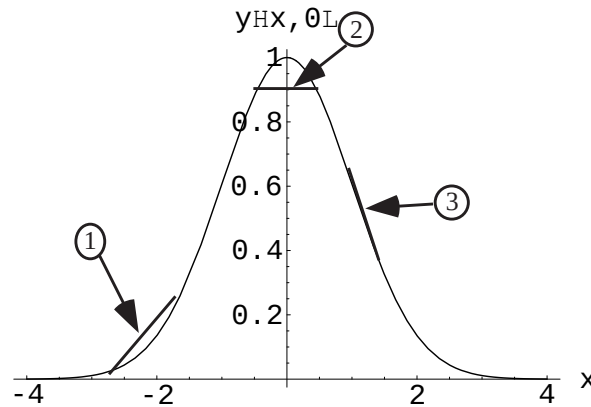


Figure 5.3: **Forces on a Pulsed String** The dynamic of the stretched string require that all points in the string be at the average of its neighbors. A simple rule for finding the force and thus the acceleration of a place on the string is to connect the neighbors with a straight line. If the string at that place is above the line, there will be a downward acceleration with magnitude proportional to the distance above. There are three examples shown. At a point on the edge of the pulse, 1, the string is accelerating upward. At the center, 2, the string is accelerating down. At a point at the midpoint of the side of the pulse, 3, the string has no acceleration.

To make this discussion more quantitative, we look at what goes on in a few small time increments. In a small time, Δt , since the velocity field is initially zero everywhere, we find that the string has not moved.

$$\begin{aligned} y(x, \Delta t) &= v(x, 0)\Delta t + y(x, 0) \\ &= y(x, 0), \end{aligned} \quad (5.13)$$

where $v(x, t)$ is the velocity of the string at the point labeled x at time t . At $t = 0$, the string is not moving and $y(x, 0)$ is known.

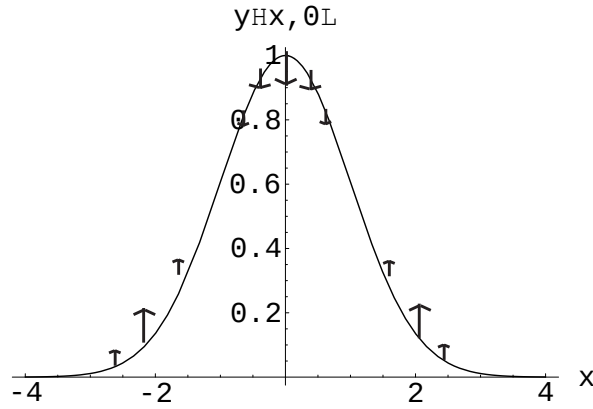


Figure 5.4: **Accelerations on a Pulsed String** Using a technique such as shown in Figure 5.4 on page 135 for the forces on the string, the algorithm in Equation 5.2 on page 130 can be applied at each point, x , and find the accelerations shown as arrows above.

We will need the velocity of the string at all times and, even in a small time, because of the forces from Figure 5.3 on page 134, the velocity changes.

$$\begin{aligned} v(x, \Delta t) &= a_{t=0}(x)\Delta t + v(x, 0) \\ &= a_{t=0}(x)\Delta t \end{aligned} \quad (5.14)$$

where we find $a_{t=0}(x)$ from an analysis such as that shown in Figures 5.3 on page 134 for each point on the string. Thus we see that after a time Δt the velocities will have the same pattern as a function of position as the initial accelerations.

Repeating the process for a second Δt using Equations 5.13 on page 135 and 5.14 on page 135 but with the time shifted another increment,

$$v(x, 2\Delta t) = a_{t=\Delta t}(x)\Delta t + v(x, \Delta t)$$

$$\begin{aligned}
&= a_{t=0}(x)\Delta t + a_{t=0}(x)\Delta t \\
&= 2a_{t=0}(x)\Delta t
\end{aligned} \tag{5.15}$$

where in the second line, I used the fact that since $y(x, \Delta t) = y(x, 0)$ and the accelerations depend only on $y(x, t)$, then $a_{t=\Delta t}(x) = a_{t=0}(x)$.

The second dynamic is handled similarly,

$$\begin{aligned}
y(x, 2\Delta t) &= v(x, \Delta t)\Delta t + y(x, \Delta t) \\
&= a_{t=0}(x)\Delta t^2 + y(x, 0).
\end{aligned} \tag{5.16}$$

We now begin to see the string moving.

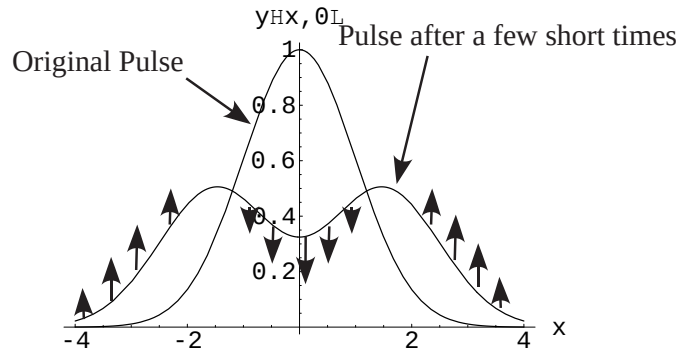


Figure 5.5: **Pulsed String after a Few Short Times** Using appropriate versions of Equations 5.13 on page 135 and 5.14 on page 135 to evolve the system, we can see the development of two pulses. Also shown are the velocities by scaled arrows. Remember the parts of the string are only free to move up and down but the pattern of up and down motion conspires to produce the effect that the pulse at negative x is moving toward greater negative x and the pulse at positive x is moving toward greater positive x . The original pulse is shown for comparison.

We can intuit that the pattern shown in Figure 5.5 on page 136 develops. The region where there is a strong bend at the edge it is pulled up and so has an upward velocity and begins to lift. The middle section is unchanged at first. The center is forced down and has a downward velocity. Because of the pattern of the upward velocity at the bends and the downward velocity at the center, the two separating pulses appear to be moving along the string away from each other. We have to remember that the all the motion of the string is transverse to its direction.

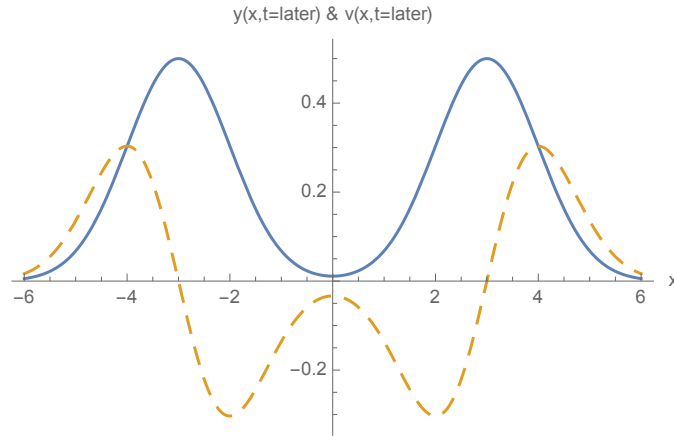


Figure 5.6: **Pulses in String Separating** After a time, the pulse initially placed on a stretched string, see Figure 5.2 on page 133, separates into two half amplitude pulses. One travels to the left with velocity $v = -\sqrt{\frac{T_e}{\rho}}$ and one travels to the right with velocity $v = \sqrt{\frac{T_e}{\rho}}$. There is also a transverse velocity field that travels along with each pulse shown as the dashed curve instead of using arrows as in Figure 5.5 on page 136.

The general pattern then develops of two distinct pulses of half the original amplitude one moving to the left and one to the right, see figure 5.6 on page 137. This transverse velocity is patterned so that the two emergent pulses are one moving to the left with speed $-\sqrt{\frac{T_e}{\rho}}$ and one moving to the right with speed $\sqrt{\frac{T_e}{\rho}}$. Each of these are called traveling waves, one to the left and one to the right. It is the pattern of traveling waves that there is both a transverse displacement field and an associated transverse velocity field with the velocity field rising in front of the motion of the traveler and falling behind the traveler. This is the pattern for field media that supports travelers. There are two fields that support each other and form the traveling configuration. For sound the two fields are the density of the air and the pressure in the air. For electromagnetic waves, it is the electric and magnetic force fields.

It is also important to realize that existence of any of these traveling disturbance do not accelerate to get to the traveler up to speed but that the speed is intrinsic and an integral part of the definition of the disturbance

which travels along the stretched string. Also remember that this speed is along the string not transverse like the velocity field.

It is worthwhile to also note that our original configuration of the displacement pulse with no velocity, Figure 5.2 on page 133 can be considered as the sum of two travelers, one going to the left and one going to the right, each of half amplitude. The addition of the displacement field gives the correct shape for the pulse and, at the instant of complete overlap, the initial instant, the two transverse velocity fields add to zero. The ability to treat the original distortion as a sum of two independent distortions is an example of superposition. This will be an important principle in many future discussions, see Section 19.8 on page 439.

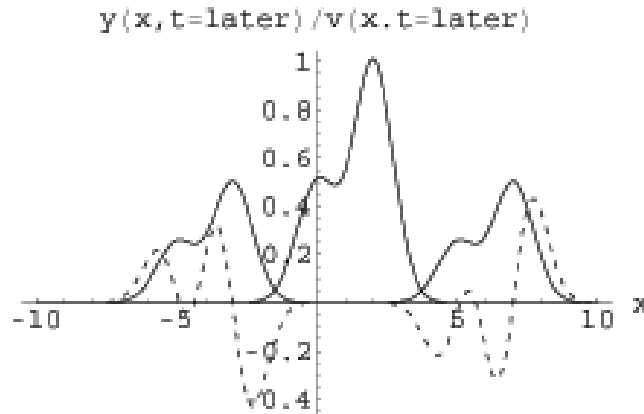


Figure 5.7: **Arbitrary Traveling Waves** Using a more general form for the initial distortion of the string, shown at the center for reference, we see at a later time the two traveling distortions, one moving to increasing x called the right traveler and one moving to decreasing x called the left traveler. The associated velocity profile for each is shown dotted. Because of the special form of the argument of the travelers, the velocity profile for the right traveler is proportional to the negative of the slope of the displacement profile of the right traveler at that instant and the velocity profile of the left traveler is proportional to the slope of the displacement profile of the left traveler at that instant.

In addition, the travelers have an interesting relationship between the displacement field and the velocity field. For a traveler that moves to increasing x , the argument of the displacement field is a single variable, $x - \sqrt{\frac{T_e}{\rho}}t$, instead of x and t as independent variables. This traveler is called a right

traveler. For waves that move to decreasing x , called left travelers, the argument is $x + \sqrt{\frac{T_e}{\rho}}t$. This is what makes them travelers; they move to increasing x or decreasing x uniformly without the shape of the disturbance changing. This is a general result and true for all one dimensional wavelike systems. We worked this out for the particular disturbance of Figure 5.2 on page 133, a simple pulse. It should be clear that this pattern of two separate travelers superposing to produce an initial distortion with no velocity field will hold for any form of distortion for the displacement field. Figure 5.7 on page 138, shows a more general initial configuration and the subsequent travelers. Because to the nature of the relationship between the x and t variables in the travelers, $x - \sqrt{\frac{T_e}{\rho}}t$ for the right traveler and $x + \sqrt{\frac{T_e}{\rho}}t$ for the left traveler, the time evolution of the displacement field which is the velocity field in this dynamic is related to the slope of the displacement of the traveler at that point.

$$\frac{\partial y_{rt}}{\partial t}(x, t) \equiv v_{rt}(x, t) = -\sqrt{\frac{T_e}{\rho}} \frac{\partial y_{rt}}{\partial x}(x, t) \quad (5.17)$$

where $y_{rt}(x, t)$ and $v_{rt}(x, t)$ are the right traveling waves displacement field and velocity field. The relationship of the velocity field and the displacement field for the left traveler is similarly:

$$\frac{\partial y_{lt}}{\partial t}(x, t) \equiv v_{lt}(x, t) = -\sqrt{\frac{T_e}{\rho}} \frac{\partial y_{lt}}{\partial x}(x, t). \quad (5.18)$$

Another feature of the travelers is that they carry energy and momentum. It takes a certain amount of work to distort the string; it has to become longer. This distortion energy is then distributed into the travelers and these then carry it off to remote regions of the string. Similarly there is momentum associated with the travelers that is transported down the string by the travelers. In a later section, we will develop a more nuanced identification for momentum and energy, see Section ?? on page ?? but for now our intuitive ideas will suffice. Notice that this is energy and momentum that moves down the string even though the string itself can only move in a transverse direction. Thus the traveler wave configurations act like a thing that moves along the string even though nothing moves down the string. Note that a superposition of the travelers constitute the original disturbance. Here we begin to see the development of a thing, something that carries energy and momentum, in the context of a field. The electromagnetic field is a wave field and will have travelers also. These are more complex and

constituted differently in the dynamic than these string travelers but they behave similarly. Since they generally operate in three spatial dimensions there is a geometric fall off in strength as they travel but they still carry energy and momentum to remote parts of the system.

In Section 8.3 on page 210, The Stretched String Revisited, we will return to the dynamics of the string. For now we are content to use it as a simple example of a field system and to have it express the basic ideas of a field theory, a construction that is a local causal dynamical system. In the next section, we will discuss Maxwell's Equations, the first fundamental theory based on a field construction.

5.4 Maxwell's Theory of Electromagnetism

The theory of the electromagnetic field is so important to the development of modern physics that it has to be treated in a separate section. It is especially important to us because of its special role in clarifying the nature of light. It was not why Maxwell started his study of the electric and magnetic forces but it was his most striking discovery.

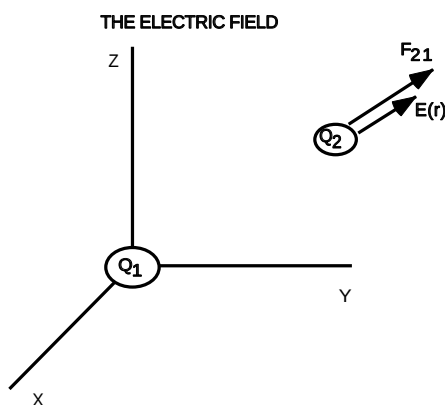


Figure 5.8: **The Electric Field and Electric Forces** Maxwell said that electric and magnetic forces were due to the presence of the electric and magnetic field. In this figure, the electric force on Q_2 is due to the presence of the field at its location, $\vec{F}_{21} = Q_2 \vec{E}(\vec{r})$. There is a similar relationship for the magnetic force.

Maxwell was interested in developing a unified description of electric and magnetic phenomena. In his time, many of the basic ideas of the electric and magnetic force systems were known. The law for the electric interaction

between charged particles had been articulated in the period 1785 and 1791 by Coulomb. The force law between magnets and the force between moving charges and magnets was known and even Faraday's Law about the relationship of changing magnetic environments and electric currents was known. In fact, Faraday had already begun to describe magnetic and electric phenomena in a field like language. What Maxwell sought was an underlying mechanical basis for all the phenomena associated with electricity and magnetism. Reducing electricity and magnetism to a mechanical basis meant that he was looking for something to push or pull but it had to do so locally. He could not believe that fundamental phenomena could take place as an action at distance phenomena like gravity was thought at the time. In order to have a thing which could push or pull locally, he hypothesized the existence of a rather rich structure for the vacuum of space, whirling vortices in an ether that produced the electric and magnetic force. Thus not only did he seek a mechanical source for electric and magnetic phenomena, he developed a field theory basis for it. His picture of electric and magnetic forces was that they were mediated by fields, the electric, $\vec{E}$, and the magnetic, $\vec{B}$, fields. It was his basic idea that the correct description of electromagnetic phenomena required a locally causal dynamic. The idea was that not only did the charges generate the fields but the fields themselves responded to the local environment of the fields themselves. In addition, the forces experienced by the charges were because of the values of the fields at the place occupied by the charges, $\vec{F} = q\vec{E} + q\vec{v} \times \vec{B}$, where q is the charge in question and $\vec{v}$ is its velocity.

In order to create the mechanical basis for the fields, Maxwell was forced to endow the ether with the correct mechanical properties of inertia and size to replicate the success of the earlier laws but now in context of a local mechanical model. The underlying idea was simple. Let's look at the simplest of the cases, Coulomb's Law. The situation is shown in Figures 5.9 on page 142 and 5.10 on page 143. A force on a charged particle took place as a two step process. A charge Q_1 is placed in empty unexcited space. This charge excites the ether next to it by creating vortices at its location. These vortices in turn excite neighboring vortices until space is full of whirling vortices. Each vortex is in dynamic equilibrium with its neighbors. There is a 'thing', the whirliness, which is a measure of the electric field at that point.

When a new charge, Q_2 , is located at some distance, $\vec{r}$, from the first charge, it detects the level of excitement of the local vortices and thus feels a corresponding force. The force is proportional to the charge Q_2 at that place and the amount of whirliness or electric field at that point.

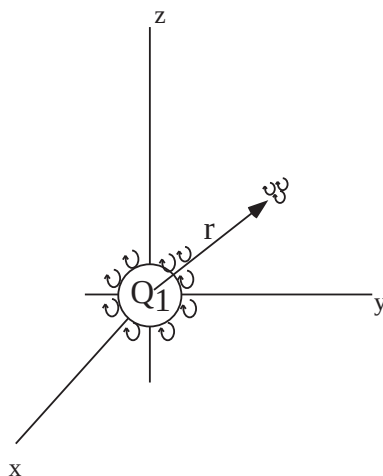


Figure 5.9: **Maxwell's Vortices** Maxwell pictured the electric force as emerging in two steps. First any charged particle would excite vortices in the ether at its location. These vortices would excite other vortices nearby and so forth until all of space would fill with whirling vortices. In a sense, the whirliness of the vortices at any place was a measure of the strength of the electric field at that point.

The mechanical properties of the ether and its vortices determine how the whirliness develops. This is set by the vortices inertia and size. These parameters for the mechanical properties of the vortices are then adjusted to accommodate Coulomb's Law.

In other words, Maxwell introduced local fields – a continuous quantity defined at all points in space and for all times – with a rule of dynamics to produce the electromagnetic forces. If an object experiences a force, there must be something at that place, the whirliness. In addition, the whirliness itself must be determined locally in both space and time. Let's go through the example of Coulomb's Law in a little more detail to see how this idea works.

The first problem is to reproduce the well known Coulomb's law of force for static situations. Coulomb's Law is an action at a distance description of interaction,

$$\vec{F}_{21} = \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{r_{12}^2} \frac{\vec{r}_{21}}{r_{12}} \quad (5.19)$$

where $\vec{r}_{12}$ is the separation between the charges. In order to simplify the discussion, let's place charge Q_1 at the origin. Since the force on the charge

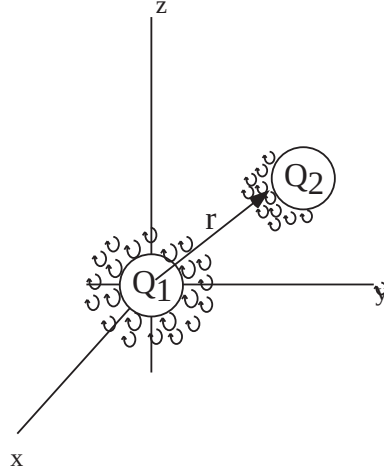


Figure 5.10: **Vortices and the Electric Force** When a charged particle, Q_2 , is positioned, the particle detects the local amount of whirliness in the vortices of the ether. This generates the electric force in proportion to the charge and amount of whirliness at its location. The local whirliness is $\vec{E}$ at $\vec{r}$.

Q_2 is supposed to be $Q_2 \vec{E}(\vec{r})$, where $\vec{r}$ is now the position at which Q_2 is located. For this case, we can identify the electric field as

$$\vec{E}(\vec{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q_1}{r^2} \frac{\vec{r}}{r} \quad (5.20)$$

around a spherically symmetric charge placed at the origin. You will reproduce the static Coulomb's Law results with the electric field if you can make a local rule about how $\vec{E}(\vec{r})$ develops that reproduces this result. It should be clear that the hard part will be to reproduce the inverse square fall off with distance in the strength of the field.

In some sense, it is really not correct to say that Q_1 is the source of this field. The field is not attached to the charge. At any point, there is a field only if there is a field or a charge in the neighborhood. The field at some point, like all things, is to be determined locally. Maxwell used his whirling vortices of the ether to discover a rule for whirliness and how whirliness effected whirliness that recovers the characteristic the inverse square fall off with distance of Coulomb's Law. Like the stretched sting, Section 5.3 on page 128, in which the transverse position of a place on the string is determined by the transverse position of the neighbors to that place,

similarly here, the idea is to find the rule on how the field arranges itself and forget about the whirlies. The following analysis reviews the process and becomes somewhat technical but the struggle to follow it is worth the effort.

Since the electric field is meant to produce a force, it must be a vector field, a directed quantity defined at every point in space and with a local rule for its construction. Basically you ask how much does the field change at a place because of what is there. For now, we are looking at a static case – no time change. But we can still ask about how the field varies as we change positions in space.

For a vector field such as the electric field since it is a vector field, you have a directed strength at each point in space and around each point you have directed strengths. At any point you can ask how much more “out pointy” these directed strengths become as you go from place to place. The analogy for our stretched string is that, at any place on the string, you can ask how “*Bendy*” is the string. On a string, “bendiness” happens when that place differs from its neighbors on each side, see ref. The string bends up when the place is lower than its neighbors and it bends down when it is higher. When there is no bend, that place on the string is at the average of its neighbors. In the static string it takes a force to maintain a bend in the string. Our case for the vector field case using “out pointiness” works in the same fashion. You can have “out pointiness” only if there are charges that are placed there, i. e. charge causes an outward directed field. Of course, we have to develop a definition, a measure, of “out pointy” and test it.

The measure of “out pointiness” is called the divergence and it is what you would have thought to define it as if you spent some time playing with the ideas of a vector field. At any point, find out how much the neighboring fields point away from where you are. That should indicate the “out pointiness”.

Since fluid flow is also a vector field it is worthwhile to think in terms of it. The vector field in this case is the velocity of the fluid. If at a point all the flow is uniform about you, you would not think of the field as becoming “out pointy”. On the other hand, if you were at a place like the drain, you would consider the surrounding flow to be “in pointy”, the opposite of “out pointy”. To be more quantitative, think of surrounding the place that you are interested in and measuring how much stuff flows in or out. By enclosing the point of interest with a surface, we can measure the incoming fluid by assessing how much stuff comes into any element of area on the surrounding surface and then adding the contribution to each part. In other words, surround the point with a surface. Cover it with

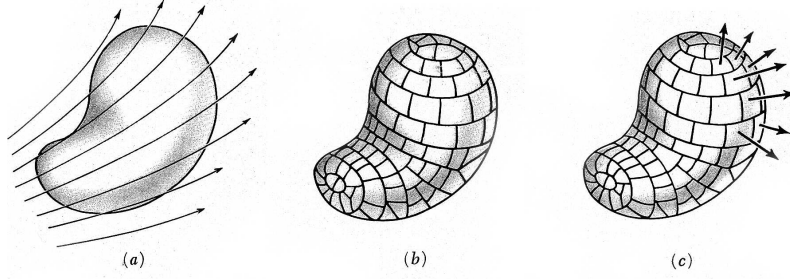


Figure 5.11: **Construction of the Divergence** To find the divergence or “out pointiness” of a vector field at a point, surround the point with a surface, step (a). Cover the surface with small elements of area so that to all intents and purposes they can be considered flat. Each element of surface will now have a normal vector. Find the magnitude of the vector field at the surface to is along the surface. Add these magnitudes for each element of surface and the total is the divergence or “out pointiness” of the vector field at the point surrounded by the surface. Then shrink the volume surrounded to a point. For a fluid, applied to the velocity field, this tells the amount of fluid that goes into a point. This series of steps is encoded in the first part of Equation 5.21 on page 145 for the case of the electric field.

elements of area, postage stamps labeled Δ^2s . Each element of area has a normal vector, $\hat{s}$, that points either outward or inward, see Figure 5.11 on page 145. Choosing the outward normal, we are defining “out pointiness”, the amount of the vector field that is along the outward normal; it is the “flow” through that element of area. Now do this for the each element of the entire surface and add up all the contribution from all the pieces. This process is summed up with the symbol $\sum_{S \supset V}$. $\sum$ is our usual symbol for adding pieces. The subscript $S \subset V$ means for the surface bounding the volume V . To reduce this analysis to a point, shrink the volume enclosed by the surrounding surface to zero, $\lim_{V \rightarrow 0}$. This same analysis holds for all vector fields. This construction at each point assesses the “out pointiness” of the neighborhood of the point and is called the divergence. Thus,

$$\text{Div}(\vec{E}(\vec{r})) \equiv \lim_{V \rightarrow 0} \frac{\sum_{S \supset V} \vec{E}(\vec{r}') \cdot \Delta^2 \vec{S}}{V} = \lim_{V \rightarrow 0} \frac{1}{\epsilon_0} \frac{Q_{\text{inside } V}}{V} = \frac{1}{\epsilon_0} \rho(\vec{r}) \quad (5.21)$$

where the first part is a mathematical statement of what is stated above but for the case of the electric field and the subsequent parts are the relationship with charge that is necessary to recover Coulomb’s Law, i. e. electric charge

is the source of “outpointiness”.

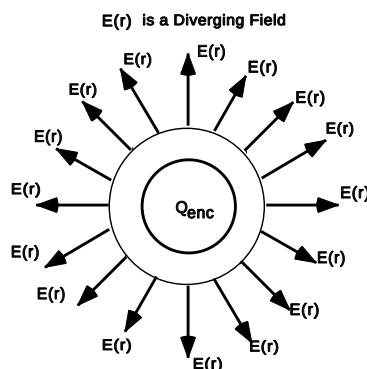


Figure 5.12: **”Outpointiness of the Electric Field”** A characteristic property of the electric field is that charge is the source of “outpointiness”. This is the idea that the electric field points away from nearby positive charges and toward nearby negative charges. This last example being negative outpointyness.

Notice that this law, Equation 5.21 on page 145, says that for a static electric field there is divergence of the field only where there is charge. Yet the picture that we all have of the static electric field around an isolated point charge is a diverging field, the electric field points outward from the origin everywhere, see Figure 5.12 on page 146. How do we reconcile this?

Consider a point away from the isolated point charge. If a surface such as that shown in Figure 5.11 on page 145 is constructed, the area nearer the charge is smaller whereas the area more distant is larger. In fact, the areas are in the ratio of the distances squared. Thus the field strength and the areas combine so that the net “out-pointiness”, actually in-pointiness, of the nearer surface balances the out-pointiness of the far surface and the net is zero out-pointiness. Thus it is because the divergence is zero at places other than the charge that the field strength falls off with distance as $\frac{1}{r^2}$.

Another property that a vector field can manifest is rotation or curl. Again you develop a definition and test it. Here the idea is to follow a closed path around the point and see how much of the vector field follow the path. The electric field does not curl.

$$\text{Curl}(\vec{E}(\vec{r})) \equiv \lim_{S \rightarrow 0} \frac{\sum_{p \in S} \vec{E}(\vec{r}') \cdot \Delta \vec{r}'}{S} = 0 \quad (5.22)$$

On the other hand, the magnetic field does curl. The magnetic field is

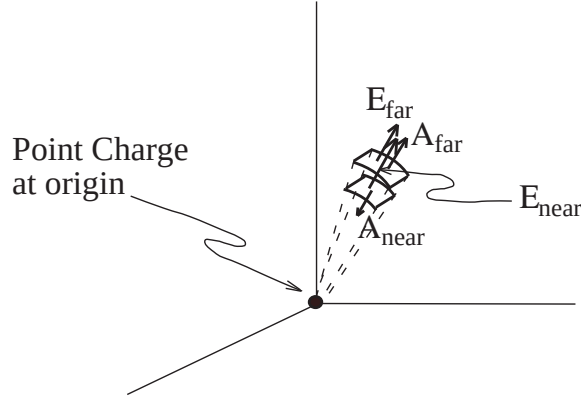


Figure 5.13: **Divergence outside Charge** A characteristic property of the electric field is that charge is the source of “outpointiness”. This is the idea that the electric field points away from nearby positive charges and toward nearby negative charges. This last example being negative outpointyness.

the force experienced by a moving charged particle.

$$\vec{F}_{\text{mag}} = Q\vec{v} \times \vec{B}(\vec{r}) \quad (5.23)$$

The magnetic field lines tend to wrap around their sources, the currents.

$$\text{Curl}(\vec{B}(\vec{r})) = \lim_{S \rightarrow 0} \frac{\sum_{p \in S} \vec{B}(\vec{r}') \cdot \Delta \vec{r}'}{S} = \lim_{S \rightarrow 0} \frac{1}{\mu_0} \frac{\mathbf{i}_{\text{enc } p}}{S} = \frac{1}{\mu_0} \vec{j} \quad (5.24)$$

and does not diverge

$$\text{Div}(\vec{B}(\vec{r})) = 0 \quad (5.25)$$

Note that we have not added a time dependence. These are all static situations.

Maxwell insisted that the field was not established everywhere at once. It was made up of whirling vortices that pushed on each other. The rate at which the vortices could push was set by the parameters of the static theory. By endowing these whirling vortices with the correct properties to reproduce the laws of static electricity and magnetism, he found how to add a local set of rules for the time evolution of the fields. These are the full set of Maxwell’s equations including time dependence:

$$\text{Div}(\vec{E}(\vec{r}, t)) = \frac{1}{\epsilon_0} \rho(\vec{r}, t) \quad (5.26)$$

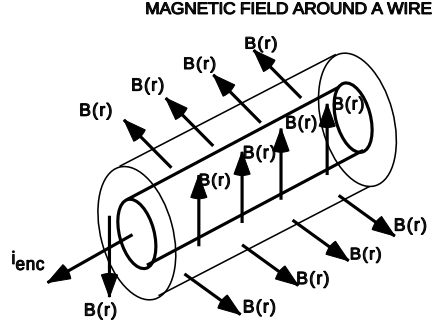


Figure 5.14: **The Curl of the Magnetic Field** In contrast to the electric field, the magnetic field wraps around or curls around its sources, the currents in the problem.

$$\text{Curl}(\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)) = \frac{\partial \vec{\mathbf{B}}}{\partial t}(\vec{\mathbf{r}}, t) \quad (5.27)$$

$$\text{Div}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) = 0 \quad (5.28)$$

$$\text{Curl}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) = \mu_0 \vec{\mathbf{j}}(\vec{\mathbf{r}}, t) - \mu_0 \epsilon_0 \frac{\partial \vec{\mathbf{E}}}{\partial t}(\vec{\mathbf{r}}, t) \quad (5.29)$$

This is the standard format for these equations. For a discussion of the field dynamics, it is important to realize that only two of these equations are a dynamic, Equations 5.27, and 5.29. The other two equations, Equations 5.26 and 5.28 are what are called constraint equations; they control the pattern of the field but not the temporal evolution. It is important to note that the last term in Equation 5.29 had not been identified in the electromagnetic system until Maxwell used his vortices to make this theory. He needed the vortices to get the standard parts but he got a new piece that, as will be seen below, is essential to having traveler solutions for the electromagnetic field.

It is apparent that the electromagnetic field is a much more complex field than the stretched string whose dynamic is Equation 5.11 on page 131. The vector nature of the field, the existence of constraints, and the sources, $\rho(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{j}}(\vec{\mathbf{r}}, t)$, obviously complicate the situation. We could have added external forces to the dynamic of the string but that would not have clarified the wave field nature of the string. Similarly, here we can discuss the electromagnetic field without the presence of $\rho(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{j}}(\vec{\mathbf{r}}, t)$. Rearranging and omitting the sources, the dynamical equations for the evolution of the electromagnetic

field become

$$\frac{\partial \vec{\mathbf{E}}}{\partial t}(\vec{\mathbf{r}}, t) = -\frac{1}{\mu_0 \epsilon_0} \text{Curl}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) \quad (5.30)$$

$$\frac{\partial \vec{\mathbf{B}}}{\partial t}(\vec{\mathbf{r}}, t) = \text{Curl}(\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)) \quad (5.31)$$

with the constraint equations

$$\text{Div}(\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)) = 0 \quad (5.32)$$

$$\text{Div}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) = 0 \quad (5.33)$$

Comparing this dynamic with the stretched string dynamic, Equation 5.11 on page 131, we can identify $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ with the displacement field of the string, $y(x, t)$, and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$ with the velocity field of the string, $v(x, t)$ and we see that the electromagnetic dynamic is more complex but similar in structure. We also realize that the identification of the traveler velocity on the string, $\sqrt{\frac{T_e}{\rho}}$, plays a similar role to $\sqrt{\frac{1}{\mu_0 \epsilon_0}}$ the speed of the travelers in the electromagnetic case, the speed of light.

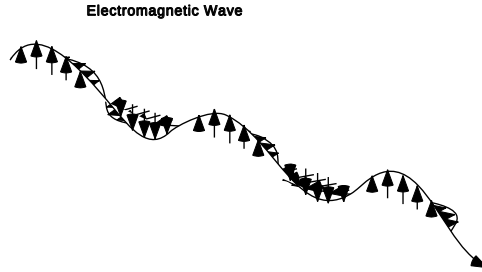


Figure 5.15: **The Field Configuration for Light** Light is a traveling wave solution of Maxwell's Equations and is composed of propagating combination of electric and magnetic fields. The direction of flow of energy and momentum is along the normal to the plane of the oscillating electric and magnetic field vectors. In the figure the upward arrows represent the electric field and the perpendicular arrows are the magnetic field.

An important feature of the electromagnetic field that can be seen from the equations above is that, if you have an electric field in a localized region of space, finite somewhere but zero elsewhere like the pulse in the stretched string, the electric field will have a curl. Thus even if there are no charges or currents, this curl is the source of a developing magnetic field, Equation 5.31.

This is like the case in the string of the displacement producing a velocity field. As the new magnetic field grows which will also be localized and thus curled, it produces a reduction in the original electric field, Equation 5.30. Thus the original field will start to reduce and there will be a growing magnetic field. This magnetic field will in turn change and produce a electric field. The relationship of the magnetic and electric fields is much like that of the velocity and displacement of the stretched string which produces traveling pulses, Section 5.3 on page 128. In fact, using Equations 5.30 on page 149 and 5.31 on page 149, in a region without charges or currents, the vacuum, you find that the electric and magnetic fields are a wavelike system and that a field configuration such as that shown in Figure 5.15 on page 149 produces a traveling wave that travels in the plane perpendicular to the plane of $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$ with a speed

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \quad (5.34)$$

which dimensionally is a speed and the only dimensional factor in the dynamic. This is the same result that Maxwell discovered with his whirlies. Putting the values of μ_0 and ϵ_0 this is the speed of light. If it walks like a duck and quacks like a duck, it is a duck and thus Maxwell concluded that light is the traveling wave solutions to the equations of electromagnetism.

It is important to realize that like in the stretched string which has only a transverse displacement and transverse velocity, the $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$ fields are not traveling but only the disturbance – changes in the field configuration. It is also important to realize that the velocity of the disturbance does not depend on the field configuration. It only depends on the dynamic of the field. Another way that this is often said is that the velocity of propagation is a function only of the medium. Since the electromagnetic field operates in the vacuum of space, it is the properties of the vacuum that determine the speed with which light propagates. A difference for the electromagnetic travels with the travelers of the field of the stretched string is that in the string any distortion will produce simply related travelers but for the electromagnetic field there are configurations of the field that do not have simply related travelers.

We now understand the amplitude that was invented by Young and Fresnel, see Section 4.6 on page 106. It is the electromagnetic field. The Fresnel construction is the general rule for the computation of the propagation of the light and holds for the traveling solutions of the electric and magnetic field dynamic.

5.5 Other Classical Field Systems

5.5.1 Temperature

5.5.2 Sound

5.5.3 The “Wave”

5.5.4 Surface Disturbances on Water

Chapter 6

Dynamics and Action

6.1 Dynamics

Dynamics, as mentioned earlier, are the rules for finding the temporal evolution of a system. In Newtonian Physics, this set of rules was succinctly summed up in the rule: $\vec{f} = m\vec{a}$, see Section 6.1 on page 153. For a while, we will forget about light and fields and the dynamics of these complex systems and just describe simple point particles that move around freely in a simple space. We will find a new way to formulate the rules of dynamics that are more general but still produce the old $\vec{f} = m\vec{a}$ when it is appropriate. The advantage will be that the new rules will work in circumstances in which Newton's Laws were inappropriate or just did not make sense. With these new rules, we will also find a more powerful understanding of the concepts of symmetry and include systems such as fields all in a single dynamical principle. We will also be able to use this new procedure to form a more solid understanding of the ideas of energy and momentum. One complication will be that in order to formulate the rule, we will need ideas about kinetic and potential energy that we formulated earlier. Before we are done, these same ideas will take on a very different and more useful form. We will be able to understand why the massless photon has momentum but first we need to build the necessary background.

6.1.1 Background on Formulation of Action

It is usually not emphasized that the original formulation of Newton's Laws applied to only a very restricted set of circumstances. In Section 6.1 on page 153, Newton's Laws were described as dealing with the effects of one system on another with the assumption that all the parts of the bodies were

basically point objects that could move freely in space. This was fine when talking about the planets but, even for some of the simplest cases, these conditions do not hold.

Consider the problem of the motion of a blackboard eraser tossed into the air in the front of the lecture hall with a twisting spinning motion. Each part of the eraser is subjected to a huge array of forces. For convenience you can think of the parts of the eraser as the atoms but, even without an atomic hypothesis, all the following considerations still hold. Each part of the eraser is subject to the force of gravity and each part is subject to internal forces from the other parts of the eraser. First, there is an absurd number of parts and forces between the parts and between the parts and the world outside the eraser. We simplify this situation somewhat by assuming that the effect of gravity is the same throughout the eraser and thus reduce these many gravitational forces to a single force acting at one point at the mass weighted center of the body. This is a good approximation for the case of a small eraser in the near vicinity of the earth.

More subtly, we know that, as the eraser twists and spins, the different parts of the eraser will effect other parts. In fact, if the eraser was not a reasonably rigid body and held together by cohesive forces, in the spinning twisting motion, the parts would fly apart. Because the eraser is rigid, there are internal forces that act to hold the respective parts in a fixed relationship to each other. These forces are very complicated. They are in a very real sense unknowable; they are what they have to be to maintain the rigid configuration. These are called constraint forces. The eraser is not an exception. A car on the highway has a constraint force from the road called the normal force that is whatever it has to be to stop the car from falling into the road. Actually, with a little thought it becomes clear that almost all systems have constraints. The direct application of Newton's laws to systems that are constrained is wrong or impossible. There are an abundance of forces – too many to handle. Worse yet is the realization that many of them are, in fact, unknowable. The forces hold the eraser as it moves through space are whatever they have to be to maintain the positional relationship between the parts of the eraser. These are generally not known and thus cannot be inserted into a simple Newtonian framework.

In many special cases, fixes were developed that allowed the use of Newton's laws for motion in the presence of constraints and it was well known that this was a problem to both Newton and his immediate followers. The general problem of the motion of systems with algebraically described con-

straints was solved by Joseph-Louis Lagrange¹. The procedure that he developed is the modern method for articulating the dynamics of any system and is the one that we will use.

6.1.2 Introduction to Action

The modern approach to dynamics is based on the use of an extremum principle like Fermat's least time theory of light. There is a physical quantity that is called the action. In some sense, this is an unfortunate name for this because we have used the word in another context, see Section 5.2.1 on page 125, and it has a connotation in the conventional usage. The action is a quantity that we will define in detail later but for now understand that is a quantity evaluated over a trajectory in space and time. Up until now, we have dealt with paths in space. Now, we deal with trajectories but the principles are the same. For instance, the Fermat principle of least time required the time of passage of the light over the entire path between two points in space. Here the action is evaluated for a trajectory on space-time between two events, an initial position and time and a final position and time. Generally, the object moves over the trajectory that has the least action. Obviously, I will need to back up a little to make this clear and to establish the terminology.

We describe the motion of anything as a connected set of events in space-time, a path in space-time called the trajectory of the particle. The events labeled by a place and a time and are the fundamental entities and a trajectory is a catalogue of the places and as time evolves where the object went. Of the infinity of trajectories that can connect two events, the naturally occurring trajectory will turn out to be the one that has the least action.

Consider a piece of chalk tossed up from my hand and returning to my hand some short time later. I am dealing with only one spatial dimension, up. The zero of up is at my hand. The motion of the chalk is a continuous series of events that start with the toss at a time selected to be the zero of time and returns to my hand at a later time T . In between, the chalk has occupied a set of places at specific times between zero and T . If you know the places for all times in that interval you have a trajectory. In Figure 6.1 on page 156, we show the trajectory in a space-time diagram.

Any trajectory is only one of several that have the same total time interval T and start and stop at the same height. Why did nature chose the

¹Joseph-Louis Lagrange was born in France in 1736 and died in 1830. In 1788, he introduced a new approach to Newton's formulation of mechanics that is the basis for our modern interpretation in 1788.

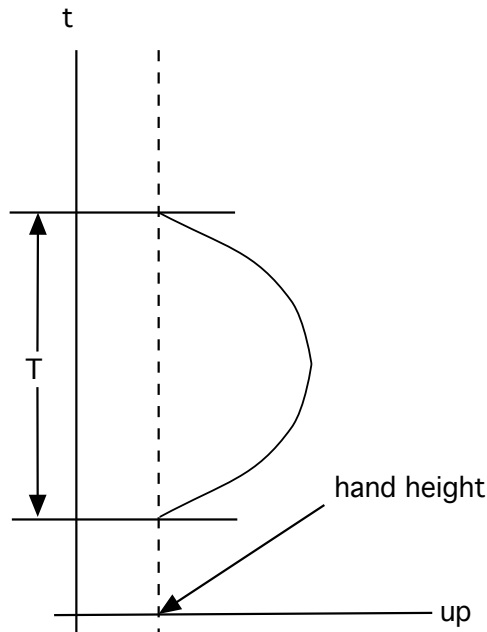


Figure 6.1: **Trajectory of a tossed piece of chalk** Chalk tossed from a height labeled zero rises with decreasing velocity until it reaches a peak and then returns to the hand after a time interval T .

one that she did? Several possible trajectories are shown in Figure 6.2 on page 157. It will turn out that our rule will be that nature chooses the trajectory from all the possible trajectories that has the least action. Since we have not yet defined the action, this is a little difficult to understand. Not only that but the approach is so different from the Newtonian that we do not have a developed intuition for this way of describing the chosen dynamic but some of the features of ‘Least Time’ articulation of paths for light carry over from our analysis of the properties of light, see Section 16 on Page 345. Like in the case of our ‘Least Time’ criteria for paths for light, this criteria is global and the minimizing property makes the ‘Least Action’ criteria produce a unique trajectory. This eliminates trajectories like the red one in Figure 6.2 on page 157 because the time flipped version has the same action. The issue of how high will require some analysis but we will find that action for this case is a combination of dependence of the maximum of the height of the trajectory above the hand height and the velocity it needs to get to the appropriate height.

If you were approaching this problem from the Newtonian point of view,

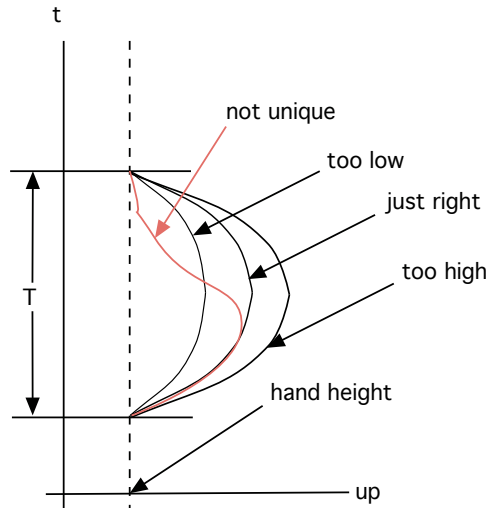


Figure 6.2: **Possible trajectories for a tossed piece of chalk** There are an infinity of trajectories that can connect the event at the start of the toss with the event at the return of the chalk to the hand at a later time T .

you would have used $\vec{f} = m\vec{a}$ and said that the chalk starts from a given place and given speed. Because there is a force, the attraction of the earth for the chalk, there is an acceleration. Since there is an acceleration, the velocity changes. The velocity changes until it is reversed at the maximum height and starts to fall. While all this is happening, the chalk is tracing out a smooth arc in space time. This description is very different than the one that we will be using for action. In the Newtonian formulation, the determination of the trajectory is done at each instant of time at the place at which the chalk is at that time. The action approach on the other hand deals with the action over the entire trajectory. This is a global approach to dynamics very similar to the ‘Least Time’ criteria of Fermat. It will be difficult to reconcile these disparate seeming approaches but you have to recover the Newtonian approach for the case in which the chalk can be treated as a point particle and free to move up and down without constraint. This will be covered in Section 6.1.5 on page 162

6.1.3 Definition of Action

Instead of $\vec{f} = m\vec{a}$ acting at each point on the body, there is now a new rule: minimize the action over the trajectory. In other words, nature chooses the least action trajectory from all the trajectories that share the same initial

and final event. This is a formulation of motion that is very much like that of Fermat's Least Time formulation for the paths of light in Section 3.2 on page 57. To determine the trajectory, you pick two events, an initial event, x_0 and t_0 , and a final event, x_f and t_f . There is a quantity called the action that is computed for every segment of the trajectory. Choose all possible trajectories and then the natural trajectory is the one that has the least action.

The action is defined from a function of the positions and velocities called the Lagrangian. In this approach to dynamics, instead of trying to figure out what forces are causing the motion, you try to find what the correct Lagrangian is. In a real sense, when a modern physicist develops a new fundamental theory of some phenomena, it is by finding the correct Lagrangian so that the trajectory that yields the least action using that Lagrangian is the one that occurs naturally.

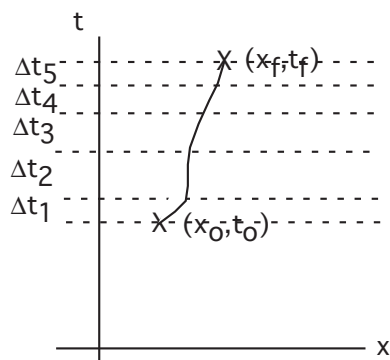


Figure 6.3: **Trajectory for the computation of the action** In order to compute the action for a given trajectory, the trajectory is divided into time slice pieces. For each time slice, the positions and the velocity can be determined. The action is then computed for that time slice and the contributions of each time slice are added to produce the overall action. The sizes of the time slices are determined by the rate of change along the trajectory.

There is a slight technical difference in this case and the case of Fermat's least time. In this case, we create our trajectory segments by creating time slices, see Figure 6.3 on page 158. For Fermat, the segments were sections along the length of the curve. As in the case of least time, the size of the time slices depends on the trajectory and the precision required. This gives a special role to the time variable. Also although we say all possible trajectories, for now, we will only deal with trajectories that advance in time

positively. We will be able to lift this condition later, Section ?? on page ??.

For a simple point object like the piece of chalk moving up and down, the Lagrangian depends on the position and velocity of the object. Given the Lagrangian, the action is

$$S(x_f, t_f, x_0, t_0; trajectory) = \sum_{trajectory, x_0, t_0}^{x_f, t_f} (L(x(t), v(t)) \Delta t \quad (6.1)$$

Action has the dimensions of an energy times a time. Although this makes the dimensions easy to remember, it is misleading. As we will learn later, the concept of energy is derivative from the action not the other way around, see Section 7.4 on page 194. It would be better to say that energy is dimensionally an action divided by a time. In terms of fundamental dimensional units, the units of action are $\frac{\text{mass} \times \text{length}^2}{\text{time}}$. From Equation 6.1 on page 159, the Lagrangian itself has the dimensions of an energy, $\frac{\text{mass} \times \text{length}^2}{\text{time}^2}$.

The rule that Lagrange found that would reproduce $\vec{f} = m\vec{a}$ for unconstrained systems and also work for more general situations is that the Lagrangian, $L(x(t), v(t))$, should be the difference in the kinetic energy and the potential energy.

$$L(x(t), v(t)) = \frac{mv^2}{2} - V(x) \quad (6.2)$$

where $V(x)$ is the potential energy. Later, Section 6.1.5 on page 162, we will show how this reproduces Newton's laws. It is important to again point out that although this approach requires that you know the kinetic energy and potential energy that these concepts are actually derived from the actions and not the other way. For now, it seems that you need to know the potential energy before you can write the Lagrangian. This is only for historical and pedagogical reasons. When a modern physicist is struggling with understanding some basic new phenomena, it is the other way around. We start with a Lagrangian and then see what the consequences are. It will also turn out that since the actions become the basis of all dynamics, it is the idea that theories that unify other earlier independent theories are considered unified when all the consequences of the theory arise from a single controlling Lagrangian. In modern language, Maxwell unified the electric and magnetic forces because the entire ensemble of equations is derivable from a single Lagrangian and the least action principle.

6.1.4 Trajectory of a Free Particle

To test our new dynamic, let's look at the simplest situation possible – a free particle. A free particle is one that has no forces acting on it. All places have the same energy value and thus $V(x) = C$, a constant, but if there is no preferred places and all places are the same then we may as well set the potential, the energy of a place, to zero. Using Lagrange's rule to get the solution for the free particle in old fashioned physics, we chose the Lagrangian that is just the kinetic energy or $L(v(t)) = \frac{mv^2}{2}$. To make it even simpler, let's require that the released particle is to return to the original position after a time T . The action is

$$S(0, 0, 0, T, traj.) = \sum_{traj., 0, 0}^{0, T} \frac{mv^2}{2} \Delta t \quad (6.3)$$

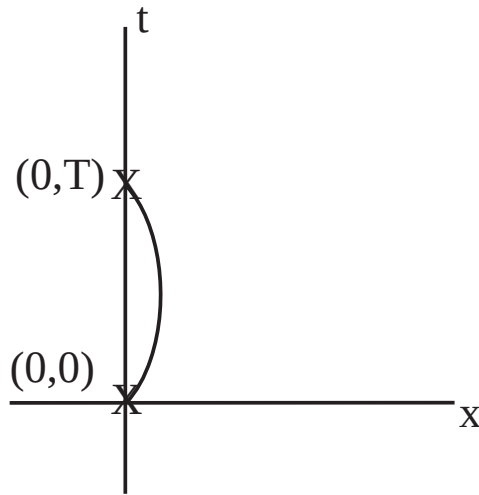


Figure 6.4: **Space-time diagrams for the action for a free particle**

A particle with no forces acting on it moves between two events, $(0, 0)$ and $(0, T)$. A possible trajectory is shown. Our experience with force free motion is that the straight line trajectory is the one that nature chooses; the particle remains at the point of release.

As was stated in the review section, Section 6.1 on page 153, a free particle at rest will remain at rest. Therefore, the natural trajectory for this case is the one that is at the starting place at all times. This is a straight

line along the t axis connecting $(0,0)$ and $(0,T)$. How do we obtain this same result using action?

Note that the action is a positive definite quantity for all velocities. Therefore any trajectory that has a non-zero velocity anywhere in the time interval will have a positive action. The trajectory that has $v(t) = 0$ for all t in the interval has an action of zero. This is clearly a minimum of the action since all other trajectories for these conditions will have a positive action. Thus this is the natural path. Actually any Lagrangian with v^2 in it will accomplish the same thing. The m is in it to give it the correct dimensions and the 2 for historical reasons. In fact, the m that is in the Lagrangian is the definition of mass. More on this later, see Section 15.3 on page 336.

Using this same result and remembering the material on Galilean invariance in Section 6.1 on page 153, we can solve a more general problem. Suppose we have a free particle that moves through the two events $(0,0)$ and (x_f, t_f) . Again, since the particle is free, the natural trajectory is the straight line connecting these events. To an observer moving by us at a speed of $v = \frac{x_f}{t_f}$, the object is at rest during the entire time interval. To that observer it is free and the initial and final events are $(0,0)$ and $(0, t_f)$ and the natural path is the straight line along the t axis as before. Thus to us the natural trajectory will be the straight line with slope $\frac{t_f}{x_f}$. Let's obtain this same result with a direct analysis.

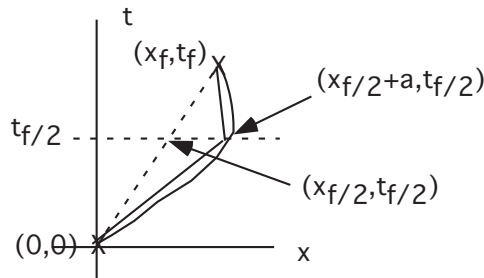


Figure 6.5: Space-time diagrams for the action for a free particle that changes position A particle with no forces acting on it moves between two events, $(0,0)$ and (x_f, t_f) . A possible trajectory is shown. The general trajectory connecting these events would be very difficult to describe. We will approximate the trajectory with a trajectory that is kinked at the mid-time and straight otherwise.

Consider a general trajectory connecting events $(0,0)$ and (x_f, t_f) , see Figure 6.5 on page 161. Our problem is to find all possible trajectories

between these events and then, for each trajectory, find the action. As we discussed about paths when dealing with the Fermat's least time approach to optics in Section 3.3.7 on page 76, path space is a rich mathematical structure. We want to do analysis. To do analysis we have to reduce the complexity of path space to something that can be described by functions. There are all these same difficulties when dealing with trajectories. To simplify our trajectory space, we reduce the trajectories that we consider to those that are "once kinked". Place the kink along the line $t = \frac{t_f}{2}$, see Figure 6.5 on page 161. In this reduced space, trajectories can be labeled by the distance, a , of the kink from the event $(\frac{x_f}{2}, \frac{t_f}{2})$ along that line. Using this trajectory in the appropriately modified Equation 6.3 on page 160 to take account of the new ending event, and the fact that the inverse slope of the line is the velocity in that segment, it is easy to compute the action for the trajectory labeled a . It is

$$S(0, 0, x_f, t_f, traj = a) = \frac{m}{2} \left(\frac{(\frac{x_f}{2} + a)^2}{\frac{t_f}{2}} + \frac{(\frac{x_f}{2} - a)^2}{\frac{t_f}{2}} \right). \quad (6.4)$$

This is an even function of a and thus has a minimum at $a = 0$. This confirms our result that the natural trajectory, the constant velocity trajectory, is the least action trajectory.

6.1.5 Proof that the Least Action Reproduces Newtonian Physics

See Feynman's famous lecture on action. It was handed out in class and is posted on Canvas.

For those who are intimidated by his vocabulary and technique, we will follow up with an approach that is hopefully more accessible. The major issue is the question of 'over all trajectories'. In Section 6.1.3 on page 157, we indicated that you could recover the usual statement of mechanics known as Newton's laws by defining the Lagrangian as the kinetic energy minus the potential energy. In this section, we will prove this. We will use the most simple and direct method for controlling the problem of the size of trajectory space by using a simple time sliced reduction of the trajectories to an $(\mathbb{R}^1 \times \mathbb{R}^1)^n$ by labeling the trajectory with the set $\{x_i, t_i\}$, where x_i is the position on the trajectory at the time slice t_i and n is the number of time slices and is a large number.

We are interested in all possible trajectories that connect an initial event (x_0, t_0) and a final event (x_f, t_f) , see Figure 6.3 on page 158. Consider the

case of n time slices. Any trajectory that connects the initial and final events is labeled by the set of numbers $\{x_i, t_i\}$. We construct the action in the prescribed fashion, $L = m\frac{v^2}{2} - V(x)$. As in the earlier examples, the velocity in any segment, which is approximated by the straight line segment between two time slices, is the change in position in the segment divided by the time interval, $\frac{x_i - x_{i-1}}{t_i - t_{i-1}}$. Similarly, the position in the segment is the average position over that segment², $\frac{x_i + x_{i-1}}{2}$. Thus following the first of Equations 6.1 on page 159, the action for any trajectory is

$$S(x_f, t_f, x_0, t_0; \{x_i, t_i\}) = \sum_{\text{seg. } \{x_i, t_i\}, x_0, t_0}^{x_f, t_f} \left(\frac{m}{2} \left(\frac{x_i - x_{i-1}}{t_i - t_{i-1}} \right)^2 - V \left(\frac{x_i + x_{i-1}}{2} \right) \right) (t_i - t_{i-1}), \quad (6.5)$$

where, as indicated, we sum over all the segments between (x_0, t_0) and (x_f, t_f) .

Assuming that, from the very large set of possible $\{x_i, t_i\}$, there exists one that minimizes the action and is thus the naturally occurring trajectory. Designating this trajectory $\{\bar{x}_i, \bar{t}_i\}$, we can calculate the action with Equation 6.5 on page 163. We can also examine the nearby trajectory which is the same as $\{\bar{x}_i, \bar{t}_i\}$ everywhere except at some particular time slice for the natural trajectory, $\bar{t}_j$, and, for this new trajectory, the position coordinate differs from the naturally occurring one by a small amount. In other words, the new nearby trajectory is the one for which $\{x_i, t_i\} = \{\bar{x}_i, \bar{t}_i\}$ except when $i = j$ and, for that case, the position time pair is $(x_j, t_j) = (\bar{x}_j(1 + \epsilon), \bar{t}_j)$ where ϵ is a small number, $\epsilon \lll 1$. ϵ is such a small number that we can neglect all powers of ϵ greater than one. As we will see the ϵ^0 terms will drop out of the change in the action and thus we need only deal with the ϵ^1 term.

From Figure 6.6 on page 164, it is clear that the difference in the action of the naturally occurring trajectory and the nearby one requires only calculating the segment of the trajectories before and after the incremented one. This not only simplifies the calculation but also shows how this use of near in trajectory space can be used to convert a global statement of dynamics to one that is local in time like Newton's laws, see Section 6.1.1 on page 153.

²Depending in the nature of the potential function a differently weighted average may be more appropriate but if the time slices are small enough this simple average is adequate.

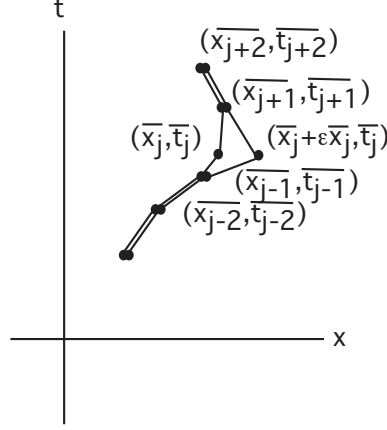


Figure 6.6: Natural Trajectory and Nearby Trajectory In the time sliced formulation of action, trajectories are labeled by the set $\{x_i, t_i\}$ where i ranges from 0 to n where n is the number of time slices. The naturally occurring trajectory is the one that minimizes the action and is designated by the $\{\bar{x}_i, \bar{t}_i\}$. The condition of minimum is established by requiring that the change in the action between the naturally occurring trajectory and any nearby trajectory vanish to first order in the measure of nearness. A measure of nearness for trajectories is to increment the space component of a trajectory at some time slice by a small amount. Above the natural trajectory which has many time slices is incremented at the j th slice to produce the nearby trajectory. In the calculation of the difference of the action over the naturally occurring and nearby trajectories only the parts of the trajectories around the j th slice are relevant. This converts a global condition for the naturally occurring one to one that is local in time like the Newtonian statement of mechanics.

Writing the difference in the actions for the naturally occurring trajectory and the nearby one,

$$\begin{aligned}
\delta S &= S_{\text{near by}} - S_{\text{naturally occurring}} \\
&= \left\{ \frac{m}{2} \left(\frac{\bar{x}_j(1+\epsilon) - \bar{x}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} \right)^2 - V \left(\frac{\bar{x}_j(1+\epsilon) + \bar{x}_{j-1}}{2} \right) \right\} (\bar{t}_j - \bar{t}_{j-1}) \\
&\quad + \left\{ \frac{m}{2} \left(\frac{\bar{x}_{j+1} - \bar{x}_j(1+\epsilon)}{\bar{t}_{j+1} - \bar{t}_j} \right)^2 - V \left(\frac{\bar{x}_{j+1} + \bar{x}_j(1+\epsilon)}{2} \right) \right\} (\bar{t}_{j+1} - \bar{t}_j) \\
&\quad - \left\{ \frac{m}{2} \left(\frac{\bar{x}_j - \bar{x}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} \right)^2 - V \left(\frac{\bar{x}_j + \bar{x}_{j-1}}{2} \right) \right\} (\bar{t}_j - \bar{t}_{j-1}) \\
&\quad - \left\{ \frac{m}{2} \left(\frac{\bar{x}_{j+1} - \bar{x}_j}{\bar{t}_{j+1} - \bar{t}_j} \right)^2 - V \left(\frac{\bar{x}_{j+1} + \bar{x}_j}{2} \right) \right\} (\bar{t}_{j+1} - \bar{t}_j). \tag{6.6}
\end{aligned}$$

In order to proceed, we will use Newton's approximation, $f(x+\epsilon) \approx f(x) + \epsilon \frac{df}{dx}(x)$ which is valid for any small ϵ and any smooth $f(x)$ for the potential function in the near by trajectory terms. Expanding and dropping terms of order ϵ^2 ,

$$\begin{aligned}
\delta S &= \epsilon x_j \left\{ m \left\{ \frac{\bar{x}_j - \bar{x}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} - \frac{\bar{x}_{j+1} - \bar{x}_j}{\bar{t}_{j+1} - \bar{t}_j} \right\} \right. \\
&\quad \left. - \frac{1}{2} \left\{ \frac{dV}{dx} \left(\frac{\bar{x}_j + \bar{x}_{j-1}}{2} \right) (\bar{t}_j - \bar{t}_{j-1}) \right. \right. \\
&\quad \left. \left. + \frac{dV}{dx} \left(\frac{\bar{x}_{j+1} + \bar{x}_j}{2} \right) (\bar{t}_{j+1} - \bar{t}_j) \right\} \right\} \tag{6.7}
\end{aligned}$$

This result can best be interpreted by first multiplying through by $\frac{\bar{t}_{j+1} - \bar{t}_{j-1}}{\bar{t}_{j+1} - \bar{t}_j}$, and reorganizing

$$\begin{aligned}
\delta S &= \epsilon x_j (\bar{t}_{j+1} - \bar{t}_{j-1}) \left[-m \frac{1}{(\bar{t}_{j+1} - \bar{t}_{j-1})} \left\{ \frac{\bar{x}_{j+1} - \bar{x}_j}{\bar{t}_{j+1} - \bar{t}_j} - \frac{\bar{x}_j - \bar{x}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} \right\} \right. \\
&\quad \left. - \frac{1}{2} \left\{ \frac{dV}{dx} \left(\frac{\bar{x}_j + \bar{x}_{j-1}}{2} \right) \left(\frac{\bar{t}_j - \bar{t}_{j-1}}{\bar{t}_{j+1} - \bar{t}_{j-1}} \right) \right. \right. \\
&\quad \left. \left. + \frac{dV}{dx} \left(\frac{\bar{x}_{j+1} + \bar{x}_j}{2} \right) \left(\frac{\bar{t}_{j+1} - \bar{t}_j}{\bar{t}_{j+1} - \bar{t}_{j-1}} \right) \right\} \right]. \tag{6.8}
\end{aligned}$$

For $\delta S = 0$, since all the factors in front of the square bracket are non-zero, the terms inside the square bracket must vanish at each t_i . This is a condition on the naturally occurring trajectory. The terms in the curly

brackets on the first line of Equation 6.8 are the change in velocity in the two relevant segments of the trajectory which when divided by the time interval over the two segments is the acceleration at t_j . The mass times this acceleration must equal the remaining terms in the curly bracket. These are position and time segment weighted averages of the quantity $\frac{dV}{dx}$ evaluated in the two relevant segments. Since in Newtonian mechanics $F = -\frac{dV}{dx}$, we have that for each i along the naturally occurring trajectory

$$m\bar{a}_i = -\frac{dV}{dx}(\bar{x}_i) = F.$$

6.2 Examples of Actions

6.2.1 Gravitation near a flat earth

As a simple example that we are all familiar with, consider the case of motion above the surface of the earth. Here the energy of position, the potential energy, is due to the gravitational interaction of a massive body with the earth. For this case, the potential energy at a height h above the earth is $V(\vec{r}) = -\frac{Gm_em}{R_e+h}$, where m_e is the mass of the earth, m the mass of the body, and R_e is the radius of the earth. For motion near the surface, a few meters up or down, from “Things Everyone Should Know,” Section 1.4.2 on page 16, we can use $(1+x)^n \approx 1+n x$ for $x \ll 1$ to reduce this to

$$V(h) = -m \left(\frac{Gm_e}{R_e} \left(1 - \frac{h}{R_e} \right) \right) = V(R_e) + mgh,$$

where we recognize $g = \frac{Gm_e}{R_e^2}$. Since this potential is to be used in an action, as we will see later in Section 7.4 on page 194, changing the action by a constant does not change the physical results in a significant way, we can drop the $V(R_e)$ term. This reduces the potential energy for objects moving in the near vicinity of the earth to

$$V(h) = mgh. \tag{6.9}$$

Another way to look at this result is to say that for motion restricted to be near the surface of the earth, the earth appears as an infinite plane. In this case, the force of gravity above the plane can not depend on anything, in particular, the height above the plane or the position sideways over the plane. Thus the force also can only be toward or away from the plane. Then realizing from the analysis above in Section 6.1.5 on page 162 that the

change in potential as you change position is the force, the only form for the potential in this case is $mgh + \text{constant}$.

For now, let us continue to consider only up and down motion, not any sideways motion. The potential energy is mgh where h is the height. Thus the action for any trajectory between an initial height, h_0 at time t_0 and final height, h_f at time t_f is

$$S(h_0, t_0, h_f, t_f; \text{traj.}) = \sum_{\text{traj.}, h_0, t_0}^{h_f, t_f} \left(\frac{mv^2}{2} - mgh \right) \Delta t \quad (6.10)$$

where the path is given by $h(t)$. Note that if you know $h(t)$, you also know $v(t)$. You can see from the form of the action that you will lower the action by having $h(t)$ to be at large h for as much time as possible. The problem is that since the initial and final position and time are given, it takes high velocity to get to large h . The high velocity increases the action $\Rightarrow$ There is a compromise between the negative contributions by going too high and the velocity needed to get there. This is the trajectory that the particle follows naturally, see Figure 6.2 on page 157.

Let's get more specific. This is again the problem of a piece of chalk tossed up in the air. First the simplest case, the chalk is released and returns to the same height after a time T .

We need to study the action for all trajectories connecting these events. Again, because of the complexity of the idea of all trajectories, we will need to reduce the number of trajectories. A first step is to use our experience to limit ourselves to simple trajectories that rise smoothly to a peak at some height a at which time the velocity is zero and then returns over a trajectory that is a reflection of the one on the rise. Our natural trajectory must be in that family. This is still a very rich family and too rich to do analysis. This is the same problem that we had with the Fermat's Least Time, Section 3.3.7 on page 76, and the free particle, Section 6.1.4 on page 160. As in the latter case, the once kinked path can be used to approximate the family of smooth trajectories that have these properties, see Figure 6.7 on page 168. Here again the variable a is the height of the approximate trajectory but more importantly now it is a label that can be used to specify the particular trajectory from the family with which we are dealing. We are labeling all possible trajectories with a single parameter—putting it on an $\mathbf{R}^1$.

Since this approximate trajectory is broken line segments, it is relatively

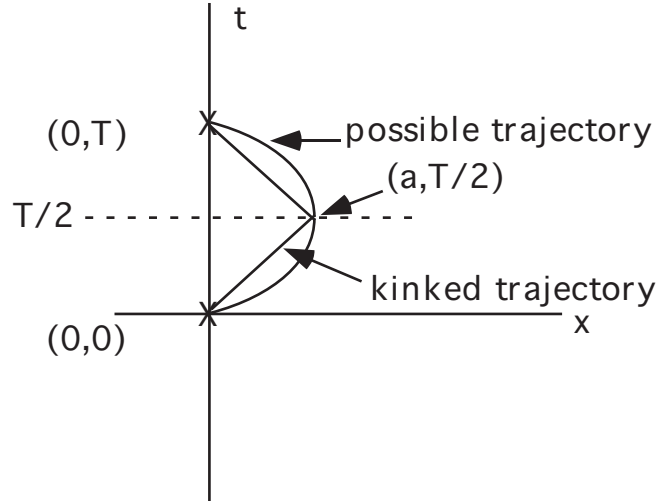


Figure 6.7: **Possible trajectory for the action for a particle in a uniform gravitational field** A piece of chalk is tossed upward and caught later at the the same height. A possible trajectory is shown. The natural trajectory is one from the family of smooth trajectories that rise to a peak at a height a smoothly and then return to a lower height on a reflected trajectory. This is still a large family of trajectories. We can approximate the members of this family with a once kinked trajectory with the same height at the time $\frac{T}{2}$.

easy to compute the action.

$$S(0, 0, 0, T; \text{traj.}) = \sum_{(0,0) \text{ traj.}}^{(0,T)} \left(\frac{mv^2}{2} - mgh \right) \Delta t. \quad (6.11)$$

For a straight line path, v is a constant and is the inverse slope of the line, and is $\frac{a}{T}$ in magnitude for both segments. The height is a more subtle question since it varies with time from 0 to a . Being reasonable, we can use the average height, $\frac{a}{2}$. For the sophisticates among you, there is the problem that the concept of average is a not trivial, see Section 6.2.6 on page 178. Thus the action for the first segment is

$$S_1(T, a) = \frac{ma^2}{2} \frac{T}{\left(\frac{T}{2}\right)^2} \frac{T}{2} - \frac{mga}{2} \frac{T}{2}. \quad (6.12)$$

Note that once I have made a mapping of the paths onto the line that S

becomes a regular function of the path label, a , instead of a functional. Although the velocity is negative, since only v^2 enters the lagrangian, the action on the second segment is the same and the total action is

$$S(T, a) = 2S_1(T, a) = ma \left(2 \frac{a}{T} - \frac{g}{2} T \right) \quad (6.13)$$

This has zero's at $a = 0$ and $a = \frac{gT^2}{4}$. The dependence of the action on the path label a is shown in Figure 6.8 on page 169. I have used dimensions in which $g = T = 1$.

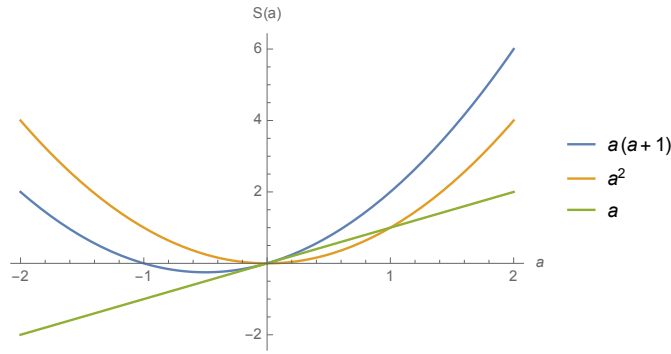


Figure 6.8: **Action as a function of a** The action as a function of the trajectory label a . This curve is a combination of a parabola, $\frac{mT^3}{6}a^2$, concave up with its vertex at the origin and a straight line, $\frac{mT^3}{6}a$, with positive slope through the origin. The units are $m = \frac{1}{6}$, $T = 1$. It is important to realize that for any polynomial adding a to it of lesser degree only shifts the plot and that in this case the two polynomials share a common zero so this point does not shift.

We can see that there is a minimum half way between the two zero's at $a = 0$ and $a = \frac{gT^2}{4}$. This implies that the trajectory from this set in this approximation that is the least action trajectory is the one with

$$a_{least\ action} = \frac{gT^2}{8}. \quad (6.14)$$

Since this is not only the path selecting parameter but is also the height of the trajectory. We get that the height is $\frac{gT^2}{8}$.

6.2.2 Same Example done another way*

I am going to do some mathematics here that I do not expect that you will be able to reproduce. I do this to show you that it can be done and that the

ideas of mathematics are useful. You are not expected to do integrals and take derivatives although you should be able to follow a development using them.

Once again, we want to examine the case of an object of mass m moving in the vicinity of the earth. We can also guess that the correct answer for the height as a function of time is a parabola, all parabolas that fit the time interval are of the form $h(t) = at(t - T) \Rightarrow v(t) = 2at - aT$, where a is label of the path in path space³. In this case, a has the dimension of an acceleration, $L \stackrel{\text{dim}}{=} a \times T^2$ or $a \stackrel{\text{dim}}{=} \frac{L}{T^2}$.

The Lagrangian is $L = \frac{1}{2}mv^2 - mgh$ and the action is

$$\begin{aligned}
 S(a) &= \int_{(x_0, t_0), \text{Path}}^{(x_f, t_f)} \left(\frac{1}{2}mv^2 - mgh \right) dt \\
 &= m \int_0^T \left(\frac{1}{2}(2at - aT)^2 - gat(t - T) \right) dt \\
 &= m \left(\frac{a^2 T^3}{6} + \frac{1}{6}agT^3 \right) \\
 &= \frac{mT^3}{6} a(a + g). \tag{6.15}
 \end{aligned}$$

To find the minimum, we can again realize that there are two zeros of $S(a)$. One at $a = 0$ and one at $a = -g$. The minimum is half way between them at $a_{\text{least action}} = -\frac{g}{2}$

Otherwise and since S is now a function, we can take the derivative of $S(a)$ with respect to a and set it equal to zero. Thus

$$\begin{aligned}
 \frac{dS}{da} &= \frac{d}{da} \left(\frac{mT^3}{6} a(a + g) \right) \\
 &= \frac{mT^3}{6} \{a + (a + g)\} \\
 &= \frac{mT^3}{6} (2a + g)
 \end{aligned}$$

or $a_{\text{least action}} = -\frac{g}{2}$ is the natural trajectory. . Another important feature of this situation is that $\frac{d^2 S}{da^2}(a) = \frac{mT^3}{3} \geq 0$. This guarantees that the minimum is a true local minimum. This is also shown by the fact that $S(a)$ is a parabola that is concave upward. Figure 6.9 on page 171, shows how the action varies with a . Again I have used units with $g = T = 1$.

³Note that since the path label a is an element of an $\mathbf{R}^1$ the action is now a function of the path label and I am now able to indicate $S(a)$ instead of the functional indication $S[a]$.

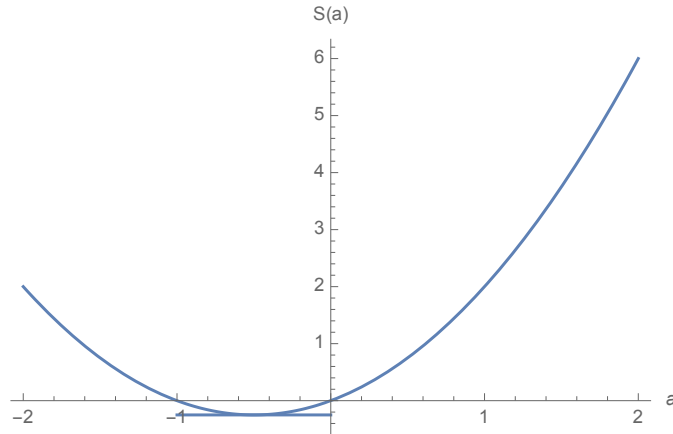


Figure 6.9: **Action as a function of a as an acceleration** Action as a function of a when the parameter a has the dimensions of an acceleration and the units of g and m are both set equal to 1. This example shows that the trajectory label does not have to be a height. You can use any convenient $\mathbf{R}^1$.

6.2.3 Simple Scattering of Two Particles

Two particles, one of mass m_1 and the other of mass m_2 collide. After the collision, the particles move away from each other, both still with masses m_1 and m_2 . This is a very special problem whose importance cannot be over emphasized. In a very real sense, when we probe the nature of the elementary constituents of matter, scattering experiments are the primary source of our knowledge. In addition, the process is so basic that it will allow us to begin to better understand many fundamental issues. For this subsection, we will make the simplest description⁴

How do we handle this process? First, we have to decide what is meant by two independent particles. Before the particles make contact, they move as if the other particle was not present, i. e. they are independent. It is reasonable therefore to assume that while they are apart or not interacting, the two particles actions add and are the usual free particle action. In other words, there is a free particle action that tells you all the properties of what is meant by a particle and its nature. For our construction of the action

⁴For now, we discuss the simplest construction of a scattering process. The subtlety will be elaborated when we have developed more tools for their articulation later in this section. Also see Section 15.1 on page 331, Section 7.4.4 on 199. For now we are pretending that we live in a pipe with an unstructured objects.

of the free particle in Section 6.1.4 on page 160, we used the Lagrangian $L(x, v) = \frac{mv^2}{2}$. The Lagrangian says the object identified as a free particle does not treat different places differently and thus there is no x dependence in the Lagrangian. If we want to recover Newton's Law, see Section 6.1.5 on page 162, we use the usual classical kinetic energy. We will find that in other circumstances, that a different free particle Lagrangian is appropriate. If we wanted to describe something more complicated than a point particle, say a small rod, we would need elements that deal with what a rod is such as moment of inertia and directional variables.

By using as the action the sum of the single particle actions, the properties of the total system will be the sum of the properties of the parts. If we did this though, and this was the end of it, nothing interesting would ever happen; the particles would merely pass through each other unchanged in their motion. We want them to scatter. Thus in addition, we need to add a part that carries the interaction. The interaction will have a Lagrangian that is made up of relationship variables such as their separation in addition to the particle labels. In other words, the action is made up of the following parts:

$$\begin{aligned} \text{Total Action} &= \text{Free Action}(\text{variables particle 1}) \\ &+ \text{Free Action}(\text{variables particle 2}) \\ &+ \text{Interaction Action}(\text{variables particle 1,} \\ &\quad \text{variables particle 2, relationship variables}). \end{aligned} \quad (6.16)$$

Of course, it is actually redundant to list the relationship variables in the interaction action since they will be composed of the relationship variables of particle 1 and 2 anyway. The importance of displaying the relationship variables separately is to be able to say that, for a scattering situation, the interaction action is zero when the relationship variables such as the separation are large. In a collision, we assume that most of the time the particles travel toward or away from each other and that the interaction terms contribute only for a short time when the particles are in contact and thus this interaction term is small and does not add significantly to the total action of the process. Another point to note is that, since the interaction terms are dominated by the relationship variables, the contribution from the interaction action should be independent of where and when the collision takes place. Thus, we can write the action for this simple one dimensional

scattering process as

$$S = \sum_{(x_{10}, t_{10}), Path}^{(x_{1f}, t_{1f})} m_1 \frac{v_1^2}{2} \Delta t + \sum_{(x_{20}, t_{20}), Path}^{(x_{2f}, t_{2f})} m_2 \frac{v_2^2}{2} \Delta t + A, \quad (6.17)$$

where A represents the contribution of the action when the particles are in contact. The scattering process is shown in Figure 6.10 on page 173.

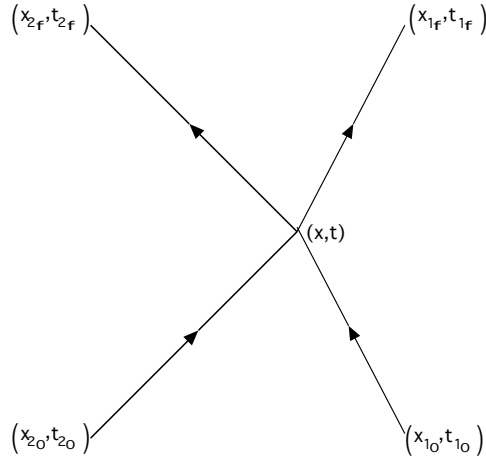


Figure 6.10: **Space-time diagram for a scattering event** Two particles of mass m_1 and m_2 free to move in one spatial dimension are directed at each other and collide at the event (x, t) and then move apart. A space-time diagram for a scattering event with particle one starting at event (x_{10}, t_{10}) and returning to (x_{1f}, t_{1f}) and particle two starting at event (x_{20}, t_{20}) and returning to (x_{2f}, t_{2f}) is shown. Although all trajectories connecting the initial and final events and the collision event should be examined, we know that free particles have a natural trajectory that is a straight line, see Section 6.1.4 on page 160.

We want to do all paths but we know that the straight path is the least action for a free particle and so all we need to do is use straight paths between the initial and collision and collision and final events. We can immediately write down the action as a function of the position and time of the collision. The coordinates of that event are the only free parameters in the problem.

Note that we are being consistent in our use of action. When you talk about collisions in the general physics class you set the initial velocities. Here we use the initial and final events. Evaluating the free particle actions,

for this system of trajectories, the action is

$$S = \frac{m_1}{2} \frac{(x - x_{10})^2}{(t - t_{10})} + \frac{m_2}{2} \frac{(x - x_{20})^2}{(t - t_{20})} + \frac{m_1}{2} \frac{(x_{1f} - x)^2}{(t_{1f} - t)} + \frac{m_2}{2} \frac{(x_{2f} - x)^2}{(t_{2f} - t)} + A. \quad (6.18)$$

We want to find the trajectory that has the least action and since we have now reduced the world of trajectories to the label of the collision point, x and t . The only term that we do not know is the contact term A . It could depend on any number of things; the hardness, compressibility and location in space and time of the collision. For our case, we assume that it is independent of every measure that we can impose; the particles are hard elastic spheres.

Thus we need to minimize this in what are now the labels, x and t . We can eliminate one of the steps by taking advantage of the requirements for uniqueness. All cases except those on the line $t = \frac{t_{1f} - t_{10}}{2} = \frac{t_{2f} - t_{20}}{2}$ have partners with the same action. You could plot this and find the minimum by hand, see Figure 6.11 on page 175, but, if you allow me to use calculus, I can find a simple analytic expression for the $x = x_{min}$ and $t = t_{min}$ that yields the least action. This means taking the derivatives with respect to x and t and finding the value of x and t that satisfy $\frac{\partial S}{\partial x} = 0$ and $\frac{\partial S}{\partial t} = 0$. This x and t label the naturally occurring trajectory.

Take my word for it. The condition for a minimum in x is

$$m_1 \frac{(x_{min} - x_{10})}{(t_{min} - t_{10})} + m_2 \frac{(x_{min} - x_{20})}{(t_{min} - t_{20})} - m_1 \frac{(x_{1f} - x_{min})}{(t_{1f} - t_{min})} - m_2 \frac{(x_{2f} - x_{min})}{(t_{2f} - t_{min})} = 0 \quad (6.19)$$

or

$$m_1 \frac{(x_{min} - x_{10})}{(t_{min} - t_{10})} + m_2 \frac{(x_{min} - x_{20})}{(t_{min} - t_{20})} = m_1 \frac{(x_{1f} - x_{min})}{(t_{1f} - t_{min})} + m_2 \frac{(x_{2f} - x_{min})}{(t_{2f} - t_{min})} \quad (6.20)$$

Realizing that momentum is mv in classical physics and that v is the difference in positions divided by the the differences in times, this is the statement that the momentum into the collision is equal to the momentum out of the collision.

The condition that there is a minimum in t gives

$$\frac{m_1}{2} \frac{(x_{min} - x_{10})^2}{(t_{min} - t_{10})^2} + \frac{m_2}{2} \frac{(x_{min} - x_{20})^2}{(t_{min} - t_{20})^2} - \frac{m_1}{2} \frac{(x_{1f} - x_{min})^2}{(t_{1f} - t_{min})^2} - \frac{m_2}{2} \frac{(x_{2f} - x_{min})^2}{(t_{2f} - t_{min})^2} = 0 \quad (6.21)$$

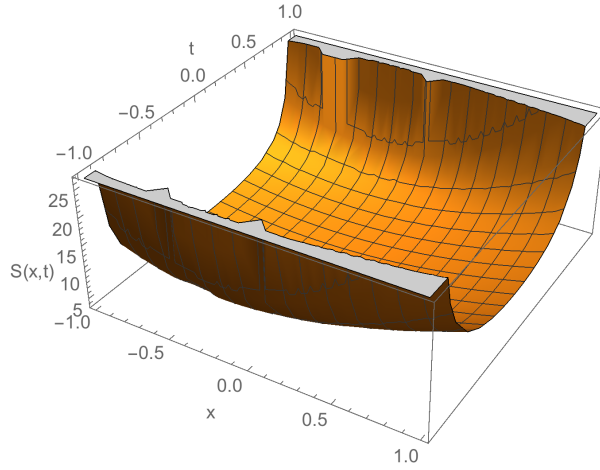


Figure 6.11: **Action for a Scattering Event** Action as a function of x and t for a scattering event shown in Figure 6.10 on page 173. There is a clear minimum and it occurs at the points at which Equation 6.20 on page 174 and Equation 6.22 on page 175 are satisfied.

or

$$\frac{m_1 (x_{min} - x_{1_0})^2}{2 (t_{min} - t_{1_0})^2} + \frac{m_2 (x_{min} - x_{2_0})^2}{2 (t_{min} - t_{2_0})^2} = \frac{m_1 (x_{1_f} - x_{min})^2}{2 (t_{1_f} - t_{min})^2} + \frac{m_2 (x_{2_f} - x_{min})^2}{2 (t_{2_f} - t_{min})^2} \quad (6.22)$$

Which is the same as the statement that the energy into the collision event is equal to the energy out of it.

Figure 6.11 on page 175 shows the action as a function of the position and time of the collision event. This is for the case that $\frac{m_2}{m_1}$ is 2 and the original and final events for particle 1 are $x_{1_0} = 1$, $t_{1_0} = -1$ and $x_{1_f} = 1$, $t_{1_f} = 1$ and for particle 2 are $x_{2_0} = -1$, $t_{2_0} = -1$ and $x_{2_f} = -1$, $t_{2_f} = 1$.

This exercise also gives us an interesting insight on what mass is. In an early assignment in this course, you were asked to devise a method for measuring mass that does not rely on gravity. Some of you came up with the idea of using collisions to define a mass scale. You can see that this analysis is directly relevant to that kind of definition. In the construction of the action, for the case of the single particle, mass is an overall factor; it is the thing you put in front of the v^2 , in the action. If the world consisted of only one particle, mass would be irrelevant since all it does is multiply the action⁵. The process of finding the natural trajectory is unchanged by the

⁵At least the concept of inertial mass. The gravitational mass which generates the

an overall scale factor on the action. Mass becomes interesting only when you have more than one particle. If there is more than one particle, you can not remove all the masses with a single scaling factor. The ratios of the mass remain. Consider a scattering event between two particles with the initial and final positions of the two particles the same before and after the collision. If the particles had equal masses, the position of the collision event is at the center. The trajectories of both particles are equally kinked. On the other hand, the higher the mass ratio of say the second particle, the less the trajectory associated with that particle will kink when it collides with another particle. In the limit of a very large mass second particle, there is no bending of the second trajectory and it looks like the first particle has hit a brick wall. This is the essence of inertia.

6.2.4 A Small Sphere Hanging from a Fixed Point

Let's use the results of the previous analysis of the up-down motion over a flat earth to look at a more complex system, the simple pendulum. An example is a rather massive small sphere called a bob connected by a flexible light cord with one end attached to a fixed point in space above a flat earth. To even start on this problem we exercise our skills in 'Spherical Cowing', see discussion on page 30. The cord is flexible and massless and the bob can be treated as a point mass. We will look at its motion for small angular displacements from equilibrium, see Figure 6.12 on page 177.

To construct an action for a classical situation our first problem is to identify the kinetic energy. We have only one mass and as usual the kinetic energy is $KE = m \frac{(\vec{v})^2}{2} = m \frac{v_x^2}{2} + m \frac{v_y^2}{2} + m \frac{v_z^2}{2}$ where the velocity is measured in the three space over the flat earth with z going up down and x - y forming the plane of the surface of the flat earth. One complication is that the velocity is a vector quantity in a three dimensional space and our mass is constrained to move in a curved two dimensional space; see Figure 6.12 on page 177. If the angle θ is small, the v_z is small compared to v_x and v_y and can be neglected⁶ of the mass's movement in three space with the motion on the projected plane of the flat earth. The potential energy is the same as in our previous case, $PE = mgh$. Putting a coordinate system on the flat

gravitational force is a totally different concept. The equivalence of these independent mass definitions is a foundational issue in the modern theory of gravity, see Chapter 16 on page 345.

⁶This is an aspect of "spherical cowing". The velocity $\vec{v}$ is constrained to be tangent to the places on the cupped surface. Therefore, at any place on the surface, the z component of the tangent is small compared to the components tangent.

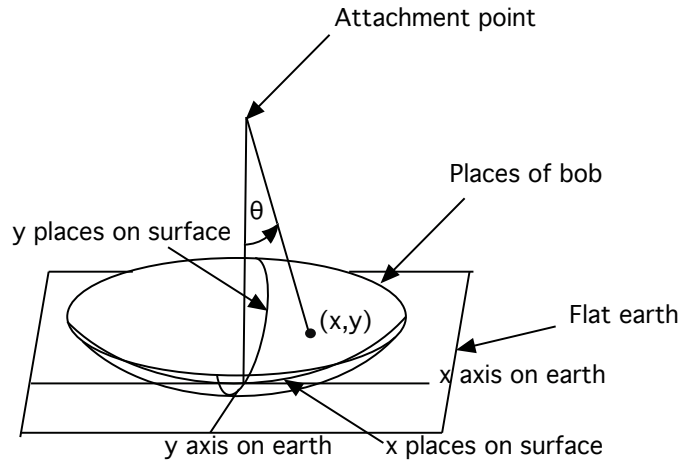


Figure 6.12: **The Simple Pendulum** A point mass, a bob, of mass m is suspended from a point over a flat earth at height h with a cord of length $L < h$. The mass is given a small angular displacement from the vertical and allowed to move freely between the place (x_0, y_0) and (x_f, y_f) in a time T .

earth with the z along the vertical and an xy plane on the earth surface. The orientation of the x axis is arbitrary since the system is rotationally symmetric for rotations about a vertical axis through the fastening point.

6.2.5 Action for the Stretched String

Although it is an unrealistic representation of a simple field, the idealized stretched elastic string is a wonderful example of a simple mechanical system which manifests field properties, see Section 5.3 on page 128. We developed a dynamic for this field which produced a broad range of motions including travelers. Since one of our principle purposes is to use the general concepts of action based physics to explain all physical phenomena our earlier discussion of action has to be extended to field phenomena. In this section we will articulate the procedures which are required to bring extended systems into an action formulation of mechanics via the example of the stretched string. We will follow the general protocols set forth by Lagrange that we used successfully for simple point masses but applied to the extended elastic system. This will provide the framework to developing the action formulation appropriate to the generic field systems.

Our basic model of the string is a connected set of point mass placed at

equal intervals each one connected to a nearest neighbor by a small elastic member, see Figure 5.1 on page 129. Following Lagrange's original prescription, the Lagrangian contains two parts; a kinetic energy component and a potential energy which, in this case, is the tension in the string times the stretch of the string associated with the configuration of the string.

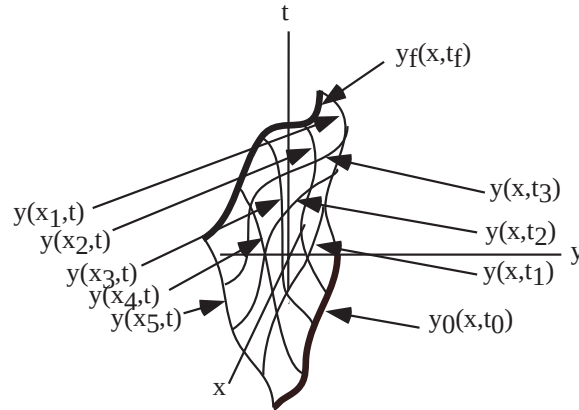


Figure 6.13: **String Fields for Action** In contrast to the case of the point particle, Figure 6.3 on page 158, the trajectory for the field between the initial field configuration and final field configuration is a two surface in a three space, two spatial directions and one time. The two spatial directions are the label for the places on the string, x , and one is the direction in which the string moves, y , assumed transverse. The naturally occurring field configuration is the surface configuration that minimizes the action. The dynamic for the field is found by looking at small variations about the naturally occurring configuration and requiring that the changes in the action vanish.

6.2.6 Digression on averages and slicing*

It should come as no surprise that most people do not think hard about what they mean by averages. This is often exemplified by the puzzle:

Consider two towns that are one hundred miles apart, for instance Austin and College Station. You want to travel between them with an average speed of fifty miles per hour. You leave Austin but get caught behind a very long funeral procession that you cannot pass that is also going to Hicksville, half way between

Austin and College Station. If the funeral procession held your speed to an average of twenty five miles per hour between Austin and Hicksville, how fast do you have to drive in the remainder of the trip to obtain your desired average of fifty miles per hour?

The accepted answer is that you have to go infinitely fast. This is because in the portion of the trip between Austin and Hicksville has taken two hours and, in order to average fifty miles per hour on a one hundred mile trip, you need two hours of travel time. Your time is all used up. Another answer that is often given is seventy five miles per hour in the second segment of the trip. Although not the accepted answer, there is a sense in which this answer is also correct.

How can there be two correct and different answers to the same question? The answer is that, as so often happens, the question is not well posed. The issue is what average is being asked for?

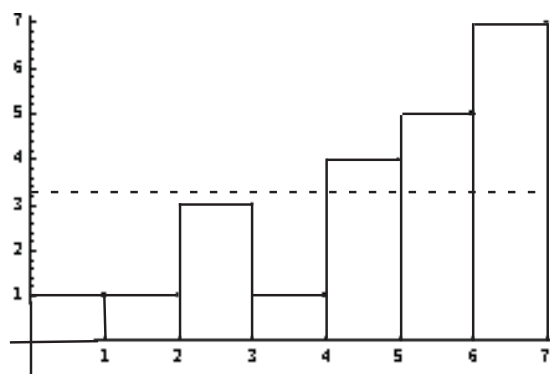


Figure 6.14: **Plot of Discrete Function for Averaging** The set of numbers, 1,1,3,1,4,5,7, are plotted as a discrete function in terms of the position of the number in the table. In addition, a bar is drawn from the next lowest location at the height of the value. Also the average, $\frac{22}{7}$ is shown as a dotted line. The area under the barred segments and the area under the dotted line are the same. This allows a more general definition of the process of averaging: the average times the interval is equal to the area under the barred plot of the discrete function generated by the set of numbers.

How do you compute an average? What is the average of the set of numbers 1,1,3,1,4,5,7. The rule is that you add up all the numbers, the sum is 22, and divide by the number of numbers which is 7. The result is $\frac{22}{7}$ or a pretty good π . Looking at this process more closely, you realize that what we have is an ordered set of numbers: the first number is 1, the second

number is 1, the third number is 3, and so forth. We have a mapping of the set of integers onto our set of numbers, a discrete function. In this language, we can say that to compute the average by sequencing through our ordered set: add the first number to the second, add that sum to the third, add that sum to the fourth, and so forth. You divide by the number of times you take a number. We can display this algorithm for this case in the form

$$\text{Average} \equiv \frac{\sum_{i=1}^n f(i)}{n} \quad (6.23)$$

where $f(i)$ is the value of our discrete function for the i element of the table and n is the number of entries or more interesting as a plot of the discrete function that we have generated, see Figure 6.14 on page 179. In addition to plotting the function as a bar graph, the average is shown as a horizontal dotted line. From the figure, it can be seen that the area under the bars of the bar graph and the area under the dotted line are the the same. This leads to an alternative algorithm for finding the average of a set of numbers: construct the bar graph for the set of numbers and calculate the area under the bar graph divide this area by the number of elements in the set. The advantage of this definition is that it is easy to extend to situations where you want the average over a continuously varying set. An algorithm for this definition is:

$$\text{Average} \equiv \frac{\sum_{i=1}^n f(i)\Delta i}{\sum_{i=1}^n \Delta i} \quad (6.24)$$

where Δi is the width of the elements of the bar graph.

From this construction, the more general definition of the average can be developed that will work for continuous functions. The integral form of this same definition is

$$\text{Average} \equiv \langle f \rangle_x \equiv \frac{\int_{x_0}^{x_f} f(x)dx}{\int_{x_0}^{x_f} dx} \quad (6.25)$$

where I have introduced a standard notation for taking the average. The subscript x indicates that the average is weighted by the variable x . The important point is that in different circumstances different weighting factors are appropriate and, although the definition looks as if it is independent of the choice of the weighting factor, it is not.

Now let's go back to our problem of the trip from Austin to College Station. To calculate an average, we need a set of numbers. How do we get the numbers? We have to decide what the weighting factor is. There are an infinity of choices but two are particularly obvious, time slicing and

space slicing. Were it not for a particular property of time slicing, space slicing is the easier because you will generally know how fast you can go at a given place. Thus to get the average velocity for space slicing choose spatial intervals and find the velocity in each. Applying this method to the Austin-College Station trip would yield the result that a speed of $75 \frac{\text{m}}{\text{hr}}$ in the second segment would give an average speed of $50 \frac{\text{m}}{\text{hr}}$.

The more accepted answer is the one that comes from using time slicing. In this case, the average is computed simply for a kinematic quantity like velocity because it is defined in terms of a time derivative. In other words,

$$\begin{aligned}
 \langle v \rangle_t &= \frac{\int_{t_0}^{t_f} v(t) dt}{\int_{t_0}^{t_f} dt} \\
 &= \frac{\int_{t_0}^{t_f} \frac{dx}{dt} dt}{t_f - t_0} \\
 &= \frac{x_f - x_0}{t_f - t_0}, \tag{6.26}
 \end{aligned}$$

and thus the average velocity is just the net displacement divided by the total time interval. You loss track of the fact that you time sliced. Unless stated otherwise it is customary to assume that what is wanted is the time averaged case.

In Section 6.2.1 on page 166, there was some question regarding the height to use in the Lagrangian since it varied in the segment. We now see that the correct choice is the time average since the action is time sliced. For cases where you replace the curved trajectory with a straight line the two averages always come out the same and thus our substitution was correct. In cases where you are using a more subtle structure such as in Section 6.2.2 on page 169, you would get the wrong answer by substituting the mean position.

It also important to note that the action principle always uses time slicing – it is a part of the definition. It could turn out that, in some applications, a different slicing is easier to understand, see Section 15.1 on page 331. In fact, when we did Fermat least time, we did segment slicing. Whatever slicing technique is chosen, the action must always be evaluated using a time slicing.

Chapter 7

Basic Principles of Physics

7.1 Symmetry

Symmetry is one of those concepts that occur in our everyday language and also in physics. There is some similarity in the two usages, since, as is usually the case, the physics usage generally grew out of the everyday usage but is more precise. Let's start with the general usage. Synonyms for symmetry are words like balanced or well formed. We most often use the idea in terms of a work of art. The following 4th century greek statue, Figure 7.1 on page 184, of a praying boy is a beautiful work of art. This is attributable to the form and balance. The figure has an almost exact bilateral, axial reflection, symmetry. A bilateral symmetry is a well defined mathematical operation on the figure: Establish a mean central axis and place a mirror there to reflect every point on the object in the plane plane of the mirror. You recover almost the same figure. In fact a Platonist would attribute the beauty in the piece to the presence of the mathematical symmetry. Of course, for this case, the symmetry is not exact but approximate.

These ideas about symmetry can be generalized and at the same time made more specific. In art and in physics, the idea is that you perform some algorithmic or well specified operation to the figure or system of interest. If you recover the same figure or system then you have a symmetry. Later on we will get very specific as to the definition of symmetry but the basic idea that you see here will endure. There is some change that you can make and if after you make the change you have basically the same thing that you started with, you say that you have a symmetry. If you recover almost the same figure or system, you have what is called a slightly broken symmetry or approximate symmetry.



Figure 7.1: **Praying Boy** In art, as it will turn out to be the case in physics, there is a sense of beauty associated with balanced or symmetric figures. This ancient greek statue of a praying boy has an approximate bilateral symmetry.

The first issue is to understand the idea of making a change. In order to differentiate the parts of this problem, we will call these changes transformations. There are obviously many transformations that you can perform both in physics and in art. Moving the figure to the side is an especially simple example. The set of operations that are shifting of the figure is an example of what is called a translation. In art, if the figure is the same after it has been translated, the figure possess translation symmetry; the transformation is a translation and there is a symmetry if the figure is the identical to the original. In most cases in art with translation symmetry, the amount of translation that reproduces the original image is an integer multiple of some fixed amount, see Figure 7.3 on page 185. This is an example of a discrete translation symmetry. Our earlier example of bilateral transformations or mirror images is also an example of a discrete family of transformations. This is an especially simple family since, if you do the transformation twice, you have not done anything. There are thus only two transformations in the bilateral set: mirror image or leave alone. The case of Figure 7.3 on page 185, there are many translations that produce a symmetry. In fact,

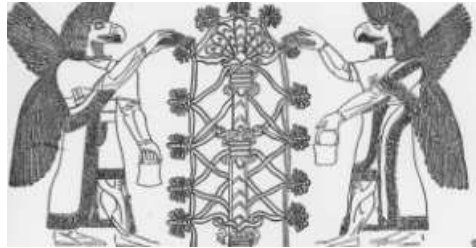


Figure 7.2: **Ancient Drawing** This ancient drawing shows an example of bilateral or reflection symmetry. Close inspection reveals that the symmetry is broken in an interesting way.

there is an infinite countable set of transformations, i. e. the transformations that produce a symmetry can be mapped onto the set of integers. Note that any combination of translations in the set of discrete translations is also a discrete translation. This is an important property of a family of transformations: they always contain in the family all combinations of the elements. In addition, they also contain the element that is no change and they also always contain an element that undoes what another element does. In the bilateral case, the only non-trivial element undoes itself if it is applied again. For the case of Figure 7.3 on page 185, you can reverse the direction of the original translation and shift the same amount.



Figure 7.3: **Borders** Note how border images tend to have discrete translation symmetry. It also has bilateral symmetry. Of course, we are assuming that the border extends indefinitely in both directions.

Another well known example of transformations in art and physics is rotations about an axis. Snowflakes are an interesting example, see Figure 7.4 on page 186. They possess a discrete rotational symmetry. Rotations of an

integer multiple of $\frac{2\pi}{6}$, reproduce the original image. Again, like the bilateral transformation, after so many of these rotations you can get back to doing nothing. This is a more interesting example of the discrete transformations with a finite number of elements than the bilateral case besides doing nothing there are five non-trivial rotations, $\frac{\pi}{3}$, $\frac{2\pi}{3}$, π , $\frac{4\pi}{3}$, and $\frac{5\pi}{3}$. In addition, the snowflake also has a bilateral. In fact since it has the discrete rotations, it actually has several bilateral transformations. These are along axis at 0 , $\frac{2\pi}{12}$, $\frac{4\pi}{12}$, $\frac{6\pi}{12}$, $\frac{8\pi}{12}$, and $\frac{10\pi}{12}$. These being combinations of the bilaterals and rotations.



Figure 7.4: **Snowflakes** Snowflakes provide an excellent natural example of a system with a discrete rotation symmetry. It also has a bilateral symmetry and, since it has a rotational symmetry, actually has several bilaterals.

As stated earlier, symmetry is a change to a system or, in the case of art, a figure or a statue that is not an important change. From these examples it is important to realize that to have a symmetry, you need a set of changes to the figure **and** then a criteria for these not being an important change. In the case of art, the criteria for not being important is that the pieces fall on top of each other. You could have a much more relaxed definition of unimportant change. For example consider the world of three sided figures whose sides are straight lines, triangles. If your criteria for unimportant change is that after the transformation you still have a triangle, then any transformation short of opening or bending one of the sides will be a symmetry. You could have a more restrictive criteria such as that the triangles be similar. In this case, rotations and rescaling all lengths would be a symmetry but changing

the size of one of the sides and not the others would not. It is important to keep in mind that the concept of symmetry is a two step process – a family of transformations and a rule about what is an important change.

Although we did not discuss it in those terms, we have already had an example of a symmetry in physics when we looked at the change in scale when we discussed dimensional analysis, see Section 2.5 on page 44. If we change the scale of length, all the numbers change but the things that happen still happen; it doesn't matter whether you make the measurements in the cgs system, the mks system, or english system, the physics is the same. We can use this as a rather loose definition of what we mean by a symmetry in physics. As we develop our vocabulary more fully, we can make this definition much more precise.

For all of the discussion so far we have defined the transformations as changes to the figure; rotate the figure by $\frac{2\pi}{6}$. With the example of change in scale, we can see a different but clearly equivalent approach. Instead of stretching the figure, we can just use a smaller length scale to discuss its size. In the old perspective, you can also look at it as if all lengths increased and the unit of length stayed the same. Here you now say that the figure stays the same and the unit of length changes. This is the difference between the active and the passive view of a transformation. In the active view, you change the figure, in the passive view the figure is left unchanged but the observers perspective is changed. In the active view, you then have another perspective

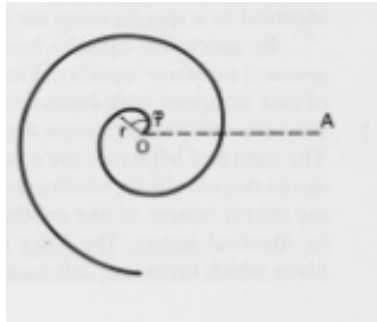


Figure 7.5: **Spiral** The spiral is generated by stretching the radius as you rotate. This is an example of a situation in which you combine two simple transformations to generate a figure with symmetry.

in symmetry. You can use the transformation to generate a figure that will automatically be symmetric. An extreme example of a symmetry is the infinitely long straight line. It satisfies bilateral symmetry about every

point. It satisfies a translation symmetry of any amount. It is homogeneous, same everywhere, and isotropic, same in both directions which are all the directions that it has. In turn, you can think of the straight line as the figure that is generated by translating a point to generate a continuous figure. Another important example is the circle. As a figure it is symmetric under rotations about the center. It can also be considered the locus of points that are equidistant from some fixed point and is generated by rotation of point at the appropriate distance from the center.

As in the snowflake example, Figure 7.4 on page 186, the family of transformations used in an active transformation includes all possible combinations of all of the elements of the family. In many cases, the resulting transformations can be a little surprising. The spiral is a shape that is generated by a compound of several simpler operations, stretch the radius as you rotate. In this case, the figure has a symmetry if as you translate in angle you stretch the distance from the origin. An interesting related example taken from biology is the shell seen in Figure 7.6 on page 188.



Figure 7.6: **Shell** The shell is an interesting example of a symmetric system. As you rotate, you translate and stretch the radius.

7.2 The Nature of Symmetry in Physics

In many respects, symmetry in physics is very similar to that in art; there are families of transformations that lead to unimportant changes in the situation. The differences deal with the things on which the transformations act and the definition of unimportant. As expected, in addition, the language that described the actions are more precise and abstract. We will also categorize the transformations of physics in a formal way and use these labels to describe important results.

7.2.1 Discrete Transformations

These are changes that can only be applied in discrete steps. Bilateral or mirror symmetry about a plane is an example from art. For the snow flakes or a particular snow flake, see Figure 7.4 on page 186, the rotations at $\theta = n\frac{2\pi}{6}$ for $n = 0, 1, 2, \dots, 5$ is an example of a family of discrete rotations that produce a symmetry for a snowflake. The rotated snow flake falls onto the image of that snow flakes original configuration exactly.

The example in physics that corresponds to bilateral symmetry is called a spatial inversion which is to replace places in one directions by their opposite. In a world with on space dimension, replace x by $-x$. In a world with three spatial directions, replace (x, y, z) with $(-x, y, z)$. This is like placing a mirror in the plane $y = 0, z = 0$. This is obviously a discrete transformation. You also note that, if it is applied twice, there is no change. It is said to be a discrete transformation of cycle two; it has two elements, do nothing, the identity transformation, and the inversion. There are many discrete transformations of cycle two: if you have identical particles, you can interchange the particles, you can invert the time, you can do a spatial inversion along the y or z axis, ...

There are, of course, discrete transformations with cycles higher than two. The snowflake example from art carries over to physics. Rotations about the origin by an angle of $\frac{2\pi}{n}$ is an example of a discrete transformation with n cycles.

You can also have a family of discrete transformations that have an infinite number of elements. In one spatial dimension, you can shift the origin by a fixed amount a , see Figure 7.3 on page 185. You can do this any number of times generating a set of transformations that has a countable infinite number of members.

It is important to realize that the method by which the members of a family of discrete transformations are labeled must itself be a discrete set of labels and that the members of a discrete set of transformations cannot be labeled by a continuous variable.

7.2.2 Continuous Transformations

Continuous transformations are changes that can be applied for arbitrarily small changes. The labeling of the transformations is a continuous parameter. Rotations about a point are a valuable example. In art, a world of concentric rings would enjoy a symmetry for rotations about the center point. These changes in angle can take any value from zero to 2π . This idea

is carried over to physics. In a three dimensional space, rotations about an axis are a family of transformations. These transformations are an example of continuous transformations. Other obvious examples are translations in space and time. Changes in the scale of length discussed in Sections 1.5.1 on page 25, and 2.5.2 on page 47 is also a continuous set of transformations. **Again it is important to realize that a continuous family of transformations can only be labeled by a continuous variable.**

It is possible to make a discrete family of transformations from subsets of continuous transformations such as the set of rotations used in the snowflake example of Figure 7.4 on page 186. Of course, the reverse process is not possible; you cannot make a continuous family of transformations from a subset of a discrete family no matter how large the set of discrete transformations.

7.2.3 Identity Transformation

Another principle that is essential to the mathematical consistency of transformation theory is that a complete set of transformations must include an identity transformation. That is one that changes nothing. For any collection of transformations, The example $n = 0$ in the discrete case above is an identity transformation. Note that $n = 6m$ where $m = 1, 2, 3, \dots$ is also the identity and we already had it in the set of transformations. In fact, any transformation in which $n > 6$ is the same as the transformation $n' = \text{mod}_6(n)$.

7.2.4 Examples of symmetry in situations like physics

You are planning a trip between Austin and College Station. There are several routes.

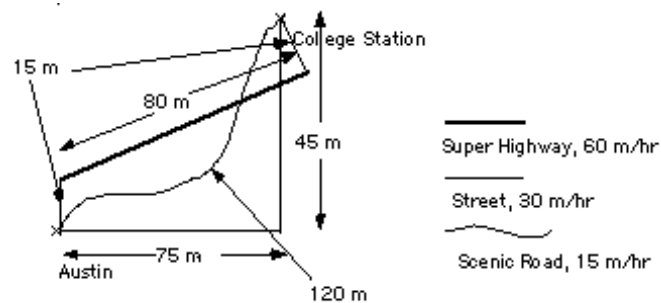


Figure 7.7: Paths to Texas A&M Miles to AM

There are several criteria that you can use to select the route: least time, least distance, see most trees and hills - one hill is worth a dozen trees. There are several changes that you can make in the system: interchange Austin and College Station, interchange super highways and streets, make the speed limit $50 \frac{\text{m}}{\text{hr}}$, measure all distances in feet. These are all discrete transformations. You could shift the entire thing a distance x to the east and we all know that as you go east there are no longer any hills. You could shift all the distances by a scaling factor α . These are continuous transformations. For all of these you can see if the transformation effects the evaluation of the criteria.

From this example you see that you need both a set of transformations and a criteria.

In physical systems, we can either change the events in the transformation process or change the measuring system that is used to identify the events. The former case is called the active view of transformations and the latter is the passive view. Obviously, they are equivalent descriptions of the effects of the transformations and which is being used is chosen by the context of the problem.

7.3 Examples of Symmetry in physics

In physics we are interested in what happens to things in space time, i. e. events. These are labeled by (x,t) . An event is a point in a space time diagram. A connected set of events is a trajectory. This is the path that a particle follows as it moves. This is often called a particles world line.

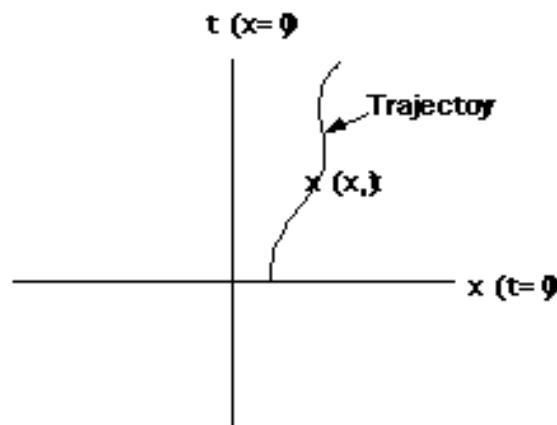


Figure 7.8: Action trajectory Trajectory 2

7.3.1 Physics transformations:

Space Reflection:

This is the transformation that corresponds to the bilateral transformation that we discussed earlier. We reflect all the events through the line $x = 0$ better known as the t axis.

$$x \rightarrow x' = -x \quad (7.1)$$

I am showing this transformation in the active view.

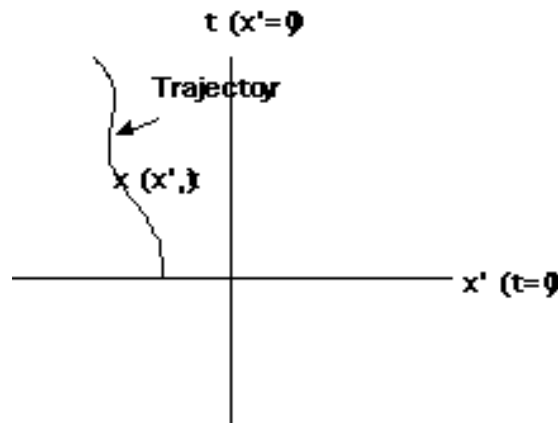


Figure 7.9: **Space Reflection** Space Reflection

Space Translation:

Shift the origin of the coordinate system.

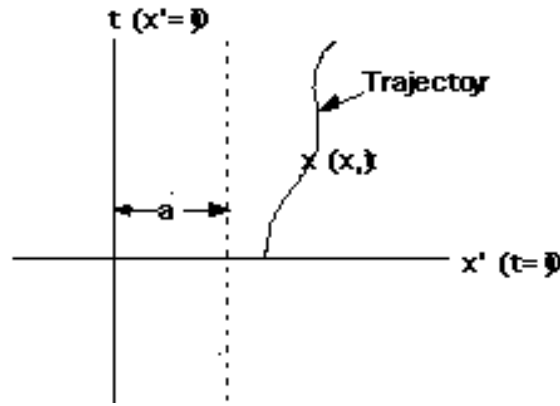
$$x \rightarrow x' = x + a \quad (7.2)$$

Time Translation:

Shift the start of the time.

$$t \rightarrow t' = t + a \quad (7.3)$$

. To be a symmetry we will require that the physics before and after the shift is the same. I have not carefully defined what I mean by "the same." I will do so shortly.

Figure 7.10: **Space Translation** Space Translation**Newton's Action at a Distance Law of Gravitation**

The law of force that describes the gravitational influence of one body, say body 2, on another body, say body 1, is

$$\vec{F}_{1,2} = G \frac{m_1 m_2}{|\vec{r}_2 - \vec{r}_1|^3} \times (\vec{r}_2 - \vec{r}_1) \quad (7.4)$$

Similarly, the gravitational force of body 1 on body 2 can be found by interchanging the labels of particles 1 and 2.

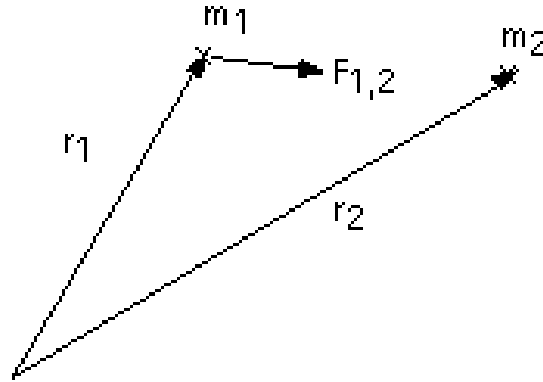
$$\vec{F}_{2,1} = G \frac{m_2 m_1}{|\vec{r}_1 - \vec{r}_2|^3} \times (\vec{r}_1 - \vec{r}_2) \quad (7.5)$$

Thus if you are operating at the level of the forces you have that if you interchange particles 1 and 2, i. e. change the labels 1 and 2, $1 \leftrightarrow 2$ and get $\vec{F}_{1,2} \rightarrow -\vec{F}_{2,1}$. This is a discrete transformation. If for some reason you are interested in the forces, this is not a symmetry. It is actually a manifestation of the Law of Action Reaction. In other words, we construct the Law of Gravitation so that it obeys the Law of Action Reaction. On the other hand, if you look at the entire set of equations without the forces, there is no change.

$$m_1 \vec{a}_1 = G \frac{m_1 m_2}{|\vec{r}_2 - \vec{r}_1|^3} \times (\vec{r}_2 - \vec{r}_1) \quad (7.6)$$

$$m_2 \vec{a}_2 = G \frac{m_2 m_1}{|\vec{r}_1 - \vec{r}_2|^3} \times (\vec{r}_1 - \vec{r}_2) \quad (7.7)$$

Some symmetries of this law:

Figure 7.11: **Gravitational Force between Two Bodies**

This is then a symmetry. When you put a shift to all the positions by some amount, $\vec{a}$, nothing changes, i. e. $\vec{r}_i \rightarrow \vec{r}_i + \vec{a}$. This is a continuous symmetry. When you replace all the positions with the reverse position, $\vec{r}_i \rightarrow -\vec{r}_i$ again nothing changes. Remember $\vec{a}_i \rightarrow -\vec{a}_i$. This is a discrete symmetry. If you change all the distances in the problem by a scale $\vec{r}_i \rightarrow \vec{r}'_i = \lambda \vec{r}_i$, then this is not a symmetry. But, if you also change the time scale by $t \rightarrow t' = \lambda^{\frac{3}{2}} t$, then you have a symmetry. This is a continuous symmetry. Note that the identity transformation is $\lambda = 1$.

7.4 Symmetry and Action

7.4.1 Introduction

You can have the situation that you make the change and the action does not change at all. Said more carefully, you have transformed end points and transformed paths and you get the same value for the action. Consider the free particle and translations in space.

$$\begin{aligned} x' &= x + a \\ t' &= t \end{aligned} \tag{7.8}$$

This implies that $v' = v$. Thus

$$\begin{aligned}
S'(x'_f, t'_f, x'_0, t'_0; path') &= \sum_{path', (x'_0, t'_0)}^{(x'_f, t'_f)} \left(m \frac{v'^2}{2}\right) \Delta t \\
&= \sum_{path, (x_0, t_0)}^{(x_f, t_f)} \left(m \frac{v^2}{2}\right) \Delta t \\
&= S(x_f, t_f, x_0, t_0; path) \quad (7.9)
\end{aligned}$$

If action is the basis of all physics, then we have a natural definition of a symmetry of a physical system. A physical system has a symmetry if there is a way to modify the system and yet there is no significant change in the action. It is important to be careful about the meaning of significant in this sentence. For most purposes the value of the action is not important. The action primary role is to select a path from the infinity of possibilities. In this sense, we can as a first step assert that the system is symmetric if the system before and after the change still selects the same path as the natural path. You again have to be careful because the same path is actually the same path as seen in the modified system. An example might help clarify this.

Harmonic Oscillator and Symmetry

The harmonic oscillator is one of the most important physical systems. We will discuss the physics of this system in greater detail in a later section, Section 8.2 on page 205, but for now will use it as another example in which to examine the role of symmetry in a physical system. For now just think of it as a physical system that goes back and forth.

The Lagrangian for the harmonic oscillator is

$$L(v, x) = KE - PE = m \frac{v^2}{2} - k \frac{x^2}{2} \quad (7.10)$$

where k is the spring constant and m is the mass and both are given constants and have the dimension $k \stackrel{\text{dim}}{=} \frac{\text{Mass}}{\text{Time}^2}$ and, of course, m is a mass. Note that, if these are the only two dimensional constants that are available, then you cannot make a length but you can make a time. If you rescale the distances by an amount λ , as follows:

$$\begin{aligned}
x &\rightarrow x' = \lambda x \\
t &\rightarrow t' = t
\end{aligned} \quad (7.11)$$

which implies that

$$v \rightarrow v' = \frac{\Delta x'}{\Delta t'} = \lambda \frac{\Delta x}{\Delta t} = \lambda v \quad (7.12)$$

The Lagrangian for the new system is

$$L'(v', x') = KE' - PE' = m \frac{v'^2}{2} - k \frac{x'^2}{2} = m \lambda^2 \left(\frac{v^2}{2} - k \frac{x^2}{2} \right) = \lambda^2 L(v, x) \quad (7.13)$$

So that

$$\begin{aligned} S'_{Path'}(x'_0, t'_0; x'_f, t'_f) &= \sum_{path', (x'_0, t'_0)}^{(x'_f, t'_f)} \left(m \frac{v'^2}{2} - k \frac{x'^2}{2} \right) \Delta t' \\ &= \lambda^2 \sum_{path, (x_0, t_0)}^{(x_f, t_f)} \left(m \frac{v^2}{2} - k \frac{x^2}{2} \right) \Delta t \\ &= \lambda^2 S_{Path}(x_0, t_0; x_f, t_f) \end{aligned} \quad (7.14)$$

where Path' is the Path that is at the rescaled distances

$$x'(t') = \lambda x(t) \quad (7.15)$$

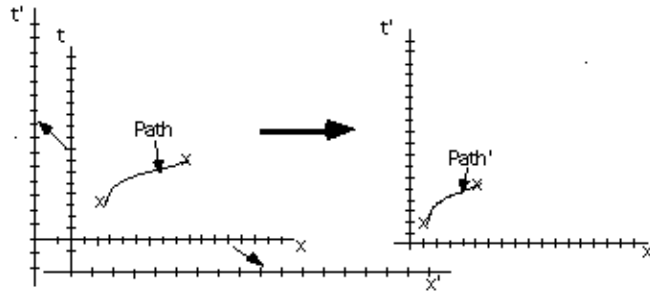


Figure 7.12: **Rescale Oscillator** Rescale Oscillator

Path	Action	Path'	Action'
1	S_1	1'	$S'_{1'} = \lambda^2 S_1$
2	S_2	2'	$S'_{2'} = \lambda^2 S_2$
.	.	.	.
.	.	.	.
.	.	.	.
natural	S_{least}	natural'	$S'_{least'} = \lambda^2 S_{least}$
.	.	.	.
.	.	.	.
.	.	.	.

You get the same path even though the calculations are all different.

7.4.2 Galilean Invariance

In order to show that the straight line was the solution to the free particle action problem I assumed that the action procedure was Galilean invariant and went to a special frame. The question is “is it.” The action is

$$S(x_f, t_f, x_0, t_0; traj.) = \sum_{traj., x_0, t_0}^{x_f, t_f} \left(m \frac{v^2}{2} \right) \Delta t \quad (7.16)$$

What happens when you make the Galilean transformation?

$$\begin{aligned} x' &= x - at \\ t' &= t \end{aligned} \quad (7.17)$$

Where a is a parameter that labels the transformations and has the dimensions of a velocity – it is actually interpreted as a velocity. With this transformation all the velocities shift, $v' = v - a$ because $v \equiv \frac{x(t+\Delta t) - x(t)}{\Delta t}$ and Inserting this into

$$\begin{aligned} v' &\equiv \frac{x'(t' + \Delta t') - x'(t')}{\Delta t'} = \frac{x(t + \Delta t) - a\{t + \Delta t\} - x(t) + a\{t\}}{\Delta t} \\ &= \frac{x(t + \Delta t) - x(t)}{\Delta t} - \frac{a\Delta t}{\Delta t} \\ &= v - a. \end{aligned} \quad (7.18)$$

Plugging this into the action,

$$S'(x'_f, t'_f, x'_0, t'_0; path') = \sum_{traj.', x'_0, t'_0}^{x'_f, t'_f} \left(m \frac{v'^2}{2} \right) \Delta t'$$

$$\begin{aligned}
&= \sum_{traj., x_0, t_0}^{x_f, t_f} \left(m \frac{(v-a)^2}{2} \right) \Delta t \\
&= \sum_{traj., x_0, t_0}^{x_f, t_f} \left(m \frac{v^2}{2} \right) \Delta t - \sum_{traj., x_0, t_0}^{x_f, t_f} (mva) \Delta t + \sum_{traj., x_0, t_0}^{x_f, t_f} \left(m \frac{a^2}{2} \right) \Delta t \\
&= S(x_f, t_f, x_0, t_0; traj.) - ma \sum_{traj., x_0, t_0}^{x_f, t_f} v \Delta t + \left(m \frac{a^2}{2} \right) \sum_{traj., x_0, t_0}^{x_f, t_f} \Delta t \\
&= S(x_f, t_f, x_0, t_0; traj.) - ma(x_f - x_0) + \left(m \frac{a^2}{2} \right) (t_f - t_0)
\end{aligned}$$

The action of any trajectory is different from the action for the transformed trajectory. Since the action changes the best that could happen is an invariance. The last two terms are independent of trajectory. Therefore the trajectory selection process that selects the least action trajectory in S will select the transformed trajectory in S' . The action changes under the transformation but in an unimportant way. This is not a symmetry and there is no associated conserved quantity. When we implement this for special relativity it will become a symmetry.

7.4.3 More on Symmetry and Action

The easiest way to guarantee that the action is symmetric under a set of transformations is to construct it only from the form invariants for that set of transformations. In fact, it is a necessary and sufficient condition that the action is symmetric that it be composed of only form invariants for that set of transformations.

As an example consider the action for a satellite of mass m in orbit around the earth. Locating the earth at the origin, the action is

$$S(\vec{x}_0, t_0, \vec{x}_f, t_f; path) = \sum_{Path, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \left(m \frac{\vec{v}^2}{2} + Gm \frac{M_{earth}}{r} \right) \Delta t \quad (7.19)$$

This action is composed of $\vec{v}^2$ which is a form invariant for rotations about the origin. r is the distance from the origin and it is also a form invariant for rotations. Obviously Δt is a form invariant for rotations. Thus this action has a symmetry that is the set of transformations that are the rotations about the origin.

7.4.4 Noether's Theorem

For every transformation family whose members are labeled by an element of an $\mathbf{R}^1$ and whose members are connected smoothly to the identity transformation that is a symmetry there exists a conserved quantity during the naturally occurring trajectory. Noether's Theorem also tells us how to construct the conserved quantity.

Agreed. This is a most opaque statement and will require some clarification. When I tell you what the question is and thus when a change is important, I can tell you how to construct the conserved quantity.

Space translation Symmetry

The conserved quantity that is associated with situations with space translation symmetry is called linear momentum. In certain cases it is $\vec{p} = m\vec{v}$ but not all the time. I will tell you when those cases are.

Rotation translation symmetry

The conserved quantity that is associated with situations with space rotation symmetry is called angular momentum. Rotations are a vector quantity. Again in certain cases it is $\vec{L} = m\vec{r} \times \vec{v}$.

Time translation Symmetry

The conserved quantity that is associated with situations with time translation symmetry is called energy. This is actually the case all the time but the form of the energy may change.

Galilean Invariance*

This is almost a symmetry classically and becomes a full blown symmetry in the modern language. First, let's discuss what the transformation is.

There is no experiment that can be performed that can measure the velocity of a uniformly moving observer. We can detect the presence of accelerations and measure the relative velocity between two bodies but we cannot measure velocities absolutely.

Another way to say the same thing is that, if you are not accelerating, you are always at rest in your own rest frame. It is a requirement for

consistency that then leads to the special category of observers¹ that have no acceleration. Their observations always must agree.

In the language of transformations, all the laws of physics must be invariant under transformations of the form

$$\begin{aligned}\vec{x} &\rightarrow \vec{x}' = \vec{x} + \vec{R} + \vec{v}t \\ t &\rightarrow t' = t\end{aligned}\tag{7.20}$$

where $\vec{R}$ and $\vec{v}$ are constants that are the parameters that label these continuous transformations. Unfortunately the parameter $\vec{v}$ is often called a velocity because dimensionally it is a $\text{dim } LT$ which has another identification as a velocity. These parameters can be interpreted in terms of two coordinate systems this can be interpreted as the measurements of two relatively displaced and relatively moving coordinate systems.

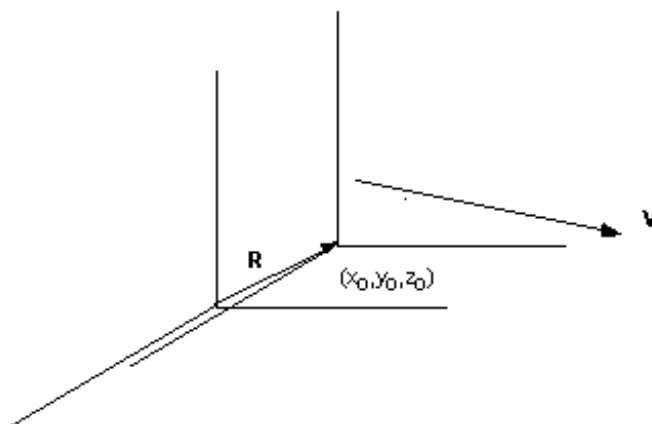


Figure 7.13: **Galilean Invariance** Galilean Invariance

Although this is a continuous symmetry that is connected with the identity, $\vec{R} = \vec{v} = 0$, it is not a symmetry classically. I will explain this later. Since this is not a symmetry, there is no conserved quantity that is the result of Galilean invariance in classical physics.

You should apply this transformation to the gravitational force above and see that neither the forces nor the equations change. If you use these as your criteria for a symmetry, this would be a symmetry. It is not so we see that we need a better criteria.

¹The idea of observer is a subtle and important one. The observer is a general title for a measuring system. Sorry for the anthropomorphism. The other important point that all observers that are at rest have the same results for their experiments; they are good physicists.

Please read the Feynman's action lecture. I do not expect that all of you will follow this material in detail. Just get the drift.

Consider a change in the system that also changes the description of initial and final events. This is what will generally happen. Here, when you do the transformations, you will get in addition to the usual terms of the integral of the Lagrangian but also terms from the end points. Our modified form of Feynman's equation

$$\begin{aligned} \delta S = & \left(\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right) \delta x \right)_{t_f} - \left(\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right) \delta x \right)_{t_0} \\ & + \int_{t_0}^{t_f} \left(\frac{d}{dt} \left(\frac{\delta L}{\delta v} \right) - \frac{\delta L}{\delta x} \right) \delta x dt \end{aligned} \quad (7.21)$$

To get the action to be stationary now we will require that as before the integrand vanish

$$\frac{d}{dt} \left(\frac{\delta L}{\delta v} \right) - \frac{\delta L}{\delta x} = 0 \quad (7.22)$$

but also that the terms from the end points vanish. This part simply selects the natural path. To understand the end points consider an example, the simple translation. In this case δx is simply a number that is added to all points in the path.

$$\delta x(t_f) = \delta x(t_0) = a \quad (7.23)$$

or

$$\left(\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right) \delta x \right)_{t_f} - \left(\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right) \delta x \right)_{t_0} = \left(\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right)_{t_f} - \left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right)_{t_0} \right) a \quad (7.24)$$

Setting this to zero, yields

$$\left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right)_{t_f} = \left(\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)} \right)_{t_0} \quad (7.25)$$

But $\frac{\delta L}{\delta v} \Big|_{x_{nat}(t)}$ is what you would **define** as the momentum. It is the momentum when you use the usual Lagrangian. Thus this is nothing more than the statement that momentum is conserved.

$$p(t_f) = p(t_0) \quad (7.26)$$

This is a special case of a general theorem called Noether's Theorem. Given any transformation that can be connected with the identity transformation, no change, by a continuous parameter. There will always be a conserved quantity. In the above example the transformation is translation. In the limit $a \rightarrow 0$ you have no translation and thus no change and the identity transformation. In this case, the conserved quantity is the linear momentum.

Another way of looking at this result is that, once you have selected the natural path and if you include the end point variations, the action is a function of the end points only. If the symmetry transformation changes the end points you have

$$\delta S_{Nat}(x_0, t_0; x_f, t_f) = \frac{\delta S_{Nat}}{\delta x_0} \delta x_0 + \frac{\delta S_{Nat}}{\delta x_f} \delta x_f + \frac{\delta S_{Nat}}{\delta t_0} \delta t_0 + \frac{\delta S_{Nat}}{\delta t_f} \delta t_f \quad (7.27)$$

In the case of translations,

$$\delta x(t_f) = \delta x(t_0) = a \quad (7.28)$$

and all the δt_i are zero.

Thus we get

$$\frac{\delta S}{\delta x_f} = -\frac{\delta S}{\delta x_0} = p = \text{constant} \quad (7.29)$$

An Example

For the free particle,

$$S_{natural} = m \frac{(x_f - x_0)^2}{2(t_f - t_0)} \quad (7.30)$$

$$p = \frac{\delta S}{\delta x_f} = m \frac{(x_f - x_0)}{(t_f - t_0)} = mv \quad (7.31)$$

since v is a constant.

We noted above that the satellite in orbit is a case that is invariant under rotations about the origin. This set of transformations is a continuous set and thus there is a conserved quantity. In this case we call it the angular momentum. The construction of this conserved quantity involves cumbersome notation because it only makes sense in a system with at least two spatial dimensions and thus involves vector notation. In addition, it is computationally difficult to find an expression for the natural path. But note that the free particle Lagrangian is also composed only of form invariants for

rotations about the origin. Thus this set of transformations is also a symmetry for this case. The analysis is still cumbersome because of the vector notation. I am aware that you will not be able to reproduce this analysis. All that I ask is that you follow it.

We will work in two spatial dimensions. For this case the action is

$$S(\vec{x}_0, \vec{t}_0; \vec{x}_f, \vec{t}_f) = \sum_{\text{NaturalPath}, \vec{x}_0, \vec{t}_0}^{\vec{x}_f, \vec{t}_f} m \frac{\vec{v}^2}{2} \Delta t \quad (7.32)$$

and as we see is composed of only form invariants not only of translations in space and time but also for rotations. The quantity $\vec{v}^2$ is invariant under rotations.

For the natural path the action is

$$S_{\text{natural}} = m \frac{(\vec{x}_f - \vec{x}_0)^2}{2(t_f - t_0)} \quad (7.33)$$

and the change in the action caused by the end point changes are

$$\delta S_{\text{Nat}}(\vec{x}_0, t_0; \vec{x}_f, t_f) = \frac{\delta S_{\text{Nat}}}{\delta \vec{x}_0} \cdot \delta \vec{x}_0 + \frac{\delta S_{\text{Nat}}}{\delta \vec{x}_f} \cdot \delta \vec{x}_f + \frac{\delta S_{\text{Nat}}}{\delta t_0} \delta t_0 + \frac{\delta S_{\text{Nat}}}{\delta t_f} \delta t_f \quad (7.34)$$

For rotations, δt_0 and δt_f are zero. The $\delta \vec{x}_0$ and $\delta \vec{x}_f$ are the displacements of the end points that result from the rotation. For a rotation through an angle θ , they are

$$\delta \vec{x}_0 \quad (7.35)$$

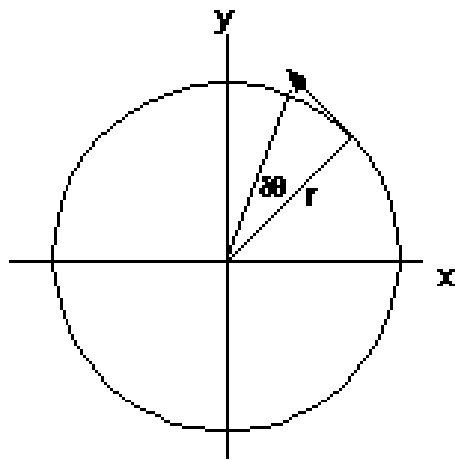
From the rule above we need the change in the S_{Nat} along this direction.

As in the translation example we see that the change in S with changes in position is the regular momentum. Thus the thing that multiplies $\delta \theta$ in the change in action is the momentum along this direction times the distance. This is what we always called the angular momentum.

Thus we get the rather complicated object

$$L_{\text{axis}} = \frac{\delta S_{\text{Nat}}}{\delta \vec{x}_0} \cdot r_0(\theta)_0 \quad (7.36)$$

The lesson of all this is that the symmetry implies that there is a conserved quantity. These are the things that we call momenta or energy etc. The form that they take depends on the nature of the Lagrangian.

Figure 7.14: **Rotation** Rotation.

Chapter 8

Special Classical Physical Systems

8.1 Introduction

In order to understand the ideas of modern physics, it is essential to understand the operations of some special classical systems. Not only do these provide a physical intuition but also a vocabulary. In the previous chapter, Chapter 7 on page 183, we dealt in some detail with two important physical systems, the free particle and the particle moving in a constant force. These were dealt with there to illustrate the principles and uses of symmetry and action. They obviously belong to the category of “Special Classical Physical Systems” but since they were treated there will not be treated here. Instead we will deal with the harmonic oscillator as an example of a more complicated but still simple system and the string as an example of a field system.

8.2 The Harmonic Oscillator

8.2.1 Importance

After the free particle, the harmonic oscillator is the most important mechanical system. Harmonic oscillators or systems that are almost harmonic oscillators are ubiquitous in nature. These are basically objects that when disturbed slightly return to their starting position but because of inertia overshoot and jiggle back and forth around the neutral position. The simplest example is the simple spring with one end attached to a fixed place

with a mass attached on the end.

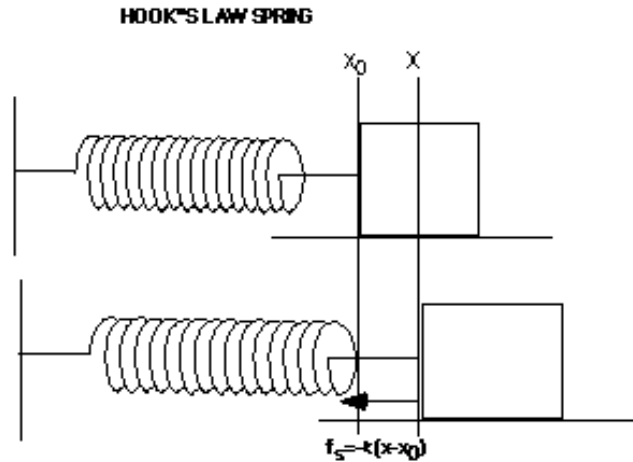


Figure 8.1: **A mass and Hook's Law Spring** A mass, m , on the end of an ideal spring is an example of a harmonic oscillator. An ideal spring or Hook's Law spring, is one in which the force at the end of the spring is proportional to the stretch of the spring, $F = k(x - x_0)$.

The general definition is that the system is a harmonic oscillator if the force on the system that emerges from movement from equilibrium is proportional to the amount of movement from equilibrium and is directed to remove the displacement from equilibrium. Defined this way, harmonic oscillators come in lots of forms. A mass on the end of a string suspended above the earth, if displaced to the side by a small amount is a harmonic oscillator. A shallow pan filled with water sloshes back and forth when disturbed and can be analyzed as a harmonic oscillator. We will discuss these examples in Section 8.2.3 on page 209. In a very real sense, any object that is held in place but still moves a little about that fixed point is generally well approximated by the harmonic oscillator system.

Even more important to our purposes, we will find that the harmonic oscillator is essential to the modern interpretation of the nature of particles. The quantum harmonic oscillator is the only system that can provide a framework for creating a quantum field theory that satisfies the requirements of having a particle interpretation

8.2.2 Dynamics

In the most general case, for a mass that can move freely in space, since acceleration and force are vector quantities $\vec{F} = m\vec{a}$, a harmonic oscillator is a system which obeys:

$$m\vec{a} = -k(\vec{x} - \vec{x}_0), \quad (8.1)$$

where $\vec{a}$ is the acceleration of the position of the block and $\vec{x}$ and $\vec{x}_0$ are the position and neutral position of the mass. k is called the spring constant. The sign is negative since we want the force to drive the system back to the neutral position. What are the dimensions of k ? Can you make a time with the dimensional parameters of this problem? Can you make a length? The dimensions of k are $\frac{\text{mass}}{\text{time}^2}$. The only dimensional parameters that involved are k and the mass, m . From these you can make a time, $\sqrt{\frac{m}{k}}$, but you cannot make a length. This lack of an intrinsic length but an intrinsic time will lead to a scaling invariance that is the basis for an interesting property of harmonic oscillators: the period of oscillation is independent of the amplitude of the oscillation.

For most of our purposes, it will be sufficient to deal with only one spatial dimension and, in this section, that is all that will be described. The results in higher spatial dimensions are easily generalized from the one dimensional case.

The Lagrangian for this system is

$$L(v, x) = \frac{m}{2}v^2 - \frac{k}{2}(x - x_0)^2. \quad (8.2)$$

Of course, this Lagrangian yields the correct one dimensional version of the dynamic for this system,

$$ma = -k(x - x_0). \quad (8.3)$$

What are the symmetries and invariances of this system? See Section 7.4 on page 194. Translation in the position coordinate: $x \rightarrow x + \alpha$? This is neither a symmetry nor an invariance for this action. Time translation: $t \rightarrow t + \alpha$? This is a symmetry and thus there is a conserved quantity, the energy, which we discuss below. A rescale of length: $x \rightarrow \lambda x$ produces an invariance. Thus systems with different lengths have the same physics. This is why the period is independent of the amplitude. A rescale of time: $t \rightarrow \lambda t$ produces neither a symmetry nor an invariance.

From the Lagrangian, we can construct the energy as the Noether conserved quantity for the time coordinate translation symmetry, see Section 7.3.1 on page 192 and Section 7.4.4 on page 199.

$$E = m \frac{v^2(t)}{2} + k \frac{\{x(t) - x_0\}^2}{2}. \quad (8.4)$$

Identifying the free particle motional energy as $m \frac{v^2}{2}$, there is a potential energy and it is $k \frac{(x-x_0)^2}{2}$. Actually most of you would have done this the other way. The potential energy is $V(x) = k \frac{(x-x_0)^2}{2}$ and the Lagrangian is $K.E. - V(x)$. I just wanted to emphasize the importance of the action approach which is the more fundamental approach.

There are two kinds of motion. If you displace the mass from the equilibrium position, x_0 , a distance d , the mass moves as:

$$x(t) = d \cos \left(2\pi \sqrt{\frac{k}{m}} t \right) + x_0 \quad (8.5)$$

It oscillates harmonically about the equilibrium position, x_0 , with a radian frequency $\Omega = 2\pi f = 2\pi \sqrt{\frac{k}{m}}$, where f is the usual cycle frequency.

If you have the mass at x_0 and give it an initial velocity, v_0 , it moves as:

$$x(t) = \frac{v_0}{2\pi \sqrt{\frac{k}{m}}} \sin \left(2\pi \sqrt{\frac{k}{m}} t \right) + x_0 \quad (8.6)$$

For the general case, you have a superposition of these two motions.

$$x(t) = d \cos \left(2\pi \sqrt{\frac{k}{m}} t \right) + \frac{v_0}{2\pi \sqrt{\frac{k}{m}}} \sin \left(2\pi \sqrt{\frac{k}{m}} t \right) + x_0 \quad (8.7)$$

The velocity is

$$v(t) = -d 2\pi \sqrt{\frac{k}{m}} \sin \left(2\pi \sqrt{\frac{k}{m}} t \right) + v_0 \cos \left(2\pi \sqrt{\frac{k}{m}} t \right) \quad (8.8)$$

This provides a wonderful example of a conserved quantity. Both $x(t)$ and $v(t)$ are changing all the time. Even the kinetic energy is changing and the potential energy is changing. Only when you take the combination of $E(t) = K.E.(t) + P.E.(t)$ do you get something that does not change. Plug

Equation 8.7 on page 208 for $x(t)$ and Equation 8.8 on page 208 for $v(t)$ into Equation 8.4 on page 208 and get that

$$\begin{aligned} E &= m \frac{(v(t))^2}{2} + k \frac{(x(t) - x_0)^2}{2} \\ &= m \frac{v_0^2}{2} + k \frac{d^2}{2}. \end{aligned} \quad (8.9)$$

8.2.3 Examples of harmonic oscillator systems

Besides being a nice simply solvable example of dynamical system, the oscillator is a very common example. Almost all systems that undergo bounded motion, act like an oscillator for small ranges of motion.

Consider the pendulum, a mass on the end of a flexible string suspended freely above the earth. This is certainly a case of bounded motion. How is it related to the harmonic oscillator system?

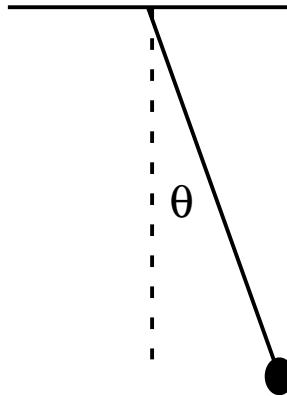


Figure 8.2: **Simple Pendulum** Simple Pendulum.

As is always the case in classical physics, the Lagrangian is $K.E. - P.E.$. For this case, the $K.E.$ is the usual $m \frac{v^2}{2}$. The $P.E.$ is our old friend mgh but, in this case, we want the dynamical variable to be the angle of the string from the vertical, θ . Using $h = l(1 - \cos(\theta))$, we have for the Lagrangian:

$$L(v, \theta) = m \frac{v^2}{2} - mgl(1 - \cos(\theta)) \quad (8.10)$$

where l is the length of the string in the pendulum. Again, there is a time translation symmetry and, with the use of Noether's Theorem, Sections 7.3.1 on page 192 and 7.4.4 on page 199, we can construct a conserved quantity

called the energy and we can identify the kinetic and potential energies. In this case, the potential energy is $V(\theta) = mgl(1 - \cos(\theta))$. Using the information from Section 1.4.2 on page 16, “Things Everyone Should Know” for small θ , $V(\theta) \simeq mgl\frac{\theta^2}{2}$. Also the kinetic energy is not directly related to how fast the angle θ is changing. Since this is our dynamical variable, we want to express the *K.E.* in terms of the rate of change. The linear speed, v is connected to the angular speed, $\frac{\Delta\theta}{\Delta t} \equiv \omega$ as $v = l\omega$. For small angles the pendulum has as its Lagrangian:

$$L(\omega, \theta) = ml^2\frac{\omega^2}{2} - mgl\frac{\theta^2}{2}. \quad (8.11)$$

Making a correspondence between v and ω and x and θ , we see that, in the limit of small θ and comparing to Equation 8.2 on page 207, the pendulum is an example of a harmonic oscillator. In other words, if we consider ml^2 to be an effective mass and mgl to be an effective spring constant the pendulum moves in exactly the same way as the harmonic oscillator. This means that the motion is harmonic about the equilibrium position, $\theta_0 = 0$, with radian frequency $\Omega = 2\pi\sqrt{\frac{mgl}{ml^2}} = 2\pi\sqrt{\frac{g}{l}}$. If you have a starting displacement θ_d and starting angular speed ω_0 ,

$$\theta(t) = \theta_d \cos\left(2\pi\sqrt{\frac{g}{l}}t\right) + \frac{\omega_0}{2\pi\sqrt{\frac{g}{l}}} \sin\left(2\pi\sqrt{\frac{g}{l}}t\right). \quad (8.12)$$

8.3 The Stretched String Revisited

8.3.1 Distributed Systems

Instead of having any masses concentrated into point, mass can be distributed. An example is the stretched elastic string, see Section 5.3 on page 128.

This is an example of a field. The disturbance of the string is defined at every point on the stretched string and the disturbance will have a dynamic, a rule for its evolution.

Let me review that physics of the string between fixed walls. The electromagnetic field has many of the same properties. It is just a more complex field and the complications do not add any to the understanding of the quantum properties of the field. The stretched string is a one dimensional field where that field variable, $y(x, t)$, is the transverse displacement of the string from its equilibrium position and x is the distance along the interval between the walls. The electromagnetic field is three dimensional.

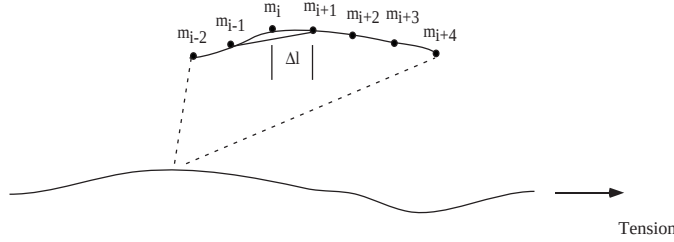


Figure 8.3: **A stretched string** The stretched string can be envisioned as a sequence of masses separated by a distance Δl which are connected by elastic elements which provide verticle force on the mass elements.

The dynamics of the string are well understood. The rule is very simple. The net vertical force on a piece of string of length Δl which equals the mass of that length of string times the acceleration of the transverse displacement is proportional to the negative of the displacement from the average of the displacements of its neighbors. The proportionality constant has the dimensions of a force per unit length and is thus the tension in the string divided by the length of the piece of string. ρ is the mass per unit length of the string.

$$\rho \Delta l a_{x,t} = -\frac{T}{\Delta l} \left[y(x, t) - \frac{(y(x + \frac{\Delta l}{2}, t) + y(x - \frac{\Delta l}{2}, t))}{2} \right] \quad (8.13)$$

You can also derive this result by cutting the string and seeing how the tension acts to stretch the string.

In the limit that Δl is zero this goes to

$$\rho \frac{\partial^2 y}{\partial t^2}(x, t) = T \frac{\partial^2 y}{\partial x^2}(x, t) \quad (8.14)$$

Note that $\frac{T}{\rho} \stackrel{\text{dim}}{=} \frac{\frac{ML}{T^2}}{\frac{M}{L}} \stackrel{\text{dim}}{=} \frac{L^2}{T^2}$ has the dimensions of a velocity squared. From this we can intuit that the disturbances in the field travel with a velocity $v = \pm \sqrt{\frac{T}{\rho}}$.

Any configuration of the displacements of a stretched string is a superposition of normal modes. When you pluck a stretched string you generally put in a localized disturbance. This excites all the modes and the higher frequency modes will damp out quickly and you are left with the fundamental.

Using the normal modes the stretched string can be considered a countable infinity of oscillators.

The quantum particle that is at its basis is called the phonon.

The photon is a state of the electromagnetic field that has a definite frequency, $\hbar\omega$. This implies that the field configuration is a normal mode. In other words, there is a photon for each of the normal modes. To understand the implications of this statement consider the stretched string.

8.3.2 Concluding Remarks

At the end of the 19th century, we had a unified physics using these action principles for particles and fields and their interactions. The name of the game was to write down the Lagrangian for the particle motions and the fields. Do the least action machinery and you knew all the conserved quantities and what was happening. There were only two fundamental forces, electromagnetism and gravitation. Both were well described by action principles, one a field theory and the other an action at a distance theory. All higher order phenomena were felt to be described by these fundamental entities. There was a feeling expressed by some that we may be near to the end of physics. This was clearly naive. Even on the face of it, there were clear problems that would require new insights. Why was the basis of physics built on such different mechanisms— field theory and action at a distance? What was the underlying machinery that could unify this physics? Despite the theoretical questions, the real basis for discovering a new physics would be the new experimental developments that took place at the turn of the century.

Chapter 9

The Special Theory of Relativity

9.1 Pre-History of concepts about light

It is interesting to note that so much of our understanding of the physical universe is based on our interpretations of the operation of vision and light and how dramatically this has changed over the centuries. The very earliest descriptions were usually attempts to understand the process of vision. As is so often the case, these early attempts to produce a theory of vision goes back to ancient Greece and is based on a simple idea of the extension of our sense of touch. We feel the location and texture of surfaces by contact. The corresponding idea of Empedocles and Euclid was that vision involved the emanation from the eye of rays that sensed the surface and returned to the eye, much like fingers. This simple picture is still with us in the form of the special vision of comic book heroes like Superman and in expressions such as “stop staring at me,” which implies the something is coming from the eye. It was the philosophical school based on atomism that lead up to Aristotle that first clearly established the vision is based on the emanations from the seen object basically by noting that there was no vision in the dark. It was also now possible to join the ideas of vision with the more general issue of light. The greek development reached a pinnacle in the ability of Ptolemy to describe reflection and measure refraction. []

After the fall of the greek nation states and during the dark ages in the west, arabian scholars not only rescued the greek texts but they continued the development of the ray theory of light. Alkindi and Alhazen bringing together the greek ideas and extending them to lenses and mirrors and Alhazen

producing what was to become the classic text on optics, *Kitah al-manazir* or *The Book of Optics*. An excellent review of the ancient contributions to optics and a layman's review of current ideas is given in the book by David Parks, [Park 1997].

With the renaissance, the primary issue became the nature of the emissions and, following Galileo, a much more refined effort to carefully measure the properties of light. Descartes filled all space with a particulate essence that was the basis for subsequent particle theories of light including Newton's. A wealth of experiments by Boyle, Hooke, and Young revealed the important properties of interference and diffraction and lead to the ideas of a particulate basis being displaced by the wave theory that originated with Huygens but reached its complete expression with Fresnel. In hind sight, it is interesting that phenomena associated with the polarization of light was the major difficulty in the acceptance of the wave theory. The development of the wave theory is very well articulated in [Buchwald 1989].

In the classical, pre-quantum, period, the next great contribution to our understanding of optical phenomena came as an addendum to Maxwell's effort to unify the electric and magnetic force systems. His development of a field theory of fundamental forces and the identification of light as the long range traveling solutions of his dynamical equations for the electromagnetic forces provided a new foundation for understanding all the phenomena associated with light. It was the anomalies associated with this dynamic and the requirement of Galilean relativity that Poincaré, Lorentz, and Einstein used to discover the basis for the special theory of relativity. A short history of these developments is given in [Born & Wolf 1999]. In Volume II of his history of the development of theories of the electric force, Whittaker provides a detailed and somewhat unique perspective of the development of special relativity, [Whittaker 1953] A more conventional history is given by Pais, [Pais 1982].

Although we are not concerned here except incidentally with the modern theory of light as expressed by quantum field theory, any complete account of our understanding of light must include the work associated with Planck and Einstein and later developed by Feynman, Schwinger, and Tomonaga, [Feynman 1985]. A detailed development of these ideas are in the later chapters of these notes.

9.2 Galilean Invariance

Almost anyone who has sat quietly waiting to depart from a bus depot or a dock and has had the bus or boat gently start to leave has had the experience of feeling that it is the depot or dock that has moved away. This simple physiological phenomena has its basis in a very general physical law that was first articulated by Galileo and thus is called Galilean Invariance. It is one of the most striking and far reaching of all of the laws of physics. It is impossible to over emphasize its importance; it is the basis of our understanding of space-time and motion. The simplest statement of the law is that there is no experiment that can be performed that can measure a uniform velocity. Since we can only know what can be measured, we can never know how fast we are moving. There is no speedometer on the starship Enterprise.

Stated this boldly, the idea is very counter to our experience. This is because what we generally observe as a velocity is not a velocity in space but is our velocity relative to the earth. Relative velocities are detectable. We note the amount of street that passes below our car or feel the flow of the air that moves over our face and infer a speed but we do not know how fast the earth is moving and thus do not know what our absolute velocity is. We do know that the earth moves around the sun and thus can determine our velocity relative to the sun. We know that the sun is moving in our galaxy and even that our galaxy is moving relative to other nearby galaxies and thus can know our velocity relative to the local cluster of galaxies. With the recent advances in astronomical detection, we are able to note our velocity relative to the place that we occupied in the early universe, our motion relative a background microwave radiation that is a detectable relic of the early universe, but again we cannot know whether that place had a velocity.

The inability to detect velocity is one of the most mysterious and counter intuitive concept that has ever been articulated. Consider a remote and empty part of the universe, no stars or galaxies nearby. Here there are no discernible forces and a released body moves in a straight line with a constant velocity. This is one of Newton's Laws and was his articulation of Galilean Invariance. Although when we start to work on General Relativity, we will have to revisit these issues, let us assume that this empty region is space. We envision this as that stable structure that Descartes and Newton needed as a background against which motion took place. In this day and age, it is generally easy to convince someone that this space obeys the Copernican Principle; it is not centered on some special place like the earth. It is also not difficult to convince someone that this idea should be extended to the

general Copernican Principle that, in an empty universe, there is no special place that could be called the center. This idea that there is a thing called space and that it is a stable structure and has no special places in it is better stated as the fact the the universe is homogeneous. Stated in a fashion that is similar to our statement of Galilean Invariance above, we can say that there is no experiment that can be performed in space that can distinguish one place from another. This is the definition of a homogeneous space. It should be obvious that if you cannot distinguish between places that you cannot have a center or a boundary. These are special places and this is contrary to the idea that all places are the same.

It might seem that these assumptions about the nature of space are so obvious that the universe must obey them. That is never the case in physics. You must test any hypothesis. On the other hand, you may want to say that this in not a hypothesis that is testable; we cannot be anywhere other than where we are. The best test of this idea is that we find that the laws of physics as we know them here on earth are found to be applicable everywhere that we apply them including distant space. Stars in remote galaxies operate in the same fashion as nearby stars. The laws of optics and electromagnetism are the same. We can also look at distributions of matter such as galaxies. Again, there is no indication that the universe is not homogeneous. A related concept is isotropy. This is the idea that space is the same in all directions. This hypothesis has been tested very precisely by the distribution of the microwave radiation that we observe.

Now consider two sets of physicists that are moving toward each other at some velocity, $\vec{v}$, and are studying the universe. If we now impose Galilean Invariance, each must have the same rules of physics and, thus, observe a universe that is homogeneous and isotropic. Yet, they are moving toward each other. It is not intuitive that space and time can be constructed consistently in this way but they are.

In other words, if we define the x direction as the line connecting the two sets of physicists, they will each measure events using space and time coordinates let's say (x, y, z, t) and (x', y', z', t') that satisfy the following relationships:

$$\begin{aligned}x' &= x - v_0 t \\y' &= y \\z' &= z \\t' &= t\end{aligned}\tag{9.1}$$

These are the Galilean transformations; the rules that indicate how to trans-

late one of the set of observer's observations to the other set of observer's observations. Each set of physicists, one set making measurements with (x, y, z) and t and the other with (x', y', z') and t' and each concluding that the universe is homogeneous and isotropic. Not only that but there is no experiment that they can perform that can yield a different result. If the physicists that differed in their measurements of events in space and time as given in equations 9.1 on page 216, found different rules for their experiments, we could tell them apart. If only one of them had Newton's Laws for the motion and the other did not, we would say that that one is at rest and the other one was moving. But since both sets of physicists have the same rules of physics and observe the same universe how can you tell which one is moving and which is at rest. **In summary: there is no experiment that can perform that can determine your velocity – all the laws of physics must be unaffected by a velocity choice.** All these relatively moving sets of observers have the same laws of physics as long as their velocity is unchanging.

By the way, it should be clear that although the different sets of observers must have the same results for any experiment, they will each describe the other's experiment differently. If one set of observers release a piece of chalk at rest relative to them, they will say that it remains at rest. Its coordinate is some (x_0, y_0, z_0) which is unchanging and its velocity is $\vec{v} = (0, 0, 0)$. The observers that are moving relative to this first set with a velocity $\vec{v} = (v_0, 0, 0)$, again choosing the x direction as the direction of relative motion, will say that the released chalk is moving uniformly in the direction of decreasing x but staying in the same place in the yz plane.

Said another way, Galilean invariance does not require that the different observers measure the same values for the things which they observe. Contrary, for the same experiment to produce the same result requires that the descriptors be different. For instance, the two observers give different descriptions of where an object is. An object that is at rest at the origin as measured by one observer will be seen as moving by the other observer. For one observer it will have non-zero kinetic energy and for the other it will have zero kinetic energy. Although places, velocities, kinetic energies are different, if the two observers do the same experiment, the same thing happens.

Consider two observers on the surface of the earth. This is not empty space and the local universe is not homogeneous and isotropic; you can tell up and down from sideways – things fall down because of gravity. But all the laws of physics must obey Galilean invariance including gravity. Have these observers move by each other at a uniform relative speed v in a horizontal

direction; both observers place a chalk on the end of their nose and release it. It falls and lands between their feet. The same experiment yields the same result. To each observer, the chalk falls along the line from his nose to his foot. Either observer when describing the others experiment sees the chalk with an initial velocity but cleverly arranged so that as it moves so does that other observers foot so that as the chalk drops past the observers foot, the foot is there also.

As stated before the principle of Galilean invariance is to state that there is no experiment that can be performed that can determine our velocity. This is true even in the presence of other fundamental forces such as gravity.

Acceleration, on the other hand, is detectable. Again consider our two observers on the surface of the earth with relative motion in the horizontal direction except in this case let there be an acceleration for one of the observers. If they drop chalk from their nose accelerated observer drops chalk, it does not land between his feet.

This is connected with the usual statement of Newton's force law. We say that force which is a push or a pull from some external agent acts to produce an acceleration according to $\vec{f} = m\vec{a}$. If there is no force the object does not change its velocity; it stays at rest to some set of equivalent observers. Not only does Galilean invariance effect the force free case, it is also operative when forces are present. In order to guarantee that all experiments have the same result, you have to make sure that $\vec{f}$ does not change when you make the connections to the other relatively moving observers, i. e. make the Galilean transformation. It is only in this case that you can have Galilean invariance. For instance, if we are talking about the Hook's law spring, $\vec{f} = -k\vec{x}$, the force changes under the transformation of equation 9.1 on page 216. But this system does not represent a homogeneous space; There is a special place, $x = 0$. The spring is attached to a fixed point. In the real world the spring is attached to another mass, m_2 . In this case the law of force on the original mass is $m_1\vec{a}_1 = -k(\vec{x}_1 - \vec{x}_2)$. Now if you apply the transformation, $x'_i = x_i + v_0t$, for $i = 1, 2$, there is no change in $\vec{a}$ and $\vec{x}_1 - \vec{x}_2$ and thus you have Galilean invariance. As you would expect the analysis of this situation using action is even more informative, see Appendix ?? on page ?? . See Section 7.4.2 on page 197. For simplicity of notation, consider a world with only one spatial dimension. the action for this case is

$$S(x_f, t_f, x_0, t_0; path) = \sum_{path, (x_0, t_0)}^{(x_f, t_f)} \left(m_1 \frac{v_1^2}{2} + m_2 \frac{v_2^2}{2} + k \frac{(x_2 - x_1)^2}{2} \right) \Delta t \quad (9.2)$$

Now applying our transformation, we get a change but it is only from the velocity terms and is thus the same case as for the free particle. In Section 7.4.2 on page 197, this case is analyzed in detail and it is seen that this family of transformations is an invariance.

In contrast to our inability to perform an experiment that can determine our velocity, it is easy to determine our acceleration. Consider a spring with a mass on the end. If we are accelerating, the stretch of the spring in equilibrium is different if we hold the spring along the acceleration direction or transverse. If we were in outer space at a distance from a massive body and held a plumb bob on the end of a string, the string would point to the massive body if we were not accelerating and would point to the side if we were accelerating. In the action analysis above, applying the transformation $x'_i = x_i + \frac{a}{2}t^2$ does not change the interaction term but does change the velocity parts and in a non-trivial way which means that there is no symmetry nor invariance. This is also consistent with the fact that even for free particles, accelerations are detectable.

9.3 Implications of and for Maxwell's Equations

All of the experiments involving electromagnetic phenomena up to the discoveries leading to quantum mechanics are described by the following local field theory and the associated force law:

$$\begin{aligned}
 \text{div}(\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)) &= \frac{1}{\epsilon_0} \rho(\vec{\mathbf{r}}, t) \\
 \vec{\text{curl}}(\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)) &= \frac{\Delta \vec{\mathbf{B}}(\vec{\mathbf{r}}, t)}{\Delta t} \\
 \text{div}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) &= 0 \\
 \vec{\text{curl}}(\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)) &= \frac{1}{\mu_0} \vec{\mathbf{j}}(\vec{\mathbf{r}}, t) - \frac{1}{\mu_0 \epsilon_0} \frac{\Delta \vec{\mathbf{E}}(\vec{\mathbf{r}}, t)}{\Delta t}
 \end{aligned} \tag{9.3}$$

and the force law:

$$\vec{\mathbf{F}} = q\vec{\mathbf{E}} + q\vec{\mathbf{v}} \times \vec{\mathbf{B}} \tag{9.4}$$

where $\rho(\vec{\mathbf{r}}, t)$ is the charge per unit volume, $\vec{\mathbf{j}}(\vec{\mathbf{r}}, t)$ is the current density or charge per unit area per unit time, $\vec{\mathbf{E}}(\vec{\mathbf{r}}, t)$ is the electric field or force per unit charge, and $\vec{\mathbf{B}}(\vec{\mathbf{r}}, t)$ is the magnetic field or force per unit charge times speed. The force, $\vec{\mathbf{F}}$ is the force on a charged particle with charge q and velocity $\vec{\mathbf{v}}$.

This system of equations, describes the electric and magnetic force system as a local field theory. Local field theory in contrast to the action at

a distance theories of the 18th and 19th century has become the vehicle of choice for the description of fundamental phenomena. The basic ideas and the procedures associated with field theory approaches are introduced and reviewed in Chapter 5 on page 123.

Like any system of forces, this set of rules articulated by Maxwell's equations, Equation 9.3 on page 219, must obey Galilean invariance or we would be able to use electromagnetic phenomena to determine a velocity in space. For instance, if you do a careful analysis of the dimensional content of these equations you will find that ϵ_0 and μ_0 have dimensions and that the combination $\frac{1}{\sqrt{\mu_0\epsilon_0}}$ has the dimensions of a speed. In fact, this speed is the characteristic speed of travel for changes in the fields and this is the speed at which disturbances of the field, light, travels. If Maxwell's equations and the associated force law are correct in all frames, the two fundamental dimensional constants must be the same and, thus, the speed of changes in the electromagnetic field must be the same to all observers.

This situation with Maxwell's equations presented quite a quandary to 19th Century physicists. Since Maxwell's equations are not Galilean invariant in the sense that they are left unchanged by the transformation law of equations 9.1 on page 216, then velocity could be measured and light could be used to do it. In other words, there was some preferred state of uniform motion in which the Maxwell's equations were true as written and in this frame the measured speed of light was $\frac{1}{\sqrt{\mu_0\epsilon_0}}$. This is analogous to the case of the stretched string in which the rest frame of the string is the preferred state in which the dynamics takes on a simple form and the speed of the waves was set simply by the parameters of the dynamic. For the case of the Maxwell system, an observer moving at any velocity with respect to the frame with the simple dynamic would not measure the same speed for light and would also have to modify equations 9.3 on page 219 and 9.4 to account for the relative velocity and in that system the equations would contain the relative velocity as additional parameters. It was still a quandary though in that all other fundamental dynamical systems were Galilean invariant but not electromagnetic phenomena. In fact, we will see that it was Einstein's genius to go the other way and insist that there were no experiments that could determine a velocity but that the simple transformation law, equation 9.1 on page 216, had to be modified and that Maxwell's equations 9.3 on page 219 and the force law 9.4 on page 219 were correct.

An interesting feature of Maxwell system and the force law is that, from the way that it operates, the magnetic force only changes the direction of a particle. It cannot do any work. From the work energy theorem it follows

that it cannot change the kinetic energy of a particle that is subject to only magnetic forces. This is a paradox. We get all of our electrical power from dynamo that are operated by magnetism. Let's take a closer look at this problem:

Side issue on Gleeson's magnetic paddle

Consider an electron and a large massive magnet. Shoot the electron into the magnet at some speed v . It is deflected and comes out at the same speed that it went in at. This is very satisfying since the kinetic energy before and after is the same.

Now consider the situation in which the electron is initially at rest and the magnet is moving at the speed v toward the electron. Initially the electron has zero kinetic energy. After it encounters the magnet, the electron is moving away from the magnet at the speed $2v$. This is how any massive paddle works. If you hit a light particle with a massive elastic paddle the light object is moving forward with speed $2v$.

The striking thing about the magnetic paddle is that like any paddle, the light particle goes from having no kinetic energy originally to one that has kinetic energy. But magnetic fields do not do any work?

If we analyze the situation in the frame of the moving magnet we see immediately the resolution for this seeming paradox. In this frame there is not only a magnetic field but also an electric field. In fact, the electric field, $\vec{E}$, is perpendicular to $\vec{B}$ and is directed along the sideways displacement that the charged particle experiences. In order to increase the kinetic energy of the charged particle to $2mv^2$ the electric field had to be $E = vB$. The important point is that you can use this paddle to convince yourself that under the Galilean transformation you not only change the coordinates but also have to change the fields $\vec{E}$ and $\vec{B}$. If they are to recover the same laws of physics, what one inertial observers says is a magnetic field will be viewed as being both a magnetic and electric field to another inertial observer.

This example shows that our ideas about what changes are necessary to have invariance between moving observers will have to go beyond just coordinate changes: it must deal with rearrangements of the elements of coupled systems.

Return to Maxwell's Equations again

In a similar fashion to the case of the stretched string, see Appendix 5.3 on page 128, Maxwell's equations predict that in a source free region there are

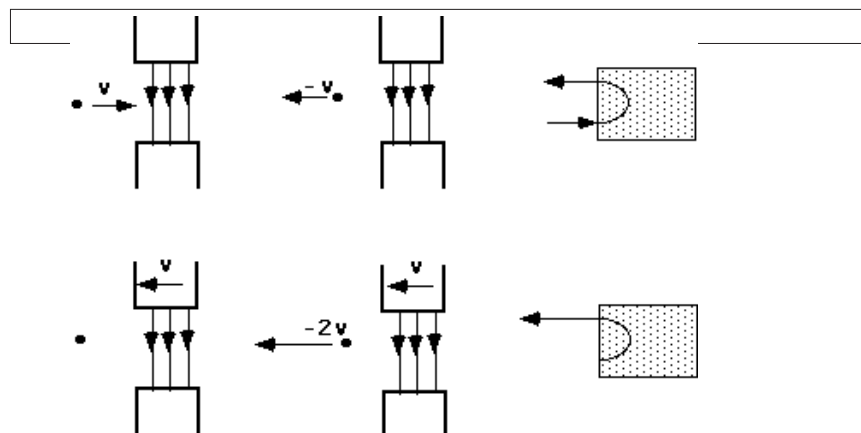


Figure 9.1: **The Magnetic Paddle:** In the upper part of the figure, a small charged particle represented by the dot is moving with speed v into a large magnet. It is deflected and comes out at the same speed v with which it entered. Now consider the same situation viewed from the frame in which the charged particle is initially at rest. Here the magnet is moving with speed v . After the magnet has passed over the original position of the charged particle, the particle is moving to the left at speed $2v$.

wavelike disturbances and the speed of these disturbances is

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0}}. \quad (9.5)$$

If these equations have to be modified to account for relative motion to the special frame in which they are true, then there should be many ways to observe these effects and measure our velocity relative to the special frame.

Actually, this is not the case. The speed of light is very large compared to speeds of terrestrial relative motion. This means that it is generally difficult to detect the small corrections caused by the relative motion. Several clever experiments were undertaken to detect motion relative to the preferred frame. These are discussed in Section 9.4 on page 223. None of these were able to detect the effects that were expected and this series of experiments were called the search for the aether that was supposed to be the underlying machinery of the electromagnetic field. This frustrating effort reached its culmination in the definitive series of special experiments carried out by Michelson and Morley in the later part of the 19th, see [Whittaker 1953].

9.4 Pursuit of a special frame

The Maxwell's equations, Equations 9.3 on page 219, implied that there were field configurations of the electromagnetic fields that traveled. In all previous examples of these traveling solutions in continuous systems the speed of travel was measured relative to that of the medium. For example, the speed of the travelers in the stretched rope, $v = \pm\sqrt{\frac{T}{\rho}}$, was relative to the rest frame of the string. In fact for Maxwell to understand the dynamics of electromagnetic system, he had to invoke the underlying structure of vortices, Section 5.4 on page 140, and the velocity of the disturbances was relative to the rest frame of these vortices.

9.5 Michelson-Morley Experiment

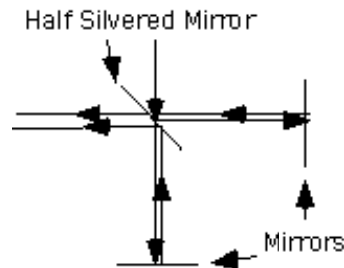


Figure 9.2: **Schematic Diagram of the Michelson-Morley Experiment:** Light enters the apparatus from above. It encounters a half silvered mirror so that one beam travels down to a reflecting mirror and returns to the half silvered mirror, reflects and leaves the apparatus to the left. The other beam from the half silvered mirror reflects to the right to a mirror and returns to the half silvered mirror to recombine with the first beam exiting to the left.

This experiment by the famous American physicist, Albert Michelson, was a search for the preferred frame for Maxwell's equations. It ultimately provided the experimental verification that there was no preferred frame and that the speed of light was the same for all relatively moving observers. It is important to point out that although this experiment is a direct verification of Einstein's postulates that are at the basis of the Special Theory of Relativity, Einstein was not aware of the experiment at the time he proposed

the theory. He based his argument on the nature of Maxwell's equations and their implications.

The fundamental idea is to try to detect an effect of the motion on the observed speed of light. It would be easy to just measure the speed of light for different states of motion and compare them. This is not possible because the speeds at which we can move an apparatus to measure the speed of light is generally negligible or well within the experimental error compared to the measurement of the speed of light itself. Michelson came up with a clever idea that would have allowed him to detect the small, actually large by most measures, speeds of celestial motion on the speed of light if they were there.

The basic idea is to compare the speed of light in two perpendicular directions at the same time, see Figure 9.2 on page 223. Since relative velocity is a vector or directed quantity, it will effect the speed of light differently in two different directions. This gets around the problem of making a direct comparison of the relative velocity to the speed of light.

To understand the experiment, let's look at a simple situation that is easier to understand.

The swimmer analogy

If a swimmer can swim at a speed v in still water and she wants to swim directly across a stream of width, D , that flows at a speed v_0 as shown in Figure 9.3 on page 224, she has to swim so that the resultant velocity, the vector sum of her velocity in the water and the velocity of the water, is directed across the stream.

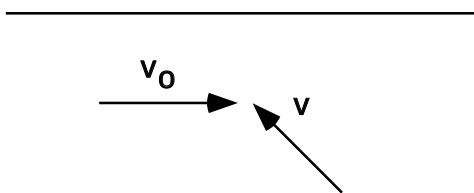


Figure 9.3: **Problem of a Swimmer in Flowing Stream:** A stream of width D is flowing with speed v_0 from left to right. A swimmer whose speed in still water is v wants to swim across and back, reaching the other bank at a point opposite the starting point.

The resultant velocity which is directed across the creek is thus $\sqrt{v^2 - v_0^2}$. If she wants to swim back again the total time is $2 \frac{D}{\sqrt{v^2 - v_0^2}}$.

If she wants to swim up the creek, with the current, a distance D and back the time is $\frac{D}{v+v_0} + \frac{D}{v-v_0}$. These two round trips cover the same distance

but the times are not the same. The difference of the round trip times is

$$\begin{aligned}\Delta t &= 2\frac{D}{v}\left(\frac{1}{1-\frac{v_0^2}{v^2}} - \frac{1}{\sqrt{1-\frac{v_0^2}{v^2}}}\right) \\ &\approx \frac{D}{v}\frac{v_0^2}{v^2},\end{aligned}\tag{9.6}$$

where I have used the relationship $(1+x)^n \approx 1+nx$ for $x \ll 1$. In the case of a swimmer, the ratio of the speeds, $\frac{v_0}{v}$ is a number a little less than one. Thus this difference in times is easily measured. It is also a fact that we could measure the speeds directly and just from the time swimming a distance D determine the drift of the stream.

Return to Michelson-Morley

For light if we say that there is no Galilean invariance and there is a special frame in which the speed of light is $\frac{1}{\sqrt{\mu_0\epsilon_0}}$. Then if we move relative to that frame, we should detect an effect on the speed of light to us. This is like the stream above. The speed of the swimmer is the speed of light in its preferred frame and the drift of the current is speed that we are moving relative to that preferred frame. The cleverness of the Michelson-Morley experiment is that it takes advantage of a special property of light to make it easy to measure the time differences for light traveling over different paths.

In the Michelson-Morley experiment, see Figure 9.2 on page 223, light enters the apparatus from above and is split into two beams by a half silvered mirror. One beam travels horizontally to a mirror and returns to the half silvered mirror and the other continues down to another mirror and returns to the half silvered mirror. The half silvered mirror allows the horizontal beam through and deflects the vertical beam so that the two beams can be combined and focused into an eyepiece. Thus if the apparatus is drifting in space at a velocity $\vec{v}_0$ relative to the frame in which the speed of light is c and also if the velocity is horizontal, we have the same circumstance as the swimmer. The net speed of the light in the two legs of the apparatus will be different and there will be a difference in time of transit through the apparatus. By using monochromatic light and the fact that the light is periodic with a very high frequency, Michelson and Morley can compare the arrival times with great precision.

From the Fresnel construction, Section ?? on page ??, and similar to the construction used to describe the Young's double slit experiment, Section ?? on page ??, the amplitude for the light at the eyepiece is the sum of the amplitudes from each leg of the apparatus. The phasors for each of these amplitudes rotate at the same frequency as the frequency of the light but will

have a different phase depending on which leg the light traveled over. In fact, the two phasors will have a difference in phase angle that is the difference in the travel time divided by the time associated with the characteristic frequency for that light or $\phi = \frac{\Delta t}{T} = \Delta t \times f = \Delta t \times \frac{c}{\lambda}$, where f is the frequency and λ is the wavelength of the light. Using “Things that Everyone Should Know”, see Section 1.4.2 on page 16, you can see that a very small Δt will produce measurable phase differences.



Figure 9.4: **Fringes of Michelson Morley Apparatus:** The pattern of bright and dark lines that are seen when viewing through the eyepiece of the a Michelson interferometer. These patterns are called fringes and it was anticipated that the fringe pattern would shift as the Michelson–Morley apparatus was rotated.

Of course in the actual apparatus, it is impossible to make the two arms the same length to the necessary precision. This would require that they be equal to within a portion of the wavelength of the light used. But also, you should realize that over the width of the beam you cannot align the mirrors that precisely. What you actually get is a pattern of lines over the width of the beam. Bright lines where the phase difference is an even multiple of π and dark lines where the phase shifts are an odd multiple of π . The bright and dark pattern of lines is called fringes, see Figure 9.4 on page 226. If you now rotate the apparatus, the role of the velocity relative to the special frame will shift in the two arms of the apparatus and the slower leg will become the quicker leg and visa versa. The fringes will shift. Michelson and Morley could detect a fringe shift as small as $\frac{\pi}{4}$. Using the results of our swimmer analogy, if the apparatus has arms of length $D \approx 10$ meters, the light can be reflected several times in each leg, the ratio of the drift velocity to the speed of light that the apparatus can detect is the order of $\frac{v_0}{c} \approx \sqrt{\frac{\lambda \times \phi}{D}}$. Using $\phi = \frac{\pi}{4}$, this apparatus can detect a relative velocity that

is about 10^{-4} of the speed of light or $3 \times 10^4 \frac{\text{m}}{\text{sec}}$. This is still a very high velocity for the apparatus but fortunately it is about the speed of the earth in its orbit. Thus since the apparatus is in orbit on the earth, Michelson and Morley should have seen fringe shifts as they rotated their apparatus. Over long periods of time, there were effectively no fringe shifts. This experiment has been repeated many times and with even greater precision than that of Michelson and Morley and still no fringe shifts. This experiment is a direct test of the postulate that regardless of your state of motion, the speed of light is the same in all directions. This is essentially Einstein's postulate about the structure of space time that is the basis of the Special Theory of Relativity. In the following Chapter, we will develop the consequences of this postulate.

There were many other attempts to detect our motion relative to the special frame in which Maxwell's equations are correct. None of them were as definitive as the Michelson – Morley experiment and none of them have contradicted the postulates of the Special Theory of Relativity.

Chapter 10

Kinematics of Special Relativity

10.1 Special Relativity

10.1.1 Principles of Relativity

Einstein postulated that there was still Galilean invariance, i. e. all uniformly moving observers had the same laws of physics; there was still no way to determine a velocity. The thing that they also agreed upon included Maxwell's equations and thus the speed of light. The problem then becomes one of defining lengths and times so that this can be done. From Section 2.6.2 on page 50, we realize that, instead of an arbitrary distance between scratches on a bar being the standard, distance can be defined from a velocity and a time. Thus, if we have a time such as the period of light from a particular atom, we can define lengths from the speed of light. If Maxwell's Equations are to be valid in all frames the speed of light, c , must be a universal constant. We will examine this concept later. We can use this so that we no longer have a fundamental unit of length. Lengths follow from this velocity and a standard to time. In other words, we use a time and c as our fundamental units and c is defined in such a way that we recover the usual meter. This change in the definition of length manifests itself in a good table of physical values by having the speed of light given as

$$c = 2.99792458 \times 10^8 \frac{\text{m}}{\text{sec}} \quad (\text{exact}). \quad (10.1)$$

In other words, we can pick the value for c since it is the standard. It is chosen so that the distance that we called the meter is what it was before.

Said another way, the meter is the length of the path traveled by light in vacuum during the time interval of $\frac{1}{299,792,458}$ of a second.

Digression on Dimensions

In olden times, the basic measured quantities were a mass, a length and a time. The standards were arbitrary and chosen for convenience. We then chose to use standards that were stable, accessible, and easy to use: the kilogram, the meter, and the second. We realized though in Section ?? on page ?? that we could use any set of algebraically independent combination of the three fundamental dimensional entities such as an energy, velocity, and a momentum. Then, you may ask, what could be more accessible and stable than the fundamental dimensional constants? The problem is to choose. There are lots of constants in physics that have dimension and could be called fundamental. One obvious example is the mass of an elementary particle like the electron. In some sense that is what was done when we chose the mass of the nucleus of carbon 12. Modern physicists would not choose this as a standard because we feel that we will calculate it in some future Theory of Everything. In fact, the hope is that the future theory will contain only the constants c , the speed of light, $\hbar$, Planck's constant divided by 2π , and G , Newton's constant in the gravitational force. These form an independent set that contain a length, mass, and a time. As indicated above, we already use c . With the increase in the precision with which we can measure $\hbar$, it will not be long before we replace the standard of mass with a standard based on $\hbar$. This will still leave time as the remaining old fashioned standard. The current standard is based on the frequency of a specific emission of the light from the cesium atom. Time can be measured with great precision and reproducibility and this is not likely to change. This is in contrast to G , Newton's constant, which because the gravitational force is so weak is difficult to measure with any precision.

Prior to Einstein's development of the Special Theory of Relativity, we had as the basis for our understanding of space time that a body unaffected by other bodies has no acceleration. The first kinematic level of detection is acceleration or:

1. There is no experiment that can detect your velocity when you are in uniform state of motion.

It means that you can **not** talk about going at a certain speed. An observer is in a special state of motion in which there is no acceleration. It is also important to realize that you can know that you are acceler-

ating. Acceleration is a measured quantity and absolutely knowable. States of motion with no acceleration are called copacetic¹.

Another way to say this is that you are always at rest in your own rest frame. You are copacetic if not being forced to have changes in velocity. Copacetics are good observers. Non copacetics are not and their measurements of phenomena have to be qualified.

2. Length and time scales are absolute. This is the statement that there are many copacetic observers who have relative velocities but, regardless of their differences of motion, their clocks run the same and the definition of length is the same.

A direct result of these postulates is that the relationship for the coordinates for an events when observed by two uniformly moving observers, copacetics, with relative speed v along the x axis is Equations 9.1 on page 216.

With Einstein, by requiring that Maxwell's equations are the same to all copacetic observers, these postulates have to change. The new postulates are:

1. There is no experiment that can detect a uniform state of motion.

Galilean invariance is retained although the transformation rule, Equation 9.1 on page 216, will have to be changed.

2. The speed of light is a universal constant.

Although Einstein came to this conclusion from his work with Maxwell's equations, it is also a direct consequence of the Michelson Morley Experiment. The implications of this postulate are far reaching. Some are obvious. It implies that the speed of light is the same in all directions and it is the same value to all inertial observers with measuring instruments that are commensurable. Others more subtle.

Reversing our thinking. Since the way that light travels is determined from Maxwell's equations, we have to find the transformation law between inertial observers that will preserve Maxwell's equations. Another way to say this is that we know the correct transformations of space and time between inertial observers must be such that Maxwell's Equations are invariant. Actually, it is even more general than that. We will have a set of transformations that leave a certain velocity, the speed of light, invariant. This is the velocity that light travels at because Maxwell's equations do

¹copacetic is defined by my computer dictionary as in excellent order.

not have any additional dimensional fundamental constants other than the speed of light.

Later, Section 11.2.2 on page 254, we will develop the set of transformations that will yield the same speed for light for all observers. For two observers with a relative speed v and choosing the positive x axis along the direction of relative motion between the second and the first observer, this set of transformations is called the Lorentz transformations and is:

$$\begin{aligned}x' &= \gamma(x - \beta ct) \\y' &= y \\z' &= z \\ct' &= \gamma(ct - \beta x)\end{aligned}\tag{10.2}$$

where

$$\beta = \frac{v}{c}\tag{10.3}$$

and

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}\tag{10.4}$$

For $\frac{v}{c} \ll 1$ these reduce to the Galilean transformations. We will derive them later, Section 11.2.2 on page 254. For now, just realize that they exist.

10.2 Harry and Sally and Space Time Diagrams

10.2.1 Introduction

The idea will be to develop an understanding of the implications about the nature of space and time that are implied by our postulates about relativity. We will do this by looking at a simple case of two relatively moving observers, Harry and Sally, and their observations. At the same time we will develop a powerful graphical analysis that will allow us to understand different situations.

10.2.2 The Paradox of Harry and Sally

Harry and Sally are two inertial observers. Harry is moving toward Sally at a high rate of speed. He is equipped with a battery pack and plug that fits

an outlet Sally is wearing and is connected to a light bulb that she has on her head. When he passes her the circuit is complete and he lights her light bulb.

A while later she writes to him. She says that she liked it when he went by and often looks out at the outgoing sphere of light that they generated together and remembers him fondly. She wishes that he was with her again at the center of that sphere of outgoing light.

He writes back that yes it was nice when he passed her but he has to inform her that he is at the center of the outgoing sphere of light and not her.

The paradox of this situation is that Harry and Sally are both correct. They both measure the light as traveling at the same speed, c . The speed of light for both of Harry and Sally is the same in all directions and thus they both see themselves as always at the center of the outgoing sphere of light. Since once they have parted, they are at different places this is a paradox. The resolution of this paradox will be at the heart of understanding relativity. In the following section we will resolve this paradox.

10.3 The Relativity of Simultaneity

In order to better understand the what is going on with Harry and Sally, let's look at another but similar situation. Consider two inertial observers. One is on a train standing in the center of one of the cars and the other is on the platform. The train is moving relative to the platform. At the instant that the train and platform observers coincide, a small firecracker explodes at their common position. There are photocells at each end of the rail car. The light from the firecracker travels to the ends of the car and triggers the two photocells. The observer on the train says that the events of triggering the photocells happen at the same time; that observer says that they are simultaneous. See Fig 10.1 on page 234. The observer on the platform, on the other hand, says that the photocell in the back of the car fired before the photocell in the front of the car. See Fig 10.2 on page 235. To that observer the events of the arrival of the light at the photocell were not simultaneous but the arrival of the light on the back of the car preceded the one on the front.

In summary, because of the constancy of the speed of light, we must conclude that the two spatially separated events that are simultaneous to one observer will not be simultaneous to a relatively moving observer.

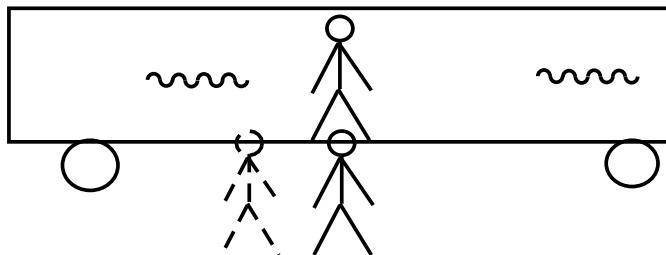


Figure 10.1: **Observer on a Moving Train:** In this case, the observer who is in the center of the car says that the light from the firecracker reaches the back of the car and the front of the car at the same time. The train is moving from left to right so we see the platform observer to the left of the original position, shown dashed, at a later time.

10.3.1 Harry and Sally's Movements in a Diagram

To understand what is going on with Harry and Sally, we will analyze the situation graphically. For simplicity of analysis and presentation, we will work in only one space dimension. Later, when we derive the Lorentz transformations, three spatial dimensions will be used, see Section 11.2.2 on page 254.

If we assign a coordinate system to Sally, we obtain the following description of what is going on. First, let's clarify some notation. In an ordinary graph, for instance plotting the xy plane, the line labeled the x axis is really the set of places that have coordinate y take the value zero or, better said, the x axis is better thought of as the $y = 0$ line. Similarly, the y axis is better thought of as the $x = 0$ line.

In space-time, we will draw the time axis vertically and the position or x axis horizontally. Again, you should think of the time axis as the place that is $x = 0$ for all times and the x axis as the time $t = 0$ for all places.

If we draw what is happening in a system based on Sally's observations, see Figure 10.3 on page 236, we will place Sally's time axis, her $x_s = 0$ line, vertically. Her x axis, the $t_s = 0$ line, will be horizontal. Harry is going by her at a relative speed of v . Therefore, the set of events that is Harry is a line with slope $\frac{1}{v}$. Don't forget, we are drawing the time axis vertically and slope is rise divided by run. Now, this set of events is what Harry would call his $x_h = 0$ line. In other words, if we choose the event of their coincidence as the origin event, $(0, 0)$, the equation of Harry's time axis on Sally's coordinate system is

$$t = \frac{1}{v}x. \quad (10.5)$$

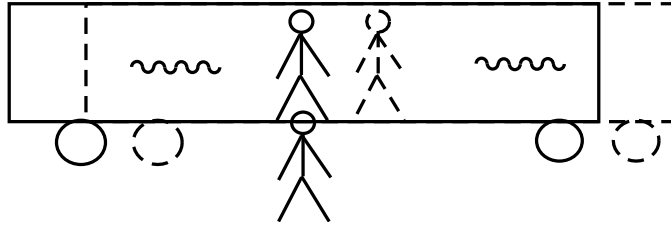


Figure 10.2: **Observer on the Platform:** In this case, at a later time, the observer who is on the platform sees the car move to the right. Since the speed of light is the same in the right and left directions, the light traveling toward the back of the car goes a shorter distance and, thus, arrives at the back of the car before the light that is sent to the front of the car. The events of the arrival of the light at the back and the front of the car are not simultaneous to the platform observer.

Of course, this is because we chose $t_s = 0$ as the time for the event when they were together. We choose this as $t_h = 0$ for Harry also. They both label the event of coincidence as $(0, 0)$. At $t_{s,h} = 0$, a light pulse emerges at $x_{s,h} = 0$ and moves away from both of them at the speed of light. On Sally's coordinate system, these events are two lines through the event $(0, 0)$ with slope $\pm \frac{1}{c}$. At some time later, $t_s = c_1$, Sally determines that she is at the center of the outgoing pulses of light and that Harry is not at the center, which is always at her place, $x_s = 0$, but instead he is at $x = v(t_s = c_1) > 0$.

We can just as well draw all of this from Harry's point of view, see Figure 10.4 on page 237. Harry is an inertial observer also. Now it is Harry's time axis that is vertical. Sally's time axis, $x = 0$ line, is now a straight line sloped at $-\frac{1}{v}$. We have picked the positive x direction to be the same for Harry and Sally. Thus, she is moving to negative position values in reference to Harry. Events at some time t to Harry are horizontal lines on this coordinate system and again at any time $t_h = C$ that Harry looks out he is at the center of the outgoing pulses of light and Sally is at the place labeled by $x = -v(t_h = C) < 0$.

Both Harry and Sally are inertial observers. There is no experimental way to distinguish them and, therefore, neither of them is to be preferred. How do we resolve this conflict?

Let's return to Sally's description of what is going on. From Section 10.3 on page 233, we realize that events that are simultaneous to Sally will not be simultaneous to Harry and visa versa. In order to understand the situation, we can endow Harry with two rods of equal length, one in front, leading, and

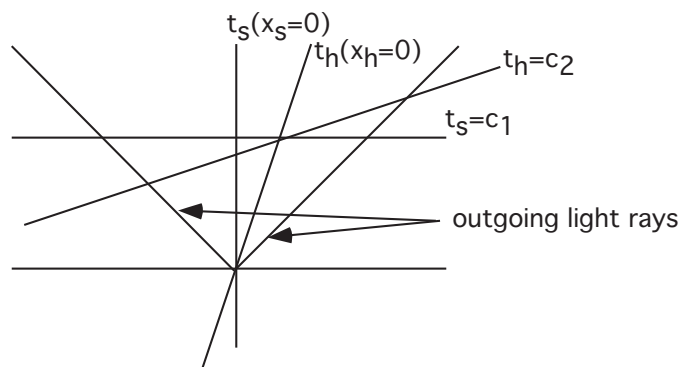


Figure 10.3: **Sally's Space-time Diagram:** Sally's space-time description of her meeting with Harry. Sally's time axis is vertical and her space axis is horizontal. Events at some time t according to Sally are horizontal lines such as $t_s = c_1$. Harry is the line $t = \frac{1}{v}x$. The events that are simultaneous to Harry are a line sloped at $\frac{v}{c^2}$ such as $t_h = c_2$. See Figure 10.5 on page 238 and the following text for details. The light rays generated at their meeting at the event $(0, 0)$ are the lines $t = \pm \frac{1}{c}$.

one in back, trailing. From the discussion of Section 10.3 on page 233, we can now find how events that are simultaneous to Harry appear on Sally's diagram. The ends of the rods are carried along with Harry and the events that are the ends of the rods have the equations $x = vt - L_0$ for the back and $x = vt + L_0$, where L_0 is a measure of the lengths of the rods. From the situation of the boxcar in Section 10.3 on page 233, we realize that the event that has the back rod coincident with the back going light ray and the event that has the front rod coincident with the forward traveling light ray are simultaneous to Harry. These lines will intersect the light lines at $(\frac{cL_0}{c-v}, \frac{L_0}{c-v})$ for the front rod and $(-\frac{cL_0}{c+v}, \frac{L_0}{c+v})$. The slope of the line connecting these two events is

$$\text{slope} = \frac{\frac{L_0}{c-v} - \frac{L_0}{c+v}}{\frac{cL_0}{c-v} + \frac{cL_0}{c+v}} = \frac{v}{c^2}. \quad (10.6)$$

It should be clear that, if Harry had been carrying a set of equal spaced confederates with synchronized clocks, the set of events that are the simultaneous reading at some time $t_h = C$ of these clocks will be a line with slope $\frac{v}{c^2}$. Realizing the lines of constant t to any observer are lines of simultaneity, we note that Harry's lines of $t_h = C$ appear on Sally's diagram as lines with slope $\frac{v}{c^2}$, see Figure 10.3 on page 236. Similarly, Sally's lines of simultane-

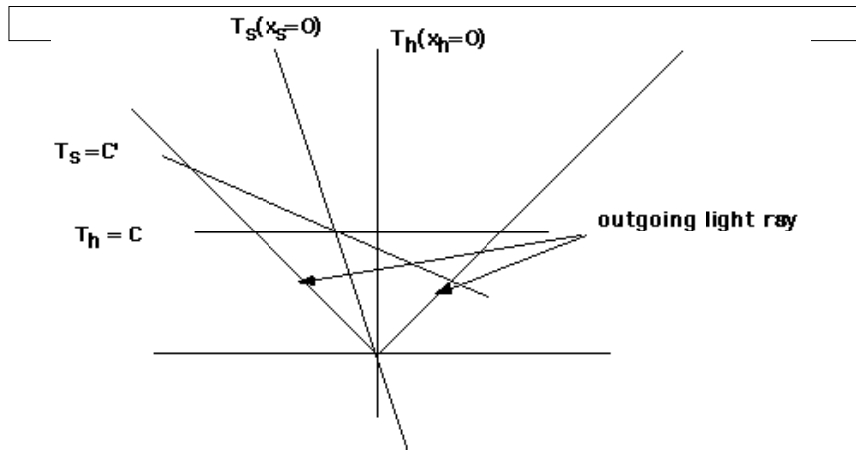


Figure 10.4: **Harry's space-time diagram:** Harry's space-time description of his meeting with Sally. In this case, Harry's time axis is vertical and Sally's is sloped $-\frac{1}{v}$. If at anytime, $t_h = C$, Harry describes the situation, he is at the center of the outgoing light pulses. She is always seen as being off center at some negative x .

ity, i. e. $t_s = c_1$, on Harry's diagram appear with slope $\frac{-v}{c^2}$ since she has a relative velocity of $-v$, see Figure 10.4 on page 236. In particular, the events on Sally's diagram that represent Harry's x_h axis, his $t_h = 0$ line, is a line passing through the event $(0,0)$ with slope $\frac{v}{c^2}$. Thus, we can now resolve the paradox of Harry and Sally. They are both right. They are both at the center of the outgoing sphere of light. They have different definitions of simultaneity, i. e. where the light is at some time t on their respective clocks. This is an important point and at the heart of many of the paradoxes associated with the Special Theory of Relativity. More importantly for our present needs, we see that we can construct a coordinate system for Harry on Sally's diagram. On Sally's diagram the coordinate axis for a Harry are no longer orthogonal.

The events that constitute where someone is at any time t are called the person's world line. This is what we called their trajectory in our earlier analysis of action, see Chapter ?? on page ??. For a uniformly moving observer like Harry, his world line is a straight line and is also his time axis. Since uniformly moving observers are inertial, we see that all inertial observers appear as straight lines. For non-inertial objects the world line is curved.

On Sally's coordinate system, Harry's space axis, his locus of events

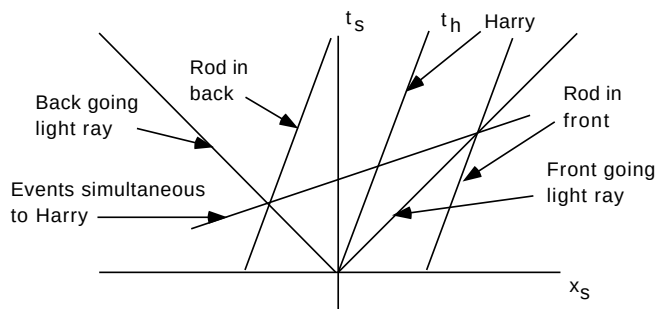


Figure 10.5: **Harry's Lines of Simultaneity:** The figure shows Harry's lines of simultaneity on Sally's diagram. On Harry's diagram these lines would be horizontal. To develop the lines of simultaneity, Harry carries equal length rods in front and in back of his position. In Sally's interpretation of this set up, Harry is at the center of an interval like the boxcar in Figure 10.2 on page 235. The event that is the coincidence of the forward going light ray and the front rod and the event that is the coincidence of the back going light ray and the back rod are not simultaneous to Sally but are simultaneous to Harry. The lines connecting these events are the lines of simultaneity to Harry and have a slope of $\frac{v}{c^2}$.

that are simultaneous with $t = 0$ to him, has slope $\frac{v}{c^2}$. This is also a general result. For any two relatively moving inertial observers, if one is chosen with the time axis vertical, the other observers lines of simultaneity will appear with slope $\frac{v}{c^2}$ where v is their relative velocity. In other words, the equation for Harry's x axis on Sally's coordinate system is

$$t = \frac{v}{c^2}x \quad (10.7)$$

From the above discussion, it should be clear that any event that will be labeled by a place and a time by Harry and Sally will have different labels for any particular event except the origin event, $(0,0)$, see Figure 10.6 on page 239. In fact as discussed in Section 10.1 on page 229, these labels for the same event are connected by the set of equations that are called the Lorentz transformations, see Equations 10.2 on page 232. If we choose the x axis along the same direction as the relative motion and if Harry carries an identical clock to Sally and has the same definition of length, these are

$$\begin{aligned} x_h &= \gamma(x_s - \beta ct_s) \\ y_h &= y_s \end{aligned}$$

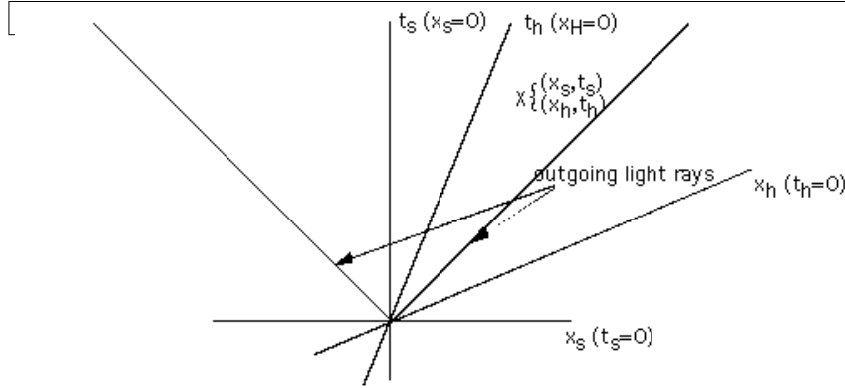


Figure 10.6: **Construction of coordinate axis for a relatively moving observer:** Harry and Sally have a relative velocity, v , with Harry moving to increasing x to Sally. They both agree to label the event of their coincidence as $(0, 0)$. His time axis, his $x_h = 0$ line, is a straight line through the origin with slope $\frac{1}{v}$ and his x axis, his $t_h = 0$ line, also passes through the origin but has slope $\frac{v}{c^2}$.

$$\begin{aligned} z_h &= z_s \\ ct_h &= \gamma(ct_s - \beta x_s) \end{aligned} \quad (10.8)$$

where $\gamma \equiv \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$ and $\beta \equiv \frac{v}{c}$.

In order to derive these equations, we will need to discuss more carefully this idea of identical clocks and the definition of length. We will do this in the next chapter. For now we can note several features of these equations. For example, if Harry carries an identical clock to Sally, then the events that are the ticks of his clock occur on his world line, his t axis or $x_h = 0$ line, at equal intervals, $t_h = n\Delta t_h$, but these equations will require that the intervals are spaced more than Sally's. This effect is called time dilation, see Section 11.3.1 on page 257. We can get the amount of the dilation from the Lorentz transformations. The coordinates of any one of these ticks according to Harry is $(0, n\Delta t_h)_h$, where n labels the tick. These same events are recorded by Sally as $(nv\Delta t_s, n\Delta t_s)_s$. Remember that all the events on Harry's time axis take the coordinate form $(vt, t)_s$ to Sally. Plugging this into the Lorentz transformations:

$$\begin{aligned} 0 &= \gamma(nv\Delta t_s - \beta cn\Delta t_s) \\ nc\Delta t_h &= \gamma(cn\Delta t_s - \beta nv\Delta t_s) \end{aligned}$$

which implies $nc\Delta t_h = \gamma(1 - \beta^2)(nc\Delta t_s) \implies c\Delta t_h = \frac{c\Delta t_s}{\gamma}$ or

$$\gamma\Delta t_h = \Delta t_s. \quad (10.9)$$

Since $\gamma < 1$, Sally says the Harry's clock runs slow compared to her clock. By the way because of the equivalence of inertial observers, Harry will also conclude that Sally's clock runs slow compared to his.

In addition, an identical length carried by Harry is shorter to Sally, see Section 11.3.2 on page 259. Here we measure the length by asking where the ends of the rods are at the same time. We will defer the derivation of the length contraction formula that section and only quote the result here. If Harry is carrying a rod of length L_0 , Sally will say that the length of the rod is

$$L_s = \gamma L_0. \quad (10.10)$$

All of these derivations require that we know the Lorentz transformations. Let's start over and carefully construct the coordinates and then derive the Lorentz transformations from our rules for constructing the coordinates.

Chapter 11

The Nature of Space-Time

11.1 The Problem of Coordinates

The basic problem of physics is to track in space and time the development of elements of a system. This requires that we have some method to communicate where and when something took place. In a three dimensional space the place is a set of three numbers; for instance, in a room you could use how far along the floor in a direction along one wall, how far along another wall, and how far up towards the ceiling. The time comes from a clock. This seems so obvious that we generally do not even think about it but, like all the things that we do, this is a subtle operation and we should understand what it is that we are doing when we make a coordinate system. In fact, the realization, that the establishment of the coordinate system is arbitrary is the key to understanding General Relativity. That will come later, Chapter 16 on page 345.

First, lets talk about places. The idea is to label the places. Think of a large parking lot, say at Disney Land. What you need is a unique label for every place. This could be done simply by going around and labeling spots on the lot with the name of a Disney character. This though is not an efficient way to label places. It is a unique label for each place which is how we started but there are many better ways to proceed. For one thing, this labeling scheme does not provide a guide for movement. If you are at Donald Duck, you do not know how far or in what direction to go to get to Goofy, the labels are not an ordered set. You could fix this by ordering the characters alphabetically. This system is nice in that it provides a guide to how to move, it does not indicate how far. It is also not extendable or divisible. An obvious solution is to use as labels the points on the real line,

create a mapping of the locations along a direction in the lot with the points of the real line. Since the real line is dense, you can always find a label for any place. If cars suddenly became smaller you would have no problem finding labels. You can also then use these labels to identify direction in the sense that from any location, increasing labels mark one direction along the lot and decreasing mark the other. In other words the sign of the difference between the labels is an indicator of direction of movement. This is a great improvement over the use of Donald Duck to label places.

There are still two problems. First, you need a distance. You can use the length that we discussed in Section 2.3.1 on page 40. In the present case, this means that we define length from how far light travels in a given time or, going back to old fashioned ideas, having some standard rod that can be placed between the points. In the simplest case, you just label the places and then come back later and measure their separation with your standard rod or whatever protocol that is defined for length. In this case the distances between places with the same label difference may have different distances. Don't forget, you just assigned labels from the real line to the places; you just labeled them. This problem though is easy to handle. You just have to measure the separations associated with the different neighboring places. In general, you will not know that all labeled places have the same separations. This process is called establishing a metric on the coordinate system. Our usual use of the cartesian coordinate assumes that when we measure the separations that they are the same in all places, i. e., the underlying manifold is assumed to be homogeneous. The separations are all independent of the labels. Sometimes and in many of the cases that follow this assumption is not warranted.

Secondly, what happens with the idea of extension. What happens when you add to the lot? You have to relabel everything. You can still cover the lot with labels but it is not convenient. By the way, this fact that you can cover a two dimensional space with a wrapped one dimensional label is also a simple proof of the size of the spaces are the same and thus that, although it might appear that a two dimensional infinite space seems bigger than an infinite one dimensional, there are as many points on the plane as there are on a line, see Section ?? on page ?? Thus since you want to extend in a direction that is not along the direction of the chosen sequence, you can improve things quite a bit by having two designators at each place and ordering each of the sets of designators so that a place is a doublet, i. e. (Goofy, Donald Duck).

If you are at the place labeled (3,1) and want to go to the place labeled (7,2) you only need to go four places in the first direction and one place in

the other, if you are at (9,0) and want to go to (7,0), you go 2 places in the backwards in the first direction. On the surface of the parking lot though, there are different ways to go between places. An obvious example is to go directly. This is because the two plane is more than two independent lines but accommodates all the paths in between.

In our parking lot, we need two measures of distance, one in each of the independent directions. If both directions are the same, we could generate a combined measure of distance, i.e. not require that all movement be along one of the coordinate directions. More than that if we assume that the space is the same in all directions at any point, isotropic, we can make a measure of distance that is independent of how we chose the directions of the coordinate system. In the case of the parking lot, if we assume that it is isotropic we can adopt for our distance measure $\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ where Δx is the displacement in the one direction and Δy is the displacement in the other direction. This distance has the advantage of being independent of the orientation of the axis system. I have to warn you that, if we were really worried about a parking lot, we would most likely not have an isotropic pattern of labels. Automobiles are longer than they are wide.

Again, if at each place the distance algorithm can be the same regardless of where you are the space is homogeneous. If the length scale is also isotropic, you really have the space as described by Descarte and the geometry will be that of Euclid. In general, it could be that, at different places, the distance between places is different or the length is different in different directions. Think about it. On the surface of the earth why should a rod that is held horizontal then turned vertical have the same length? This idea of the distance being the same at all places is also an important simplification, an essential symmetry. When you think about it though it may not be possible. The space may not be homogeneous. Each place may be special. Length may depend on where you are or your orientation. Different directions may have different length scales. In our considerations of General Relativity, Chapter 16 on page 345, these issues will become important.

What is it that we want to get out of this rather extended discussion of the process for labeling a place. The most important thing is the realization that in contrast to what was our original ideas about labeling places, there is a great deal of choice. The choice, as is often the case, is arbitrary and cannot influence important issues. Later, when we discuss the General Theory of Relativity, Chapter 16 on page 345, we will use this ambiguity as a part of the basis for understanding the theory. Suffice it to say, that we must develop a method for labeling places that must be consistent for all observers. It is the consistency requirement that allows us to derive the relationship

between the different observers labeling of places and times.

Let us now go into the standard construction of the coordinate system. There are two general methods: the use of confederates at each place and the single observer method. We will start with the confederate method and then show its equivalence to the single observer method which is the one that we will subsequently use.

We begin by defining the spatial coordinates. We assume that we can fill space with a confederate at every place and that the distance between the origin observer and each of the confederates remains fixed. I must warn you about the intrinsic anthropomorphism of this action. Please be assured that the use of words like “confederate” and “observer” which is common to this business imply a humanity that is not really intended. In actually, by confederate or observer, we mean a measuring system – a clock and recording devices – not necessarily a person. It may appear that this assumption about our ability to fill space with fixed and uniform confederates must be true. In fact, one of the insights from general relativity is that this is the case only in the absence of gravity. Since they are fixed in space, we will label the confederate by how far away he/she is in each of the three coordinate directions. Obviously, if the space is homogeneous and isotropic, the location of the origin and the directions of the coordinate axis are arbitrary. For the definition of the distance, we will use the length defined earlier, Section ?? on page ??, a defined speed of light and a time to label all distances. This speed will be universal for any observer establishing a coordinate system. This means that we need a standard clock and we choose the frequency of given emission line of a Cesium atom. In other words, our second is 9,192,631,770 oscillations of the light. To find the distance to any confederate, we send a light ray to that confederate who reflects it back and, with the standard clock, the observer at the origin can determine how far away that confederate is, $d = \frac{c\Delta t}{2}$, where Δt is the time interval for the round trip of the light.

We have not discussed the problem of labeling the time. The situation is similar to the problem of labeling places. We need some ordered system at each place. What order do a series of events occur in? By endowing each confederate with a clock, we will have at each place a reference set of events to compare with the events being labeled. We use our standard clock. We tell each confederate to make a standard clock. Since the space is assumed to be homogeneous, all the clocks must run at the same rate for each confederate. This is the first step in getting the time of an event that we want to label, to coordinatize. Since we have now endowed each confederate with a clock, we can use as the space and time label for any event as the time recorded on the nearest confederate’s clock and the location of the

nearest confederate. You should realize that it is not enough to use the same clock at each place but we have to deal with the problem of synchronizing the several clocks; the confederates must synchronize their clocks – at some time agreeing on the time. It must also be consistent with our understanding that the speed of light is the same in all directions regardless of the velocity of the observer. Of course, this leads to the problem of the relativity of simultaneity and makes it important that we understand the process by which any observer synchronizes clocks. For now since we are dealing with only one frame, we do not need to worry about the relativity of simultaneity but it will cause some concern when we compare the coordinate systems constructed by two relatively moving observers. This is discussed in the next section, Section 11.2.2 on page 254. For now, we can accomplish the synchronization by having a burst of light at some very early time released from the origin and, since we know the speed of light and that it is isotropic and we know the location of each confederate, we will know when it passes each confederate and they can set their clocks appropriately.

Let me summarize the confederate scheme for coordinatizing any event, see Figure 11.1 on page 246. An observer establishes a lattice of confederates with identical synchronized clocks and the label of any event in space-time, for that observer, is the reading of the clock and the location of the nearest confederate to that event.

There is a scheme that is equivalent to the confederate scheme that can be accomplished in a less elaborate way by the simple mechanism of having a single clock at the spatial origin and requiring that the observer continuously send out light rays in all directions keeping track of the time of emission. At any event, the incoming light ray is reflected back to the observer. Therefore, the observer has two times and a direction that are associated with any event: the time the reflected ray left and the time of return of the reflected ray and the direction of the reflected light. To yield a spatial coordinatizing that is consistent with the confederate scheme, the spatial distance to the event is the difference in the two times times c divided by 2 or

$$|\vec{x}| = \frac{c(t_2 - t_1)}{2} \quad (11.1)$$

where t_2 is the later time and t_1 is the earlier time. The distance is resolved along the coordinate directions according to the direction of the incoming light ray. To be consistent with the time labeling of the confederate scheme, the time coordinate is

$$t = \frac{t_2 + t_1}{2}. \quad (11.2)$$

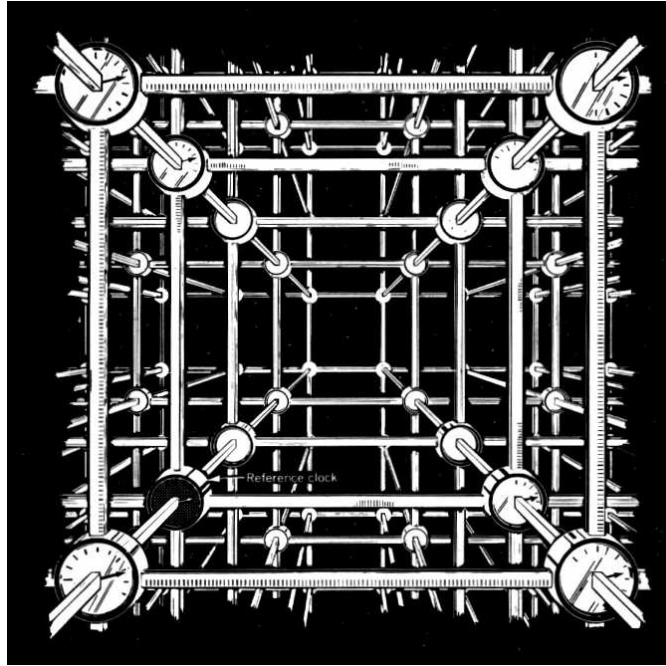


Figure 11.1: **General Construction of a Coordinate System:** Fill all of space with identical clocks. The location of each clock is given and all the clocks are synchronized. An event is given coordinates by assigning the position as the location of the nearest clock and the time on that clock when the event took place.

This protocol for coordinatizing is shown in Figure 11.2 on page 247.

If we accept this protocol for coordinatizing, in order to maintain the equivalence of the inertial observers, all observers should use identical clocks and this protocol. Our problem now becomes the problem of insuring that the clocks are identical and the comparison of results for different inertial observers. These comparative coordinates are related by the Lorentz transformations.

11.2 The Lorentz Transformations

Now that we have developed a protocol for coordinatizing events, we need to find the transformation rules that one inertial observer must use to compare observations with another moving at relative velocity, $\vec{v}$. Actually, this is a special case of the more general problem of finding the transformation rules

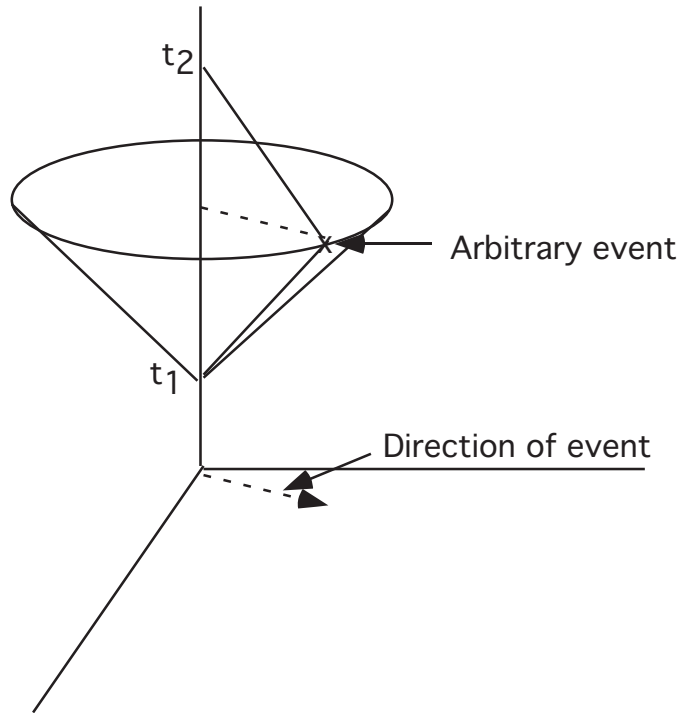


Figure 11.2: **Protocol for Coordinatizing an Event:** The distance of an event for an inertial observer with a clock is $\frac{c(t_2 - t_1)}{2}$ and the direction is along the direction of the return signal. The time coordinate is $\frac{t_2 + t_1}{2}$.

between any two coordinate systems. Since all inertial observers will see force-free motion as also inertial and as a straight line, you can convince yourself that the most general set of transformation rules between inertial observers is a set of transformations that is linear; it must transform straight lines, one inertial observer, into another straight line, the other inertial observer. Each observer sees his/her time axis as his/her $x = 0$ line.

Before dealing with the case of velocity differences, consider the particularly simple case of two coordinate systems that differ from each other only in the location of the origin. This is the case of two observers that have zero relative velocity, the same orientation of their coordinated axis and it is just the observer one says that observer two has her origin at the location (x_{20}, y_{20}, z_{20}) . An event measured at the coordinate, $(x, y, z, t)_1$. Will have the label $(x - x_{20}, y - y_{20}, z - z_{20}, t)_2$ to observer two or

$$x' = x - x_{20}$$

$$\begin{aligned}
y' &= y - y_{20} \\
z' &= z - z_{20} \\
t' &= t
\end{aligned} \tag{11.3}$$

This family of transformations has the general name of space translations and is labeled by the values (x_{20}, y_{20}, z_{20}) . It is an example of a linear transformation between the coordinates; the coordinates enter on both sides of the equations linearly. This example is also inhomogeneous. It has terms that are also independent of the coordinates. The Lorentz transformations that we will deal with here will be linear and homogeneous. Later we will add the inhomogeneous terms which will again deal with translations. The translation transformation were discussed extensively earlier in Section ?? on page ?. We could also develop the transformation rules for two observers that are at rest with respect to each other, share the same origin, but have different coordinate axis directions. These are the rotation transformations but like the translations these can be incorporated in the family of transformations that are developed here.

Our process for finding the Lorentz transformations will be to use specific rules for the establishment of a coordinate system, see Section 11.1 on page 245, and then to require that the same procedure be used in any inertial system. This process will lead to the fact that for two relatively moving systems, the same event will have two different coordinate designations. This should not come as a surprise since even prior to Einstein's Theory of Special Relativity, the Galilean transformation, see Equation 9.1 on page 216, gave different coordinates for an event when measured by two different inertial observers.

$$\begin{aligned}
x' &= x - v_{x_0}t \\
y' &= y - v_{y_0}t \\
z' &= z - v_{z_0}t \\
t' &= t,
\end{aligned} \tag{11.4}$$

where v_{x_0} , v_{y_0} , and v_{z_0} are the x , y , and z components of the relative velocity of the second observer as measured by the first observer. This family of transformations is labeled by these velocities. As a consequence of these coordinate transformations, the velocities of objects as measured by these observers are also transformed.

$$v'_x = v_x - v_{x_0}$$

$$\begin{aligned}v'_y &= v_y - v_{y0} \\v'_z &= v_z - v_{z0}\end{aligned}\tag{11.5}$$

These changes also imply that many significant dynamical variables such as momentum and energy are also transformed.

In the case of the Special Theory of Relativity, the rules connecting the different labels for a pair of relatively moving inertial observers that have the same origin and share coordinate axis directions are called the Lorentz transformations. We will derive them in this section. The full family of transformations that include the rotations, translations, and velocity transformations are called the Poincaré transformations.

To construct the Lorentz transformations, we will need to construct two independent inertial coordinate system. It should be clear that each inertial observer must have the same protocol for establishing their coordinate system, the same standard clock, and the same definition of the speed of light. In the previous discussion of Harry and Sally and the the story of the observers in the box car and on the station, Section 10.3 on page 233, we worked for simplicity with only one spatial dimension. Here we will treat the full complication of three space dimensions. Later in many applications, we will return to the case of one spatial dimension where the simplicity allows the point to be made more clearly. You should realize that the primary criteria of the extension to all three spatial dimensions will be that, to each inertial observer, the world should have the usually assumed symmetries of mirror symmetry, inversion symmetry in any direction, and isotropy, no preferred direction.

For definiteness, we will assume that there is an event at which the two relatively moving observers are at the same place and this event will be used as the origin of both coordinate systems. Since as mentioned in the beginning of this section, inertial observers are straight lines in space-time diagrams and thus in three spatial dimensions, only in special cases, will this coincidence occur. Regardless, If this were not the case, a simple spatial coordinate translation of one of the observers, see Equation 11.3, on page 248 will relocate the spatial origin.

We set both observer's clocks to $t = 0$ at this event. Since there are two straight lines that meet, there is a plane in space time. The two observers agree that their relative velocity, v , is in that plane and designate the spatial axis in that plane as the positive x axis of observer one. This is the one spatial dimension that will require special attention in the following. Whenever we deal with one space and one time direction, it will be this direction unless stated otherwise. This direction is called the longitudinal direction.

The second observer has chosen the same orientation for his/her x axis.

Firstly, note that the requirement for universal agreement among observers about the speed of light requires that that for both observers light advances by equal distances in equal times. We also use commensurate time and space units. If the spatial intervals are defined the times are in the time that it takes light to travel that distance and *visa versa*, if the time is the defining unit the distances are the distance that light travels in that time; an example is years and lightyears.

As we saw in Section 10.3 on page 233 the requirement that all observers measure the same speed of light, implies the relativity of simultaneity. Thus although there is agreement about the origin event, $(x = 0, t = 0)$, the locus of events which are straight lines with $t = 0$ to the different observers are different sets of events. This relativity of simultaneity is at the heart of the interpretational difficulties of special relativity.

Since the orientation of the two sets of spatial coordinates is the same, the second observer will say that the first observer has relative speed v directed toward the negative x axis or a velocity of $-v$. This is the direct consequence of the fact that both observers have front back symmetry in this direction and the same speed for light.

First, consider the nature of the agreements and disagreements about measurements that the two observers can have. Both observers are equivalent; neither is preferred. For instance, whatever of substance observer one says about observer two, two must also conclude about one. For instance, if one says that two's standard clock runs the same as one's, then two says that one's clock runs the same. This is the case of Galilean transformations. The two observers would still be equivalent if one said that two's standard clock ran slower if, at the same time, two also said that one's standard clock ran slower. They both disagree in the same way. It would not work that one said that two's clock ran slower and two agreed that his clock ran slower than one's because then they would not be equivalent; one would have the faster clock. An analogy that I like to use is that in the class, all the students are equivalent even in John says that he is sane and the rest of the class is crazy if then Emily is also allowed to conclude that the rest of the class, including John, is crazy and she is sane.

Some coordinates are the same between the two relatively moving observers. Coordinates transverse to the direction of motion are the same. This can be argued this way. Consider two observers as shown in Figure 11.3 on page 251. As stated earlier, the coordinate transformation between these must be linear so that $z' = Bz$, where B is some function of the relative velocity. Now consider the configuration if the two observers had chosen

instead a coordinate orientation that is obtained by a rotation about the z axis of π radians and invoking the principal that if one sees two moving along at v along the positive x axis then two sees one as moving along the negative x axis at speed v . This reverses the roles of one and two and thus if the transformation was $z' = Bz$ it is now $z = Bz'$ which implies that $B^2 = 1$. We can dismiss the $B = -1$ solution so that we have $z = z'$. A similar argument can be made for the other transverse direction, the y direction.

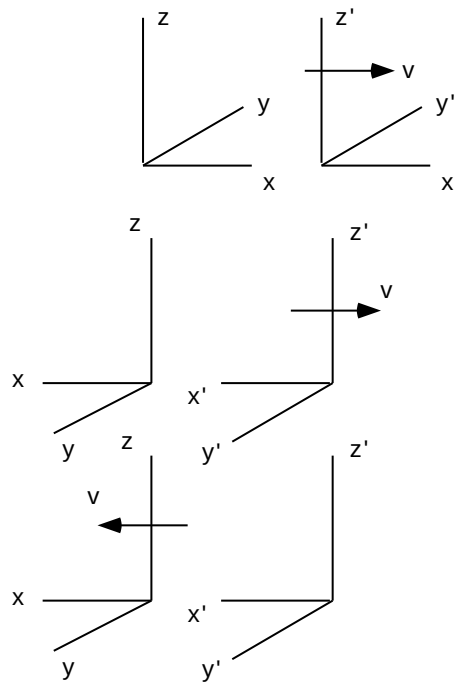


Figure 11.3: **Proof of Agreement on Transverse Direction Coordinates:** At the top of the figure are the coordinate frames for two observers moving relatively along the x axis. Below that are the same observers using frames rotated π radians about the z axis. In the lowest configuration, is the equivalent realization with the first observer moving to the left. This final configuration is the same as the original configuration with the roles of observers one and two reversed.

With the coordinates in the transverse directions the same, we can now show that the relatively moving observers will disagree about the rate at which the standard clock runs.

11.2.1 The Relatively Moving Clock

As discussed in Section ?? on page ?? there is an atomic basis for the standard clock. Regardless, if we can make a system that repeats periodically this system will also be a clock. We will now use the agreement about the transverse lengths to construct a clock that proves that a moving clock must run slower than its identical cousin at rest. Since all observers will agree on the speed of light, we will use the speed of light and an agreed upon distance to make a clock.

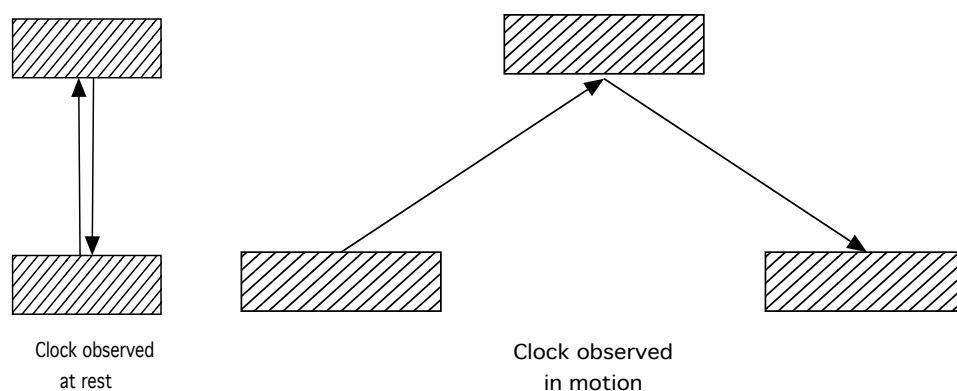


Figure 11.4: **A Clock Using Light:** Using the fact that the speed of light is the same to all inertial observers, we can use light as the basis for a clock. Setting two mirrors a distance, D apart, light bounces back and forth and the interval between passes is the unit of time. Since the light travels a longer distance, this same clock when observed by a relatively moving observer is seen to run slower.

We construct our clock by placing two mirrors a distance D apart and let a burst of light bounce between the two mirrors. The time that passes as the light travels from one mirror to the other and returns is the unit of time. Each observer constructs an identical clock; two mirrors set a distance D apart and held transverse to their relative motion so that they can agree that the mirrors are, in fact, the same distance apart. Consider Harry and Sally again. On her clock, Sally says that the interval between returns of the light is $\Delta t_0 = 2\frac{D}{c}$ but when she observes the operation of Harry's clock, she says that the interval between ticks is longer since the light has to travel a greater distance. Said in another way, only the component of the velocity of the light perpendicular to the mirrors, $v_{\perp}$ matters. Remember that the speed of light is the same in all directions and that both Sally and Harry

have the same speed for light. Thus she says that his clock takes

$$\Delta t = 2 \frac{D}{\sqrt{c^2 - v^2}} = 2 \frac{D}{c \sqrt{1 - \frac{v^2}{c^2}}} = \frac{\Delta t_0}{\sqrt{1 - \frac{v^2}{c^2}}}. \quad (11.6)$$

To Harry though, it is his clock that has a time interval of $\Delta t_0 = 2 \frac{D}{c}$ and her clock that is running slow and has the interval $\Delta t = \frac{\frac{2D}{c}}{\sqrt{1 - \frac{v^2}{c^2}}}$. Remember that $\sqrt{1 - \frac{v^2}{c^2}}$ is the same for v or $-|v|$.

Look at this situation on space-time diagrams. First we draw the situation as represented by Sally. Here Sally's time axis, her $x = 0$ line is vertical and Harry's time axis is a line with slope $\frac{1}{v}$. If she has a clock that reads at time t_0 , she will record a time of $\frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$ for an identical clock carried by him when it reads t_0 to him. She will also record that the moving clock was located at v times that time, $\frac{vt_0}{\sqrt{1 - \frac{v^2}{c^2}}}$, since the clock is traveling along Harry's time axis.

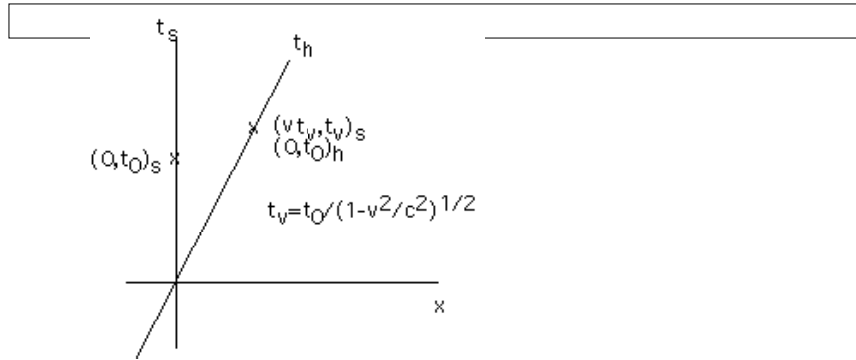


Figure 11.5: **Operation of Mirror Clock:** Sally's time axis is vertical. Harry's time axis has slope $\frac{1}{v}$. If each observer carries an identical clock that to them ticks after a time t_0 , the event of the tick on Sally's clock has the coordinates $(0, t_0)_s$ and, since the clocks are identical, the tick of Harry's clock is labeled by Harry as $(0, t_0)_h$. This same event though is labeled by Sally as $(vt_v, t_v)_s$, where $t_v = \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$.

But a similar discussion is appropriate for Harry. He labels the event of that reading on his clock at $(0, t_0)_h$. His coordinates for the event of the reading of t_0 on her clock is at $(\frac{vt_0}{\sqrt{1 - \frac{v^2}{c^2}}}, \frac{t_0}{\sqrt{1 - \frac{v^2}{c^2}}})_h$. Remember that, to

him, Sally's speed is $-|v|$, a negative number, see Figure 11.6 on page 254. The slope of her time axis in his space-time diagram is a negative number, $\frac{1}{v} = \frac{1}{-|v|}$.

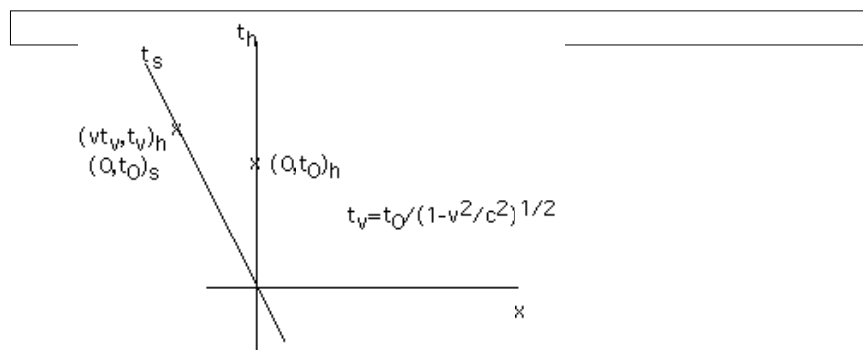


Figure 11.6: **Operation of Mirror Clock in Harry's Frame:** The same pair of related events as in Figure 11.5 on page 253 except as recorded on a space-time diagram based on Harry's time axis being vertical.

11.2.2 Derivation of the Lorentz Transformation

Coordinates of events

As stated in Section 11.1 on page 245, Each observer is to send out a light ray that hits the event and one that returns. Record the times that the first ray is sent out and the time that the second ray comes back and the space coordinate and time coordinate are given by

$$\begin{aligned} x &\equiv \frac{c(\tau_2 - \tau_1)}{2} \\ t &\equiv \frac{\tau_1 + \tau_2}{2}. \end{aligned} \tag{11.7}$$

This rule must be the same for all inertial observers.

When two relatively moving observers label an event, it is important to note though that all observers will use the same two light rays for any particular event, see Figure 11.7 on page 255. In other words, any event is characterized uniquely by the two light rays that pass through it; all observers that are finding the labels of a particular event use the same transmitted and received rays. This apparent coincidence is actually a reflection of the

fact that all observers agree on the speed of light and that the intersection of two light rays is an event and thus a unique label of an event.

11.2.3 Details of the Derivation of the Lorentz Transformations

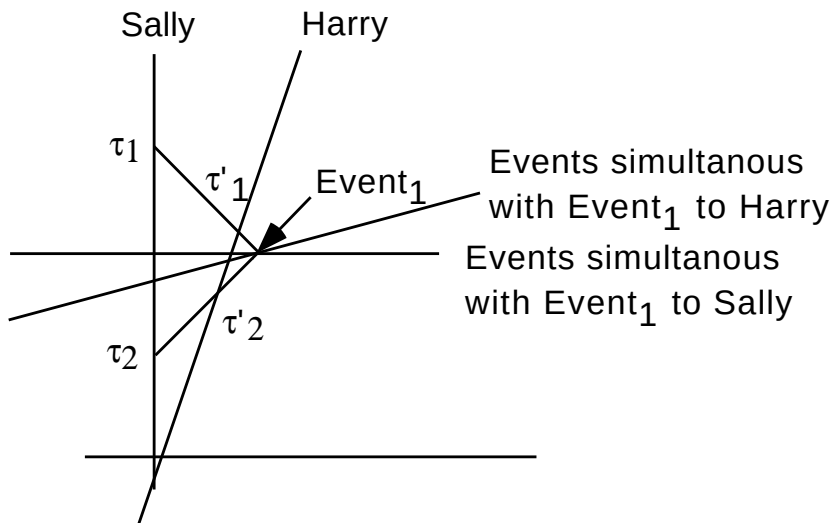


Figure 11.7: **The Rules for Coordinatizing an Event for Two Relatively Moving Observers:** Note that the times t'_1 and t'_2 are the times read on each of the observer's clocks.

Now consider two observers, Sally and Harry, that share the same origin and want to coordinatize the same event. We have shown that the transverse coordinates must be the same for Harry and Sally, Figure 11.3 on page 251, and, in fact, used this information to construct our clocks. Let us now show that this requirement is also obtained in the signaling method of coordinatizing.

In Figure 11.7 on page 255, event 1 is coordinatized by Sally as (x_s, t_s) . By definition, Harry would label it (x_h, t_h) . The Lorentz transformations are the relationship between (x_s, t_s) and (x_h, t_h) .

This is a rather tedious derivation, but a worthwhile exercise. Start by finding the coordinates of the events labeled τ'_1 and τ'_2 in terms of the coordinates of event 1 in Sally's coordinates.

Event τ'_1 has the form (vt_1, t_1) in Sally's coordinates since it is on Harry's time axis and he is moving at a speed v with respect to her. This event is

also on a light ray with event 1. The equation of that light ray is

$$x - x_s = c(t - t_s). \quad (11.8)$$

Putting in the coordinates of the event τ'_1 which is on this line,

$$vt_1 - x_s = c(t_1 - t_s). \quad (11.9)$$

Solving for t_1 ,

$$t_1 = \frac{ct_s - x_s}{c - v}. \quad (11.10)$$

Because of time dilation, see Section 11.2.1 on page 252 and Figure 11.5 on page 253,

$$\tau'_1 = t_1 \sqrt{1 - \frac{v^2}{c^2}}. \quad (11.11)$$

Combining these:

$$\tau'_1 = \sqrt{1 - \frac{v^2}{c^2}} \frac{ct_s - x_s}{c - v} \quad (11.12)$$

Similarly for event τ'_2

$$\tau'_2 = t_2 \sqrt{1 - \frac{v^2}{c^2}} \quad (11.13)$$

and

$$\tau'_2 = \sqrt{1 - \frac{v^2}{c^2}} \frac{ct_s + x_s}{c + v} \quad (11.14)$$

Inserting this into the definitions, Figure 11.7 on page 255 and Equation 11.7 on page 254, and doing some straightforward algebra, we have

$$\begin{aligned} x_h &= \frac{x_s - vt_s}{\sqrt{1 - \frac{v^2}{c^2}}} \\ t_h &= \frac{t_s - \frac{v}{c^2}x_s}{\sqrt{1 - \frac{v^2}{c^2}}}, \end{aligned} \quad (11.15)$$

which are the appropriate Lorentz transformations for this case.

Adding the fact that the transverse directions are unaffected by the velocity transformation, we get the usual Lorentz transformations, Equation 10.8

on page 239, or written out more fully,

$$\begin{aligned}
 x_h &= \frac{x_s - vt_s}{\sqrt{1 - \frac{v^2}{c^2}}} \\
 y_h &= y_s \\
 z_h &= z_s \\
 t_h &= \frac{t_s - \frac{v}{c^2}x_s}{\sqrt{1 - \frac{v^2}{c^2}}}
 \end{aligned} \tag{11.16}$$

An interesting feature of these relations is that the combination $(ct_h)^2 - (x_h)^2 - (y_h)^2 - (z_h)^2$ does not involve the velocity and is therefore equal to Sally's coordinates for the same event,

$$(ct_h)^2 - (x_h)^2 - (y_h)^2 - (z_h)^2 = (ct_s)^2 - (x_s)^2 - (y_s)^2 - (z_s)^2. \tag{11.17}$$

This is a special case of the general form for the invariants of the Lorentz transformations, see Section 12.3 on page 280. We will take advantage of this simple relationship in our subsequent analysis of these relationships.

It is worthwhile checking to see if the Lorentz transforms effect lines of simultaneity and observer time axis as expected. For instance, Harry's line of simultaneity with the origin is the set of events at $t_h = 0$ which is also the events $t_s - \frac{v}{c^2}x_s = 0$; a line sloped at $\frac{v}{c^2}$ through the origin. Harry's time axis is his $x_h = 0$ line. This is a line through the origin with slope $\frac{1}{v}$ on Sally's space-time diagram.

11.3 Using Lorentz Transformations

11.3.1 Time Dilation

Time dilation is the general term for difference in the time interval recorded on two relatively moving but otherwise identical clocks. We had already treated the problem of time dilation in Section 11.2.1 on page 252 using a light clock with mirrors but this is a general phenomena and not limited to light clocks and, using invariants of Equation 11.17 on page 257, the formula for the time difference is direct and intuitive.

If you are moving with a clock and it reads an interval of time Δt_0 . Say this is time of a tick on Sally's clock. At the instant of the tick of Sally's clock, an identical clock which is moving uniformly at a speed v relative to her and synchronized with her clock at the start of the interval by Harry, will read a time Δt which is less than Δt_0 , see Figure 11.8 on page 258 This

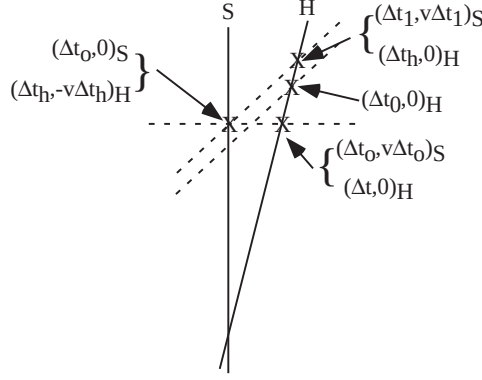


Figure 11.8: **Time dilation in a Moving Clock** Two observers, Sally and Harry, with identical clocks are moving relative to each other at a speed v . At the time that the one observer, say Sally, notes the time Δt_0 on her clock she would assign the coordinates of the simultaneous event on the other clock as $(\Delta t_0, v\Delta t_0)_S$. The observer moving with that clock, Harry, records the event as $(\Delta t, 0)_H$. Since the invariant form must take on the same value for all Lorentz equivalent coordinatizings of the same event, $(\Delta t)^2 - (\frac{0}{c^2})^2 = (\Delta t_0)^2 - (\frac{v\Delta t_0}{c^2})^2$ or $\Delta t = \sqrt{1 - \frac{v^2}{c^2}} \Delta t_0$. Similarly, setting Harry's time for the tick of Sally's clock as Δt_h . his coordinates for the tick of Sally's clock is $(\Delta t_h, -v\Delta t_h)_H$. The invariant relationship requires $\Delta t_h = \frac{\Delta t_0}{\sqrt{1 - \frac{v^2}{c^2}}}$.

same event has two Lorentz equivalent coordinate descriptions, $(\Delta t_0, v\Delta t_0)_S$ and $(\Delta t, 0)_H$. Therefore the invariant requires that

$$\Delta t = \sqrt{1 - \frac{v^2}{c^2}} \Delta t_0. \quad (11.18)$$

Thus since $\Delta t < \Delta t_0$, Sally says that Harry's clock runs slower. To Harry his clock ticks after a time Δt_0 and thus occurs on his time axis after this event which Sally says is simultaneous with her clock tick.

The inverse problem of when Harry says that Sally's clock has ticked requires that we find the event on Harry's time axis simultaneous with the tick of Sally's clock. Harry, assigns a coordinate time of Δt_h to this event, see Figure 11.8 on page 258. Since Sally is moving with a speed v in the negative position direction, Harry assigns the coordinate designation of $(\Delta t_h, -v\Delta t_h)$ to the event of Sally's clock ticking at her clock. Again the invariant requires

that $\Delta t_0 = \sqrt{1 - \frac{v^2}{c^2}} \Delta t_h$ and, in this case, $\Delta t_h > \Delta t_0$. Since

$$\Delta t_h = \frac{\Delta t_0}{\sqrt{1 - \frac{v^2}{c^2}}} > \Delta t_0 = \frac{\Delta t}{\sqrt{1 - \frac{v^2}{c^2}}} > \Delta t,$$

Harry says that Sally's clock has not yet ticked when his clock reads Δt_0 ; it runs slower and yet Sally says that her clock is the first to tick.

Although at first this seems to be an anomaly, with some thought it is clear that this is the way it has to be. Either all identical clocks indicate the same time intervals which is the Newtonian case or as, in this case, all relatively moving clocks run slow but each clock unto itself is correct. It is like a world in which I am sane and everyone else is crazy. This is an equivalent relationship as it holds for everyone. Thus others would conclude that, although I think otherwise, they are sane and I am among the crazies. Of course, a situation with moving clocks running faster would be equivalent but this is not what nature chooses.

11.3.2 Length contraction

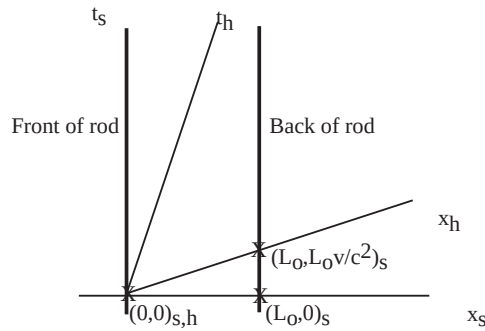


Figure 11.9: **Length Contraction:** Sally carries a rod of length L_0 . The ends of the rod are indicated by the vertical lines $x_s = 0$ and $x_s = L_0$. Harry's time axis is labeled t_h . Each observer says that the length of the rod is the separation of the ends at the same time. Due to the relativity of simultaneity, they use different event pairs to measure a length. It should therefore not come as a surprise that they get different lengths.

In the transverse direction there is no ambiguity about length. In the direction of the motion we have to be careful. Sally holds a rod of length

L_0 . To Harry who is moving relative to Sally at speed v along the same direction as the extended rod, how long is the rod?

In order to understand the situation, let's look at a space-time diagram, Figure 11.9 on page 259.

To any observer, the length of a rod is where the ends are at the same time to that observer. In the frame that is commoving with the rod, Sally, the ends of the rod are at the two lines $x_s = 0$ and $x_s = L_0$. Thus two events at the ends of the rod that are simultaneous to Sally are $(0, 0)_{s,h}$ and $(L_0, 0)_s$ and thus the length is the difference in the space coordinates or L_0 . To Harry, the event that is simultaneous with $(0, 0)_{s,h}$ is at the other end of the rod and is $(L_0, \frac{L_0 v}{c^2})_s$. Remember that Harry's line of simultaneity has slope $\frac{v}{c^2}$. Using the Lorentz transformations to get Harry's coordinate assignment for this event, $(L_0, \frac{L_0 v}{c^2})_s$ is transformed to $(L', 0)_h$, where $L' = \sqrt{1 - \frac{v^2}{c^2}} L_0$. Thus to Harry the length of the rod is the difference of the space coordinates at the same time or he says that the length is

$$L' = L_0 \sqrt{1 - \frac{v^2}{c^2}} \quad (11.19)$$

Another way to see the same result is to realize that to Harry, the coordinates of the end of the rod at $t_h = 0$ must take the form $(L', 0)$. Sally's coordinates for that same event are $(L_0, L_0 \frac{v}{c^2})$. Harry's coordinates for any event differ from Sally's by a Lorentz transformation. Using the invariant of the Lorentz transformations, $(0)^2 - (L')^2 = (c \frac{L_0 v}{c^2})^2 - (L_0)^2$ which then gives Equation 11.19 on page 260 for L' .

11.3.3 The Doppler Effect

We are all familiar with the classical Doppler effect. An approaching fire truck is racing to the chemistry building and the siren is at a high pitch. When the fire truck passes and is moving away from us, the pitch of the siren drops. In other words, an approaching sound source sounds at a higher frequency than the frequency that it produces and a receding sound source has a lower frequency than that of the source.

Consider the case of Sally moving by Harry at a speed v . Sally sends a ray of light to Harry at a time τ_e after the time of their coincidence. The event of arrival of the light on Harry's space-time diagram of the emission is at some time t_e at location $x_e = vt_e$ since it is on Sally's time axis which goes through the origin event and has slope $\frac{1}{v}$. Since she would coordinatize the event as

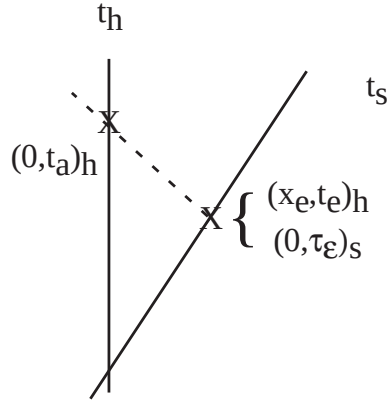


Figure 11.10: **Doppler Effect:** After passing a light signal is sent between two relatively moving observers, Harry and Sally with Sally transmitting. The time interval, t_a , between their passing and the arrival of light signal from the other observer who transmitted at a time τ_e after passing is $t_a = \sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}} \tau_e$

$(0, \tau_e)$, we can use the invariant $c^2 \tau_e^2 = c^2 t_e^2 - v^2 t_e^2$ to find the relationship between τ_e and t_e or as expected from time dilation $\tau_e = t_e \sqrt{1 - \frac{v^2}{c^2}}$.

We can find the time of arrival of the light ray emitted by Sally to Harry, t_a . Note that from Figure 11.10 on page 261, Sally is moving away from Harry. We use the equation of the light ray going through the emission event. The equation of this line is $(x - x_e) = -c(t - t_e)$. Thus $ct_a = ct_e + x_e = (c + v)t_e$ or $t_a = \frac{(1+\frac{v}{c})}{\sqrt{1-\frac{v^2}{c^2}}} \tau_e = \sqrt{\frac{(1+\frac{v}{c})}{(1-\frac{v}{c})}} \tau_e$. τ_e could be considered the first of a sequence of periodic signals and t_a the interval of between the reception of a pair of the signals. Thus the frequency of emission, $f_e = \frac{1}{\tau_e}$ and the frequency of reception, $f_a = \frac{1}{t_a}$ are related by

$$f_a = \sqrt{\frac{1 - \frac{v}{c}}{1 + \frac{v}{c}}} f_e, \quad (11.20)$$

which is the relativistic Doppler effect for the frequency of a signal sent from a receding transmitter to a receiver. The non-relativistic limit, $\frac{v}{c} \ll 1$, of this expression yields the usual Doppler formula, $f_a = (1 - \frac{v}{c}) f_e$. The case of the approaching emitter is simply found by replacing v by $-v$.

There may be some concern about the fact that, in a situation in which

there is more than one spatial dimension, two inertial observers may not meet and this derivation used their coincidence event as a basis. Remember that in any number of spatial dimensions, there is always an event pair that are events of closest approach between the observers. If a commover to one of the observers, O_1 , is located at the event of closest approach on the other observer, O_2 , the above analysis works for that commover. That commover sees the frequency given by Equation 11.20 on page 261. That commover can then merely retransmits the received signals to O_1 . Of course, there is no difference in the time interval for signal between the commover and O_1 . Thus O_1 will see that the interval given by Equation 11.20 on page 261. When you think about this problem you realize that the resolution is in the translation symmetry of the individual inertial observers.

11.3.4 Addition of velocities

Given the Lorentz transformations, Equation 11.16 on page 257, it is now easy to get the formula for the addition of velocities. Consider Harry, Sally, and Tom. Harry moves by Sally to increasing x at v_{hs} , where v_{hs} is Harry's velocity as measured by Sally. For simplicity of the analysis, first let's consider that case that from Sally's point of view Tom is also moving in the positive x direction and he moves by Sally at v_{ts} , where v_{ts} is Tom's velocity as measured by Sally. How fast does Harry say that Tom is moving? The situation is shown Graphically in Figure 11.11 on page 262.

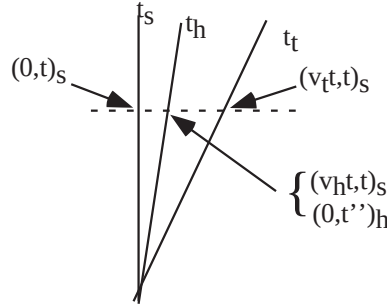


Figure 11.11: **Addition of Velocities.** To determine how colinear velocities add, consider three inertial observers, Harry, Sally, and Tom moving in the same direction. If we know Harry's and Tom's velocities relative to Sally, we can find Tom's velocity relative to Harry by transforming to Harry's frame.

$$\text{Although we do not need it, note that } t'' = \sqrt{t^2 - \frac{v_{hs}^2}{c^2} t^2} = \sqrt{1 - \frac{v_{hs}^2}{c^2}} t.$$

Also note that the v_t and v_h in the figure should be v_{t_s} and v_{h_s} . The graphics package does not allow for stacked subscripts.

Drawing this same set of events in terms of a coordinate system based on Harry is given in Figure 11.12 on page 263

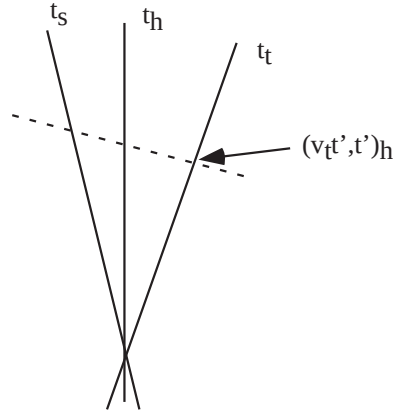


Figure 11.12: **Addition of Velocities from Harry's Point of View**
Three inertial observers moving colinearly as described in Figure 11.11 on page 262 in a coordinate system using Harry as the vertical time axis.

Similarly to above the v_t in the figure should be v_{t_h} . Using the Lorentz transform for this event in Harry's coordinates.

$$\begin{aligned} v_{t_h} t' &= \frac{v_{t_s} t - v_{h_s} t}{\sqrt{1 - \frac{v_{h_s}^2}{c^2}}} \\ t' &= \frac{t - \frac{v_{h_s}}{c^2} v_{t_s} t}{\sqrt{1 - \frac{v_{h_s}^2}{c^2}}} \end{aligned}$$

Dividing these equations

$$v_{t_h} = \frac{v_{t_s} - v_{h_s}}{1 - \frac{v_{h_s} v_{t_s}}{c^2}} \quad (11.21)$$

In the limit of $\frac{v}{c}$ small, this result reduces to the usual Galilean result, Equation 11.5 on page 249.

There should be some concern about how general this result is. Even in a situation with one spatial dimension, there is no need for an event of coincidence between the three observers. There could be situations with a

coincidence of Harry and Sally and a different event for the coincidence of Harry and Tom and of Sally and Tom. The above proof will still work for a comover of Sally at the event of coincidence of Harry and Tom and, since v_{h_s} and v_{t_s} is the same for this comover, the result for the relative velocity of Tom to Harry, v_{t_h} , the desired result, is still that given by Equation 11.21 on page 263. Another approach to the resolution of this problem is to realize that we enjoy space and time translation symmetry. Even if the events of coincidence of Harry and Sally and Harry and Tom and Tom and Sally are not the same event, we can translate such that all three events of two party coincidence are the same. The alert reader should understand that these two explanations of the velocity addition argument are actually the same.

A more substantive concern is that in higher dimensions there is no need for any coincidences at all but also that the velocities need not be co-linear. To be specific, at some time t_{0_s} to Sally, Harry and Tom are at some distance, $\vec{x}_{h_s}$ and $\vec{x}_{t_s}$, with velocities $\vec{v}_{h_s}$ and $\vec{v}_{t_s}$ relative to Sally. There exist comovers of Harry and Tom at Sally at this time. We can now call this the event of coincidence and find the relative velocity of these comovers. The relative velocity between these co-movers will be the same as the relative velocity for Tom relative to Harry. To find this relative velocity, Sally can now do a similar exercise to the one above: pick a time, t , and label where Harry's co-mover and Tom's co-mover are, Lorentz transform to the frame that moves Harry's comover to origin for all times and then find Tom's co-movers relative velocity from his coordinates. The difference between this case and the above is that, after identifying the co-movers at the coincidence point, the co-movers velocities relative to Sally, $\vec{v}_{h_s}$ and $\vec{v}_{t_s}$, are not necessarily co-linear. Using the isotropy of Sally, we can orient the x axis along the velocity of Harry's co-mover. Similarly, we can orient the frame so that the Tom's co-mover velocity is in the $x - y$ plane. Thus the general case is reduced to one requiring only two spatial dimensions and an analysis which is similar to the one in Figure 11.11 on page 262 and Figure 11.12 on page 263 but now in two spatial dimensions. For generality, let's call the x direction the longitudinal direction and y the transverse direction. This construction is shown in Figure 11.13 on page 265.

Again, in the figure $(v_{Lt}t', v_{Tt}t', t')_h$ should be $(v_{Lt_h}t', v_{Tt_h}t', t')_h$ but due to limitations of the graphics package could not be double subscripted.

Using the appropriate Lorentz transformation,

$$v_{Lt_h}t' = \frac{v_{Lt_s}t - v_{h_s}t}{\sqrt{1 - \frac{v_{h_s}^2}{c^2}}}$$

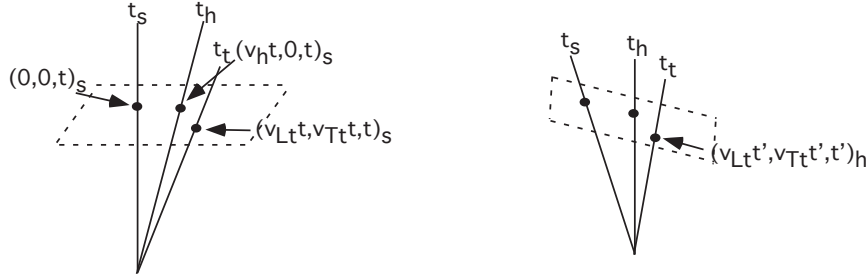


Figure 11.13: **Addition of Velocities for Non-Co-linear Case:** Three inertial observers moving non-co-linearly in two coordinate systems, one using Sally as the vertical axis and one using Harry as the vertical time axis.

$$\begin{aligned} v_{Tt_h} t' &= v_{Tt_s} t \\ t' &= \frac{t - \frac{v_{h_s}}{c^2} v_{Lt_s} t}{\sqrt{1 - \frac{v_{h_s}^2}{c^2}}} \end{aligned} \quad (11.22)$$

Thus, we see that the longitudinal component transforms as in the one space dimension case, Equation 11.21 on page 263, or

$$v_{Lt_h} = \frac{v_{Lt_s} - v_{h_s}}{1 - \frac{v_{h_s} v_{Lt_s}}{c^2}}. \quad (11.23)$$

The transverse component also changes and is given by

$$v_{Tt_h} = \frac{v_{Tt_s} \sqrt{1 - \frac{v_{h_s}^2}{c^2}}}{1 - \frac{v_{h_s} v_{Lt_s}}{c^2}}. \quad (11.24)$$

Despite the added complications, in the limit of $\frac{v}{c} \rightarrow \infty$, this result reduces to the usual Galilean result, Equation 11.5 on page 249.

11.3.5 Time for Different Travelers

Sally and Dorothy are inertial and are at rest with respect to each other, co-moving. They are separated by a distance of one light year. Harry is traveling at $\frac{3}{5}c$ toward Sally and Dorothy. He passes Sally and continues to Dorothy, turns around instantly and at the same speed goes back to Sally. How long is the trip from Sally and back to Sally according to Sally? According to Harry? How far apart are Sally and Dorothy according to Harry? When he is at Sally, how far away does he say that Dorothy is?

Sally says that Harry reached Dorothy in $\frac{5}{3}$ years. Dorothy was one light year away and he was traveling between them at $\frac{3}{5}c$. Similarly for the return trip. So she says that the round trip takes $\frac{10}{3}$ years. By Sally's coordinatizing, the event of Harry meeting Dorothy is $(1, \frac{5}{3})_s$. Using the Lorentz transformations, this same event is labeled by Harry as $(0, \frac{4}{3})_h$. By the way, although Sally says that she and Dorothy are one lightyear apart, he says that they are $\frac{4}{5}$ of a lightyear apart, see Section 11.3.2 on page 259. (He says that Dorothy is coming at him at $\frac{3}{5}c$ and it takes $\frac{4}{3}$ of a year for her to get there.) On a space time diagram, the event of his being at Sally for the first time and the event of where Dorothy is are different for Harry and Sally because of the relativity of simultaneity, Section 10.3 on page 233. To Harry the return trip to Sally will also take $\frac{4}{3}$ of a year and thus the round trip time is $\frac{8}{3}$ of a year. In other words, Harry and Sally disagree about the elapsed time of the trip.

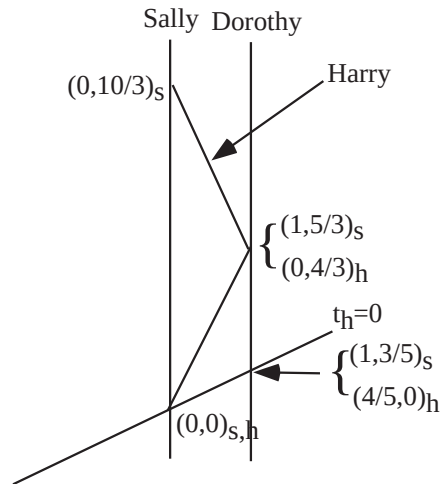


Figure 11.14: **Harry and Sally over Different Trajectories:** Harry travels between Sally and Dorothy. Sally and Dorothy are at rest relative to each other and separated by one light year. Harry departs from Sally and returns after meeting Dorothy. He travels at $\frac{3}{5}c$ to and from Sally. If Harry and Sally measure the elapsed time, they disagree about the total time of the trip.

This difference between elapsed times on different trajectories is a general feature of Special Relativity. Before we can discuss this issue in general terms, we will need to develop an appropriate vocabulary.

11.3.6 Visual Appearance of Rapidly Moving Objects

In order to find the apparent length or the length as it is seen, we must realize that seeing involves the light that enters the eye at any instant. Thus the events of interest are those that leave the extended body at different times and are thus on light-like intervals, on the light cone from the observation event. The following figure, again showing only the ends of the rod, for our case of the relative velocity of $\frac{3}{5}c$, indicates the event at the far end of the rod that is seen at the same time as the origin event at the near end.

The space-time diagram is shown in Figure 11.15 on page 267. This diagram makes clear what is meant by the apparent length of the longitudinally moving rod. Of course, for longitudinal orientation, the rod is always seen as a point. Its apparent length is the range of coordinates covered by the rod at the time the near end is observed. We can find this length directly using

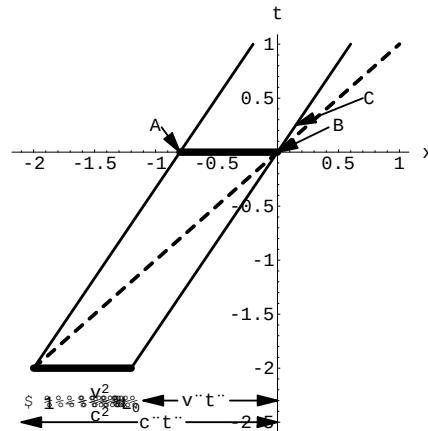


Figure 11.15: **Space-time diagram of moving rod** Space-time diagram in frame of original observer showing the ends of the rod moving at $\frac{3}{5}c$ and the events on the light ray, shown dotted, that are on the same ray as the origin event. The two horizontal lines at $t = 0$ and $t = -2$ show the position of the rod at those times to the rest observer. Shown below that is the length of the rod, $\sqrt{1 - \frac{v^2}{c^2}}$, the distance to the front of the rod, $v|t|$, and the distance to the back of the rod, $c|t|$, to the rest observer.

the space-time diagram. Doing the general case, the equation of the light ray linking with the origin event, B, is $t = \frac{x}{c}$. The time axis of the far side of the rod is $t = \frac{1}{v}(x + \sqrt{1 - \frac{v^2}{c^2}}L_0)$. Finding the simultaneous solution to these

two equations, this event is thus $(-\sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}}L_0, -\sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}}\frac{L_0}{c})$. The definition of the apparent or visual length is the spatial coordinate difference between these two light like related events, B and this event, or $L_{vis} = \sqrt{\frac{1+\frac{v}{c}}{1-\frac{v}{c}}}L_0$. Thus the longitudinally moving rod actually can appear longer than the rod at rest. A similar analysis for the rod once it has passed the observer yields for the visual length $L_{vis} = \sqrt{\frac{1-\frac{v}{c}}{1+\frac{v}{c}}}L_0$. In this case, the length is contracted.

This is a rather striking result. First, since it is first order in $\frac{v}{c}$, it is a large effect. The Lorentz-Fitzgerald contraction is second order. Also since it is first order, it is sensitive to the sign of v . Thus rods moving toward the observer are stretched and rods moving away are contracted. This is consistent with our understanding of the basis for the effect. Because of the finite speed of light, we see farther parts of an extended object at times that are earlier than the times that we see the near parts. Thus, for the rod moving toward you, the farther part is seen when it is further away. Whereas, for the receding rod, the farther parts are not as far away. With this observation and from the form of the equation, we realize that this effect is the spatial correspondent to the well-known Doppler effect for temporal differences. In that case, the approaching light intervals are shortened and the frequencies go up and for a receding source the intervals get longer and the frequencies go down. Here the expansion and contraction are reversed. There should be no surprise that there is a spatial correspondent to the Doppler effect.

The case of the transverse rod can be analysed in a similar fashion.

The visual appearance of a rapidly moving object was discussed first by Penrose [?] and elaborated by Terrell [?]. They discuss the case of an object that is not moving longitudinally toward the observer and restrict the analysis to objects with a small viewing aperture. A very clear presentation of their arguments is give by Weisskopf [?]. The case of the longitudinally oriented rod is discussed by Weinstock [?] but not using space-time diagrams.

Chapter 12

Events, Worldlines, Intervals

12.1 Introduction

As should have become clear in all the previous discussion, the primary unit in Special Relativity is the event. An event is at a place and a time. It is the problem of coordinatizing to label events. Although we have developed a coordinatizing scheme that all inertial observers can agree on, the labeling of any specific event will vary from one inertial observer to the next. For instance, in Section 11.3.5 on page 265, when discussing when Harry is at Dorothy. Note that even though there is one event it has different coordinate descriptions depending on the observer. Harry says that when he is at Dorothy they are zero distance apart but that this occurred $\frac{4}{3}$ years he left Sally. Sally says that when Harry is at Dorothy they are one light year away and it was $\frac{5}{3}$ years after she and he were together. This should not be a surprise. It was the case in Newtonian physics; position labels depended on the observer even in Galilean transformations, Equation 11.5 on page 249. The big difference in the case of the Lorentz transformations is the changes in the time coordinates. It is in this sense that people often talk of the space-time continuum when talking about Special Relativity. There is an intrinsic mixing of space and time labels. It may be worthwhile to make a brief excursion into a discussion of place in the two dimensional plane to remind us of what can be done here.

12.2 Place and Path in the Two Dimensional Plane

The material of this subsection is a trivial diversion from the general development of the kinematic effects of special relativity. It is being set here to

provide a background to the development of a more intuitive basis for ideas that are important to understanding the implications of the relativity.

Consider an unmarked plane of places. Our problem is the establishment of a labeling system that is efficient and easy to use. A more general discussion of this problem of labeling was in Section 11.1 on page 241 and is extended to General Relativity in Section 17.7 on page 372. We will make a pair of often unstated assumptions about the nature of our plane. All places are the same in the sense that there is nothing that you can do at any place that would differentiate that place from any other place. In addition, we assume that at any place all directions are the same. These assumptions allow us to say that the space of places is homogeneous and isotropic.

In this case, all the places can be labeled simply by choosing two orthogonal directions each with a length scale and a special location called the origin. From our assumptions about the structure of our space, it is clear that the choice of an origin is arbitrary and that this choice cannot play an important part in any analysis of the properties of places. You cannot tell where you are and all that can matter is difference in the labels of the places. Another way to say this is to say that, although you can talk about where you are with the coordinate of a place, all important concerns do not involve the coordinates but involve only differences in coordinates, $(\vec{x}_2 - \vec{x}_1)$. This form is unchanged by a translation of the coordinate origin. If you replace the coordinates $\vec{x}_{1,2}$ by the new coordinates $\vec{x}'_{1,2} = \vec{x}_{1,2} - \vec{a}$,

$$(\vec{x}'_2 - \vec{x}'_1) = (\vec{x}_2 - \vec{x}_1);$$

the combination of variables $(\vec{x}_2 - \vec{x}_1)$ is unchanged by the translation. It is called a form invariant for translations. A form invariant is a combination of coordinates that when transformed, although all the coordinate terms change, is itself not changed. In more formal language, the transformation of the coordinates, $\vec{x}' = \vec{x} + \vec{a}$, is a family of transformations called translations. The elements of this family are labeled by the parameters $\vec{a}$. In the form invariant, $(\vec{x}_2 - \vec{x}_1)$, the transformation $\vec{x}' = \vec{x} + \vec{a}$, changes all the elements in the form $(\vec{x}_2 - \vec{x}_1) \rightarrow (\vec{x}'_2 - \vec{x}'_1)$ but, because the label of the transformation, $\vec{a}$, drops out of the form $(\vec{x}_2 - \vec{x}_1) = (\vec{x}'_2 - \vec{x}'_1)$ that particular form is unchanged.

An important issue in the plane is the distance between two points. In the above paragraph, a distance scale has been defined for each coordinate direction. These need not be the same. This may seem to be a bizarre choice but it does happen. I was born and raised in Philadelphia, a city with row houses. The primary problem was that movement within the grid

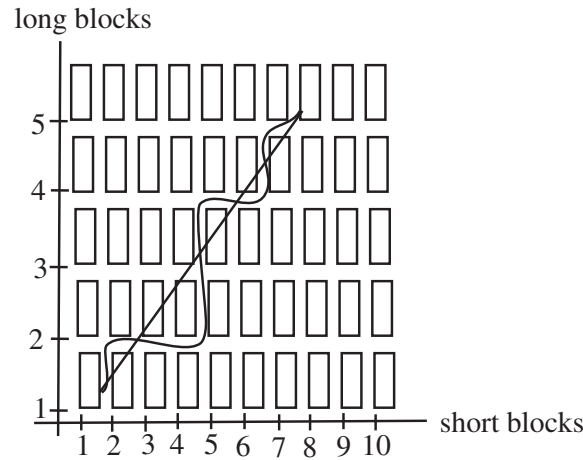


Figure 12.1: **A Path in the City** For a trip from my house to a nearby friends house in a city grid, two paths are shown. One is a realistic one using the city grid and one as a crow would fly.

was restricted to be along the row houses or the ends of the block, see Figure 12.1 on page 271. The unit of length was what was called a “block”. The trouble with the “block” was that in the two directions the actual block had different lengths. The simplest measure of distance was the “block” and the length in units of blocks was the total number of blocks traversed,

$$d_{blocks} = n_s + n_l$$

where n_s is the number of short blocks in the path and n_l is similarly defined for long blocks. In fact, the shorter direction was about a quarter of the long direction. There was another distance that was used called the “city block” which was the same length as the long direction block. When we talked seriously about how far something was, we used “city blocks”. In other words,

$$d_{city\ blocks} = \left(\frac{n_s}{4} + n_l \right).$$

You can make this more sophisticated and they did by adding a coordinate grid that identified the blocks and then the distances n_s and n_l were given as coordinate differences, see Figure 12.1 on page 271. In this way, the distance in “city blocks” was

$$d_{city\ blocks} = \left(\frac{|\Delta x_s|}{4} + |\Delta x_l| \right).$$

Actually, the situation was more bizarre than that because this is the distance in "city blocks" as walked and not as the crow would fly. For people, the distance is the total number of intervals over the grid between the places because that is how you have to move in this system. As the crow flies reflects the distance defined in a fashion that reflects the underlying homogeneous isotropic nature of the plane,

$$d_{crow} = \sqrt{\frac{(\Delta x_s)^2}{16} + (\Delta x_l)^2},$$

still in "city blocks". This crow path was not available to the walker. The walker distance is real and leads to interesting geometry. What is the straight line path? What does a circle look like?

Let's now follow the more usual construction of using the same distance scale in each of the coordinate directions. In fact, we can go one step further and use the same distance scale for all directions. Then the distance between two points can be found by using the distance scale along the direction set by the two points, as the crow flies. Saying this we now realize that it is assuming the rotation of the distance scale to different orientations does not change it. This is an expression of the underlying isotropy of the space. An alternative approach to finding the distance is to use the coordinate differences. To reproduce the effects of the reorientation of the rod, the measure of distance must reflect the translation and reorientation invariance of the distance measure. This rotational and translation invariance in the definition of distance is expressed as

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{\Delta x^2 + \Delta y^2}. \quad (12.1)$$

In other words, the transformation of the coordinate system produces changes in the coordinates which for rotations and translations are $(x, y) \rightarrow (x', y')$ with $x' = x \cos(\theta) + y \sin(\theta) + a_x$ and $y' = y \cos(\theta) - x \sin(\theta) + a_y$ where θ is the angle of rotation and the label for the elements of the family of rotations and $\vec{a}$ is the labels for the translations. Equation 12.2 on page 273, is a form invariant for these transformations.

Let's now follow the more usual construction of using the same distance scale in each of the coordinate directions. In fact, we can go one step further and use the same distance scale for all directions. Then the distance between two points can be found by using the distance scale along the direction set by the two points, as the crow flies. Saying this we now realize that it is assuming the rotation of the distance scale to different orientations does

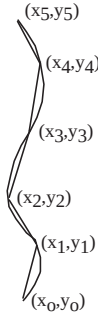


Figure 12.2: **A Path in a Plane** For a curved path, a cumulative distance can be assigned by adding the straight line distance for each segment of a sensibly rectified approximation to the curve, $d[path] = \sum_{i=0, \text{Path}}^{f-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$. In the limit of small segments, this cumulative distance is the total distance over the path, $d[path] = \int_{(x_0, t_0), \text{Path}}^{(x_f, t_f)} \sqrt{dx^2 + dy^2}$.

not change it. This is an expression of the underlying isotropy of the space. An alternative approach to finding the distance is to use the coordinate differences. To reproduce the effects of the reorientation of the rod, the measure of distance must reflect the translation and reorientation invariance of the distance measure. This rotational and translation invariance in the definition of distance is expressed as

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{\Delta x^2 + \Delta y^2}. \quad (12.2)$$

In other words, the transformation of the coordinate system produces changes in the coordinates which for rotations and translations are $(x, y) \rightarrow (x', y')$ with $x' = x \cos(\theta) + y \sin(\theta) + a_x$ and $y' = y \cos(\theta) - x \sin(\theta) + a_y$ where θ is the angle of rotation and the label for the elements of the family of rotations and $\vec{a}$ is the labels for the translations. Equation 12.2 on page 273, is a form invariant for these transformations.

Distance not only depends on the two points but also the path connecting the points. In our discussion above, we used the word distance for the separation of the points which is the distance over the straight line path between the points, as the crow flies. In the more general case, there can be an arbitrary path connecting the points. Since the number of paths

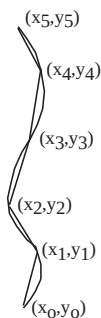


Figure 12.3: **A Path in a Plane** For a curved path, a cumulative distance can be assigned by adding the straight line distance for each segment of a sensibly rectified approximation to the curve, $d[path] = \sum_{i=0, \text{Path}}^{f-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$. In the limit of small segments, this cumulative distance is the total distance over the path, $d[path] = \int_{(x_0, t_0), \text{Path}}^{(x_f, t_f)} \sqrt{dx^2 + dy^2}$.

between two points is an infinite class that is larger than the class of real numbers, you cannot perform ordinary analysis on the path labels. For this case, the name functional is used instead of function. This extension of the idea of functions to functionals leads to a wealth of new and very powerful mathematics.

For our purposes it is sufficient to consider paths that can be sensibly rectified into a sequence of straight line segments and the total distance is the sum of these intervals. We could measure the length of each segment by taking our length definition and placing it along the straight line segments measuring each length by aligning the length along the segment. This is a place where our assumption of homogeneity and isotropy come into play. The length of the rod is the same no matter how we orient it and where we place it. Alternatively, we can use the coordinate difference method but then we have to be sure the coordinate system reflects these symmetries. Our definition of length, Equation 12.2 on page 273 accomplishes this if the coordinate directions use the same length. In this case, the path length is

the sum of the appropriate straight line separations or

$$d[\text{path} : (x_0, y_0; x_f, y_f)] = \sum_{i=0, \text{Path}}^{f-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (12.3)$$

or in the limit of small intervals

$$d[\text{path} : (x_0, y_0; x_f, y_f)] = \int_{(x_0, y_0) \text{Path}}^{(x_f, y_f)} \sqrt{(dx)^2 + (dy)^2}. \quad (12.4)$$

Using the Equation 12.3 on page 275 and rotational and translational symmetry it is easy to show that the straight line path is the shortest. This takes advantage of another idea that is worth discussing at this point. Our previous discussions of the rotations and translations dealt with changes to the coordinate system. These same ideas can also be applied to the points themselves. You can view the transformation as shifting all the points. When the transformation is on the coordinate axis the transformation is called passive. When it is applied to the points, it is called active. In the case of proving that the straight line is the shortest path between two points, we can use a passive transformation to move the origin to one of the points. Then we can use an active transformation to rotate the the second point so that it is on the y axis. Using Equation 12.3 on page 275, the path length of any arbitrary path other than a straight line will include δx contributions which can only increase the sum above that of the path with no δx contributions. There for the straight line path is the shortest path.

Although we have used the idea of angle in the above discussion of rotations, we have assumed the usual measure of angle and not discussed the exact nature of the relationship of angle to the rotation transformation. Our only real criteria when viewing the coordinate transformation was to preserve the form of the distance measure,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{\Delta x^2 + \Delta y^2}.$$

As stated above, rotations and translations are the family of transformations that preserve this form,

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 = (x'_2 - x'_1)^2 + (y'_2 - y'_1)^2, \quad (12.5)$$

where x' and y' are the new coordinate labels for the point at (x, y) . A more general transformation would be

$$\begin{aligned} x' &= ax + by + c \\ y' &= dx + ey + f. \end{aligned} \quad (12.6)$$

The form invariant is preserved by all values of c and f but the other parameters that label the transformation are constrained by the requirement of Equation 12.5 on page 275. These constraints are

$$\begin{aligned} a^2 + d^2 &= 1 \\ b^2 + e^2 &= 1 \\ 2ab + 2de &= 0. \end{aligned}$$

and the solutions can be written in terms of the single parameter b as $a = \sqrt{1 - b^2} = e$ and $d = -b$. In other words, our family of transformations is a three parameter group; two for translations, c and f , and one for rotations, b . Putting this family of transformations into a group adds the requirement that sequential operation of the transformation is a one of the members of the family. There is an added requirement that the family of transformations have an identity transformation. Our family of transformations satisfy these additional requirements. Thus the rotation segment of our transformations can be written as a matrix operating on the doublet (x, y) with the matrix given by

$$\begin{pmatrix} \sqrt{1 - b^2} & b \\ -b & \sqrt{1 - b^2} \end{pmatrix} \quad (12.7)$$

This is not the only solution to the constraint equations but it is a nice one in that the identity element, do nothing element, is the $b = 0$ element. A complication with this form for describing rotations is the requirement that two rotations are a rotation. This requires that

$$\begin{aligned} &\begin{pmatrix} \sqrt{1 - b_2^2} & b_2 \\ -b_2 & \sqrt{1 - b_2^2} \end{pmatrix} \begin{pmatrix} \sqrt{1 - b_1^2} & b_1 \\ -b_1 & \sqrt{1 - b_1^2} \end{pmatrix} \\ &= \begin{pmatrix} \sqrt{1 - b_3^2} & b_3 \\ -b_3 & \sqrt{1 - b_3^2} \end{pmatrix}, \end{aligned}$$

or

$$b_3 = b_1 \sqrt{1 - b_2^2} + b_2 \sqrt{1 - b_1^2}. \quad (12.8)$$

Using the parameter b to label the rotations, you can see that they do not add when rotations are combined. This is an unfortunate statement because it is not really true. The b s add but not simply like numbers.

By now, I would guess that many of you smell the rat in this analysis. If we had just used the good old fashioned idea of the angle to label the rotations, the rotation matrix would simply have been

$$\begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \quad (12.9)$$

or $b = \sin \theta$ and the condition of Equation 12.8 on page 276 becomes simply $\theta_3 = \theta_1 + \theta_2$ or usual numeric addition. When you think about it, you realize that the property of simple addition comes from the definition of the angle. The angle θ between two straight lines is

$$\theta \equiv \frac{S}{R} \quad (12.10)$$

where S is the arc length of the circle generated by a distance R as the line of length R sweeps from the one line to the other.

A further digression on this will be useful. First let's clarify an idea about transformations. Up until now, we have treated transformations as coordinate label changes with the points fixed. An alternative is to leave the coordinate system unchanged but shift all the points. For example in the case of translations, all the places in the plane are shifted to new places by the rule that the new place labeled by $\vec{x}$ is the old place that had been labeled $\vec{x} + \vec{a}$. When a transformation is used to change the coordinate labeling, the transformation is called passive. When the transformation is used to change the places, the transformation is called active.

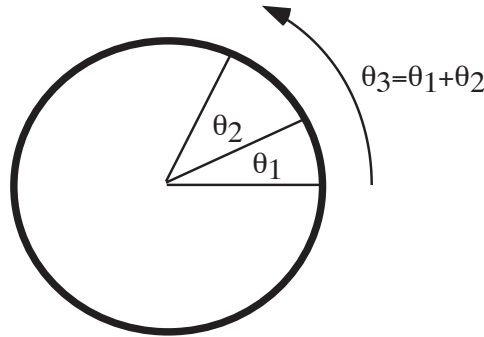


Figure 12.4: **Addition of Angles** The usual definition of angle is $\theta \equiv \frac{S}{R}$ where S is the arc length of a circle. Since this definition of angle uses the length of a segment of the invariant curve for rotations, the circle, and since the arc lengths of curves add simply, this "angle" adds simply.

In our discussion of the label of rotations above a certain curve in the two plane, a circle, played a special role in the definition of angle. The circle acquires its special role because it is the locus of places that is generated by rotations when the transformations are viewed actively. This definition uses a construction that is based on the fact that the active transformation of points generates a set of points on what is called the invariant curve or

surface in higher dimensions. In other words, the circle is the invariant curve for our rotations. Using the arc lengths along these curves which are clearly additive in the numeric sense, we obtain an additive measure of rotations.

Our b in the previous analysis was a definition of amount of rotation, "angle," that was $\theta' \equiv \frac{h}{H}$ where h is the height and H is the hypotenuse of a right triangle constructed between the lines generating the "angle." Of course, we recognize this as $b = \sin \theta$ and then the complex addition formula, Equation 12.8 on page 276 is a reflection of the fact that

$$\arcsin b_1 + \arcsin b_2 = \arcsin \left(b_1 \sqrt{1 - b_2^2} + b_2 \sqrt{1 - b_1^2} \right).$$

Let's take this analysis of rotations as active transformations one step further. In old fashioned two space, (x, y) , if we consider rotations about the origin, we had a form invariant $x^2 + y^2 = r^2$, i. e. if the coordinate system is rotated, in the new coordinates that same point is now labeled as (x', y') and the combination, $x'^2 + y'^2$ takes on the same value, $x'^2 + y'^2 = r^2$. Now viewed actively, for every point, rotations generate a locus of places with the same distance from the origin satisfy this form invariant and are circles centered on the origin. A rotation will map one point on a circle onto another point on that same circle.

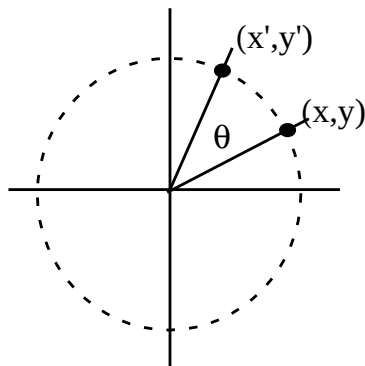


Figure 12.5: Rotations can be treated as a mapping of the points in the plane. Here the rotation θ maps (x, y) onto (x', y')

In the above analysis of the addition of angles, Equation 12.9 on page 276, the special functions $\cos(\theta)$ and $\sin(\theta)$ did neat things for us. Another related property of these functions is that they satisfy the constraint, $x^2 + y^2 = r^2$, by having $x = r \cos(\theta)$ and $y = r \sin(\theta)$. This constraint is satisfied for any θ since $\cos^2(\theta) + \sin^2(\theta) = 1$. We also can describe the location of a

place, (x, y) , as a distance and an angle, (r, θ) . It is a trivial observation that the rotations connect different places with the same distance. Consider three places, $(r, 0)$, (x_1, y_1) and (x_2, y_2) , that are the same distance from the origin, i. e. on the same circle, $r = \sqrt{x_1^2 + y_1^2} = \sqrt{x_2^2 + y_2^2}$. The rotation that maps place $(r, 0)$ on to (x_1, y_1) is labeled by an angle θ_1 , $\theta_1 = \arctan(\frac{y_1}{x_1})$, and the rotation that maps $(r, 0)$ onto (x_2, y_2) is labeled by an angle θ_2 , $\theta_2 = \arctan(\frac{y_2}{x_2})$. A rotation with the angle labeled $\theta_2 - \theta_1$ maps (x_1, y_1) onto (x_2, y_2) – and again the angles are additive, see Figure 12.5 on page 276.

We can use this idea to find the general transformation law for rotations, Equation 12.9 . Consider the point $(r_1, 0)$, it is obvious that under the rotation θ this point is mapped to $(r_1 \cos(\theta), r_1 \sin(\theta))$. Similarly, a rotation of the same angle θ maps $(0, r_2)$ into the point $(-r_2 \sin(\theta), r_2 \cos(\theta))$.

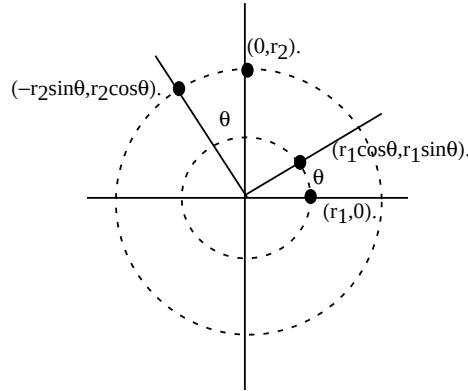


Figure 12.6: Rotating points on the coordinate axis to find the form of the transformations under rotations.

Since the transformations are linear, we can make the general transformation for any point (x, y) by combining these:

$$\begin{aligned} x' &= x \cos(\theta) - y \sin(\theta) \\ y' &= x \sin(\theta) + y \cos(\theta) \end{aligned} \quad (12.11)$$

You can easily derive the addition formula for the trigonometric functions by having two rotations. Starting with (x, y) and transforming by a rotation, θ_1 to (x', y') and then transforming (x', y') by a rotation θ_2 to (x'', y'') , we have the sequence of equations:

$$\begin{aligned} x' &= x \cos(\theta_1) - y \sin(\theta_1) \\ y' &= x \sin(\theta_1) + y \cos(\theta_1) \end{aligned} \quad (12.12)$$

and

$$\begin{aligned}x'' &= x' \cos(\theta_2) - y' \sin(\theta_2) \\y'' &= x' \sin(\theta_2) + y' \cos(\theta_2)\end{aligned}\tag{12.13}$$

and since the angles are additive,

$$\begin{aligned}x'' &= x \cos(\theta_1 + \theta_2) - y \sin(\theta_1 + \theta_2) \\y'' &= x \sin(\theta_1 + \theta_2) + y \cos(\theta_1 + \theta_2)\end{aligned}\tag{12.14}$$

Substituting the first equation into the second and reorganizing, we also have

$$\begin{aligned}x'' &= x (\cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)) \\&\quad - y (\sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2)) \\y'' &= x (\sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2)) \\&\quad + y (\cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2))\end{aligned}\tag{12.15}$$

Equating the coefficients of x and y , we have the usual formulas for the addition of angles of the trigonometric functions

$$\sin(\theta_1 + \theta_2) = \sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2)\tag{12.16}$$

$$\cos(\theta_1 + \theta_2) = \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2)\tag{12.17}$$

The requirement of additivity is a linear one and thus does not fix the scale of the angles. Since θ is dimensionless, and we require that additivity hold for all r , a natural measure of angle is the radian, $\theta = \frac{\text{arc length}}{r}$. With this definition, the full circle has angle 2π .

12.3 Minkowski Space-time

In the discussion of Harry, Sally, and Dorothy, Section 11.3.5 on page 265, we studied the fact that Sally and Harry had different trajectories in space-time. A trajectory is the connected set of events that represent the places and times through which an object moves. Trajectories of material objects and observers are called world-lines. Sally's time axis is her world-line. Sally is also an inertial observer; she experiences no acceleration in the course of her motion. Harry's world-line, on the other hand, has a bend. He has an acceleration and that is knowable by him, see Section 9.2 on page 215;

he spills his martini on his shirt. Note that any other inertial observers coordinatizing this situation cannot be differentiated from Sally and would have her world-line as straight and his would still have a bend. Thus the idea of an inertial world-line, straight, is the same for all inertial observers and the straightness of the inertial observer world-line is unchanged, the Lorentz transforms, Equation 11.16 on page 257, map straight lines into straight lines. The space-time that is coordinatized by Sally is an example of a Minkowski space-time. This is a space-time that has a global coordinate system such as Sally's and is also invariant under the Lorentz transformations and space and time translations. This large group of transformations is called the Poincaré transformations. The basic assumption of special relativity is that the events take place in a four dimensional structure that contains a three dimensional Euclidean space and a time like dimension. A $(3, 1)$ space that has an invariant measure,

$$\Delta s^2 \equiv \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2 \Delta t^2, \quad (12.18)$$

for the Poincaré transformations. It is easy to show by substitution of the Lorentz transformations, Equation 11.16 on page 257, that the interval, Equation 12.18 on page 281 is invariant under Lorentz transformations. Since it is defined by differences in coordinates, it is invariant under translations in space and time.

This $(3, 1)$ is different from the Euclidean space plus time of Newtonian physics in that the group of transformations that govern it, the Poincaré transformations, preserve this different measure. The Galilean transformations, Equations 11.4 on page 248, preserve the usual distance measure of the Pythagorean Theorem, $\Delta l^2 = \Delta x^2 + \Delta y^2 + \Delta z^2$, which is invariant under rotations and spatial translations, see Section 12.2 on page 269. This is why the three spatial dimensions are a Euclidean space. In the case of the Galilean transformations, the time coordinate is unchanged. That is why the Newtonian world is a space plus time world not a $(3, 1)$ world.

12.3.1 Future, Past, and Elsewhere

The geometry of this four dimensional, $(3, 1)$, manifold is important to our understanding of the kinematics of Relativity. Although there are similarities with a four dimensional Euclidean manifold, the differences are important and often at the heart of what seems to be a paradox of Relativity. Since this manifold has translation symmetry, the structure is contained in the relationship between event pairs; since this is a space-time, events are

the fundamental element and not points. This is the first important difference since we tend not to think of ourselves as lines, a connected set of events - a sequence of heart beats.

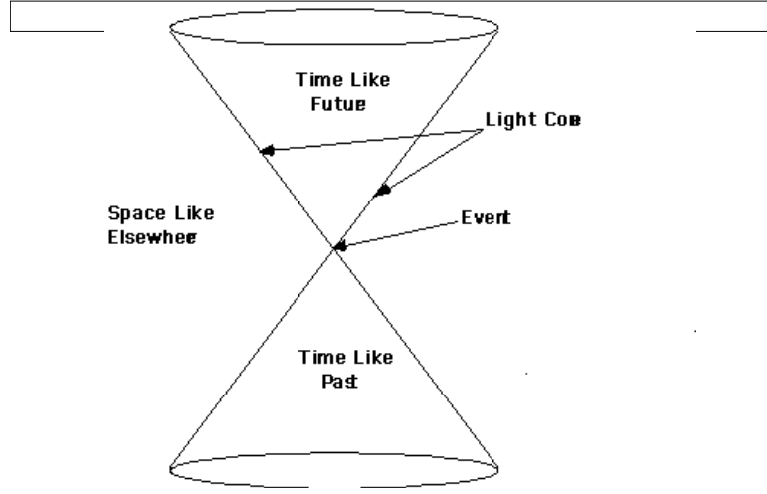


Figure 12.7: Future Past and Elsewhere. For any event, in this case the event at the vertex of the two cones, all the other events in space-time can be categorized into a future, a past, and an elsewhere. Since the trajectories of light rays are unchanged by Lorentz transformation, this classification of the relationship between two events is the same for all inertial observers.

In a $(3,1)$ space the events are labeled with the four coordinates $(\vec{x}, ct)$ and these coordinates are articulated by the procedures of Section 11.2.2 on page 254. We now reverse our approach and define a Poincaré transformation as any transformation of the labels of events such that Equation 12.18 on page 281 is a form invariant, $\Delta x'^2 + \Delta y'^2 + \Delta z'^2 - c^2 \Delta t'^2 = \Delta x^2 + \Delta y^2 + \Delta z^2 - c^2 \Delta t^2$, where

$$\begin{aligned} x' &= a_{xx}x + a_{xy}y + a_{xz}z + a_{xt}ct + b_x \\ y' &= a_{yx}x + a_{yy}y + a_{yz}z + a_{yt}ct + b_y \\ z' &= a_{zx}x + a_{zy}y + a_{zz}z + a_{zt}ct + b_z \\ ct' &= a_{tx}x + a_{ty}y + a_{tz}z + a_{tt}ct + cb_t. \end{aligned} \tag{12.19}$$

Where the b_i are the translations in space and time. Any set of the sixteen elements, a_{ij} that satisfy this is called a Lorentz transformation. Obviously, with this definition, the translations are a subset of the Poincaré transformations. Among the a_{ij} are the rotations.

The nature of these requirements are much simpler but not different in nature when looking at our text book example: the (1,1) world. Here there are no rotations. In fact, it is legitimate to consider the (3,1) case the (1,1) case with rotations added. In this case, Equation 12.20 on page 283 is replaced by

$$\begin{aligned}x' &= a_{xx}x + a_{xt}ct + b_x \\ ct' &= a_{tx}x + a_{tt}ct + cb_t.\end{aligned}\tag{12.20}$$

and the condition

In space-time as in space, we have the concept of a continuous connected set of events. This is called a trajectory. At any event on a trajectory, the slope is the inverse of the velocity relative to some inertial observer; the one that we choose to have a time axis, the $x = 0$ line, straight up and with a perpendicular set of lines of simultaneity, lines of constant t . The trajectories of light rays are straight lines with slope one. The trajectories of inertial observers are straight lines with slopes greater than one.

Space-time around any one event is divided into regions separated by the trajectories of light rays emanating from that event, see Figure 12.7 on page 282. This separation of events is the same for all Lorentz observers since the light ray trajectories are unchanged by the Lorentz transformations. All the events in the upper light cone are the future of the event in question. This is in the sense that, from the origin event and any event in the future, there exists an inertial observer for whom the interval between these two events is a pure time, $(0, \tau)$, i. e. no spatial separation, and that the time of the other event is after the now of our original event, $\tau > 0$. Any other inertial observer would give the two events labels $(\vec{x}_0, t_0)$ as the original event and $(\vec{x}_1, t_1)$ as the other event and, using the form of the invariant, Equation 12.18 on page 281, we have

$$c^2\tau^2 = -\Delta s^2 = c^2(t_1 - t_0)^2 - (x_1 - x_0)^2 - (y_1 - y_0)^2 - (z_1 - z_0)^2 > 0.\tag{12.21}$$

Similarly, events in what is called the backward light cone from our original event are in the past. There exists an inertial observer for whom the second event is a pure time, $(0, \tau)$, but in this case $\tau < 0$. Again, any other inertial observer would give the two events labels $(\vec{x}_0, t_0)$ and $(\vec{x}_1, t_1)$ and, using the form of the invariant, Equation 12.18 on page 281, we have $c^2\tau^2 = -\Delta s^2 = c^2(t_1 - t_0)^2 - (x_1 - x_0)^2 - (y_1 - y_0)^2 - (z_1 - z_0)^2 > 0$.

The union of the events in the past and of the events in the future to our origin event is the set of time-like events relative to our origin event. This is all events relative to the original event with intervals in any inertial coordinate system such that the negative of the interval squared, $-\Delta s^2 =$

$c^2(t_1 - t_0)^2 - (x_1 - x_0)^2 - (y_1 - y_0)^2 - (z_1 - z_0)^2 > 0$. If the event under discussion is in the upper light cone or future of our origin event, we chose the positive sign for the square root of the negative of the interval squared and if the event is in the lower or past light cone we chose the negative of the square root of the interval squared. This is called the proper time between the events although we have to be careful because, just as is the case in Euclidean spaces where distance is the corresponding concept and we now realize that distance is path dependent, convention calls this the proper time between the events, see Section 12.2 on page 269 and discussion later in this section.

There are clearly a large number of events that are not time-like relative to our origin event, see Figure 12.7 on page 282. These are called elsewhere or space-like events relative to our origin event. Similar to our construction of future and past, for any elsewhere event there exists a Lorentz observer for whom the events are separated by a spatial interval, $(\vec{x}, 0)$. Again, any other inertial observer would give the two events labels $(\vec{x}_0, t_0)$ and $(\vec{x}_1, t_1)$ and, using the form of the invariant, Equation 12.18 on page 281, we have $d^2 = \vec{x}^2 = \Delta s^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2 - c^2(t_1 - t_0)^2 > 0$.

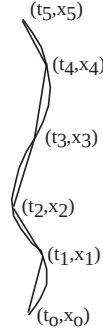


Figure 12.8: **A Time-Like Trajectory in Space-Time** For a curved trajectory to be time-like each segment must be time-like. A cumulative time can be assigned to a time like trajectory by adding the proper time for each segment of a sensibly rectified approximation to the curve,

$$\tau [traj.] = \sum_{i=0, \text{Traj.}}^{f-1} \sqrt{(t_{i+1} - t_i)^2 - \frac{(x_{i+1} - x_i)^2}{c^2} - \frac{(y_{i+1} - y_i)^2}{c^2} - \frac{(z_{i+1} - z_i)^2}{c^2}}. \text{ In the}$$

limit of small segments, this cumulative time is the proper time over the trajectory,

$$\tau [traj.] = \int_{(x_0, t_0), \text{Traj.}}^{(x_f, t_f)} \sqrt{dt^2 - \frac{dx^2}{c^2} - \frac{dy^2}{c^2} - \frac{dz^2}{c^2}}.$$

Another important geometric concept deals with trajectories. It makes sense to describe a cumulative interval along a trajectory. Depending on the bending, it is sensible to approximate the cumulative interval by sensibly rectifying the trajectory and adding the intervals of each segment, see Figure 12.8 on page 284,

$$\tau(traj.) = \sum_{i=0}^f \sqrt{(t_{i+1} - t_i)^2 - \frac{(x_{i+1} - x_i)^2}{c^2} - \frac{(y_{i+1} - y_i)^2}{c^2} - \frac{(z_{i+1} - z_i)^2}{c^2}}. \quad (12.22)$$

This is the same procedure that is used in the case of paths in an two, three, or even an n dimension Euclidean space, see Section 12.2 on page 269. The complication here is the fact that intervals squared, Equation 12.18 on page 281, comes in two varieties, time-like and space-like, negative and positive intervals squared respectively. Although it is possible to have trajectories with some time-like segments and some space-like segments, it does not make any sense to assign a cumulative interval to them. We will see shortly that, in addition, trajectories with space-like segments have problematic causal structure. For these reasons, we require that all trajectories of sensible material objects have time-like trajectories. A time-like trajectory is one in which all the segments are time-like intervals. An equivalent definition is that a trajectory is time-like if for every event on the trajectory all subsequent events are in the future light cone of that event and all previous events are in the past light cone. A special simple case of a time-like trajectory is the trajectory of an inertial observer. The proper time over this trajectory is to within an origin time the coordinate time for that inertial observer; it is the time on his clock. This idea is carried over to the case of any time-like trajectory. The cumulative proper times of the segments is called the proper time over the trajectory and is the time that would be recorded on a clock carried along that trajectory. This is actually what we did in our analysis of the travel time of Harry and Sally in Section 11.3.5 on page 265. Without comment, we added the times of the segments of Harry's time as recorded on a clock used by an inertial observer that would be comoving with him; it was without comment because it was so eminently plausible. We will look at this question in more detail in a later section, Section 12.7 on page 299. Note also that the slope of the segment is the inverse of the average velocity in the segment. In this sense the velocity along any time-like segment is always less than the speed of light or any segment that is space-like has an average velocity that is greater than c .

There are also trajectories in space-like directions, i. e., all segments

space-like, and a cumulative distance can be assigned. This cumulative distance is called the proper distance of the trajectory. Since there are three spatial directions, there are also space-like surfaces, generated by two non-parallel space-like directions, and space-like volumes, all three sides space-like. These are not possible constructions for time-like situations since there is only one time coordinate.

As stated above the slope at any event is the inverse of the instantaneous velocity of the trajectory at that event. For time-like segments, this velocity can be used to reference a family of inertial observers that have that velocity. These are called comovers. The comover that shares the event is called the local comover. In our example of time differences for two traveler in Section 11.3.5 on page 265, Sally and Dorothy were comovers. They were not local comovers since they were apart. Harry had two families of comovers, one for the first segment and another one for the second segment. The comover that moved with Harry on the either segment is the local comover for that segment. The anomaly of Harry's travel time is that he uses the clocks of two non-identical comovers. This is also the signature that he is accelerated. We will discuss this situation in more detail in Section 12.7 on page 299.

As in our discussion of points in space, Section 12.2 on page 269, transformations of a Minkowski space-time can be viewed both as active or passive. As in that case, the passive view identifies the transformation with a change in the coordinate system, how a relatively moving inertial observer would label the same event, and in the active view the transformation is between related events that have similar properties to an different inertial observer. When viewed actively, the Lorentz transformations are often referred to as Lorentz boosts. For example, consider the event that one observer says is simultaneous with his/her origin event and a distance d away along the positive x axis. Now consider a Lorentz transform with the label v . When viewed as a passive transformation, the event which was labeled $(d, 0)$ is now

labeled $\left(\frac{d}{\sqrt{1-\frac{v^2}{c^2}}}, \frac{-\frac{v}{c^2}d}{\sqrt{1-\frac{v^2}{c^2}}} \right)$ Whe this transformation is viewed actively, the

second label is an event that is at the same proper distance from the origin as the event $(d, 0)$ and is one that would be labeled as $(d, 0)$ to the inertial observer moving at a speed v in the negative x direction. Similarly, for the time-like separated event from the origin $(0, \tau)$, the Lorentz transformation

labeled v produces the label $\left(-\frac{v\tau}{\sqrt{1-\frac{v^2}{c^2}}}, \frac{\tau}{\sqrt{1-\frac{v^2}{c^2}}} \right)$ and is the label for that event for an inertial observer moving with speed v along the positive x axis.

When viewed actively, this is an event at the same proper time from the origin event as the original event and one that would be labeled as $(0, \tau)$ for a Lorentz observer who is moving at a speed v in the negative x direction.

12.4 Causality and Trajectories

An obviously important issue is the idea of causality. In Newtonian physics, causality expressed itself as preceding events could influence subsequent events but not visa versa. In relativistic physics, there are more subtle points. Influence is achieved by being able to get an object or message from one event to another. This was actually the basis for our designation of the events in the forward light cone from any event as that events future. These later events are ones that can be connected by a time-like trajectory. In other words, an observer at the origin event could throw a rock or a light ray and it would get to the location of the event before the event happened. This causal relation does not hold for events in the elsewhere of our origin event. There is no material or light signal that can go from our original event to the place of elsewhere events before they occur. Events in each others elsewhere are not causally related.

There is a more interesting example of causality breakdown associated with trajectories with space-like intervals. First consider a simple example with three events that are all space-like with respect to each other and simultaneous to some inertial observer. We could take the time of these events to be $t = 0$ for that observer, see Figure 12.9 on page 288. The line of simultaneity of these events can be taken as $t = 0$ for that observer. There are two other inertial observers moving toward the original observer and all three coincide at the central event and synchronize their clocks to $t = 0$ at that time. For the two later observers, the events are no longer simultaneous. In fact, for the relatively moving observer moving to increasing position the event at positive position occurs before $t = 0$ and the event at negative position is after $t = 0$. For the observer moving to decreasing position, the equivalent situation obtains; the event at negative position occurs before $t = 0$ and the event at positive position occurs after $t = 0$. In other words, for events in the elsewhere of an origin event, the sign of the time of those events will depend on the inertial observer who coordinatizes the events. This is not the case for events that are future or past of the origin event. These are either positively signed for events in the future or negatively signed for events in the past to all inertial observers.

The situation becomes more complex when we connect these space-like

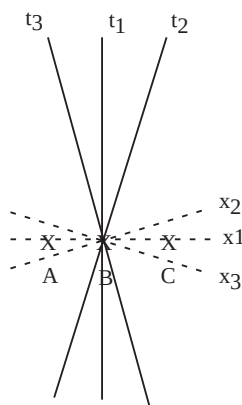


Figure 12.9: **Temporal Order of Space-like Events** Three events labeled A B and C are simultaneous to some inertial observer, Observer 1. Two other inertial observers move relative to 1 one to increased position, observer 2, and one to decreasing position, observer 3, and coincide with 1 at event B. Their lines of simultaneity for event B are shown as x_2 and x_3 . To observer 2 event A is after B, $t_{A_2} > 0$, and event C is before B, $t_{C_2} < 0$. On the other hand, observer 3 has event A is before B, $t_{A_3} < 0$, and event C is after B, $t_{C_3} > 0$.

related events in a trajectory. Consider a trajectory that has the three events in the previous paragraph and Figure 12.9 on page 288. The other segments of the trajectory are time-like and comove with the observers 2 and 3. In Figure 12.10 on page 289, we show this trajectory. These two co-moving inertial observers have a very different interpretation of the trajectory. To the co-mover of the first segment, the trajectory unfolds as a single trajectory with a uniform sense of flow. For the comover to the second time-like segment, the trajectory folds onto itself with an interval of time in which there are three events on the trajectory at any one time. This bizarre behavior makes no sense in classical physics. Suppose the trajectory represented in Figure 12.10 on page 289 was a message being sent from each of the co-movers. Co-mover 3 says that he/she is sending the message to co-mover 2 but conversely co-mover 2 says that he/she is sending the signal. Suppose that at event B the message is destroyed. Which co-mover did not get the message? In light of this causality problem, we make it an axiom that the trajectories of objects or messages must travel by time-like or for light light-like trajectories. In order to guarantee a coherent idea of causality, there is no signal that travels faster than the speed of light.

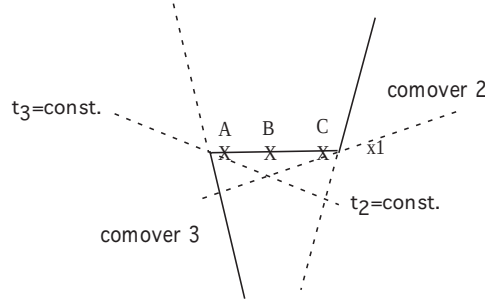


Figure 12.10: **A Trajectory with a Spacelike Interval** The Three events labeled A B and C of Figure 12.9 on page 288 are part of a continuous trajectory. The inertial observer that is co-moving with the first segment of the trajectory would indicate that the trajectory advances smoothly through the space-like interval and it makes sense to assign a direction. To the co-mover to the second time-like segment, the region has a flow direction that is the reverse of the usual. In fact as seen by the lines of simultaneity for co-mover 2, $t_2 = \text{const.}$, that for times slightly before event C to a time slightly after event A, there are three events that are on the trajectory.

12.5 The Hyperbolic Hangle*

In Section 12.2 on page 269, we found that for rotations, the proof that distance from the origin to a place (x, y) , $d^2 = x^2 + y^2$ is a form invariant for rotations about the origin is easily proven, if we have the harmonic functions, $\sin \theta$ and $\cos \theta$ which satisfy $\sin^2 \theta + \cos^2 \theta = 1$. Can we find a set of functions similar to the sin and cos and an additive measure like θ that satisfy the form invariant for Lorentz transformations, $x^2 - c^2 t^2 = d^2$? The answer to this question will be yes. We will do this analysis in a space with only one space and one time dimension for notational simplicity. The extension to higher dimensions is trivial.

Define

$$\cosh(\phi) \equiv \frac{e^\phi + e^{-\phi}}{2} \quad (12.23)$$

and

$$\sinh(\phi) \equiv \frac{e^\phi - e^{-\phi}}{2} \quad (12.24)$$

Also, define $\tanh(\phi) \equiv \frac{\sinh(\phi)}{\cosh(\phi)} = \frac{e^\phi - e^{-\phi}}{e^\phi + e^{-\phi}}$. Then $\cosh^2(\phi) - \sinh^2(\phi) = 1$ for all ϕ . These functions are not the only pair of functions that satisfy this relationship but, as we will see, they are the only pair that do that and

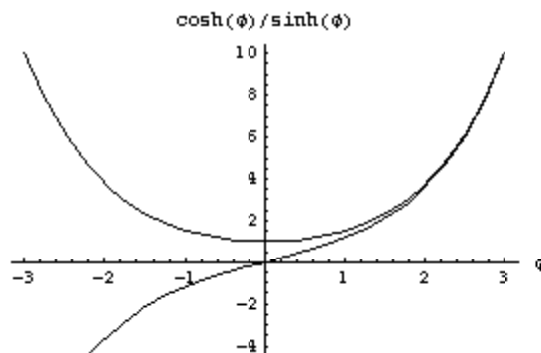


Figure 12.11: The functions $\cosh(\phi)$, topmost, and $\sinh(\phi)$, lower are plotted on the same graph.

also satisfy the additivity requirement for Lorentz transformations when they are parameterized by ϕ . Using ϕ and calling it the hyperbolic angle or “hangle,” we can develop a set of relations for Lorentz transformations much like that which was accomplished in the previous section for rotations, Section 12.2 on page 269. Another name for ϕ which we will occasionally use is the “rapidity.” Hangle reminds us of the relationship to angles and rapidity reminds us of the relationship to velocity. This second relationship will be clarified later.

First let’s find the invariant surfaces for Lorentz transformations. This development follows the same pattern as the two space case, Section 12.2 on page 269. These surfaces must be the values of (x, t) that satisfy the invariant form, $x^2 - c^2 t^2$, events that have the same proper time and distance from the origin event. This is better expressed using the four hyperbolas $x = \pm\sqrt{d^2 + c^2 t^2}$ for events that are space-like from the origin and $t = \pm\frac{1}{c}\sqrt{d^2 + x^2}$ for events that are time-like, see Figure 12.12 on page 291.

In particular treating the Lorentz transformations with label $-v$ actively, the event $(d, 0)$ is mapped onto (x_1, t_1) where x_1 and t_1 are:

$$\begin{aligned} x_1 &= d \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \\ t_1 &= \frac{d}{c} \frac{\frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}}. \end{aligned} \quad (12.25)$$

Obviously $d^2 = x_1^2 - c^2 t_1^2$ and, as v varies from $-c$ to c , (x_1, t_1) moves along the hyperbola, $x = \sqrt{d^2 + c^2 t^2}$. Also note that for any (x_1, t_1) , the proper

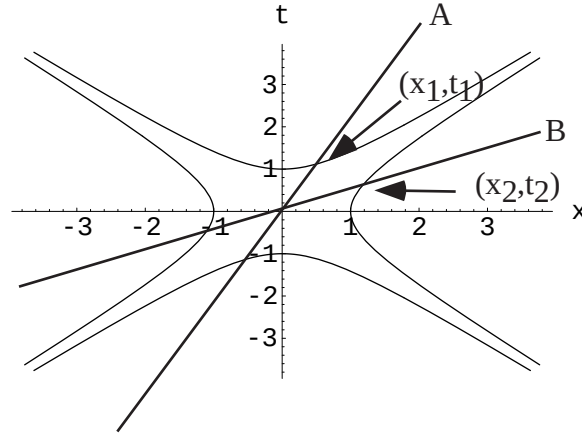


Figure 12.12: **Lorentz Invariant Surface** The four curves considered in a counter clockwise loop starting at the extreme right, $x = \sqrt{d^2 + c^2 t^2}$, $t = \frac{1}{c}\sqrt{d^2 + x^2}$, $x = -\sqrt{d^2 + c^2 t^2}$, and $t = -\frac{1}{c}\sqrt{d^2 + x^2}$ form the invariant surface for the Lorentz transformations. The events on the two hyperbolas, $x = \sqrt{d^2 + c^2 t^2}$ and $x = -\sqrt{d^2 + c^2 t^2}$, are in the elsewhere of the origin event. The events on the hyperbola, $t = \frac{1}{c}\sqrt{d^2 + x^2}$, are in the future of the origin event and the hyperbola, $t = -\frac{1}{c}\sqrt{d^2 + x^2}$, are in the past of the origin event. The inertial observer that shares the origin event and (x_1, t_1) , Equation 12.25 on page 290, has as the locus of events that are simultaneous with the origin event the line B passing through the origin and the event (x_2, t_2) , Equation 12.26 on page 292. The use of the term surface for these curves stems from the fact that in a three space one time world these are surfaces. For the figure the value of $d = 1$ was chosen.

distance from the origin is d . In this sense, the events along the hyperbola, $x = \sqrt{d^2 + c^2 t^2}$, are the locus of events that are the same proper distance, d , from the origin. Similarly, the event $(-d, 0)$ generates the hyperbola $x = -\sqrt{d^2 + c^2 t^2}$.

A similar argument holds for the event $(0, \frac{d}{c})$ which is on the upper leg of the invariant hyperbola, $t = \frac{1}{c}\sqrt{d^2 + x^2}$, and is mapped onto (x_2, t_2) on that hyperbola as follows:

$$x_2 = d \frac{\frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$t_2 = \frac{d}{c} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (12.26)$$

It is important to note that, for the inertial observer that passes through the origin and the event (x_1, t_1) , the line of events that contains (x_2, t_2) and the origin event is the locus of events that are simultaneous with the origin event to that observer.

We can also identify x_1 and t_1 by

$$\begin{aligned} x_1 &= d \cosh(\phi) \\ t_1 &= \frac{d}{c} \sinh(\phi) \end{aligned} \quad (12.27)$$

and as ϕ develops the point (x_1, t_1) moves up along the hyperbola $x = \sqrt{d^2 + c^2 t^2}$. Again, we identify x_2 and t_2 as;

$$\begin{aligned} x_2 &= d \sinh(\phi) \\ t_2 &= \frac{d}{c} \cosh(\phi) \end{aligned} \quad (12.28)$$

and as ϕ increases the point moves outward along the hyperbola. If we identify the line between the events $(0, 0)$ and (x_1, t_1) as the line of simultaneity and the line between the events $(0, 0)$ and (x_1, t_1) as the worldline of the inertial frame moving at speed v relative to the original frame, we have a new labeling for the Lorentz transformations.

In other words, and very similarly to the case of rotations, we can identify the Lorentz transformations of velocity v with a hangle ϕ as:

$$\begin{aligned} x' &= x \cosh(\phi) + ct \sinh(\phi) \\ ct' &= x \sinh(\phi) + ct \cosh(\phi) \end{aligned} \quad (12.29)$$

where $\tanh(\phi) \equiv \frac{v}{c}$.

The great advantage of this labeling of the Lorentz transformations is the additivity of the labeling in ϕ . To show this, we follow the same pattern that was used for the rotations. Consider two subsequent Lorentz transformations:

$$\begin{aligned} x' &= x \cosh(\phi_1) + ct \sinh(\phi_1) \\ ct' &= x \sinh(\phi_1) + ct \cosh(\phi_1) \end{aligned} \quad (12.30)$$

and

$$\begin{aligned} x'' &= x' \cosh(\phi_2) + ct' \sinh(\phi_2) \\ ct'' &= x' \sinh(\phi_2) + ct' \cosh(\phi_2) \end{aligned} \quad (12.31)$$

and, if we want the hangles to be additive, the compounding of these transformations should yield:

$$\begin{aligned}x'' &= x \cosh(\phi_1 + \phi_2) + ct \sinh(\phi_1 + \phi_2) \\ ct' &= x \sinh(\phi_1 + \phi_2) + ct \cosh(\phi_1 + \phi_2)\end{aligned}\quad (12.32)$$

Inverting the defining relations, Equation 12.23 on page 289 and Equation 12.24 on page 289, we have $e^{\pm\phi} = \cosh(\phi) \pm \sinh(\phi)$. Expanding the definition of $\sinh(\phi_1 + \phi_2)$ and $\cosh(\phi_1 + \phi_2)$ it is easy but tedious to show that these functions satisfy the correct addition relations so that they are equal to two successive Lorentz transformations of magnitude ϕ_1 and ϕ_2 . The addition formula is:

$$\sinh(\phi_1 + \phi_2) = \sinh(\phi_1) \cosh(\phi_2) + \cosh(\phi_1) \sinh(\phi_2) \quad (12.33)$$

$$\cosh(\phi_1 + \phi_2) = \cosh(\phi_1) \cosh(\phi_2) + \sinh(\phi_1) \sinh(\phi_2) \quad (12.34)$$

Although the formula for the addition of velocities is rather cumbersome, the addition of hangles is simple.

As with the case of rotations, the hangle is dimensionless and the defining rules do not set a scale. The “natural” scale for ϕ is the ratio of the c times the proper time along the hyperbola $x = \sqrt{d^2 + c^2 t^2}$, remember that it is a timelike curve, to the proper distance from the origin to that hyperbola. Calling this unit of hangle the hradian, we have

$$\phi(\text{in hradians}) \equiv \frac{c \times \text{proper time along given proper distance curve}}{\text{proper distance}} \quad (12.35)$$

Notice that the hangle goes to infinity as the relative velocity goes to c .

From the previous material and realizing that the commover at the event (x, t) has a relative velocity $\frac{v}{c} = \frac{ct}{x} = \tanh(\frac{c\tau}{d})$, we have:

$$x = d \cosh\left(\frac{c\tau}{d}\right) \quad (12.36)$$

$$ct = d \sinh\left(\frac{c\tau}{d}\right) \quad (12.37)$$

where τ is the proper time from the event $(d, 0)$ to the event (x, t) on the trajectory $x = \sqrt{d^2 + c^2 t^2}$.

This rather extended diversion served two purposes. Firstly, it clarified the complex addition formula for collinear velocities. Here the problem was that velocity was not a good label for the family of transformation that is identified as the Lorentz velocity transformations. The additive label is the

hangle not the velocity. This is similar to the case in two spatial dimensions discussed in Section 12.2 on page 269. The second reason will be that the time-like trajectory that is generated by the invariant curve at a distance d from the origin will be seen to be the trajectory of the a uniformly accelerated observer, see Chapter 14 on page 309. This is a special case of motion but it has great interpretive power and is valuable as an exact analytic solution to a non-trivial state of motion.

We could have defined the hangle using the time-like lines through the origin. In Figure 12.12 on page 291, the line through (x_1, t_1) and the origin is the world line of an inertial observer moving at speed $\frac{1}{v}$ relative to our coordinating inertial observer. The event (x_1, t_1) is on a curve that is the locus of events at the same proper time from the origin event. The events on this curve are space like and are separated from the vertex event of the space like curve by a proper distance $s_{(x_1, t_1)}$. Similarly to Equation 12.35 on page 293, the hangle can be defined as

$$\phi(\text{in hradians}) \equiv \frac{\text{proper distance along given proper time curve}}{c \times \text{proper time}} \quad (12.38)$$

shinola Finish this

12.5.1 The same result directly using calculus*

Unfortunately, the standard symbol for the derivative is lower case d ; this is also our symbol for the proper distance to the trajectory of the uniformly accelerated object from the origin event, $(0,0)$. In order to avoid confusion, in this section, I will use the symbol D for the proper distance and keep d for the derivative. It should be clear from the context the role the symbol d is playing.

From the definitions of $\sinh$ and $\cosh$, Equations 12.24 and 12.23 on page 289, we can derive

$$\frac{d(\sinh(\phi))}{d\phi} = \cosh(\phi) \quad (12.39)$$

$$\frac{d(\cosh(\phi))}{d\phi} = \sinh(\phi) \quad (12.40)$$

From the definition of the proper time along the trajectory between the event $(D, 0)$ and the event (x, t) ,

$$c\tau = \int_{\text{traj}, (D, 0)}^{(x, t)} \sqrt{c^2 (dt)^2 - (dx)^2} \quad (12.41)$$

and the equation of the trajectory in terms of ϕ , $ct = D \sinh(\phi)$ and $x = D \cosh(\phi)$,

$$\begin{aligned} c\tau &= \int_0^{\phi(x,t)} D \sqrt{\cosh^2(\phi') - \sinh^2(\phi')} d\phi' \\ c\tau &= D \phi(x,t) \end{aligned} \quad (12.42)$$

where $\phi(x, t) \equiv \tanh^{-1}(\frac{ct}{x})$ is the hangle from the origin to the event (x, t) .

Although this approach seems to be much simpler than the previous derivation, it must be kept in mind that the derivative and integral relations used above depend on the additivity properties that were the central part of the previous discussion. When you think about it you realize that the arc length along the trajectory is additive and therefore must be proportional to ϕ , the measure that is additive along the invariant curve.

12.6 Four Vectors and Invariants*

In the previous sections, we developed the idea of a Minkowski space, Section 12.3 on page 280. In this section, we want to develop an efficient formalism for expressing ideas in Minkowski space. As in Euclidean space, a vector formalism is possible. Given an origin event and inertial observer, a coordinate system can be established. An event is a place and a time, a set of four numbers, $(\vec{x}, t)$, that specifies that event in that coordinate system. We can designate the coordinates with an index x^μ with $x^0 \equiv ct$, $x^1 \equiv x$, $x^2 \equiv y$, and $x^3 \equiv z$. In this notation, the Lorentz transformations are expressed as

$$x'^\alpha = \sum_{\mu=0}^3 \Lambda^\alpha_\mu x^\mu \quad (12.43)$$

with

$$\Lambda^\alpha_\mu = \begin{pmatrix} \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} & 0 & 0 & \frac{-\frac{v}{c}}{\sqrt{1-\frac{v^2}{c^2}}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{-\frac{v}{c}}{\sqrt{1-\frac{v^2}{c^2}}} & 0 & 0 & \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \end{pmatrix} \quad (12.44)$$

for a Lorentz transformation along the positive z direction with speed v . Other Lorentz transformations are implemented similarly. The rotations which are a subgroup of the Lorentz transformations are the usual rotation elements operating in the bottom three by three spaces in this four by four

object. There is a broadly accepted convention that simplifies the notation considerably called the Einstein convention that eliminates the summation symbol for cases in which the same index appears up and down in the same equation. In this notation, Equation 12.43 on page 295 appears simply as

$$x'^{\alpha} = \Lambda^{\alpha}_{\mu} x^{\mu}. \quad (12.45)$$

Given two events we can talk about the interval between them. In this language, there is a four vector interval

$$s^{\mu} = (c(t_2 - t_1), (x_2 - x_3), (y_2 - y_3), (z_2 - z_3)). \quad (12.46)$$

The invariant interval squared, Equation 12.18 on page 281, is now expressed as

$$\Delta s^2 = s^{\alpha} g_{\alpha\mu} s^{\mu}, \quad (12.47)$$

where

$$g_{\alpha\mu} \equiv \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad (12.48)$$

$g_{\alpha\mu}$ is called the metric in Minkowski space. This terminology and the index operations are a special case of the general formalism called indexology in Appendix ?? on page ??. We will not need the full power of the indexology formalism until we get to gravitation as geometry in Chapter 17 on page 363. For our purposes now, it will be sufficient to deal with four vectors defined as any quartet of numbers that transform under Lorentz transformations in the same pattern as Equation 12.45 on page 296 and Lorentz scalars which are quadratic forms of four vectors such as Equation 12.47 on page 296 that are invariant under Lorentz transformations. There are many examples of four vector quantities such as a four velocity, defined below, and the relativistic energy and momentum defined in Chapter ?? on page ??.

The condition that the interval squared be an invariant places a condition on the form of the Lorentz transformations,

$$\Delta s^2 = \Delta s'^2$$

which implies

$$s^{\alpha} g_{\alpha\mu} s^{\mu} = s'^{\rho} g_{\rho\gamma} s'^{\gamma} = s^{\delta} \Lambda^{\rho}_{\delta} g_{\rho\gamma} \Lambda^{\gamma}_{\omega} s^{\omega}$$

for all s^{μ} . The reader must realize that all the indices in this expression are summed and are thus dummies and can take any greek letter. Thus we have the condition that

$$g_{\alpha\rho} = \Lambda^{\mu}_{\alpha} g_{\mu\gamma} \Lambda^{\gamma}_{\rho}. \quad (12.49)$$

This condition can be used as the defining equation for the Lorentz transformations. The sixteen numbers, Λ_ν^μ , are a Lorentz transformation if they satisfy Equation 12.49 on page 296. Equation 12.49 on page 296 is not sixteen equations since $g_{\mu\nu}$ is symmetric. This is ten independent equations which leaves six free parameters. That is just what is needed – three parameters to label a velocity and three parameters to label rotations in a three space.

For a time-like trajectory, we can define a four vector velocity by using the proper time over the trajectory to calculate a rate of change. In other words, a trajectory which is a connected set of events which would be coordinatized by some inertial observer as $(\vec{x}(t), t)$, can be parametrized by the elapsed proper time of the time-like trajectory, $(\vec{x}(\tau), t(\tau))$, where

$$\begin{aligned} \tau[\text{trajectory} : (\vec{x}_0, t_0; \vec{x}, t)] &\equiv \int_{\text{traj}, \vec{x}_0, t_0}^{\vec{x}, t} \sqrt{dt^2 - \frac{d\vec{x} \cdot d\vec{x}}{c^2}} \\ &= \int_{\text{traj}, \vec{x}_0, t_0}^{\vec{x}, t} \sqrt{1 - \frac{\frac{d\vec{x}}{dt} \cdot \frac{d\vec{x}}{dt}}{c^2}} dt \\ &= \int_{\text{traj}, \vec{x}_0, t_0}^{\vec{x}, t} \sqrt{1 - \frac{\vec{v} \cdot \vec{v}}{c^2}} dt. \end{aligned} \quad (12.50)$$

where $\vec{v} \equiv \frac{d\vec{x}}{dt}$ is called the coordinate velocity and is the usual definition of velocity. This may look like a rather complex object but this construction is much like the parametrizing of a curve in two space with distance along the curve, see Section 12.2 on page 269. The notation is also the same as that in Section 12.2. The elapsed proper time is a functional of the trajectory but is a function of the labels of the events at the end points of the integral. Since it is a function of the time on the trajectory we can derive a differential form for Equation 12.50. Differentiating with respect to the time of the event on the world-line,

$$\frac{d\tau}{dt} = \sqrt{1 - \frac{\vec{v} \cdot \vec{v}}{c^2}} \quad (12.51)$$

Putting all this together, we can construct a four vector velocity,

$$u^\mu \equiv \frac{dx^\mu}{d\tau} = \frac{\frac{dx^\mu}{dt}}{\frac{d\tau}{dt}}, \quad (12.52)$$

which, since the Lorentz transforms are linear and constant, transforms the same way as x^μ in Equation 12.45 on page 296.

The construction of other kinematic four vectors such as a four acceleration follows the same pattern,

$$a^\mu \equiv \frac{du^\mu}{d\tau} = \frac{d^2x^\mu}{d\tau^2}. \quad (12.53)$$

By construction of the proper time, it follows that the four velocity vector length is always the same,

$$u^\mu g_{\mu\nu} u^\nu = \frac{\frac{dx^\mu}{dt} g_{\mu\nu} \frac{dx^\nu}{dt}}{\left(\frac{d\tau}{dt}\right)} = \frac{\left(-1 + \frac{\vec{v} \cdot \vec{v}}{c^2}\right)}{\left(1 - \frac{\vec{v} \cdot \vec{v}}{c^2}\right)} = -1. \quad (12.54)$$

Any four vector with a negative length squared such as the four velocity is called time-like four vector; there always exists a Lorentz frame in which the four vector takes the form $(c, \vec{0})$. In the case of the four velocity, the constant $c = 1$. In the general frame, the four velocity takes the form

$$u^\mu = \left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}, \frac{\vec{v}}{\sqrt{1 - \frac{v^2}{c^2}}} \right). \quad (12.55)$$

Differentiating Equation 12.54 on page 298,

$$\frac{d}{d\tau} (u^\mu g_{\mu\nu} u^\nu) = 0 = 2a^\mu g_{\mu\nu} u^\nu. \quad (12.56)$$

In the frame in which $u^\mu = (1, \vec{0})$, the four acceleration must take the form $a^\mu = (0, \vec{c})$. Since the length squared is a Lorentz invariant,

$$a^\mu g_{\mu\nu} a^\nu > 0. \quad (12.57)$$

The acceleration is a space-like four vector. Differentiating Equation 12.55 on page 298, general form of the acceleration four vector is

$$\begin{aligned} \frac{du^\mu}{d\tau} &= \left(\frac{\vec{v} \cdot \frac{d\vec{v}}{d\tau}}{\left\{1 - \frac{v^2}{c^2}\right\}^{\frac{3}{2}}}, \frac{\frac{d\vec{v}}{d\tau}}{\sqrt{1 - \frac{v^2}{c^2}}} + \vec{v} \frac{\vec{v} \cdot \frac{d\vec{v}}{d\tau}}{\left\{1 - \frac{v^2}{c^2}\right\}^{\frac{3}{2}}} \right) \\ &= \left(\frac{\vec{v} \cdot \frac{d\vec{v}}{dt}}{\left\{1 - \frac{v^2}{c^2}\right\}^{\frac{3}{2}}}, \frac{\frac{d\vec{v}}{dt}}{\sqrt{1 - \frac{v^2}{c^2}}} + \vec{v} \frac{\vec{v} \cdot \frac{d\vec{v}}{dt}}{\left\{1 - \frac{v^2}{c^2}\right\}^{\frac{3}{2}}} \right) \frac{dt}{d\tau} \\ &= \left(\frac{\vec{v} \cdot \frac{d\vec{v}}{dt}}{1 - \frac{v^2}{c^2}}, \frac{d\vec{v}}{dt} + \vec{v} \frac{\vec{v} \cdot \frac{d\vec{v}}{dt}}{1 - \frac{v^2}{c^2}} \right) \end{aligned} \quad (12.58)$$

In the frame in which $u^\mu = (1, \vec{0})$, i. e. $\vec{v} \rightarrow 0$, the acceleration four vector is $(0, \vec{a})$, where $\vec{a}$ is the coordinate acceleration as measured by a comoving inertial observer. This is the acceleration of Newtonian physics.

Further manipulations of four vectors can be found in Section 14.2.1 on page 311, Chapter ?? on page ??, and Appendix ?? on page ??.

12.7 Harry, Dorothy, and Sally Revisited

With the insights gained from the previous discussions, it is worthwhile to revisit Harry, Dorothy and Sally of Section 11.3.5 on page 265. In that section, we found that Harry, on his return to Sally, had not aged as much as she had and, in addition, he was accelerated. With the definitions of Section 12.3 on page 280, Equation 12.22 on page 285, we can see that this effect is a special case of a general result that in space-time, for time-like trajectories, the straight line trajectory, one possessed by an inertial observer, is the longest trajectory and all other time-like trajectories are shorter. The proof of this statement follows from the same reasoning as the proof that in space the straight line is the shortest line. The difference here is the negative contributions of the spatial intervals in the sum of the segments that contribute to the total proper time. In other words, given any two time-like separated events and a connecting time-like trajectory, there always exists an inertial observer for whom the initial and final events are at the same place. Of all the trajectories that can connect these two events, the one with the longest proper time is the straight one since all others will have some contributions from segments with spatial contributions which will reduce the proper time below that of the coordinate time difference.

Another interesting question is what is the nature of a coordinate system that would be developed by Harry. In other words, how would Harry describe the events around him? Until Harry met Dorothy, he was an inertial observer. The co-mover with him is always inertial and obeys all the requirements to establish a valid coordinate system. This co-mover could develop a coordinate system that is a Lorentz transform of Sally's, and this would be the same as Harry's at least until he turns around. Once he turns around, he is again inertial and has a new co-mover who could be used to establish a coordinate system. If we argue that this dual co-mover coordinate system is the coordinate system that Harry would use, there is an anomaly associated with the events that are between the lines of simultaneity before and after the turn around. If we consider the turn around to be instantaneous, these events must all be at the same time, see Figure 12.13

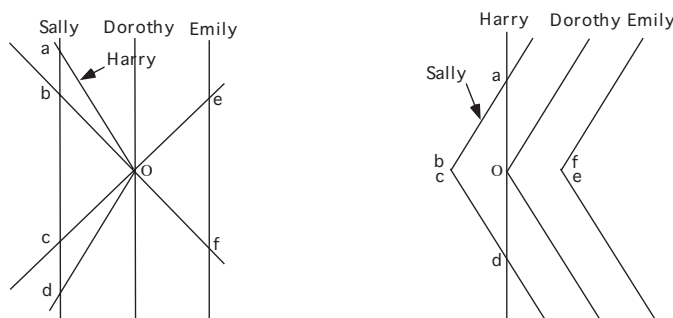


Figure 12.13: **Harry's Turn Around** At event d , Harry leaves Sally moving relative to her to increasing x at a speed of $\frac{3}{5}c$. On reaching Dorothy who is co-moving with Sally and one light year away according to them, the event o , Harry turns around and starts back toward to Sally, again, at $\frac{3}{5}c$, meeting her at event a . The line eoc is the line of simultaneity to a Harry co-mover just before turning around and the line bof is the line of simultaneity to a Harry co-mover just after turning around. These co-movers are inertial and thus can develop consistent coordinatizing of space-time. If Harry were to define his coordinate system as that of the co-mover with him before turn around up until turn around and that of the co-mover moving with him after turn around after turn around, since the turn around takes place very quickly all the events within the cones boc and eof are basically at the time that he would label as $\frac{4}{3}$ year. This difficulty is not relieved by spreading the turn-around out in time. Although Sally's events b and c would now be separated and in the correct order, a co-mover with her such as Emily who is one light year further from her than Dorothy would have her events e and f recorded in a sequence the opposite from that which she records.

on page 300. In other words, events on Sally's world-line that are between the events labeled b and c in Figure 12.13, and are clearly separated in time are all recorded by Harry as at the same instant. This problem is not solved by having the turn-around spread over a small but non-zero interval of time. Although the events on Sally's world-line are now separated and reasonable, the events on a world-line of another inertial observer, say an Emily who was co-moving with Sally and Dorothy and further from Sally than Dorothy on the same side as Dorothy, would have events on her world-line that are not in the correct temporal order; those that Emily say occurred earlier are timed later by Harry.

Another approach to coordinatizing by Harry could be with comoving

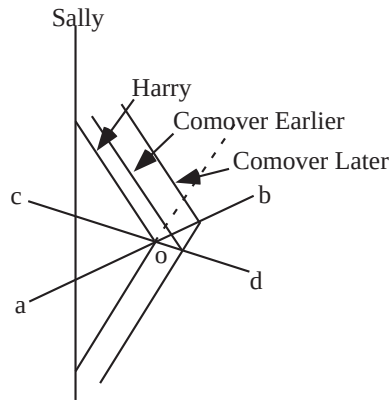


Figure 12.14: **Cohort Coordinatizing by Harry** Harry has a cohort that initially is a distance L from him. The problem of coordinatizing with the cohort is that the cohort does not have a determined time to accelerate. There are three choices. The cohort can accelerate simultaneously with Harry at the event that is simultaneous with Harry's acceleration event just after acceleration, shown as comover earlier above, or simultaneous with Harry's acceleration just after, shown as comover late. As is clear from the figure, the cohort separation from Harry is no longer L . For the case of the cohort earlier, the separation is $\sqrt{L^2 - c^2 \Delta t^2}$.

confederates. Of course, he if he uses two sets of confederates, he will reproduce the situation described above. If he requires his confederates to actually reproduce his motion, have an acceleration, there are several immediate ambiguities. The first is what event on the confederate's worldline should be the turn-around event? Using Figure 12.14 on page 301. There are two obvious choices; the line of simultaneity before and after turn-around. Using either of these places the cohort that was at a distance L

Chapter 13

Paradoxes of Relativity

The conceptual complications of the special theory of relativity are often expressed through stories whose outcomes are counter intuitive, paradoxes. The following three are a representative sample.

13.1 The Twin Paradox

13.1.1 The Problem

Alphonse and Gaston are twins and they are authors. Alphonse writes advertising copy and has to travel to town every day and Gaston writes novels and stays home. Each day when Alphonse is on the train going to town he is observed by Gaston. Due to their relative motion, Gaston sees Alphonse's clock running slower and thus Alphonse is aging slower than he does. At the end of the day, when Alphonse has returned home he has not aged as much as Gaston and is therefore younger. The problem is that, during the trip, Alphonse observes Gaston. He notes that Gaston's clock is the one that runs slow. He expects that, when they get back together, Gaston will be younger. When they get back together are they the same age? If there is a difference in their ages, who is younger. The clue to the problem is that Alphonse spills a drink on his shirt every day.

13.1.2 The Solution

Actually, we have already solved this paradox. This is Harry and Sally of Section 11.3.5 on page 265 and 12.7 on page 299. The supposed paradox here is that it seems that Alphonse and Gaston are identical. Not only are they twins but they both see the others clock run slow. The simple

fact is that they are not identical. This clue is the answer to the paradox; Alphonse spills his drink on his shirt because he is accelerated. Gaston never spills his drink; he is not accelerated. Acceleration is knowable, velocity is not, see Section 9.2 on page 215. Now that we understand that they are no longer identical, one was accelerated, they can be different and it can be that one can now be older than the other one when they get back together. From Section 12.7 on page 299, since the straight line time-like trajectory is the longest, the non-accelerated twin is always the oldest. The exact age difference can be computed from the trajectories of each twin in any convenient frame. The example of Harry and Sally, Section 11.3.5 on page 265, is straight forward.

13.2 The Boy in the Barn

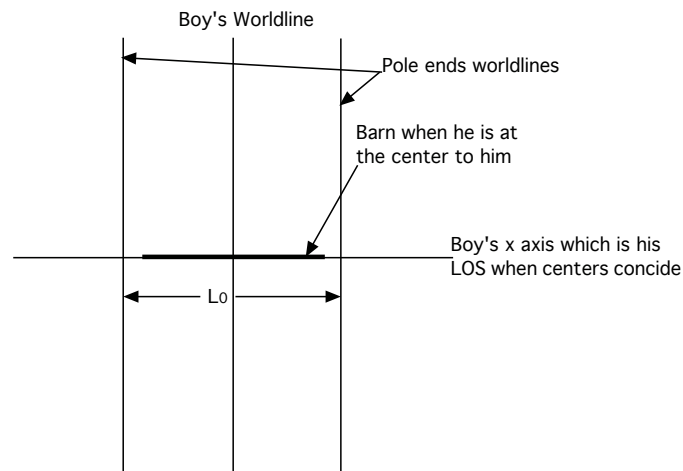
13.2.1 The Problem

A boy is a pole vault freak. He runs around a track all day to practice. He also can run very fast, $\frac{3}{5}c$. He has to pass through a barn. In fact, the pole that he practices with is taken from the roof beam of the barn and is the same length as the barn when they are at rest together. He practices all day and his parents worry about him. They want to stop him and make him come in for dinner. He agrees that, if he and his pole are ever entirely in the barn, they can close the front and back doors. His reasoning is that since his pole is much longer than the barn, there is no problem. They will never get him. They agree to do as he says. The clue here is that parents are always right. Do they get him?

13.2.2 The Solution

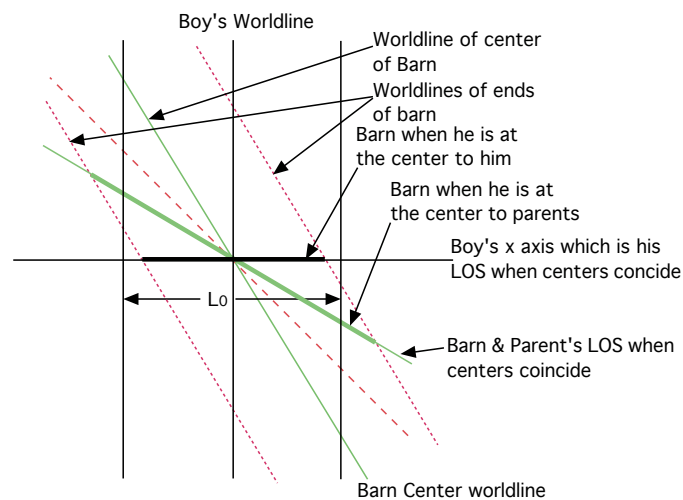
This is a classic example of the problem of simultaneity for space-like event intervals, Section 10.3 on page 233. Let's draw a space time diagram from the boy's perspective to understand his argument. The rest length of the pole and the barn are L_0 . He says that barn length is $\sqrt{1 - \left(\frac{3}{5}\right)^2} L_0 = \frac{4}{5} L_0$ which means that if, he measures it when he is at the center of the barn, his assertion is clearly true; both ends of his pole are outside the barn when he

is in the center.



The ends of the pole's worldlines are the thin vertical black lines.

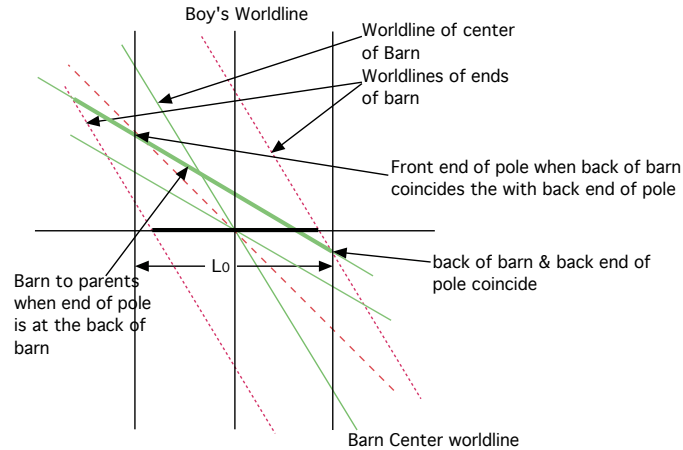
Now, let's add the information about barn's worldlines to this. The center of the barn and boy worldlines meet at the origin. The worldline of the center of the barn is the thin green line sloped at $\frac{5}{3}$ with similar parallel lines with shown in dotted red at the same slope for the ends of the barn.



A light line passing through the origin is dashed red. Also shown by a thin green line is the LOS for the parents for the event of the coincidence of the center of the pole and the center of the barn and a thick green line between the ends of the barn for the barn at the time of this coincidence for the

parents. It is clear the pole is entirely inside the barn to them; two pole end worldlines intersect the barn, the thick green line.

The parents can close the barn doors when the back end of the pole arrives at the back of the barn since the front of the pole is well short of the end of the barn at that time to them.



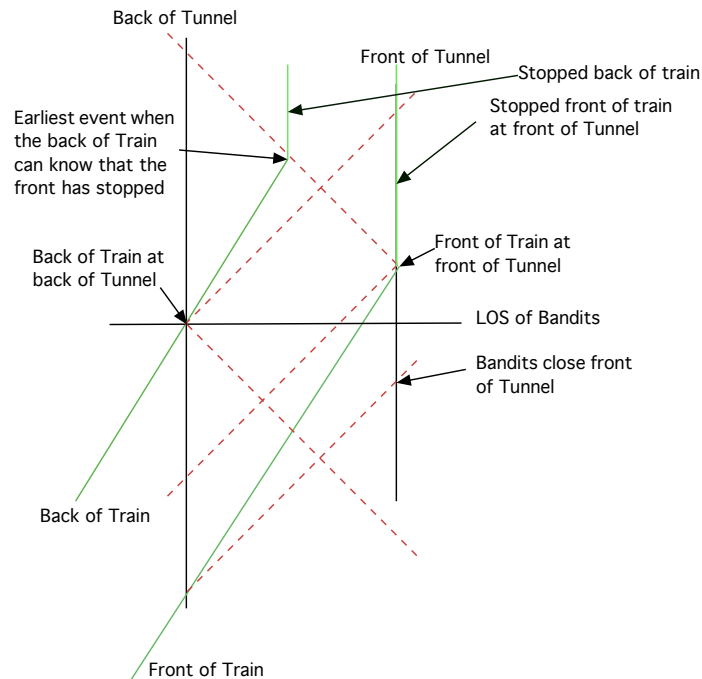
13.3 The Bandits and the Bullet Train

13.3.1 The Problem

The Japanese bullet train travels at $\frac{3}{5}c$. It carries gold bullion from Osaka to Tokyo. It passes through a tunnel that is exactly as long as the train when each are measured at rest. A group of bandits want to capture the train and get the gold. In their frame, the train is shorter than the tunnel and thus can be captured inside of it. The owners of the train know about the tunnel and are concerned about the gold. They decide that, since the train is longer than the tunnel, it will be sufficient to post a guard at the front and the back of the train and thus there will always be one of the guards outside the tunnel. The bandits attempt to capture the gold. They have two ways to accomplish this. One is to rely on the train schedule and close the tunnel gate at the far end from the train right after the the train back enters the tunnel. Another is to wait at the entrance to the tunnel and when the back of the train passes close the back and send a signal to the baddies at the front exit end to close the gate. The train people are not worried about a capture of the train since they can put guards at both ends of the train and there is never a time when the entire train is in the tunnel. Will

13.3.2 The Solution

It is more realistic to do the capture in the other scenario. The bandits wait until the front of the train is at the back of the tunnel. They signal the bandits at the front of the tunnel to close the entrance to the tunnel. Later, because the train is traveling at less than the speed of light, the front of the train arrives at the front of the tunnel. It crashed to a stop. This is after the back of the train is at the back of the tunnel. Thus any guard at the end of the train is captured in the tunnel and there is no chance to send an alarm. Bandits win. All this is shown below.



13.3.3 Consequences*

It is interesting to speculate on the fate of the train. Let's get concrete or, at least, as concrete as makes sense in this context. As usual, we will 'Spherical Cow' the problem. One extreme version of a train is like a linked chain that can only be pulled; In physics language, a continuous system which can only sustain a tension stress. You cannot push on the front end of a chain, you can only pull. Pushing piles the links into a mess. The other is to treat the train as a system of connected cars that are themselves rigid finite lengths. A system of these can sustain a compression or extension stress. This is generally called an elastic solid rod.

Use a rest length for the train as L_0 . It's length to the bandits is $\frac{4}{5}L_0$. Using the event of the back of the train at the back of the tunnel as the origin event. The front of the train is at the LOS of the bandits at the event labeled $(\frac{4}{5}L_0, 0)$. This makes the trajectory of the front of the train until it stops at the gate, $ct = \frac{5}{3}(x - \frac{4}{5}L_0)$. The front of the train therefore comes to the front of the tunnel at $x_{front} = L_0$ or $ct_{front} = \frac{5}{3}(1 - \frac{4}{5})L_0 = \frac{1}{3}L_0$. I am assuming the front of the train comes abruptly to a stop at the front of the tunnel. The earliest that the back of the train can know about the stopping event of the front of the train is from a light signal. The trajectory of the light going to the back of the train is $ct = -x + L_0 + \frac{1}{3}L_0$. The trajectory for the back of the train is $\frac{3}{5}ct = x$. The coincident event for these trajectories is $(\frac{1}{2}L_0, \frac{5}{6}L_0)$ and we anticipate abrupt stopping. To the bandits at the time the back of the train stops, it is only half as long as it was and now it is comoving with them. How do you get a train to shrink to half its size? Two answers depending on what you think a train is. If it is a chain, it is a large mess of links. The inertia of the links cause them to pile on top of each other or crush the links. If it is an elastic rod, the rod would experience a compression wave from the abrupt stopping at the front. This wave would propagate back along the rod at the speed of sound. The speed of sound in solids is much lower than the speed of light. In fact each part of the train is forced to stop faster than this process allows. This is what is called a shock wave. The shock wave front creates high temperatures and will actually leave the rod as a pool of molten stuff; another way to say a mess. Another way to say this is that interatomic forces that were responsible for us to call it a solid or a set of links cannot do the job of maintaining the properties that we ascribe to the train. Of course, all of this assumes the gate at the front of tunnel can stop the train. It is clearly a very strong gate.

Chapter 14

Uniform Acceleration

14.1 Events at the same proper distance from some event

Consider the set of events that are at a fixed proper distance from some event. Locating the origin of space-time at this event, the equation for this set of events is:

$$x^2 - c^2t^2 = d^2 \quad (14.1)$$

The parameter, d , is the proper distance of these events from the origin event. The origin event and the events on the curve are related by this distance d and thus for the set of events on the curve the origin is called the magic event and d is the distance from the magic event to the curve.

In space-time, this is a two branch hyperbola with light cones emanating from the origin as the asymptotes. If we now consider only the branch that has $x > 0$, $x = \sqrt{d^2 + c^2t^2}$, we have a single curve. In Figure 14.2 on page 311, We plot several of them for different d . Since this equation is a form invariant under the Lorentz transformations, all inertial observers will have the same curve and Lorentz transformations will map points on the curve to points on the curve.

By locating a light cone on the event at $(d, 0)$, we can see that all the events on the curve at later times are in the future; the curve is monotonically asymptotic to a light cone that is later in space-time. Thus all the events at later times on the curve are in the future of $(d, 0)$. Similarly all the events that are before $t = 0$ are in the past of $(d, 0)$. Thus the curve is time-like and is therefore a candidate for the motion of a material particle. In the next section, we will see that this is the trajectory of the uniformly accelerated object.

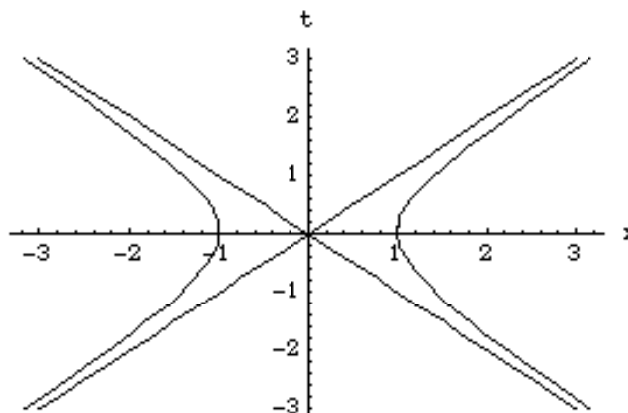


Figure 14.1: The locus of events that are at the same proper distance from the origin. These curves all points of which are elsewhere related to the origin are pairwise time-like related.

14.2 Uniformly accelerated motion

Since this curve is time-like, it is a possible state of motion for a material particle. It is certainly a case of motion that is not uniform, not a straight line in space-time. For any observer in uniform motion, an object following this trajectory will appear to be approaching at a very rapid rate, almost $-c$, and slowing down until at some event it is as close as it will ever get and at rest with respect to the observer and then moving away so that at long times later it is receding at almost c .

Since the Lorentz transformations are homogeneous and linear, lines through the origin are transformed into lines through the origin and space-like lines are transformed into space-like lines and similarly for time-like lines. Thus if you pick an event, say (x_0, t_0) , on this curve, the line through it and the origin which is space-like can be transformed to the space-like line through $(d, 0)$ and $(0, 0)$ by the Lorentz transformation with $v = c^2 \frac{t_0}{x_0}$. This is also the transformation that brings the tangent to the curve to the vertical which means that the instantaneous relative velocity at (x_0, t_0) is v . Or said another way, an observer with relative velocity, $v = c^2 \frac{t_0}{x_0}$, is a comover to the this trajectory at the event (x_0, t_0) . Thus we see that the instantaneous relative velocity at (x_0, t_0) is $v = c^2 \frac{t_0}{x_0}$. More significantly, to the respective comovers, the acceleration at (x_0, t_0) is the same as the acceleration at $(d, 0)$. Therefore, as measured by comovers, the instantaneous acceleration at any event is the same and this is the acceleration that the object experiences in

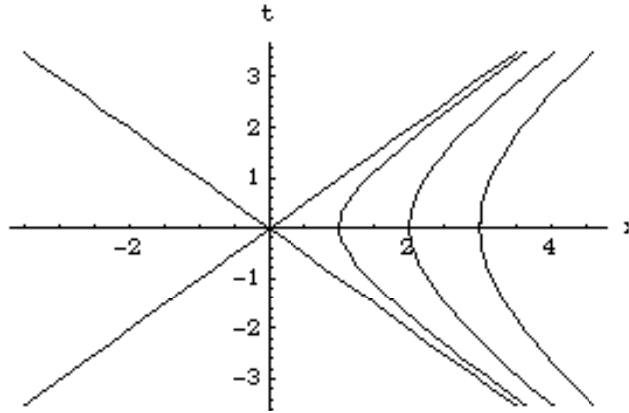


Figure 14.2: The locus of events with $x > 0$ that are at the same proper distance from the origin for different values of the proper distance, d .

its motion. On simple dimensional grounds, the acceleration at the event $(d, 0)$ must be

$$a = \frac{c^2}{d}. \quad (14.2)$$

Also note that it follows from the previous argument that the line from (x_0, t_0) to the origin is the line of simultaneity for the commover at the event (x_0, t_0) .

14.2.1 Details of the calculation of the acceleration

The easiest way to calculate the acceleration is to use calculus. For a uniformly accelerated curve the instantaneous co-mover's velocity is

$$\begin{aligned} \frac{dx}{dt} &= \frac{d}{dt}(\sqrt{d^2 + c^2 t^2}) \\ &= \frac{1}{2} \times \frac{2c^2 t}{\sqrt{d^2 + c^2 t^2}} \\ &= c^2 \frac{t}{x} \end{aligned} \quad (14.3)$$

which we already knew. The acceleration is

$$\frac{d^2 x}{dt^2} = \frac{d(c^2 \frac{t}{x})}{dt}$$

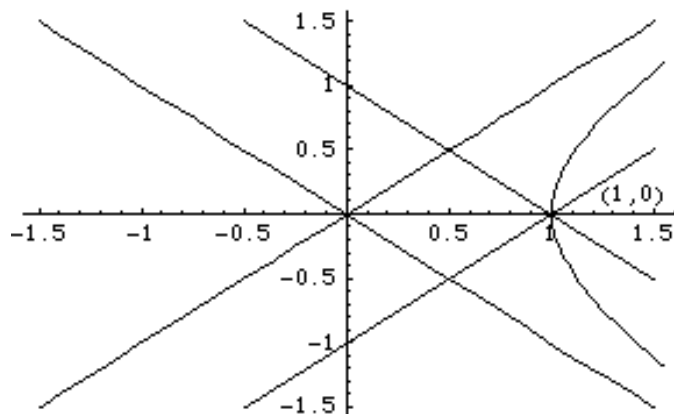


Figure 14.3: Placing a light cone at the event $(1, 0)$ shows that the locus of events with $x > 0$ that are at the same proper distance, $d = 1$, from the origin is a timelike trajectory.

$$\begin{aligned}
 &= c^2 \left(\frac{1}{x} - \frac{t}{x^2} \frac{dx}{dt} \right) \\
 &= c^2 \left(\frac{1}{x} - c^2 \frac{t}{x^2} \times \frac{t}{x} \right) \\
 &= c^2 \left(\frac{x^2 - c^2 t^2}{x^3} \right) \\
 &= c^2 \frac{d^2}{x^3}
 \end{aligned} \tag{14.4}$$

which, at the event $(d, 0)$, means that $v = 0$ and $a = \frac{c^2}{d}$, which was our result from dimensional arguments in Equation 14.2 on page 311.

If you have an aversion to calculus, you can look at the motion for small times near the event $(d, 0)$. It must reduce to the expression for the position for the constant acceleration that we know from classical physics, $x_{cl}(t) = x_0 + v_0 t + \frac{a}{2} t^2$ which should be valid for $\frac{at}{c} \ll 1$. Expanding our $x(t)$ for small t and using the fact, $(1 + x)^n \approx 1 + nx$ for $x \ll 1$, that everyone should know from Section 1.4.2 on page 16, we have

$$\begin{aligned}
 x(t) &= \sqrt{d^2 + c^2 t^2} \\
 &= d \sqrt{1 + \frac{c^2}{d^2} t^2} \\
 &\approx d \left(1 + \frac{c^2}{2d^2} t^2 \right).
 \end{aligned} \tag{14.5}$$

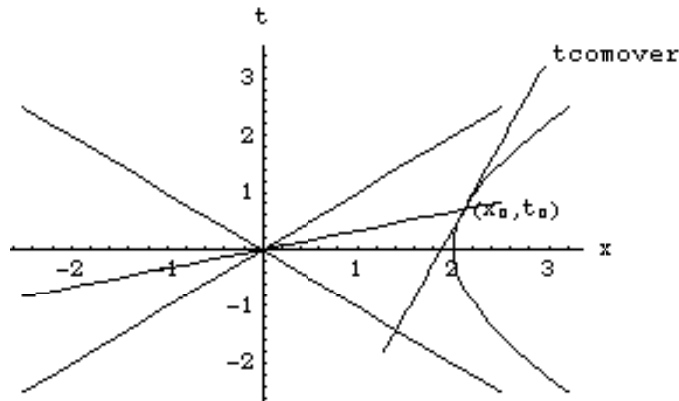


Figure 14.4: The uniformly accelerated observer with the world line and the line of simultaneity of the commover for the event (x_0, t_0) .

Comparing this with $x_{cl}(t)$, we see that, for small times near the event $(d, 0)$, the velocity is 0 and the acceleration is $\frac{c^2}{d}$, again our result Equation 14.2 on page 311.

It is very important to point out that this is the acceleration that the accelerated object “feels”. Consider an accelerated rocket with a pair of identical springs and masses, one mass-spring system mounted on a frictionless surface horizontally and the other mass-spring suspended vertically. Vertical in the rocket is along the line from front to back and horizontal is one of the transverse directions. We also calibrate our springs so that we know the force that is required to stretch them a given amount, i. e. we know the spring constant, k , of the springs. The horizontal mass-spring will have one equilibrium position and the vertical one will have a different one. If we now carefully adjust the thrust of the rocket so that the stretch of the springs does not change with time, our rocket when observed by someone who was initially at rest with us will register it at $x(t) = \sqrt{d^2 + c^2 t^2} - d$ where $d = \frac{c^2}{\frac{k \times \text{stretch}}{m}}$ where m is the mass, k is the calibrated spring constant, “stretch” is the difference in the length of the vertical and horizontal springs. The extra d is in $x(t)$ to make the rocket and the original commover coincident in space at $t = 0$. At later times, the rocket has moved away from the original commover but the mass-spring system still measures the same acceleration, the acceleration that is measured by the new instantaneous commover.

This is another case of a term which is dimensionally the same but whose

physical interpretation is different. Acceleration is generally defined kinematically as $a_k \equiv \frac{d^2x(t)}{dt^2}$. Through Newton's laws, we have an equivalent definition in the form $a_s \equiv \frac{f}{m}$ where f is the effect of external objects on a body of mass m . It is this a_s that is "sensed" by the accelerated system that informs it that it is not inertial. This is the essence of Galilean invariance. A free body has no acceleration. The equality of a_s and a_k expressed in Newton's law can be required only in the case of a world of low relative velocities. Since the kinematic definition is not a constant in this motion although the sensed acceleration is constant, we have an interpretation problem. It is required that all inertial observers of this motion agree on its sensed acceleration and from the previous discussion all events on the trajectory have the same sensed acceleration to a local commover and this acceleration is the same as the kinematic one as evaluated or measured for small times around the events when the object is commoving with that observer. For all other times, the kinematic and sensed acceleration are different. The kinematic acceleration is the acceleration evaluated by one of the commover inertial observers for all time and it varies from $\frac{c^2}{d}$ the small time value to zero at large times when the object is distant. The kinematic acceleration is $\frac{d^2}{dt^2}x(t)$ where both x and t are coordinates for the specific inertial commover. An alternative might be to call this motion not uniformly accelerated motion but uniformly effected motion.

14.3 The proper time along the trajectory

As was stated in Section 12.3 on page 280, the proper time between two events is a trajectory dependent concept. As the accelerated object moves along its trajectory, its coordinate position and time are given by $x(t) = \sqrt{d^2 + c^2t^2}$. This same motion can be conceived of as both x and t both evolving as a function of the proper time, $x(\tau)$ and $t(\tau)$. Our problem is to find these relationships. Noting that because of the definition of the trajectory as the locus of events with the same proper distance from the origin event that for all τ the two functions $x(\tau)$ and $t(\tau)$ satisfy $(x(\tau))^2 - c^2(t(\tau))^2 = d^2$.

14.3.1 Timelike Trajectories and Accelerated Motion

Although it does not constitute a proof, we can use accelerated motion to justify the often heard comment that there is no force that can boost a material particle to speeds greater than the speed of light. As stated in

Section 14.2.1 on page 311, the acceleration a that labels this trajectory is the acceleration that a material particle moving along that world line “feels.” In other words, the force that accelerates the particle to move it along this trajectory is a constant as measured by the sequence of commovers and these are the suitable observers of the force of acceleration. In this case of constant force, we see that no matter how long the force operates, the velocity of the particle that is subjected to this force moves relative to its initial velocity at a speed that is less than c ; the trajectory remains timelike for all times. Also in any finite time interval, there is no acceleration and thus no force that can change the trajectory from timelike to space like.

14.4 Examples using accelerated motion

With the tools developed in the previous sections, we can now analyze all kinds of simple uniform acceleration problems. In fact, just about any of the usual uniform acceleration problems that are encountered classical physics can be studied. In this section, I will go through the details of three typical problem types.

14.4.1 Deceleration

Sally is moving toward a wall with a relative speed of $\frac{3}{5}c$. When she is one lightyear away from the wall, she decides to decelerate. What is the minimum deceleration that she can use so that she just comes to rest at the wall?

We can find the answer in the frame in which the wall is at rest. Firstly, we should diagram the motion.

From this we can see that the problem can be stated in a simpler fashion. At any event, (x_0, t_0) , on the uniformly accelerated trajectory, we know the relative velocity at that point, $\frac{v}{c^2} = \frac{t_0}{x_0}$. For the case shown in Figure 14.5 on page 316, note that t_0 is negative and x_0 is positive so that v is negative. Thus we can ask given an acceleration, a , how far from the event (x_0, t_0) on that trajectory is the vertex of the hyperbola? In the usual coordinate system, the vertex is at $(d, 0)$ and thus the stopping distance for that case is $\delta = x_0 - d$. Remember the d is related to the acceleration, a , as $d = \frac{c^2}{a}$. The event (x_0, t_0) satisfies $t_0 = \frac{v}{c^2}x_0$ where v is the relative velocity at that event and $x_0 = \sqrt{d^2 + c^2 t_0^2}$ or $x_0 = \frac{d}{\sqrt{1 - \frac{v^2}{c^2}}}$. Thus the general formula for

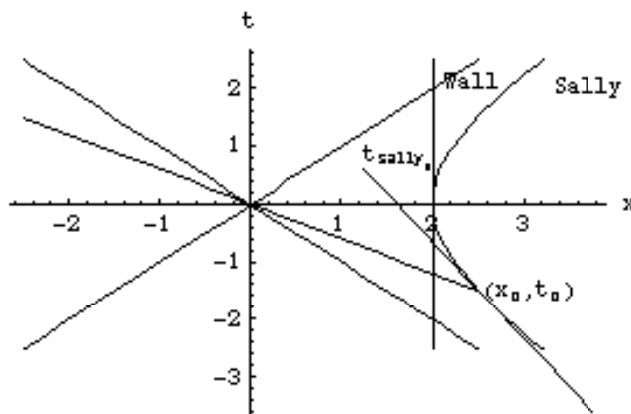


Figure 14.5: Sally turning from the wall. The event (x_0, t_0) is the event at which she decelerates. The line labeled “Sally” is her trajectory. The line labeled “ t_{Sally_0} ” is her worldline before decelerating

the stopping distance for a given velocity and acceleration is

$$\delta = d \left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1 \right) = \frac{c^2}{a} \left(\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1 \right).$$

The next problem is to decide what δ is. From the problem setting, I would argue that the one light year distance is the coordinate distance in her frame at the instant that she starts the acceleration. This δ is the distance in the wall’s frame. This is not Sally’s distance. That distance is the proper distance between the event (x_0, t_0) and the intersection of the line of simultaneity of the commover at (x_0, t_0) and the worldline of the wall. The equation of the line of simultaneity is $\frac{t-t_0}{x-x_0} = \frac{v}{c^2}$ and the line of the wall is $x = d$. The event at the intersection of these two lines is $(d, \frac{v}{c^2}(d-x_0) + t_0)$ and the proper distance between this event and the event at the start of the acceleration, (x_0, t_0) , is $\sqrt{1 - \frac{v^2}{c^2}}(x_0 - d)$. Calling her distance to the wall δ' , we now have $a = \frac{c^2}{\delta'} \left(1 - \sqrt{1 - \frac{v^2}{c^2}} \right)$.

How does this compare to the classical result, stopping distance $= \frac{v^2}{2a}$? From “Things”, Section 1.4.2 on page 16, for large c , $\sqrt{1 - \frac{v^2}{c^2}} \approx 1 - \frac{v^2}{2c^2}$. Plugging this in we have the classical result exactly.

For our specific problem, we have $v = -\frac{3}{5}c$ and $\delta' = 1$ ltyr and $a = \frac{1}{5} \frac{\text{ltyr}}{\text{yr}^2}$ or $2 \frac{\text{m}}{\text{s}^2}$.

14.4.2 Accelerated Rocket

A rocket of length $\frac{1}{2}$ lightyear is accelerated at a constant acceleration of $\frac{1}{2} \frac{\text{lightyear}}{\text{year}^2}$. At $t = 0$, the rocket starts to accelerate. When a clock at the bottom reads a time τ_{bottom} , what is the time for a clock in the top of that rocket?

Again, we have to determine what is being told to us in the problem. We have to decide where the parts of the rocket are, i. e. their world lines. The top of the rocket is rigidly connected to the bottom so that as the rocket accelerates the distance as measured from the bottom of the rocket to the top is unchanged. Under stress but unchanged. The world line of the bottom which is accelerating at a rate a_{bottom} in the standard coordinate system is

$$\begin{aligned} x(\tau_{\text{bottom}}) &= \frac{c^2}{a_{\text{bottom}}} \cosh\left(\frac{a_{\text{bottom}}}{c} \tau_{\text{bottom}}\right) \\ t(\tau_{\text{bottom}}) &= \frac{c^2}{a_{\text{bottom}}} \sinh\left(\frac{a_{\text{bottom}}}{c} \tau_{\text{bottom}}\right) \end{aligned} \quad (14.6)$$

or, using $d = \frac{c^2}{a_{\text{bottom}}}$, where d is the proper distance from the origin event, $(0, 0)$, to any event on the world line,

$$\begin{aligned} x(\tau_{\text{bottom}}) &= d \cosh\left(\frac{c}{d} \tau_{\text{bottom}}\right) \\ t(\tau_{\text{bottom}}) &= d \sinh\left(\frac{c}{d} \tau_{\text{bottom}}\right). \end{aligned}$$

Note that the commover to any event, $(x(\tau_{\text{bottom}}), t(\tau_{\text{bottom}}))$, has a line of simultaneity that goes from that event through the origin event, $(0, 0)$.

A second set of events that are all at a proper distance $d + h$ from the origin event, $(0, 0)$, see Figure 14.6 on page 318 would be at

$$\begin{aligned} x(\tau_{\text{top}}) &= (d + h) \cosh\left(\frac{c}{d + h} \tau_{\text{top}}\right) \\ t(\tau_{\text{top}}) &= (d + h) \sinh\left(\frac{c}{d + h} \tau_{\text{top}}\right). \end{aligned}$$

Also since the lines of simultaneity are the lines through the origin event, the distance between these world lines when measured by the commover at the bottom of the rocket is h . The trajectory of the top of the rocket is

$$\begin{aligned} x(\tau_{\text{top}}) &= \frac{c^2}{a_{\text{top}}} \cosh\left(\frac{a_{\text{top}}}{c} \tau_{\text{top}}\right) \\ t(\tau_{\text{top}}) &= \frac{c^2}{a_{\text{top}}} \sinh\left(\frac{a_{\text{top}}}{c} \tau_{\text{top}}\right) \end{aligned} \quad (14.7)$$

Thus these are the world lines of the top and the bottom of the rocket.

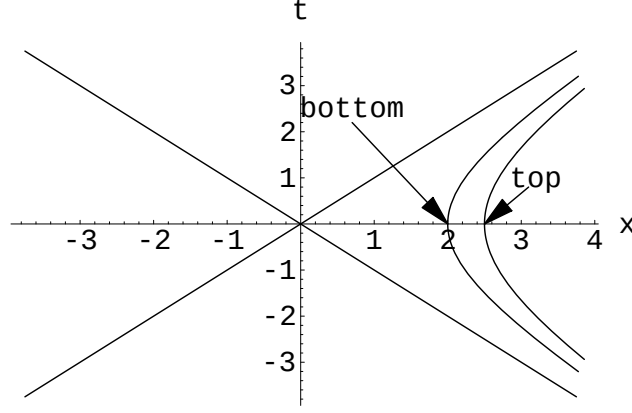


Figure 14.6: **The world lines of the top and bottom of an accelerating rocket.** The bottom of the rocket has an acceleration of $\frac{1}{2} \frac{\text{lightyear}}{\text{year}^2}$. The top of the rocket is at a distance $\frac{1}{2}$ lightyear from the bottom.

We see immediately that the top of the rocket does not have the same acceleration as the bottom. Using $d = \frac{c^2}{a_{\text{bottom}}}$, we get that

$$a_{\text{top}} = \frac{a_{\text{bottom}}}{1 + \frac{ha_{\text{bottom}}}{c^2}}. \quad (14.8)$$

In Figure 14.6 on page 318, we also see that, since the world lines of the top and the bottom of the rocket share the same asymptotes, the hangle to the line of simultaneity to any event is the same and thus that

$$\phi = \frac{c\tau_{\text{bottom}}}{d} = \frac{c\tau_{\text{top}}}{d + h}$$

or writing this in terms of the accelerations of the rocket,

$$\phi = \frac{a_{\text{bottom}}\tau_{\text{bottom}}}{c^2} = \frac{a_{\text{top}}\tau_{\text{top}}}{c^2}$$

or

$$\tau_{\text{top}} = \left(1 + \frac{ha_{\text{bottom}}}{c^2}\right)\tau_{\text{bottom}}. \quad (14.9)$$

Thus, clocks at the top and bottom of a rocket run at different rates. This situation can be made a little more baffling by noting that although the top and bottom of the rocket have clocks that run at different rates, the top and bottom share the same lines of simultaneity. They just differ about the time of these simultaneous events.

14.4.3 John Bell's Problem

The next example is the problem of two identical rockets and John Bell's Problem. Although I am not able to vouch for this story directly, I have been told the following fascinating story about John Bell. Yes, the same John Bell of Bell's Theorem, see Chapter 21 on page 471. When a new theoretical physicist would come to the world famous laboratory, CERN, where Bell was employed, Bell would go to lunch room and look up the new person and, as a part of the getting-to-know-you chit chat, ask the new person the following question: If two identical coasting rockets were connected by a string and the rockets then given identical uniform accelerations would the string between them break after some time?

Without making a careful analysis, usually without even thinking about it carefully, the unsuspecting innocent would quickly answer that the string would not break. The quick argument being that, if the two rockets were moving at the same velocity originally and had identical accelerations, they would always stay the same distance apart. We are now enough informed about the interesting effects of relativity and particularly uniform acceleration in special relativity to be a little more careful. If identical clocks at the top and bottom of a rocket can drift apart in time, then it is plausible that identical rockets can begin to separate, see Section 14.4.2 on page 317 above. The proof that the string will break is easily shown graphically, see Figure 14.7 on page 320.

Well, at least in principle, it is simple even if the figure is rather complex. Two identically uniformly accelerated rockets have trajectories that are shifted from each other. Consider two rockets that are separated by a distance h and have an acceleration a , their trajectories are

$$\begin{aligned} x_{top} &= \sqrt{c^2 t^2 + \left(\frac{c^2}{a}\right)^2}, \\ x_{bottom} &= \sqrt{c^2 t^2 + \left(\frac{c^2}{a}\right)^2} - h, \end{aligned} \quad (14.10)$$

where the top rocket is the one to the side of the acceleration.

The end of a string of length h suspended from the top rocket has the trajectory

$$x_{string} = \sqrt{c^2 t^2 + \left(\frac{c^2}{a} - h\right)^2}. \quad (14.11)$$

It is clear that the $x_{string} - x_{bottom} > 0$ for all t . In fact, we can easily calculate the separation for small $\frac{ah}{c^2}$, physically not an unreasonable criteria

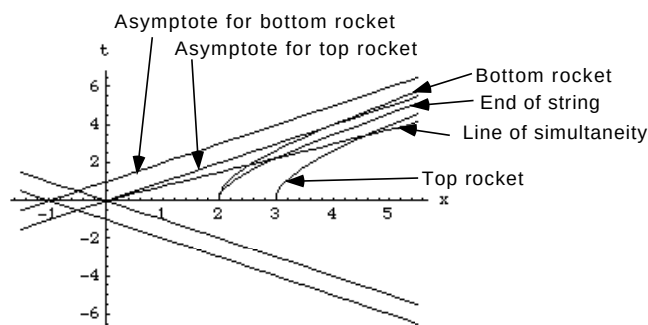


Figure 14.7: **John Bell's Problem** Two identical rockets have trajectories that follow each other. We define bottom and top as in the earlier example, Section 14.4.2 on page 317, by the direction of the acceleration. If a string is suspended from the top rocket that just reaches the bottom rocket at $t = 0$, it will have the trajectory shown. Since the end of the string moves so that it is a fixed distance from the top rocket as measured by the top rocket, it shares the same asymptote as the top rocket. The bottom rocket has a different asymptote and, in fact its trajectory crosses the top rockets asymptote. Thus it is clear that it is further than the end of the string from the top rocket. Since the string and top rocket share the same line of simultaneity, you can see along that line that at any time t to the top rocket the bottom rocket is further than the end of the string. The parameters for this figure were $a = \frac{1}{3} \frac{\text{lyrs}}{\text{yr}^2}$, $h = 1 \text{ lyr}$.

for the size of the rocket and the acceleration. In this limit and after a couple of applications of the result from “Things”, Section 1.4.2 on page 16 and some rather tedious algebra,

$$x_{string} - x_{bottom} = h \left(1 - \frac{1}{\sqrt{1 + \frac{a^2 t^2}{c^2}}} \right). \quad (14.12)$$

Again, it is clear that this is positive for all t . The problem is that this is not the length of interest if the question is when the string will break. Equation 14.12 on page 320 is the separation of the end of the string and the bottom rocket to the original co-mover at some time t according to that inertial observer's clock. We really want the distance the string realizes at any time τ to the string. Of course, we realize from the previous example, Section 14.4.2 on page 317, that different parts of the string have different times. Fortunately though, the elements of the string all share the same

line of simultaneity and it is, of course, the same as that of the top rocket. This quandary about clocks along accelerated systems will be examined in more detail in the Section 14.5 on page 322 where we discuss the problem of allowing an accelerated observer to create a coordinate system. It is also discussed in the development of General Relativity on the implications of the Equivalence Principle, see Section 16.4 on page 350.

Using as our time, the time τ of the top rocket, we can determine the events at the end of the string and bottom rocket that are simultaneous with τ on the top rocket. The equation for the line of simultaneity to the top rocket for any event, (x_0, t_0) , and the string at a time τ on the top rocket is

$$\frac{t}{x} = \frac{t_0}{x_0} = \tanh\left(\frac{a\tau}{c}\right) \quad (14.13)$$

and the event at the end of the string simultaneous with τ at the top rocket is

$$\begin{aligned} x_{string_\tau} &= \left(\frac{c^2}{a} - h\right) \cosh\left(\frac{a\tau}{c}\right) \\ t_{string_\tau} &= \left(\frac{c^2}{a} - h\right) \frac{\sinh\left(\frac{a\tau}{c}\right)}{c}. \end{aligned} \quad (14.14)$$

The event on the bottom rocket trajectory that is simultaneous to the string and the top rocket satisfies

$$(x_{bottom_\tau} + h)^2 - \left(\frac{c^2}{a}\right)^2 = \tanh^2\left(\frac{a\tau}{c}\right) x_{bottom_\tau}^2. \quad (14.15)$$

Of the two roots of this equation, the physically acceptable one yields

$$x_{bottom_\tau} = \left[\sqrt{\left(\frac{c^2}{a}\right)^2 - \left(\frac{c^2}{a}\right)^2 \tanh^2\left(\frac{a\tau}{c}\right) + \tanh^2\left(\frac{a\tau}{c}\right) h^2} - h \right] \cosh^2\left(\frac{a\tau}{c}\right) \quad (14.16)$$

with the t_{bottom_τ} given by

$$t_{bottom_\tau} = \tanh\left(\frac{a\tau}{c}\right) \frac{x_{bottom_\tau}}{c}. \quad (14.17)$$

The stretch of the string, δ , is the proper distance between the events at the end of the string and the bottom rocket,

$$\begin{aligned} \delta &= \sqrt{(x_{string_\tau} - x_{bottom_\tau})^2 - c^2 (t_{string_\tau} - t_{bottom_\tau})^2} \\ &= (x_{string_\tau} - x_{bottom_\tau}) \sqrt{1 - \tanh^2\left(\frac{a\tau}{c}\right)} \\ &= (x_{string_\tau} - x_{bottom_\tau}) \cosh^{-1}\left(\frac{a\tau}{c}\right). \end{aligned} \quad (14.18)$$

Plugging in for $x_{string\tau}$ and $x_{bottom\tau}$, and doing considerable algebra and using the hyperbolic function identities,

$$\delta = \frac{c^2}{a} - h \left(1 - \cosh \left(\frac{a\tau}{c} \right) \right) - \sqrt{\left(\frac{c^2}{a} \right)^2 + h^2 \sinh^2 \left(\frac{a\tau}{c} \right)} \quad (14.19)$$

Using the same parameters as in Figure 14.7 on page 320, the stretch as a function of τ is shown in Figure 14.8 on page 322. Given an elasticity and

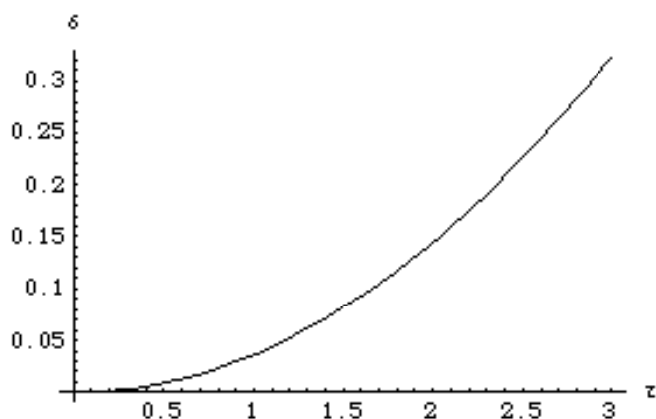


Figure 14.8: **Stretched String between Rockets** The stretch of a string connected between two identical rockets as a function of the time of the top rocket, see Figure 14.7 on page 320. The parameters for this figure were $a = \frac{1}{3} \frac{\text{ltyrs}}{\text{yr}^2}$, $h = 1 \text{ ltyr}$.

breaking tension, we could calculate the τ at which the string breaks but that would get us into a problem in materials engineering.

14.5 The Accelerated Reference Frame

Although we know that an accelerated observer does not have the same laws of physics as an inertial observer, there are often circumstances in which it is advantageous to make observations from an accelerating system. In addition, we will find that the General Theory of Relativity will have a very close and important connection with accelerated observers and the intuition that is developed here will be valuable there, see Section 16.2 on page 346.

We can proceed to construct the reference frame for an accelerating system in the same way that we did for inertial observers, see Section 11.1 on

page 241. Immediately, there are several problems. If we use the confederate procedures, i. e. placing confederates by some rule and endowing them with a clock to label events. There are actually several choices. At some time t , we could set at a fixed distance from each other a set of confederates with the same acceleration. This is not reasonable. As time goes on the confederates would find themselves drifting apart and, worst still, they would not have common lines of simultaneity, see Section 14.4.3 on page 319. Another choice would be to place them at a fixed distance but give them suitably adjusted accelerations so that they maintain their separations. In this case, all the confederates experience different accelerations, see Section 14.4.2 on page 317. Not only do they experience different accelerations, If we endow them with identical clocks, these clocks will run at different rates, again see Section 14.4.2 on page 317. Of course, we can see that since they share the same magic event, they will agree on simultaneity. Thus

$$\begin{aligned} x_{h,\tau'} &= \left(\frac{c^2}{g} + h \right) \cosh \left(\frac{\frac{g\tau'}{c}}{1 + \frac{gh}{c^2}} \right) - \frac{c^2}{g} \\ ct_{h,\tau'} &= \left(\frac{c^2}{g} + h \right) \sinh \left(\frac{\frac{g\tau'}{c}}{1 + \frac{gh}{c^2}} \right) \end{aligned} \quad (14.20)$$

could be used to label events where $(x_{h,\tau}, t_{h,\tau})$ are the event labels provided by the inertial commover of the origin confederate. In Equation 14.20 on page 323, h designates a position of the confederate and τ' is the time on that clock. g is the acceleration of the confederate at the origin. These expressions are simplified if we refer all clock readings to the origin confederate's time, i. e. the nearest confederate records the event time on their clock and then translates to the origin confederate's time using Equation 14.9 on page 318. This implies that one of the origin confederate plays a special role and is "in charge." With this change, we have

$$\begin{aligned} x_{h,\tau} &= \left(\frac{c^2}{g} + h \right) \cosh \left(\frac{g\tau}{c} \right) - \frac{c^2}{g} \\ ct_{h,\tau} &= \left(\frac{c^2}{g} + h \right) \sinh \left(\frac{g\tau}{c} \right) \end{aligned} \quad (14.21)$$

We can invert this system to yield the equations of h and τ in terms of the inertial coordinate labels,

$$h = \sqrt{\left(x_{h,\tau} + \frac{c^2}{g} \right)^2 - c^2 t_{h,\tau}^2} - \frac{c^2}{g}$$

$$\tau = \frac{c}{g} \tanh^{-1} \left(\frac{ct_{h,\tau}}{x_{h,\tau} + \frac{c^2}{g}} \right) \quad (14.22)$$

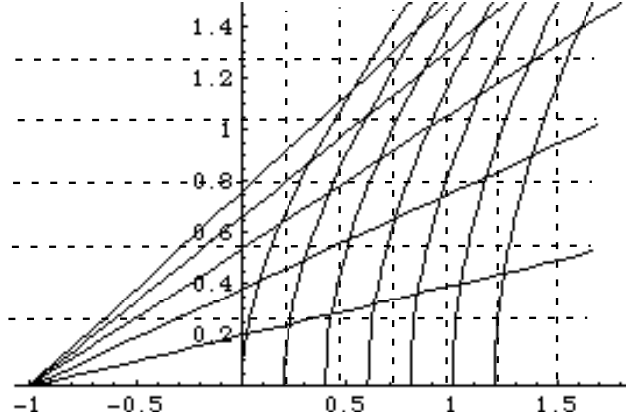


Figure 14.9: **Coordinate grid for a uniformly accelerated observer by means of confederates.** The time-like world line passing through the origin event is that of an observer that has an acceleration of $1 \frac{\text{lyr}}{\text{yr}^2}$. This is the reference observer for this coordinate system composed of confederates at fixed distances from the reference observer. The space-like lines are the locus of events coordinatized at the same time in this coordinate system. Shown dotted are the lines of constant time and place as determined by an inertial observer that is comoving with the accelerated observer at the initial event.

This coordinate scheme still has very serious drawbacks. The farthest confederate below the reference observer is at the magic event, $h = -\frac{c^2}{g}$ and that confederate has an infinite acceleration. The range in τ is $-\infty < \tau < \infty$. In fact, no events outside the forward elsewhere of the magic event has a nearby confederate. The forward elsewhere from any event is all the space-like events with positive position from that event bounded by light lines emanating from that event. An event near the magic event light trajectory although at finite times in the inertial coordinates is at plus or minus infinity in τ . This feature of not being able to cover all of space time with confederates and bounded times will be intrinsic to accelerated coordinate systems and we will not be able to repair it. The infinite acceleration is problematic but not easy to overcome except to realize that these confederates are hypothetical.

A simpler coordinatizing scheme which was identical to the confederate method in the inertial case is achieved by using a protocol like the one in Section 11.1 on page 241 in which there is only one observer and that observer uses a clock and records the travel times of light to and from the event in question and then sets the coordinates as we did in the inertial case,

$$\begin{aligned} x &= \frac{c\tau_2 - c\tau_1}{2} \\ t &= \frac{\tau_2 + \tau_1}{2}. \end{aligned} \quad (14.23)$$

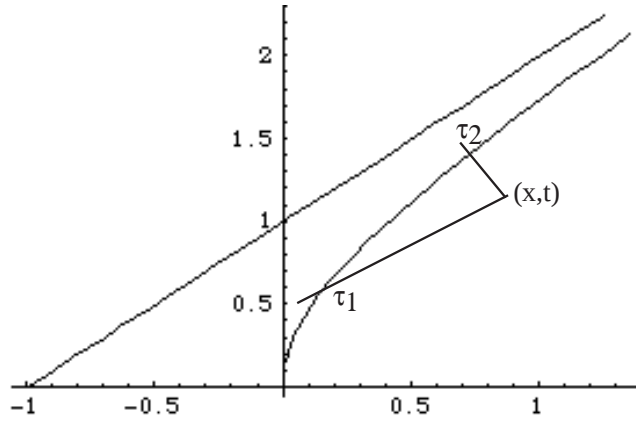


Figure 14.10: **Protocol for using an accelerated observer to coordinatize space-time.** The event that an inertial observer would label as (x_0, t_0) would be labeled as $x = \frac{c\tau_2 - c\tau_1}{2}$ and $t = \frac{\tau_2 + \tau_1}{2}$.

This coordinatizing is shown in Figure 14.10 on page 325. This method of coordinatizing also has the advantage of not assuming that the underlying space is homogeneous. More will be made of this later, see Chapter 18 on page 373.

For a uniformly accelerated observer with acceleration g and setting the origin event at the zero velocity event of the observer, we can find the new coordinates, (x, t) , in terms of the inertial observers coordinates, (x_0, t_0) , by following the procedure in Section 11.2.3 on page 255 and Figure 11.7 on page 255. The equations of the two light cone lines from (x_0, t_0) are $\frac{t-t_0}{x-x_0} = \pm \frac{1}{c}$. Thus τ_1 and τ_2 satisfy

$$\frac{c^2}{g} \cosh\left(\frac{g\tau_1}{c}\right) - \frac{c^2}{g} - x_0 = \frac{c^2}{g} \sinh\left(\frac{g\tau_1}{c}\right) - ct_0$$

$$\frac{c^2}{g} \cosh\left(\frac{g\tau_2}{c}\right) - \frac{c^2}{g} - x_0 = -\frac{c^2}{g} \sinh\left(\frac{g\tau_2}{c}\right) + ct_0. \quad (14.24)$$

These can be solved for τ_1 and τ_2 and inserted into Equations 14.23 on page 325 to find (x, t) .

$$\begin{aligned} x &= \frac{c^2}{g} \ln \left[\left(1 + \frac{(x_0 + ct_0)g}{c^2} \right) \left(1 + \frac{(x_0 - ct_0)g}{c^2} \right) \right] \\ t &= \frac{c}{g} \ln \left[\frac{\left(1 + \frac{(x_0 + ct_0)g}{c^2} \right)}{\left(1 + \frac{(x_0 - ct_0)g}{c^2} \right)} \right]. \end{aligned} \quad (14.25)$$

Once again, note that these coordinates are singular on the light cone boundaries, $-\frac{c^2}{g} = (x_0 \pm ct_0)$, of the forward elsewhere from the magic event, $(x_0 = -\frac{c^2}{g}, t_0 = 0)$. In this coordinate, the range of x is $-\infty < x < \infty$ and similarly for t . This looks more like a distance and a time. Despite this range in x and t , you should realize that this range of coordinates does not cover the entire range of (x_0, t_0) but only the forward elsewhere from the magic event. We can get a better feel for the shape of this coordinate system by removing those pesky $\ln$ functions. Redefining distance and time by

$$\begin{aligned} \eta &\equiv \exp\left(\frac{gx}{c^2}\right) \\ \zeta &\equiv \exp\left(\frac{gt}{c}\right). \end{aligned} \quad (14.26)$$

Plugging in and doing a little algebra,

$$\begin{aligned} \eta^2 \left(\frac{c^2}{g}\right)^2 &\equiv \left(x_0 + \frac{c^2}{g}\right)^2 - c^2 t_0^2 \\ \zeta^2 &\equiv \frac{1 + \frac{(x_0 + ct_0)g}{c^2}}{1 + \frac{(x_0 - ct_0)g}{c^2}}. \end{aligned} \quad (14.27)$$

Note that, in the forward elsewhere from the magic event, η^2 and ζ^2 are positive with η equal to zero on both of the edges and ζ equal to zero at the lower edge and plus infinity at the upper edge. From Equation 14.27 on page 326, it follows that events at the same distance, same x or η , are hyperbolas with the common magic event $(-\frac{c^2}{g}, 0)$ in the inertial coordinate. In the new coordinate, (x, t) , the magic event is at spatial minus infinity or in (η, ζ) at $\eta = 0$. The events at the same time, same t or ζ , are straight

lines passing through the magic event. In the (x, t) coordinates, the lower edge is at minus infinity and the upper edge is at plus infinity. Thus this coordinate system looks like the system with confederates at fixed separation and adjusted accelerations, Figure 14.9 on page 324, with just a relabeling of distances and times. Obviously, lines of constant time, t , are lines of simultaneity to the special observer and the lines of fixed separation are the various suitably accelerated time-like curves. It is easy to show that this system of coordinatizing is the same as the one with the confederates with adjusted accelerations and with corrected clocks by merely reidentifying (h, τ) in terms of (η, ζ) or (x, t) .

$$\begin{aligned}\eta &= \frac{g}{c^2} \left(h + \frac{c^2}{g} \right) \\ \zeta &= \exp \left(\frac{g\tau}{c} \right)\end{aligned}\tag{14.28}$$

It is interesting to note that now that, although the relevant times are the same, $t = \tau$, the relevant distances are not the same,

$$h = \frac{c^2}{g} \left(\exp \left(\frac{gx}{c^2} \right) - 1 \right)\tag{14.29}$$

Confederates placed at equal spacing as measured in h will not be equally spaced in x even though the scale of length at the origin Δh and Δx are commensurate. At any place labeled by either h or x , the scales of distance are related by Equation 14.29 on page 327 and increments are related by

$$\Delta h = \exp \left(\frac{gx}{c^2} \right) \Delta x.\tag{14.30}$$

This is an example of a metric relationship. We will come upon this problem later in General Relativity, Section 17.7 on page 372. Which distance is the separation, h or x ? The Δh was constructed to be the proper distance between local confederates. The distance Δx is the incremental distance as measured by light travel time. Either can be used as the distance but practically speaking the light travel time method is the one that is utilized and thus makes sense as our measure although we will have to correct for the local distortion using the metric. This is one of the complications of accelerated systems.

We can complete the construction of our accelerated coordinate system in (x, t) by inverting Equation 14.25 on page 326,

$$x_0 = \frac{c^2}{g} \left(\exp \left(\frac{gx}{c^2} \right) \cosh \left(\frac{gt}{c} \right) - 1 \right)$$

$$t_0 = \frac{c}{g} \exp\left(\frac{gx}{c^2}\right) \sinh\left(\frac{gt}{c}\right). \quad (14.31)$$

Our interpretation of the distance measures can now be verified by using the metric that is provided by the inertial coordinate system. The interval, see Section 12.6 on page 295, between nearby events with differences in their coordinates of $(\Delta x_0, \Delta t_0)$ is given by

$$\Delta x_{prop}^2 = \Delta x_0^2 - c^2 \Delta t_0^2 \quad (14.32)$$

where x_{prop} is the proper distance, if the separation is spacelike, and

$$\Delta \tau_{prop}^2 = \Delta t_0^2 - \frac{\Delta x_0^2}{c^2} \quad (14.33)$$

where t_{prop} is the proper time, if the separation is timelike. Using Equation 14.31 on page 328, these become

$$\Delta x_{prop}^2 = \exp\left(\frac{2gx}{c^2}\right) \{ \Delta x^2 - c^2 \Delta t^2 \}, \quad (14.34)$$

if the separation is spacelike, and

$$\Delta \tau_{prop}^2 = \exp\left(\frac{2gx}{c^2}\right) \left\{ \Delta t^2 - \frac{\Delta x^2}{c^2} \right\}, \quad (14.35)$$

if the separation is timelike. These same relations in the (h, τ) coordinates are

$$\Delta x_{prop}^2 = \Delta h^2 - \left(h + \frac{c^2}{g}\right)^2 \frac{g^2}{c^2} \Delta \tau^2, \quad (14.36)$$

if the separation is spacelike, and a similar expression for the timelike case. Using the hangle, see Section 12.5 on page 289, between the magic event and the events in question, $\Delta \phi \equiv \frac{g \Delta \tau}{c}$, Equation 14.36 on page 328 becomes

$$\Delta x_{prop}^2 = \Delta h^2 - \left(h + \frac{c^2}{g}\right)^2 \Delta \phi^2. \quad (14.37)$$

The similarity between this form and the usual form for the distance in polar coordinates is striking and consistent with our interpretation of the hangle. See Figure 12.4 on page 277 and Figure ?? on page ??.

Can this system, particularly in (x, t) , generate a reasonable coordinate system? Will it? It should be obvious that there are some serious problems here. Before we go into all the problems, let's look at how our

friend the accelerated observer would indicate events. Not thinking that he or she is particularly different, he/she would use a conventional grid for the labels of the events that are recorded. He/she would think that his/her measures of time and space are like those of an inertial observer and thus prepare an orthogonal grid to represent events. There is a clear and obvious

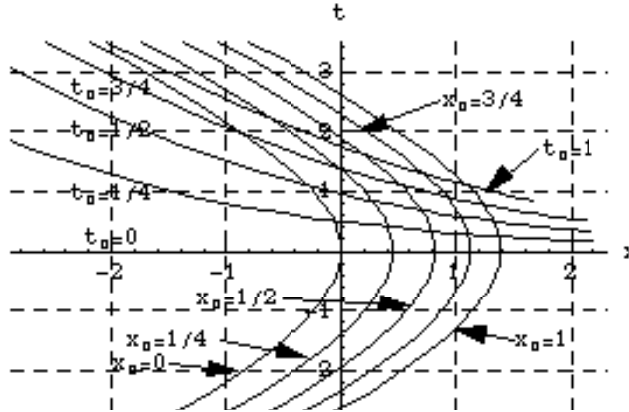


Figure 14.11: **Lines of Constant position and time in an accelerated coordinate system** In Figure 14.9 on page 324, the dashed lines represent events at either constant position, vertical dashed lines, or constant time, horizontal dashed lines, as designated by the inertial observer. In this figure these lines are the solid curves and the lines of constant position and time as designated by the accelerated observer are shown as dashed. Again in this figure lengths are in units of $\frac{c^2}{g}$.

distortion for the accelerated observer. Several features should be noted. It was noted above that, even though the range of position and time are the same as for the inertial observer, the events that are coordinatized are those in the forward light cone from the magic event and that points on these light lines, although finite to the inertial observer are mapped to infinity in these coordinates. In particular, note that the lines of constant t_0 for $t_0 = 1$ and x_0 for $x_0 = 0$ never cross and move off to ∞ together. This is, of course, a reflection of the fact that the event $(x_0 = 0, t_0 = 1)$ is on the light line from the magic event. Thus the accelerated observer thinks the all events are coordinatized but, as already discussed, the only events that can be coordinatized are in the forward elsewhere from the magic event. A further ramification is that, since lines of constant x_0 are inertial and commoving with the accelerated observer at $t = 0$, these inertial observers experience a finite time between the events that bound the forward elsewhere from the

magic event and yet the accelerated observer says that this same observer experiences an infinite time interval between these events. Also note that, if the inertial observer should chose to pursue the inertial observer by accelerating in that direction, once the inertial observer passes the events bounding the forward elsewhere from the magic event, there is no acceleration that can accomplish this goal. This situation is very similar to the case of the black hole, see Section 18.1 on page 373, in which there is an event horizon and, in fact, the underlying physics is very similar.

All of these problems with the coordinatizing by the accelerated observer are also similar to those that emerge when attempting to coordinatize a curved space with a single flat map. Atlas maps of the earth are all distorted and some points such as the north pole are even topologically distorted, a point on the earth appears in the atlas as a line. As we will see in Section 17.7 on page 372, these similarities are not accidental.

14.6 Events at the Same Proper Time

Now consider the set events that are the same proper time from the origin event. These are the events that satisfy

$$c^2t^2 - x^2 = c\tau_0^2$$

where τ_0 is the proper time between the origin event and all the events on this curve.

Chapter 15

Relativistic Dynamics

15.1 Relativistic Action

As stated in Section ?? on page ??, all of dynamics is derived from the principle of least action. Thus it is our chore to find a suitable action to produce the dynamics of objects moving rapidly relative to us. For a starter, we will consider only the action that would be associated with point particles but even more simply freely moving particles. Later we can discuss the action for relativistic fields and actions that combine particles and fields, see Section ?? on page ??.

As we saw in Section 7.4 on page 194, it is advantageous if the action possess the maximum amount of symmetry. This will produce the largest number of conserved quantities which in turn will simplify the analysis. In other words, in addition to having the usual symmetries of space and time translation, it would be nice to have the action be symmetric under Lorentz transformations. Remember that the classical actions are not symmetric under Galilean transformations but are invariant instead, see Section 7.4.4 on page 199. Having an action that is symmetric under the Lorentz transformations will expand the set of conserved quantities available for the solution of dynamical problems. In the following sections, we will be more careful in our handling of the notation and remember that there are three spatial directions, i. e. the position is $\vec{x}$. Where it is unimportant for the interpretation, we will suppress the vector designation.

15.1.1 The Action for a Free Particle

In order to discover the action for rapidly moving particles, we should look at simple situations. For the free particle, we know what the natural trajectory

in space time is – a straight line. We want this to be the trajectory with the least action. In addition, if we want the set of Lorentz transformations to be a symmetry for for this action, we should construct it from form invariants of the Lorentz Transformations, see Section 7.4.3 on page 198, and Section 12.6 on page 295. For time-like trajectories, the form invariant that characterizes the trajectory is the proper time. The action for the free particle should be dependent on the proper time and only on the proper time. The simplest possibility is that the action depend on the proper time linearly. Since action has the dimensions of an energy times a time, we have to multiply the proper time by something with the dimensions of an energy. Fortunately, the relevant dimensionful parameters are available. One is the mass of the particle. In fact, when you think about it this will be the definition of mass. Well, actually only that mass that is called the inertial mass. We will expand on this idea in Chapter 16 on page 345 and below in Section 15.3 on page 336. Also, since in the case of the free particle, we know that the trajectory is a straight time-like line. This is because there must exist a Lorentz observer who has the particle at rest in his/her frame. Since the straight trajectory is the longest world-line between two events, see Section 12.3 on page 280, and we want the action to be a minimum for the naturally occurring trajectory, the action should be proportional to the negative of the proper time. In this way, the greatest proper time will correspond to the least action. The unique combination that we have been led to is

$$S(\vec{x}_0, t_0, \vec{x}_f, t_f; trajectory) = -mc^2 \tau_{(\vec{x}_0, t_0, \vec{x}_f, t_f; trajectory)} \quad (15.1)$$

$$= -mc^2 \sum_{trajectory, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \Delta\tau_i \quad (15.2)$$

where the $\Delta\tau_i$ are the proper time intervals in each segment, see Figure 15.1 on page 333.

This form is inappropriate for the interpretation of an action since it is not time sliced. To transform from segment slicing which is what we have in Equation 15.2 on page 332 to time slicing, we use the fact that we can relate proper time intervals to coordinate time intervals as $\Delta\tau_i = \sqrt{(t_i - t_{i-1})^2 - \frac{(\vec{x}_i - \vec{x}_{i-1})^2}{c^2}}$. If we now factor out the $(\Delta t_i)^2$ and realize that the velocity in space time is the inverse slope to the trajectory, $\frac{\Delta\vec{x}_i}{\Delta t_i} \equiv \vec{v}_i$, or

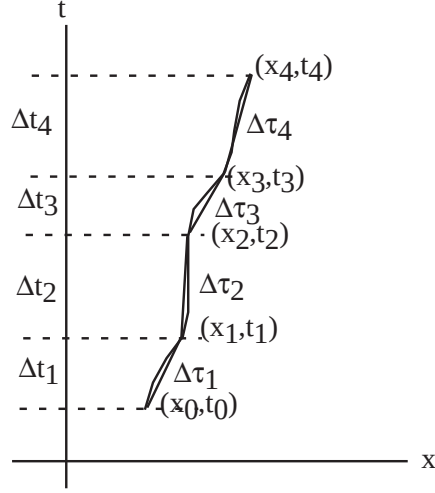


Figure 15.1: **Segmented Relativistic Action** The action for a relativistic particle is naturally expressed in terms of the proper time intervals as $-mc^2 \sum_{trajectory, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \Delta\tau_i$, where the $\Delta\tau_i \equiv \sqrt{(t_i - t_{i-1})^2 - \frac{(x_i - x_{i-1})^2}{c^2}}$ are the proper time intervals in each segment. Actions are best interpreted in terms of coordinate time, $\Delta t_i = (t_i - t_{i-1})$.

$$\begin{aligned}
 S(\vec{x}_0, t_0, \vec{x}_f, t_f; trajectory) &= -mc^2 \sum_{trajectory, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \sqrt{(\Delta t_i)^2 - \frac{(\Delta \vec{x}_i)^2}{c^2}} \\
 &= -mc^2 \sum_{trajectory, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \sqrt{1 - \frac{(\Delta \vec{x}_i)^2}{(\Delta t_i)^2}} \Delta t \\
 &= -mc^2 \sum_{trajectory, \vec{x}_0, t_0}^{\vec{x}_f, t_f} \sqrt{1 - \frac{\vec{v}_i^2}{c^2}} \Delta t \quad (15.3)
 \end{aligned}$$

With this time slicing, we identify the Lagrangian for the free particle as $L(\vec{v}, \vec{x}) = -mc^2 \sqrt{1 - \frac{\vec{v}^2}{c^2}}$. We should compare this result with the classical lagrangian for the free particle, $L_{Class}(\vec{v}, \vec{x}) = m \frac{v^2}{2}$. Remember, $v \equiv |\vec{v}|$ and $v^2 = \vec{v}^2$.

In the limit the $\frac{v^2}{c^2} \ll 1$, $L(\vec{v}, \vec{x}) = -mc^2(1 - \frac{v^2}{2c^2} \dots) = (m \frac{v^2}{2} - mc^2 \dots)$. Thus, the relativistic lagrangian is the same as the classical Lagrangian to

within an additive constant. An added constant in the Lagrangian adds a term in the action that is

$$- \sum_{\text{trajectory}, \vec{x}_0, t_0}^{\vec{x}_f, t_f} mc^2 \Delta t = -mc^2(t_f - t_0)$$

which does not depend on the trajectory and thus does not effect the path selection process. Therefore, the physics is the same for these two Lagrangians in the low velocity limit.

We can also calculate the relativistic free particle action between two events over the natural path since we already know that this is the straight line trajectory; that was how we decided what the action was. The action for the naturally occurring trajectory is

$$S(\vec{x}_0, t_0, \vec{x}_f, t_f; \text{natural}) = -mc^2 \sqrt{1 - \frac{(\vec{x}_f - \vec{x}_0)^2}{(t_f - t_0)^2 c^2}} (t_f - t_0) \quad (15.4)$$

15.2 Energy and momentum of a single free particle

Using the fact that the spatial and temporal translations are a continuous symmetry for this action, we have energy and momentum conservation. Using Noether's theorem, Section 7.4.4 on page 199, the energy is the change in the action when the final time is shifted or

$$E = \frac{\delta S}{\delta t_f} \quad (15.5)$$

$$= \frac{mc^2}{\sqrt{1 - \frac{(\vec{x}_f - \vec{x}_0)^2}{(t_f - t_0)^2 c^2}}} \quad (15.6)$$

$$= \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}. \quad (15.7)$$

since for the straight line trajectory, $\frac{\vec{x}_f - \vec{x}_0}{t_f - t_0} = \vec{v}$.

Similarly, the momentum is the change in the action when you translate the final position.

$$\vec{p} = \frac{\delta S}{\delta \vec{x}_f} \quad (15.8)$$

$$= \frac{m \frac{(\vec{x}_f - \vec{x}_0)}{(t_f - t_0)}}{\sqrt{1 - \frac{(\vec{x}_f - \vec{x}_0)^2}{(t_f - t_0)^2 c^2}}} \quad (15.9)$$

$$= \frac{m \vec{v}}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (15.10)$$

Note that, for a massive particle observed in its rest frame, $\vec{v} = \vec{0}$, the momentum is zero and the energy is mc^2 . Thus this energy is not necessarily an energy of motion like the classical kinetic energy, see Section 15.4 on page 337. On the other hand, note that, for a massive particle that is moving relative to some frame at a relative speed $\vec{v}$, the energy and momentum are dependent on the relative motion of the observer. This is how it was in old fashioned classical physics; the momentum was $\vec{p} = m\vec{v}$ and the energy of motion or kinetic energy was $KE = \frac{m\vec{v}^2}{2}$ and the values of the momentum and energy depended on the velocity with which the particle is observed. The difference in this case is that there is still an energy term even in the case of zero relative motion.

What is this energy? This energy is the energy in the famous formula $E = mc^2$. We now realize that this formula is not completely correct as written. It is more properly written as

$$E_{v=0} = mc^2. \quad (15.11)$$

For a system that is basically not moving relative to an observer, i. e. a commover, this is the energy that is necessary to form the system; its rest energy, Section 15.9 on page 340. This will make more sense when we talk about many particle systems, Section 15.8 on page 340.

In that regard, please note that by dividing Equation 15.10 by Equation 15.7,

$$\frac{\vec{p}c^2}{E} = \vec{v}. \quad (15.12)$$

For a single particle, with mass m , this is an interesting observation and provides another way to measure the relative velocity of a particle. For the multi-particle case, Section 15.8 on page 340, it will become the definition of the system velocity.

We should also note that with these formula's for the energy and momentum, we have a new interpretation of c . It is a conversion factor from momentum to energy units and even to mass units. For example, for Equation 15.12 on page 335 to be true $pc \stackrel{\text{dim}}{=} E$ or, for Equation 15.11 on page 335,

$E \stackrel{\text{dim}}{=} mc^2$. The most common way that this conversion is seen is when you see momenta expressed in $\frac{\text{MeV}}{c}$. MeV is an energy unit, the energy scale of nuclear reactions. It is 10^6 eV where the eV is the energy that an electron gains by moving through a voltage of one volt and is the energy scale of chemical reactions, see Section 1.4.2 on page 16, or an eV is 1.6×10^{-19} Joules. A momentum of $1 \frac{\text{MeV}}{c} = 1.6 \times 10^{-13} \frac{\text{Joules}}{c} = \frac{1.6 \times 10^{-13}}{3 \times 10^8} \frac{\text{kg m}}{\text{sec}} = 5.3 \times 10^{-22} \frac{\text{kg m}}{\text{sec}}$. Sometimes it is even worse than this. The factors of c will be suppressed as when you see someone write that the mass of the electron is 0.51 MeV. What the more careful person means is $0.51 \frac{\text{MeV}}{c^2} = 9.1 \times 10^{-31} \text{ kg}$.

15.3 Mass

In formulating the appropriate action for the relativistic particle, we needed to include the mass of the particle, see Section 15.1.1 on page 331. It was indicated that this is the inertial mass, the mass that resists changes in the state of motion. For a free particle, the more the trajectory deviates from a straight trajectory, the more action that it costs, the proper time is shorter, and the mass is the weighting factor; the larger the mass, the higher the action for the same deviation in the trajectory from the straight trajectory.

We also saw that to the co-mover, the energy of the system is simply related to the mass, $E_{v=0} = mc^2$. This mass was interpreted as the energy needed to create the system, Section 15.2 on page 334. This will be better understood when we discuss multi-particle states, Section 15.8 on page 340.

Note that the following combination of the energy and momentum does not have the velocity of the particle in it.

$$E^2 - \vec{p}^2 c^2 = \left(\frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}} \right)^2 - \left(\frac{m\vec{v}}{\sqrt{1 - \frac{v^2}{c^2}}} \right)^2 c^2 \quad (15.13)$$

$$= (mc^2)^2 \quad (15.14)$$

Even though E and p have different values depending on the relative motion to the observer, no matter what your relative motion to the particle, the combination $E^2 - p^2 c^2$ is the same and is $m^2 c^4$. In fact, since p and E are the dynamical entities, they are the things that are generally measured in an experiment on elementary particles. You measure the momentum, p , by seeing how the particle is effected by a known force and the energy, E , by a direct energy transfer measurement. This is how the mass of elementary particles is actually measured. You independently measure E and p and then

form the combination $E^2 - p^2c^2$ to determine the mass. Although mass is not the only identifying characteristic of an elementary particle, it is the most important one. The miracle of this operation is that you discover that with all the experiments that we have performed measuring a tremendous range of p and E that the masses observed are always in a small group of fixed values; these are the masses of the elementary particles. All electrons have a mass of 9.1×10^{-31} kg, all protons have a mass of 1.7×10^{-27} kg and so on. This even works for systems that we know are composite. All carbon twelve nuclei, composed of six protons and six neutron, have a mass of 1.9932×10^{-26} kg, which is **not** the mass of the six separated isolated protons and six separated isolated neutrons when measured carefully. The care that must be maintained if the mass difference is to be detected is the reason for the high precision. Similarly, the mass of a hydrogen atom which is a proton and an electron is ???? and, again, is **not** the mass of an proton and an electron when measured very carefully. We will discuss this problem when we discuss multi-particle states, Section 15.8 on page 340.

15.4 Kinetic Energy of a Single Particle

In non-relativistic physics the kinetic energy is zero if you are moving with the particle. It is in this sense that we define the kinetic energy for a relativistic particle as the energy of motion and thus the energy above the rest energy.

$$KE \equiv E - E_{v=0} \quad (15.15)$$

$$= mc^2 \left\{ \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1 \right\} \quad (15.16)$$

If $\frac{v^2}{c^2} \ll 1$, using $(1 + x)^n \approx 1 + nx \cdots$ for $x \ll 1$.

$$KE = mc^2 \left\{ \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - 1 \right\} \quad (15.17)$$

$$\approx mc^2 \left\{ 1 + \frac{v^2}{2c^2} \cdots - 1 \right\} \quad (15.18)$$

$$= \frac{mv^2}{2} + \cdots \quad (15.19)$$

Thus, in the small $\frac{v}{c}$ limit, we recover the usual kinetic energy of classical physics.

Later, in the section on applications, Section 15.10 on page 340 Item ?? on page ??, we will discuss how in interactions of particles, in particular nuclei, energy is conserved, mass is reduced, and kinetic energy is produced.

15.5 Transformations of Momentum and Energy

In Section 15.3 on page 336, we discovered that the combination $\frac{E^2}{c^4} - \frac{p^2}{c^2} = m^2$. In other words, the combination of variables $\frac{E^2}{c^4} - \frac{p^2}{c^2}$ although both E and p depend on the relative velocity v , does not depend on the relative velocity. Since the relative velocity is one of the ways to label the Lorentz transformations, the inference is that the energy and momentum combine to form a form invariant for the Lorentz transformations.

From the definitions of the energy, Equation 15.5 on page 334, and the definition of the momentum, Equation 15.8 on page 334, and the fact that the action, S , was made to be symmetric under the Lorentz transformations, Section 15.1.1 on page 331, you can show that the momentum, $\vec{p}$, and the energy, actually $\frac{E}{c^2}$, transform like the position, $\vec{x}$ and the time, t , see Section ?? on page ?? and Equations ?? on page ?? :

$$\vec{p}' = \frac{\vec{p} - \vec{v} \frac{E}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (15.20)$$

$$\frac{E'}{c^2} = \frac{\frac{E}{c^2} - \frac{v}{c^2} p}{\sqrt{1 - \frac{v^2}{c^2}}}. \quad (15.21)$$

For instance, as we saw in Section 15.2 on page 334, a particle moving at a speed v , $p = \frac{mv}{\sqrt{1 - \frac{v^2}{c^2}}}$ and $E = \frac{mc^2}{\sqrt{1 - \frac{v^2}{c^2}}}$, is seen to be at rest, $p = 0$ and $E = mc^2$, by an observer moving at v relative to the original observer. In the active view of transformations, Section ?? on page ??, we would say that the transformation brings the particle to its rest frame.

The same set of ideas can be reversed. To the observer that is at rest with respect to the particle, the energy, E , is mc^2 . To an observer moving at a velocity v relative to that observer the particle has momentum $p' = \frac{-\frac{E}{c^2}v}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{-mv}{\sqrt{1 - \frac{v^2}{c^2}}}$ and energy $\frac{E'}{c^2} = \frac{\frac{E}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{m}{\sqrt{1 - \frac{v^2}{c^2}}}$. Remember that this observer sees the particle moving with a velocity of $-v$. In fact, this could be

the another way to derive the addition of velocities formula, Section 11.3.4 on page 262.

In this sense, p and $\frac{E}{c^2}$ form a transforming set like x and t . We call anything that transforms like this a four vector. This nomenclature comes from the fact that in the real world there are three space coordinates and one time.

15.6 The Energy, Momentum, and Mass of Light

When you discuss light in the classical sense of Maxwell, it is not clear what is meant by the energy and momentum of light as discussed above. For now though, think of light in the modern context – a particulate transfer of energy and momentum, Section 19.4 on page 426. Note that for a particle the ratio

$$\frac{\vec{p}c}{E} = \frac{\vec{v}}{c} \quad (15.22)$$

is independent of the mass that we start with and thus holds for all masses including zero. For particles that travel at the speed of light, this implies that the ratio is 1 and that $p = \frac{E}{c}$. In other words for these particles, $E^2 - p^2c^2 = 0$ which implies that the mass is zero. Particles that travel at the speed of light are massless and the converse also holds that massless particles travel at the speed of light. Note also that massless particles, for example light, have energy and momentum.

Another way to see that is the for any particulate system that carries energy and momentum,

$$E = \sqrt{m^2c^4 + p^2c^2}. \quad (15.23)$$

In the limit as $m \rightarrow 0$, $E = pc$, which implies that $v = c$.

In addition, the momentum and energy of light transform under the Lorentz transformations in the same way as they do for massive particles, Equations 15.20 on page 338 and 15.21 on page 338. You can easily convince yourself that for massless particles, you cannot find a transformation that can yield $p' = 0$ in Equation 15.20 on page 338. Since $m^2 = \frac{E^2}{c^4} - \frac{p^2}{c^2}$ is the same for all Lorentz related E and p , if you had $p = 0$, and $m = 0$, you will also have $E = 0$. A massless particle with $E = 0$ and $p = 0$ is not there. Conversely, If a light beam, a beam of energy momentum transferred by massless particles, has energy E , it has momentum $p = \frac{E}{c}$ and another observer moving relative to the observer that measures that for the beam will measure an E' and p' given by Equations 15.20 on page 338 and 15.21

on page 338. Notice how my language here has segued into beam energetics regardless of the particulate nature of the energy¹.

15.7 Interactions

15.8 Multi-particle Systems

15.9 Rest energy of composite and elementary systems

When you are at rest with respect to a particle the p is zero and

$$E_{v=0} = mc^2 \quad (15.24)$$

This non-zero rest energy is an interesting aspect of special relativity. As stated above, Section 15.2 on page 334, it is this result that is the basis of the statement that there is an equivalence of mass and energy, $E = mc^2$.

Note that you can never bring light to rest and thus it is consistent to say that light has energy and still meaningless to talk about rest mass.

15.10 Applications of Energy Momentum

15.10.1 Nuclear Reactions

If we had redone our collision problem with relativistic kinematics we would still have found that the energy and momentum are conserved. Said more generally, since we want all fundamental processes to be time translation invariant we want energy to be conserved. Prior to relativity we also assumed that mass – the amount of stuff – was also conserved. When you think about it this is just a prejudice. We do not have a symmetry requiring mass conservation. In other words, you can now think of processes in which the mass is changed. A popular example is

¹In keeping with the spirit of our theme that Action is fundamental construct for all evolution and we assume that all fundamental systems have time and space translation symmetry, we should use Noether's Theorem to derive the constructions for the energy and momentum. The Maxwell Action which produces the Maxwell Equation evolution for the Electric and Magnetic fields, Equations 5.26 through 5.28 on page 148, we would also construct the field energy, E , and momentum densities, P , from Noether's Theorem and these would satisfy the requirement $E^2 - P^2 c^2 = 0$.

$$D + T \rightarrow He + n + 17.6 \text{ MeV.} \quad (15.25)$$

This example is popular because it is the basis for potential commercial fusion energy. A deuterium nucleus which is a heavy hydrogen nucleus, one neutron and one proton, and a triton, an even heavier hydrogen nucleus, which happens to be radioactive would serve as fuel for the fusion reactor.

The D and T are basically at rest. The incoming energy is $m_D c^2 + m_T c^2$. The outgoing energy is $\frac{m_{He} c^2}{\sqrt{1 - \frac{v_{He}^2}{c^2}}} + \frac{m_n c^2}{\sqrt{1 - \frac{v_n^2}{c^2}}}$.

Since there is no momentum coming in the momentum of the He and n are equal and opposite, $p_{He} = p_n \equiv p$. Writing the energy in terms of the momentum,

$$m_D c^2 + m_T c^2 = \sqrt{m_{He}^2 c^4 + p^2 c^2} + \sqrt{m_n^2 c^4 + p^2 c^2} \quad (15.26)$$

Solving for $\frac{p}{c}$

$$\frac{p}{c} = \sqrt{\frac{\{(m_D + m_T)^2 + m_{He}^2 - m_n^2\}^2}{4(m_D + m_T)^2} - m_{He}^2} \quad (15.27)$$

Looking up the values,

$$m_D = 2.01474 \text{ AMU} \quad (15.28)$$

$$m_T = 3.017 \text{ AMU} \quad (15.29)$$

$$m_{He} = 4.00387 \text{ AMU} \quad (15.30)$$

$$m_n = 1.00898 \text{ AMU} \quad (15.31)$$

These masses are in atomic mass units or AMU and the conversion is

$$1 \text{ AMU} = 931 \frac{\text{MeV}}{c^2}. \quad (15.32)$$

$$\frac{p}{c} = 0.174961 \text{ AMU}$$

This is one case where you need a calculator. You have to compute small differences.

Thus the value of $pc = 163 \text{ MeV}$.

Note that the kinetic energy of the He and n are:

$$\frac{KE_{He}}{c^2} = \sqrt{m_{He}^2 + \frac{p^2}{c^2}} - m_{He} = 0.0038209 \text{ AMU} \quad (15.33)$$

and

$$\frac{KE_n}{c^2} = \sqrt{m_n^2 + \frac{p^2}{c^2}} - m_n = 0.0150571 \text{ AMU} \quad (15.34)$$

Thus the KE of the He is 3.56 MeV and the KE of the n is 14.01 MeV. The total energy that goes into KE or motional energy is 17.57 MeV

15.10.2 A Relativistic Rocket

Another important application of relativistic dynamics is the case of the relativistic rocket. The simplest case is a rocket propelled by the emission of light. There are several reasons why this case is the simplest and yet important. Light propellants are the most efficient available. The other special feature is that no matter what speed the rocket is going the propellant has a momentum energy relationship that is simple and universal; no matter the frame of the observer $E = pc$.

It is worth drawing the process of emission in a space time diagram. Figure 15.2

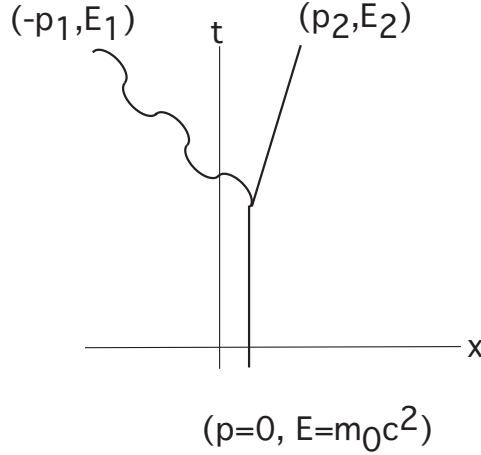


Figure 15.2: **Relativistic Light Rocket** A rocket with fuel has a mass m_0 initially co-moving with the observer. It emits light propellant of energy momentum $(-p_1, E_1)$. The rocket recoils and has momentum p_2 and energy E_2 .

The values of the energy and momentum of the recoiling rocket are constrained by energy momentum conservation. The mass and remaining fuel of the rocket is no longer m_0 but is $m_2 = \sqrt{\frac{E_2^2}{c^4} - \frac{p_2^2}{c^2}}$. Note that p_1 and p_2 are positive quantities. Momentum balance requires $p_1 = p_2$ and energy balance requires $m_0 c^2 = E_1 + E_2$. The fact that it is light that is emitted requires that $p_1 c = E_1$. The energy constraint becomes

$$\begin{aligned} E &= E_1 + E_2 \\ m_0 c^2 &= p_1 c + \sqrt{m_2^2 c^4 + p_1^2 c^2}. \end{aligned} \quad (15.35)$$

A little algebra shows the relationship between the original rocket and fuel mass and final rocket and fuel mass and the momentum of the rocket and fuel after the emission,

$$p_2 = p_1 = \frac{(m_0^2 - m_2^2) c}{2m_0}. \quad (15.36)$$

Combining this with Equation 15.35 on page 343 and doing a little bit of algebra

$$E_2 = \frac{(m_0^2 + m_2^2) c^2}{2m_0}. \quad (15.37)$$

The relative velocity of the rocket after emission is simply

$$\begin{aligned} \frac{v_2}{c} = \frac{p_2 c}{E_2} &= \frac{m_0^2 - m_2^2}{m_0^2 + m_2^2} \\ &= \frac{\left(1 - \left(\frac{m_2}{m_0}\right)^2\right)}{\left(1 + \left(\frac{m_2}{m_0}\right)^2\right)}. \end{aligned} \quad (15.38)$$

This equation is more interesting when it is inverted,

$$\frac{m_2}{m_0} = \sqrt{\frac{1 - \left(\frac{v_2}{c}\right)}{1 + \left(\frac{v_2}{c}\right)}}. \quad (15.39)$$

There are a few features of these results that require comment. Consider the case of galactic exploration. A typical galaxy, M31, is 7×10^{17} km or 7×10^4 ly. Our galaxy is actually just a little smaller. Suppose you want to explore the galaxy.

Chapter 16

Introduction to General Relativity

16.1 The Problem

After 1905 and the success of the Special Theory of Relativity, Einstein turned his attention to the problem of making the other known fundamental force of his time, gravitation, consistent with Special Theory of Relativity. Remember that the electromagnetic theory of Maxwell was consistent with the Special Theory from the start. The other force systems that we now know about such as the strong or nuclear force and weak force had not yet been identified. At this time, gravitation was still described by the action at a distance formulation identified with Newton, see Section 5.2 on page 123. This theory was intrinsically inconsistent with speed of light restrictions on the propagation of energy and momentum. In Newtonian Gravity, the acceleration of the moon due to the presence of the earth as $a_{moon} = -G_N \frac{M_{earth}}{r} \vec{r}$, where $\vec{r}$ is the separation vector between the moon and the earth. If, for some reason, the mass of the earth would change, the acceleration of the moon is instantly changed to accommodate the new mass. The moon instantly changes its orbit to a new one to accommodate the change. In essence, there is momentum and energy transferred to the moon. This implies that the information about the earth's mass in the form of energy and momentum is propagated to the moon faster than the speed of light. This violates the basic premises of the Special Theory.

The theory that he developed was rather long in gestation. It was not until 1916–1917 that he was finally able to articulate the basic principles of what is now called the General Theory of Relativity. This name is both a

misnomer and yet an insightful appellation. It was a modern theory of the effects of gravitation and thus should be called by that name – The Modern Theory of Gravity. But it was only after he took the fullest advantage of the underlying concepts of relativity that he was able to find the correct formulation of the theory and, in fact, it was through a generalization of the principles of relativity that he was able to develop the theory. We will follow this thread of development. The problem is that it is rather abstract and there is some tendency to lose track of the fact that it is a theory of gravity. On the other hand, it has the advantage of making it clear that a modern theory of gravitation is, in fact, a theory of the structure of space-time.

16.2 Free Fall Observers and the Equivalence Principle

In Section 9.2 on page 215, we discussed the physical implications of Galilean invariance. One of the ways of describing the meaning of this invariance was that you were always at rest in your own rest frame. In other words, there was an infinite set of related observers all of whom thought that they were at rest. Their world was isotropic. An object held out and released would remain there. If the object was given an initial relative velocity, it maintained that velocity. Yet these observers were moving relative to each other. On the other hand, the accelerated observer finds that a released piece of chalk will drift in some direction. The space is no longer isotropic. There are any number of experiments that the inertial, uniformly moving observer, and accelerated observer can perform to note their difference. It is in this sense that we say that, although you cannot measure velocity, you can measure acceleration. There are no speedometers on the starship Enterprise but it can have an accelerometer. It can even integrate the accelerations over time to find a velocity relative to some initial velocity but it cannot know its velocity in any absolute sense. Beside noticing the important point of the unmeasurability of absolute velocity, it is important to appreciate the fact that being inertial is a knowable fact. If you hold out a piece of chalk and release it. It will stay fixed in position. If it suddenly begins to move, you can know that you accelerated. Even more fundamentally, you feel a jolt. We should be a little more careful here. How do you know that it was you and not the chalk that was suddenly accelerated away from you? Putting us again, back in the box of knowing only relative effects, in this case acceleration. It is the jolt that is relevant here. Not only does the chalk start accelerating but a mass and spring held by you changes its configuration –

a jolt. In other words, you can build an Inertiality Maintenance Detector. For instance, using identical masses and springs build a three axis stretch meter, see Figure 16.1 on page 347. To an unaccelerated observer, these six mass-spring systems are all identical. If there is a difference between them, there is an acceleration. This is what is meant by a “jolt.” Thus we can tell if it is us or the chalk that is accelerating. Thus inertiality is an experimentally determined state.

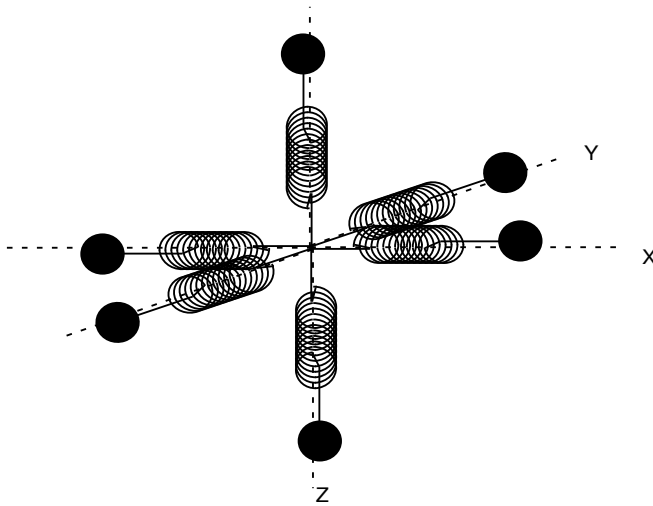


Figure 16.1: **Inertiality Maintenance Detector** Using three pair of springs and masses each pair arranged along each of three axis, we can construct an instrument to detect whether or not we are accelerating or, better said, whether or not we are inertial. Differences in the configuration of the springs will indicate the magnitude and direction of an acceleration.

There is another situation in which there is a detection of inertiality but there is acceleration. If an observer with an Inertiality Maintenance Detector, IMD, is in a gravitational field but is also in “free fall.” This is, for instance, the case when near the surface of the earth an observer is falling or an astronaut is in near earth orbit. For all these cases, an IMD would not show any preferred direction. This statement is actually not quite true and we will have to clarify it later, see Section 16.5 on page 355. Note, that in any of these free fall situations, there are actually an infinite family of “free

fall” observers. In fact, all observers connected by a Lorentz transformation are equally free fall. This is the difference between the astronaut and the observer that is just falling, a Lorentz boost. These free fall observers are interestingly like the inertial observers that we dealt with in Special Relativity. In the absence of gravity, these free fall observers are the same as our inertial observers of special relativity. We can now make the statement of the first principle of General Relativity. Free fall observers have the same laws of physics as the inertial observers of Special Relativity. This principle is called the Equivalence Principle.

16.3 The Equivalence Principle

The Equivalence Principle states that locally the effects of gravity are indistinguishable from those of an acceleration. This is the same as the observation in the previous section that an observer with an IMD that registers inertiality has the same laws of physics as an inertial observer in Special Relativity.

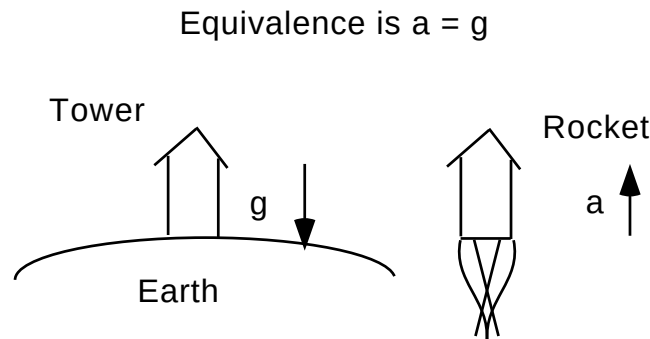


Figure 16.2: **Equivalence Principle** The Equivalence Principle states that locally there is no experiment that can differentiate between the effects of gravity and the a rocket ship with $a = g$.

The Equivalence Principle allows us to now identify some of the important effects of gravity. Using our knowledge of the the physics of accelerated motion in Relativity, see Chapter 14 on page 309 and particularly Section 14.5 on page 322. These will be examined in more specific contexts later, see Chapter 18 on page 373, but for now we will review the simplest implications. Before we get into these cases, let’s look at a very popular lecture/demonstration, The Monkey and the Hunter.

16.3.1 The Monkey and the Hunter

There is a very popular demonstration that is performed in most high school and college introductory physics classes. There is a gun of some type that launches a projectile and a target object, usually a toy monkey, that can fall some distance. The gun and the monkey are rigged so that at the instant the gun fires the monkey is released to start to fall.

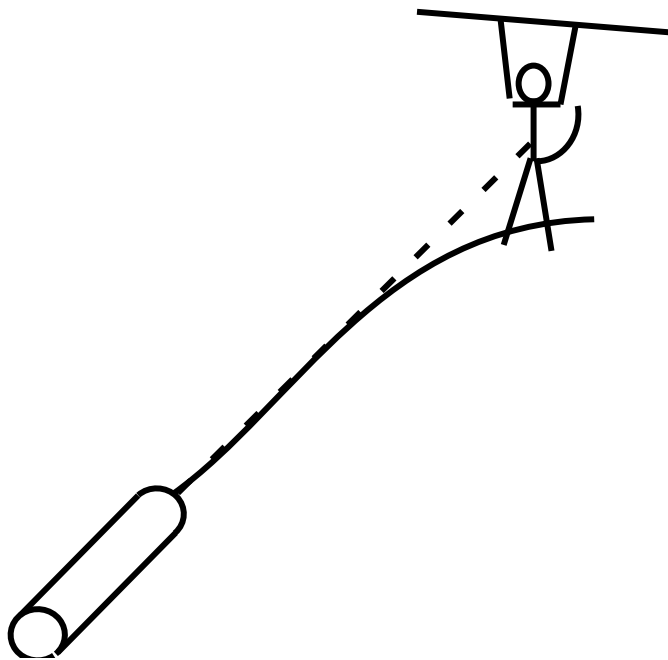


Figure 16.3: **Monkey and the Hunter** A popular lecture demonstration is to fire a projectile at a hanging toy monkey. The monkey is released at the instant that the gun is fired.

The class is usually asked where does the hunter aim. Since the monkey is falling, there is an argument that the hunter should aim below the initial position of the monkey to compensate for the finite time of flight of the projectile. On the other hand, the projectile has an arced trajectory and thus the aim should be above the current position. The correct answer is that the hunter should aim at the present position of the monkey. This is because once the gun is fired both the projectile and monkey are falling with an acceleration of g . In the frame accelerating down at a rate g , the effects of gravity are cancelled and thus neither the monkey nor the projectile have accelerated motion. In that frame, the projectile travels in a straight line

and the monkey never moves. Note that if the aim is correct, no matter how small the projectile muzzle velocity, it will ultimately hit the monkey. This is an interesting pre-relativity example of the equivalence principle. **Also it is important to note that in the free fall frame, the one in which the effects of gravity are removed, the projectile and the monkey have trajectories that are straight lines in space-time.**

16.4 Direct Effects from the Equivalence Principle

The statement of the Equivalence Principle above that the effects of gravity are indistinguishable from those of an acceleration is valid only locally. Measurements over extended regions of space and time can and as we will see show a difference between an acceleration and gravity but the Equivalence Principle provides a basis for some of the more direct effects of gravity. In real situations of mass distributions leading to gravitational effects there are two things that make the following discussion approximate. First, gravity is a field and thus takes on values at all points in space and time. It is just a fact that the dynamics of the gravitational field, called Einstein's Equations, do not admit solutions that are uniform in space and time. There is a similar circumstance in the case of the electromagnetic field. Maxwell's Equations do not admit solutions that are uniform in space and time. For applications of the Equivalence Principle since there is only one acceleration that the frame can have, it can only match a gravitational field at some point. Nearby points will have different values of g and thus will not be eliminated. We will see a case of this in our discussion of the Gravity Detector, see Section 16.5 on page 355. Regardless, there will be many cases when the gravitational system of interest can be well approximated by a uniform field and we will do so in the following. Second, any measurement apparatus will have some extension and thus the effects will have to take into account the extended effects of gravity. Again, in many circumstances, the measuring apparatus is small in extent compared to the region of interest and the measurement can be considered local. Clearly, a legitimate approximation. With these provisos, we proceed to look at some of the simple direct effects of gravity.

16.4.1 Universality and Eötvös–Dicke

One of the most striking features of Newton's Theory of Gravity is its universality. The great idea that behavior of apples falling from trees and the

moon in orbit were two aspects of the same law was one of its first significant philosophical and phenomenological successes of the theory. Not only does it effect all things, it effects them in the same way. Again, an interesting lecture demonstration is almost always performed in high school physics classes. A penny and a feather are enclosed in an evacuated glass tube. Inside the tube where the only significant forces on the penny and the feather are gravity, they fall together. The Equivalence Principle gives an immediate explanation to these two simple aspects of the universality of the Newton's Theory. All objects move the same in gravity because it is the observer that is accelerated.

Interestingly, Newton achieves universality in an indirect way and in several steps. First, gravity sees only mass and no other attribute of the object. Then it identifies two distinct roles for mass and then arbitrarily equates them. This issue of the relationship between gravitational and inertial mass was discussed earlier, see Section 2.2 on page 34. Let's be more specific. Newton first ascribes the force of gravity to an action at a distance force law, see Section 5.2 on page 123, that is based on the identification of mass as the source of the strength of the force. The gravitational force between two bodies labeled 1 and 2 is

$$\vec{F}_{Grav_{12}} = -G \frac{m_{gr1} m_{gr2}}{r_{12}} \vec{r}_{12}, \quad (16.1)$$

where $\vec{r}_{12}$ is the displacement vector from body two to body one. The two masses in this equation are called gravitational masses, indicated by the subscript gr , and are the source for the gravitational force. These masses are measured for instance in a balance scale. Body one reacts to the force by having an acceleration according to Newton's Laws as

$$m_{i1} \vec{a}_1 = \vec{F}_{Grav_{12}} = -G \frac{m_{gr1} m_{gr2}}{r_{12}} \vec{r}_{12}. \quad (16.2)$$

where the m_{i1} in the first part of Equation 16.2 on page 351 is the inertial mass of body one. One way in which this mass could be measured is by collision with a standard mass. The next step is to invoke the magic idea that these two very different concepts of mass are identical, $m_{i1} = m_{gr1}$, and cancel them from the two sides of the equation so that the acceleration no longer depends on the mass of body one. As is emphasized in Chapter 2 on page 33, the standards and protocol for measuring something is its definition. Here we have two very different definitions of mass that would have two very different protocols for measurement. This equality of the two masses is even more striking in light of the mass energy relationship, Equation ??

on page ??, and the constituent nature of matter. Yet these two different things are the same, strikingly the same.

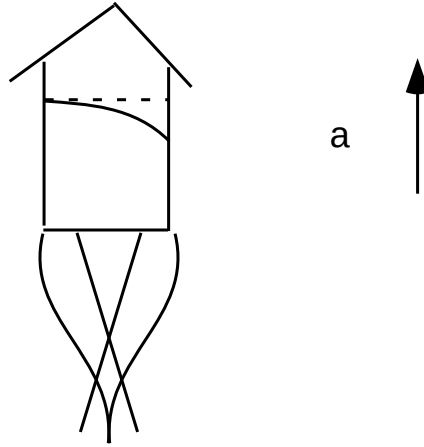
The equality of the gravitational and inertial masses was tested in a classic experiment in 1889 by Roland von Eötvös and recently improved by R. H. Dicke, [Eötvös 1922, Dicke 1967]. The idea is that the measurement of the gravitational force in a rotating system is influenced by the non-inertiality of the laboratory and the effects of each are proportional to the gravitational to inertial masses and thus the corrections are proportional to the ratio of the inertial mass to the gravitational mass. Newtonian universality required that this ratio be one. Using different materials in each leg of a torsion bar, the experiment could detect differences between the ratio for these materials. Eötvös found that the difference between wood and platinum was less than 10^{-9} and Dicke improved this limit for aluminum and gold to 10^{-11} . These small differences are very impressive especially in light of our new understanding about mass and energy as discussed in Section 15.3 on page 336. Gold or aluminum atoms have very different atomic and nuclear structures and thus different energies of binding. These differences are well within the measured precision of this experiment. Thus even if protons, neutrons, and electrons have identical inertial and gravitational masses, in these systems, the binding would manifest a detectable difference.

The Equivalence Principle directly requires the Universality of Newton's gravity and includes the result of the Eötvös–Dicke experiment without further assumptions.

16.4.2 Bending of Light Rays

Consider a rocket accelerating in a region of space that has no nearby masses and thus is free of gravitational effects. In Figure 16.4 on page 353, it is clear that a light beam entering one side of the rocket perpendicular to one wall will have a bent trajectory as measured in the rocket.

Using the Equivalence Principle, light in the neighborhood of a massive body must also bend. We can even be more quantitative. The time of passage of the beam across a rocket of width L is $\frac{L}{c}$. If the acceleration is g , the deflection on the far side of the rocket is $\frac{g}{2} \frac{L^2}{c^2}$. This calculation can be carried out with more care using the information that we have on accelerated observers, see Section 14.5 on page 322, but this result is certainly the correct order of magnitude. For the earth, the deflection in a one kilometer size room, a big room, is 5×10^{-11} meters, too small to be measured. This effect though has been measured for the case of the bending of star light by the sun. A classic experiment using a total eclipse of the sun was among



Light Rays in an accelerating rocket

Figure 16.4: **Bending of Light** A light ray entering one side of an accelerating rocket will be seen in the rocket as bending down. The Equivalence Principle then requires that a beam of light bend in the presence of a massive body.

the first verifications of the General Theory of Relativity of Einstein. It is important to note that the bending of star light predicted by the equivalence principle alone does not produce the full bending but, to get the correct value, will require that we use the full metric theory that is developed later, see Section 17.7 on page 372, [Will 1986], and [Weinberg 1972].

16.4.3 Clocks and Accelerations in Towers

In Section Item:Rocket on page 317, we study the behavior of clocks in an accelerated rocket. There we find that clocks at the top and bottom run at different rates and that the relationship between them is given in Equation Eq:AcceleratedTimes on page 318 as

$$\tau_{top} = \left(1 + \frac{ha_{bottom}}{c^2}\right) \tau_{bottom} \quad (16.3)$$

where h is the length of the rocket and a_{bottom} is the acceleration. It is important to realize that the top of the rocket is a fixed proper distance from the bottom. This keeps the rocket a fixed length as measured on the rocket. Because of this requirement is also important to note that the acceleration of the top of the rocket is not the same as that measured at

the bottom of the rocket. Using the relationship between the acceleration of the uniformly accelerated observer and the distance to the magic event, Equation 14.2 on page 311, these accelerations are related by

$$a_{top} = \frac{c^2}{d_{top}} \quad (16.4)$$

but d_{top} is $d_{bottom} + h$ so

$$\begin{aligned} a_{top} &= \frac{c^2}{d_{bottom} + h} \\ &= \frac{\frac{c^2}{d_{bottom}}}{1 + \frac{h}{d_{bottom}}} \\ &= \frac{a_{bottom}}{1 + \frac{h a_{bottom}}{c^2}} \end{aligned} \quad (16.5)$$

These phenomena associated with an accelerated rocket are nicely summarized in the examples on accelerated rockets and Bell's Problem, see Section 14.4.2 on page 317 and Section 14.4.3 on page 319.

Thus, again invoking the Equivalence Principle, we have that, in a tower near or on a massive body, clocks at the top and bottom of the tower run at different rates. Setting $a_{bottom} = g$ the local gravitational field, these rates follow directly from Equation 16.3 on page 353 where h is the height of the tower and g is the gravitational field at the location of the tower. A simple insertion of values into this equation would seem to indicate that it is not testable in an earth based laboratory. For a tower of height h in meters, the fractional change in rate between the bottom and top clock is $\frac{\Delta t}{t} = \frac{10 \frac{\text{m}}{\text{sec}^2}}{9 \times 10^{16} \frac{\text{m}^2}{\text{sec}^2}} h \approx 10^{-16} \text{ m}^{-1} h$. This would appear to be a forbidding shift to measure but since precision time sources are available and, with a clever trick to identify the signal from the background, Pound and Rebka have measured this shift, [Pound & Rebka 1959].

Of course, if we could have towers several kilometers tall, there would be no problem in conducting these experiments. The trouble is that our formula is valid only in the cases in which the gravitational field strength is a constant. The effect though is universal. In the presence of gravity, clocks run slower at the bottom than clocks at the top. If we use the full power of the Einstein Equations, Section 17.8 on page 372, cases of a varying field strength can be treated and this shift to lower frequencies called a red shift is observed in radar ranging experiments to the moon. In addition, with the advent of earth satellites in low earth orbit, these effects will also be

realized. In fact, the GPS positioning system has to be corrected for these effects. An application of General Relativity in everyday life.

16.5 Intrinsic Effects of Gravity

Consider an observer near a massive body, a free fall observer above the surface of the earth for instance. Using an IMD, Inertial Maintenance Detector, see Section 16.2 on page 346 and Figure 16.1 on page 347, the observer concludes that he/she is in free fall. There is no distortion of the masses in the IMD that would indicate an unbalanced force. A piece of chalk released by the observer at his/her location hovers where it is released. This is the essence of the Equivalence Principle. The acceleration has removed the effects of gravity. Despite this, the IMD is not uniform in all six axis. If the IMD is oriented so that one of the axis is along the line to the massive body, the the two masses along that axis are slightly further apart then the mass pairs in the two other axis directions that are in the plane parallel to the surface of the massive body. The elongation is twice the compression. There is no state of motion that the observer can carry out that eliminates this distortion. Even if he/she accelerates, there will be a distortion identified with the unbalanced gravitational force but there will also be this unusual distortion. Thus we conclude that the Equivalence Principle cannot remove all the effects of gravity. There always remains a distortion which stretches along the axis directed at the mass and compresses halfthat amount in the plane perpendicular to that axis. The magnitude of the distortion is proportional to the mass ofthe gravitating body, inversely proportional to the cube of the distance from the gravitating body, and the size of the IMD. A distortion of a elastic system in which the system is stretched in one direction and compressed in the other two is called a tidal distortion.

16.5.1 Distortion of Elastic Bodies

In an elastic mechanical system, there can be distortions of the system in which is little net motion but only relative motion between parts. The system is deformable. An elastic rubber band stretches, see Figure 16.5 on page 356.

Another simple distortion is shear. A common and simple way to produce this distortion is by placing a large phone book, not really elastic since the phone book will retain the distortion, on a table face up and on the top of the phone slide your hand across the flat top surface. The cross section

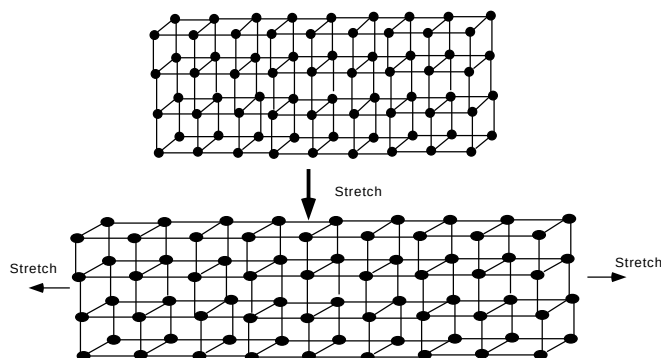


Figure 16.5: **Stretch of an Elastic Solid** The stretch distortion of an elastic solid. A pulled rubber band is an example of a stretched elastic solid

of the phone book will change from a rectangle to a rhombus. This is shear, see Figure 16.6.

A general property of a shear distortion is that although there is relative displacement of the parts the enclosed volume is retained as the distortion takes place. In deformable bodies, shear is a very common phenomena. An interesting example is that a Pascal or perfect fluid in hydrostatics can be defined as one which will not sustain a shear forming distortion. This is the reason that you cannot pile water. This is the direct manifestation that pressure is a scalar quantity,

$$\vec{F} = P\vec{A} \quad (16.6)$$

and that the hydrostatic force is directed along the normal to the area.

The stress that leads to shear deformations is called a shear stress. Generally, these are couples, a pair of equal and opposite pair of forces acting at a slight separation, not a single force.

The distortion that is manifest in our IMD is a tidal distortion, see Figure 16.7 on page 358. In this case, the body extends along one axis and compresses on the other two orthogonal axis. As shown below, Section 16.5.2 on page 357, to first order in the stretch, this distortion also has the property that it is volume preserving. It stretches twice the contraction but there are two contraction directions so that the total effect does not change the volume.

Again, this distortion is reasonably common and the most well known manifestation is the oceanic tides of the earth and thus the name. Although most of the explanations of the origin of the tides is a complex analysis of

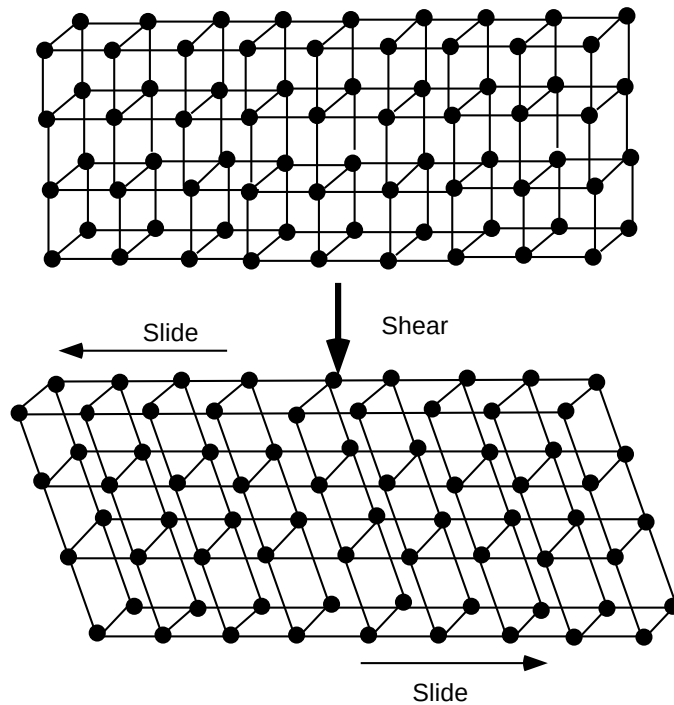


Figure 16.6: **Shear of an Elastic Solid** The shear distortion of an elastic solid. In shear two parallel planes are displaced relative to each other. Volumes are preserved.

the the gravitational attraction of the moon and the center of mass motion of the earth due to the earth moon orbit, it is really that the ocean is an IMD, a bunch of water—an incompressible perfect fluid, for the earth and that the earth is in free fall in the gravitational field of the moon. Thus the direct acceleration effects of the moons gravity are eliminated but the tidal distortion of the gravitational field remains. You can find the shape of the tides on the earth by combining the gravitation from the earth to the tidal force. The shape of the surface of the liquid is the one that everywhere has its normal along the net gravitational force and has the correct volume.

16.5.2 Gravitation and Tidal Forces

Returning to our main theme, we now understand that the Equivalence principle provides a means for the elimination of gravitational forces but does not eliminate the tidal stress that is the intrinsic signature of the presence

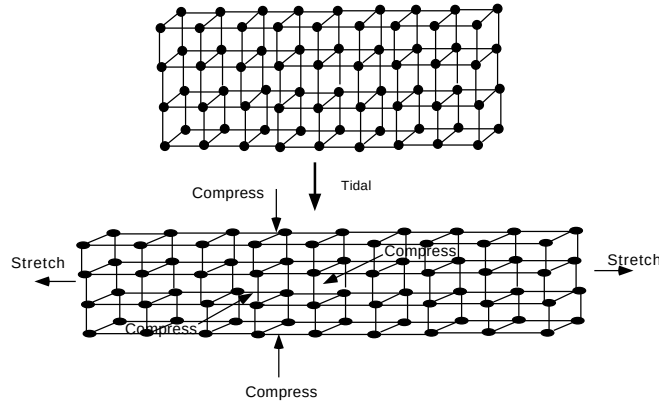


Figure 16.7: **Tidal Distortion of an Elastic Solid** The tidal distortion of an elastic solid is an extension along one axis and compressions in the other two. Volumes are preserved.

of gravity. No motion based or any other type of coordinate relabeling can eliminate this aspect of gravity. More on this later, see Section 17.7 on page 372. In order to understand its implications better, let's look at a system like our Inertiality Maintenance Detector but actually a little simpler, basically no springs, and do this a bit above the surface of the earth. A free fall observer places several independent masses in a sphere that surrounds him/her and one at his/her location, see Figure 16.8 on page 359. The released masses are independent in the sense that, once placed, they are in free fall. There are no external forces acting on them except gravity but they are released so that they are not moving relative to the original observer. They are co-moving at $t = 0$.

As time develops, there is a tidal distortion of the sphere with the expanding axis along the line to the center of the earth. It is easy to understand the basis for this distortion in the conflict between nature of the gravitational field and the direct application of the Equivalence Principle. First consider the three masses along the axis to the center of the earth, called the in/out axis. For definiteness, let's use δ as the initial radius of the sphere. Since the gravitational field is different at the three locations, the corresponding free fall accelerations are different and there is a relative acceleration between them. The three gravitational field strengths are $g_{top} = g \left(\frac{R_e}{R_e + h + \delta} \right)^2$, $g_{center} = g \left(\frac{R_e}{R_e + h} \right)^2$, and $g_{bottom} = g \left(\frac{R_e}{R_e + h - \delta} \right)^2$ where R_e is the radius of the earth and h is the height above the earth of our free fall observer, the

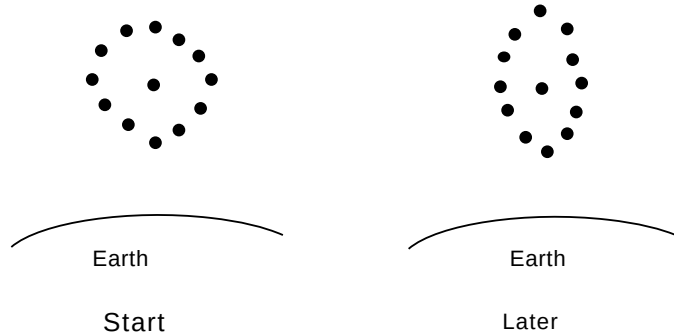


Figure 16.8: **Tidal Distortion of Free Fall Masses** A free fall observer arranges a sphere of identical free fall masses that are commoving. After a time, the masses are no longer spherical but the masses along the line to the center of the earth begin to separate and the masses in the plane tangent to the line to the center begin to come closer together. As the motion develops, the volume of the sphere is preserved.

center of the sphere. Since all three are in free fall these must be their accelerations. Thus the two relative accelerations between the center and top and bottom are

$$a_{rel\ top} = 2g_{center} \frac{\delta}{R_e + h} \quad (16.7)$$

and

$$a_{rel\ bottom} = -2g_{center} \frac{\delta}{R_e + h} \quad (16.8)$$

to first order in $\frac{\delta}{R_e + h}$. In this case the masses move apart.

Thus a plot of these trajectories as observed by the central free faller is shown in Figure 16.9.

The important point to note is that all three of these trajectories are for free fall objects, objects that have simple physics and thus these trajectories are of objects that are “inertial” and are straight line trajectories. These are three straight lines that start out parallel and as time develops drift apart. This is clearly not the geometry of Euclid. In Section 17 on page 363, we will discuss the meaning of this non-Euclidean geometry.

Picking an axis in the plane tangent to the earth and calling it the sideways axis, the three masses have their free fall accelerations directed differently. All three point to the center of the earth, see Figure 16.10 on page 361.

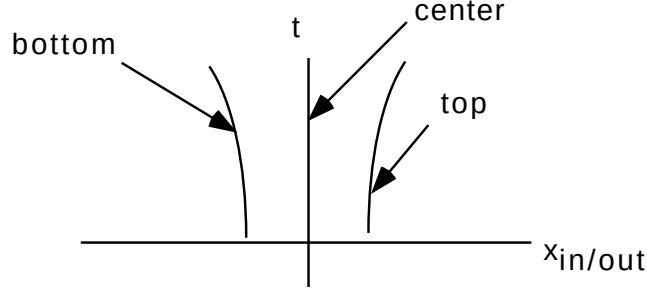


Figure 16.9: **Trajectories of Top and Bottom Free Fall Masses** Over time, the three free fall masses along the axis to the center of the earth, the $x_{in/out}$ axis, move apart. They are initially commoving.

Thus although the magnitudes of the free fall accelerations are the same, there is a relative acceleration between the masses. Projecting along the in/out and sideways direction and using the fact that θ is small and that $\sin(\theta) \approx \tan(\theta) = \frac{\delta}{R_e + h}$, the relative acceleration to first order in $\frac{\delta}{R_e + h}$ of either sideways mass relative to the center mass is

$$a_{rel \pm sideways} = \mp g_{center} \frac{\delta}{R_e + h}, \quad (16.9)$$

where, again, δ is the radius of the initial sphere of free fall masses. Thus, the sideways masses move toward the center, see Figure 16.11 on page 361. Again, we have a situation in which three initially commoving free fall objects, straight line trajectories, are moving toward each other. Again, a direct violation of Euclid's axioms and thus the geometry that is non-Euclidean.

Thus, we see that in all three space-time two planes the geometry is non-Euclidean. To make progress, we will have to understand a little bit of geometry.

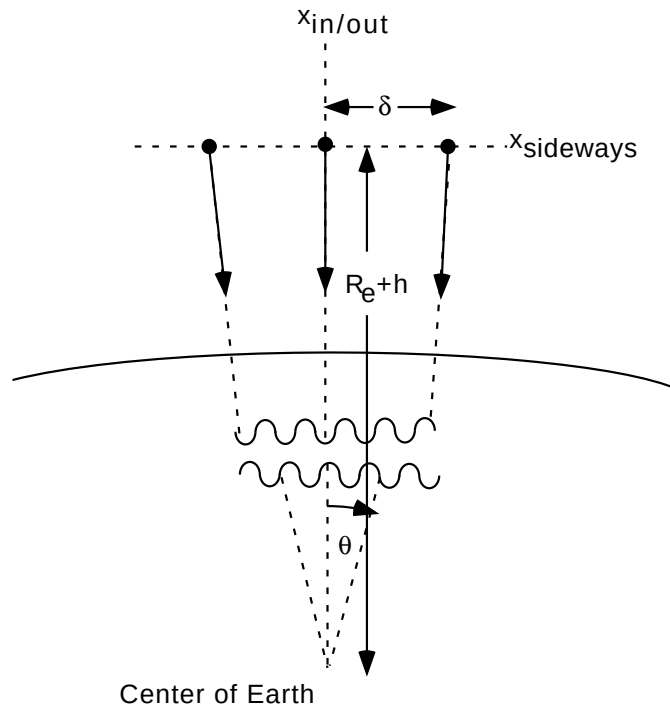


Figure 16.10: **Free Fall Accelerations in the Sideways Direction** For masses released along an axis in the plane tangent to the earth's surface, the free fall accelerations which are directed to the center of the earth are in different directions. Thus there is a relative acceleration among the masses.

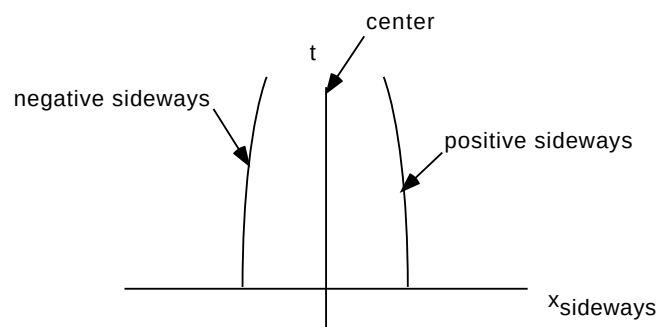


Figure 16.11: **Trajectories of Sideways Free Fall Masses** Over time, the three free fall masses along the axis in a plane tangent to the surface of the earth, the $x_{Sideways}$ axis, move apart. They are initially commoving.

Chapter 17

Geometry and Gravitation

17.1 Introduction to Geometry

Geometry is one of the oldest branches of mathematics, competing with number theory for historical primacy. Like all good science, its origins were based in observation and, with historical hindsight, we realize that the evident truths discovered by early geometers were really a result of limited perspective. But like them, for our discussion, we will take certain ideas as evident and as the basis for what we understand. The idea of the point and connected sets of points and particularly the idea of the straight line. As is evident from our discussion of Special Relativity, see Sections 11.3.5 on page 265 and 12.3 on page 280, we take the straight line to be the shortest distance between two points in space and the longest distance between two events in space-time.

Geometry developed from the need to measure land surfaces for agricultural purposes. The geometry that developed was what we now call plane geometry and the basis for it was first clearly articulated by Euclid and thus the name Euclidean geometry. Euclid set the foundation for plane geometry by means of a set of axioms, evident truths. Modern formulations of geometry realize that there are consistent systems that do not have the same set of axioms. The question then becomes one of choice or appropriateness. In fact, if the early geometers had considered the geometry that is appropriate to large distances on the earth, they would have developed a geometry that was not Euclidean. This alternative geometry is well known and is called spherical geometry. It differs from the Euclidean with the replacement of one axiom, the axiom of parallels. In Euclidean geometry, the axiom of parallels states that given a straight line and a point not on that line that there

is one and only one straight line through that point that never touches the original line no matter how far the lines are extended and that line is called parallel. In spherical geometry, the straight lines are the arcs of great circles, circles on the surface whose center is the center of the sphere. A point to note is that the center of the sphere is not on the surface. In the case of the sphere, all straight lines through a point not on the original line meet the original line, in fact twice. There is a line through a point not on the original line that requires the greatest distance to the nearest intersection of extension before meeting. This line at that point is said to be locally parallel to the original line and this line is unique.

Because in spherical geometry, the axiom of parallels is no longer valid; many of the usual rules of Euclidean geometry no longer hold. The sum of the interior angles of a triangle do not add to π but is always greater than π . Think of a triangle on the sphere of the earth formed by the equator and two lines of longitude. At the equator the two lines are locally parallel and the angle between them and the equator is $\frac{\pi}{2}$. They will meet at the north or south pole at some non-zero angle and thus the sum of all three angles is greater than π . Make a square, a four sided figure of equal length sides with all sides meeting at right angles, on the surface. In contrast to the Euclidean case, it does not stop and start at the same point but over-closes, two of the legs of the square meet before the full side length is achieved. A third test is to make a circle, a set of points that are equidistant from some point, on the earth. The ratio $\frac{\text{circumference}}{2r}$, where r is the distance from the point to the circle defined as the radius, is less than π . To most people this is trivial. The problem is that we are measuring on the surface of the sphere. In the underlying three dimensional space in which the sphere is imbedded, the geometry is Euclidean and the world makes sense. For instance, if, instead of the distance as measured from the center on the sphere, the distance used, r' , is the distance to the axis that is perpendicular to the plane of the circle passing through the center, the usual result that the ratio $\frac{\text{circumference}}{2r'}$ is π . Because this first identification of a non-Euclidean geometry was on an imbedded sphere, these non-Euclidean geometries are now called curved spaces. This is an unfortunate accident of history as we will discuss shortly but it is so prevalent that everyone uses these terms and we will continue to use this nomenclature. Geometries are flat, Euclidean, or curved, non-Euclidean, with an example being a two dimensional spherical surface imbedded in a flat, Euclidean, three space.

17.2 Gaussian Curvature*

The next significant step in the development of modern geometry was taken by the great mathematical physicist Carl Friedrich Gauss. Gauss was interested in the general problem of the shape of a two dimensional surfaces in our three dimensional space. Instead of a plane, the basis for Euclidean geometry, or a sphere the basis for spherical geometry, consider a two dimensional surface in the shape of a pear imbedded in three space. At a point on the surface there are various curvatures, using an intuitive idea that will be articulated with greater care shortly. At the points near the bottom or top of the pear the surface is much like that of a sphere while in the neck region there is a another type of bend. Also at any point, if the region of examination is small enough, the geometry acts as if it is Euclidean or flat, i. e. for a small enough triangle, the sum of the interior angles of triangles is π .

In order to proceed, Gauss needed a definition of curvature. It had to be local, at a point, and agree with our intuitive notions about curvature. The basic idea is that, on a curved surface, as you move through nearby points on the surface, the normal to the surface changes direction. Thus he produced the following construction: as you move over an element of area on the surface, the tip of the unit normal will paint an area on the unit sphere, see Figure 17.1 on page 365. the curvature at a point on the surface

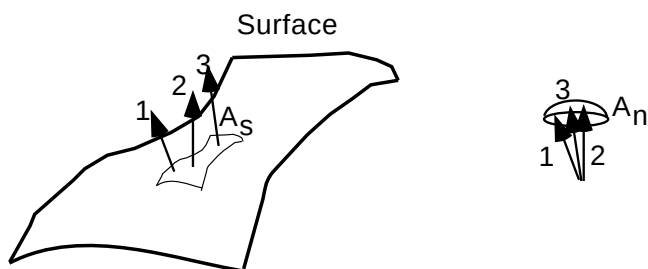


Figure 17.1: **Gauss's Definition of Curvature** Gauss defined curvature as the ratio of the area generated by the tips of the unit normals, A_n , for an element of area, A_s , on the surface as the area on the surface, A_s , goes to zero, $K_G \equiv \lim_{A_s \rightarrow 0} \frac{A_n}{A_s}$.

is the ratio of the area generated by the tips of the unit normals, Area_n , for an element of area, Area_s , on the surface as the area on the surface goes to

zero,

$$K_G \equiv \lim_{\text{Area}_s \rightarrow 0} \frac{\text{Area}_n}{\text{Area}_s}. \quad (17.1)$$

In order to appreciate the subtlety of this construction, let's consider several examples. A flat surface has no curvature since the normal is always the same and thus the Area_n that is generated is that of a point and thus the Area_n is zero. On a sphere of radius r , using the usual spherical coordinates, θ and ϕ , a patch of $\text{Area}_s = r^2 \delta\theta \delta\phi$ and the normal which is the radius vector generates an $\text{Area}_n = \delta\theta \delta\phi$. Thus the curvature is $\frac{1}{r^2}$. This construction shows that this idea of curvature makes sense and that the limit defining it exists for reasonably shaped surfaces¹. Also note that in the limit of large r the curvature is zero. Now consider a point on the neck of the pear mentioned above. Another example and probably easier to visualize is a Pringle potato chip, see Figure 17.2 on page 366.

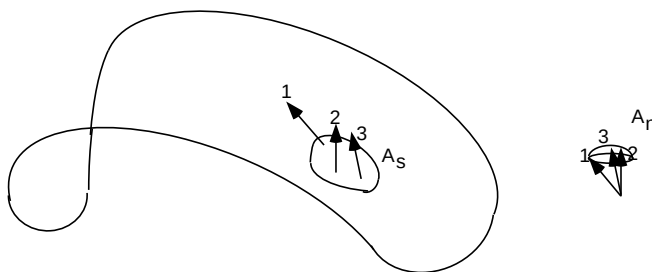


Figure 17.2: **Curvature of a Pringle** A Pringle is an example of a negatively curved surface. The area, A_n , generated by the normals to the surface, A_s , at any point is not zero. The difference between this case and the sphere though is that the area, A_n , is oppositely oriented from that of the area on the surface, A_s , i. e. a right hand coordinate plane on A_s generates a left handed coordinate system on A_n , see Section 17.3 on page 366.

17.3 Example of negative curvature: the Pringle*

I have no idea how Pringles are manufactured, but I will construct my Pringle-like surface by taking a circle of radius R_1 centered on the origin in the two plane, (x, z) , displacing it by R_2 , $R_2 > R_1$, and then making this circle a surface of revolution about the z axis. This generates a torus

¹The definition of reasonably shaped is stated as ‘the surface is locally flat’, flat in the sense of the usual idea of the existence of the derivative.

or donut shape. We can take a segment of the inner surface, the surface toward the z axis, as our Pringle.

The advantage of this construction is that the labeling of points on the surface and the properties of the normal vector can be determined easily. For example, a point on the surface can be determined from the angle around the original circle as measured from the top most point, θ , and the angle of rotation of the circle around the z axis, ϕ both ranging from zero to 2π .

Using these coordinates, a point on the surface is at

$$\begin{aligned} x &= [R_2 - R_1 \sin \theta] \cos \phi \\ y &= [R_2 - R_1 \sin \theta] \sin \phi \\ z &= R_1 \cos \theta, \end{aligned} \tag{17.2}$$

and the area, A_s , generated by incrementing the two coordinates which are orthogonal is $[R_2 - R_1 \sin \theta] R_1 \delta\theta \delta\phi$. The unit normal vector is along the line from the center of the circle at ϕ and the point on the surface or $\hat{n} = (-\sin \theta \cos \phi)\hat{x} + (-\sin \theta \sin \phi)\hat{y} + \cos \theta \hat{z}$. As the area A_s is swept out, the change in the unit normal is $\delta\hat{n} = (-\cos \theta \cos \phi \delta\theta + \sin \theta \sin \phi \delta\phi)\hat{x} + (-\cos \theta \sin \phi \delta\theta - \sin \theta \cos \phi \delta\phi)\hat{y} + (-\sin \theta \delta\theta)\hat{z}$. Again the lines swept out by the coordinate increments are orthogonal and the area, A_n , generated is $\sin \theta \delta\theta \delta\phi$. The Gaussian curvature is $|K_G| = \frac{\sin \theta}{(R_2 - R_1 \sin \theta) R_1}$.

I have put absolute value signs on this result because the curvature in this case is actually negative. You should realize that, if we choose the coordinate directions in A_s to be right handed in the sense that the normal is outward and generated by rotating directed lines at constant θ into lines of constant ϕ , then the area A_n is left handed in the sense that the image traces of constant θ and ϕ are now left handed. This change in orientation of the areas is the indicator that this curvature is negative and thus

$$K_G = -\frac{\sin \theta}{(R_2 - R_1 \sin \theta) R_1}. \tag{17.3}$$

There are other features of this result that are worth commenting on. The obvious result that the curvature is independent of ϕ is expected. More intriguing is the θ dependence, $K_G(\theta)$. Note that, had we done the analysis for the region $\pi < \theta < 2\pi$, the orientation of the image plane would have been the same as the original element of surface and thus, as given by Equation 17.3 on page 367, the curvature is positive. At $\theta = \frac{\pi}{2}$, the curvature is $K_G(\frac{\pi}{2}) = \frac{1}{(R_2 - R_1) R_1}$. The square root of the inverse of the curvature is the geometric mean radius of the two circles that make up the surface at this point, the radius of our original circle and the radius of the surface from

the axis of symmetry, the z axis. This same observation is also valid for the $\theta = \frac{3\pi}{2}$. This is a general result that we will deal with in more detail in the next section, Section 17.4 on page 368. The other interesting set of points is at $\theta = 0$ and $\theta = \pi$. Here, the curvature is zero. This can be looked at in two ways. These points are the transition points from the region of negative curvature, the inside of the torus, and the region of positive curvature, the outside of the torus. Since we expect the curvature to smooth, it is required that the curvature vanish at these points. More significantly, This region really is flat in the sense that it is Euclidean.

Think of a cylinder. The curvature of a cylinder is zero – the normal moves along a line as you move around the cylinder but does not change as you move along the axis of the cylinder. Thus, the area, A_n is zero. It is also important to note that the geometry of the cylinder is the same as that of a flat plane; you can unroll the cylinder onto a flat plane. You can do your geometry in the flat plane with the straight lines being the same as usual and the geometry is Euclidean, interior angles of triangles add to π . Thus the cylinder can be covered entirely by a single flat map. You cannot cover a curved surface entirely with a single flat map. You can cover it locally but at some places the distortion caused by the mapping becomes so severe that points are mapped to lines and visa versa. Think of a map of the earth. The usual atlas projection treats the poles, points, as lines. If you exclude the anomalous points by restricting the range of the coordinates you do not cover the earth with a single map but need more than one flat map. This is also a general property of non-Euclidean spaces. Is a cone flat or curved?

17.4 Curvature and Geodesics

In order to proceed further, we will have to examine the general issue of curves in the surface. An arbitrary path connecting two points in the surface can have lots of turns and bends. There are two sources of these, the bends of the surface and the bends of the path within the surface. We can eliminate the bends within the surface by considering only straight line paths between the points. These, by definition, are the shortest distance paths between the points. Since these may be very curved instead of calling them straight lines a better name is geodesic. One of many theorems of the theory of surfaces is that these are unique. These geodesic paths thus contain the bends of the surface and only those bends. In Section 17.5 on page 370, we will develop a specific differential condition for geodesics that is valid in any coordinate system. For now, we will continue with the more intuitive notions of their

properties.

Remembering that our two surface is imbedded in a flat three space, we can identify three directions at any point on the path, the direction along the local tangent to the path, the direction in the surface perpendicular to that direction (Don't forget that, at a point on the surface for a small enough region, the surface is flat and thus this direction is known. To find it, pick another point on the surface not on the original straight line and draw another geodesic through it. These two paths determine a plane, the tangent plane. All geodesics through p share this tangent plane.), and the direction that is perpendicular to these two. This last direction is locally perpendicular to the surface in the sense that the two other directions have generated the tangent plane at the point. This direction is called the normal direction. We already took advantage of these ideas in the identification of the normal to the surface in the previous section, Section 17.2 on page 365, in which we constructed the Gaussian curvature.

In the neighborhood of the point, the original geodesic is contained in the plane formed from the normal direction and the tangent direction of the geodesic. In the neighborhood of the point p , pick two other points on the original geodesic on opposite sides of p but near p , which will all be in that plane. As is well known from analytic geometry, three points determine a circle. This circle is called the osculating circle. Osculating is from the latin word for "kissing." In some sense, the idea of the osculating circle is the next step up from the tangent. The tangent is determined by two nearby points, determines a magnitude and a direction, and in the limit leads to the concept of the derivative. The osculating circle is determined by three nearby points and utilizes the second derivative, the difference in two tangents, the tangents formed from the original point and the other two points. The inverse of the radius of this osculating circle is called the curvature of the original geodesic. Remember that by using geodesics, there is no bending in the surface. All the bending is due to the surface. There is another geodesic through p that is orthogonal. On that geodesic, construct an osculating circle. Thus at p , for a pair of orthogonal geodesics, there are two osculating circles, one for each of the mutually orthogonal geodesics. As the orientation of this orthogonal pair of geodesics is varied, there will be a direction in which the curvature for each of the orthogonal geodesics will be an extremum. There is no other orientation of the geodesics that have extremum curvatures except trivial variations on this orientation. This last result is called Euler's Theorem. Gauss showed that the Gaussian curvature of the surface as defined in Section 17.2 on page 365 is the product of these

two extremum curvatures,

$$K_G = \frac{1}{R_1 \times R_2}, \quad (17.4)$$

where R_1 and R_2 are the radii of the osculating circles. In addition, the sign of the curvature is determined by the relationship of the two osculating circles. The curvature is positive if both the osculating circles are on the same side of the surface. This is the case for the sphere as discussed earlier. For the Pringle, Section 17.3 on page 366, on the inner edge, the osculating circles are on opposite sides of the surface and this is the signature of negative curvature.

As is always the case with the Gaussian curvature, this curvature is a basic property of the surface and does not depend on the coordinate system that we used to make the construction. Granted that the construction of the curvature is most readily done in a coordinate system that is based on a system of orthogonal geodesics, it is still clear from the nature of the Gauss map and Equation 17.1 on page 366 that the coordinates make the construction possible by staking out the grid but that the local value of the curvature is the same regardless of the coordinate system used. In fact the coordinate system that was used for the torus, Equation 17.2 on page 367, are not geodesic coordinates; the lines of constant ϕ are geodesics but the lines of constant θ are not. This issue will be discussed in much greater detail later, see Sections 17.5 on page 370, 17.6 on page 371.

17.5 The Theorema Egregium and the Line Element

As is clear from Section 17.4 on page 368, Gauss made an extensive study of the nature of surfaces imbedded in a Euclidean three space. He is responsible for many of the insights and theorems that govern understanding of these surfaces. He was, of course, interested in two surfaces imbedded into the larger three space. He recognized the important role of curvature in defining the nature of the surface; to within an orientation and a translation, the surface is determined by its curvature. His most famous theorem in the theory of surfaces was so striking to him that, when he recognized its implications, he gave it the title of the Theorema Egregium. A direct translation of the latin would call this the egregius theorem. The modern sense of egregius: outstandingly bad is not the original meaning. The original use of the word was in the sense of outstandingly good and is what is intended in

the latin. It was later usage that lead to the current interpretation of egregious as outstandingly bad, see [OED 1971]. It seems that modern young people are not the first ones to reverse the meaning of bad and good when describing things. Regardless, the point of Gauss' name for the theorem was in the sense of outstandingly good. Maybe a better translation would be the Extraordinary Theorem.

This theorem proved that all the important properties of the surface could be developed from information that is intrinsic to the surface and did not need to use properties that were determined by the imbedding of the surface in a Euclidean three space or the coordinate system that was used to do the construction of the Gauss map. The only element that is needed to construct curvature is the length of the line element in whatever coordinate system is being used to identify places on the two plane.

The other amazing fact is the realization of Riemann that these techniques developed by Gauss carry over to manifolds of any number of dimensions, Section 17.6 on page 371. The theorem's proof is rather tedious and not really enlightening except in its use of intermediate elements that are very important in our later study of geometry in higher dimensions. We will proceed to look at the situation in higher dimensions directly introducing the concepts as required.

In other words, if when you begin to label points on the surface with some set of coordinate labels and, if at the same time, you determined the actual lengths separating nearby coordinate points, you would have all the information that you need to determine the curvature. For an n -dimensional space, each place is designated by an n -vector, his object is called the metric and is designated g_{ij} and the length of a line element the two nearby places $(x_{10}, x_{20}, \dots, x_{n0})$ on the surface

17.6 Geometry in Four or More Dimensions*

Riemann showed that the methods of analysis used by Gauss for two planes embedded in a three space could be extended to higher dimensional spaces and, in particular, that a space of any dimension could possess non-Euclidean features and that, like the case of the two plane, the non-Euclidean aspects can be manifest by measurements contained to the two planes that constitute the higher dimensional space.

The fundamental construction element in the analysis of the higher dimensional case is called the Riemann tensor. For our purposes it is sufficient to define

The metric is g_{ij} is called the first fundamental form and is defined as

$$g_{\mu\nu} \equiv \frac{\partial x^\gamma}{\partial q^\mu} \frac{\partial x_\gamma}{\partial q^\nu} \quad (17.5)$$

The second fundamental form is $b_{\mu\nu}$ is

$$b_{\mu\nu} = \frac{\partial^2 x^\gamma}{\partial q^\mu \partial q^\nu} n_\gamma \quad (17.6)$$

The mixed Riemann curvature tensor is

$$R_{\nu\gamma\rho}^\mu = \frac{\partial \Gamma_{\nu\rho}^\mu}{\partial q^\gamma} - \frac{\partial \Gamma_{\nu\gamma}^\mu}{\partial q^\rho} + \Gamma_{\nu\rho}^\eta \Gamma_{\eta\gamma}^\mu - \Gamma_{\nu\gamma}^\eta \Gamma_{\eta\rho}^\mu \quad (17.7)$$

where

$$\Gamma_{\nu\gamma}^\mu = g^{\mu\eta} \Gamma_{\nu\gamma,\eta} \quad (17.8)$$

where

$$\Gamma_{\nu\gamma,\eta} = \frac{1}{2} \left\{ \frac{\partial g_{\eta\gamma}}{\partial q^\nu} + \frac{\partial g_{\eta\nu}}{\partial q^\gamma} - \frac{\partial g_{\nu\gamma}}{\partial q^\eta} \right\} \quad (17.9)$$

and the Riemann curvature tensor is

$$R_{\mu\nu\rho\gamma} = R_{\nu\rho\gamma}^\eta g_{\eta\mu} \quad (17.10)$$

The Ricci tensor is

$$R_{\mu\nu} \equiv R_{\mu\rho\nu}^\rho \quad (17.11)$$

The symmetries of the Riemann curvature tensor are:

$$R_{\lambda\mu\nu\kappa} = R_{\nu\kappa\lambda\mu} \quad (17.12)$$

$$R_{\lambda\mu\nu\kappa} = -R_{\mu\lambda\nu\kappa} = -R_{\lambda\mu\kappa\nu} = R_{\mu\lambda\kappa\nu} \quad (17.13)$$

The cyclicity condition:

$$R_{\lambda\mu\nu\kappa} + R_{\lambda\kappa\mu\nu} + R_{\lambda\nu\kappa\mu} = 0 \quad (17.14)$$

17.7 Coordinate Labels in General Relativity

17.8 Einstein Equations

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = -\frac{8\pi G}{c^4} T^{\mu\nu} \quad (17.15)$$

Chapter 18

Effects of Gravitation

18.1 Curvature around a Massive Body

18.2 The Universe

18.2.1 Background Ideas

After 1916, Einstein and others applied the General Theory of Relativity, the modern theory of gravity to the entire universe. This is the subject called cosmology, the study of the physics of the universe. There were earlier attempts to apply the physics known at the time to clarify observations of the world at large but the observational information was rather scant, structures such as galaxies had not yet been seen, and the observations that lead to the idea of the expanding universe had not yet been made, see Section 18.2.4 on page 380. In addition, the conclusions using Newton's theory of gravity suffered from technical difficulties, see Section 18.2.6 on page 389. In that section, we discuss a simple expanding Newtonian expansion since the physics is closer to our experience helps our understanding of the mechanisms of gravitation in determining the nature of the large scale universe. One thing was clear from the beginning. The force system that set the properties of the large scale structure of the universe was gravity. The electromagnetic force was negligible on large scales since it was so strong that matter in the large had to be electrically neutral and thus all the residual electric and magnetic fields were small¹.

¹If there were pockets of one type of charged matter, the repulsion of the like charges would cause it to spread rapidly and dissipate. If there were separated pockets of differently charged matter, there would be significant recombination radiation at the interfaces.

The basic ideas of the nature of the large scale structure are so simple and compelling that it seems that they must be correct and, fortunately, most of the observational data are in complete concordance. Despite this simplicity, the history of the subject is full of surprising turns and it is worthwhile telling some of this history so that we can understand the context of our current understanding and why this is still an exciting and active research field – hardly a week goes by without some new article in the newspapers indicating some controversial measurement. Like all good science, cosmology is now being driven by new experimental results.

Also, It is important to realize that some of the current controversies in our understanding of the operation of the universe are all really at the interface of General Relativity and micro-physics. In this section, we will deal only with the broadly accepted aspects of the subject and leave the issues that emerge from the interaction of the large scale universe with microphysics to a later chapter, see Chapter 23 on page 491. Because of this, in this chapter, we will treat the matter in universe very simply and accept forms of matter that are currently not completely understood.

Einstein had a rather simple outlook on the nature of the universe and its origin. Like Descartes and others before him, he felt that the universe has always been present or at least reasonably stable. This desire was tempered though by the observation that, although the ages of the sun and planets were quite large, there were certainly dynamical processes taking place in the cosmos. This balance between perpetuity and evolution meant that he wanted solutions for the space-time structure of the universe that had stationary or at least quasi-stationary solutions, i. e. solutions that were stable over long periods of time. We should realize that the astronomy of the period was not nearly as advanced as it is today and the observational situation was that, at all distances, the night sky looked the same. Due to the fact that the speed of light is finite, looking at longer distances was the same as looking back in time. It is just that the distances that we being observed were small compared to what we now know are relevant to cosmological questions. Also, we have been observing the universe seriously for only the last few hundred years and on the lifetime of stars and things like that this is but an instant.

As they were originally proposed the equations for the evolution of space-time, the Einstein equations ?? on page ??, did not possess any stationary solutions; there were not enough dimensional parameters to define a time. He realized that there was a simple way to modify the equations and he

added the term now called the cosmological constant.

$$R^{\mu\nu} - \frac{1}{2}Rg^{\mu\nu} - \Lambda g^{\mu\nu} = -8\pi GT^{\mu\nu} \quad (18.1)$$

where Λ is the cosmological constant. With this term added, he was able to construct solutions that were stable over long times. Note that the cosmological constant has the same dimensions as the curvature, R , which is an inverse length squared. The equations now have three fundamental dimensional constants, c , G , and Λ .

Two things changed the situation. The great astronomer Hubble observed that the distant galaxies were receding and the rate of recession was proportional to the distance. We will discuss this observation in more detail in Section 18.2.4 on page 380. This observation freed Einstein from the illusion that the universe had to be stationary and he dropped the added terms. In 1922, Avner Friedmann produced a set of solutions for the structure of space-time for the universe without the use of the cosmological constant, that were very compelling. In a sense, the Hubble observation allowed Einstein to accept the Friedmann solutions as a basis for studies of the structure of the universe. There was another reason that it was easy to accept an expanding universe. Olber predicted that in a stationary universe the night sky should be bright, Section 18.2.3 on page 378. It is not. Thus with the availability of the Friedmann solutions of his equations without the cosmological constant, Einstein dropped the cosmological constant term from his equations and considered his addition of it to them “his greatest mistake” since they were not a natural geometric element.

From the beginning, the Friedmann model of the universe was ambiguous about some of the important features of the universe such as its general geometry. Observational data was not only insufficient to resolve these questions, it was also ambiguous. The primary issue centered on whether or not the expansion was slowing down. The acceleration of the universe is hard to observe directly. We have been observing the universe seriously for only a small fraction of its lifetime. The nature of the acceleration of the universe is determined by the energy/matter terms in the Einstein Equation, Equation ?? on page ?. The density of matter in the universe is also difficult to measure and what measurements were available were not consistent with the dynamics of galaxies and clusters of galaxies, see Section 18.2.6 on page 399. Again, through the Einstein Equations, whether or not the expansion was slowing down or speeding up was connected to the question of whether the average curvature was positive or negative. Neither question could be answered.

As would be expected, the Freedman-Hubble expanding universe was not the only candidate for a model of the universe but its theoretical basis was so compelling that it was widely accepted. Not only was the average energy/matter density of the universe important for acceleration, its makeup was determined by the early thermal history of the universe. Speculation on the nature of the matter in the universe was, of course, determined by the micro-physics of the period. Although the nature of the cosmic distribution of matter was difficult to determine observationally, it was clear early on that the matter in the universe was dominantly light nuclei, electrons, and photons. Using information about this mix, it became possible to assign a temperature to the universe. The observation of the 3^0K background radiation in 1964 by Penzias and Wilson which was predicted within the context of a hot Freedman-Hubble model confirmed this family of models but, because it also contained this requirement that the universe be hot, also led to the acceptance of the unfortunate name – Big Bang Cosmology, see Section 18.2.8 on page 401.

Despite its great success, Big Bang Cosmology had several disturbing features, see Section 18.2.9 on page 401. Although it is natural to expect that the universe was reasonably homogeneous initially, the observational data was simply too good. There were also predictions from micro-physics of particle species that should have been formed in the early universe and are not observed. Not surprisingly, advances in micro-physics now called the Standard Model, see Section 18.2.11 on page 401, implied a mechanism for the initiation of the expansion. This is the Inflationary Theory of the early universe, see Section 23.2 on page 491. More recently, a great deal of observational data of large scale systems has produced more questions and even reopened the question of the role of the cosmological constant, see Section 18.2.10 on page 401. In fact, the experimental situation with the large scale features of the universe is so compelling that many people are turning to questioning our understanding of the micro-physics that we are using. This is an exciting time to be dealing with cosmological physics. There has emerged a Standard Model of Cosmology that has such a secure observational basis that it is now a seriously accepted challenge to the simple Big Bang Theory.

18.2.2 Copernican Principle

It got Galileo into a great deal of trouble with the church but, today, we have no trouble convincing anyone that the earth is not the center of the universe. Not only that, we have no trouble convincing most people that

the sun is also not the center of the universe. You can also get people to accept the idea that the universe is homogeneous, i. e. the laws of physics are the same everywhere or, said another way, the universe is symmetric under space translations. Despite the acceptance of this statement, it is still difficult to convince people that there is no center and no boundary to the universe. The lack of a center or a boundary is an immediate consequence of homogeneity.

It is a fundamental assertion of cosmology that the universe is homogeneous and isotropic. Homogeneous means that there is no experiment that can be carried out that can distinguish one place from any other. This is not to say that there are not special circumstances that exist at certain places; we living on the earth are in a place that is complex, fruitful, and supports life. This is considered a special consequence of the history of this place and an accidental small fluctuation in the vast emptiness of the universe. Similarly, isotropic means that at any point all directions are the same. It is easy to think of spaces that are homogeneous and not isotropic, a cylinder for instance. Regardless, if all points are the same, there can be no point being distinguished as a center or a point on an edge. Regardless, the idea of a homogeneous universe has to be tested experimentally. The idea that the same laws of physics hold everywhere is well tested. Galaxy counts and the background glow of the universe agree with the assertion of homogeneity and isotropy to within expected limits, see Section 18.2.3 on page 378.

In a very real sense, this name of Big Bang does not help. Most explosions have a center and certainly all have an edge. This is contrary to our expectations for the universe. The length scale for which the homogeneity holds is one in which the galaxy is a point and even the fact that we are in a small cluster of galaxies is a local density fluctuation that, although intergalactic, is still on a small scale.

The homogeneity assumption implies that the important physical variables, such as density and so forth, must be independent of position. As stated above, it also means that the laws of physics hold at all places. This is probably the best general test of homogeneity. At large distances, stars work the same way as they do in our galaxy. In addition, all deep sky surveys are consistent with homogeneity at the largest distances. Otherwise, it is hard to make a direct test of homogeneity since we have only occupied this small piece of the cosmos. We are in an awkward situation. We are trying to construct a theory of the universe and we have little experience in it both spatially and temporally.

Isotropy is the statement that at any point all directions are the same. Again, there is no experiment that can differentiate one direction from an-

other. Here we can at least test this hypothesis locally by examining phenomena in all directions. The strongest test of isotropy is the 3^0K background radiation, see Section 18.2.8 on page 401, which can be tested in all directions. Other than expected small fluctuations, it is shockingly isotropic, maybe too much so, see Section 18.2.9 on page 401.

Another important test of the homogeneity and isotropy assumption is the pattern of the Hubble Expansion, see Section 18.2.4 on page 380. The requirement of homogeneity and isotropy restricts the form of the relationship between velocity and distance for remote systems that was observed by Hubble. In fact, the assumptions of homogeneity and isotropy can be used to predict its form uniquely. The fact that it is consistent observationally is verification of these principles.

18.2.3 Olber's Paradox

The Paradox

This was one of the earliest indications that a permanent unchanging universe was not tenable. Basically it is the observation that, in a homogeneous steady universe, the night sky should be bright. Since it is not, there is a problem.

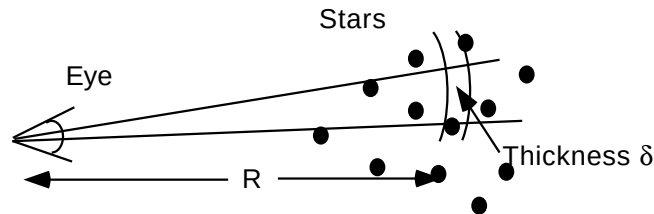


Figure 18.1: **Olber's Paradox** The number of stars that are in a shell of thickness δ in the field of vision at a distance R is proportional to the distance squared. The brightness from each star at the eye falls off as R^{-2} . In a homogeneous universe, the density of stars is the same everywhere and the brightness is the same. Thus the brightness received at the eye is independent of distance and thus the sky should be bright.

The basis of this prediction is that as you look out at the night sky, since you see into some finite opening angle, the number of stars that are in your field of view from some distance R , grows as R^2 , see Figure 18.1 on page 378. At the same time, the brightness of the light from a star at a distance R falls off with distance as R^{-2} . Therefore, in a homogeneous steady universe, the

net light, the number of stars times the brightness per star, received from the stars is independent of the distance. Adding up the contributions from all distances leads to a very large intensity, a bright night sky. Another way to look at it is to realize that in a homogeneous infinite universe along any direction your sight line must ultimately hit a star. This is Olber's Paradox – why is the night sky dark?

Of course, this picture has to be modified in modern times by our realization that the stars that we see are residents in our local galaxy and that, on the large scale, the points of light in the sky are identified not with stars but with galaxies. Substitute the word galaxy for star in the above explanation and you have the modern version of Olber's paradox.

The earlier explanation was that since there was dust or gases in the cosmos, the light falls off faster than R^{-2} and thus we see the dark patches caused by the extra absorption from the intervening material. In a unchanging universe, this explanation will not work. The intervening dust would absorb the light and heat up and glow until its glow balanced the light being absorbed, see Section ?? on page ???. Thus, if the universe is infinite and forever, there should not be a dark night sky.

Although our original interpretation of the dark spaces that are between the bright spots are nothing, we will find out that they are a very interesting and rich structure that providing us with a great deal of information about the early universe. Remember that since the speed of light is finite that seeing further is seeing earlier. With this in mind, we now can ask this dark is and begin to understand the way it came about.

The Modern Resolution

We will get ahead of our story but it is good to understand the modern resolution of Olber's Paradox. The resolution of the paradox is that the universe is expanding and dynamic, see Section 18.2.4 on page 380. When looking out, we are really looking back in time and, at these earlier times, the stars and galaxies have not yet formed. Thus looking out and back in time in most directions we are seeing between the stars and galaxies; directions in which in the universe has no glowing object has yet formed. This is a place that has only the widely scattered matter and radiant energy and by definition is not glowing. The issue is that, if the universe is expanding and we are looking back, this portion of the universe that was once very dense. Not only very dense but also very hot. In the our model, this is light from a hot dense homogeneous aggregation of matter and radiation. In fact, what we see in the interval between the galaxies is the light from the universe

when it was about 300,000 years old. At this time, the universe was a hot sea of matter and mostly photons. The light that comes into our detectors now is the light of last scatter off the surface of this hot body. We cannot see any earlier because that light does not stream out. This is very similar to what we see coming from the sun. Sun light is the light from the outer surface. The interior of the sun is much hotter than the surface but we see light only from the outer most layer which is the surface of last scattering for the light. The interior of the sun is so hot that the atoms are completely ionized and the media is a plasma. The light in the hotter interior layers continues to scatter and thus is thermalized with the ions and electrons. At the surface, the temperature has cooled enough that neutral atoms can form and in this layer the medium is transparent and these photons emerge. This same scenario applies to the universe as a whole. Looking out is looking back. The early universe was not only dense but very hot. The evidence for this today is the high entropy of the universe. There are $10^5 \frac{\text{photons}}{\text{nucleon}}$. Thus, we see only the surface at about 300,000 years because before that the photons and matter are thermal at temperature too high to have atomic forms. After that time, they are sufficiently soft and the matter can combine to neutral atoms and they no longer scatter and these are the ones that come into our detectors. Obviously, this is when the universe has cooled adiabatically to a temperature about the same as that of the surface of the sun, about 3000 °K. If this were the end of the story we would still have a bright sky between the galaxies. In addition, as the light from the surface of last scatter at age 300,000 years travel to the detectors, the universe has expanded and the light has been red shifted to longer wavelengths. The light thus appears to be from a body that has cooled adiabatically to a very low temperature and is identified as the 3°K background radiation, see Section 18.2.8 on page 401. Thus in the modern interpretation, there is no paradox. We do not see the glow of an infinity of stars. Instead, we see the a low temperature glow that is the remnant of the young universe. To the order expected this glow is isotropic, see Section 18.2.9 on page 401.

18.2.4 Hubble Expansion

Originally realized observationally, the Hubble Law was the statement that remote galaxies are moving away from us and that the recession velocity is directly proportional to the distance of the galaxy from us,

$$\vec{v}_{gal} = H\vec{R}. \quad (18.2)$$

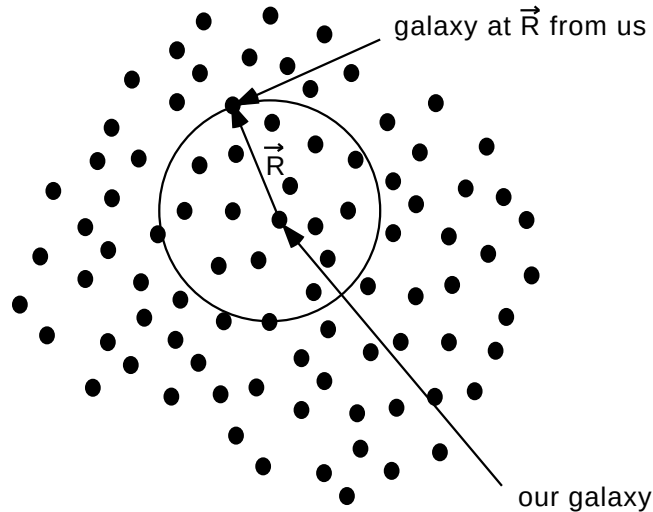


Figure 18.2: **Hubble's Expansion** Hubble's observation that the galaxies are systematically moving away from us. As observed from our galaxy, a galaxy at a relative position $\vec{R}$ from us, has a velocity $\vec{v} = H\vec{R}$. Galaxies at the same distance R have the same speed. The velocity is directed along the relative position away from us. The figure is drawn with our galaxy near the center. You should realize that this is artistic license and does not imply that our galaxy is located in a special part of the universe, see Section 18.2.2 on page 376.

This simple relationship is the basis of all modern cosmology. The original observations were not very compelling, see Figure 18.3 on page 382. Not only were there few data points but the uncertainty in measuring the distances were rather large. In addition, for nearby galaxies, there may be local motion that distorts the effect. Only on really large distances does the cosmic expansion dominate the velocity. In order to verify this relationship, you need separate measures of distance and velocity. The velocity is actually the easier to measure because of the Doppler shift. The distances are more difficult. Using standard stars, such as variable stars which have a very small range of luminosities, the luminosity can be used to gauge the distance. In fact, now a days, the Hubble Law is now one of the best measures of distance for objects far enough away that the local relative motion is negligible when compared to the cosmic motion. The Hubble plot is a convincing affirmation of the Law, See Figure 18.3.

It took a great deal of faith to base a theory of the universe on this data

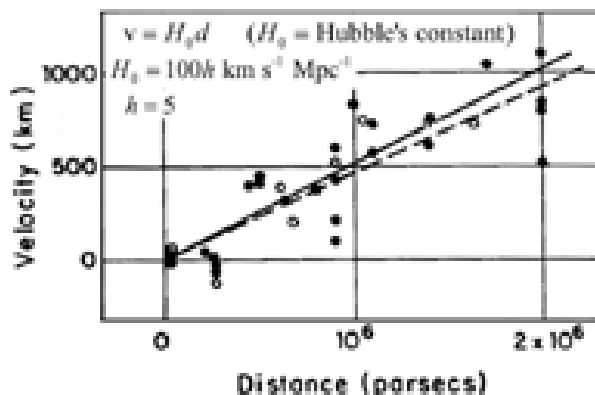


Figure 18.3: **Hubble's Original Plot** The data on the expansion of the universe as presented by Hubble in his original paper in 1929. Note the error in the units on the left axis. Subsequent observations have confirmed the conjecture about the expansion of the universe, see Figure ?? on page ??.

but subsequent analysis has confirmed the conjecture. Note that the Hubble Constant, H , by its definition, is independent of relative displacement, $\vec{R}$, but it can be a function of time. At the time of the laws original formulation, the Hubble Constant was thought to be constant in time but it should be clear that, in any dynamical model of the universe, it will depend on time and in all current models of the universe it does. Of course, if it is a function of time, it is changing at a rate set by the time scale of the universe and, thus, very slowly varying to us. We will thus follow the accepted convention and call it the Hubble Constant despite our anticipation that it varies with time.

A very important fact to note about the Hubble Law is that the Hubble Constant is a scalar; all galaxies at the same distance have the same magnitude of velocity and the direction of that velocity is along the line of sight from us to the galaxy in question, see Figure 18.2 on page 381.

This is a strong conformation of the isotropy of the universe. The only directed quantity that enters the law is the relative displacement. There is no directionality coming from the properties of the universe. The universe is acting like a Pascal fluid, see Section 16.5.1 on page 355. We will take advantage of this fact in preparing simple models of the expansion, see

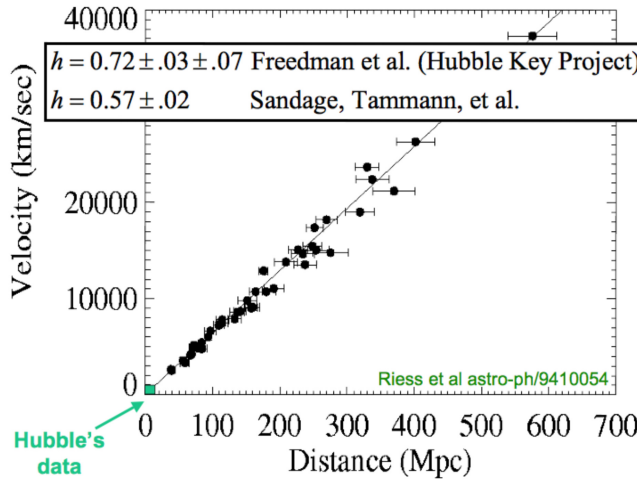


Figure 18.4: **Current Data on Recession of Galaxies** Data from 1994 on the recession of the galaxies. The range of distance is more the seventy times that which was available to Hubble and confirms a linear recession law.

Section 18.2.7 on page 400.

In addition, the Hubble Law is an important confirmation of the homogeneity of the universe. First, we have to make a small technical correction to the form of the law. As discovered by Hubble, the law applied only to reasonably nearby galaxies and the measured velocities were small compared to the speed of light. From our discussion of the hangle, see Section 12.5 on page 289, it should be clear that the hangle is a better measure of relative velocity, $\chi \equiv \tanh^{-1}(\frac{v}{c}) \approx \frac{v}{c}$. It is the additive measure of the Lorentz transformations. Thus,

$$c\vec{\chi} = H\vec{r} \quad (18.3)$$

where $\vec{\chi}$ has magnitude $\chi \equiv \tanh^{-1}(\frac{v}{c})$ and direction along $\frac{\vec{v}}{v}$. The real motivation for this change will be clear as our argument develops. For simplicity of argument, consider a one dimensional universe and an expansion pattern that is arbitrary, $v(R)$ or better $\chi(R)$. Consider the universe and Hubble relationship that would be obtained from a galaxy that is displaced from us by an amount d . Call this galaxy's relationship $\chi_d(R_d)$ where R_d is the distances as measured from that galaxy. We can obtain this pattern with a shift from our galaxy to the new observer galaxy by translating from

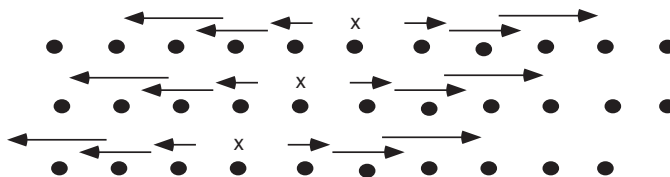


Figure 18.5: **Hubble Law in a Homogeneous Universe** Galaxies distributed uniformly in space are seen in three different reference frames. For example the top universe is viewed from our galaxy. The second down is viewed from our nearest neighbor galaxy. To arrive at the this view, the system is translated and a Lorentz boost is used to bring that galaxy to rest. Our galaxy is now receding. The universe from this new view must be the same as the universe as we view it. Similarly, the bottom line is the universe as viewed from the next galaxy over from our neighbor and looks exactly like our universe.

our location to galaxy d away and doing a Lorentz transformation of $\chi(d)$ to come to rest on that galaxy. In order to have the same physics and thus the same Hubble Law at the original and the new location, $\chi_d(R_d) = \chi(R_d)$. The translation yields a hangle field $\chi_{1_d}(R_d) = \chi(R_d - d)$. The Lorentz transformation yields $\chi_d(R_d) = \chi(R_d - d) + \chi(d) = \chi(R_d)$ or using $R_d = R + d$, $\chi(R) + \chi(d) = \chi(R + d)$. Seeking solutions of the form $a_n R^n$, the only solutions are $n = 0, 1$. The $n = 0$ case is eliminated with the requirement $\chi(0) = 0$. Thus we see that homogeneity, translation symmetry, and Lorentz invariance implies the Hubble relationship that distance which is additive for the translation and χ which is additive for the Lorentz transformation have to be in a linear relationship, Equation 18.3 on page 383.

Probably the most significant feature of the Hubble Law is that it provides for the idea of a finite age for the universe. Reverse all the velocities of expansion and the universe compresses into a dense system, ultimately infinite density in a finite time, see Section 18.2.6 on page 385. This is a particularly simple model for the dynamics of the universe but not overly unrealistic. The fact that the Hubble Law provides us with an dimensionful constant that characterizes the universe is enough to infer a finite lifetime for the universe. The dimension of H is t^{-1} . Thus, $\frac{1}{H}$ is a time. As stated earlier, Section 18.2.1 on page 373, if gravitation is the determining force for the large scale structure of the universe and the universe is homogeneous so that there is only a mass density, there is no time scale in the theory. Thus H^{-1} provides that time scale and, in any reasonable model of the universe,

the age of the universe will be of the order of H^{-1} . In fact, when people quote an age for the universe, they are reporting on the latest estimate of H^{-1} . H^{-1} is difficult to measure precisely but observations are settling around a number of the order of 10^{10} years. This is a very satisfying number in the sense that we have not been able to find anything significantly older, see Section 18.2.5 on page 385.

18.2.5 The Age of the Universe

18.2.6 Models of Expanding Universes*

Milne Universe

The simplest model of the universe that incorporates the expansion is called the Milne Universe. Fill the forward light cone of an origin event of Minkowski space with galaxies at all relative velocities. At first, we will discuss a $(1, 1)$ universe but the generalization to $(1, 3)$ is direct. The space-time is represented in Figure 18.6 on page 386. The set of trajectories are $x_v = vt_v$ or

$$\begin{aligned} t_v &= \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \tau = \tau \cosh \chi \\ x_v &= \frac{v}{\sqrt{1 - \frac{v^2}{c^2}}} \tau = c\tau \sinh \chi \end{aligned} \quad (18.4)$$

where the parameter v or even better the handle or rapidity, χ , see Section 12.5 on page 289, designate the respective trajectories. In Equation 18.4 on page 385, $-\infty < \chi < \infty$ and $0 < \tau < \infty$. In a comoving coordinate system these trajectories would carry the galaxies. In other words, a coordinate system labeled by (χ, τ) would look like an expanding Hubble universe.

The surfaces, in this case curves, of constant τ are space-like and infinite,

$$ds^2 = (dx^2 - c^2 dt^2)|_{\tau=\tau_0} = \left\{ \left(\frac{\partial x}{\partial \chi} \right)^2 - \left(c^2 \frac{\partial t}{\partial \chi} \right)^2 \right\} d\chi^2 = c^2 \tau_0^2 d\chi^2, \quad (18.5)$$

or for fixed τ , $s = c\tau\chi$.

In this universe, the Hubble expansion is obvious, $HS = v$. This follows simply from $\frac{s}{c\tau_0} = \chi$, see Figure 18.7 on page 386. For low velocities, $\chi = \tanh^{-1} \left(\frac{v}{c} \right) \approx \frac{v}{c}$ which implies $H = \frac{1}{\tau_0}$ or in general for a universe at age τ , $H = \frac{1}{\tau}$. Thus, the Hubble constant is not a constant in time.

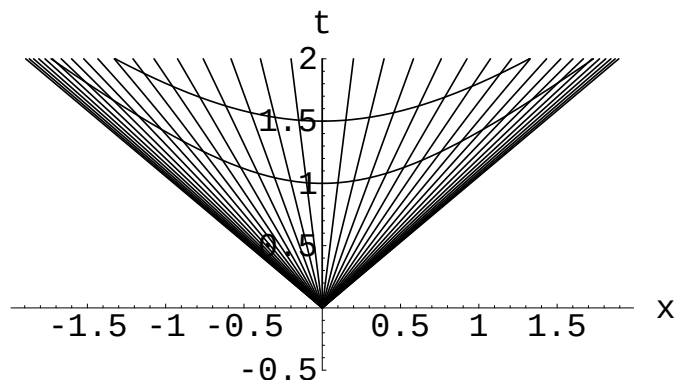


Figure 18.6: **A Milne Universe** The simplest relativistic model of an expanding universe in $(1,1)$. Galaxies carry local coordinates and these are distributed homogeneously in space and hangle. Two space-like surfaces are also shown.

Actually, the Hubble law is better expressed in terms of the hangle since it is the additive measure for the Lorentz transformations, see Section 12.5 on page 289,

$$Hs = c\chi. \quad (18.6)$$

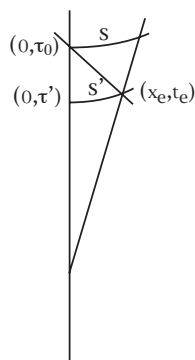


Figure 18.7: **Galaxy Observation in a Milne Universe** A galaxy at a distance s has a hangle $\chi = \frac{s}{c\tau_0}$. This galaxy is viewed at a distance s' and at the same hangle, $\chi = \frac{s'}{c\tau'}$.

In this form, it is also important to note that the Hubble law is then the only velocity or hangle law consistent with the homogeneity of the universe, see Section 18.2.4 on page 380, and Figure 18.5 on page 384. In a $(1,3)$ we

will also see that it is an important indicator of isotropy.

Because for large v , $\frac{v}{c}$ near 1, which implies large s which in turn implies an earlier time and place, see Figure 18.7 on page 386, the original Hubble form must be corrected even further. Calling our galaxy the trajectory at $v = 0 \Rightarrow \chi = 0$ and thus the coordinatizing galaxy, and τ_0 as now, a galaxy currently a distance s from us is seen to be at a distance s' at a universe age of τ' as shown in Figure 18.7 on page 386. The relationship between the times τ_0 and τ' is the Doppler shift times discussed in Section 11.3.3 on page 260 and thus is given by Equation ?? on page ?? where t_a , the receiving time, is τ_0 and τ_e , the emitting time, is τ' or

$$\tau_0 = \tau' \sqrt{\frac{1 + \frac{v}{c}}{1 - \frac{v}{c}}}$$

or in terms of the hangle, $\chi = \frac{s}{c\tau_0}$,

$$\tau_0 = \tau' \sqrt{\frac{1 + \tanh(\chi)}{1 - \tanh(\chi)}}. \quad (18.7)$$

Since the hangle to the current position and time and the observed position and time is the same,

$$\chi = \frac{s}{c\tau_0} = \frac{s'}{c\tau'}. \quad (18.8)$$

Plugging all this together

$$Hs' = \chi \sqrt{\frac{1 - \tanh(\chi)}{1 + \tanh(\chi)}}. \quad (18.9)$$

Today, the observed objects are at such a distance that the recession velocities are very close to c and thus these corrections need to be included.

Obviously, at any time τ_0 , the observer galaxy will have all the other galaxies trajectories in its past at some time. A more interesting and relevant question is how much of any earlier universe is in the past of the observer universe now. Again, Figure 18.7 on page 386 and Equations 18.8 on page 387 and 18.7 on page 387 are relevant. At any time τ_0 , the observer galaxy sees the galaxy labeled χ at age τ' and at distance s' . The relevant point is that although throughout this discussion, we have used the phrase galaxy at distance s' and hangle χ what was really intended has been a patch of an evolving universe. Although a certain patch may at some time contain

a galaxy, because galaxy formation takes time, this patch of universe as seen now may have only cosmic dust and not be seen in the sense that the patch is not luminous and thus conversely is transparent to light from further earlier patches of the universe viewed in that same direction. What we see is the first glowing object in any direction. Fortunately, there do not seem to be many visibly glowing elements now, see Section refSec:OlbersParadox, and in some directions we have a clear view of earlier universes.

To discuss this problem, we need to expand our discussion of Milne universes to $(1, 3)$ spaces so that we can have different directions. In a $(1, 1)$ space you could never see past the nearest galaxy. The definition of a $(1, 3)$ universe is similar to the $(1, 1)$ case with the extra condition that the space portion of the space-time is homogeneous and isotropic. In other words, the observer universe is at the center of a sphere and all directions are identical. Along any direction, a (θ, ϕ) , from the observer universe a galaxy or patch of universe is moving away at a hangle χ and $\chi = Hs$ where H is the Hubble constant and s is the distance from the observer universe now. As we saw earlier, this universe is not only isotropic but symmetric under translations along the direction (θ, ϕ) . A galaxy at a distance s and direction (θ, ϕ) is made the observer universe by boosting by $\chi = \frac{Hs}{c}$. In this case, the coordinates are transformed by

$$\begin{aligned} t &= \tau \cosh\left(\frac{Hs}{c}\right) \\ r &= c\tau \sinh\left(\frac{Hs}{c}\right), \end{aligned} \quad (18.10)$$

or

$$\begin{aligned} dt &= \cosh\left(\frac{Hs}{c}\right) d\tau + \frac{H\tau}{c} \sinh\left(\frac{Hs}{c}\right) ds \\ dr &= \sinh\left(\frac{Hs}{c}\right) c d\tau + H\tau \cosh\left(\frac{Hs}{c}\right) ds. \end{aligned} \quad (18.11)$$

Thus usual underlying Minkowski metric is transformed into

$$\begin{aligned} c^2 dT^2 &\equiv c^2 t^2 - dr^2 - r^2 (d\theta^2 + \sin^2(\theta) d\phi^2) \\ &= c^2 d\tau^2 - H^2 \tau^2 ds^2 \\ &\quad - c^2 \tau^2 \sinh^2\left(\frac{Hs}{c}\right) \{d\theta^2 + \sin^2(\theta) d\phi^2\} \end{aligned} \quad (18.12)$$

With this coordinate system in hand, we can seriously discuss a simple $(1, 3)$ universe. In this context, Figure 18.6 on page 386 is still relevant but

with the x label on the horizontal axis interpreted as the r coordinate in Equations 18.10 on page 388. The figure obviously includes the r coordinate in the direction antipodal to (θ, ϕ) . *shinola*

More relevantly, as described in Section 18.2.3 on page 379, if the universe is expanding and has any entropy, there should be a time when the entire universe is at a very high temperature and dense so that it glows like the surface of a star, the surface of last scatter of the early universe. This universe is one that has a temperature of about 3000°C . The observant reader may wonder why if these patches of universe are the co-moving matter and energy located at that place we do not also include the effects of thermal pressure on the patches. In the Milne universe the patches are inertial by definition and, if we insist that the trajectories are co-moving with the matter and radiation and we have a normal thermal system of matter and radiation, we have an internal inconsistency in the model. This is not the only inconsistency of the Milne model, it has no gravity, but the model is an useful start and subsequently we will add features to complete our model of the universe. For now it is useful to identify the relationship between the galaxy now, τ_0 , at distance s and when it

An Expanding Newtonian Cosmology

For the analysis of this section, we will use non-relativistic physics. This can always work in the sense that we keep the distances and thus the relative velocities small. In addition, we are considering the current epoch of the universe and the energy density is dominated by matter. This analysis will also allow us to separate the effects of ordinary Newtonian Gravitation from the geometric effects of General Relativity. In Section 18.2.6, we will examine a simple General Relativistic Cosmology.

As a measure of the expansion, we will keep track of the distance to some ring of galaxies which are currently at a distance $R(t)$. Following the Hubble Law this ring of galaxies is moving away from us at a speed $\dot{R}(t) = HR$ and these galaxies are gravitationally bound by the sphere of matter contained inside that radius. In the sense of a General Relativistic analysis, we are tracking the expansion in a commoving coordinate system attached to the galaxies at $R(t)$.

Let us examine the dynamics of the expansion as they arise naturally from the Newtonian force law on a galaxy of mass m at $R(t)$. Define the quantity M_{inside} as the mass inside the sphere of radius $R(t)$. Obviously,

the mass density ρ is

$$\rho = \frac{3}{4\pi R(t)^3} M_{inside}, \quad (18.13)$$

Newton's force law yields

$$\begin{aligned} m a(t) &= m \ddot{R}(t) = -\frac{GmM_{inside}}{R(t)^2} \\ &= -\frac{Gm\rho \frac{4\pi R(t)^3}{3}}{R(t)^2} \end{aligned} \quad (18.14)$$

or

$$\ddot{R}(t) = -\frac{4\pi}{3} \rho G R(t). \quad (18.15)$$

Note that this equation is negative definite, gravitation is universally attractive. It is clearly the case that in a homogeneous universe ρ can only depend on time. Thus the first integral of this equation with respect to time and usually identified with the energy has to be handled with care. Replacing ρ with M_{inside} , Equation 18.13, and assuming that M_{inside} is a constant in time, we get

$$\frac{d}{dt} \left(\dot{R}(t)^2 \right) = 2M_{inside} G \left(\frac{d}{dt} \left(\frac{1}{R(t)} \right) \right).$$

Integrating and putting ρ back

$$\dot{R}(t)^2 = \frac{8\pi}{3} \rho G R(t)^2 + K \quad (18.16)$$

where K is a constant of integration. This result has a simple interpretation in terms of the energy of the galaxies at the edge of a sphere of radius R , see Figure 18.2. For a galaxy of mass m , the potential energy is

$$PE = -\frac{GmM_{inside}}{R} = -\frac{4\pi R^2 \rho m G}{3}. \quad (18.17)$$

The kinetic energy for a galaxy of mass m at this distance is

$$KE = \frac{1}{2} m v^2 = \frac{1}{2} m \dot{R}(t)^2. \quad (18.18)$$

Thus, the total energy of the galaxy at $R(t)$ is

$$\begin{aligned} E(R) &= KE + PE \\ &= \frac{1}{2} m \dot{R}(t)^2 - \frac{4\pi R^2 \rho m G}{3} \\ &= \frac{1}{2} m K. \end{aligned}$$

Using the definition of the Hubble constant, $v = HR$, the

$$KE = \frac{1}{2}mH^2R^2. \quad (18.19)$$

Thus the total energy of galaxies at the distance R is

$$\begin{aligned} E(R) &= KE + PE \\ &= mR^2\left\{\frac{1}{2}H^2 - \frac{4}{3}\pi\rho G\right\} \\ &= \frac{mR^2H^2}{2}\left\{1 - \frac{8\pi\rho G}{3H^2}\right\} \end{aligned} \quad (18.20)$$

Note that, because of the homogeneity assumption, H and ρ are independent of position. This energy is positive or negative at all R and is the same sign no matter what the value of R . Thus the sign of this energy is a measure that is universal in the universe. We will find later, Equation 18.26, that, if the energy is negative, the galaxies will stop expanding and later start to fall back. Thus if E is positive, the galaxies will continue to expand indefinitely. Thus, there is a critical mass density of the universe that denotes the boundary between continued indefinite expansion and slow down and ultimate collapse.

Using the dimensional content of H and G , we can define a mass/energy density

$$\rho_{crit} \equiv 3\frac{H^2}{8\pi G}. \quad (18.21)$$

Since H is universal this is the critical density everywhere as expected on the basis of homogeneity. Also since $H = \frac{dR}{dt}$ where as stated above R is a commoving coordinate, if there is acceleration in the commoving coordinate, H and thus the critical density changes with time.

The energy of a galaxy currently at distance R_N from us is

$$E(R_N) = \frac{mH_N^2R_N^2}{2}\left(1 - \frac{\rho_N}{\rho_{crit\ N}}\right), \quad (18.22)$$

where the subscripts N indicate that we are using the current value.

Defining

$$\Omega \equiv \frac{\rho}{\rho_{crit}}, \quad (18.23)$$

where both densities are taken at the same time, this energy is

$$E(R_N) = \frac{mH_N^2R_N^2}{2}(1 - \Omega_N). \quad (18.24)$$

The criteria for the positivity of the expansion energy of the universe in the current epoch is simply whether or not $\Omega_N > 1$.

Equation for evolution of the scale factor

The energy expression, Equation 18.20, can be used to calculate the evolution of $R(t)$. It is interesting to note that we have been calculating a Newtonian Cosmology. There is no field theory of gravity with finite propagation effects or general or special relativistic corrections. This turns out to be okay because of the judicious choice of the comoving coordinate system. Later we will look at the General Relativistic approach, see Section ?? and compare that approach with this one. The advantage of this Newtonian analysis besides its conceptual simplicity is the references to our usual intuition of dynamics. The three things that we are doing that would not have been appropriate to a true Newtonian cosmology is identifying the evolutionary nature of the universe associated with the cosmological expansion, identifying the space time with the galactic expansion, and using as the source of gravity the mass/energy. In addition, none of the current analysis treats issues of geometry of space let alone space time.

Using Equation 18.20, the energy per unit mass of a galaxy on the shell at R_N is

$$\frac{E(R_N)}{m} = \frac{1}{2}H_N^2 R_N^2 - G\frac{4\pi}{3}\rho_N R_N^2 \quad (18.25)$$

In the same notation, the energy for the galaxies in the same shell at a latter time is

$$\begin{aligned} \frac{E(R(t))}{m} &= \frac{1}{2}\left(\frac{dR}{dt}\right)^2 - G\frac{4\pi}{3}\rho_N \frac{R_N^3}{R(t)} \\ &= \frac{1}{2}\left(\frac{dR}{dt}\right)^2 - \frac{H_N^2 R_N^2}{2}\Omega_N \frac{R_N}{R(t)} \end{aligned} \quad (18.26)$$

where the mass/energy contained within the shell, $M_{inside R_N} \equiv \frac{4\pi}{3}\rho_0 R_N^3$, has been conserved.

Equation 18.26 has the same dependence as the one for an object of unit mass being projected to a height, $h = R(t)$, on a body of mass $M_{inside R_N}$. Thus if we require conservation of energy for comoving elements for all time, $E(R(t)) = E(R_N)$, then, if $E(R_N)$ is positive, $\frac{dR}{dt}$ will increase indefinitely and, in a sense, escape the massive body. If $E(R(t))$ is negative, the projected body would have slowed and eventually turn around and start to fall back.

For instance, setting $E(R(t)) = E(R_N)$, or, better said the energy of expansion, Equation 18.24, we find that, if Ω_N is greater than one, the greatest distance that a galaxy, which is currently at distance R_N , will be

from us is

$$\begin{aligned} R_{max} &= \frac{8\pi G\rho_N R_N}{3H_N^2(\Omega_N - 1)} \\ &= R_N \frac{\Omega_N}{\Omega_N - 1}. \end{aligned} \quad (18.27)$$

Similarly, if $\Omega_N < 1$, $\frac{dR}{dt} > 0$ for all time.

The expansion energy can be used to find the general expression for $\frac{dR}{dt}$,

$$\left(\frac{dR}{dt}\right)^2 = H_N^2 R_N^2 \left(1 - \Omega_N \left(1 - \frac{R_N}{R(t)}\right)\right). \quad (18.28)$$

Since $\frac{dR}{dt} > 0$, the positive root is the appropriate choice.

$$\left(\frac{dR}{dt}\right) = H_N R_N \sqrt{1 - \Omega_N \left(1 - \frac{R_N}{R(t)}\right)}. \quad (18.29)$$

Both for reasons of simplicity and ease of interpretation, it is best to use rescaled variables, the distance in units of R_N , $\alpha \equiv \frac{R(t)}{R_N}$ and times in units of H_N^{-1} , $\tau \equiv H_N t$, Equation 18.29, takes the particularly simple form

$$\frac{d\alpha}{d\tau} = \sqrt{1 - \Omega_N \left(1 - \frac{1}{\alpha}\right)}. \quad (18.30)$$

α is often called the scale factor of the universe.

Two features of this result are important to note. Firstly, we have a one parameter, Ω_N , family of universes. Depending on the value of Ω_N , and only on Ω_N , the universe will either forever expand or reverse expansion and collapse. If $\Omega_N > 1$, the term in the square root is always positive and the system will expand forever. If $\Omega_N < 1$, the term with the square root can vanish and the universe will collapse back onto itself. Secondly, the acceleration is easy to compute,

$$\frac{d^2\alpha}{d\tau^2} = -\frac{\Omega_N}{2\alpha^2}. \quad (18.31)$$

There is no surprise in this result. This is Newton's Law of Gravitation applied to the commoving galaxy in these new variables. In fact, the first integral of this expression is the energy of expansion, Equation 18.20. This acceleration is negative definite. Gravity is the only force operating and it is

always attractive. In fact, measurement of a positive acceleration is a special problem for this approach to cosmology. Recent observations indicating the presence of a positive acceleration, [?], present a special problem for this approach. We will see that, in the General Relativistic approach, there is the possibility of positive accelerations but that it will require a form of matter that is not consistent with our current understanding of microscopic physics or an uncomfortable value for the cosmological constant, see Section ??.

In addition, Equation 18.30 is easy to integrate although the closed form solution is not particularly useful. The boundary condition is obviously $\alpha(\tau = \text{now}) = 1$. Choosing the origin of time such that $\tau = \text{now} = 1$, we can plot the evolution of the scale factor of for times earlier than now, see Figure 18.8 and in Figure 18.9 for longer times for three values of the Ω_N ; $\Omega_N = 0.5, \Omega_N = 1$, and $\Omega_N = 1.5$. In Figure 18.8, The universe starts from the time that the scale factor vanishes. It can be seen from Figure 18.8 that the current age of the universe is not strongly dependent on Ω_N and is the order of the inverse Hubble constant as expected.

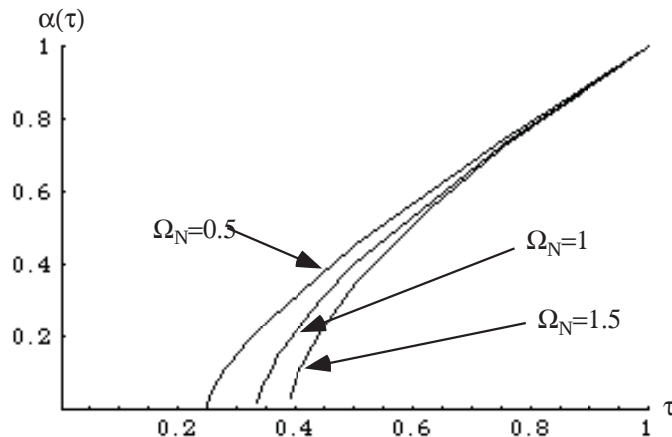


Figure 18.8: **Evolution of the scale factor for early times** The evolution of the scale factor depends only on the mass/energy in the universe. Three cases for the mass/energy density are shown: $\Omega_N = 0.5$ which is an ever expanding universe, $\Omega_N = 0.5$ which is at the transition between collapsing and ever expanding, and $\Omega_N = 1.5$ which is collapsing universe

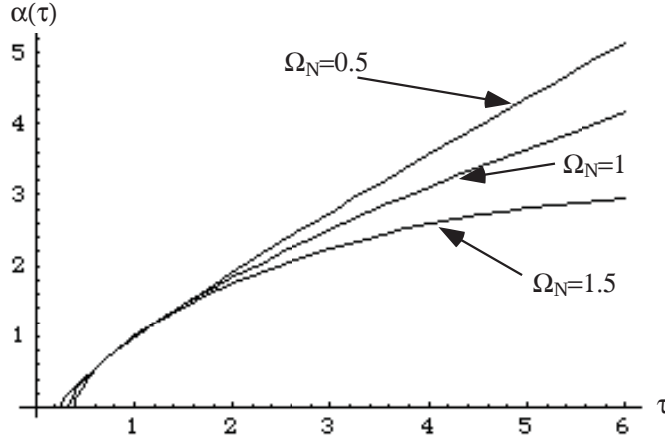


Figure 18.9: **Long time dependence of the scale factor** The evolution of the scale factor depends only on the mass/energy in the universe.

Evolution of Density

Using the fact that the mass/energy in any comoving shell is conserved, $M_{inside R(t)} = M_{inside R_N}$, the density scaling law becomes

$$\rho = \frac{\rho_N}{\alpha^3}. \quad (18.32)$$

Putting this expression into Equation 18.31, the acceleration of the scale factor becomes

$$\frac{d^2\alpha}{d\tau^2} = -\frac{4\pi}{3}\rho\alpha\frac{G}{H_N^2}. \quad (18.33)$$

This result shows the Newtonian gravitational basis for the acceleration of the scale factor, it is not as useful as it may appear since we need to find the evolution of the density to integrate it. From the density scaling law, Equation 18.32,

$$\frac{d\rho}{d\tau} = -3\frac{\rho}{\alpha}\frac{d\alpha}{d\tau} \quad (18.34)$$

$$= -3\rho\frac{H}{H_N}. \quad (18.35)$$

Again this expression is not as useful as it seems. We require the solution for $H(\tau)$ in order to integrate it.

Similarly, the evolution of the density in terms of α follows from the scaling law, Equation 18.32 and Equation 18.34 as

$$\begin{aligned}\frac{d\rho}{d\tau} &= -3\frac{\rho_N}{\alpha^4}\frac{d\alpha}{d\tau} \\ &= -3\frac{\rho_N}{\alpha^4}\sqrt{1 - \Omega_N\left(1 - \frac{1}{\alpha}\right)}.\end{aligned}\quad (18.36)$$

Given the solution of Equation 18.30, this equation can be integrated to give the evolution of the density.

Evolution of H

Given the acceleration of the scale factor, Equation 18.31, it is straight forward to get the equation for the evolution of H

$$\begin{aligned}\frac{d}{d\tau}\left(\frac{H}{H_N}\right) &= \frac{d}{d\tau}\left(\frac{\frac{d\alpha}{d\tau}}{\alpha}\right) \\ &= \frac{\frac{d^2\alpha}{d\tau^2}}{\alpha} - \frac{\left(\frac{d\alpha}{d\tau}\right)^2}{\alpha^2} \\ &= -\frac{\Omega_N}{2\alpha^3} - \left(\frac{H}{H_N}\right)^2 \\ &= -\frac{1}{\alpha^2}\left(1 - \Omega_N + \frac{3\Omega_N}{2\alpha}\right).\end{aligned}\quad (18.37)$$

which is manifestly negative definite as expected.

This model contains all of the large scale features of what is termed the “Big Bang” cosmology. There are features of this model that have not been dealt with such as the nature of the mass/energy in the universe. These will be dealt with later when microphysics has been included. Suffice at this point to say that the matter considered is ordinary matter that obeys all the usual rules of macroscopic and microscopic matter physics such as thermodynamics and our latest discoveries of elementary particle physics. These matters will all be discussed in Section 19.2.1 on page 404. In addition, there has been no discussion of the space/time geometry. This will require the use of General Relativity which is dealt with in Section 18.2.7 on page 400.

One property of the mass/energy that is clearly important is the amount. Ω_N is the only parameter that labels our models of the universe and thus determines whether the universe will expand forever or will eventually fall back on itself and collapse, see Section 18.2.6

Friedman Robertson Walker Space-Time

A Friedman Robertson Walker space-time is homogeneous and isotropic in space and obeys the Einstein equation in a $(1, 3)$ space. We have experience with homogeneous isotropic two spaces in three space. Embedding the three generic examples, we have the two sphere constrained by $x^2 + y^2 + z^2 = R^2$, the flat plane with the constraint $z = 0$, and the hyperboloid of revolution with the constraint $x^2 + y^2 - c^2 t^2 = R^2$ in a $(1, 2)$ Minkowski space.

These can be unified into a simple single form for the metric. Starting with the two sphere, $x^2 + y^2 + z^2 = R^2 \Rightarrow 2x dx + 2y dy + 2z dz = 0 \Rightarrow dz = -\frac{xdx+yd y}{\sqrt{R^2-x^2-y^2}}$,

$$\begin{aligned} ds^2 &= dx^2 + dy^2 + dz^2 \\ &= dx^2 + dy^2 + \left(\frac{xdx + ydy}{\sqrt{R^2 - x^2 - y^2}} \right)^2, \end{aligned}$$

where R is the constant radius of the two sphere.

Going to polar coordinates in the xy plane, $x = r \cos \theta$, and $y = r \sin \theta$,

$$dl^2 = \frac{R^2 dr^2}{R^2 - r^2} + r^2 d\theta^2, \quad (18.38)$$

or defining a dimensionless radius, $r_0 \equiv \frac{r}{R}$,

$$dl^2 = R^2 \left\{ \frac{dr_0^2}{1 - r_0^2} + r_0^2 d\theta^2 \right\}. \quad (18.39)$$

The homogeneous isotropic negative curvature two surface is the hyperboloid of revolution $x^2 + y^2 - c^2 t^2 = -R^2$ embedded in a $(1, 2)$ Minkowski space-time. This is the surface at fixed proper time $c\tau = R$. Following the same pattern as before, $cdt = \frac{xdx+ydy}{R^2+x^2+y^2}$ or

$$dl^2 = dx^2 + dy^2 - \left(\frac{xdx + ydy}{R^2 + x^2 + y^2} \right)^2.$$

Again going to polar coordinates,

$$dl^2 = \frac{R^2 dr^2}{R^2 + r^2} + r^2 d\theta^2, \quad (18.40)$$

Changing to the dimensionless radius,

$$dl^2 = R^2 \left\{ \frac{dr_0^2}{1 + r_0^2} + r_0^2 d\theta^2 \right\}. \quad (18.41)$$

The flat case is obtained by taking the limit $R \rightarrow \infty$ in either Equation 18.38 or 18.44 or

$$dl^2 = dr^2 + r^2 d\theta^2. \quad (18.42)$$

Using the fact that the curvature

$$\kappa = \begin{cases} \frac{1}{R^2} & : \text{ for the positively curved case} \\ 0 & : \text{ for the flat case} \\ -\frac{1}{R^2} & : \text{ for the negatively curved case} \end{cases}, \quad (18.43)$$

all three cases are given by

$$dl^2 = \frac{dr^2}{1 - \kappa r^2} + r^2 d\theta^2. \quad (18.44)$$

or

$$dl^2 = R^2 \left\{ \frac{dr_0^2}{1 - \kappa r_0^2} + r_0^2 d\theta^2 \right\}, \quad (18.45)$$

where

$$k = \begin{cases} 1 & : \text{ for the positively curved case} \\ 0 & : \text{ for the flat case} \\ -1 & : \text{ for the negatively curved case} \end{cases} \quad (18.46)$$

is the sign of the curvature.

The extension to an isotropic homogeneous three space in a $(1, 3)$ space is found by replacing the angular measure from the unit one sphere, the circle, with the usual measure on the unit two sphere. Thus our metric is

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left\{ \frac{dr^2}{1 - \kappa r^2} + r^2 d\theta^2 \right\}, \quad (18.47)$$

where $R(t)$ is called the scale factor of the universe. For this metric, the non-zero curvature components are for the Ricci tensor

$$\begin{aligned} R_{00} &= -3 \frac{\ddot{R}}{R} \\ R_{ij} &= - \left[\frac{\ddot{R}}{R} + 2 \frac{\dot{R}^2}{R^2} + 2 \frac{k}{R^2} \right] g_{ij} \end{aligned} \quad (18.48)$$

and the curvature scalar is

$$R = -6 \left[\frac{\ddot{R}}{R} + \frac{\dot{R}^2}{R^2} + \frac{k}{R^2} \right] \quad (18.49)$$

Shinola

Missing Mass

As can be expected, it is very difficult to measure the mass/energy density of the universe. There are several reasons for this. We are not in a region of the universe that is typical. Our planet is in a solar system about a star that is in a galaxy that is a part of a local cluster of galaxies. The star that we orbit is at least a second generation star and thus the matter that is around us is not cosmic in origin. Most significantly, until very recently, the only observable tool was the light from or absorbed by the matter. In fact, all that you can directly observe is the luminous matter. You have to infer the mass from the nature of the light.

Luminous Matter

The standard procedure is to look at the glow of standard objects whose mass can be inferred from other properties of the object. Models of stellar structure provide a tight relationship between the glow of stars and their mass. Galaxies are made of stars and thus we can infer the mass of the glowing material of the galaxies. Thus a ratio of luminosity to mass and assumed proportionality can be established for the mass associated with all the luminous objects observed in the universe and from this a density of matter. In all cases, for the systems in consideration, the mass dominates the mass/energy density. Of course, there could be cool dark objects and often you will hear arguments for their contribution to the mass density of the universe. The occurrence of these kinds of things at a rate sufficient to contribute significantly to the mass density provides theoretical astronomers with lots of speculative freedom and opportunities to publish. It should also be clear that this estimate is at best correct to within a factor of two. The current best estimate is that the mass associated with luminous matter is

$$\Omega_{N_{lum}} \approx 0.01 \quad (18.50)$$

or less.

Gravitational Mass

Besides using the luminous matter, we can infer mass from its gravitational effects. Assuming that the stars in galaxies are gravitationally bound and If you look at the speed of stars then you can estimate the mass that is the source of the gravity that is binding them.

Figure of rotation curves

The mass required to provide dynamic equilibrium is approximately 10 times the luminous mass. This increases the critical density to

$$\Omega_{N_{grav}} \approx 0.1 \quad (18.51)$$

In addition, the galaxies are clustered. We are in group called the local cluster. If you assume that these clusters are not accidental combinations but are also gravitationally bound, there is dark mass between the galaxies. Adding in this mass increases the critical density to

$$\Omega_{N_{clus}} \approx 0.2 \quad (18.52)$$

Einstein had a theoretical prejudice for a universe with $\Omega_N > 1$. We have not yet discussed the space time structure of the universe, see Section 18.2.7, but in the same way that the values of Ω_N determines the collapse or expansion of the universe, it determines the nature of the geometry. This should be no surprise since a collapse would imply a finite timelike geodesic. In a fully relativistic treatment, a finite timelike world line implies finite space-like geodesics and, thus, a finite universe. In this case, there is no need for boundary conditions on the universe at its start. Thus, there was a reason to feel that there should be more matter in the universe than that which was observed by the these two methods. This became known as the “missing mass” problem. More recently, there has been a theoretical prejudice for the case $\Omega_N = 1$. This is driven by the need for an inflationary phase at the start of the universe, see Section 23.2. Regardless, there was a strong desire to find more matter than could be seen, luminous, or felt, gravitational. The problem now is that positive accelerations have now been observed and the best description of the large scale structure of the universe, the “Standard Model” , Section 18.2.11 requires dark matter and dark energy. Neither of these seem to be consistent with our current understanding of the nature of matter as developed in microphysics.

shinola

18.2.7 The Space Time Structure

Shinola

Before elaborating further on the difficulties with a simple expansion model of the universe, we will redo the analysis of the above section, Section 18.2.6, using the tools of general relativity still restricting ourselves to a simple picture of the nature of the matter in the universe. This will enable us to understand the geometry of the universe and to better understand the role of the dark energy.

Using the arguments of homogeneity and isotropy you can show that the general form of the metric is

$$c^2 d\tau^2 = c^2 dt^2 - R^2(t) \left\{ \frac{dr^2}{1 - kr^2} + r^2 d^2\Omega \right\} \quad (18.53)$$

where $R(t)$ is a function of time and is determined by Einstein's equation if you know the energy and momentum densities. $R(t)$ is called the scale factor of the universe. k is a constant that takes on the values 1, 0, or -1.

Using this metric you can get all the curvatures. The three three space curvatures are equal and are $\frac{k}{R^2}$. Thus the three space is positively curved for $k = 1$. It is flat if $k = 0$ and negatively curved for $k = -1$. For $k = 1$ the geodesics are all finite in length and thus have finite volume. The other two spaces have infinite geodesics and thus infinite volumes. We can thus identify the three cases that we have here with the values of the critical density that we had above. $\Omega_N > 1$ is the closed positively curved universe. $\Omega_N = 1$ is the case of the flat space and $\Omega_N < 1$ is the negatively curved universe. These last two cases have infinitely long straight lines.

Whether or not the universe is finite or infinite is determined by the mass density of the universe. It is clear that the value of Ω_N is an important parameter.

18.2.8 Black Body Background

18.2.9 Problems with the Expanding Universe

18.2.10 The Cosmological Constant

18.2.11 The Standard Model of the Universe

Chapter 19

Introduction to Quantum Theory

19.1 Introduction

Toward the later part of the 19th century, several new observations caused people to question the basic ideas that were the cornerstone of the physics of the time. Primary among these was the growing acceptance of an atomic theory of matter. Successful predictions in chemistry and the development of a statistical particle basis for thermal phenomena being key. To us today, the atomic basis of matter is so obvious that we do not question it. On the other hand to the physicist of the early 19th century, the continuous nature of matter was obvious. Given the technology of the day, any attempt to measure the size of the atom was impossible. The scale of phenomena at which the discreteness of the atoms could be observed was inconceivably small, see Sec 1.4.2 on “Things That Everyone Should Know.” Even Dalton, the father of Chemistry, had his doubts about the atomic nature of matter. Although his model of atoms described with great success the rules of chemical composition, he could not understand chemical structures like gaseous O₂. If two oxygen atoms were attracted and bound to one another to form a stable O₂, why wouldn’t two O₂’s be even more attracted to one another and form an O₄? Continuing this line of reasoning, he would believe that oxygen should be a solid and not a gas. Regardless of these conceptual difficulties, by the later part of the 19th century, because of its success in chemistry and statistical mechanics, the atomic theory became dominant and, with it, the idea that the atom had definite properties and a definite size. At this same time, the discover of the electron provided an opportu-

nity for a model of atoms based on bound electrons. The development of a model of atoms became a dominant effort of the period. We now know that all efforts based on a classical model were destined to failure. It would take the development of quantum mechanics to solve this problem and it would be some time after the first successes of quantum mechanics before a satisfactory model of the atom was possible. Today, the success of quantum mechanics in “explaining” chemistry is phenomenal. We can compute complex reactions and propose molecular configurations before they are seen in the laboratory.

Although it was the successful application of quantum mechanics to atomic theory in the early parts of 20th century that lead to its acceptance, the conceptual development of quantum mechanics begins much earlier and deals with a much simpler system – light. We will start with this pre-atom development not only because it is the historically correct approach to the study of quantum mechanics but also because, in its simplicity, it makes the conceptual basis of the theory most clear. For this we will need some understanding of thermal processes. The fact that the application of quantum mechanics to atoms was so successful makes the description of the quantum mechanics of contained massive objects such as electrons.

19.2 Blackbody Radiation

19.2.1 Background on Thermal Processes

Before discussing the phenomena associated with what we call black body radiation, we will have to understand some of the basic issues involved in thermal phenomena. The official name for the field is thermodynamics, the study of processes (dynamics) developed by temperature (thermo) differences. The classic example is the aeolipile of Hero of Alexandria which converted heat energy to rotational energy in the first century, see Figure 19.1. Thus ideas about the use of heat to do useful work began in the earliest recorded times. As is often the case, the understanding of the phenomena required a more careful articulation of the processes and more refined development of the vocabulary of the processes. This also almost always involves a qwerty use of words with implied meanings. The earlier sentence on the aeolipile is typical since we have been rather cavalier about the words heat and energy. Hopefully, we will be able to clarify these usages.

At some time in prehistory, humans became conscious of the sense of hot and cold. With the advent of the scientific era, there was a systematic effort to measure the primary physiological response to what today we call



Figure 19.1: **The Aeolipile of Hero** A small fire boils water and the pressure builds up to ultimately force steam out of angled jets. The flow of steam from the jets rotate the sphere.

temperature difference. Although originally crude, by the 17th century, the development of precise instruments for the measurement of temperature became reasonably widely accepted. It is worthwhile to point out that this idea of temperature is not to be identified with the physiologically phenomena of hot and cold. The deviation of hot and cold from the measured entity called temperature is observed in the statement that the outside temperature is 95⁰ F but feels like 100⁰; a strong indication that something subtle was going on. This scientific epoch in which the measurement of temperature was established is best called the period of thermometry.

It was because of these careful measurements in thermometry and the development of an instrument called the calorimeter that it became clear that beside temperature there was an important concept called quantity of heat. It with the invention of the calorimeter that the effects of heat flow could be carefully measured. The independence of temperature and this quantity of heat idea was most obvious in cases in which the material in question went through a phase change, i. e. water to steam or ice to water, without a temperature change. The idea being that the same matter, say water stuff, had different heat quantities that was intrinsic to their manifest

nature. Ice had less heat in it than liquid water that had less heat content than steam. At first this quantity of heat was thought to be something special created to change the system either by a change in temperature or a change in state, liquid or solid. The next major development in our understanding of the thermal behavior was brought about by the experiments of J. P. Joule, the namesake for our SI unit of energy, which lead to the realization that heat, the unit of which was the calorie, defined to be the amount of heat that is needed to change the temperature of liquid water at 0°C by one degree, was, in fact, just another form of process energy such as mechanical or motional energy. In other words, if carefully done, the aeolipile's rotational energy would increase in direct proportion to the rate of heat flow from the flame once the conversion of units from calories per second to Joules per second. Of course, this is in an idealized world. In the realistic situation, there is heat flow to the surrounding environment and to mechanical friction. If the friction forces at the rotational axis of the aeolipile are a constant, the work done against friction will increase as the flame continues to provide heat energy. In fact, the aeolipile's rotational energy does increase when first exposed to the flame. This increase in the rotation rate continues until the energy loss to friction and the surrounding environment consume all the heat input from the flame and, thus, the aeolipile's equilibrium state is a constant rotational energy established at this balance point of dissipation. There are two important ideas here. The first is that heat has meaning only during process. It is not in the state of the system. Note that this is contrary to the ideas expressed in the previous paragraph in which heat content or latent heat was identified. The other important point is the idea that heat flow was the same as an energy flow as long as the correct units were used. This is the basis for the entry in 'Things Everyone Should Know' titled the 'energy content of food'. These two ideas are combined and are expressed as the first law of Thermodynamics: The energy processes manifests themselves in many forms, mechanical, thermal or heat, material flow, etc. that can morph into each other but in all these processes once suitable units are identified the total energy is always conserved. For our purpose of establishing the origins of quantum physics, it would be sufficient to understand thermodynamics at the level of the first law but thermal processes are so ubiquitous and important and they are so subtle that it is worth a full articulation of all of the laws that we will continue to develop the history and introduce the other three laws of thermodynamics.

The development of the steam engine in the late 18th and early 19th centuries provided a strong stimulus to articulating a complete theory of

heat and heat processes and their relationship to mechanical processes. The development of a this theoretical framework started its full articulation in this later period when Sadi Carnot, a French military engineer, published his book, *Reflections on the Motive Power of Fire and on Machines Fitted to Develop that Power* in 1824. Unfortunately at the time, Carnot's ideas were generally ignored until picked up later in the 19th century by R. J. E. Clausius and Lord Kelvin who articulated what is now called the second law of thermodynamics. The primary goal of Carnot was an understanding of the efficiency of heat engines such as the steam engine. To do so, he developed a theory of heat flow as energy falling from a higher to a lower temperature. Since he was working on the engines, he also showed the importance of cyclic, return to their original state, machines such as steam engines. The final piece of his contribution was showing the universal importance for thermodynamics of a specific cyclic process now called the Carnot engine or cycle. This observation of Carnot provide the basis for the 2nd Law of Thermodynamics, The statements of Clausius and Kelvin as modified by Planck of the second law is

Kelvin-Planck: No process is possible whose sole result is the extraction of energy from a heat bath, and the conversion of all that energy into work.

Clausius: No process is possible whose sole result is the transfer of heat from a cooler to a hotter body.

In more careful language than that used by Clausius at that time, we would use for the word cooler the phrase lower temperature and for hotter the phrase higher temperature. It is important to note that the Kelvin-Planck implies an intrinsic inefficiency in processes that are driven by heat. We will see in the 1st Law that energy is conserved but regardless all the heat energy is not available to do useful work. This intrinsic inefficiency is often quite large and as implied above cannot be reduced or eliminated.

Modern treatments of thermodynamics are generally based on the realization that matter is composed of small entities called atoms. This is historically backwards. Thermodynamics was fully developed and later Maxwell and others showed that the atomic hypothesis could be used to “explain” the thermal properties of gases; the atoms act in simple mechanical ways as articulated by Newton but macroscopic collections of these microscopic constituents have emergent behaviors such as temperature that emerge from the complications of underlying system [Atkins 2007]. This realization is correct but misses the point that the laws of thermodynamics were established prior

to the acceptance of the atomic theory of matter and these laws constitute a complete, coherent and correct description of all processes using or generating heat flow which constitutes almost all evolutions of macroscopic systems. **The behavior of these macroscopic systems is simply summarized in four laws and these laws like all laws in physics are found to hold in all cases.**

19.2.2 Basic Thermodynamics

Thermodynamics emerged as a formal system from early studies of the use of heat energy to produce useful work, basically the effort to catalogue the operations of the steam engine. What emerged was a beautiful complete set of theoretical constructions that could then be used to describe operations of huge classes of phenomena including biology and chemistry. In fact, the ubiquitous nature of thermal processes is one of the great steps forward in our understanding of natural phenomena. For that reason and because of the subtlety of the subject, we will spend more time on this than is needed for our establishment of basis for modern physics. This is just something that all people should know. Also the formal constructs of the theory developed after some of the vocabulary of the phenomena had become a part of the vernacular and, as often happens, the words of thermodynamics have a vernacular connotation that can at times be misleading. Also as stated above, thermodynamics was developed classically and is a consistent construction that matured before the establishment of the atomic description of matter. The atomic theory of matter provides a valuable example of a basis for matter that provides a picture of the “why” for thermal phenomena but is actually only an example of a possible system that behaves thermally. This leads to an interesting anomaly in the use of thermal descriptions that the underlying theoretical constructions are “explained” by the picture of how the atoms behave. This is misleading since the concepts of thermal systems are actually very general and stand on their own. This leads to the interesting speculation that the success of classical thermodynamics was among the first examples of verification of the atomic nature of matter before this theory was articulated. The interesting point though is that the constructions of thermodynamics are general enough that the radical transformation of our picture matter after the discovery of the atom and the subsequent development of quantum mechanics does not weaken the edifice that thermodynamics gave us about heat transfer processes. In the following, we will deal with the ideas of thermodynamics classically and without the crutch of the idea of the disorder of systems of “atoms”.

19.2.3 –1st Law: Equilibrium & Variables

The usual exposition of Thermodynamics is the articulation of the four basic laws, 0 through 3. but it is very important to understand the time scale and meaning of processes in Thermodynamics. Although the greek word thermodynamics is translated as heat processes, thermodynamic processes take place on an long time scale compared to the usual processes of mechanical and electrical phenomena. This is due to the fact that thermodynamic descriptors such as volume, temperature, pressure... have to be articulated consistently for system or subsystems under consideration. For example, a macroscopic volume of gas confined to a space with a piston is said to be at temperature T or pressure P. The piston is moved to change the volume. if this change is done too rapidly there will not be enough time for the system, the volume of gas, to have the same temperature or pressure throughout. But thermodynamic analysis requires the state variables to be defined for the entire system under consideration.¹ This need to have the state variables of the system defined requires time on some usually rather long time scale. This is generally summarized as allowing all the fast things to have happened before the slow processes of the thermodynamic changes take place. The processes of thermodynamics thus has two time scales. When the system is manipulated on time scales that are too short to allow ‘thermalization’ the process is irreversible. When the time scale is long enough for the ‘thermalization’ to be obtained at each infinitesimal step the process is said to be reversible; so slow that at any step in the process the change can be reversed back to the original configuration. In this introduction to thermodynamics, we will deal with reversible processes.

It is also important to have some understanding of thermodynamic variables. These are organized in two different ways, state and process. For specificity, you should consider the simplest system, a simple gas in a containing volume. For Thermodynamic purposes the situation is described by state labels; pressure, P, volume, V, amount of gas², N, and the tempera-

¹In cases like this, there are two ways around the problem. One is to subdivide the system into smaller parts so that in each part equilibrium is meaningful. The other is to allow irreversible processes but to determine the final state of the process by means of conservation laws. The 1st Law discussed later is a good example of these.

²Since for gases in reasonable circumstances, the most useful measure of how much gas is in the volume is measured by the number of atoms in the volume. In the case of gases, the mass unit is the gram-mole which using Avogadro’s number, $N = 6 \times 10^{23} \frac{\text{atoms}}{\text{gm-mole}}$, sets the number of atoms in the volume. The gram mole is the mass in grams divided by the atomic mass number of that atomic specie, i.e. for helium whose atomic number, the number of nucleons in a helium nucleus, is 4, the mass of N grams is $\frac{N}{4}$ in gm moles.

ture T . These variables are constrained by an equation of state which for the simple gas such as helium at modest pressures is $PV = NRT$ where R is a universal constant called the ideal gas constant³. A system with these state variables is called a PVT system. Many students in their first exposure to thermodynamics have the impression that the only systems that are dealt with are these ideal gases. This is not the case. We could be analyzing the energetics of electrical storage batteries which in the simplest case is a Voltage, V , charge Z , and temperature T system. We could be dealing with elastic extensive systems more commonly called rubber bands. Here we would have a tension, t , length, l , temperature, T , system. For all these, we would identify a constraining arrangement of the variables called the equation of state.

The other important aspect of the thermodynamical variables is how the variables scale with the system size. Simply stated, the variables come in two categories: extensive and intensive. A thermodynamic variable that scales with the system size is an extensive variable. If you make an exact copy of the system with which you are dealing, remove the separator between them and combine them into one system, extensive variables will double and intensive variable will remain the same. In the ideal gas case, the volume and the number of atoms are extensive and the pressure and temperature are intensive. In the battery case, the charge is extensive and the voltage and temperature are intensive. In the rubber band, the length is extensive and the tension and temperature are intensive. It is important to note that since the classifications, extensive and intensive, are based on a scaling law of size of system that the ratio of two extensive variables is an intensive variable. Specifically, in the ideal gas case, the density of gram moles is $\rho \equiv \frac{N}{V}$ and thus ρ is an intensive variable.

Remember that the basic role of thermodynamics is to investigate processes involving changes in the variables of the system to produce things like work and heat. We started our discussion of thermodynamics with a statement that pressure, an intensive variable, differences between two systems is what leads to changes volumes, an extensive variable. The thermodynamical variables all come in pairs one extensive and one intensive; in gases pressure and volume, in elastic systems tension and length ... and processes come about by the relationship because of changes in the value of a member of a pair. The usual example is work done by changing the volume of a gas that is contained in a volume that is variable by having a movable wall element, a piston between two systems. The relevant phrase is that the pressure brings

³ R is basically the energy content of one gram mole of a gas at a temperature T .

about the change in volume and in doing so work was done or energy in the form work was transferred to or from a member of the system. Shinola Shinola

19.2.4 0th Law: Temperature Exists

Depending on the macroscopic system under study, there are small number of properties of the system that can be measured, see Chapter 2 on page 33, and the values of these attributes define the state of the system. These are things like the volume, pressure, concentration of species, magnetic field strength and so forth. Among these is the temperature, T . The 0th Law of Thermodynamics is basically the statement that temperature exists; it is a measured quantity whose values can be put on an objective scale. Like all measurement processes, there is an instrument or set of instruments that are standards and a protocol for use. The temperature is measured through the process of contact with a system whose response has been calibrated. The home mercury thermometer is a simple and classic example of a calibrated system that serves as the basis for measurement of temperature; a quantity of mercury is contained in a volume under zero pressure and with elongated space for expansion and the volume measured and calibrated to predetermined situations that allow us to define a temperature scale.

The protocol for measuring the temperature follows from the 0th Law of Thermodynamics which states that **two or more bodies held in thermal contact for a long enough time will come to the same temperature**. There are several things that need clarification. What do you mean by contact and what constitutes a long enough time? In many cases, contact means literally what it says – touching. A more general and more appropriate definition is best stated as that somehow it is reasonable to talk about the two systems separated by a permeable membrane that allows an exchange between the systems. Of particular interest here is the exchange of energy as driven by a temperature difference. By the way, we will see that this sentence also provides a definition of heat, the energy that is exchanged between two systems in contact with the separating wall that allows temperature driven energy exchange. Note also that the systems under discussion have no other contact; they are isolated from everything else. The exchange can carried out in different ways but for now we restrict our discussion to contact by touching, better said enveloping, but not allowing any other transfers. In fact, this particular form of exchange is called conduction. Shortly, we will expand our contact regimens beyond touching but for now let's leave it at that. The other issue with the statement of the 0th Law

is how long is long enough. The -1^{st} Law discussion above clarifies this issue. If the objects have very different temperatures when brought into contact, they will begin to show changes in other measures of their state such as density, length, color, etc. These or, at least, one of these other measures will start to change. Long enough is when these changes stop. We say that the systems have achieved thermal equilibrium. In equilibrium, all state variables are unchanging including the temperature. This is exactly how the home thermometer is used to measure temperature in a person. In this case, when the expansion of the mercury stops, the thermometer and the person or more precisely the part of the person in contact with the thermometer bulb are at the same temperature; the temperature is indicated on the thermometer by the current volume of the mercury. We would then establish a temperature scale based on these volumes.

I should expand on some of the subtleties of the previous discussion. As stated above temperature is one among several measures that are required to specify the state of the system. Note that I have not been precise about the idea of “a system.” Often in introductory discussions of thermodynamics, the system is a finite volume of gas but this is much too restrictive to contain all the physical cases. Some examples and their state variables are: the box of gas case described by the pressure, temperature, volume, species content; a rubber band suspended between two fixed posts with temperature, tension, and density specifying the state; the mercury thermometer discussed above with its vacuum volume and liquid mercury bounded by a glass envelope. These constitute systems. A physical entity that is described by an appropriately complete set of state variables and with specified boundaries. In some cases, the system is made of subsystems. The mercury in class case is a simple example. At a higher level of analysis it can be considered three connected subsystems, the volume of mercury, the volume of the vacuum and the volume of the glass enclosure all with their state variables. This is because we have idealized the parts. The mercury is assumed to be incompressible and if so there is no need for a discussion of the pressure. We also assumed consistently that the vacuum does not have pressure. In fact the mercury is slightly compressible and the vacuum is actually filled with mercury vapor and the spatial separation of the mercury and the mercury vapor allows mercury atoms to move between the two. For now and since each case is unique and needs its own considerations let's leave it that there are system and systems are specified by contents at a certain state labeled by measurable variables and in some sense isolated by proscribed boundaries. The boundaries may or may not allow passage of something between the system. For example in our thermometer, the boundary between the

mercury vapor and the liquid mercury can or cannot allow the passage of mercury atoms and thus allow for changes in the amount of mercury in either subsystem. When there is passage of matter or energy between connected systems this is a process.

It is very important to realize that Thermodynamics was developed to describe processes utilizing temperature and heat flow to produce useful work. We also have to return to our earlier observations about measurement and ‘what is’ in Section 2.2 on page 34 in which we state that a thing is what is measured. The example used there of distance and length is important to our understanding of temperature. In the example of length, we articulated that the identification of a standard and a protocol defined the meaning of distance. Similarly here we define temperature through the establishment of standard and protocol. For our purposes it is adequate to consider our earlier discussion of the mercury example as our standard and protocol, the extension of the mercury in the thin tube in contact with the system whose temperature is being assessed. We must keep in mind that our goal is to describe thermal processes which do mechanical work or some other useful activity. This brings about a reinterpretation of our state labels. The comparison of temperature with another more familiar state variable, pressure for example, is worth pursuing. Pressure was an established concept and its role in energy transfer clearly indicated by Newton. If two isolated systems, labeled 1 and 2, are enclosed by a movable partition and the pressure in 1 is greater than the pressure in 2, The volumes of both systems are changed and, if we do this process slowly, and wait a long time to allow the pressures to equilibrate, the volumes will be changed and, although this will sound like a strange way to say it, both volumes will change. Said another way, pressure differences drive volume flow. Actually from simple mechanics we see that the pressure driven volume changes exchanges energy between the systems. The higher pressure does work on the lower pressure system; the work done during a volume change is $P\delta V$. Temperature is a state variable of a system like pressure. These variables come in pairs. Pressure is paired with volume and volume changing processes are work flow. Pressure is a state variable that is intrinsic. It is the same throughout the system and two identical systems have the same pressure but two identical systems have twice the volume. Volume is an extrinsic variable. Similarly two identical systems have the same temperature; temperature is intrinsic. We have not defined the partner extrinsic variable which is related to the heat flow. That is the problem of the 2nd Law. We will get to that after we deal with the 1st Law.

The example above using a thermometer to determine of body temper-

ature measurement is also an illustration of another important property of thermodynamic systems. There are systems which for some reason or another do not change important thermal variables in processes. These are called baths. In the example above of measuring body temperature, we assume that the body temperature does not change as the process of thermal equilibrium established within the mercury in glass thermometer. Heat flows from the body to the thermometer but the body's temperature does not change. For this process, the body is said to be a temperature bath. In fact, this experience is an indicator that physiologically we sense heat flow not temperature since as the thermal equilibrium is established the thermometer goes from 'feeling' cool to okay. The human body achieves its thermal bath status by two means. Firstly, the body is so large compared to the thermometer that the heat flow needed to thermalize the thermometer is insufficient to change the body temperature. Also even if we place the human body in contact with another large system, our metabolism will function to maintain our internal body temperature. In a room at some nominal temperature, the body does thermalize in the sense that our surface temperature becomes almost the same as the room temperature which is acting as the temperature bath. Since this is different from our internal body temperature maintained by our metabolism, the average person transfers about 70 Watts or 70 Joules per second of heat energy into the room. In this sense, we never achieve thermal equilibrium with the room but this amount of heat flow is sensed as normal. More than this is a cold room and less is hot. Another useful example of a thermal bath is a pot of boiling water at standard atmospheric pressure. If you increase the heat flow from the stove all that happens is that the water boils more vigorously but its temperature does not change. In fact, this thermal bath is used to fix one of the standard temperature designations. In this case, 212°F . The other end of the Fahrenheit temperature scale is defined from a temperature bath of ice, water, and salt defined to be at 0°F . With these two points, the temperature scale can be set.⁴

This brief discussion of the definition of temperature is characteristic of all thermodynamic discussions. Although we characterize systems with labels such as pressure, temperature and volume, the important issues deal with the processes of change. In this discussion of the 0th law, we defined

⁴The current temperature standard is not the simple mercury in glass simple thermometer using only two temperature baths. The National Institute of Science and Technology, NIST, follows the International Temperature Scale of 1990, ITS-90. ITS-90 sets the temperature scale using several temperature baths in the range from 83.4°K to 962°C to calibrate a platinum resistance thermometer.

temperature as a measurable quantity but used the idea of process occurring manifest by change to the each of the systems labels due to an initial temperature difference between systems in thermal contact. The basic issue is change brought about by differences in state variables and the ability of all systems to have the changes ultimately stop. This condition is called equilibrium. The existence of a state of equilibrium between systems is what allows us to define the temperature scale but temperature is not the only state variable that is measured by establishing equilibrium. Two cylinders of gas connected by a movable wall but at different pressures will change by moving the wall toward the smaller pressured cylinder. Change stops occurring when the pressures equilibrate. We say that the pressures are equal. Note that this discussion of pressure does not use the concept of force or area. We can define a pressure scale using pressure baths and equilibrium just as we did with temperature. As in the case of temperature, there are systems that can be considered pressure baths. A huge volume of gas such that the movement of the piston of a small system connected to it does not change the volume significantly enough to modify the pressure or any of the other thermal variables of the large system. The atmosphere is a pressure bath to all intents and purposes for man-made processes on the earth. The connection between this purely thermal discussion of pressure and its usual force and area description will be clarified by the other laws of thermodynamics.

There is also an important point to note and that is that the process under consideration can take place in two ways. We can bring into contact two systems with very different temperatures. As required these systems will ultimately come to equilibrium as indicated by the stability of their state variables. Another approach is to bring about the temperature change by means of a system of thermal baths that makes the process a continuous set of very small changes. This later case is called reversible. The reversible process has the advantage that the system is in a certain state as manifest by its state labels at all steps in the process. The first case is labeled an irreversible process. Both can be analyzed by thermodynamics but the the outcomes of the processes will be different if the process is irreversible or reversible. How there can be a difference in the final states of reversible and irreversible processes will be clarified in the 2nd law. In our original discussion of temperature equilibrium, we just discussed making the contact membrane between the systems permeable to the flow of heat. In some sense, all partitions are permeable to the flow of heat. What really matter is the time scale over which the heat flow operates. Any material wall with a finite temperature difference will pass some amount of heat on some time

scale, this is based on the thermal conductivity, $\tau \equiv \frac{\delta Q}{\Delta \text{time} \cdot \Delta T}$ of the material. The rule for reversible processes is that the rate of heat transfer must be small for any time scale for thermalization of each of the systems. Another way of saying that the time scale for heat flow through the wall is infinitely long.

This discussion of thermal equilibrium above is based on two systems in thermal contact. The idea is that the systems are both contained in some sense; enclosed so that there is no communication between them. In our discussion of the process of thermal equilibrium above the two systems initially with different temperatures are brought into contact by joining them at a surface of each one and that the surface allows heat to flow between them and it will flow until the two system temperature are the same. This simple set up is the epitome of the operation of processes in thermodynamics. Two or more systems are connected by a membrane which is porous to some particular one or more important thermodynamic labels. The contact wall allows flow of something and that flow is driven by the difference in the relevant state variable. In the case of the 0th Law, the contact wall allows heat to pass if there is a difference in temperature between the systems.

Before closing this section on temperature equilibrium processes, we should look at the situation of two like systems connected by a movable wall but well insulated to heat flow between them. We know that volume will flow to the low pressure system from the high pressure system and that equilibrium will be reached. Using our ideas of mechanical energy we would say that there was energy flow between the systems with the high pressure system losing mechanical energy to the low pressure system. In this way of looking at this, there is in complete analogy with the temperature difference system. Is this volume exchange the same as the energy concept in mechanical systems? This is the subject of the next section.

1st Law: Conservation of Energy

The separation membrane between two systems can effect each other in different forms than conductive heat flow driven by temperature differences. For example, consider two volumes of gas with differing pressures and differing temperatures. The barrier between them can be movable, a piston, and heat conducting. In language that follows our discussion above about heat flow, we can describe the process as pressure equalization by movement of the piston and the movement continues until the pressure difference vanishes. In the process of pressure equilibration the mechanical measures, volumes, of the component systems change. Again, as in the case of two

systems in contact and having a temperature difference other differences emerge as the equilibration takes place. In our simple glass thermometer as the equilibration comes about the volume of the mercury changes and we conclude that the equilibration process is taking place until all measures on the system stop changing.

In this case, the energy transfer is called work and the idea that it is done by one system on the other, work flows from the higher pressure system to the lower pressure system. In the case of temperature equilibration, the heat flows from the higher temperature body to the lower temperature body.

Another example a thermodynamic process would have the two systems separated by barrier that allow matter to pass but nothing else, no movement and not heat flow and so forth.

In our thermometer example the mercury expands as heat is added but because the other space in the thermometer cavity has no pressure, $P_{vac} = 0$, another system state variable, In this case, there is no work done since work done for a volume change is the $P\Delta V$. P is called an intensive state variable since its defining statement that in a process which changes the volume, the work done is the energy exchange driven by pressure differences. The First Law of Thermodynamics relates the various forms of energy change in a thermodynamic process. It is simply the identification that, in any process, the total energy is conserved. As applied to the example of the mercury in the thermometer, the heat flow into the mercury that changed the temperature did not $P\Delta V$ work but went into energy content of the mercury which manifested as the temperature increase. If there was some pressure in the ‘empty’ part of the thermometer, there would be work done by the expansion and more heat would have been needed to raise the temperature the required amount.

2nd Law: Increase of Entropy

The next two laws, two and three, are much less well known and thus a bit more subtle to discuss. The 2nd Law clarifies the nature of heat. It is important to realize that heat is defined only in a process.

3rd Law: Absolute Zero

19.2.5 Radiation in a Cavity

Now that we have a vocabulary, let’s examine an especially simple thermal system. Consider a massive block of stuff, say aluminum, that is in thermal equilibrium at some temperature T_1 . We can change the temperature by

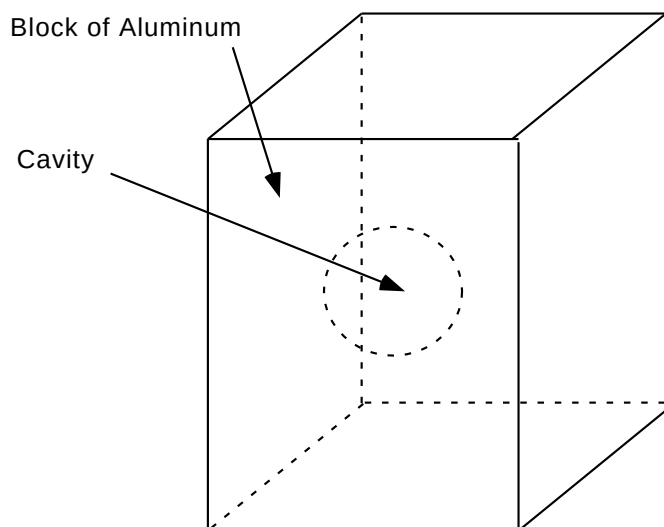


Figure 19.2: **Black Body Cavity** Inside a block of material an empty cavity absorbs heat. The amount of heat needed to raise the temperature of the cavity scales as the volume of the cavity.

placing the block of aluminum into contact with a succession of temperature baths to study the thermal properties of the aluminum. These temperature baths have very closely spaced temperatures so that we can consider the heating to be a slow or equivalently reversible process. We can measure the heat flow into the aluminum when we incrementally change the temperature of the aluminum. The heat that is required to raise the temperature a small amount scales with the mass of the aluminum block in use. We construct the quantity, $C \equiv \frac{\delta Q}{\Delta m \Delta T}$ called the specific heat.⁵ Also note that there are now two different symbols for the changes brought about by the thermal processes. Lower case δ is a change in something that is not a state variable. Heat is an energy but it is not a state variable, see the discussion in Section 19.2.4 on page 417. It is only defined during a change; there is no such thing as a Q . Make a hole in the center of the stuff. The hole is empty, a perfect vacuum or at least as near as we can get, see Figure 19.2. We have studied all the thermal properties of the stuff, the aluminum in this

⁵There is a technical issue here. Adding heat in the open air is different than adding the heat at constant volume; the block of aluminum expands against atmospheric pressure, whereas at fixed volume there is no mechanical work done on the atmosphere. Therefore, we always specify the process in which the heating is done. In this case, we are using C_V , the constant volume specific heat.

case, and completely understand it. In order to raise the temperature of the aluminum, we have to add heat or energy to it including the hole. If we take into account the heat to raise the temperature of the stuff, we find that it takes energy to raise the temperature of the nothing that is in the hole. In the case of the empty cavity, the heat to increase the temperature scales as the volume of the hole.

This has to be a surprising result. Even though the hole is empty, a vacuum, it takes heat input to raise its temperature. If you make a bigger hole, you need more heat input for the same temperature change. The amount of heat required for a given temperature change scales as the volume of the cavity; for a spherical cavity of radius R as R^3 . If the energy being put in the cavity were a surface phenomena, the heat input to raise the temperature would scale as R^2 .

Put a hole in the side and look at what comes out. If the temperature is high enough, light that you can see comes out. As you raise the temperature, the light gets whiter. We now know what is going on. We are filling the cavity with light. The heat energy added to raise the temperature goes into making the light. There are several simple observations that confirm this. The nature of the light as indicated by the spectrum is independent of the material bounding the cavity. On the other hand, the spectrum of the light is universal because there is nothing in the cavity, the vacuum. The spectrum depends only on the temperature. We can verify that the energy delivered to the cavity scales as the volume and not the surface area of the cavity. Again, this makes sense since there is nothing in the hole but the light that is produced. Hypothesizing a particulate nature for that light, what we now call the photon, Planck developed the following formula for the the light intensity at any wavelength in a wavelength interval $\Delta\lambda$:

$$\rho(\lambda, T)\Delta\lambda = 8\pi c \frac{h}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1} \Delta\lambda \quad (19.1)$$

where λ is the wavelength, c is the speed of light, k is Boltzman's constant, T is the temperature and h is a constant that he used to fit the spectrum. That constant, h , is now known as Planck's constant and considered to be one of the fundamental dimensional constants of nature⁶

⁶The form of the function $\rho(\lambda, T)\Delta\lambda$, 19.1 on page 419, is a shorthand representation for $\int_{\Delta\lambda, \Omega} I(\lambda, T) d\lambda d^2\Omega$ where $I(\lambda, T)$ is the light intensity at a wavelength λ . Ω is the directions of the beam of light being analyzed. If the light is going in all directions $\Omega = 8\pi$. This is an angle and is dimensionless. The integration range is $\Delta\lambda$ is taken to be infinitesimally small so the λ is the wavelength at that value.

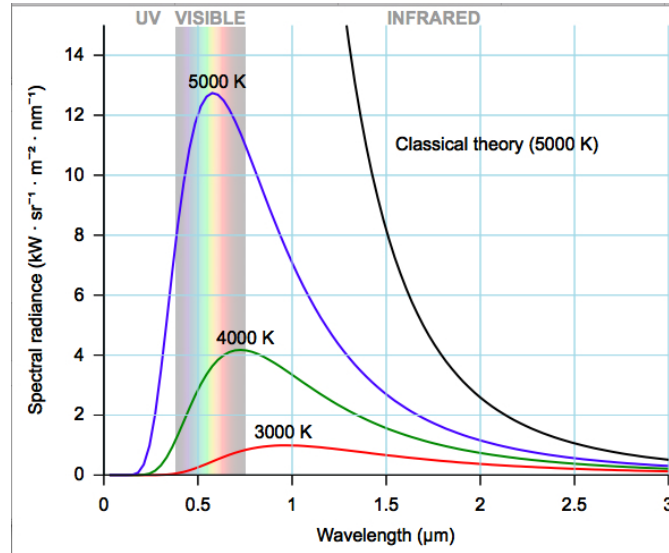


Figure 19.3: **Intensity of Light from a Cavity as a Function of Wavelength.** Plots of the intensity of the light as a function of the wavelength for three temperatures, 3000 °K, 4000 °K and 5000 °K. These curves are plots of Planck's Law, Equation 19.1 on Page 419. The higher the temperature the lower the peak and the greater the area under the curve. Also shown is the location of the visible window on the spectrum and the classical Rayleigh Jean's predicted spectrum for $T=5000$ °K

The primary features of the observed energy density spectrum that Planck found, see Figure 19.3 on page 420, had sharp peaks at a characteristic wavelength that increase and shift to lower wavelength, λ , as the temperature increases. In addition, the area under the intensity curve grows larger as the temperature increased. Equation 19.1 on page 419 is the function that Planck found that fit the data and it is in excellent agreement with the observed data. He managed to get this formula through a process that is common in physics research labeled phenomenological. He picked a rather simple formula that had the pattern of development that looked like the then known observed data. His reasoning was functional simplicity. As more data was becoming available, he was impressed with how closely the new data matched his simple function. He immediately realized that he had to develop a physical picture of the thermally developed light to 'explain' the data.

It is worthwhile to point out at this point that the dimensional content

of Planck's fitting parameter, h , is action $h \stackrel{\text{dim}}{=} \frac{ML^2}{T}$.⁷

19.2.6 Attempts to describe the spectrum

The first effort to predict the spectrum of the light in the black body cavity from first principles was by Lord Rayleigh and Sir James Jeans. They followed what would be natural to us. We treat light as a field phenomena. The earlier descriptions of light as particulate based phenomena developed by Newton were discredited even in his own time.

The only trick in developing the prediction of Rayleigh-Jeans is to get the form of the intensity of the thermal light from dimensional analysis. Intensity is energy per unit area per unit time in a wavelength interval or $I \stackrel{\text{dim}}{=} \frac{E}{T \cdot L^2} L \stackrel{\text{dim}}{=} \frac{ML^2}{T^2 \cdot L^2 \cdot T}$. Using only temperature and wavelength which is all you have for a cavity, dimensionally the spectral intensity must take the form

$$\rho_{RJ}(\lambda, T) \Delta\lambda = 8\pi \frac{kT}{\lambda^4} \Delta\lambda, \quad (19.2)$$

where k is the Boltzman's constant and the 8π is a geometric factor.

From Figure 19.3 on page 420, the Rayleigh-Jeans law looks awful. It does well only at very long wavelength, the far infrared. More difficult was the fact that energy at small wavelengths diverged. This was referred to as the ultraviolet catastrophe and was a clear indication that this pure field approach to explain the spectrum was inadequate.

Another early effort to derive the spectrum was by Wilhelm Wien. Studying the thermodynamics of a field filled cavity, Wien found that, if instead of plotting the spectrum against the wavelength for fixed temperature you scaled the wavelength with the temperature $\lambda_T \equiv \lambda T$, the curves other than a scaling with temperature were universal. This was called the Wien displacement law. He then modeled the spectrum following the pattern of the well known earlier success of Maxwell who had developed a velocity spectrum for gas atoms confined to a fixed temperature volume. It is worthwhile to review Maxwell's descriptions of the thermodynamics of atomic system to understand both the limitations and the success of the treatment that Planck ultimately applied to the thermal cavity. Wien attempted to describe the

⁷The dimensional content of h is determined either by the overall scale of the spectral intensity, $\rho(\lambda, T) \Delta\lambda = 8\pi \frac{ch}{\lambda^5} \Delta\lambda$ or the requirement that the power of the exponential, $\frac{ch}{\lambda kT}$, be dimensionless. Also note that in these sections on thermodynamics the symbol T is used for temperature but in all the discussions of dimension the symbol T is reserved for the dimension of time. The reader will have to distinguish the usages by the context.

light in the cavity as Maxwellian like but satisfying his displacement rule⁸. Wien's formula has two fitting parameters, c_1 and c_2 but still could not match the data in the infrared.

$$\rho_{Wien}(\lambda, T) \Delta\lambda = \frac{c_1}{\lambda^5} \frac{1}{e^{\frac{c_2}{\lambda T}}} \Delta\lambda \quad (19.3)$$

This was not a rigorous derivation but it did fit the small wavelength part, the peaks and the motion of the peaks with changes in temperature.

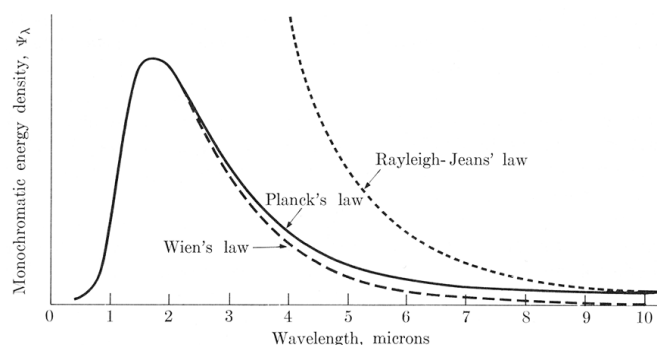


Figure 19.4: **Classic Attempts to Predict Spectrum** Plots of the intensity of the light as a function of the wavelength for a fixed temperature for the case of the Wien Law and Rayleigh-Jeans Law. It is no surprise that Planck's formula took the form that it did. In a sense all that it had to do was interpolate between the two classical laws.

19.2.7 Planck's Explanation of the Spectrum

The issue for Planck was not to get the relationship, equation 19.1 on page 419, but to justify it from some fundamental understanding of the nature of light. To derive his formula, Planck had to assume that thermal energy was distributed in a system of discrete entities whose energy was proportional to the frequency. His reasoning followed from Wien's displacement condition and the Maxwell form. The $\lambda_T \stackrel{\text{dim}}{\approx} E \cdot L \stackrel{\text{dim}}{\approx} \frac{E}{cT}$ was a term

⁸Maxwell's rule for speed distribution for a system of atoms of mass m and energy $m\frac{v^2}{2}$ where v is the speed is

$$\rho_{Maxwell}(v, T) \Delta v = 4\pi v^2 \left(\frac{m}{2\pi kT} \right)^{\frac{3}{2}} e^{-\frac{mv^2}{2kT}} \Delta v$$

that naturally could fit into an exponential, but that the statistics, the way that multiple copies of the same thing are counted, were not the normal ones. He had to count the physically independent identical states in a very unusual and inexplicable way. The proportionality constant between the energy and the frequency that he had to use was the fitting constant h that he used in Equation 19.1 on page 419 is now called Planck's constant.

$$\epsilon = h\nu = \hbar\omega, \quad (19.4)$$

where ν is the cycle frequency, ω is the radian frequency and $\hbar = \frac{h}{2\pi}$.

The second aspect of Planck's attempt to completely describe the phenomena of thermal radiation by articulating how a collection of fundamental structural entities dealt with the issue of identity of these elements. In order to get his function, he had to count the possible configurations in an unusual way. To get some understanding of the counting issue, consider a collection of n coins with no other distinguishing features except a two labeled degree of freedom such as H and T for heads and tails. For example consider 5 coins. The configuration HHHHT and the configuration HTHHH would be considered distinct. For his counting, he had to consider these the same configuration. For a collection of n of his entities, he had to require that be only n unique identifiable configurations. The basis for this strange result was clarified by S. N. Bose and Albert Einstein some time later. The resulting counting scheme is called Bose-Einstein statistics and is one of the cornerstones of our understanding of modern physics [Wali 2009].

19.3 Photo-Electric Effect

We are all well aware of the use of light to control electronic devices. Photoelectric devices are used to open doors, send optical signals, and operate computers. Today, we usually see these devices as small solid state elements. This was not always the case. The early photoelectric devices were large vacuum tubes. Light hitting the clean metal surface inside the tube would cause an electric current to flow. This phenomena was just being discovered shortly after the turn of the 19th century, see Figure 19.5.

Einstein clarified the conflict with the observational data available at the time by saying that Planck was right and that light is composed of countable entities that have energy in proportion to the frequency of the light, the color. He suggested that the process of the photo cell was that the electrons in the metal of the anode absorb the light to have the energy

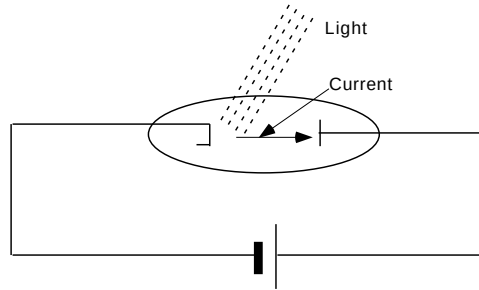


Figure 19.5: **Photo Electric Effect** Light shining on a metal plate in a vacuum enclosure releases electron into the evacuated space which form a current.

they need to move out of the metal into the space between the cathode and anode and move across the gap to complete the current.

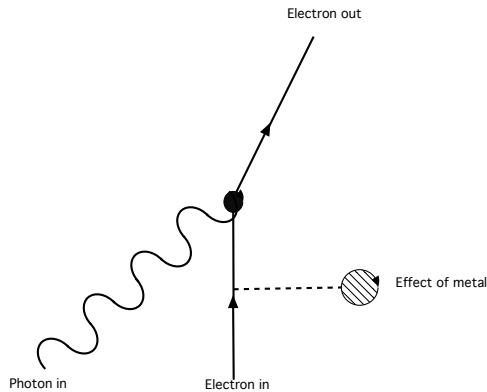


Figure 19.6: **Einstein Photo Electric Effect** A single photon of light interacts with a single electron. The energy of the photon removes the electron from the metal and gives it kinetic energy, see Equation 19.5.

The energy of the photon, $h\nu$, is transferred to an electron which gives it both kinetic energy and the energy to overcome the weak binding of the metal, see Figure 19.6 on page 424.

$$h\nu = m\frac{v^2}{2} + \phi \quad (19.5)$$

Where ϕ is the energy required to move the electron out of the metal.

This picture predicts that the number of electrons, the current through the tube, is equal to the number of photons. Thus, for a given frequency,

color, the current is proportional to the intensity of the light. He also noted that you could measure the velocity of the electron by back-biasing the tube and seeing what voltage just stops the current⁹. This phenomena is like the ability to measure the velocity of a vertically thrown ball by measuring how high it rises against the earth's gravitational force; $\frac{mv^2}{2} = mgh$.

$$\frac{mv^2}{2} = eV_{stop} \quad (19.6)$$

In other words, by measuring V_{stop} and the frequency you can measure h as the slope of the straight line.

$$eV_{stop} = h\nu - \phi. \quad (19.7)$$

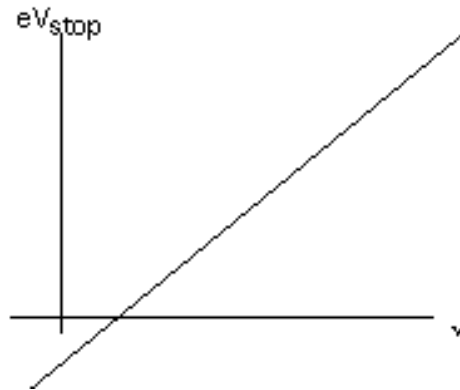


Figure 19.7: Plot of Stopping Potential Versus Frequency The stopping potential versus frequency of the light curve as predicted by Einstein, Equation 19.7 on Page 425. Thus the relationship between the frequency and stopping potential was linear with slope h . The experiment was successfully executed by Millikan and confirmed the prediction of Einstein.

Thus Einstein, by extending Planck's analysis of the black body experiments, makes several predictions. When you plot the stopping potential against the frequency of the light, you get a straight line and that the slope of that line is the value of $\frac{h}{e}$ where e is the charge of the electron and that the h is the same as Planck's value, see Figure 19.7. The value of the

⁹Einstein was awarded the Nobel Prize in 1922 for this work. Interestingly to modern physicists, he was never awarded the Prize for his development of the General Theory of Relativity.

charge of the electron was already measured in an earlier experiment by J. J. Thompson.

Another prediction from Einstein's construction is that the current in the phototube illuminated by monochromatic light is proportional to the intensity of the light but that no matter the intensity of the monochromatic light there was a cutoff for a phototube independent of the intensity and depends only on the frequency.

All of these predictions were satisfied by a series of very difficult experiments carried out by the great american physicist, Robert Millikan. These energy packets of light were subsequently called photons and are now universally recognized as the fundamental constituents of light.

Subsequent experiments importantly also showed that you could run photo-detectors at such a low intensity of light that the current that you could count time separated photon events.

It is the combination of the well understood properties of light as manifested by the Young's double slit experiment and the result of the Einstein analysis of the photo-electric effect that is the foundation of quantum mechanics. In addition, when these are combined with the understanding of the black body experiment that was articulated by Planck, the emergence of particles in the vacuum and the counting scheme for identical particles, we have the basis for a modern theory of matter that is manifest in the quantum theory of fields, see Chapter 22 on page 483.

19.4 Young's Double Slit Experiment Revisited

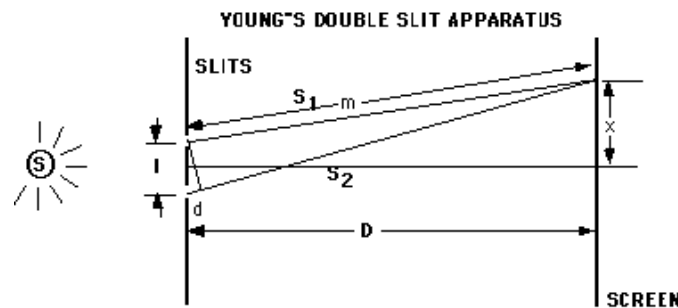


Figure 19.8: **Double Slit Revisited** Light from a single source illuminates two slits. When both slits are open, there is an interference pattern at the screen.

In Section 4.4 on page 88 while discussing light, we introduced the double slit experiment of Thomas Young. Monochromatic light passing through two narrow slits and illuminating a distant screen produced a pattern of illumination that at some places produced a brightness that was four times what would be present if only one slit was exposed. There were intervening places that had no illumination. The spatial average of the brightness of the illumination was twice that of one slit being open. The only working description that was possible was that the light's causal agent was not the brightness but an underlying descriptor, the amplitude, whose square was the brightness. Not only that but using Newton's observation that light had an underlying structure that manifested itself as color and that the label for the color was associated with a label that was periodic. Fresnel, Section 4.6 on page 106, extended this analysis to diffraction phenomena which lead to our discovery that light is a phenomena that travels over all paths in all of space when going between two places. The only construction that would describe the pattern of bright spots and dark places on the illuminated screen exposed to monochromatic light was that light was an amplitude definable at every point in space, a field which was not directly measurable but whose square was the brightness at that place. Figure 19.9 on page 427 shows the

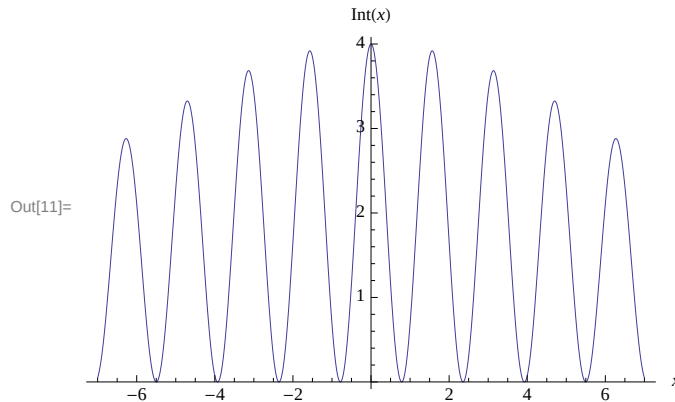


Figure 19.9: **Two Slits** When monochromatic light from two narrow slits illuminates a screen, a pattern of bright and dark lines are produced, see Figure 19.8. The figure above shows the brightness of the light, $\text{Int}(x)$, for the positions x measured up the screen of the light as you move up the screen. There is small decrease in the brightness of the peaks as you move away from the central maximum due to diffraction.

intensity pattern for positions varying as you move up the screen in Fig-

ure 19.8 on page 426. These developments in our understanding of light was further extended by Maxwell, see Section 5.4 on page 140, when he unified the electric and magnetic force system to include the observation that disturbances of the electric or magnetic field would propagate as a wavelike disturbance in the fields traveling at the speed of light; light had become the electric and magnetic field effects operating at very high frequencies.

Now let us replace the remote screen on which we displayed the illumination pattern with an array of of very small Einsteinian photo detectors¹⁰. What happens now? Instead of intensity, you count the photons at each location. There are clicks all over the surface of the array. You can run the intensity down so low that, for that color of light, you get only one count a week. Actually, this is an exaggeration. You can run the experiment at such a low rate that you are confident that there cannot be two photons in the system at the same time¹¹. When you do this, you do not see two half hits at the same time. You also do not see hits in places where the Fresnel/Young analysis indicated that there was no brightness. Ultimately, as the counts accumulate, the number of counts at a place x on the screen, $n(x) \equiv \frac{\text{counts in patch of area } A}{A}$ per unit time follow the Young/Fresnel pattern seen in Figure 19.9 on page 427.

$$\langle n(x, t) \rangle_T \hbar\omega = I(x) \equiv \langle A_{\text{Tot}}^2(x, t) \rangle_T = \langle \{A_1(x, t) + A_2(x, t)\}^2 \rangle_T, \quad (19.8)$$

where $I(x)$ is the intensity or more correctly the energy per unit area per unit time of the light at a place on the screen and $\langle \rangle_T$ indicates that a time average over times T large compared to the period of the light and also large with respect to the inverse photon arrival rate.

Earlier in our discussion of the double slit, we used a Fresnel construction that the light was a traveling wave phenomena to describe the brightness pattern. Now we need to incorporate the knowledge from the photoelectric effect that the detection of a photon by interaction is a local probabilistic effect. In fact, any attempt to correlate the time and place on the screen

¹⁰Actually we do this all the time. The CDD plane that is at the heart of current digital cameras is an array of Einsteinian photo detectors. Since the back plane of the camera is a couple of inches square, an array capable of producing Mega pixel images would use photo detectors that have an exposure face of about 10^{-3} inches in width and height.

¹¹Just to be specific, you can have the distance from the source of the monochromatic light to the screen be tens of meters, say D . If this is the case and you associate the constituent that makes the clicks to be traveling with the speed of light, it is reasonable to say the photon that made the click is in the apparatus for $\frac{D}{c} \approx 10^{-8}$ seconds and, if the interval between clicks is about a second, there is no time when any two photons are in the apparatus at the same time.

of the event in relation to the previous event fails. There is no correlation on an event by event basis. **At the single event level the relationships between events are purely probabilistic, stochastic in both space and time**¹². Only the long time accumulation of many events show a pattern; in this case, the classical double slit pattern. The photons like classical field phenomena travel over all paths although individual photons are always sensed locally. The pattern developed by $n(x, t)$ requires the photon locations detected at the array of Einsteinian photo detectors must be determined by a field that satisfies a Fresnel construction. In a very real sense, it is impossible to talk about individual photons. Although they may be widely separated in space and time, they are effectively entangled.

This was a difficult problem for Einstein to come to terms with. In fact, he never really did. He believed that our descriptors of elementary objects, causal agents in my vocabulary, had to possess 'local reality'. On the other hand, this phenomena is real and has become very important in applications to what is today called quantum information systems, see Chapter 21 on page 471.

It is also worthwhile reminding ourselves that the pattern on the screen or array of photo detectors is determined by the wavelength of the light. The Planck condition on the energy of the photons is determined by the frequency. It is also important to point out that, when the Maxwell field system or any pure wavelike field system is developed from an action perspective and has both time translation symmetry implying a conserved quantity called energy and space translation symmetry implying through Noether's Theorem a field energy and momentum. This construction for the Maxwell field requires that the energy and momentum are related by $\frac{Pc}{E} = 1$ where P is the field momentum density, E is the field energy density and c is the velocity of the disturbances or in this case the speed of light¹³ Using this condition for the photon requires that

$$p_{\text{photon}} = \frac{E_{\text{photon}}}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}. \quad (19.9)$$

It should be noted that since c is huge compared to our usual velocities and that the photon momentum is very small compared to its energy. Thus

¹²Stochastic is the simple idea that the system has no memory. In a sequence each event is independent of its predecessors. After tossing five heads in a row, the probability that the sixth toss is a head is still one half; fair coin tosses are stochastic.

¹³For the free particle of mass m with the usual non-relativistic free particle action, $L = \frac{mv^2}{2}$ the momentum is $P = mv$ and the energy is $E = \frac{mv^2}{2}$ this ratio $\frac{Pv}{E} = 2$ clearly a constant of the motion.

an energetic disturbance on a stretched rope or water surface produces a noticeable momentum whereas an intense beam of light heats the surface but produces little pressure on it. A beam of n photons of light transfers an energy $n h \nu$ and a momentum transfer of $n \frac{h}{\lambda}$ or the ratio of the momentum transferred to the energy transferred is $\frac{1}{c}$ which in usual MKS units is $\approx 3 \times 10^{-9} \frac{\text{sec}}{\text{m}}$. This is why, although you sense the warming of sunlight, you do not sense the pressure of the sunlight. The sunlight does not blow on your face.

19.5 Electrons and Young

We now make one more change to our Young's double slit apparatus. We replace the light source s in Figure 19.8 on 426 with a cathode to provide a current of electrons¹⁴ and we replace the screen with a plane of electron counters. The pattern on the screen follows the same general rule as it does for the light; the electrons come to the detectors in the same way as the photons of the light. The electrons are detected discretely. There are never half counts. As the counts are accumulated a pattern emerges with definite minimum and maximums. In the case of light, we used a field property of light to describe the pattern. The immediate implication is that electrons as elementary countable entities have field properties also. Some how we have to identify a measure of the electron beam that describes the pattern.

The wavelength of the light which determines its pattern is given by $e_{\text{photon}} = h \nu_{\text{photon}} = \frac{hc}{\lambda_{\text{photon}}} = p_{\text{photon}} c$. Similarly, setting the wavelength on the momentum of a non-relativistic free electron is $p_e = m_e v = \frac{h}{\lambda_e}$. Thus the energy¹⁵ of the electron will be determined by its wavelength through

$$E_e = \frac{p_e^2}{2m_e} = \frac{h^2}{2m_e \lambda_e^2}. \quad (19.10)$$

Even when the beam of electrons is very sparse, for example one electron per second, the pattern predicted by the electron wave length above develops over a long time at a rate of a click per second.

¹⁴Actually this is not what is actually done. It is impossible to realize the usual Young set up for electrons. The slit separations would have to be on the scale of atom sizes. Instead, we use X-ray diffraction from a crystal to measure the distance between atoms in the crystal. We then send a beam of electrons onto the same crystal and from the resulting diffraction pattern determine the wavelength of the electrons in the beam.

¹⁵The appropriate form for the energy in this case is the non-relativistic kinetic energy, $\frac{m_e v^2}{2} = \frac{p_e^2}{2m_e} \approx \frac{m_e^2 c^2}{\sqrt{1 - \frac{v^2}{c^2}}} - m_e c^2$, for $\frac{v^2}{c^2} \ll 1$.

The conclusion is that all things elementary, in this case electrons and photons, have both constituent and field properties. Although we have only dealt with electrons and photons, since they are primary constituents of matter, modern physics assumes that all matter and energy is quantum in nature and that the mechanics of the constituents is determined by descriptors that were capable of manifesting this. It was also important to develop these from an approach based on the principle of least action and follows the usual rules mechanics in which Noether's Theorem implies conservation laws based on symmetries. It will turn out that this formulation will lead to a single descriptor set known today as a quantum field degree of freedom that is the basis for the theory of Quantum ElectroDynamics, QED, see Chapter 22 on page 483.

19.6 Action and Quantum Mechanics

To understand the two results, the photo-electric effect and the double slit experiment, we need to have light be different than our earlier constructions. To Maxwell, light was purely a field phenomena. There was nothing discrete to be counted and the non-zero field values traveled following the construction of Fresnel. With Planck and Einstein the picture has changed drastically – all interactions are discrete transitions that take place instantly and locally and are stochastic in both space and time. nature and that light travels over all paths and generates an interference pattern from a superposed amplitude whose square is the probability that you will find energy and momentum deposition at that place at that time, see Equation 19.8 on page 428. The implication being that light is composed of discrete entities that are at the place of the transition at that time. if you make a position measurement at that place you will find the constituent there. These points lead to all the quandaries that are associated with quantum mechanics. It is the combination of superposition and localization of interaction that is at the heart of such conundrums such as Schrödinger's cat.

In the case of the double slit, we have a superposition of the two sources, light from slit 1 and light from slit 2, and the local transition that says the the light hits a point on the screen.

We will also follow a parallel development that we saw in the case of Fermat's Least time and the Fresnel/Huygens's construction. We will find that the simple principle of Least Action as the process of selection of the natural trajectory from all possible trajectories is replaced by a process of adding an amplitude to each trajectory and calculating the phase advance

over that trajectory and then at the final event adding the phases from all the trajectories to get the amplitude at the final event.

In order to describe quantum phenomena, we will require that the light travels over all paths and as it travels over all paths it generates a field that is composed of the sum of the phases from each path. In the language of the double slit, it comes from both the slits. The square of the resulting amplitude is the probability that there will be a photon at that point. In contrast to the Fresnel/Huygens construction, the phase that is advanced over the trajectory is the action in units of $\hbar$.

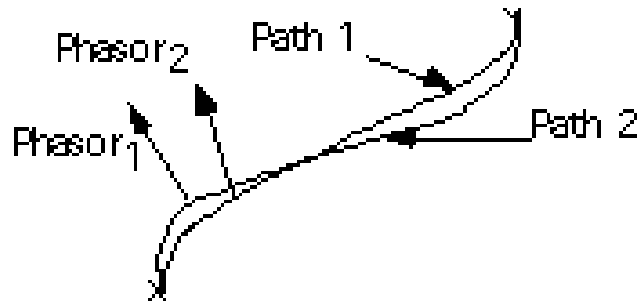


Figure 19.10: **All Paths Formulation of Quantum Mechanics** Light and, in fact, all things travel over all possible paths. Along the path, a phasor is advanced by computing the action over the path in units of $\hbar$, Equation 19.11. The resultant amplitude is the superposition of all the path phasors at that point.

$$\Delta\phi = \frac{\Delta S_{path}}{\hbar} \quad (19.11)$$

The situation here is the same as that which was obtained from the analysis of the Fresnel/Young construction and the Fermat Least Time goes; the paths around the least action path reinforce $\Rightarrow$ that the particles are found around the classical path. There is a wavelength in the field that the path forms just like in Huygens/Fresnel case. Once there is a natural trajectory, a connected region that has the phasors reinforcing, you have all the usual ideas of a particle mechanics. For a path made of phasors, you have a wavelength. But also from Neother's theorem, $\frac{\delta S_{nat}}{\delta x_{nat}}$, is the momentum. Momentum is not just a particle concept. It is related to any thing with a dynamic.

$$\lambda = \Delta x_{nat} \quad (19.12)$$

for which $\Delta\phi = 2\pi$ or

$$\hbar\Delta\phi_{path} = \frac{\delta S_{nat}}{\delta x_{nat}}\delta x_{nat} = p\lambda \quad (19.13)$$

or

$$p = \frac{h}{\lambda} = \hbar\frac{2\pi}{\lambda} = \hbar k \quad (19.14)$$

where k is called the wave number; the number of complete waves per unit length. For light this is consistent with our identification of $\epsilon = \hbar\omega$ since for light even in the classical wave theory, a light beam with energy density E has a momentum density $p = \frac{E}{c}$.

19.7 Constructing the Amplitude.

Our problem is to find a closed form for the amplitude. Our technique will be to follow the procedures of the Fresnel construction using phasor clocks that advance as the action accumulates on the trajectory. The way that this is described is to say the field or particle propagates from (x_0, t_0) to (x_f, t_f) . As it propagates the phase and the magnitude of the amplitude changes. There are a few essential differences between this and our earlier algorithms. In this case, we are dealing with trajectories, connected events in space-time, not paths in space. An additional complication is that the phasor clock advances as the action not just the time advances. A third difference is that, since actions are time sliced, we will not rectify the segments of the trajectory, see Figure 19.11.

Divide the time interval into small segments each of size ϵ . There are $\frac{(t_f - t_0)}{\epsilon}$ of these time slices. At each time on a time slice, you let x take on all values. Between the ends of intervals let the path be a straight line. In this sense, we are saying that the set $\{x_i\}$, of positions at each time slice is the designation of the trajectory. Between each time slice, given the $\{x_i\}$, or each little leg of the trajectory, we can calculate the the position and velocity and thus the action and thus know how much to advance the phasor. We then do this for all sets $\{x_i\}$ to obtain all possible trajectories. You could say that there is propagation between each time slice and that the final propagation is the effect of the total of all these.

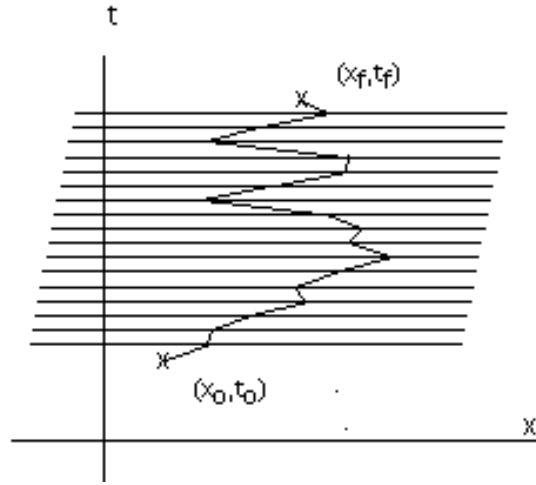


Figure 19.11: **Time Slicing** As light or any particle travels from event (x_0, t_0) to event (x_f, t_f) , it travels over a path that is time sliced in small segments of size ϵ . The path is given by the set of values of x at each time slice.

$$\begin{aligned}
 Prop(x_0, t_0; x_f, t_f) = & \\
 & Prop(x_0, t_0; x_1, t_1) Prop(x_1, t_1; x_2, t_2) Prop(x_2, t_2; x_3, t_3) \dots \\
 & \dots Prop(x_{n-1}, t_{n-1}; x_f, t_f)
 \end{aligned} \tag{19.15}$$

In each time slice the phase advances by the change in the action in that time slice in units of $\hbar$. Note that by time slicing in this way guarantees that we only deal with trajectories that are always advancing in time. There are not trajectories that have segments that run backwards in time.

Using Equation 19.15 on page 434, we have a different picture of the Fresnel construction. At each time slice multiply the propagators. Before the problem was to add the hands of the little clocks for each possible path. Now we seem to be multiplying. From Equation 19.15, we see that we need something that multiplies and yet adds as you advance through the time slices. This can be reconciled if we understand the exponential function a little better. Remember that $x^a \times x^b = x^{a+b}$. This gives me an excuse to make an excursion into some useful information about the exponential function that everyone should know. It is also a great source of Fermi problems.

19.7.1 A Mathematical Aside – The Population Equation The Exponential Function

Any system that develops or decreases at a rate proportional to the size of the system at any time is an example of a population system. Compound interest is an example. If you have a bank account that pays back 5% interest per year and you just leave the money there but do not add other money, the change in value of the account, in any year, is $\Delta P|_{year} = 0.05P_{year_0}$, and the value at the end of the year is P_{year_0} at the beginning of the year plus $\Delta P|_{year}$ so that the principle for the next year is $P_{year_0} + \Delta P|_{year} = P_{year_0}(1+0.05)$. After n years, the value of the account is $P_n = P_0(1+0.05)^n$. Now instead of compounding it once in a given year, you compound it α times per year. Then the value after n years is $P_n = P_0(1 + \frac{0.05}{\alpha})^{\alpha n} = P_0((1 + \frac{0.05}{\alpha})^\alpha)^n$. The function, $(1 + \frac{0.05}{\alpha})^\alpha$ can be plotted.

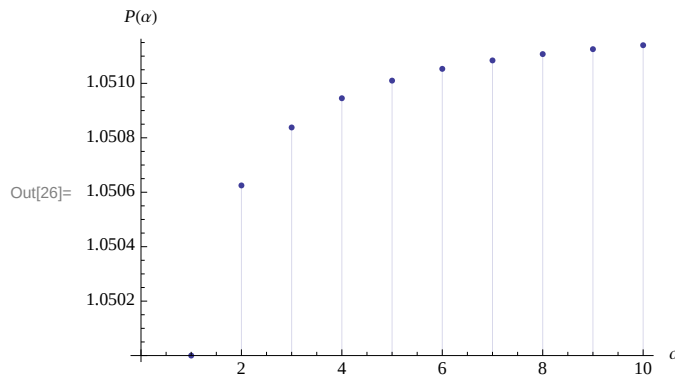


Figure 19.12: **Compound Interest** When interest of 5% per year is compounded frequently, α times per year, the value of an initial investment P_0 after n years is $P_0((1 + \frac{0.05}{\alpha})^\alpha)^n$. As α increases, the function $P(\alpha) \equiv (1 + \frac{0.05}{\alpha})^\alpha$ quickly rises to the value $e^{0.05} = 1.05127$ and thus for α large the principle is $P_0e^{0.05n}$ after n years.

There appears to be a finite limit at large α . In fact, this is the definition of the exponential function

$$\exp(x) \equiv \lim_{\alpha \rightarrow \infty} (1 + \frac{x}{\alpha})^\alpha. \quad (19.16)$$

You can simply show that $\exp(\beta x) = (\exp(x))^\beta$ and $\exp(x+y) = \exp(x)\exp(y)$, so that it is convenient to write $\exp(x)$ as e^x . The value of $\exp(x)$ at $x = 1$, e , is the number that is the base of the natural logarithms. Putting all this

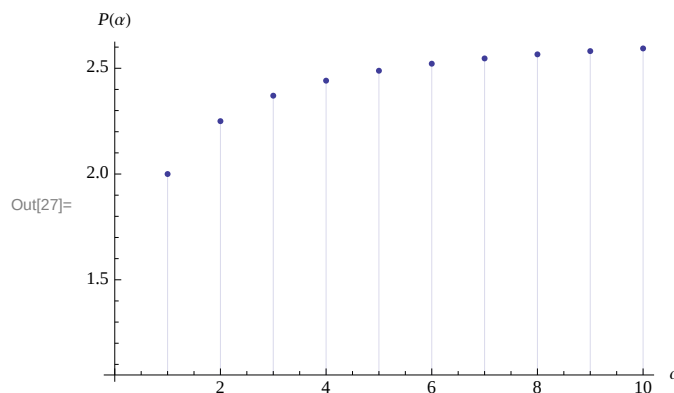


Figure 19.13: **Definition of e** A more dramatic example occurs for the case $P(\alpha) \equiv (1 + \frac{1}{\alpha})^\alpha$ which is the definition of e.

together, we can say that if we have a population that grows or decays at a rate that is proportional to its current population,

$$\frac{\Delta P}{\Delta t} = kP, \quad (19.17)$$

then the population at some time t is

$$P(t) = P_0 e^{kt}. \quad (19.18)$$

In the language of derivatives, the population equation is

$$\frac{dP}{dt} = kP \quad (19.19)$$

and Equation 19.18 is the solution to that differential equation¹⁶.

For any $k > 0$, the exponential function grows very rapidly, faster than any power of t . This implies that there is no such thing as a small rate of growth for a population that has any positive rate of growth no matter how small. There are an incredible number of applications of the population equation. We will look at some interesting examples.

¹⁶It is worthwhile making this result and Equation 19.18 more mathematical, i. e. dimensionless. P is a number and therefore already dimensionless. t is a time and thus $k \stackrel{\text{dim}}{=} \frac{1}{T}$. Rescaling t to $\beta \equiv kt$, Equation 19.19 is reduced to $\frac{dP}{d\beta} = P$ and the solution to $P(t) = P_0 e^\beta$. It is in this sense that physicists say that the times scale of this system is $\frac{1}{k}$.

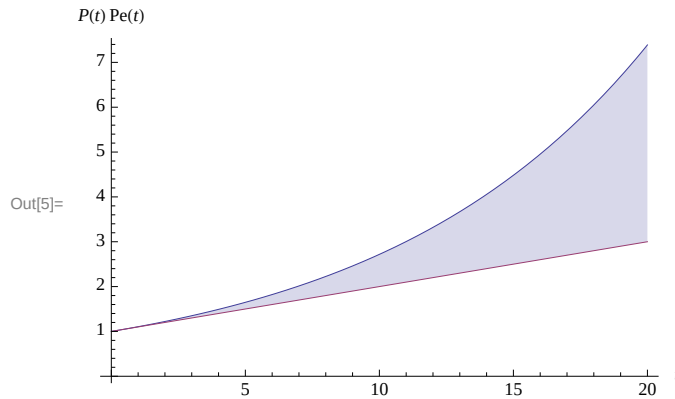


Figure 19.14: **Comparison of Compound Interest and Uniform Growth** A comparison of the rate of growth of a compounded interest and uniform growth. The lower curve is $P(t) = (1 + 0.1 \times t)P_0$, the uniform or linear growth curve, at 10% per year and the upper curve is the $P_e(t) = P_0e^{0.1t}$ a growth of 10% compounded instantaneously. You can see here the genesis of the statement that when a growth is very large it is called “exponential.” In the figure $P_0 = 1$ for both cases.

In figure 19.14, there is a comparison of a linear growth at 10% per year, $P(t) = (1 + \frac{1}{10}t)P_0$ with the exponential growth resulting from instantaneous compounding with the same growth rate of 10% per year, $P(t) = P_0e^{0.1t}$. The effects of compounding are significant and dramatic.

The best mnemonic for the use of the exponential growth is with the idea of doubling time. There is really nothing special about the value of the natural logarithms and any base can be used. A very convenient base is 2. At a given rate of growth, how long is it before you double the population.

$$P(t_2) \equiv 2P_0 = P_0e^{kt_2} \Rightarrow t_2 = \frac{\ln(2)}{k} = \frac{0.69}{k} \quad (19.20)$$

If the growth rate is k per year, in $t_2 \equiv \frac{0.69}{k}$ years, the amount will double. For convenience, most people do two things. They express k as a percent instead of a fraction and they round out the 0.69 to 0.70, see 1.4.2. Combining these two things, you get the **Rule of Seventy** or

$$t_2 = \frac{70}{k_{\text{per cent}}}. \quad (19.21)$$

The population equation becomes

$$P(t) = P_0 2^{\frac{t}{t_2}} \quad (19.22)$$

A population example: Some strains of bacteria, if given adequate food, will divide every minute. This is a doubling time of one minute or a rate of $0.69 \text{ minutes}^{-1} \simeq 0.7 \text{ minutes}^{-1}$. If you start the population with one cell at 11 how many cells do you have at noon? $e^{\frac{60}{0.7}} \simeq 6 \times 10^{37}$

If this number of cells fills a bottle, at what time is there half a bottle? – one minute before noon. You also have two bottles at 12:01 and so forth.

A second example: Electrical power has a growth rate of 7% per year. This is a doubling time of 10 years. In 100 years you have 10 doubling times or 2^{10} times the number of power plants in the country if we do not change things. Estimate the number of power plants now and how many there will be in 100 years. Estimate when there will be a power plant for every person. Consider the SUV phenomena. By ratcheting up the gasoline consumption per mile and combined with a population increase, the rate of gas consumption is increasing at about 5% per year. If the oil reserves are 100 years worth of consumption at current consumption rates, how long will that supply last in the face of the SUV phenomena and population growth. Suppose we discovered new reserves to extend that supply to 1000 years. How long will it last in the face of the SUV phenomena and population growth.

A third example: Your parents are giving up about \$10,000 per year to send you to college. For 5 years that is \$50,000. If they had not spent that money, they could have put it into their retirement account. That account earns 7%. If they are in their late 40's and retire in the late 60's, this allows two doubling times. Thus at retirement this is worth \$200,000. If they continue to get 7% per year and expect to live another 15 years this is worth about \$25,000 per year in their golden years.

On the other hand if you have negative growth, $k < 0$, the population disappears asymptotically and fairly rapidly, $P = P_0 e^{-|k|t}$. Instead of doubling times, you now have halving times, often called the half life of the sample.

$$t_{1/2} = \frac{0.69}{|k|} \quad (19.23)$$

Using this you can write

$$P(t) = P_0 e^{-\frac{0.69}{t_{1/2}} t} = P_0 2^{-\frac{t}{t_{1/2}}}, \quad (19.24)$$

similar to the relationship for positive rates of growth, $P(t) = P_0 2^{\frac{t}{t_2}}$.

Another important feature of the population equation follows from the defining equation:

$$\frac{dP}{dt} = kP \rightarrow \frac{d(e^x)}{dx} = e^x \quad (19.25)$$

The slope of the exponential function is equal to the exponential function. This is the special case of the more general case,

$$\frac{d(e^u)}{dx} = e^u \frac{du}{dx} \quad (19.26)$$

19.8 New Clocks or Even More on Phasors*

Manipulation of the clocks in doing the analysis of Fresnel's construction is often difficult and unintuitive. Now that we have the exponential function there is a simple way to handle these clocks or better called phasors. Let's summarize the use of the Feynman clocks. At each place on the initial place on a path, we identified a clock as designated by a two dimensional arrow. The clock was then carried over the path of interest to the end place on the trajectory. On each segment of the path, the clock hands were rotated by the time of travel over that segment. At the end of the path, the carried clock was added with the clock hands of all the other clocks from the other contributing paths to the signal at the place. The addition was a 'tip to tail' addition of the two dimensional vectors the clock hands represent.

We can accomplish the same end by representing Feynman's clocks as $A_0 e^{i\theta}$ where A_0 is called the magnitude and is an ordinary number, θ is called the phase, hence the name phasor, and i is the famous, infamous?, number whose square is $i^2 = -1$ usually denoted the unit imaginary number. The analysis of phasors as clock hands is the same as that of the complex numbers in the form $A_0 e^{i\theta}$ except that the conventions are different. For the clock hands, the angle was measured from the vertical and advances in a clockwise direction. For the complex numbers the angle is measured from the horizontal and is positive in the counterclockwise direction.

In the quantum mechanics case, the angle θ is set by the action

$$\Delta\theta = \frac{\Delta S}{\hbar} \quad (19.27)$$

The best way to get A , the amplitude, is $A_0 e^{-i\theta} A_0 e^{i\theta} = A^2$ where the process of replacing i by $-i$ is called conjugating so that this process is to

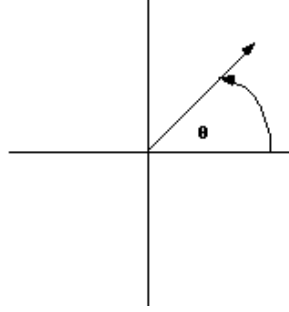


Figure 19.15: **Phasor as a Complex Number** A phasor can be interpreted as a complex number. In this figure the horizontal represents the real part and the vertical the imaginary part of the complex number $A_0 e^{i\theta}$ where A_0 is the amplitude or length of the vector and θ is the phase.

take the amplitude, conjugate it, and multiply the conjugate and the original amplitude. That result is the amplitude squared.

$$\psi(x_f, t_f; x_0, t_0) = \sum_{\text{traj.}} e^{\frac{iS}{\hbar}} = \sum_{\text{Path}} e^{\frac{i \sum_{\text{traj.}(x_0, t_0)}^{(x_f, t_f)} L(V, x) \Delta t}{\hbar}} \quad (19.28)$$

where a path is designated by some set $\{x_i\}$ of coordinate values at each time slice. All paths are achieved by now allowing each $\{x_i\}$ to take on all values.

To be specific let's do the free particle. In between any time slice, we use the straight line path to evaluate the action. This implies that the velocity is a constant and is the inverse slope between the end points of the segment, $v = \frac{x_f - x_0}{t_f - t_0}$, for that interval. Thus,

$$L(v, x) = \frac{m}{2} v^2 \rightarrow S = \sum_{(x_0, t_0)}^{(x_f, t_f)} L(v, x) \Delta t = m \frac{(x_f - x_0)^2}{2(t_f - t_0)} \quad (19.29)$$

Using this for each time slice and setting the interval as ϵ with $n\epsilon = t_f - t_0$ and using the path designation as the set $\{x_i\}$ with each x_i allowed to take on all values:

$$\psi(x_f, t_f; x_0, t_0) = \prod_{i=1}^{n-1} \int_{-\infty}^{\infty} e^{im \frac{(x_i - x_{i-1})^2}{2\epsilon \hbar}} dx_i \left(2 \frac{\pi i \hbar \epsilon}{m} \right)^{-\frac{n}{2}} \quad (19.30)$$

This is a series of gaussian integrals. Using the following and a great deal of patience you get

$$\psi(x_f, t_f; x_0, t_0) = [2\pi i \hbar \frac{(t_f - t_0)}{m}]^{-\frac{1}{2}} e^{im \frac{(x_f - x_0)^2}{2\hbar(t_f - t_0)}} \quad (19.31)$$

This object is called the propagator. It is like the Fresnel construction, see Sec. 4, in optics tells you how things get from (x_0, t_0) to (x_f, t_f) . The product of it and its conjugate is the probability that you will find the particle at (x_f, t_f) .

By direct substitution you can show that

$$-\frac{\hbar}{i} \frac{\partial \psi(x_f, t_f; x_0, t_0)}{\partial t_f} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x_f, t_f; x_0, t_0)}{\partial x_f^2}. \quad (19.32)$$

This is the free particle Schrödinger equation. For the more general case,

$$\begin{aligned} \psi(x_f, t_f; x_0, t_0) &= \sum_{\text{traj.}} e^{\frac{iS}{\hbar}} \\ &= \sum_{\text{traj.}} e^{\frac{i \int_{(x_0, t_0)}^{(x_f, t_f)} L(V, x) dt}{\hbar}} \\ &= \sum_{\text{traj.}} e^{\frac{i \int_{(x_0, t_0)}^{(x_f, t_f)} \left(m \frac{v^2}{2} - V(x) \right) dt}{\hbar}} \end{aligned} \quad (19.33)$$

and, using the time slicing, you can show that

$$-\frac{\hbar}{i} \frac{\partial \psi(x_f, t_f; x_0, t_0)}{\partial t_f} = \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x_f^2} - V(x_f) \right) \psi(x_f, t_f; x_0, t_0). \quad (19.34)$$

This the full Schrödinger equation. A general state is propagated from an initial configuration and, in general, is not at some point but distributed

$$\psi(x, t) = \int \psi(x, t; x_0, t_0) \psi(x_0, t_0) dx_0 \quad (19.35)$$

The probability that the object will be at the place x is

$$\psi^*(x, t) \psi(x, t) dx = P(x, t) dx \quad (19.36)$$

where $P(x, t)$ is the probability that an experiment at time t will find the particle at the position x . Note that, because of the manner in which it is constructed, this state also satisfies the Schrödinger equation for any starting state.

19.9 The Uncertainty Relations

Consider the problem of finding the location of something in a microscope. Light shines on the specimen and enters the microscope lens and is deposited

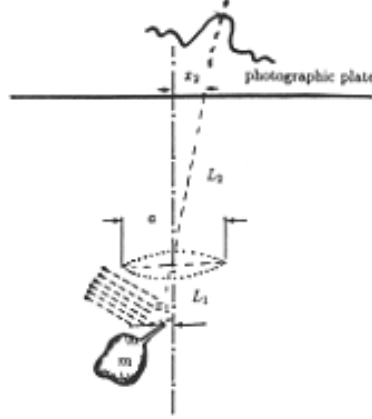


Figure 19.16: **Uncertainty Microscope** The position of a small object is recorded on a photographic plate with the use of a microscope.

on a photographic plate at a position x_2 .

Given the wave nature of light, the position of a bright spot behind a lens on the photographic plate is uncertain. There is diffraction. The light is focused by a lens. Since the brightness from the entering lens is spread you can only know the position of the grain of silver by

$$\Delta x_2 \approx \frac{1}{6} \frac{\lambda}{a} L_2 \quad (19.37)$$

Using $\frac{\Delta x_1}{\Delta x_2} \approx \frac{L_1}{L_2}$, we get

$$\Delta x_1 \approx \frac{1}{6} \frac{\lambda}{a} L_1 \quad (19.38)$$

Since the light that hit the electron in the material has a momentum $p = \frac{h}{\lambda}$, and it entered the lens, we know that the electron now has a random momentum

$$\Delta p \approx \frac{h}{\lambda} \frac{a}{2L_1} \quad (19.39)$$

or

$$\Delta x_1 \Delta p \approx \frac{h}{6 \times 2} \approx \frac{\hbar}{2} \quad (19.40)$$

This the famous Heisenberg Uncertainty Relationship.

It a special case of a very general set of relations in quantum systems. Variables are paired in sets that have an incompatibility. You can measure either of them with great precision but if you measure one with a certain dispersion, the other will have a dispersion also and the products of these dispersions are related – you cannot measure both of the variables with precision simultaneously. Two variables such as momentum and position that have this relationship are considered incompatible. Note that momentum and space translation, unimportance of position, are related through Noether's Theorem and symmetry.

19.9.1 The Uncertainty Principle and the Quantum Mechanical Harmonic Oscillator

$$V(x) = \frac{1}{2}\kappa x^2 \quad (19.41)$$

We can use that uncertainty principle, Equation 19.40, to determine the energy of the lowest configuration.

$$\Delta p = \frac{\hbar}{2\Delta x} \quad (19.42)$$

Using $p \approx \Delta p$ and $x \approx \Delta x$ and plugging into the energy relationship for the harmonic oscillator,

$$E = \frac{p^2}{2m} + \frac{1}{2}\kappa x^2 = \frac{1}{2m}\left(\frac{\hbar}{2\Delta x}\right)^2 + \frac{1}{2}\kappa\Delta x^2 \quad (19.43)$$

To find the Δx that minimizes the energy look at plot of $\frac{1}{\Delta x^2}$, Δx^2 , and the sum.

For our problem the minimum occurs at

$$\Delta x = \frac{\sqrt{\hbar}}{\sqrt{2}m^{1/4}\kappa^{1/4}}. \quad (19.44)$$

or defining $\omega \equiv \sqrt{\frac{\kappa}{m}}$,

$$E_0 = \frac{\hbar\omega}{2} \quad (19.45)$$

Even in the lowest energy state, the particle still has energy – no surprise. Thinking in terms of our particle traveling over all paths, even though we are in the lowest energy state, the particle has some spread in position and momentum and thus has some energy.

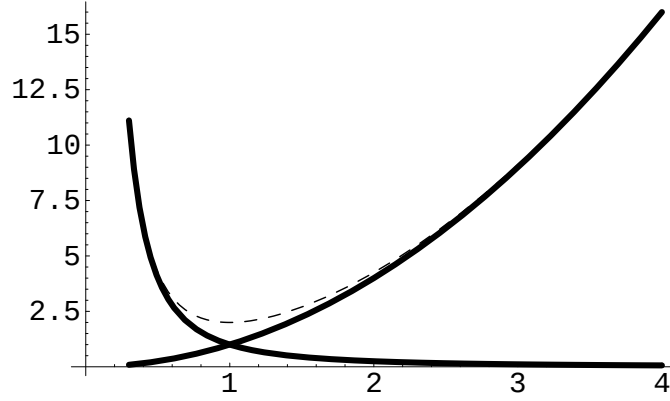


Figure 19.17: **Ground State of Oscillator** A plot of equation 19.43, the energy of the quantum oscillator, in the lowest energy state as a function of the uncertainty in position, Δx . The term that comes from the kinetic energy and goes like $\frac{1}{\Delta x^2}$ is large and dominates for small Δx . Where as, the term from the potential energy that goes as Δx^2 dominates for larger Δx . Thus there is a minimum somewhere between these two domains.

19.9.2 Oscillator Ground State Wave Function

In our interpretation of the Feynman path integration, the field that is associated with the position of the mass at the end of the spring is spread out. Again, this occurs even in the lowest energy state. If we guess that the field, called the wave function, is a gaussian, which it is in fact,

$$\psi(x) = N e^{-\frac{x^2}{2\sigma}} \quad (19.46)$$

Then $P(x)$ is $\psi^2(x)$ or

$$P(x) = N^2 e^{-\frac{x^2}{\sigma}} \quad (19.47)$$

The condition on N is found from the fact the $\sum_{\Delta x=-\infty}^{\infty} P(x)\Delta x = 1$. This is the statement that the particle must be found somewhere. This satisfies

$$N^2 \int_{-\infty}^{\infty} e^{-\frac{x^2}{\sigma}} dx = N^2 \sqrt{\pi\sigma} = 1 \quad (19.48)$$

or the lowest energy state wave function is

$$\psi(x) = \frac{1}{\sqrt[4]{\pi\sigma}} e^{-\frac{x^2}{2\sigma}} \quad (19.49)$$

to get the $\Delta x = \sqrt{\frac{\hbar}{m\kappa}}$, then $\sigma = \frac{\hbar}{\sqrt{m\kappa}}$.

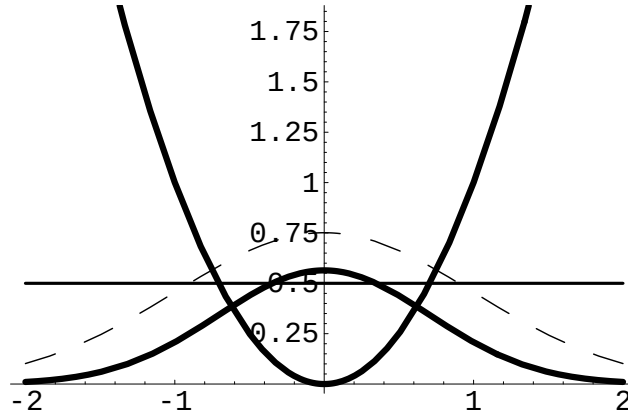


Figure 19.18: **Wave Function and Energy of Ground State** The thick curve that is concave down is a plot of the wave function for the lowest energy state of the quantum oscillator, Equation 19.49. The thick concave up curve is the potential energy of the oscillator. The dashed curve is the probability the the mass in the oscillator will be found at the position x , Equation 19.47. The horizontal line is the energy of the ground state, $\frac{1}{2}\hbar\sqrt{\frac{\kappa}{m}}$.

The thick curve that is concave up is the potential energy, the thin horizontal line is the energy of the lowest energy state on this energy scale. The thick curve that is concave down is the amplitude, ψ , and the dashed curve is the probability distribution.

Notice how the solution 'leaks' into the classically forbidden region. There are places at which the potential is greater than the total energy.

19.10 Localized States

19.10.1 An Aside on the Particle in the Box

Consider a system that has the following potential energy.

$$V(x) = \begin{cases} \infty, & x \leq 0; \\ 0, & 0 < x < L \\ \infty, & L \leq x \end{cases} \quad (19.50)$$

A particle free to move in this potential is the best model of a particle in a box. To be consistent, we require that $P(x)$ be zero everywhere outside

the region, $0 < x < L$, the inside of the box. If the particle could be found outside the energy would be infinite. It is in this sense that this is the model for a particle in a box. Since $P(x)$ must be zero outside the box, ψ must be zero and since it is continuous, ψ must be zero at the edges of the box or $\psi(0)$ and $\psi(L)$ must be zero. The simplest function that does that is

$$\psi(x) = \begin{cases} 0, & x \leq 0; \\ N \sin\left(\frac{\pi}{L}x\right), & 0 < x < L \\ 0, & L \leq x \end{cases} \quad (19.51)$$

The probability requirement, $\int_{-\infty}^{\infty} P(x)dx = \int_{-\infty}^{\infty} \psi^2(x)dx = 1$ is that

$$\int_{-\infty}^{\infty} \psi^2(x)dx = \int_0^L N^2 \sin^2\left(\frac{\pi}{L}x\right) dx = 1 \Rightarrow N = \sqrt{2\frac{\pi}{L}} \quad (19.52)$$

In the box, the particle has a wavelength which implies that we know the wavelength, $\lambda 2L$ and thus we know the momentum in the box, $p = \frac{h}{2L}$. This makes the energy of this state

$$E_{low} = \frac{p^2}{2m} = \frac{h^2}{2m4L^2} = \pi^2 \frac{\hbar^2}{2mL^2} \quad (19.53)$$

This is the lowest energy state. There are other states that have $\psi(0)$ and $\psi(L)$ equal to zero.

$$\psi_n(x) = N_n \sin\left(n\frac{\pi}{L}x\right) \quad (19.54)$$

for $n = 1, 2, 3, \dots$. It turns out that N_n is independent of n and thus is the same as before. In this case, $\lambda = \frac{2L}{n}$ and the $p_n = h\frac{n}{2L}$. The energy is

$$E_n = \pi^2 \frac{\hbar^2}{2mL^2} n^2, \quad n = 1, 2, 3, \dots \quad (19.55)$$

Several of the wave functions are shown in Figure 19.19. These states are separable and thus are stationary, the time dependence is trivial canceling out of the probability distribution functions.

The probability distribution functions are the square of the wave function. For these separable or stationary states there is no time dependence. These are shown in Figure 19.20.

The energies are discrete. These are the states that have a fixed energy.

We add energy by adding nodes, places where ψ is equal to zero.

You can make states of almost any energy, none less than the lowest. You can make states where the particle is found at least initially in some region

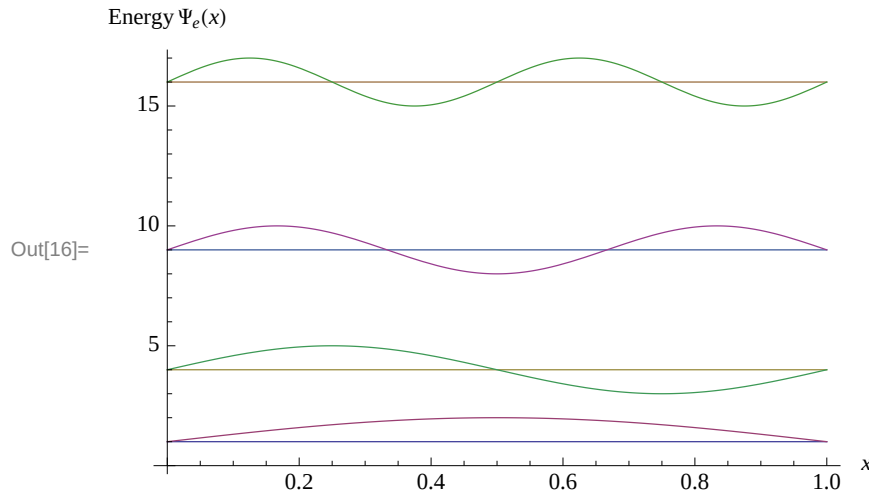


Figure 19.19: **wave Functions for Energy States** A plot of the wave functions for stationary or definite energy states of the particle in the box. The plots are a combination of the fixed energy values, represented by horizontal lines and the wave functions of the states with that energy raised so that each one is plotted on the horizontal energy value associated with that state.

of the box. These are just like the normal modes in the stretched string. You can start the string with an arbitrary pluck. It will be a superposition of the normal modes. Here you superimpose the definite energy states. These definite energy states are called stationary states since they have a given energy they can be interpreted as having a given frequency and they just oscillate like the normal modes did.

19.10.2 Returning to the Oscillator*

Since we now know that the higher energy states are constructed by adding nodes we could guess that there are higher energy definite states and they are finite polynomials times the gaussian that we have discovered above. We also know that at large x , the energy value is negligible and thus all solutions at large x have the same fall off.

$$\psi_n = N_n P_n e^{-\frac{x^2}{2\sigma}} \quad (19.56)$$

The condition that the energy be definite forces the polynomial to take on a very specific form.

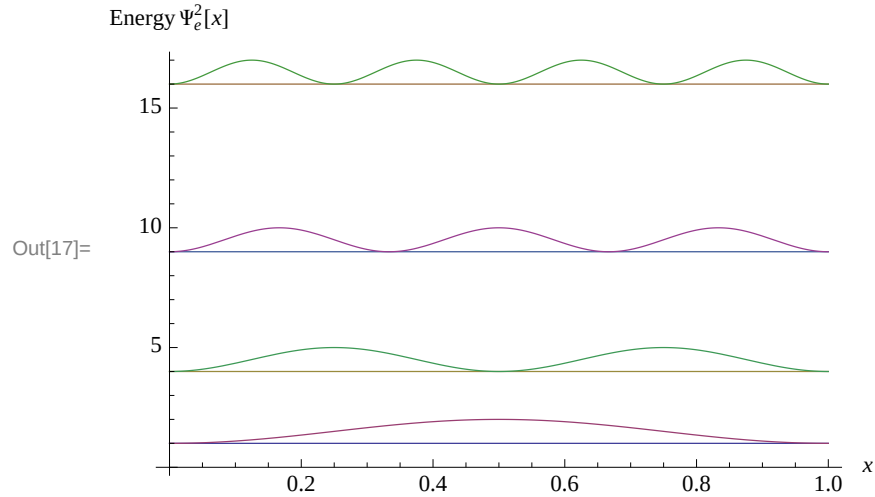


Figure 19.20: **Probability Distributions for Definite Energy States**

A plot of the probabilities for finding the particle at the position x for stationary or definite energy states of the particle in the box. The plots are a combination of the fixed energy values, represented by horizontal lines and the probability distributions of the states with that energy raised so that each one is plotted on the horizontal energy value associated with that state. See Figure 19.19 for the wave functions.

We find that the energy is

$$E_n = \frac{1}{2}(2n+1)\hbar\sqrt{\frac{k}{m}} \text{ for } n = 0, 1, 2, 3, \dots \quad (19.57)$$

We recover our lowest or zero point energy.

We can plot the first few wave functions, ψ_n , for states with definite energy, $E_n = \frac{1}{2}(2n+1)\hbar\sqrt{\frac{k}{m}}$, Figure 19.21.

Here I have raised the height of each ψ so that it is at its energy level and can be seen. Note that all the energies differ by same amount.

Plotting the probabilities, Figure 19.22

The fact that the energy of the state is linear in n cannot be over emphasized. **The energies of the simple harmonic oscillator count.** There is a minimal excitation $\epsilon = \hbar\sqrt{\frac{k}{m}}$ and that the n th state is one with n of these excitations. The unit of energy, $\hbar\sqrt{\frac{k}{m}}$, is the energy of a particle and the state has n particles in it.

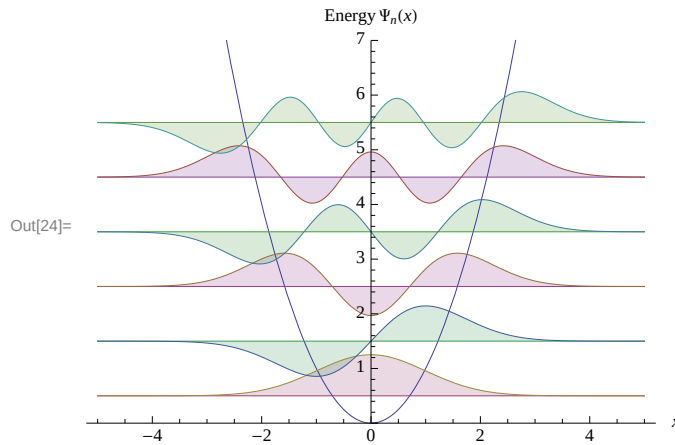


Figure 19.21: **Oscillator Energy States** A plot of the wave function for definite energy states of the quantum oscillator. Also shown is the potential energy and, in order to show them more clearly, the wave functions are raised so that each one's zero is at the corresponding height for that energy.

In addition, there is this magical lowest energy that is not zero energy and is $\frac{\hbar}{2} \sqrt{\frac{\kappa}{m}}$. That the lowest possible energy state has non-zero kinetic and potential energy is a direct reflection of the uncertainty principle, Section 19.9. In order to have zero potential energy, the oscillator mass would have to be located at the origin. But if it localized to just the origin, the uncertainty principle, Equation 19.40, would require that the state have a huge range in momentum and thus a huge kinetic energy. Clearly this is not the lowest energy state. Similarly, the state with no kinetic energy would have a huge position uncertainty and thus a huge potential energy. As we saw in Section 19.9.1, the lowest energy state is achieved with a compromise of spread in position and momentum.

19.11 Time Development of Quantum Systems*

19.11.1 Motion in Quantum Mechanics

For energy states, the time development of the state is especially simple. So far the $P(x)$ that we generate for the energy states are time independent; the ψ are time independent. How do the ψ develop in time? In the stretched string we know that each normal mode just oscillates and that the motion

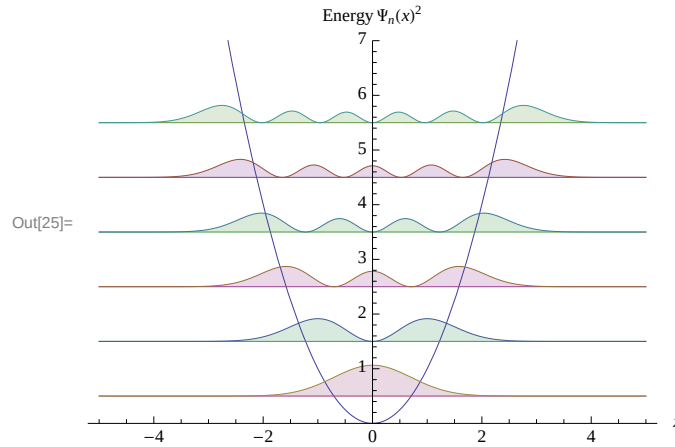


Figure 19.22: **Oscillator Probability Amplitudes** A plot of the probability that you will find the mass at the position x for the quantum oscillator. Also shown is the potential energy and, in order to show them more clearly, the probability curves are raised so that each one's zero is at the corresponding height for that energy.

comes from the fact that the normal modes have different frequencies and that as time advances the amplitudes of the normal modes change.

What we have found above are the states of definite energy. We treat these as normal modes in the sense that, since they have a definite energy, they have a definite frequency, $\epsilon = \hbar\omega$. In the case of quantum systems, we can use the all paths arguments to show that advance in phase by the rule that $\Delta\theta = \frac{\Delta S}{\hbar}$, to show that the states of definite energy develop in time with a phase that advances at the rate of this frequency, $\omega = \frac{\epsilon}{\hbar}$.

For any definite energy state, the $P(x, t)$ is time independent. In order to get any motion, we need to have the system in a superposition of several definite energy states. Then you get $P(x, t)$ in which there is motion.

So that we see that the quantum oscillator changes with time by the interference of the phasors associated with the definite energy states.

$$\psi(x, t) = e^{i\Delta\theta_0(t)}\psi_0(x) + e^{i\Delta\theta_1(t)}\psi_1(x) + \dots \quad (19.58)$$

where each phasor advances in angle as $\Delta\theta_i = \frac{\epsilon_i}{\hbar}t$.

19.11.2 Relation between the Quantum and the Classical Oscillator

It is interesting to compare the classical oscillator with the quantum mechanical one. In section 19.11.3, we will actually try to find how we can make a quantum oscillator act like a classical oscillator. In this section, we will go the other way. What is the corresponding classical configuration that looks like the quantum case. Remember that in the quantum case, you are dealing with small systems. In fact, it is only in the last few years that experimentalists have been able to manipulate few or single atom systems. Since the quantum mechanics is intrinsically probabilistic, see Chapter ??, we need to look at a configuration of the classical system that has a probabilistic interpretation. Since in the early days of quantum mechanics, everyone had been trained in only classical systems, there was a tendency to interpret the newly emerging quantum phenomena from a classical perspective, this will allow us to better understand the earlier interpretations of quantum mechanics. Using a classical oscillator with a random start with the same energy as a very excited quantum oscillator, say $\epsilon = \hbar\omega(n + \frac{1}{2})$ with $n = 20$.

The classical probability of seeing the mass at point x is inversely proportional to the time spent in that interval, i. e. $P_{cl}(x) \propto \frac{1}{\text{speed}}$. This should be compared with the quantum case.

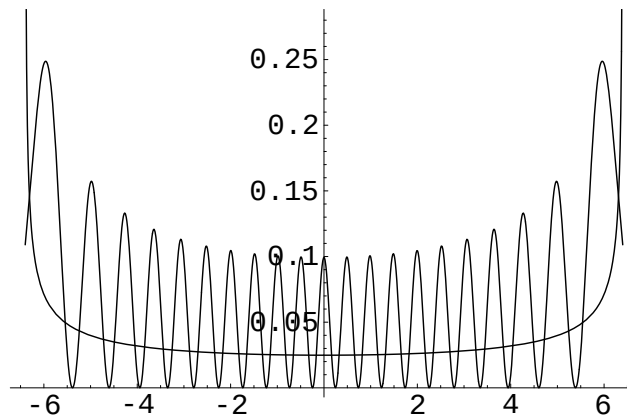


Figure 19.23: **Large Energy Quantum Oscillator** The probability amplitude for quantum oscillator that is in a large n state.

A more interesting question is “What is the state that has the mass pulled to the side and released?”. We can construct this state. Like the stretched string, it is a superposition of the definite energy states. Each en-

ergy state will evolve in time as its phasor advances and it is the interference of the states that determines how the probability distribution changes.

19.11.3 Classical Motion of the Quantum Oscillator*

How do we recover the classical limit? How do we get something that oscillates back and forth? If we displace the ground state solution a large distance compared to the spread in the ground state wave packet, we should have a solution that moves back and forth like the classical mass and spring. This state should walk, quack, and act like the classical oscillator.

$$\psi(x, 0) \equiv \psi_d(x) = \psi_0(x - d) = \frac{1}{\sqrt[4]{\pi}\sigma} e^{-\frac{(x-d)^2}{2\sigma}} \quad (19.59)$$

where $\sigma \equiv \frac{\hbar}{\sqrt{m\kappa}}$.

We want to find the time development of this state. The separable or definite energy states of the oscillator are the states with the simple time development; a separable state with energy E_n has the time evolution $\psi(x, t) = \psi_x(x)\psi_t(t) = \psi_x(x)e^{-iE_nt}$. Expand our stretched ground state as a superposition of the definite energy states.

$$\psi_d(x) = \sum_{n=0}^{\infty} \alpha_n \psi_n(x). \quad (19.60)$$

There is a definite procedure for finding α_n . Because the $\psi(x, t)$ satisfy the Schrödinger equation, you can show that the energy eigenstates satisfy $\int_{-\infty}^{\infty} \psi_{n'}^*(x)\psi_n(x)dx = \delta_{n',n}$ where $\delta_{n',n} \equiv 1$ if $n' = n$ and 0 otherwise. Thus $\alpha_n = \int_{-\infty}^{\infty} \psi_n^*(x)\psi_d(x)dx$. Using equation 19.59 and a table of integrals, we can show that

$$\alpha_n = \frac{\left(\frac{d}{\sqrt{2}\sigma}\right)^n}{\sqrt{n!}} e^{-\frac{d^2}{4\sigma}} \quad (19.61)$$

The probability of finding the displaced state with n excitations is α_n^2 . This is $P_n = \frac{\left(\frac{d^2}{2\sigma}\right)^n}{n!} e^{-\frac{d^2}{2\sigma}}$. This is a well known probability distribution, the Poisson distribution, see Section 19.11.4. The mean of this distribution is $\frac{d^2}{2\sigma}$ and the standard deviation is $\frac{d}{\sqrt{2}\sigma}$.

19.11.4 An Aside on the Binomial & Poisson Distribution

Both the related Binomial and Poisson Distributions are very common distributions and you should know about them independently of their importance

in quantum mechanics. These are the distributions that arise when you select a sample from a population. The classic example is the large bag of socks that are half red and half black. What is the chance of getting five red socks in a sample of ten socks? What is the chance of getting four red and six black socks. Although it may not seem to be important when we are dealing with socks and the condition that the population be huge may seem artificial when applied to socks, this distribution is extremely important in many cases. It is a special case of the binomial distribution which applies to sock sampling when you have finite or smaller bag of socks. The Poisson distribution is the limit that you obtain from the binomial distribution when the bag of socks becomes infinitely large. If you have a large population, preferably infinite, and want to draw samples from it, and if you expect to draw N , the probability that you will draw m is

$$P_m(N) = \frac{N^m}{m!} e^{-N}. \quad (19.62)$$

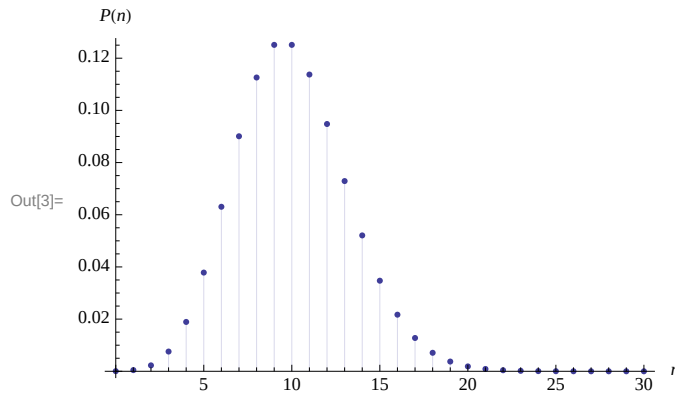


Figure 19.24: **Poisson Distribution with Expected Value of Ten** A plot of the Poisson Distribution for the case when the mean or most likely value is 10. For example in the case of the huge bag containing half red and half black socks, if you sample the bag by removing 20 socks, you expect that most of the time you will draw ten red and ten black socks. This is the distribution of number of red socks that you will get for a sample size that you expect will draw 10 red socks.

A dramatic feature of Figures 19.24, 19.25, and 19.26 is how much the spread of the distribution narrows as the mean gets larger. This is an important property of the Poisson distribution. The mean, N , which is also its

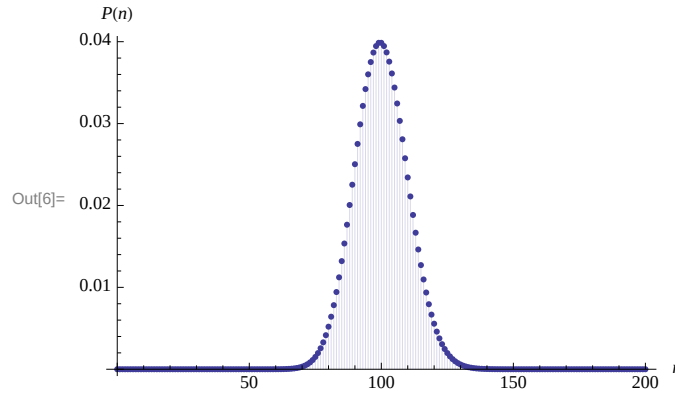


Figure 19.25: **Poisson Distribution with Expected Value of One Hundred** This is the distribution for a sample of a population in which you expect to draw 100 red socks.

peak or most likely value and the width are related. The width of the distribution is the range of values that have a certain likelihood. For instance, σ_{99} is the range of values in a sample of the population that have a 99% likelihood. You pick your comfort range and that sets the size of the σ with which you will live. The range of values set by σ_{99} says that, if your sample value was outside the range $N \pm \frac{1}{2}\sigma_{99}$, your chances of selecting that sample was one less than one in a hundred. σ_{99} , is

$$\sigma_{99} = 2.6 \times \sqrt{\text{mean}}. \quad (19.63)$$

This is often referred to as the square root of N rule. The rule is that in general the width of the distribution is proportional to $\sqrt{N}$ where N is the most likely value or mean for samples selected from that population. You can have weaker criteria for satisfaction than σ_{99} . A useful rule of thumb is to use the usual definition of the width called the standard deviation which holds for about two thirds of the cases. In that case, the rule is simply that

$$\sigma = \sqrt{N}. \quad (19.64)$$

The fact that the distributions, Figures 19.24, 19.25, and 19.26, narrow as N increases is a consequence of the square root of N rule. As N increases the range of values around N that are likely increases but the fraction of values that are likely divided by N gets very small as N gets large.

$$\frac{\sigma}{N} \rightarrow \frac{1}{\sqrt{N}} \quad (19.65)$$

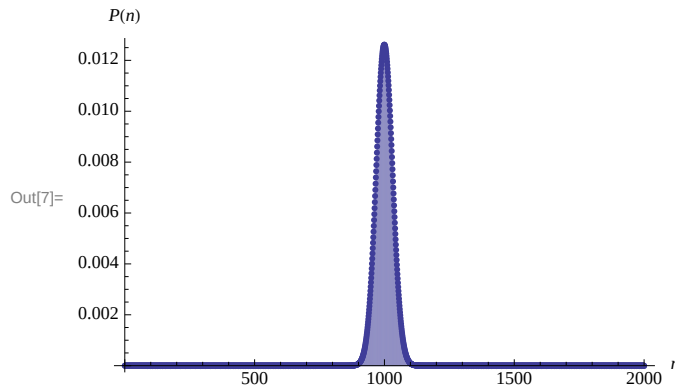


Figure 19.26: **Poisson Distribution with Expected Value of Thousand** This is the distribution for a sample of a population in which you expect to draw 1000 red socks. Note the dramatic narrowing of the width as the expected number of red socks increases. This is result of the $\sqrt{N}$ rule, equation 19.64.

for large N . The fact that there is a range of possible values around the expected value is called statistical fluctuations about the expected value. For these ideal sampling distributions, the square root of N rule means that as the size of the sample grows, the fractional size of the statistical fluctuations shrink. This is the basis of the fact that although there are always sampling errors, as the size of the sample grows, these go to zero. Out of chaos comes certainty.

Let's consider the simple case of political polling. Suppose that there are two candidates and 60,000,000 voters. This is a bag of red and green socks with a million socks, a really big bag. Since the bag is so big, we can use the Poisson distribution and its associated $\sqrt{N}$ rule. Suppose the real preference of the voters is about 60% for candidate A and 40% for candidate B. How we know that before sampling is a interesting problem that will be discussed shortly but let us just blissfully proceed and see what happens. If we take a sample of 100 voters, we would expect around 60 for candidate A and 40 for candidate B. But any set of 100 that we pick is a sample of the real population and we only have a finite chance of getting the most likely mix. Since this is a sampling problem, we realize though that our chances of any result have a distribution whose width is set by the expected value. Suppose we made our 100 calls and we found that there were 57 for candidate A. Do we conclude that 57% of voters like A and 43% like B. That's what the pollsters do. How do they handle the problem of the possibility that

they had a non-representative sample? They assume that they did not but then quote a margin of error for the poll. Using the standard deviation as our level of confidence, $\approx \frac{2}{3}$ of the samples will fall within one standard deviation, they have that $\sigma = \sqrt{57} = \sqrt{64-7} \approx 8 - \frac{7}{16} \approx 7.5$ or a % of $\pm \frac{4}{57} \times 100 \approx \pm 8\%$. If they want to use a more certain basis, they will use a two standard deviation error which is a probability of success of 95 %.

Of course, you see the circularity of the argument. They have to assume that their sample value for candidate A divided by the sample size is the correct fraction for candidate A for the population. They then base their error estimate on the sample value because there is nothing else available. This is better than nothing although what they should give is the sample size and the population size. The trouble with this approach is that most people would consider it ridiculous to call a hundred people to determine the preferences in a population of one million. From our square root of N rule, Equation 19.65, we see that they should increase their sample size to reduce the statistical fluctuations. If they went to a thousand people in their sample, they would reduce their fractional uncertainty by $\frac{1}{\sqrt{10}} \approx \frac{1}{\pi}$. If they went to a sample of 10,000, they would reduce the fractional error by a factor of $\frac{1}{10}$. There is a certain point in which it is not worth it to reduce your statistical error below what may be systematic errors in your sampling. In this case of polling, because they are using the phone, they may have a bias in their sample. The people with or at phones may be more likely to prefer one candidate over the other.

The Poisson is the distribution that you get when you look at rare events or background in a large sample. The famous example of the number of deaths due to horse kicks in corps of the German Cavalry recorded in the period 1875 to 1894.

Deaths per year	Number of Corps
0	144
1	91
2	32
3	11
4	2
5 or more	0

There are a total of 280 corps. The average number of deaths in a corps is $\frac{0 \times 144 + 1 \times 91 + 2 \times 32 + 3 \times 11 + 4 \times 2}{280} = 0.7$. Each corps is a sample of the population of all cavalry soldiers and thus the number of deaths should be distributed as

m	$P_m(0.7)$	Observed Fraction
0	0.5	0.51
1	0.35	0.33
2	0.12	0.11
3	0.03	0.04
4	0.005	0.01
5	0.007	0

Three mile island – downwind cancers. down wind population $\approx 25,000 \Rightarrow$ 250 deaths per year $\Rightarrow$ 50 cancer deaths per year. In three years there were 144 deaths. The expected rate was 142. Are there 2 excess deaths per year?

19.11.5 A Return to Classical Motion of the Quantum Oscillator*

Following Section 19.11.1 and particularly Equation 19.58 on Page 450, the time development of this state, 19.59 on Page 452, is

$$\psi_d(x, t) = \sum_{n=0}^{\infty} \frac{\left(\frac{d}{\sqrt{2}\sigma}\right)^n}{\sqrt{n!}} e^{-\frac{d^2}{4\sigma}} e^{-i\sqrt{\frac{\kappa}{m}}(n+\frac{1}{2})t} \psi_n(x) \quad (19.66)$$

Here the term for the phasor is written out explicitly as $e^{-i\sqrt{\frac{\kappa}{m}}(n+\frac{1}{2})t}$. What is the $\langle x \rangle_d(t)$?

$$\langle x \rangle_d(t) = d \cos\left(\sqrt{\frac{\kappa}{m}}t\right) \quad (19.67)$$

Problem: Show that

On the other hand if you calculate the expected value of x in a state with a definite energy you find that it is zero. **In any state with a definite number of excitations the expected position is 0.**

On the other hand, the expected value of the position squared is related directly to the energy and thus is not zero.

Some examples of superposition

An example of superposition that we will need later on. Up till now we have considered light of only one frequency and superimposed multiple sources and had it interfere. Now consider three sources with almost the same frequency at the same point, say one at some average frequency and one at

a slightly lower frequency and one higher by the same amount, $\Delta\omega$. Over time what do you get at the point?

If you look on a short time scale compared to $\frac{1}{\Delta\omega}$:

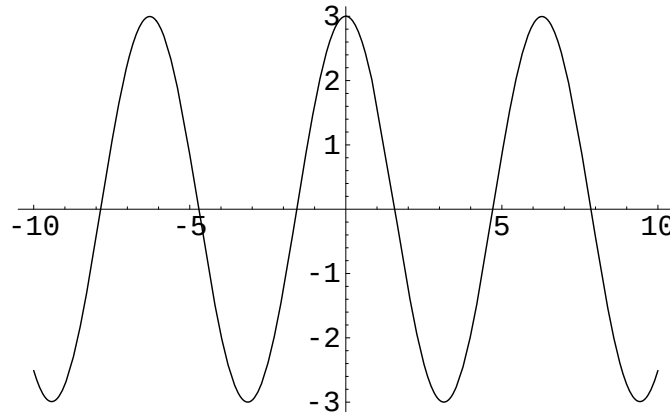


Figure 19.27: **Three Harmonic Amplitudes** The superposition of three harmonic amplitudes with close frequencies looked at on a short time scale. In this case, the three signals have a radian frequency, ω , of sec^{-1} , 1.01sec^{-1} , and 0.99sec^{-1} . On the time scale of seconds, the resultant is a signal with frequency of about sec^{-1} and an amplitude that varies on a larger time scale. To view the resultant on a longer time scale, see Figure 19.28

In this case, $\Delta\omega = 10^{-2}$ in inverse units of what ever the time unit is. We are looking over a 20 second time period.

If you look on a large time scale compared to $\frac{1}{\Delta\omega}$.

The signal becomes localized in time. There are periods of large amplitude and periods of small amplitude.

In the double slit that had two spatially separate sources that we superimposed. That lead to a pattern in space for the signal, here we get a pattern in time.

The more pieces you include the better. This case has five.

This seems to be a trivial idea and yet it leads to many of the quandaries of quantum mechanics.

The effect of adding sources is very dramatic. In Figure 19.30, shows the superposition of fifty sources with a total spread in the frequency that is $\frac{\Delta\omega}{\omega} = 10^{-3}$.

Again though, if you look at times short compared to $\frac{1}{\Delta\omega}$ it looks okay.

For times long compared to $\frac{1}{\Delta\omega}$ you get it active and then inactive for a long time. Notice how tightly you can pack the cluster when you have lots

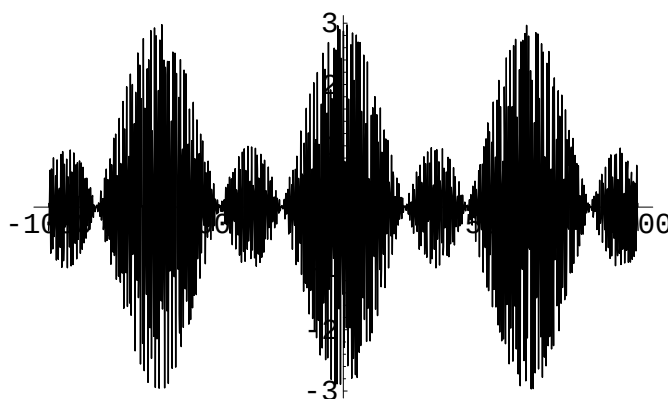


Figure 19.28: **Superposition of Three Harmonic Amplitudes** The superposition of three harmonic amplitudes with close frequencies looked at on a short time scale. The same situation as in Figure 19.27. On the time scale of hundreds of seconds, the resultant is a signal with frequency of about sec^{-1} and an amplitude that varies on this larger time scale.

of terms.

This is part of a general pattern. If you spread in frequency, you localize in time. You can localize only if you spread in a related variable. In other words we get a tight localization in t if we spread out in ω . If you use a tight value of ω you are spread out in t . The spread in t , Δt , and the spread in frequency, $\Delta\omega$ are related by $\Delta t \Delta\omega \approx 1$.

Since we know that there is a momentum associated with motion in quantum mechanics and that the momentum is proportional to the wavelength, $p = \frac{h}{\lambda}$, we get a similar relation with wavelength and position as with frequency and time.

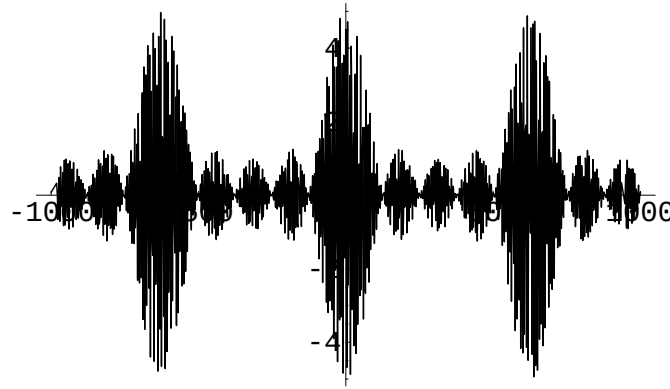


Figure 19.29: **Five Harmonic Amplitudes** The superposition of five harmonic signals with close frequencies on time scales that are long compared with $\frac{1}{\Delta\omega}$.

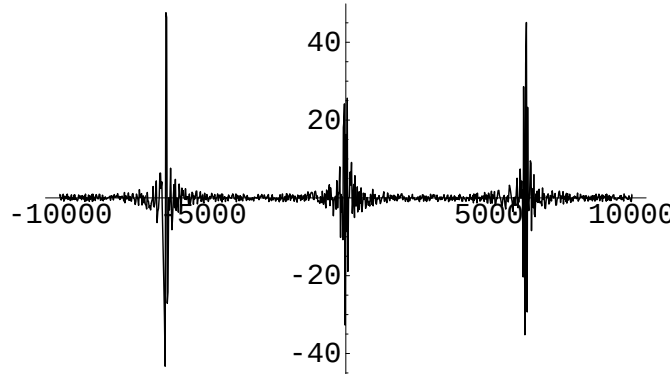


Figure 19.30: **Fifty Harmonic Amplitudes** The superposition of fifty harmonic signals with close frequencies on time scales that are long compared with $\frac{1}{\Delta\omega}$.

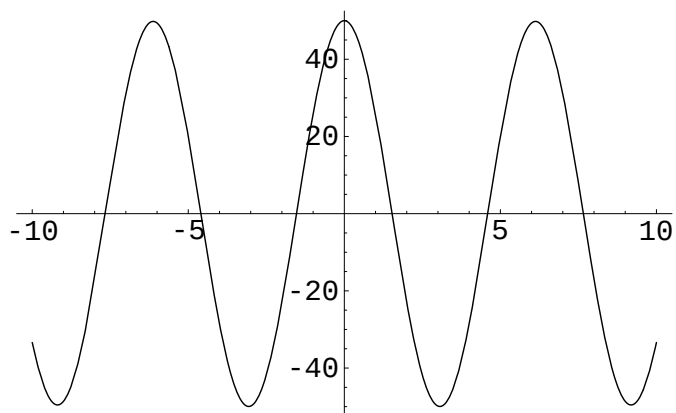


Figure 19.31: **Five Harmonic Amplitudes Revisited** The superposition of five harmonic signals with close frequencies on time scales that are small compared with $\frac{1}{\Delta\omega}$.

Chapter 20

Single Particle Quantum Mechanics

20.1 Single Particle Quantum Mechanics for Massive Particles

20.1.1 Introduction

The implications of Planck's and Einstein's inferences when dealing with the interactions of electrons and light were incredibly far reaching. Planck's realization that the structure of the spectrum of the emissions of a black body cavity required that light had a discretely countable constituent that was identified by its energy content which was given by the light's frequency, $\epsilon = hf$, and that these constituent elements were strangely indistinguishable. The person who recognized the far reaching implications of Planck's work was Einstein. He connected Planck's black body processes with any energy exchange processes between electrons and light. He realized that these process could be observed in the photoelectric process described in Section 19.3 on page 423. Analysis of the operation of the photo cell showed that the process of energy transfer from the light to the electrons was local but stochastic in both space and time within the illumination pattern of the light.

It is important to realize that this will not be our final theory of energy and matter. In this section, we are restricting ourselves to the case of low energy processes with a fixed number of massive objects. In fact, our entities are basically light and electrons. These are processes in which the scales of energies are small compared to a few $m_e c^2$ where m_e is the mass of the

electron, see Section 15.2 on page 334. The theory that is applicable at all energies will be constructed using the insights of the few electron low energy models.

20.1.2 Implications of Young's Double Slit with Electrons

From the observation that both light, historically considered a field object, and electrons, historically considered a particle, exhibit interference phenomena in Young's double slit experiment; the patterns of bright and dark occur whether the beam is electrons or light.¹

When dealing with light, we were convinced that this pattern of bright and dark spots of illumination can only be interpreted by having the light be composed of causal agents that add on superposition but can not be directly observable, their squares representing the brightness defined as the energy per unit area per unit time. Light was a field object. Replacing the screen with Einsteinian photo cells and running the system at very low detection rates, the pattern of clicks on the detectors built up ultimately to follow the illumination pattern of the classical Young experiment with the relation $I(x)A = n(x)hf$, where $I(x)$ is the brightness of the light, A is the area of the individual photo cells and $n(x)$ is the number of photons detected at the place x on the screen.

For light, following Maxwell, we can develop an action that produces via the process of minimization Maxwell's equations for the evolution of the light. Since the Maxwell equations are time and space translation symmetric, there are conserved energy and momentum constructs. It is a traveler field and the frequency and wavelength are related by $\lambda = \frac{c}{f}$ where c is the speed of light. This action also will be space and time translation symmetric and thus there has to be a classical momentum and energy. For a beam of light, the classical momentum, p and the classical energy, E of the beam satisfies $p = \frac{E}{c}$. If $p = \frac{E}{c}$ for light and the constituents have an energy content $\epsilon = hf$ then the momentum of each constituent is $p = \frac{hf}{c} = \frac{h}{\lambda}$. Thus since the pattern on the screen is determined by a wavelength for the Einsteinian photodetector case we can say that it is determined by the momentum of the constituents.

¹Actually, it is not true that we can do a Young's double slit experiment with electrons. The electron in this context is a non-relativistic object whereas light is intrinsically relativistic at all energies including the low energies available to Young. The wavelength which we assign to electrons is so small there is no material that could serve as a slitted screen. Regardless, for the sake of simple pedagogy, we will use this gedanken or thought experiment.

For light, Maxwell discovered that the field is the electromagnetic field. Since we have similar patterns for electron or light we need that field construct for electrons which could account for the pattern in our Young's experiment with electron beams and a detector that indicates the presence of an electron. This new causal agent again has to work similar to the one we have for light; it is additive on superposition and it is local. Following the Einstein's interpretation of the process by which the photo cell operates its square is the probability of finding the electron at a place $\vec{x}$ at a time t . The field is designated $\Psi(\vec{x}, t)$.

20.1.3 The Schrödinger Probability Amplitude Field*

The $\Psi(\vec{x}, t)$ field has a local dynamic. The dynamic contains an evolution, a part telling how as time advances the field changes, and the local dynamic which states how this evolution develops from the conditions at any place. The dynamic for particles and, at low energy large mass, appropriate to a low energy electron, is the Schrödinger equation of tee shirt fame² or

$$i\hbar \frac{\partial \Psi}{\partial t}(\vec{x}, t) = -\frac{\hbar^2}{2m} \nabla^2 \Psi(\vec{x}, t) + V(\vec{x}) \Psi(\vec{x}, t). \quad (20.1)$$

This certainly appears to be a rather large leap in technical complications but we are now equipped to at least interpret it since we have been indoctrinated into the mysteries of field theory. Many of the symbols are obvious. m is the mass of the electron. $\hbar$ is $\frac{h}{2\pi}$ where h is Planck's constant. $V(\vec{x})$ is the energy at the place $\vec{x}$. i is that magic number whose square is -1 . There are three positional dimensions and thus the $\vec{x}$ is our place labels. Of course, Ψ is our field; the amplitude for the measurable probability of the electron being at the place $\vec{x}$ at time t , $n(\vec{x}, t) = \Psi^*(\vec{x}, t) \Psi(\vec{x}, t)$. The idea of the amplitude squared being the measured entity is expanded by the use of the complex number system. The ∇^2 is short hand for $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$. Finally, ∂ is what we have been writing as Δ . In other words, if we reduce our space to one dimension, x , and revert to our notation the tee shirt equation reduces to

$$i\hbar \frac{\Delta \Psi}{\Delta t}(x, t) = -\frac{\hbar^2}{2m} \frac{\Delta}{\Delta x} \left(\frac{\Delta \Psi}{\Delta x} \right) (x, t) + V(x) \Psi(x, t). \quad (20.2)$$

²This is the usual tee shirt example. Actually, there are two fields and two evolution equations. Equation (20.1) has a partner equation that is the equation for the complex conjugate field. Complex conjugation is the process changing all the i to $-i$. For reasons that I do not understand, many physicists act and think that there is only one relevant field in this system.

Except for the magic i , we have seen field dynamics like this one before. The first term in the dynamic contains the construction $\frac{\Delta}{\Delta x} \left(\frac{\Delta \Psi}{\Delta x} \right) (x, t)$. Read directly in our vocabulary, this is the slope of the slope of Ψ at (x, t) . This is what for want of a better name we have called the ‘Bendy’, see 5.3 on page 128 operator on $\Psi(x, t)$. In this sense our dynamic is now

$$i\hbar \frac{\Delta \Psi}{\Delta t}(x, t) = -\frac{\hbar^2}{2m} (\text{Bendy}(\Psi))(x, t) + V(x)\Psi(x, t). \quad (20.3)$$

We can handle the i problem by realizing that this is actually a two field problem and this single evolution equation evolves both of these fields. If we define $\Psi_R(x, t) \equiv \frac{\Psi(x, t) + \Psi^*(x, t)}{2}$ and $\Psi_I(x, t) \equiv \frac{\Psi(x, t) - \Psi^*(x, t)}{2i}$. Both of these fields are real, $\Psi_R^* = \Psi_R$ and $\Psi_I^* = \Psi_I$. The final form of our dynamic is this coupled field dynamic

$$\begin{aligned} -\hbar \frac{\Delta \Psi_I}{\Delta t}(x, t) &= -\frac{\hbar^2}{2m} (\text{Bendy}(\Psi_R))(x, t) + V(x)\Psi_R(x, t) \\ \hbar \frac{\Delta \Psi_R}{\Delta t}(x, t) &= -\frac{\hbar^2}{2m} (\text{Bendy}(\Psi_I))(x, t) + V(x)\Psi_I(x, t). \end{aligned} \quad (20.4)$$

This separation argument is true in any number of spatial dimensions. In other words, the use of the i is a convenient notational tool for combining two real fields into one complex field.

The Schrödinger field is the amplitude for the probability of finding the electron at the place x at time t . The probability of finding the electron at x at time t is the absolute value of the amplitude which in the case of a complex field is

$$n(x, t) = \Psi^*(x, t)\Psi(x, t) = \Psi_R^2(x, t) + \Psi_I^2(x, t). \quad (20.5)$$

We can find the mean position of the electron by multiplying the place x by the probability of the electron being at the place x and summing over all places,

$$\langle x \rangle(t) = \sum_{\text{all places}} x n(x, t) \Delta x \quad (20.6)$$

$$= \sum_{\text{all places}} \Psi^*(x, t) x \Psi(x, t) \Delta x \quad (20.7)$$

$$= \sum_{\text{all places}} x (\Psi_R^2(x, t) + \Psi_I^2(x, t)) \Delta x. \quad (20.8)$$

Beside the convenience of a single evolution equation there is an important symmetry that is at the heart of the nature of quantum mechanics. Any complex $\Psi(x, 0)$ can be replaced by

$$e^{i\theta}\Psi(x, 0) = \sin\theta\Psi_R(x, 0) + i\cos\theta\Psi_I(x, 0) \quad (20.9)$$

and this new field has the same evolution as the original. In our notation, we are saying that initial $\Psi_R(x, 0)$ and $\Psi_I(x, 0)$ that we start with that is consistent with observables like $n(x, 0)$ can be replaced by $\Psi_R(x, 0) = \sin\theta\Psi_R(x, 0)$ and $\Psi_I(x, 0) = \cos\theta\Psi_I(x, 0)$ for any θ and this will not change the physics of the problem. This result is a consequence of the fundamental process for the evolution of massive quantum objects³. It is also the source of the ability to expand computational capabilities as in quantum computation, quantum cryptography, and quantum teleportation.

20.1.4 The Field as a Causal Agent

The idea of Causal Agent was first introduced in Section 6.1 and described in an “Aside” on page 6. The Schrödinger field is the probability amplitude for finding an electron at the position x at time t .

20.1.5 The Magic of Separable Field Configurations

For any field evolution, there are special solutions called separable. These are solutions of the evolution equations that take the form $\Psi(\vec{x}, t) = \Psi_x(\vec{x})\Psi_T(t)$; the field solution $\Psi(\vec{x}, t)$ splits into the product of two functions, $\Psi_x(\vec{x})$ and $\Psi_T(t)$ each of which depend only on the spatial variables and time respectively.

20.1.6 Dimensional Considerations in Single Electron Quantum Mechanics

As expected the Schrödinger equation contains $\hbar$. The presence of $\hbar$ is the signal that we are dealing with a quantum phenomena. Also since it is the appropriate field for a massive particle. From these two parameters, it is impossible to establish a complete set of measurable constants; a mass

³Note that in Equation 20.1 and its derivative equations you cannot put $m = 0$. Massless particles cannot be described by Schrödinger evolution. The reason is that for those particles there is no non-relativistic domain of motion. Note also that this means that for energetic electrons we cannot use the Schrödinger equation for evolution and must use relativistic quantum field theory, a theory that allows particle creation.

M, a length L and a time T. Of course, $V(x)$ could and likely will contain parameters with a length dimension. Notice the for the free particle, $V(x) = 0$, and thus there is no length or time and, particularly, no velocity and thus the evolution of the system with $V(x) = 0$ is not a pure wave system with travelers independent of the shape of the initial disturbance. The particular case of the localized electron does not have $V(x) = 0$ everywhere but has non-zero potential for all points outside the box. It is hard walled if that potential is very large compared to the energy range of the electron being localized, basically infinite, or soft walled for cases in which the electron can be found outside the box. Also note that $\frac{\hbar}{E} \stackrel{\text{dim}}{=} T$

20.1.7 More than One Low Energy Electron

Our first realization is that the field for the case of m electrons must now be described by a

20.2 Examples

20.2.1 Free Particle

20.2.2 Particle in the box

Another simple system that we can deal with is a localized particle. In the extreme ‘spherical cow’ sense, the ideal particle in an ideal box. Let’s go even one step further by working in one spatial dimension, the real line. It is obvious that if we want to restrict the electron to a segment of the line, we can do so by having $V(x) = \infty$ for all the places outside that region. This will force $\Psi(x, t) = 0$ outside that region and thus $\Psi^*(x, t)\Psi(x, t) = 0$ for all places outside the allowed region. This is the idea of a hard walled box. It can only be considered an approximation to any realistic situation such an electron localized in an atom. Our purpose here is to show the intrinsic features of quantum systems and using simple confinement protocols to do so. For now, we are also requiring that we have one electron in the box. This requires that the sum on $n(x, t)$ over the length of our box equal 1 for all times since we must find our electron in the box or

$$\sum_0^L n(x, t) \Delta x = \sum_0^L \Psi^*(x, t) \Psi(x, t) \Delta x = \sum_0^L (\Psi_R^2(x, t) + \Psi_I^2(x, t)) \Delta x = 1$$

There are two approaches to this analysis. Have some $\Psi(x, 0)$ that is normalized, $\sum_0^L n(x, t) \Delta x = 1$ and then evolve it forward in time. This

would be the standard field approach. The alternative is to find the separable solutions and use their completeness to represent the original field configuration

Chapter 21

EPR, Quantum Measurement and Bell's Theorem

The combination of the observation that there is a probability amplitude that is constructed by superimposing states by adding contributions of all paths, a wavelike property, and the interactions being instantaneous and stochastic leads to what are often interpreted as paradoxes in the quantum behavior of things. These ideas are treated under the heading of measurement theory.

Historically we could not do experiments on individual fundamental particles and could only deal with ensemble systems and the language was developed in that context. We are now entering an era in which we can manipulate individual fundamental systems. We are finding that all the rules that were developed in the ensemble language work for the fundamental systems.

In the following we deal with light and photons as our fundamental entities. This is a choice of convenience. Everything that I do here goes through for electrons. or any fundamental entity. The photon is a particularly simple system to deal with .

21.1 A Two Level System

In order to understand the essence of quantum mechanics and the measurement process in particular, lets study the simplest system possible. We will work with a system that has only two states and thus can appear as only

a superposition of these two possible states. The double slit is an example. The light had to come from either slit one or slit two.

It turns out that light itself offers us an example of a two level system, the two polarizations of light. In the classical wave picture of light, the light is oscillations in the value of the electric field, $\vec{E}(x, t)$, and the magnetic field, $\vec{B}(x, t)$. Maxwell's equations determine that nature of the behavior of the electric and magnetic field. For light, the wavelike emissions, these equations require that the electric field be transverse to the direction of the motion of the light and that the magnetic field be perpendicular to both the electric field and the direction of propagation. Thus if we are given a direction for the light to travel, the electric field can only point in some direction in a plane, a two dimensional space.

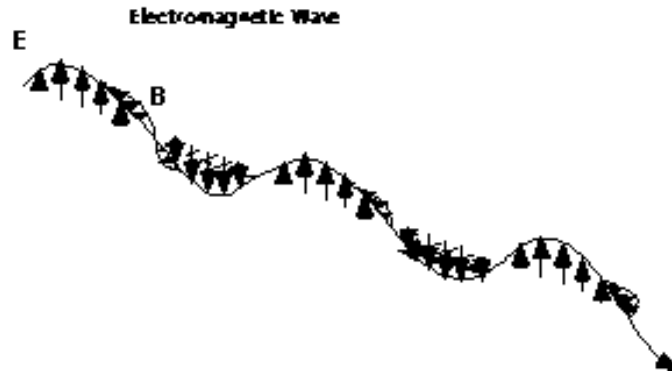


Figure 21.1: **Electric Wave** Light is high frequency, $5 \times 10^{-14} \frac{1}{\text{sec}}$, oscillations of the electric and magnetic fields that are traveling waves. Maxwell's equations require that the traveling waves have the electric and magnetic fields perpendicular to the one another and also to the direction of travel.

In one of our home experiments, we played with polarizers. These are sheets that absorb the light that has polarization transverse to a given direction and allows light polarized in the given direction to pass. In other words, light traveling in given direction comes in two varieties, let's say horizontal and vertical, which are at right angles with respect to each other.

If we had measured carefully in our home experiment with polarizers, we see that the light is a vector disturbance and that the vector amplitudes add.

If we insert an angled polarizer in the gap, we get some light.

For our purposes, the isolation of pure polarization states of light is

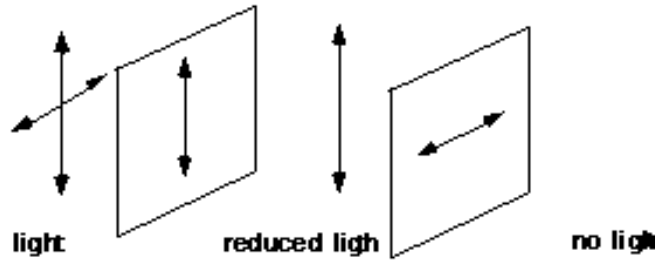


Figure 21.2: **Two Polarizers** Polarizers are a plastic material arranged in thin sheets that selectively absorbs light which is polarized in a given direction. Light from a thermal source is an incoherent jumble of light that is polarized in all directions. A polarizer placed in the beam removes all the light polarized in a given direction. In other words, a polarizer sheet has a hard direction, the direction of the polarized light that it absorbs and another direction perpendicular to that hard direction called easy that passes all light with that polarization. Unfortunately, polarizing sheets are identified by the easy direction. In the Figure above, thermal light enters from the left. The emergent light is now polarized in the easy direction. Since the thermal light has all polarizations, the intensity of the polarized light following the polarizer is half that of the incident beam. If a second polarizer is placed in the beam and oriented with the same orientation of the easy direction, the beam passes through it unaffected. On the contrary as shown above, if the second polarizer is rotated $\frac{\pi}{2}$ so that its easy direction is perpendicular to the first polarizer, no light passes.

better accomplished using birefringent materials¹. The advantage is that neither of the polarization states of the light are absorbed. They are just separated.

All of these properties are simple to understand when we examine them from the wave picture. It becomes difficult when we combine with how that interactions happen. We will use the calcite crystal system to make a series of measurements on this two level system.

¹Birefringence is the process by which a single beam of light is split into two separated beams each with a given polarization. It was discovered in the 17th century in calcite crystals and is now known to occur in many materials.

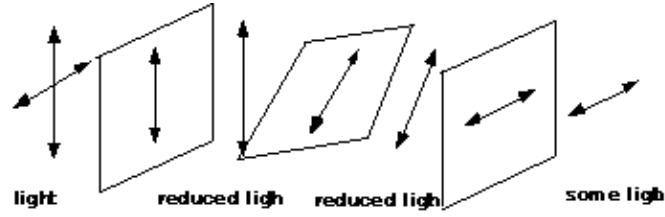


Figure 21.3: Three Polarizers

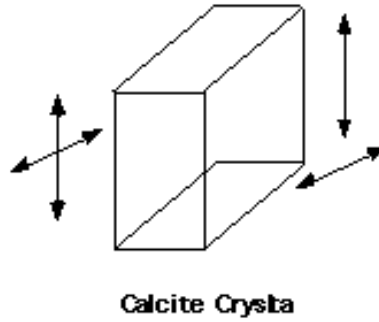


Figure 21.4: **Calcite Crystal** Calcite crystals have the property that they are birefringent. They separate a mixed polarization light beam into two spatially separated uniformly polarized beams of light.

21.2 More on polarized light as a two level system

We can use the calcite to divide a beam of light into the two polarization states:

The initial beam is a superposition of the two polarizations. The two emergent beams are in pure states of each polarization.

$$E_{in} = E_V + E_H = a\hat{E}_V + b\hat{E}_H \quad (21.1)$$

where if there are n photons per second coming in that are polarized equally between the two choices,

$$E_{in}^2 = n \quad (21.2)$$

and

$$a^2 + b^2 = n \quad (21.3)$$

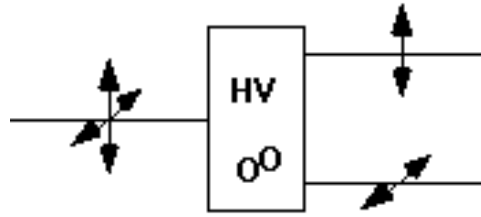


Figure 21.5: **Calcite Analyzer** The calcite crystal can cut so that the orientations of the two separated polarization states are along the vertical and horizontal directions. When this is the case, it is designated a 0° degree analyzer.

and I have defined $\hat{E}_V$ and $\hat{E}_H$ are the one photon state per second with the horizontal and vertical polarization, i. e. $\hat{E}_V^2 + \hat{E}_H^2 = 1$.

It is important to realize that you can have a calcite crystal that can separate out the two polarizations of light at different angles. In fact, any angle. Let's work with $\frac{\pi}{4}$ or 45° .

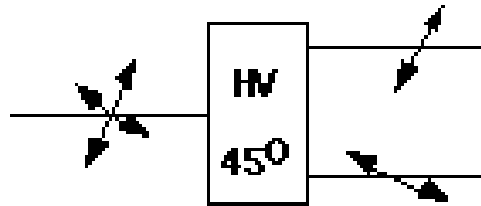


Figure 21.6: **A 45° Analyzer** By cutting the calcite crystal in different ways, the analyzer can be made to separate the beam into two beams with orthogonal polarizations in any angular orientations relative to the vertical.

What happens if you stack these things? As expected the second stage is consistent with the comment that the first stage in measuring the polarization has all the photons in it with the right polarization.

If you stack even more of the same type of polarizers you keep getting the same thing.

What happens if you mix angles?

Now you get light on all four channels.

What is the state of the light in the gap? In the upper leg, it is $a\hat{E}_V$ and the analyzer is at 45° . Calling the two relevant directions $+45$ and -45 , the state after the 45° analyzer is also described as

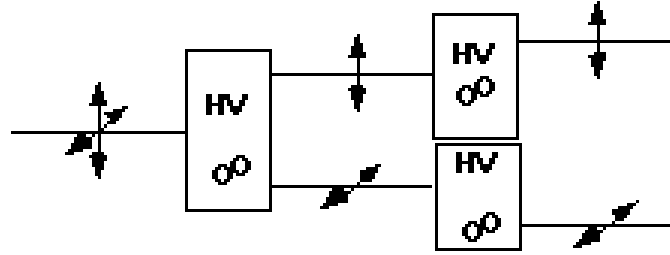


Figure 21.7: Stacked Analyzer

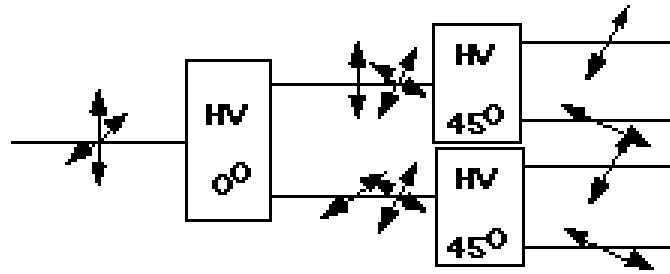


Figure 21.8: Stacked Turned Analyzer

$$a\hat{E}_V = c\hat{E}_{+45} + d\hat{E}_{-45} \quad (21.4)$$

with $a^2 = c^2 + d^2$ so that we have the correct number of photons. In other words, the state is a superposition of the $+45$ and -45 states. After the analyzer, we have c^2 photons in the upper most leg and d^2 photons in the second leg. What is the state of the photon in between the two analyzers. It is vertically polarized and it is a coherent mixture of $+45$ and -45 . From our experience with the polarizers or if you like from the wave description of polarization for an arbitrary orientation, θ , we have

$$\hat{E}_V = \cos\theta\hat{E}_\theta + \sin\theta\hat{E}_{\perp\theta} \quad (21.5)$$

In our case, we have the $c = a \cos(45)$ and $d = a \sin(45)$.

You can also reconstruct the state of polarization that has been split. What happens if we look to see if the photon is in the upper branch of the middle legs or in the lower branch?

All of this leads to the question of what is the state of the photon. Again I have to emphasize our basic rule. Everything goes over all paths and has instantaneous local interactions that are stochastic.

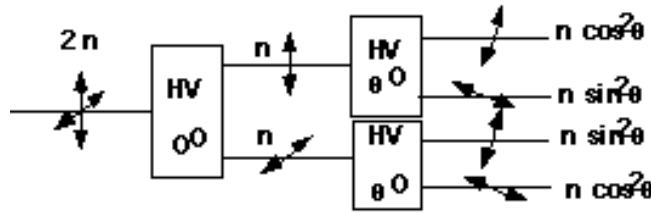


Figure 21.9: Intensity Analyzer

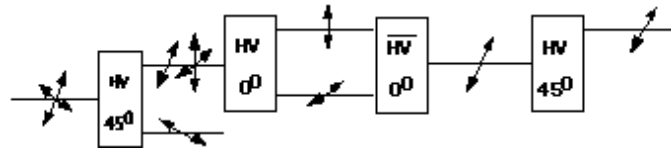


Figure 21.10: Tree of Analyzers

The state of the photon is in a superposition of polarization states. If we start with a vertically polarized beam and put it through a 45° analyzer, after the analyzer, it is half vertical and half horizontal. If we follow with a 0° analyzer, then we can then have, in one leg, a beam that is vertical. In the view of the individual photons that make up that last beam, when did they become vertical. Where they always vertical? If so where do we get the 45° beam come from. Did the vertical and horizontal photons interfere to produce the 45° beam. If we do this one photon at a time what happens? In our picture, we say that the effect of the analyzers is an interaction and we agree that interactions are stochastic and local in space and time. We say that the measurement changes the state, prepares the state. Another phrase that is used is that the superimposed state is collapsed into the measured state.

21.3 EPR and Bell's Theorem

Einstein always had trouble accepting the probability and field interpretation of quantum mechanics. In 1935, with B. Podolsky and N. Rosen, he proposed an experiment now called EPR that he thought made it clear that quantum mechanical predictions were implausible [EPR 1935]. The experiment involved the measurement of two separated photons that had been produced by a single decay. The apparatus is shown in Figure 21.12. At

the center is a source². At the end of the two decay pipes are two identical calcite crystals with photo detectors on each leg. If we modify the EPR

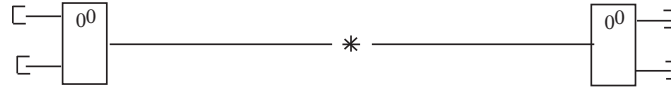


Figure 21.11: **EPR Apparatus** The apparatus discussed by Einstein Podolsky and Rosen

apparatus by putting an analyzer with an arbitrary orientation on the end

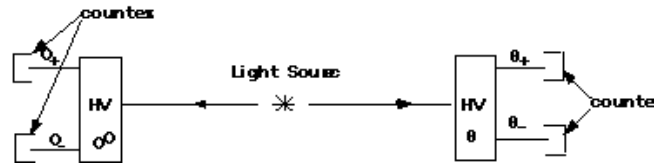


Figure 21.12: Modified EPR The apparatus discussed by Einstein Podolsky and Rosen

Using the properties of the analyzer, if we have n photons in this system we expect $\frac{n}{2}$ in the horizontal, 0_- , and $\frac{n}{2}$ in the vertical, 0_+ , counter of the left side of the apparatus. If the photon in the left is in 0_+ , we know what the photon in right is. The problem is that we have reoriented the analyzer. On the left we get

$$\begin{aligned}
 n(0_+, \theta_+) &= \frac{1}{2}n \cos^2 \theta \\
 n(0_+, \theta_-) &= \frac{1}{2}n \sin^2 \theta \\
 n(0_-, \theta_+) &= \frac{1}{2}n \sin^2 \theta \\
 n(0_-, \theta_-) &= \frac{1}{2}n \cos^2 \theta
 \end{aligned} \tag{21.6}$$

Define the correlation coefficient C

$$C \equiv \frac{\{n(0_+, \theta_+) + n(0_-, \theta_-) - n(0_+, \theta_-) - n(0_-, \theta_+)\}}{n} \tag{21.7}$$

²One such source would be neutral pi zero mesons. The π^0 s decay spontaneously into two identical photons traveling back to back. In addition, since the π^0 s have no angular momentum the two photons must have opposite polarizations

If θ is zero then $C = 1$, they are correlated. If θ is $\frac{\pi}{2}$, $C = -1$, they are anticorrelated. Halfway, $\frac{\pi}{4}$, $C = 0$, they are not correlated at all. For us the correlation coefficient is

$$\begin{aligned} C &= \frac{\{\frac{1}{2}n \cos^2 \theta + \frac{1}{2}n \cos^2 \theta - \frac{1}{2}n \sin^2 \theta - \frac{1}{2}n \sin^2 \theta\}}{n} \\ &= \cos^2 \theta - \sin^2 \theta \\ &= \cos 2\theta \end{aligned} \tag{21.8}$$

$$\tag{21.9}$$

Now consider three detections 0, ϕ , and θ . You can form lots of combinations, 0_+ , ϕ_- , θ_+ . Make a table of random combinations

$$\begin{array}{ccc} 0 & \phi & \theta \\ + & - & + \\ - & + & - \\ + & + & + \\ \cdot & \cdot & \cdot \end{array} \tag{21.10}$$

Let $n(\phi = +, \theta = -)$ be the number of sets with that configuration and so forth. Using figure 21.13, you can show that

$$n(0 = +, \phi = +) + n(\phi = -, \theta = +) \geq n(0 = +, \theta = +) \tag{21.11}$$

In figure 21.13, the slices of the pie represent the number of triplets of each type. Note that $n(0 = +, \theta = +)$ is represented by sector AOC. Similarly, $n(\phi = -, \theta = +)$ is given by sector COE and $n(0 = +, \theta = +)$ by BOD. Clearly AOC + COE must be greater than or equal to BOD so it follows that

$$n(0 = +, \phi = +) + n(\phi = -, \theta = +) \geq n(0 = +, \theta = +) \tag{21.12}$$

You do randomly different EPR experiments

Using the first set up we can measure $n(0_{\pm}, \phi_{\pm})$ and so forth. We can measure all the parts of the inequality

$$n(0_-, \phi_+) + n(\phi_-, \theta_+) \geq n(0_-, \theta_+) \tag{21.13}$$

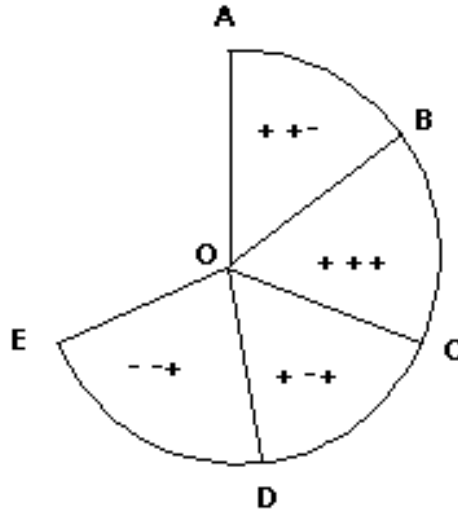


Figure 21.13: **Pie Chart of Probabilities** Pictorial representation of the table of randomly generated two valued choices for three things assignments. The area of the slices represent the fraction of the events with that assignment. Any probabilistic system has to be expressible in a chart of this kind.

Using equations 21.6

$$\cos^2 \phi + \sin^2(\theta - \phi) \geq \cos^2 \theta \quad (21.14)$$

Pick $\phi = 3\theta$. We will obtain

$$\cos^2 3\theta + \sin^2 2\theta - \cos^2 \theta \geq 0 \quad (21.15)$$

But this should always be greater than zero. So quantum mechanics predicts things that can not happen with local random labels. The data follows the quantum mechanical prediction. Thus there can be no hidden variables theory consistent with these measurements.

Add material on incompatible measurements.

You are stuck with the measurement problems of quantum mechanics.

role of the observer

collapse of the wavepacket

Schrödinger's cat

many worlds

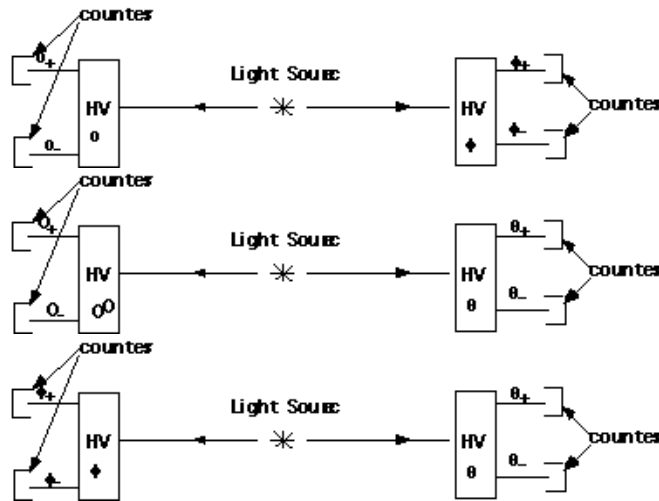


Figure 21.14: **Bell's configuration of Three EPR Apparatus** Bell suggested that three different EPR set ups with three different angles be run. Third of three configurations used in Bell's inequality. This apparatus measures the correlation between the angle ϕ and θ .

21.3.1 What is a particle and what is the field ?

We now know quite a bit about the photon. This is the object that carries the energy and momentum of the electromagnetic field. Yet we know that electromagnetism has field properties. How do we reconcile these observations. We should realize that we observe the field nature when there are many photons present, i. e. in cases in which the energy is many times $\hbar\omega$. How do we make a quantum theory of the electromagnetic field?

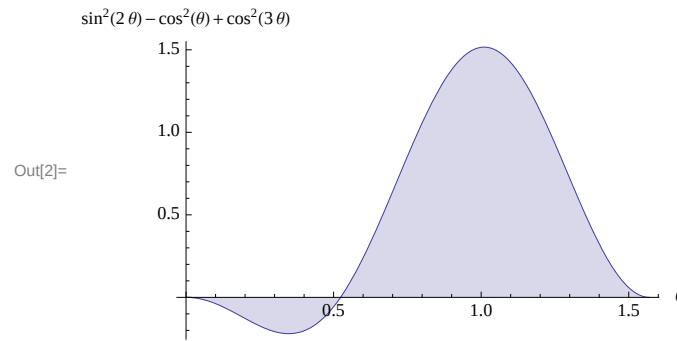


Figure 21.15: **Bell's Inequality**, Equation 21.13, which must be satisfied by any local theory of probabilistic transmission is not satisfied by quantum mechanics. When the appropriate amplitudes produced by a quantum mechanical system is used in the inequality it predicts that $\cos^2 3\theta + \sin^2 2\theta - \cos^2 \theta$ must always be greater than zero. As can be seen above, for angles less than about 0.5 radians it does not satisfy the inequality. Experiment agrees with the quantum mechanical results.

Chapter 22

Quantum Field Theory

22.1 Introduction

So far we have studied the properties of photons and electrons primarily as non-interacting single particles. The many particle states that we discuss are just collections of non-interacting single particle quantum systems. It was Einstein's great discovery to realize that these constituents of light are the substance of things we call light but again his detectors deal with only the single photon interactions. Granted, given Bell's Theorem and the Young's Double Slit experiment, these are entities that are very different from those that are encountered in a Newtonian world. At the same time, we realize that we have to develop a picture based on photons that adequately describes the many wavelike properties that are associated with light, the field properties. The theory that does that is quantum field theory. We want to make a quantum theory of the electromagnetic field which preserves its classical success and yet meets the requirements imposed by Planck and Einstein.

The electromagnetic field is a rather complex field; it is a combination of two vector fields with a rather complex dynamics. For this reason, we will first discuss a simpler field, the stretched string. It possesses the essential features of a field theory and, at the same time, we can understand the dynamical properties in sufficient detail to be able to construct a quantum theory. The problem though is that nature has chosen not to have any known system manifest this exact pattern of behavior. We will construct it by realizing that the phenomena that we identify with the field nature of light is characterized by energies that are large compared to $\hbar\omega$ and therefore states with many photons. This is consistent with our study of the quantum

oscillator, see Sections 19.9.2 and 19.11.2, which indicated that to recover classical motion, we needed states composed of several stationary states. These are the principle goals of this Chapter.

Our first problem will be to describe the many photon state. This is actually a subtle construction and will lead us in to a new definition of the identity of these particles. Actually, we are laboring to develop a formalism that simply leads to the strange counting that Planck originally discovered. Once we have a coherent description of light, we will add the fundamental charged particles, electrons, and review the theory called quantum electrodynamics. This is a complete theory of the world of photons and electrons and describe successfully all the phenomena that emerges in systems with only these constituents. It is the most successful theory ever developed. It agree with experiment to one part in 10^{15} , an incredible precision. This is the theory that is called Quantum Electrodynamics or QED. A detail look at this theory requires that we understand processes at a fundamental level. The most complete language for describing this theory is based on an analysis using Feynman diagrams. With the experience of using these diagrams, we can develop the current language for the description of all the fundamental processes that have as yet been observed. We will cover also one of the great theorems of modern physics, the spin statistics theorem.

22.2 The Many Photon State

Many things locally transfer different amounts of energy and momentum and other things. They do this locally in both space and time. The example that we have been dealing with is light and, from what we know from Einstein, the transfer, when we use monochromatic light, of some physical property is done discretely. For example, the energy is an integer multiple of $\hbar\omega$ where ω is the radian frequency of the light. Similarly for the momentum which comes in units of $\frac{\hbar}{\lambda}$ where λ is the classical wavelength. For the angular momentum the unit is $\hbar$.

The energy is related to the time evolution of the state and thus there is a frequency identified, $\omega = \frac{\epsilon}{\hbar}$. This frequency is related to the classical frequency. I remind you though that in the definite energy state of a quantum system nothing is moving back and forth.

From the classical relationships we know that there is a relationship between the energy and momentum, $\epsilon = |\vec{p}| c$.

The polarization of the light was known from the classical case to be related to the angular momentum of the light. The photon is said to have

an intrinsic angular momentum $L = \hbar$. In fact we can do experiments that measure the angular momentum transferred by the absorption of photons. I remind you that classically the polarization comes from the vector nature of the field.

Let us consider the case of light of only one frequency, ω , and therefore the photons have energy $\hbar\omega$, and thus also a momentum $p = \frac{\hbar}{\lambda}$, and some intrinsic angular momentum state. We want the states to be amplitudes and the multiphoton state comes from putting several photons in the state.

22.3 The Quantum Stretched String

In the case of the stretched string, we saw that the string can be describes as an infinity of independent oscillators, one for each of the normal modes. Each of these modes has a frequency of the normal mode, $\sqrt{\frac{T}{\rho}} \frac{\alpha\pi}{L}$, where T is the tension in the string, ρ is the mass per unit length, L is the length of the string, and α is an integer from 1 to ∞ and also labels the mode.

We saw that a quantum oscillator has definite energy states and that these have a definite frequency and the energy is an $(n + \frac{1}{2}) \hbar\omega$, where ω is the frequency of the oscillator.

The general state is thus a system in which for each mode there is a number of excitations, $\{n_i\}$.

$$| \rangle = | n_1, n_2, n_3, \dots, n_m, \dots \rangle \quad (22.1)$$

Each state like this will have a definite energy

$$\begin{aligned} E &= \left(n_1 + \frac{1}{2}\right) \hbar\omega_1 + \left(n_2 + \frac{1}{2}\right) \hbar\omega_2 + \left(n_3 + \frac{1}{2}\right) \hbar\omega_3 + \dots \\ &\quad + \left(n_m + \frac{1}{2}\right) \hbar\omega_m + \dots \\ &= \sum_{m=1}^{\infty} \left(n_m + \frac{1}{2}\right) \hbar\omega_m \end{aligned} \quad (22.2)$$

The state that has an ambiguity. All the states that have the excitations in different orders are the same. All excitations of the same mode are identical. This is a new definition of identical. Using this definition of identical you recover the magic counting that Planck needed to get the black body distribution

These states are orthogonal

22.4 The field

In the oscillator, we saw that the displacement of the mass required a superposition of many excitations. The state with a definite amplitude is not a state with a definite number of excitations. This is similar to the problem that we had with the states of polarization in light. These are incompatible measurements.

You can show that the expected value of the field in any definite energy state is zero. This is the same situation that we had in the oscillator and thus makes sense.

Also we have the same situation as in the oscillator. The definite energy state has a field² that is not zero.

22.5 Elementary Particles

These are the things that transfer discrete amounts of energy and momentum and other things. The example that we have been dealing with is the photon. It has a definite energy and momentum. It also has some intrinsic directional information as is shown by the polarization. The energy is related to the time evolution of the state and thus there is a frequency identified, $\omega = \frac{E}{\hbar}$. I remind you though that in the definite energy state nothing is moving back and forth.

In an empty space the thing that we have called the mode label is the momentum.

From the classical relationships we know that there is a relationship between the energy and momentum, $E = |\vec{p}| c$.

The polarization was known from the classical case to be related to the angular momentum of the light. The n is said to have an intrinsic angular momentum $L = \hbar$. I remind you that the polarization comes from the vector nature of the field. If we had a given polarization and then separated the different values of the polarization at a new angle θ relative to the original direction, the probabilities shifted to $n \cos^2 \theta$ etc..

The field takes on non-zero values when you have a carefully arranged state of many excitations. Systems with a classical field can have states with a large number of excitations in the same mode. These are things like ns and phonons.

An example of another particle is the electron. There is an associated field and the excitations are the particle. One difference is that the electron has mass. The mode label is the again the momentum. For a slowly moving

electron, we have $\epsilon = \frac{p^2}{2m}$.

The electron also has a polarization and there is even a device like the calcite that separates the states. The difference is that if you reorient the apparatus the probabilities go as $n \cos^2\left(\frac{\theta}{2}\right)$. Thus we say that the electron has an intrinsic angular momentum and the value is $L = \frac{\hbar}{2}$.

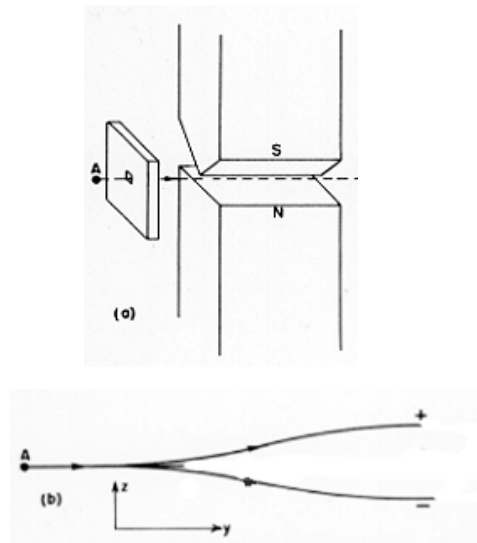


Figure 22.1: A diagrammatic representation of the Stern Gerlach apparatus. In the upper figure a beam of electrons passes from point A through an aperture and between the poles pieces of a magnet with an inhomogeneous field. In the lower part of the figure is how the beam is split into two beams one with spin up, labeled +, and the other with spin down, labeled -.

22.6 Fundamental Processes

Not only does the particle follow all possible paths, it undergoes all basic processes. This is best illustrated diagrammatically. Before going into the real processes relevant to electron and photons, it is worth examining some simple illustrative processes. Consider a drunk going home after an evening of partying. If he was a quantum drunk, his probability of getting home would be given by an amplitude from the party to home, $\Psi(\text{home, party})$, and his probability of being home after the party is $\psi^*(\text{home, party})\Psi(\text{home, party})$. But the evolution of his trip home is to stop at all distracting places on the

way. Thus the amplitude for getting home from the party is

$$\begin{aligned}\Psi(\text{home, party}) &= \Psi_0(\text{home, party}) + \Psi_0(\text{home, Alice's})g_0(\text{Alice's})\Psi_0(\text{Alice's, party}) \\ &\quad + \Psi_0(\text{home, Bob's})g_0(\text{Bob's})\Psi_0(\text{Bob's, party}) \dots \\ &\quad + \Psi_0(\text{home, Alice's})g_0(\text{Alice's})\Psi_0(\text{Alice's, Bob's})g_0(\text{Bob's})\Psi_0(\text{Bob's, party}) \dots \\ &\quad + \Psi_0(\text{home, Alice's})g_0(\text{Alice's})\Psi_0(\text{Alice's, Alice's})g_0(\text{Alice's})\Psi_0(\text{Alice's, party}) \dots \\ &\quad + \dots,\end{aligned}$$

where $\Psi_0(\text{anywhere, anywhere})$ is the fundamental amplitude for going from anywhere to anywhere and $g_0(\text{newplace})$ is the strength of getting into the place once there.

For instance, in the action for a charged particle there has to be a term that has both the particle terms, now a field for the electron and the electromagnetic field. This is because a charged particle is the source of an electromagnetic field and the electromagnetic field also produces changes in the motion of the electron. It is just a fact of life that all matter is made up of these fundamental constituents and these quantum properties are the basic operating procedures.

In the action formulation of physics, you have to introduce all effects through an action term. Therefore there is a generalization of the action to allow for an interaction. All interactions come from a term in the action. The neat thing about this is that it is a generalization of the old action reaction law.

$$\begin{aligned}\text{Action}_{\text{Total}} &= \text{Action}(\text{variables particle 1}) + \text{Action}(\text{variables particle 2}) \\ &\quad + \text{Action}(\text{variables particle 1, variables particle 2})\end{aligned}\quad (22.3)$$

For example, the electron and the electromagnetic field have to have a term in the action that looks like

$$S = -mc^2 \int_{t_1}^{t_2} d\tau - q \int_{t_1}^{t_2} [\phi(x, y, z, t) - \vec{v} \cdot \vec{A}(x, y, z, t)] dt \quad (22.4)$$

See Feynman lecture on action.

In the graphical language that we are developing there is a fundamental process in which an electron becomes an electron and a photon, see Figure 22.2. If you have that process you also have a photon and an electron turning into an electron and an electron and a positron, an anti-electron,

turning into a photon, see Figure 22.3, and a photon turning into an electron positron pair see Figure 22.4. I will return to this issue and the positrons when we have more of the material developed.

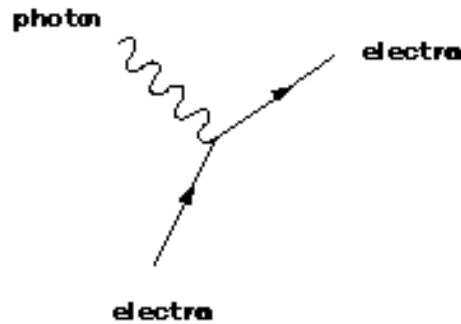


Figure 22.2: A space-time or Feynman diagram of the fundamental electromagnetic interaction. An electron enters at some time at the bottom of the figure. At a later time, it changes its velocity and emits a photon.

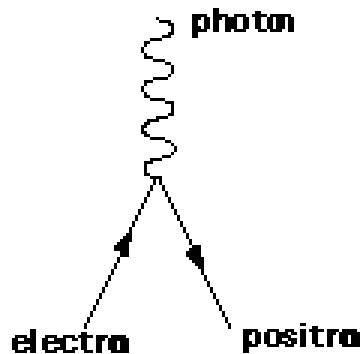


Figure 22.3: A space-time diagram depicting the annihilation of an electron positron pair into a photon.

None of these processes can occur and conserve energy momentum. They are virtual. We had the first virtual reality!

The basic point is that all fundamental processes occur locally, stochastically and instantaneously. In addition to following all paths, all processes occur also. Each of these enter through the action. All interactions have an effect on the action.

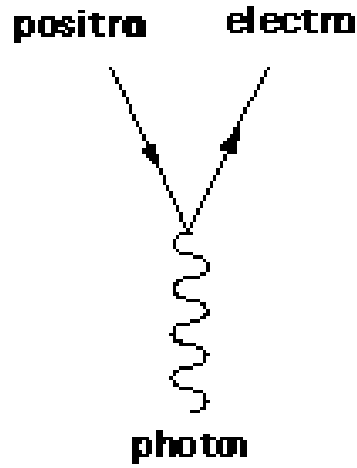


Figure 22.4: A space-time diagram depicting the process by which a photon is converted into an electron positron pair.

22.7 Feynman Diagrams

22.7.1 Introduction

These are the techniques that have been developed to compute the full evolution of quantum systems. The basic rules are the same as in our original discussion of the Einstein photo electric effect, 19.3 on page 423

Chapter 23

Interface of Large Scale and Micro-physics

23.1 Structure in the Universe

23.2 The Inflationary Universe

The dynamical equations are

$$\frac{\ddot{R}}{R} = -\frac{1}{6}(\rho + 3p) \quad (23.1)$$

$$\left(\frac{\dot{R}}{R}\right)^2 = \frac{1}{3}\rho - \frac{k}{R^2} \quad (23.2)$$

with k as usual the sign of the curvature.

Fill the space with a scalar field,

$$\mathcal{L} = \frac{1}{2}g^{\mu\nu}(\partial_\mu\phi)(\partial_\nu\phi) - V(\phi). \quad (23.3)$$

The Euler-Lagrange equation is

$$\square^2\phi + \frac{dV}{d\phi} = 0, \quad (23.4)$$

and the energy momentum tensor is

$$T_{\mu\nu} = (\partial_\mu\phi)(\partial_\nu\phi) - g_{\mu\nu}\left\{\frac{1}{2}(\partial_\sigma\phi)(\partial^\sigma\phi) - V(\phi)\right\}. \quad (23.5)$$

Comparing this to a Pascal fluid,

$$T_{\mu\nu} = (\rho + p) u_\mu u_\nu - p g_{\mu\nu} \quad (23.6)$$

where u_μ is the fluid four velocity field, the $T_{\mu\nu}$ in a local rest frame is

$$T_{\mu\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p & 0 \\ 0 & 0 & 0 & p \end{pmatrix}, \quad (23.7)$$

where the ρ and p have the usual interpretation of energy density and pressure respectively. Thus the energy density and pressure of the scalar field are

$$\begin{aligned} \rho_\phi &= \frac{1}{2} \dot{\phi}^2 + V(\phi) + \frac{1}{2} (\vec{\nabla} \phi)^2 \\ p_\phi &= \frac{1}{2} \dot{\phi}^2 - V(\phi) - \frac{1}{2} (\vec{\nabla} \phi)^2. \end{aligned} \quad (23.8)$$

In a spatially homogeneous universe, $\vec{\nabla} \phi \approx 0$ and, if it is also temporally slowly varying,

$$\rho_\phi = V(\phi) = -p. \quad (23.9)$$

The scalar field produces an effect that is the same as that of a cosmological constant term in the Einstein equation, Equation ???. This same observation can be seen directly from a comparison of Equation 23.7 with $\lambda g_{\mu\nu}$

It is consistent with the cosmological principal to use spatial homogeneity to require that $\phi(t)$ only. Thus the scalar field is dynamically the same as a point particle with a potential $V(\phi)$.

The equations of motion for ϕ follow from the hydrodynamics of the Pascal fluid,

$$\partial_\mu T^{\mu\nu} = 0. \quad (23.10)$$

This reduces to

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} = 0, \quad (23.11)$$

where H is the FRW Hubble constant $\frac{\dot{R}}{R}$. This is the classical mechanics problem of a point particle sliding down a potential hill with the “friction” set by H .

Assuming that at some time t , ρ is dominated by ϕ

$$\left(\frac{\dot{R}}{R} \right)^2 = H^2 = \frac{1}{3} \rho_\phi = \frac{1}{3} \left[\frac{1}{2} \dot{\phi}^2 + V(\phi) \right] \quad (23.12)$$

and thus we know H

To get $\ddot{R} > 0$, we need $\dot{\phi} < V(\phi)$, see Equation 23.8. Inflation in the “slow roll” approximation

23.3 String Theory

23.4 Quantum Gravity

Bibliography

- [Born & Wolf 1999] Max Born & Emil Wolf, *Principles of Optics* 7th Edition, Cambridge University Press, Cambridge, UK (1999)
- [Buchwald 1989] Jed Z. Buchwald, *The Rise of the Wave Theory of Light*, University of Chicago Press, Chicago, Ill. (1989).
- [Park 1997] David Park, *The Fire within the Eye: A Historical Essay on the Nature and Meaning of Light*, Princeton University Press, Princeton, NJ (1997).
- [Whittaker 1953] E. T. Whittaker, *A History of the Theories of Aether and Electricity, Vol. II* Humanities Press, New York, NY (Reprinted 1973)
- [Pais 1982] Abraham Pais, *Subtle is the Lord...The Science and Life of Albert Einstein*, Oxford University Press, New York, NY (1982)
- [Feynman 1985] R. P. Feynman, *QED: The Strange Theory of Light and Matter*, Princeton University Press, Princeton NJ (1985).
- [Rae 1994] Alastair Rae, *Quantum Physics: Illusion or Reality?*, Cambridge University Press, Cambridge England, (1994).
- [Moore 1998] Thomas A. Moore, *A Traveler's Guide to Spacetime: An Introduction to the Special Theory of Relativity*, Mc-Graw Hill, Inc., New York (1995).
- [Paulos 1988] John Allan Paulos, *Innumeracy*, Vintage Books, New York, NY (1988).
- [Newton 1730] Sir Isaac Newton, *Opticks: A Treatise of the Reflections, Refractions, Inflections, & Colours of Light* Dover Publications, Inc., New York, NY (1952)

- [Alder 2002] Ken Alder, *The Measure of All Things*, The Free Press, New York, NY (2002)
- [Giancoli] Giancoli's book
- [Pennycuick 1992] C. J. Pennycuick, *Newton Rules Biology: A Physical Approach to Biology Problems*, Oxford University Press, Walton Street, Oxford OX2 6DP (1992).
- [Devlin 1988] Keith Devlin, *Mathematics: The Golden Age*, Penguin Books, 375 Hudson Street, New York NY (1988).
- [Jammer 1966] Max Jammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill Book Company, New York NY (1966).
- [Popper 1973] Karl R. Popper, *Objective Knowledge: An Evolutionary Approach*, Oxford University Press, Ely House, London W. 1 (1973).
- [Will 1986] Clifford M. Will, *Was Einstein Right? Putting General Relativity to the Test*, Basis Books, New York, (1986).
- [Weinberg 1972] Steven Weinberg, *Gravitation and Cosmology: Principles of the General Theory*
- [Ronchi 1970] Vasco Ronchi, *The Nature of Light: An Historical Survey; [revised &] translated [from the Italian] by V. Barocas*, Harvard University Press, Cambridge Mass. (1970)
- [OED 1971] *The Compact Edition of the Oxford English Dictionary*, Oxford University Press, (1971).
- [Pound & Rebka 1959] R. V. Pound and G. A. Rebka, Phys Rev Letters, **3**, 439 (1959) and R. V. Pound and G. A. Rebka, Phys Rev Letters, **4**, 337 (1960).
- [Eötvös 1922, Dicke 1967] R. v. Eötvös, D. Pekár, and E. Fekete, Ann. Phys., **68**, 11 (1922) and P. G. Roll, R. Krotov, and R. H. Dicke, Ann. Phys. (NY), **26**, 442 (1967).
- [Weber 1973] R. L. Weber, *A Random Walk in Science*, The Institute of Physics, London England, (1973).
- [Popper 2002] K. Popper, *The Logic of Scientific Discovery*, Routledge Classics, London England, (2002)

- [EPR 1935] A. Einstein, B Podolsky, N. Rosen, Phys. Rev. **47**, 777 (1935).
- [Atkins 2007] P. Atkins, *Four Laws that Drive the Universe*, Oxford University Press, Oxford England, (2007)
- [Cohen,1958] B. Cohen and R. E. Schofield, *Isaac Newton Papers and Letters on Natural Philosophy and Related Documents*, Harvard University Press, (1958)
- [Wali 2009] K. C. Wali , *Satyendra Nath Bose: his life and times (selected works with commentary)*, Singapore: World Scientific, ISBN 981-279-070-5 (2009)