Project Physics 2

Travis Norsen

March 26, 2009

Preface

Let me try to explain what this book is, why I wrote it, and how I think it can and should be used. To begin with, it is a text designed to be used in the second semester of a more-or-less standard "freshman physics" course. It presupposes a familiarity with the standard topics covered in the first semester of a college physics course – kinematics, projectile motion, circular motion, Newtonian dynamics, momentum, and energy. The book uses calculus and is therefore probably best used in a "University" (calculus-based) as opposed to "College" (non-calculus-based) Physics course, at least for institutions large enough to have such a distinction. But I have used this material extensively and successfully at a small college where there is only one "freshman physics" course and many of the students have not had calculus.

Many of the topics covered in this text are standard ones for this kind of course – e.g., rotational kinematics and dynamics, gravitation, the physics of gases, and some thermodynamics. But this is not just another ordinary textbook, with its boring and (to the student) seemingly arbitrary progression from one chapter to the next. Instead, the standard topics have been tightly integrated into two broad historical arcs, covering the origins, development, and applications of two great theories of pre-20th century physics: Newton's Theory of Universal Gravitation and the Atomic Theory of Matter. The standard topics are therefore supplemented and integrated with significant additional material that might normally be found in a history of science book, but is (too) rarely seen in science textbooks.

The original purpose of this re-organization of the curriculum was simply to make the standard material more interesting, by placing it in a historical context and giving the course as a whole a sense of drama and mystery – one might say, a plot. I think it succeeds on this front. But the inclusion of historical material serves additional goals as well, most importantly the inculcation in students of a realistic understanding of science and scientific method. In virtually all other disciplines (in the humanities, arts, and social sciences) it is taken for granted that literacy in that field requires a firsthand knowledge of and appreciation for the important historical figures of that field. A proper education in philosophy, for example, simply requires that one has read Plato, Aristotle, Descartes, Hume, and Kant. What well-educated literature student has never read Shakespeare? There is a kind of irony (or perhaps tragedy) in the fact that the natural sciences – where contemporary work is most obviously and most hierarchically grounded on earlier historical discoveries – tend to educate students un-historically. This means that science students are largely asked to accept the claims they are taught without

understanding their historical origins – i.e., without understanding the *evidence* which makes it scientifically rational to accept those ideas. In other words, science – too often – is taught as a kind of dogma which students are asked to accept on faith from an authority (the textbook or the teacher). The tragedy is that the sciences are precisely the area where such an appeal to authority and faith is unnecessary. Actually, the real tragedy is that students taught this way will never fully appreciate the difference between science and the (irredeemably dogmatic) ideas that also vie for influence over their lives.

Research in science education in recent decades has revealed that traditional lecture classes tend to increase students' (often already-existing) sense that science is all about memorizing equations without thought or question or understanding – i.e., that science is dogma. We have thus seen an explosion in the use of more inquiry-based and project-based approaches to structuring classrooms and class time. We have not, however, seen similarly radical restructuring of the *content* of such courses, as manifested most obviously in textbooks.

This isn't to say that there haven't been some important improvements. For example, Priscilla Laws' *Workshop Physics* curriculum is specifically designed to allow a course in which a traditional textbook plays, at best, a secondary role. And other modern texts (e.g., *Understanding Physics* by Cummings et al.) make some valuable attempts to connect the material in the text with experiments and projects which may be performed by students in class. Also, the "Tutorials" pioneered by the PER group at the University of Washington have made significant headway in encouraging and allowing teachers to spend some class time (often what used to be spent on "problem sessions") having students work in small groups to confront and master challenging qualitative and conceptual issues.

Still, for the most part, even the (in this sense) best current textbooks have tables of contents that are virtually indistinguishable from those of many decades ago. So one of the things I am after in this book is to try a more radical re-structuring of (at least this part of) the curriculum. This is not merely a small tweak to the standard sequence of topics, pertaining only to how class time is spent or whether end-of-chapter exercises are phrased in the third or second person. It is rather the result of stepping as far back as possible from the standard curriculum and asking: what should students at this level actually be learning and doing? And perhaps more importantly: what should students know and be able to do before they go on to more advanced coursework in physics (or elsewhere)? And perhaps most importantly: what should students know and be able to do as preparation for life in the real world (whether as scientists or not)?

My answers, in outline, are as follows. Students should know something about the broad historical development of the important theories in physics, and this historical knowledge should be tightly integrated with their technical, mathematical knowledge of those same theories. That is, they should genuinely understand why it is scientifically rational to accept and use the theories – and they should know some of the practical benefits, whether to technology or to subsequent science, of doing so. Students should be spending time actually doing real physics – working with raw data, participating actively in important derivations, performing experiments, thinking creatively about how to set up challenging problems, and occasionally making (or reproducing) important

discoveries. Students should understand that plugging numbers into formulas is not physics. They should understand that science is fundamentally a way of knowing in which all claims are *ultimately* traceable back to empirical evidence. And they should begin to see the virtue in applying the methods of science to every aspect of their lives.

The historical structure of the text is, I think, an indispensable means to the achievement of these goals. This, however, is no history (or history of science) book. I am no historian, and (although I have made significant efforts to read all of the primary literature that is discussed in the text) the historical accounts found here are still heavily based on secondary sources and will no doubt suffer from many of the vices (such as Whiggishness and an unrealistic emphasis on the importance of theorists) often attributed to "bad history." My only defense against such charges is that this is not intended as history, but rather as *physics* – at a level that can and should be understood by every "freshman physics" student.

There is a more important sense in which this is not a history of science. History texts are typically written for humanities students, whose technical math skills are not strong. This is a book written for scientists and science students. Hence, our goal is not to teach history per se, but to teach the *technically rigorous* physics ideas (as they have evolved historically). Thus, instead of shying away from technical discussions and resorting to loose qualitative analogies, we tackle mathematically rigorous material head-on and expect students to do the same. If anything, the level of mathematical rigor in this text will be seen as higher than a standard freshman-level text. The equation-to-word ratio is probably a bit lower than in normal texts, but the fact that each equation and derivation plays some important role in the evolving plot raises the stakes: individual results cannot be taken (memorized, applied) in isolation, but must be fully digested as part of a coherent whole.

Let me say a few things about how I think this book should be used. It is designed to be used in – or more precisely, is designed to help bring about – a more "inquiry" based or Problem- (or Project-) Based-Learning (PBL) classroom environment. This has several aspects. First, the text itself is written in a way that students should be able to follow and understand. I see no need to spend valuable class-time having the teacher lecture on the material covered in the text (much as it would be regarded as preposterous for a literature professor to spend class-time reading Shakespeare to his students). Students should read the text outside of class – *before* class – thus freeing up class-time for more focused and purposeful Q&A periods, discussion of difficult concepts or derivations, working through derivations and discoveries that are merely sketched in the text, etc. I have included a number of "Questions for Thought and Discussion" at the end of each chapter to stimulate productive, open-ended discussions.

Each chapter also contains a number of end-of-chapter "Projects." These are typically very different from the end-of-chapter Exercises or Problems found in most texts. Since this book is designed for use in a physics course, a necessary pre-requisite for which is students' ability to do algebra, I simply take for granted that students can algebraically manipulate given equations and plug numbers into them. Those activities are appropriate for an algebra course, but they are not physics. The Projects in this book, by contrast, ask students to engage in real physics. For example, students are given raw data and

asked to interpret it from the perspective of various theoretical models. They are asked to invent their own theoretical models. They are asked to fill in steps that were missing from important derivations sketched in the text. They are asked to use computers to solve equations which are too complicated to solve analytically. And they are asked to participate in and reproduce important moments of discovery.

These Projects are generally somewhat open-ended. They almost always require some amount of creative integration, by the student, of different pieces that were covered in the text. And very often they are designed to put the student in the shoes of actual scientists working with actual data and making actual discoveries. Thus, not just the main text of the book, but also the end-of-chapter assigned work, is designed to convey to students a much more realistic picture of science and its methods. Traditional plug-and-chug exercises teach students that deep understanding and creative thought are not required in science – that doing science consists of blindly grabbing pre-formulated "magic formulas" and then going through a rote procedure. None of the Projects in this book can be done blindly – without creative thought, genuine understanding, and hard work.

Note that the open-endedness, complexity, and sheer difficulty of (many of) the Projects in this book may require a shift in perspective when it comes to assessment and grading. Good students who genuinely understand the material, work in earnest, and actually think about what they're doing, may nevertheless not get "the right answer" at the end of the day. (Indeed, in some cases, there is no clear right answer.) That's OK. The shift in focus urged by this text – from physics as out-of-context dogmas, to physics as fundamentally a *method* of knowing – must extend all the way to how student homeworks and exams are graded. What matters most is that students understand what they are doing and have a coherent, rational, scientific justification for their approach and their conclusions. One of the crucial lessons of the history of science is that people on the losing side of scientific debates – those who had the wrong answers, as judged by future hindsight – should not necessarily be judged fools, and may not even have done anything wrong. (Often it is only *later discoveries* which show, finally and conclusively, that a given claim is definitely false.) Having learned this important lesson from history, we should surely extend the same courtesy to our students in the present.

I've already mentioned that I like to spend (at least part of) one class period each week in a kind of open discussion format, typically using students' answers to the "Questions for Thought and Discussion" as a jumping-off point. I've also explained that I envision the book being used with a project-based classroom structure. What that means specifically is that significant class-time each week should probably be spent working through the end-of-chapter Projects.

Sometimes the Projects are similar enough to material that was covered in the text that students (perhaps working in small break-out groups) could be immediately set to the task of working through them. Others are perhaps best approached in a whole-class-demonstration format, with the instructor taking the lead in working through the Project (but with lots of active participation from students). Weekly homework assignments should probably consist of several additional Projects. (If there is a particularly challenging or complex Project that I want to assign as homework, I sometimes give

students time in class to start working on it – by themselves, or in groups, or with some assistance from me.) I also like to reserve a little bit of class time at the end of each week (or whenever the homework is due) for students to ask questions about the assigned Projects, to compare solutions with one another, etc.

Let me finally say something about the book's (still tentative) title. It was only long after embarking on this curricular project that I learned that, in some ways, I was re-inventing the wheel. Many before me have had the idea of spicing up science education by incorporating historical material. In physics, the best and best-known and most systematic attempt in this direction was Harvard's *Project Physics* (Cassidy, Holton, Rutherford, et al.) which began in the early 1960s (growing out of some earlier, related work, e.g., that leading to the splendid "Harvard Case Histories in Experimental Science") and produced textbooks and supplementary materials that stayed in print until at least 2002. I have benefited greatly from this work, and the current text owes much to it. But (despite its title), Project Physics was in no way project-based. Its scripted content was, I think, a significant improvement over standard texts, by virtue of its historical, inductive approach to the subject matter. But the work assigned to students was largely standard, plug-and-chug type exercises that gave students little opportunity to think creatively, use contemporary tools and develop contemporary skills, or participate in the discovery process in a genuinely first-hand way. Actually, the earlier editions of the curriculum (from the 1960s) are pretty good on this front. The more recent incarnations (with title *Understanding Physics*) are not only significantly dumbed-down relative to the original, they also lost the spirit of creative scientific inquiry that was, at the beginning, the whole point. (It is also sad and rather telling that the far more rigorous original text from the 1960s was intended for high school students, while the contemporary dumbed-down incarnation is intended for college students, even if not scientists and engineers.)

In any case, the (tentative) title is in some ways an homage to the original Project Physics course. It is particularly appropriate since one of my central motivations was to design a curriculum that left significant creative work to be done by students, in the form of the Projects. I also wanted a title that would make obvious that this is a book for use not in the first, but in the second, semester of "freshman physics." Hence: Project Physics 2.

Let me finally mention some of the other texts that I have learned from or leaned on in the preparation of this book. Malcolm Longair's *Theoretical Concepts in Phsyics*, which is intended for junior or senior physics majors, is perhaps the most similar book I've found to the current one and also one of the most inspirational to my project. In some ways, my whole project is a response to Longair's book: the topics he covers and the way he covers them are *so important and right* that they, I thought, should and must be done at the introductory level. The *Feynman Lectures on Physics* has also been inspirational to me... though what physicist wouldn't claim that? Feynman's legendary course was not organized historically and did not include open-ended Projects to be tackled by students in class and in homework. But I have attempted to imitate, as much as possible, Feynman's casual-yet-penetrating style, and the way he tightly integrates technical, conceptual, and historical material. The works of Stephen Brush, including especially

his (and Gerald Holton's) *Physics: The Human Adventure* and his *Statistical Physics and the Atomic Theory of Matter*, have been particularly helpful. Thomas Kuhn's classic tome on *The Copernican Revolution* first sparked my serious interest in the history of science and subsequently helped me navigate through some otherwise-impenetrable primary texts. Cohen's *Birth of a New Physics* was a particularly influential example of how to explain the discoveries of Galileo and Kepler at an appropriate level. The *Harvard Case Histories in Experimental Science* were probably the most useful secondary sources for the second half of the book. And lots of other things too...

Let me finally say that this is very much a work in progress at this point. And so your feedback will be greatly appreciated!

Contents

Part I

Newton's Theory of Universal Gravitation

Chapter 1

Greek Astronomy

Isaac Newton first published his theory of universal gravitation in his most important book, the *Principia Mathematica*, which came out in 1687. This book presented the entire structure of Newtonian mechanics (the three laws of motion, momentum, etc.) that was covered in the first half of your course. But it also extended the previously-familiar terrestrial notion of gravity out into the heavens and presented a detailed theory of how objects in the universe attract one another gravitationally. The primary evidence for this new theory was that it correctly accounted for a wide range of astronomical and terrestrial phenomena such as: the motion of the planets around the Sun, the motion of moons around planets, the motion of comets, the twice-daily rising and falling of the ocean tides on earth, and the fact that the earth's rotation axis changes its direction over a period of thousands of years. We'll discuss all of these effects as the semester progresses.

If we want to understand Newton's theory of universal gravitation, though, our first job must be to understand how the theory came about. And given the central role played in his thinking by the motion of the planets around the Sun, it makes sense to first try to understand how and why Copernicus proposed (way back in the first half of the 1500s) that the planets (including the earth) moved around the Sun. And in order to understand *that*, it makes sense to first try to understand how astronomers prior to Copernicus thought about what was going on in the heavens. So, for this first week, we'll need to do our best to forget everything we think we know about Newton and gravity and Copernicus and the fact that the planets go around the Sun, and start, as it were, from the beginning. One should think of this as an attempt to understand clearly the earliest seeds out of which Newton's theory of gravitation eventually grew.

1.1 Basic Observations

If you open your eyes and *just look* at how various things move, there are a couple of things that jump out at you. First, familiar terrestrial objects (like rocks and cars and people and trees) pretty much just sit there at rest unless something applies a force to make them move. A rock will just sit there on the ground – until or unless someone

comes along and kicks it, in which case it will move some distance but then come again to rest. And it's clearly the same with lots of other things. Even people seem to have stillness as a kind of natural state: you can of course move around, but it takes some effort or energy to do so – and eventually (when you die) you'll lose the ability to exert that needed effort and hence stay at rest permanently. To summarize, it seems that things on earth have *rest* as their natural state, and that the unnatural state of *moving* requires some kind of *force* to be exerted. A budding physicist might try to summarize this with something like the following law:

$$F \sim v \tag{1.1}$$

where F represents the applied force, and v represents the velocity of an object. The idea is: if no force acts on an object, its velocity will be zero, and the faster you want it to move, the harder you have to push. Of course, we (who know about $F = ma$) know this is wrong. But if we forget about that and try to just consider familiar sorts of observations in a naive, unbiased way, Equation 1.1 is at least not crazy.

Let's then consider things up in the heavens such as the Sun, Moon, and stars. How do these things move? Probably the motion of the Sun is most familiar to you: it rises every morning in the east, travels slowly in a big arc across the sky during the course of the day, and then sets in the west. It seems plausible to guess that the Sun doesn't go out of existence each night and get re-born again in the morning. Rather, it just keeps going in a big circle around the earth – we just can't see it at night.

You may or may not already know that the stars move in a very similar way. First of all, the stars move in some sense *as one* – that is, the stars all move in the same way such that their relative positions with respect to each other are the same, night after night after night. (This is why there are recognizable collections of them such as constellations.) If you pick some one star and follow its motion throughout the night, you'll find that it does pretty much what the Sun does: it will rise in the east, travel in a big arc across the sky during the night, and then set in the west. And, just as with the Sun, it seems pretty obvious that the stars are "out there" during the day, too: the ones you can see at night are not visible during the day because they're below the horizon, and the ones that are in principle "out" during the day are simply too dim to see when the Sun is also out.

Actually, it's a bit misleading to say that the stars move the same way the Sun moves. Lots of them do. But stars in the extreme southern part of the sky just barely come up over the horizon to the south – they don't so much rise in the east and set in the west, as rise *just east* of directly south, and then set *just west* of directly south – and they're only up for a short period of time around midnight. And then the stars in the north behave rather differently, too: they go in the same kind of big circle we talked about the Sun and other stars going in, but (unlike the Sun and other stars) their circular paths *never dip below the horizon.* See Figure 1.1.[1]

An important point here is that there is a particular star – the *north star* or Polaris – which doesn't move at all. And all the other stars move in circles *centered on* the

[1] Much of what we're saying here applies only for observers in the northern hemisphere.

Figure 1.1: A several-hour-long exposure taken from Hawaii, facing north, during the night. Note that the stars move in concentric circular paths – and hence leave circular "star trails" – centered at a certain point on the sky (the north celestial pole) that is very close to Polaris, the north star.

north star. The north star (assuming you live in the Northern Hemisphere – if you don't, you won't be able to see the North Star at all) is above the horizon to the north by an angle equal to your latitude. Hence, if you live down in the South, the north star will be $25 - 30^\circ$ above the horizon, while if you live up near the Canadian border it will be $45 - 50^\circ$ above the horizon. (If you lived at the North Pole, the north star would be directly up – 90° above the horizon!) Any star whose angle from the north star is less than the north star's angle above the horizon will never rise or set, but will always be "up." (Of course, you won't be able to see it during the daytime!) And those stars which are *farther* than this from the north star will rise and set like the Sun. And presumably there are even stars which are so far south that they *never* rise above the horizon. Indeed, if one were to travel south – to Mexico, say – there would be new stars visible in the southern part of the sky which weren't visible from Vermont.

We'll come back shortly to talk in more detail about the precise motion of the stars, Sun, Moon, and planets. But it is helpful at this stage to introduce the so-called "two sphere model" of the cosmos, to summarize what has been said so far. The model is just this: the Earth is a sphere, and then all the stars are (so to speak) painted on the interior of another big sphere which rotates around and around the earth. See Figure 1.2.

We'll develop this model in more detail shortly. For now, the main point is just this:

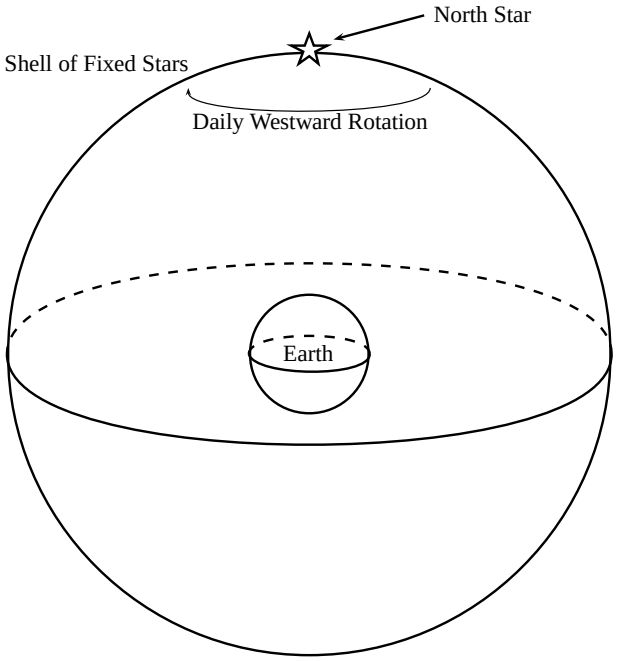

Figure 1.2: The two sphere model of the universe. The outer sphere is the shell of the fixed stars, which rotates westward once per day (or technically every 23 hours 56 minutes) around the Earth, which is at rest in the middle.

in contrast to terrestrial objects like rocks and people which seem to have *rest* as their natural state, the stars (and Sun and Moon) seem to incessently, "naturally" *move in circles*. This is a pretty fundamental difference, and it became, for the Ancient Greeks, the basis for the idea of a fundamental dichotomy between the earthly realm and the astronomical or heavenly realm.

It is worth briefly summarizing here what we might call Aristotle's Cosmology. This is just the basic worldview (including the heavens as part of the world) that was more or less accepted by Ancient Greek thinkers around the time of Aristotle (circa 350 BC) and was then orthodoxy until the Copernican revolution some 2,000 years later.

The whole cosmology is best illustrated with a picture: see Figure 1.3. The first thing to note is the radial lines and concentric circles, which are meant to be a kind of polar-coordinates "cosmic graph paper." The earth sits at its center, and the other (outer) sphere contains the stars. The Greeks had figured out that the Sun was closer to Earth than the stars, and that the Moon was closer to Earth than the Sun, so the Sun and the Moon occupy intermediate positions between the Earth and stars. Indeed, the sphere of the Moon functioned as a kind of dividing line between heaven and Earth.

The Greeks believed in four terrestrial elements: earth, air, fire, and water. One can think of these as basically standing for what we now think of as the three phases of matter (solid, liquid, and gas), plus fire, which didn't seem to fall under any of those

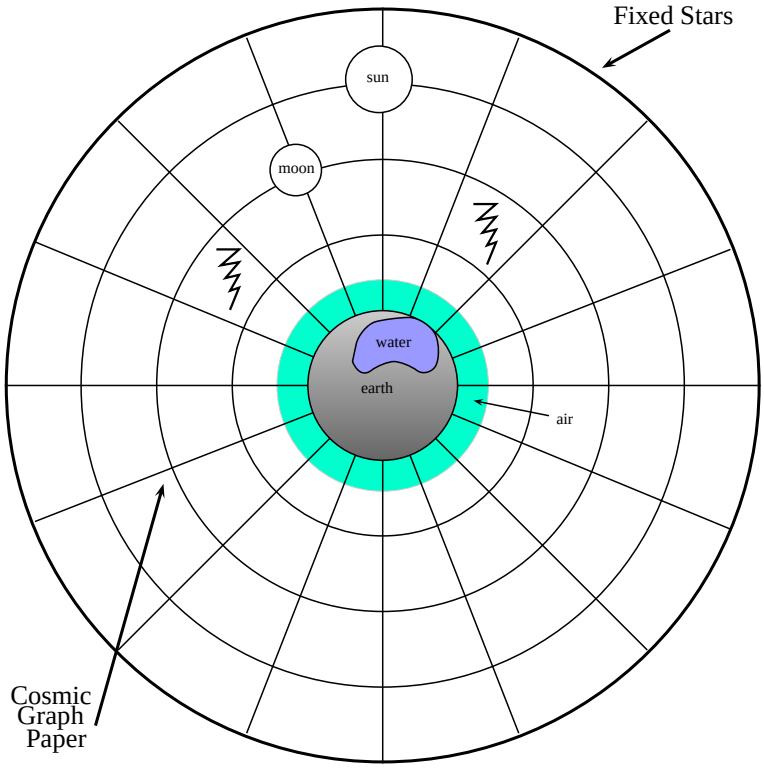

Figure 1.3: Simple sketch of Aristotle's universe. The matter (earth, water, air, fire, and aether) arranges itself relative to a fixed background space (the "cosmic graph paper") with the earth clumped up nearest the center, then water, then air, then fire, and finally (above the sphere of the Moon) the aether making up the Sun, stars, and other heavenly bodies (not shown).

categories. (Actually, sometimes people say that "plasma" is a fourth state of matter, and fire is indeed a kind of plasma.) We talked above about the concept of a "natural state of motion" for various kinds of things. The Greeks also inferred, from familiar sorts of observations, that in addition objects had natural *places*. For example: rocks sink in water, whereas air bubbles up through water to the surface. Water poured from a pitcher falls down toward the ground, whereas fire tends to move up (and pull other things with it). It is not crazy to infer that everything spontaneously "wants" to move down, but that there's a kind of hierarchy: earthy things like rocks are, so to speak, more desperate to "get down" than water, which is in turn more desperate to "get down" than air, which is in turn more desperate than fire.

Positing something like that as a basic physical law can explain why the Earth has the basic structure that it has. It is essentially a big ball of (what else?) earth – a ball being the most efficient way for as much as possible of the earthy material to get as close as possible to the center of the universe (where the radial lines of the "cosmic graph

paper" all converge). The water also likes to be as close as possible to the center of the universe (but gives way to earth) and so it tends to pool up in low spots on the surface – hence oceans, lakes, etc. And then the air also wants to be as close as possible to the center, but it gives way to both earth and water and so forms a fairly uniform blanket around the earth called the atmosphere. And evidently there should be some fire up in the outer regions of the atmosphere – e.g., lightning!

Of course, all these materials aren't in equilibrium. Things are, apparently, constantly getting churned up. Water somehow gets pulled up into the atmosphere for a while (in the form of clouds) before eventually falling back down to earth; there is evidently some fire trapped down here in things like trees, which gets released back up toward its natural place in the process we call burning; volcanoes occasionally spew earthy and watery and fiery junk up into the atmosphere where it lingers for a while before slowly moving back toward its natural place; etc. But as a rough account of why the materials should more or less arrange themselves (from the bottom up: earth, water, air, fire) as they appear to do, this all makes good sense.

As already mentioned, the idea is that the outer sphere with the stars on it just rotates around and around and around. The Greeks actually believed in a *fifth element* called "aether" – basically the stuff that stars and other heavenly bodies were supposed to be made of. And just as the natural, unforced state for the terrestrial elements is rest (namely, rest in as close to its "natural place" as it can get to, given the constraint that all the other hunks of stuff are also trying to achieve their natural places, too), the natural state for aether is circular motion. So that explains why the sphere of stars just turns around and around.

The other heavenly bodies (Sun, Moon, and planets) move in a rough way just like the stars move – around and around in circles. But their motion isn't quite as *perfectly* circular as that of the stars. They drift, slowly, relative to the stars. The Greeks thought of this as a sort of imperfection, and thought that, as one moved in from the outer sphere of stars, the heavenly bodies were increasingly corrupted with a little bit of the terrestrial elements (e.g., it does seem like maybe the Sun has a lot of fire in it since it's so warm and so bright), resulting in motion that is roughly circular (due to the aether) but imperfectly so (due to the fire, earth, etc.).

In Aristotle's cosmology, there is also an interesting dynamical connection between the heavens and the Earth. He believed that the incessant rotation of the outer sphere of stars was *mechanically* responsible for pretty much all the other motion in the universe. The rotation of the stars pulled the planets around, which in turn pulled the Sun around, which in turn pulled the Moon around, and then the Moon in turn communicated this motion down into the Earthly (sub-lunary) region. This also has a kind of plausibility to it: for example, the sloshing of ocean water we call tides correlates with the motion of the Moon and Sun, as do the daily warming and cooling of the Earth, and yearly progression through the seasons. So the "churning" of the Earthly regions of the cosmos discussed above is maintained, according to the Greeks, by the motion of objects up in the heavens.

This is all really interesting, but for us it's just background. So let's move along and discuss now in more detail the motion of the Sun, Moon, and planets.

1.2 Astronomical observations in more detail

Let's jump off from the two sphere model mentioned above, and see how the Sun, Moon, and planets can be incorporated in a way that is consistent with their actual motion. Note, to begin with, that in this model the sphere of stars rotates in about a day (which we can define as the amount of time between noon on one day and noon on the next day). But it is not *exactly* a day. If one measures carefully the amount of time between the rising of some bright star (Sirius, say) on two subsequent nights, it will be about 4 minutes less than 24 hours: 23 hours and 56 minutes. This amount of time is called a "sidereal day." So, in this model, the outer sphere of stars rotates around and around, always and uniformly toward the *west*, once every 23 hours and 56 minutes.

Now let's discuss, in turn, the Sun, Moon, and planets.

1.2.1 The Sun

The first thing to say about the Sun is that it moves with the stars. If one only watches for a few days and/or doesn't watch too carefully, one would probably think that the Sun is just sitting on some particular spot on the sphere of stars, and hence rotating around with it as it rotates. But more careful observation reveals that this isn't quite right. The Sun moves just a bit each day relative to the stars – in particular, it slides just about 4 minutes to the *east* each day. Here's what that means: compared to (say) the rising of Sirius, the Sun will rise (on average) 4 minutes *later* each day than it did the day before. Or one can turn it around the other way. Since, in fact, we measure time in terms of the Sun, we could describe the relative motion of the Sun and stars by saying that the stars are all a little bit further *west* each night, compared to the previous night at the same time. So if, say, Sirius was exactly to the south at midnight on some particular night, it would be just a bit to the west of south at midnight on the next night. And how far to the west exactly? Four minutes to the west – meaning, the distance that Sirius moves across the sky in 4 minutes.

Note here an amazing numerical coincidence:

$$4 \text{ minutes} \times 365 \approx 24 \text{ hours} \tag{1.2}$$

What does this mean? Well, if one prefers to think of the Sun as moving (each day) a little bit east relative to the stars, it means this: after an entire year, the accumulated slightly-late-rising of the Sun (relative to Sirius) will have it rising late by exactly a day – which means, it'll actually be rising in exactly the same *place* relative to the stars, and at the same sidereal *time* as at the beginning of the year. In effect, the Sun will have "gone all the way around" and come back to its original location in the stars.

This periodic motion of the Sun is what defines the year: astronomically speaking, the year is the amount of time it takes the Sun to slowly wander all the way around a big circle through the stars (through the 12 constellations of the zodiac, actually) and come back to where it started.

So, during a year, the Sun moves through a closed circular path through the stars. There is a technical name for this path: it is called the *ecliptic*. Thus, the Sun is always

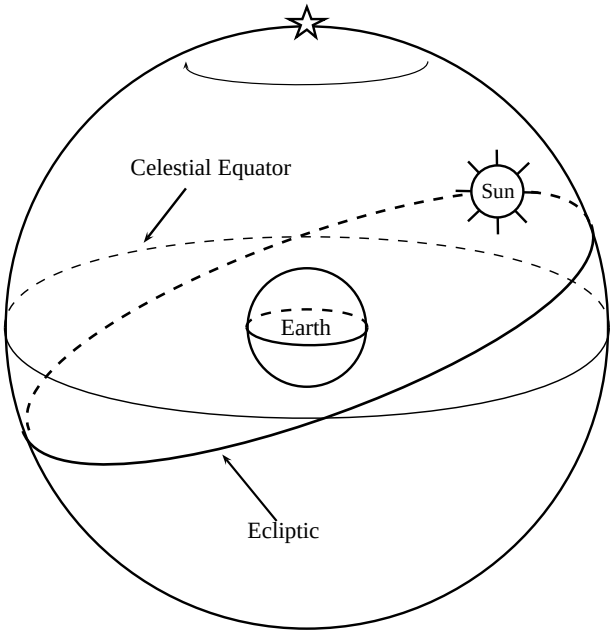

Figure 1.4: The two-sphere model of the universe, now including the Sun and the ecliptic (the Sun's path through the fixed stars). Note that the ecliptic and the celestial equator can be thought of as circles – specifically, intersections of the sphere of fixed stars with a plane passing through the center of the earth. The plane that cuts through the celestial equator and the plane that cuts through the ecliptic make approximately a 23.5° angle with one another. Thus, during the course of the year, the Sun is as much as 23.5° north of the celestial equator, and as much as 23.5° south of the celestial equator. It reaches these two extremes on the Summer and Winter Solstices, respectively, and goes back and forth between them in between, crossing the equator on the spring and fall equinoxes.

located somewhere on the ecliptic, but slowly moves to the east along the ecliptic during the course of a year.

It is useful to define another, related path through the stars called the "celestial equator." This is defined as all the points on the star-map that are exactly 90° away from the north star. If you are at the north pole of the earth, the north star will be straight above you and you'll just be able to see the stars on the celestial equator at the horizon. If you are at the earth's equator, the celestial equator will form a big arc across the sky, from directly east, to straight up above you, to directly west. If you're in the continental United States, the north star is about halfway between the horizon to the north and straight up, and the celestial equator forms an arc going from directly east, through a point about 45° up from the horizon to the south, and then over to directly west. I say all of this mostly to encourage you to get some kinesthetic feel for how the diagram in Figure 1.4 relates to the real world around you.

It is very important to understand that the ecliptic and the celestial equator aren't

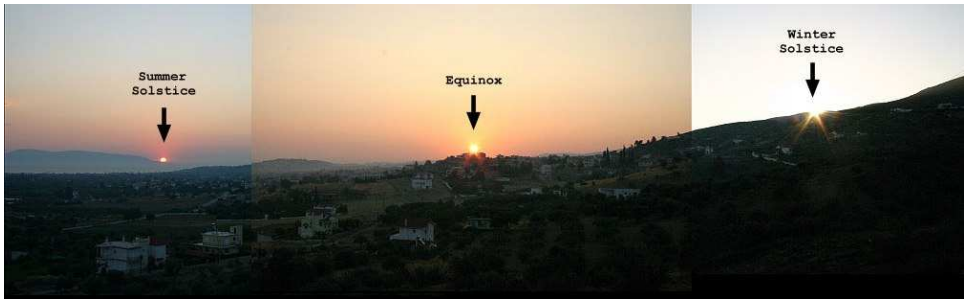

Figure 1.5: This is a composite of three photos taken at sunrise from the exact same location, at three different times during the year: the summer solstice, the winter solstice, and one of the equinoxes. The Sun is rising just to the east at the equinox, a bit north of east at the summer solstice, and a bit south of east at the winter solstice.

the same. If they were, then the Sun would basically follow the same path through the sky each day throughout the year. But they're not, so it doesn't. The ecliptic is "tilted" by about 23.5° relative to the equator. So one day each year, the Sun is a whole 23.5° south of the celestial equator – which means it will be very low on the horizon even at noon, and will rise and set considerably south of (respecitvely) east and west. This day is called the winter solstice. Since the arc of the Sun between rising and setting is considerably less than half a circle, the length of the daytime (the time between rising and setting) is considerably less than 12 hours.

There is then a corresponding day about 6 months later (the summer solstice) when the Sun is 23.5° further *north* than the celestial equator. On the summer solstice, the Sun rises and sets considerably north of east and west, and its arc takes it almost directly overhead at noontime. And since its arc is considerably greater than half a circle, the time between the rising and setting of the Sun is considerably greater than 12 hours.

During the rest of the year, the daily arc of the Sun slowly interpolates back and forth between these two extremes. The midpoints of this yearly cycle – when the Sun is at one of the two points where the ecliptic *crosses* the celestial equator – also have special names: the vernal (or spring) and autumnal (or fall) equinoxes. On these two days, the Sun rises and sets *exactly* to the east and west, and there is exactly half a day (12 hours) between Sunrise and Sunset.

To summarize: it is the motion of the Sun along the ecliptic which gives rise to the *seasons*. During the summer, the days are long and the Sun is close to directly overhead at noon, so it tends to be warm. And during the winter, the Sun is low on the horizon even around noon and the days are short, so it tends to be cold. The yearly progression of temperatures and weather (and all of the biological and ecological phenomena this cycling gives rise to) can be explained by the fact that the ecliptic is tilted relative to the celestial equator!

(Just for the record, some of what I've said here only applies to the middle latitudes of the northern hemisphere. The seasons work a bit differently if you go far enough south

or north. But I'll leave that for you to puzzle out.)

The take-home point here is pretty simple: the Sun basically moves each day the same way the stars do, but not exactly. We can abstract out the extra motion of the Sun by thinking about its motion *relative to the stars*. And this extra, relative motion is also pretty simple: the Sun moves around in a circle, the ecliptic, with a period of exactly one year.

1.2.2 Moon

In all but a few details, the motion of the Moon is exactly like the motion of the Sun. The Moon shares the (rough) daily motion of the Sun and stars, and also shares with the Sun a more subtle motion with respect to the stars: it, too, moves along the path through the stars called the ecliptic. The only difference is that the Moon moves eastward along the ecliptic *faster* than the Sun. Where the Sun takes a whole year to complete its circuit around the ecliptic, the Moon takes only about a month (27.3 days to be precise).

The Moon exhibits one other unique feature, too: phases. During its monthly circuit around the ecliptic, the Moon alternates between "full" (when the entire circular disk of the Moon is illuminated) and "new" (when the entire circular disk is dark). These phases are readily explainable by the assumption that the Moon's light is not intrinsic, but reflected light from the Sun. The Moon presents as full when it is just opposite the Sun in the sky, such that, from here on earth, it is precisely the bright side of the Moon that is visible. New Moon occurs when the Moon is very close to the Sun in the sky, meaning that the side of the Moon that is illuminated faces away from earth, with only the dark side being "visible" from here. And so forth for all of the intermediate (crescent, half, gibbous) phases. Note that this explanation requires that the Moon be closer to the earth than the Sun.

It is worth mentioning here another occurrence involving the Moon: eclipses. There are two types. A lunar eclipse happens when the Moon passes through the shadow cast by the earth and thus appears dark for a short period of time right around full Moon. (Do you see why a lunar eclipse can only happen at full Moon? It doesn't happen *every* full Moon because the Moon's path isn't *exactly* along the ecliptic – it is rather within a couple of degrees of the ecliptic, but this is enough that most of the time it doesn't pass directly through the earth's shadow.) The other type of eclipse is a solar eclipse. This occurs when the Moon gets right between the earth and the Sun, so that the view of some or all of the Sun is blocked. And, again, this doesn't happen *every* time there is a "new" Moon, because the Moon's path is only roughly along (within a couple of degrees of) the ecliptic.

1.2.3 Other Planets

The Moon and Sun have several things in common as against the stars. First, unlike the stars, they are not *fixed* in their positions relative to (other) stars. Rather, they move (more or less) slowly through the stars, along the ecliptic. And second, the Sun and Moon just look different than stars: stars look like little points of light, while the Sun and Moon both present a large disc.

Careful observation of the heavens, however, reveals several additional objects which *look* like stars (in the sense of the second point just mentioned) but which have the first point in common with the Sun and Moon. That is, these objects look like little points of light (though they are typically as bright as some of the brightest stars), yet their positions are not fixed. Like the Sun and Moon, they *wander*. There are five such "planets" (from the Greek word for "wanderer") that are visible to the naked eye and hence were known about by the Greeks: Mercury, Venus, Mars, Jupiter, and Saturn. Actually, not surprisingly, the Greeks tended to think of all *seven* of the wandering objects we've talked about as "planets."

Not only do these additional five planets, like the Sun and Moon, wander – they wander in much the same way. Each of them (in addition, of course, to sharing the daily rotation of the stars) moves with a roughly-steady eastward drift along the ecliptic. Mercury and Venus each move around the ecliptic in (on average) one year, just like the Sun. The other three planets take longer: about two years for Mars, about twelve years for Jupiter, and about thirty years for Saturn.

The Greeks basically just assumed that the correlation between distance-from-earth and ecliptic-period which held for the Moon and Sun, continued to hold for the other planets as well. So they inferred that the seven planets had distances from earth in the following ascending order:

- Moon
- Mercury, Venus, Sun (order ambiguous!)
- Mars
- Jupiter
- Saturn

Note that Mercury, Venus, and the Sun cannot be placed unambiguously on this list because they all take, on average, exactly one year to go around the ecliptic.

To understand why it is necessary to say "on average" we must clarify a further important detail about the observed motion of the planets. Whereas the Sun and Moon *always* move eastward along the ecliptic, the five planets only do this most of the time. They also occasionally stop their eastward drift, move for a short period of time *to the west*, and then stop and return to their normal eastward motion. This bizarre behavior is referred to as "retrograde" (backward) motion. Each planet retrogrades at regular, periodic intervals, but the period varies from planet to planet. Saturn does it once every 378 days, Jupiter does it every 398 days, Mars does it every 779 days, Venus every 584 days, and Mercury every 116 days. So there is no obvious correlation here between the (distance) order of the planet and their frequencies of retrograding.

Another curious feature is that the planets are not uniformly bright. A given planet (say, Mars) is sometimes brighter and sometimes dimmer than its average brightness, and (curiously) the planets Mars, Jupiter, and Saturn achieve their maximum brightness just as they retrograde.

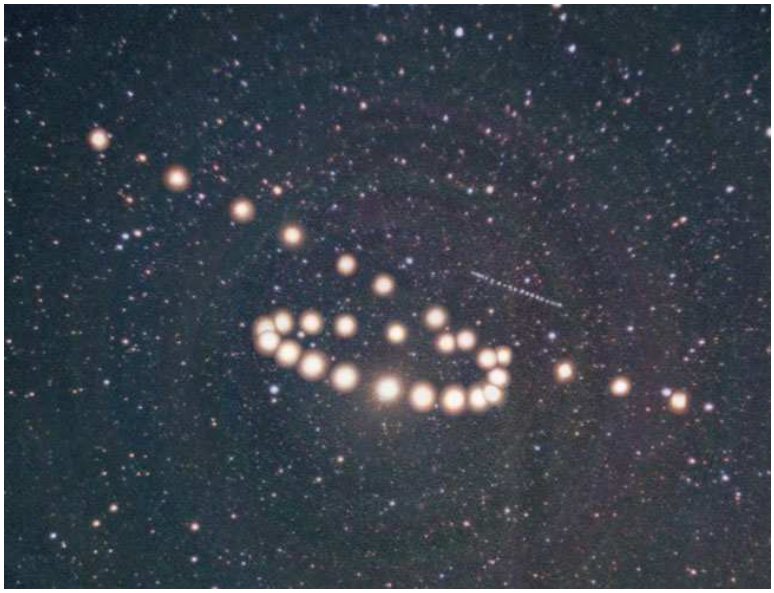

Figure 1.6: A sequence of images of Mars, stacked so that the stars line up in each frame. Mars begins on the right, moving left (to the east). But over the course of several days, Mars reverses direction, moving for a while westward, only to eventually return to is "normal" eastward drift along the ecliptic. Notice that the motion relative to the stars here is not *exactly* along the ecliptic – there is some motion in the orthogonal direction in the figure. The discussion in the text thus over-simplifies things a bit by ignoring this other aspect of the motion. One can also observe in the figure that Mars achieves maximum brightness during its retrograde motion. Finally, notice that there is some other planet in the background, which also leaves a trail through the stars.

It is a little harder to determine the brightnesses of Venus and Mercury, since both planets are always near the Sun in the sky. Think of this in terms of their motion along the ecliptic. Most of the time, Venus moves eastward along the ecliptic at a rate just a little faster than the rate of the Sun. But then, when it gets about 45° ahead of the Sun, Venus reverses direction and moves for a time *westward* along the ecliptic, until it is about 45° behind the Sun, at which point it resumes its eastward motion. It's important that the motion is "centered" on the Sun: the "normal" eastward motion of Venus along the ecliptic always has it catching up to and then overtaking the Sun, and then it passes it again in the other direction as it retrogrades, only to start over again (584 days later). Mercury does basically the same thing, only it goes back and forth faster, and doesn't get as far away from the Sun on either side.

One of our main projects for the week will be to examine some more detailed data about all of these things, and figure out how to incorporate it into a theory about how these things move. The point is: if you're a little confused and fuzzy (or just plain overwhelmed by the confusing complexity of these planetary motions), it's OK. We'll

spend the week trying to get clear on all this.

1.3 Measuring the distance to the Sun and Moon

It is rather amazing that the Ancient Greeks had figured out the sizes of and distances to the Sun and Moon. (It's amazing because it seems like, and is, a pretty sophisticated kind of discovery which, one might naively guess, people would only have figured out in the last couple hundred years. But it is also amazing in a kind of opposite way: it is interesting that one could discover such geometrical facts about the heavenly bodies when one was nevertheless so wrong about such fundamental facts as whether the Sun moved around the earth or vice versa!) For that reason alone, it's worth spending a few minutes to understand how they figured these things out. But the fact that (especially) the distance to the Sun was known, becomes an important part of the story here – as we'll see in due course.

So how did the Greeks figure this stuff out? Well, to begin with (and contrary to what one sometimes hears in the context of Christopher Columbus), the Greeks knew that the earth was round and they even knew how big it was. There are several pieces of evidence for the earth's round shape that they knew. Here is the Greek astronomer Ptolemy's summary:

> "the more we advance towards the north pole, the more the southern stars are hidden and the northern stars appear. So it is clear that here the curvature of the earth covering parts uniformly in oblique directions proves its spherical form on every side. [Also], whenever we sail towards mountains or any high places from whatever angle and in whatever direction, we see their bulk little by little increasing as if they were arising from the sea, whereas before they seemed submerged because of the curvature..." (I.4)

Another piece of evidence has to do with lunar eclipses: the earth always casts a *circular* shadow, and so must itself be shaped like a ball.

Anyway, once it is established that the earth is spherical, it becomes possible to measure the size of the sphere. This was first accomplished by the Greek astronomer Eratosthenes (273-192 BC). Assuming (not arbitrarily) that the Sun is far enough away that its rays can be treated as parallel when they reach the earth, Eratosthenes arranged the experiment illustrated in Figure 1.7. The city of Syene (now Aswan, Egypt) is on the Tropic of Cancer (23.5° north latitude), which means that each year on the summer solstice, the Sun at noon would be directly overhead – as evidenced by the observable fact that a vertical pole would cast no shadow at that moment. By contrast, 500 miles to the north, in Alexandria, a vertical pole *does* cast a shadow at noon on the summer solstice. It is easy enough to measure the height of the pole (h) and the length of the shadow (s), and hence determine the angle θ according to

$$\sin(\theta) = \frac{s}{h}. \tag{1.3}$$

Eratosthenes found that $\theta = 7.2° = 0.126$ radians.

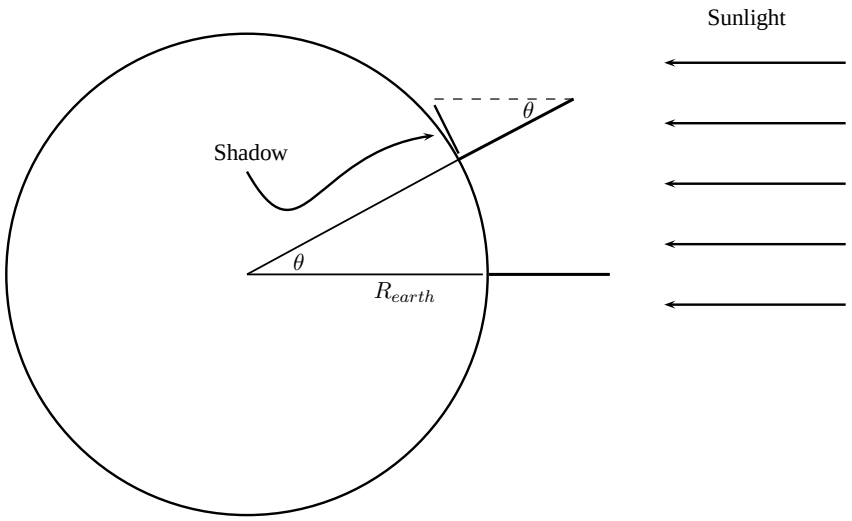

Figure 1.7: Schematic diagram of Eratosthenes' method of measuring the size of the earth. It is assumed that the Sun is far enough away that the incoming light rays can be treated as parallel. Then, one can determine the angle θ by measuring the height of a pole and the length of the shadow it casts (and using some trigonometry). And then, as is clear from the geometry in the figure, the angle θ is also the difference in latitude between the two locations. And so, since the distance (along the surface of the earth) between the two locations can also be measured, the radius of the earth R_{earth} can be calculated.

Also, as the geometry illustrated in the Figure makes clear, the angle θ which the Sun's rays make with the pole in Alexandria, is the same as the angle between Syene and Alexandria as measured from the center of the earth. And so the radius of the earth could be computed from what is essentially the definition of angle in radians:

$$\theta = \frac{D}{R_{earth}} \tag{1.4}$$

where $D = 500$ miles is the distance between Syene and Alexandria along the (curved) surface of the Earth). The result is

$$R_{earth} = \frac{D}{\theta} = 4,000\,\text{miles} = 6.4 \times 10^6\,\text{meters}. \tag{1.5}$$

Once the radius of the earth is known, the sizes and distances to the Moon and Sun can be inferred from a series of observations which relate them to the size of the earth. Let's go through these in turn.

First, the distance to the Moon can be determined by observing what is called the "parallax" of the Moon – i.e., its slightly different apparent position (relative to the background fixed stars) as seen from two different points on earth, or, equivalently, from the same point on earth at two different times (after the heavens have rotated a bit).

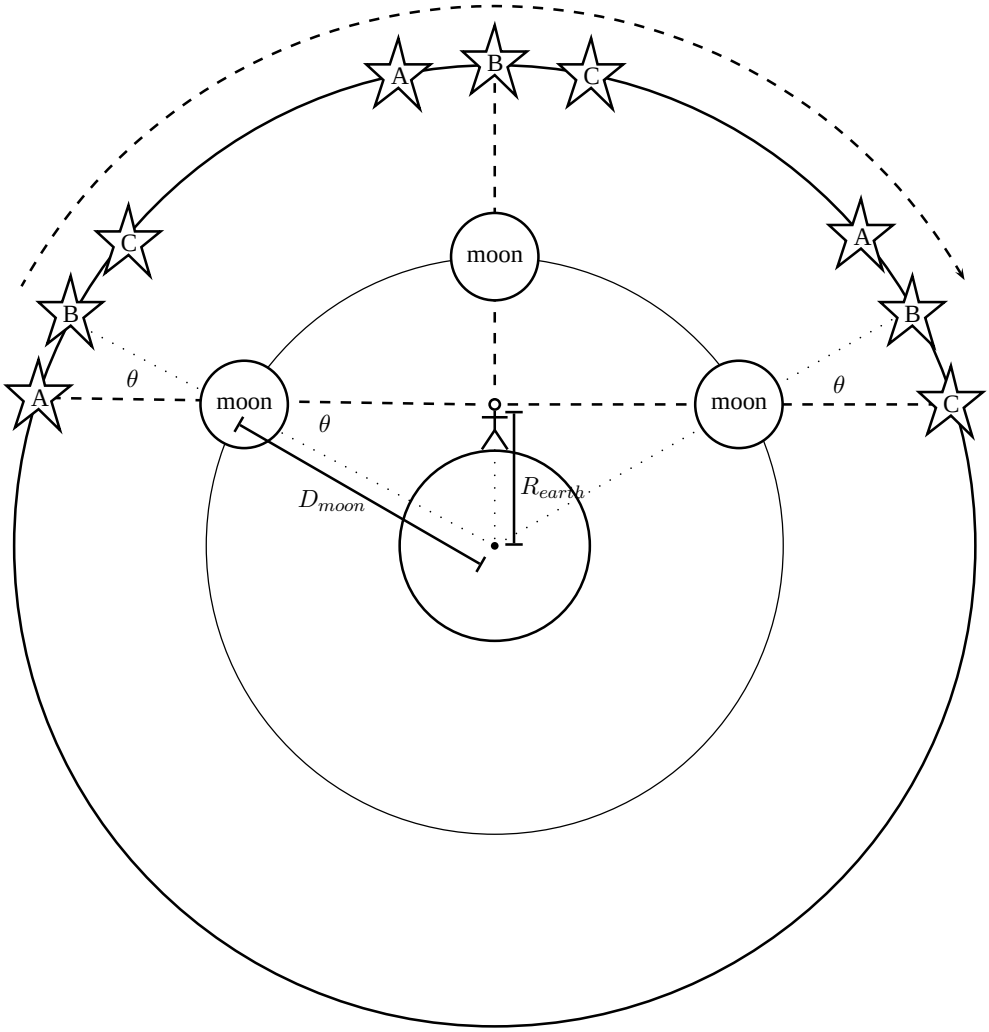

Figure 1.8: Due to the observer's location on the surface (as opposed to the center) of the earth, the Moon occupies three distinct apparent positions (relative to the fixed stars) at three times during the night. (Note that to highlight the phenomenon of parallax, the figure shows the Moon's "true position" – i.e., its apparent position as would be seen from a hypothetical observer at the center of the earth – as constant throughout the night, which in fact it isn't.) The parallax angle θ is the maximum deviation of the apparent position from the true position. And this measurable parallax angle relates the Earth's radius (R_{earth}) to the previously unknown distance to the Moon (D_{moon}) by simple trigonometry: $\sin(\theta) = R_{earth}/D_{moon}$.

Figure 1.8 shows schematically how the apparent position of the Moon will change over the course of the night due to the observer's position on the surface of, as opposed to in the center of, the earth.

Here is Ptolemy's description of the Moon's parallax:

> "since the distance from the earth's centre to the lunar sphere is not as that to the ecliptic circle which is so great that the magnitude of the earth is in the ratio of a point to it, therefore the straight line drawn from the Moon's centre to sections of the ecliptic, according to which the true courses of all the stars are conceived, necessarily does not everywhere sensibly coincide with the straight line according to which its apparent course is observed – that is, the straight line drawn from some part of the earth's surface or rather from the observer's eye to the Moon's centre. But when the Moon is directly above the observer, then only are the straight lines drawn from the earth's centre and the observer's eye to the Moon's center and the ecliptic one and the same straight line." (IV.1)

The parallax angle (θ in the figure) is the maximum deviation from the "true position" (i.e., what an observer at the center of the earth would see) due to the real observer's position on the surface of the earth. Careful observations revealed that the Moon displays a (maximum) parallax angle of just a little less than a degree: $\theta \approx 1^\circ$. The geometry of the figure – in particular the triangle formed by the Moon, the observer, and the center of the earth – then makes it clear that

$$\sin(\theta) = \frac{R_{earth}}{D_{moon}} \tag{1.6}$$

so it is possible to solve for

$$D_{moon} = \frac{R_{earth}}{\sin\theta} = 60R_{earth} = 240,000 \text{ miles} = 3.8 \times 10^8 \text{ meters} \tag{1.7}$$

You should be wondering: how is the parallax angle θ actually measured, since this is (as shown in the figure) an angle between two stars *as seen from the Moon.* The answer is: we are assuming here that (contrary to the not-to-scale figure!) the shell of fixed stars is very large compared to the circle representing the Moon's orbit. So the angle between stars A and B in the figure as seen from the Moon (i.e., the angle θ shown in the figure) will be the same as the angle between those stars as seen from here on earth. And that, of course, is readily measurable.

There is one other slight "cheat" here which is worth mentioning. In the figure, the Moon's "actual position" (namely: on top of Star B) doesn't change during the 12 hours or so between the three moments pictured. But in fact the Moon moves slowly and steadily relative to the fixed stars, along (roughly) the ecliptic. So you couldn't in fact measure the parallax of the Moon just by comparing the Moon's apparent position when it rises, and then later when it is overhead. You'd have to already have measured the Moon's *average* rate of motion along the ecliptic, so one could subtract off the part of the difference angle due to the motion of the Moon relative to the stars, leaving the parallax angle. One of the Projects at the end of the chapter steps you through this.

It is worth making a couple other comments about this method of determining the distance to the Moon.

First, the method relies on already knowing the size of the earth. In determining the size of the earth, we had to assume the Sun is very far away compared to the size of the earth. And then in measuring the Moon's parallax angle, we had to assume that the stars are very far away compared to the Moon. Is there any evidence to support these assumptions? Yes. First of all, the Moon displays an easily-noticable parallax of about one full degree. But no such parallax is observed in the Sun. (Well, it can be observed today, but the angle is so tiny the Greeks never noticed it.) So the Sun must be much further away than the Moon, and hence *very* far away compared to the size of the earth. And of course, since the shell of fixed stars is in this model the outside edge of the universe, the stars must be further away than the Sun! Or more directly: the stars also display no observable parallax. And so they, like the Sun, must be very far away compared to both the distance to the Moon or the size of the earth. Indeed, "that the Earth has the ratio of a point to the heavens" is one of the first points stressed in Ptolemy's book. He writes that "in all parts of the earth the sizes and angular distances of the stars at the same times appear everywhere equal and alike, for the observations of the same stars in the different latitudes are not found to differ in the least." (I.6)

Second, the parallax angle is in fact quite difficult to measure accurately, in part because the Moon is big and bright. A precise measurement of its position relative to the fixed stars requires some particular point on the Moon to be identified and tracked, and also requires that the background stars right next to the Moon can be seen – which is difficult because the Moon's brightness tends to overwhelm the surrounding stars. So it can be surprisingly tricky to measure the apparent position of the Moon, relative to the fixed stars, to an accuracy significantly better than a degree.

Given these sources of uncertainty, it is impressive that the parallax of the Moon can be measured at all. The implication is of course that the uncertainty on θ is pretty small compared to the reported value of θ. Actually, although Ptolemy reports a value for the parallax of the Moon and uses it to infer the distance to the Moon by essentially the argument described here, his description of the methods of determining the parallax angle (and hence that angle's uncertainty) leaves much to be desired by modern standards. In fact, he rather races through the argument presented here, and then proudly presents the reverse argument *from* the (now "known") distance to the Moon *to* the observable parallax angle. It is as if he is embarrassed to have to use observation to figure out how far away the Moon is, and so hurries through this, then lingering on what he considers more logically sound: using facts about the world to calculate appearances. We mention this here only to stress the very different approaches to science taken by the Ancient Greeks, as compared to modern *empirical* science – which of course is not at all embarrassed to base its conclusions about the world on observation.

So much for the determination of the distance to the Moon. Once this is known, it is relatively easy to determine the *size* of the Moon by measuring the angle ϕ subtended by the Moon, i.e., the angle between two opposite points on its edge, i.e., its angular diameter. The angle is easily measured to be about half a degree ≈ 0.0087 radians. Using the small angle approximation (according to which $\sin(x) = \tan(x) = x$), we then

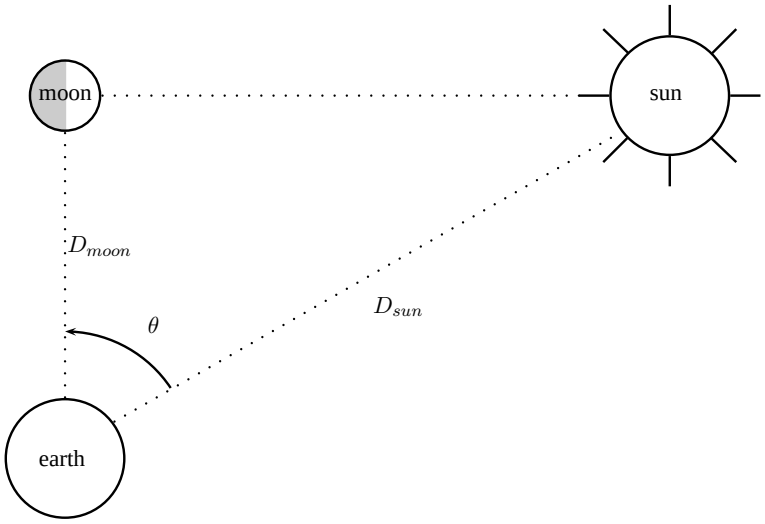

Figure 1.9: At half-moon, the earth, Moon, and Sun form a right triangle. Careful measurement of the angle between the Sun and Moon (θ in the figure) can therefore reveal via trigonometry the relative lengths of the sides of the triangle, i.e., the Sun's distance relative to the Moon's distance. This allows an absolute determination of the Sun's distance, since the Moon's distance is already known.

have

$$\phi = \frac{2R_{moon}}{D_{moon}} \tag{1.8}$$

where R_{moon} is the moon's radius. Plugging in the numbers,

$$R_{moon} = \frac{1}{2} D_{moon} \times 0.0087\,\text{radians} = 1,000\,\text{miles} = 1.7 \times 10^6\,\text{meters}. \tag{1.9}$$

So the moon is about a quarter the (linear) size of the earth. (Its volume is thus the earth's volume times a quarter *cubed*, i.e., it is about a sixty-fourth as big as the earth in that sense.)

Now what about the Sun? In principle, one could determine its distance, and then its size, by following the method just outlined for the Moon: first measure its parallax angle relative to the fixed stars, and then measure its angular diameter. But as already noted, the Sun turns out to be considerably farther away than the Moon, giving it a considerably smaller parallax angle that was simply too small for the Greeks to measure. (Plus, it's *really* hard to see which stars are right next to the edge of the Sun!) So a different approach is needed.

The simplest approach, first used by Aristarchus (310 - 230 BC), is to measure the angle between the Moon and the Sun at half-moon. In order for the Moon to appear precisely half-illuminated, the angle between the Moon-earth line and between the Moon-Sun line must be precisely 90°. And so the three bodies form a right triangle. See Figure

Figure 1.10: This is a composite of several photos shows the Moon passing in front of the Sun during a solar eclipse. Note in particular that when the Moon is just over the Sun ("totality"), its disc just covers the disc of the Sun (revealing the fuzzy solar corona, which is too dim compared to the ordinarily-blinding solar disc to see except during a solar eclipse).

1.9. It is clear that

$$\cos(\theta) = \frac{D_{moon}}{D_{sun}} \tag{1.10}$$

or, equivalently, that

$$D_{sun} = \frac{D_{moon}}{\cos(\theta)}. \tag{1.11}$$

Aristarchus reported that $\theta = 87°$, which then implies that $D_{sun} = 20D_{moon} = 1200R_{earth}$.

It is clear, though, that since the angle is very close to 90°, the calculated distance to the Sun will be very sensitive to the measured angle. For example, using instead a value of 88° gives a distance ratio of about 30 instead of 20. In fact, accurate contemporary measurements reveal that $\theta = 89.85°$, which gives

$$D_{sun} = 400D_{moon} = 24,000R_{earth} = 93,000,000\,\text{miles} = 1.5 \times 10^{11}\,\text{meters}. \tag{1.12}$$

And finally, once the distance to the Sun is known, its size can be calculated from its apparent (angular) diameter. Just as with the similar calculation for the Moon, we have that the angular diameter ϕ is

$$\phi = \frac{2R_{sun}}{D_{sun}}. \tag{1.13}$$

It turns out that the angular diameter of the Sun is just equal to that of the moon, as evidenced most dramatically during a solar eclipse. See Figure 1.10. Thus, plugging in the same $\phi = .0087$ radians used above yields

$$R_{sun} = 6.9 \times 10^8\,\text{meters} \tag{1.14}$$

if one uses the correct, modern value for D_{sun}. The Greeks underestimated this value by about a factor of 20, and hence also underestimated the size of the Sun by about a factor of 20.

The actual radius of the Sun is thus about 100 times the radius of the earth. The Greeks thought it was a mere 5 times bigger. Even with their underestimate for its size, then, the Greeks (correctly) believed that the Sun was more than a hundred times bigger (in volume terms) than the earth. (In fact it is about a million times bigger.)

1.4 Ptolemy's Theory

Claudius Ptolemy (85 - 165 AD) was the most important of the Greek Astronomers, partly because he systematized and cataloged many of the things that had been done earlier by Eudoxus, Aristarchus, Eratosthenes, and Hipparchus (whose writings have largely been lost). But Ptolemy also helped develop and improve the kinds of observations we were cataloging above, and he systematically developed a theory that had been posited earlier to to explain and integrate in particular the observed motions of the planets.

The Ptolemaic theory basically starts with the two-sphere model described previously, and incorporates the Sun, Moon, and the other planets in roughly the way we've already suggested. So, according to Ptolemy's theory, the earth is at rest at the center, with a big rotating sphere of fixed stars on the outside. The Sun, Moon, and planets are placed in the region between the earth and the stars, in the order we've already mentioned. To begin with, each of these seven planetary objects is pulled around (some way or other, either mechanically or just mathematically) by the rotating sphere of stars. This accounts for the shared gross daily motion of all the heavenly bodies.

1.4.1 Epicycles

Ptolemy's major innovation was a clever scheme for accounting for the details of the planets' motions – namely, their average eastward drift along the ecliptic, punctuated (for all but the Sun and Moon) by occasional retrograde motions. But to understand why this scheme was clever, one must first appreciate what Ptolemy was trying to do. Remember here the apparent ubiquity of *circular motion* for the heavenly bodies: not only do they all move in circles each day, but the extra motion of the planets relative to the stars is (at least on average) also circular. Plus, these objects were conceived to be made of a substance (aether, the fifth element) whose natural motion was circular motion. So the problem – the assignment, if you will, which is usually attributed to Plato – was to figure out a way of explaining the detailed, observed motions of the planets *in terms of circles*.

Here is Ptolemy's own statement of the guiding principle of his work:

> "it is first necessary to assume in general that the motions of the planets in the direction contrary to the movement of the heavens are all regular and circular by nature, like the movement of the universe in the other direction. [....] But

> the cause of this irregular appearance [i.e., deviations from the just-stated first assumption!] can be accounted for by as many as two primary simple hypotheses. For if their movement is considered with respect to a circle in the plane of the ecliptic concentric with the cosmos so that our eye is the cenre, then it is necessary to suppose that they make their regular movements either along circles not concentric with the cosmos, or along concentric cricles; not with these simply, but with other circles borne up on them called epicycles. For according to either hypothesis it will appear possible for the planets seemingly to pass, in equal periods of time, through unequal arcs of the ecliptic circle which is concentric with the cosmos." (III.3)

To summarize, the idea is to account for the observed motions of the planets – including especially the fact that they *don't* just move uniformly along the ecliptic circle relative to the stars – by compounding or otherwise fiddling with circular motions. Ptolemy here mentions two devices for achieving this. The first, called the "eccentric", involves using circles whose centers are shifted away from the earth. We will explore this later, in the Projects at the end of the chapter. The second and more immediately important device is called the "epicycle" and involves letting the planet move around a *smaller, second circle* which is itself pulled around a circular orbit centered at the earth. There is also a third device, called the "equant", which we will also come back to later.

For now let's focus on the epicycles and see how these are useful in accounting for retrograde motion. The idea, to repeat, is to *compound* two circular motions for each planet. The first would account for the average easterly drift along the ecliptic, while the second would account for the occasional retrograde motion. The way it works is sketched in Figure 1.11. It should be clear (looking at the figure) how this compounding of two circular motions (one circular motion relative to another point which is itself undergoing circular motion) can give rise to precisely the sort of behavior observed for the planets. In particular, by adjusting the relative sizes of the two circles and the two speeds involved for each planet, one can match pretty well the observed motions.

A few things are worth noting. First, since the Sun and Moon never retrograde, the secondary circle ("epicycle") is only needed for the 5 other planets: Mercury, Venus, Mars, Jupiter, and Saturn. And second, although this scheme works pretty well to explain the observed gross motions of the planets along the ecliptic, it doesn't quite get all the details exactly right. So in fact Ptolemy's full theory for each planet involved all three of the devices mentioned before: not just an epicycle, but also an eccentric and an equant.

See the Projects at the end of the chapter for some more information about these extra devices; they do play an important role in understanding what was and what wasn't immediately seductive about Copernicus' heliocentric theory, when he proposed that a millenium and a half later. But for the moment we'll ignore these other two devices and work with the simple one-deferent-one-epicycle construction.

The basic idea is that there is a special point (the "deferent") which moves in uniform circular motion around the ecliptic, and then a second point (actually occupied by the planet) which moves around a second, smaller circle (the "epicycle") with uniform

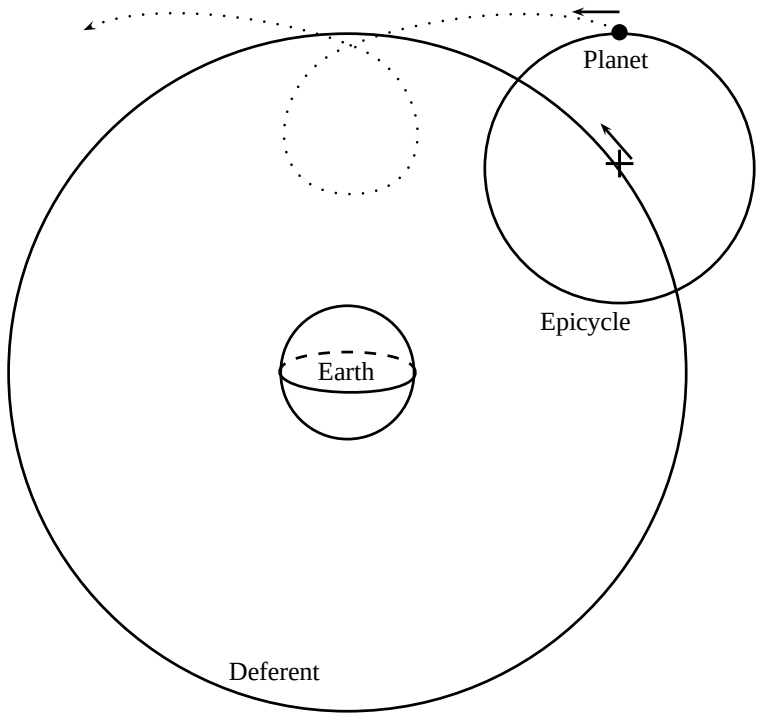

Figure 1.11: Sketch of the basic deferent-epicycle combination in Ptolemy's theory. The point marked + moves uniformly around the deferent circle, while the planet moves uniformly around the epicycle (which is centered at the + and pulled around the deferent as it moves). Both circles – the deferent and the epicycle – lie (approximately) in the plane of the ecliptic. This compounding of two circular motions gives rise to a trajectory like that sketched in the dotted line. As seen from the earth, the motion is generally counter-clockwise (which here means eastward along the ecliptic), but the occasional retrograde motion is also accounted for. Note too that the theory automatically explains the observed correlation between retrograde motion and brightness: the planet retrogrades when it is *closest* to the earth, which accounts for its increased apparent brightness.

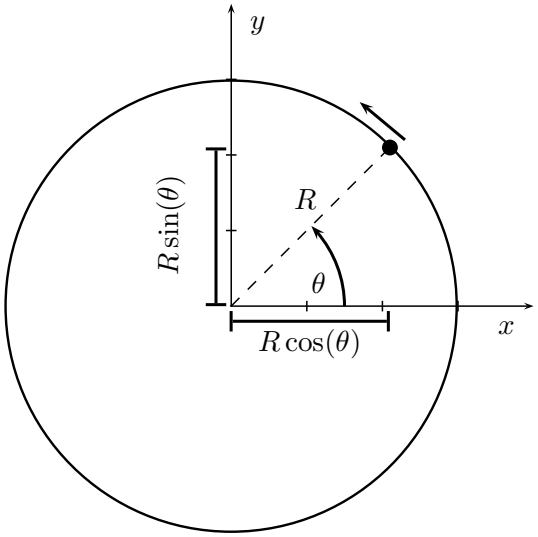

Figure 1.12: An object (the black dot) moves along a circle of radius R. Its position can thus be specified by the angle θ that the object makes with (say) the x-axis. The object's x and y coordinates are then given by $x = R\cos(\theta)$ and $y = R\sin(\theta)$. The object's angular velocity ω is given by the rate at which θ increases in time: $\omega = d\theta/dt$.

circular motion relative to the deferent point. It is worth developing some mathematical technology for dealing with all of this.

1.4.2 Angular Kinematics

Let's begin with simple uniform circular motion, described in terms of angle. For an object moving in a circle, its position can be described in terms of the angle as shown in Figure 1.12. For *uniform* circular motion the angle just increases linearly in time:

$$\theta(t) = \theta_0 + \omega t. \tag{1.15}$$

This should remind you of a corresponding equation from 1-D translational kinematics: $x(t) = x_0 + vt$. In accordance with this analogy, the quantity ω – which evidently describes the rate at which the angle θ increases – is called the *angular velocity*. Not surprisingly, it can be defined (generally, i.e., not necessarily assuming that the angular motion is uniform) as follows:

$$\omega = \frac{d\theta}{dt}. \tag{1.16}$$

This is just parallel to the familiar definition of (translational) velocity in terms of position: $v = dx/dt$.

For an object moving with constant angular velocity, its motion will be *periodic* – that is, it will repeat itself over and over again. Since a complete circuit comprises 2π radians and occurs in an amount of time we'll call T (the period of the motion), the

angular velocity for uniform angular motion can also be written: $\omega = 2\pi/T$. And this is equivalent to

$$T = \frac{2\pi}{\omega}. \tag{1.17}$$

Note that this last formula only makes sense if ω is constant in time.

Following the analogy with 1-D translational kinematics, we may also define the *angular acceleration*

$$\alpha = \frac{d\omega}{dt} = \frac{d^2\theta}{dt^2}. \tag{1.18}$$

as the rate of change of the angular velocity. Since the mathematics is all completely parallel, we can immediately steal some familiar results from 1-D kinematics. For example, if an object moves with *constant* angular acceleration α, its angular coordinate will evolve in time according to:

$$\theta(t) = \theta_0 + \omega_0 t + \frac{1}{2}\alpha t^2 \tag{1.19}$$

where θ_0 and ω_0 are, respectively, the angle and angular velocity at $t = 0$.

1.4.3 Angular and Rectangular Coordinates

Let's now consider how to relate the angular coordinates of a moving object to its rectangular coordinates. Assume (again) that the object is moving in uniform circular motion with angular velocity ω and initial angle θ_0. And let's choose the origin of our coordinate system to lie at the center of the circle the object moves along. Then simple trigonometry gives

$$x(t) = R\cos[\theta(t)] = R\cos[\omega t + \theta_0] \tag{1.20}$$

$$y(t) = R\sin[\theta(t)] = R\sin[\omega t + \theta_0] \tag{1.21}$$

where R is the radius of the circle.

Probably the simplest way to think about Ptolemy's scheme for explaining retrograde motion in terms of epicycles, is in terms of *vector addition*: the position of a given planet relative to the earth is found by adding two vectors – one representing the position of the deferent point, and the other representing the position of the planet relative to the deferent point. And, of course, if we want to add these two vectors to figure out where the planet is relative to earth, the easiest way to do that is by adding (separately) the x and y components of the two vectors.

Let's work through this in detail. Suppose the deferent for a given planet has radius R_d and moves with angular velocity ω_d. Then the x, y coordinates of the deferent (as functions of time) will be

$$x_d(t) = R_d\cos(\omega_d t + \theta_0^d) \tag{1.22}$$

$$y_d(t) = R_d\sin(\omega_d t + \theta_0^d). \tag{1.23}$$

Likewise, suppose the *epicycle* for this planet has radius R_e and angular velocity ω_e. Then the x, y coordinates of the planet (relative to the deferent point!) will be

$$x_e(t) = R_e \cos(\omega_e t + \theta_0^e) \quad (1.24)$$
$$y_e(t) = R_e \sin(\omega_e t + \theta_0^e). \quad (1.25)$$

And so the x, y coordinates of the planet *relative to the earth* will be given by the vector sum:

$$x(t) = x_d(t) + x_e(t) = R_d \cos(\omega_d t + \theta_0^d) + R_e \cos(\omega_e t + \theta_0^e) \quad (1.26)$$
$$y(t) = y_d(t) + y_e(t) = R_d \sin(\omega_d t + \theta_0^d) + R_e \sin(\omega_e t + \theta_0^e) \quad (1.27)$$

Of course, the Greeks only knew how to measure the *angle* of a planet along the ecliptic. And this angle is related to the x and y coordinates as follows:

$$\tan[\theta(t)] = \frac{y(t)}{x(t)} = \frac{(R_d/R_e)\sin(\omega_d t + \theta_0^d) + \sin(\omega_e t + \theta_0^e)}{(R_d/R_e)\cos(\omega_d t + \theta_0^d) + \cos(\omega_e t + \theta_0^e)} \quad (1.28)$$

So this is, in a way, the fundamental equation of the Ptolemaic theory. It connects the observational angle $\theta(t)$ – the angular coordinate of the planet on the ecliptic at time t – with the parameters in the theory (R_d, ω_d, θ_0^d, R_e, ω_e, and θ_0^e) which give rise to the planet's detailed motion.

Note that this basic equation has been written so that only the *ratio* R_d/R_e appears. This makes it clear that, from data about the ecliptic angle $\theta(t)$, one is never going to be able to find the absolute sizes of the deferent and epicycle – one can only find the relative size of the one circle relative to the other. This makes sense, since all the angles will be the same if both circles are (say) doubled in size. (Draw a picture if that isn't clear.)

If these parameters were known (for each planet), the above equation makes it clear that it would be possible to predict exactly how the observational ecliptic angle would vary in time. Of course, in reality, it works the other way: the problem is not to figure out how the planet will appear to move relative to the stars given the detailed parameters about its deferent and epicycle; rather, the problem is to figure out how big the epicycle and deferent radii and angular velocities need to be in order to account for the observed angular position, $\theta(t)$. We'll be spending quite a bit of time this week (in the Projects) figuring out how to do this. And the numbers that come out have some interesting surprises hidden in them.

1.5 The Precession of the Equinoxes

So far we have discussed the daily rotation of the heavens as a whole, and then also the motions (more or less along the ecliptic) of the Sun, Moon, and the other 5 planets. But there is one additional motion that was first discovered by Hipparchus. Here is Ptolemy's summary:

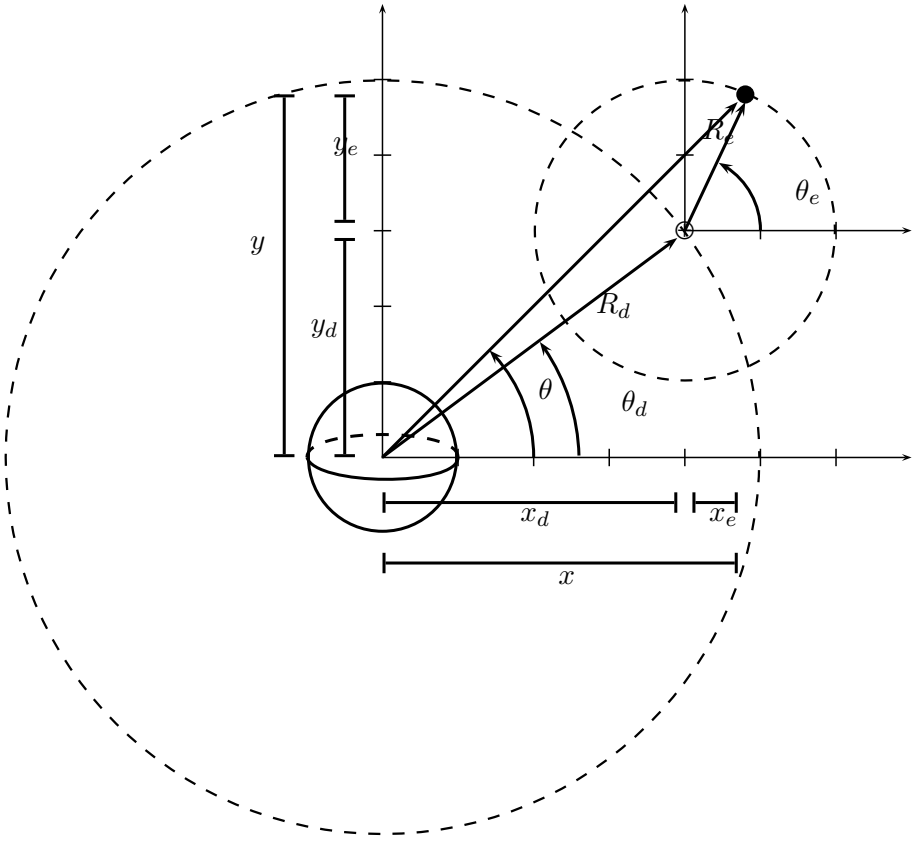

Figure 1.13: How the observable angle θ of a planet on the ecliptic relates to the radii (R_d and R_e) of the deferent and epicycle, and the angles θ_d and θ_e.

> "Hipparchus [noted that] for the apparent returns of the sun with respect to the tropics and equinoxes, the length of the year is found to be less than 365 1/4 days, but for its returns observed with respect to the fixed stars it is found to be more. And from that he conjectures that the sphere of the fixed stars also has a very slow movement, and like that of the planets is in the direction contrary to that of the prime movement which revolves the circle that passes through the poles of the equator and the ecliptic." (III.1)

Let us try to unpack what he is saying here.

There are two subtly different ways of defining a year, and they turn out not to be precisely equal. The first way – the so-called "sidereal year" – is defined as the amount of time it takes the Sun to complete a full circuit around the ecliptic and return to precisely the same spot relative to the fixed stars. The other way – the "solar year" – is defined as the amount of time between (say) two subsequent spring equinoxes. Based on what has been explained so far, one would expect these to be equal, since the ecliptic and the celestial equator have been described as *fixed* paths through the fixed stars, and the

equinoxes occur when the Sun is at the intersection of these two paths. So equivalent to the claim that the sidereal and solar years are not quite equal, is the claim that either the ecliptic or the celestial equator *moves*!

And that is just right. It turns out that the ecliptic really is fixed (in the sense that, even over long periods of time, the Sun always moves along the same path, i.e., in front of the same set of fixed stars). But the celestial equator *moves* (very slowly) through the fixed stars. And since the celestial equator is just all those points on the sky that are perpindicular to the *celestial pole*, a simpler way to understand this claim is that the celestial pole moves! We described this at the beginning of the chapter by saying that there is a particular star (the "north star" or Polaris) at (or, really, near) the celestial pole, i.e., the point that all the other stars move in circles around. So the point here is that this is true today, but was not true in the past, and will not be true in the future. Instead, the celestial pole is a moving target, that moves around (of course!) a *circle*, which is concentric with the ecliptic.

This slow motion of the celestial pole means that the sidereal and solar years will be slightly different: the Sun will be at a certain point on the ecliptic on the spring equinox one year, but the next spring equinox will occur *slightly before* the Sun returns to that same particular spot on the ecliptic in the following year. That is, the solar year is slightly shorter than the sidereal year. Or put the same point again this way: the location of the Sun on the ecliptic *at the moment of the spring equinox* slowly moves (westward) along the ecliptic. And this is also equivalent to saying that the sphere of fixed stars as a whole not only rotates once per day about an axis perpindicular to the plane made by the earth's equator, but that it also "wobbles" or "precesses" such that the intersection of this axis (passing through the geographic poles of the earth) with the fixed stars moves slowly to the east around a circle 23.5° down from the pole of the ecliptic.

How fast is this precessional motion?

> "having sighted Spica and the brightest stars about the ecliptic .... we find their distances with respect to each other very nearly the same as those observed by Hipparchus [i.e., the arrangement of the stars themselves has not changed in the several hundred years between Hipparchus and Ptolemy] but their distances with respect to the tropic and equinoctial points to have shifted eastward very nearly 2 1/2° compared to the record of Hipparchus." (VII.2)

This 2 1/2° of precession in the 2 1/2 centuries between Hipparchus and Ptolemy makes for an overall precession rate of approximately one degree per century. So this is indeed a small effect, noticable only with a long span of accurate measurements. With several additional centuries of observation to work with, the number has been more recently pegged at about 1.4° per century. Thus, in about 26,000 years, the north celestial pole will again be near the star Polaris (having drifted quite far – 47° – away from it over this period).

This subtle motion will play an interesting role in the story to come.

1.6 Arguments against heliocentrism

That pretty much wraps up what there is to say about Greek Astronomy and in particular Ptolemy's geocentric theory for the motion of the planets. It should come as no surprise that, in the next chapter, we are going to discuss the transition to the alternative heliocentric theory proposed in 1543 by Copernicus. But in understanding that transition, it will be helpful to know that Copernicus wasn't the first to propose a heliocentric theory. In all but some details, Copernicus' theory had been proposed already in Ancient Greece by Aristarchus, who is thus sometimes referred to as the Greek Copernicus!

An account of Aristarchus' proposal is given by another great Greek scientist, Archimedes, in his book *The Sand Reckoner*:

> "You King Gelon are aware the 'universe' is the name given by most astronomers to the sphere the centre of which is the centre of the earth, while its radius is equal to the straight line between the centre of the Sun and the centre of the earth. This is the common account as you have heard from astronomers. But Aristarchus has brought out a book consisting of certain hypotheses, wherein it appears, as a consequence of the assumptions made, that the universe is many times greater than the 'universe' just mentioned. His hypotheses are that the fixed stars and the Sun remain unmoved, that the earth revolves about the Sun on the circumference of a circle, the Sun lying in the middle of the orbit, and that the sphere of fixed stars, situated about the same centre as the Sun, is so great that the circle in which he supposes the earth to revolve bears such a proportion to the distance of the fixed stars as the centre of the sphere bears to its surface."

That is, in this heliocentric theory, it is not the sphere of fixed stars but the *earth* which (gasp!) rotates around once per day. Likewise, the yearly motion of the Sun around the ecliptic is accounted for by a yearly orbit of the earth around the Sun. It is sometimes speculated that Aristarchus proposed the heliocentric system not only because he saw it was an alternative (and arguably simpler) way of accounting for the observed apparent motion of the heavens, but also because he knew that (as discussed earlier) the Sun was significantly larger than the earth – and somehow it seemed more plausible for the smaller object to orbit around the bigger object, than vice versa.

Why does the heliocentric theory require the universe to be "many times greater" than implied by the geocentric theory? Recall that the only direct observational evidence bearing on the size of the sphere of fixed stars was the lack of observed parallax.

We discussed above how the parallax of the Moon could be used to measure the distance to the Moon. In principle, one could use the same effect to measure the distance to the stars, if only their parallax could be detected. Or, since no parallax was detected, this placed a lower limit on the distance to the stars. Assuming that, using the whole diameter of the earth as a baseline, the stars' parallax is less than (say) a hundredth of a degree (i.e., 0.000175 radians, which would definitely prevent it from being noticed by the Greeks), one can infer that the distance to the stars must satisfy

$$D_{stars} > \frac{2R_{earth}}{.000175\,\text{radians}} = 50,000,000\,\text{miles}. \tag{1.29}$$

Remember that (due to not-very-accurate measurement of the angle between Sun and Moon at half-moon) the Greeks thought it was only about $5,000,000$ miles to the Sun. So it was conceivable to them that the outer edge of the whole universe – the sphere of fixed stars – was only about one order of magnitude farther out than the Sun. This maybe seemed about right, as it left just about the right amount of room for the shells of Mars, Jupiter, and Saturn.

But if the earth orbits the Sun (instead of vice versa), the baseline for parallax increases dramatically! Instead of just being able to look at the stars from opposite sides of the earth, one can look at the stars from opposite sides of the earth's orbit – which the Greeks thought to have a diameter of some $10,000,000$ miles. And with *that* baseline, the condition on the size of the shell of fixed stars (required to account for their lack of observable parallax) becomes

$$D_{stars} > \frac{10,000,000\,\text{miles}}{.000175,\text{radians}} = 57,000,000,000\,\text{miles}. \tag{1.30}$$

This is really an incredible distance, about a thousand times bigger than the earlier geocentric estimate (because they thought the distance to the Sun was about a thousand earth radii). To the Greeks, it seemed impossible that the universe could be so vast, so much bigger than the biggest other things known. And hence it seemed to them impossible that the earth could orbit the Sun. So this was one major reason why, despite being proposed much earlier, the heliocentric worldview proposed by Copernicus was not taken seriously by the Greeks.

Another reason has to do with the several *motions* of the earth required by the heliocentric system. If the earth is spinning around on its axis once per day, then the surface of the earth near the equator must be moving at an incredibly large speed:

$$v = \frac{2\pi R_{earth}}{24\,\text{hours}} \approx 1,000\,\text{miles/hour}. \tag{1.31}$$

The Greeks thought: wouldn't we *notice* this? Shouldn't we have to hold on for dear life just to keep up with the earth as it moves? And, for example, wouldn't this imply a constant westerly wind of about this same speed? And if you dropped a rock, wouldn't it fall straight down (as rocks are known to do) *while the earth raced out from under it to the east* such that the rock would hit the ground a mile or more to the west of where it was dropped? As Ptolemy summarizes,

> "never would a cloud be seen to move toward the east nor anything else that flew or was thrown into the air. For the earth would always outstrip them in its eastward motion, so that all other bodies would seem to be left behind and to move towards the west." (I.7)

Yet none of these fantastic implications are in accordance with actual experience. So the earth cannot possibly be rotating in the way suggested by the geocentric model.

And note that all of these same objections can be made again – and with even more force – in relation to the alleged yearly motion of the earth around the Sun, which

evidently requires the earth as a whole to be moving with a speed

$$v = \frac{2\pi D_{sun}}{1\,\text{year}} \approx 20\,\text{miles/second}. \tag{1.32}$$

Thus, for example, a ball thrown high up in the air such that it stays up for 5 seconds should evidently land 100 miles away. Or more precisely: *we'd* be 100 miles away when the ball came back down to the same place it had been thrown from. And so forth.

Another argument against the heliocentric system makes (even more explicit) reference to the Aristotelian physics ideas we began with: even if the earth were somehow displaced from the center of the universe, it would simply return to the center. That, after all, is what the laws of physics sketched in the beginning of this chapter require. In order to perform the sorts of circular motions attributed to it in the heliocentric model, the earth would have to be made of aether, not earth – which is clearly preposterous.

It is easy to laugh at this sort of objection. It's maybe less easy to take it seriously and answer it clearly. So it's something to puzzle over for yourself before we cover the arguments of Copernicus and his followers in the following chapter.

Questions for Thought and Discussion:

1. What other observations can you think of that are consistent with – and might have been taken as evidence for – the Greek cosmology sketched in Figure 1.3? Can you think of any observations that are definitely contrary to this picture?

2. Orient yourself spatially, i.e., figure out which way is north. Now indicate with a sweep of the arm the path that the Sun will take across the sky today. Now indicate (again with your arm) how this path changes over the course of the year.

3. Describe how the daily trajectory of the Sun across the sky would vary over the course of the year if you were in Ecuador (which, of course, is on the equator). How about at the north pole? What is the significance of the Tropic of Cancer and the Arctic Circle (in the northern hemisphere, and correspondingly the Tropic of Capricorn and Antarctic Circle in the southern)? What are the latitudes of these "special" points, and why are those numbers significant?

4. Suppose it is the Winter Solstice, so the Sun's arc across the sky during the day is low and short. What can you say about the path taken by the Moon across the sky – and/or about the duration of time between the Moon's rising and setting – *that night*? What if in addition you are told that it happens to be full Moon? In general, how does the path/duration of the Moon – around full Moon – relate to the path/duration of the Sun, over the course of the year?

5. Everybody knows that it's hotter in the summer because the Earth's orbit takes it closer to the Sun during those months. Which Ptolemaic corrective device allows for this? What about the fact that it's winter in the southern hemisphere when it's summer in the northern? How is *that* explained?

6. It was mentioned in passing in the text that the Moon can only exhibit the full range of observed phases (new, crescent, half, gibbous, full) if the Moon is closer to the earth than the Sun. Explain why. For example, suppose to the contrary that the Moon and Sun both orbited earth in circles, but the Sun's orbit was smaller/closer than the Moon's. What range of lunar phases would then be observed?

7. It was mentioned in passing in the text that, unlike the Moon, the stars do not exhibit any parallax (that would have been noticable to the Greeks). Explain, by reference to Figure 1.8 or some diagram of your own, what it would mean for the stars to display parallax. (Parallax for the Moon or Sun means a change in the apparent position of that body *relative to the background fixed stars*. How could the fixed stars appear to move relative to themselves?)

8. You should understand how the deferent-epicycle device sketched in Figure 1.11 can account for the retrograde motion of a planet. But look more carefully at Figure 1.6. What aspects of the actual motion of Mars *cannot* be accounted for by the deferent-epicycle device as shown in Figure 1.11? How could this be fixed in Ptolemy's theory?

9. As you probably know, lunar and solar eclipses are somewhat rare. They certainly don't happen as often as every month. What does this imply about the deferent circles for the Sun and Moon in Ptolemy's system? In particular: does the deferent circle for the Sun lie in the plane of the ecliptic? Does the Moon's?

10. Can observers at different locations on earth disagree about whether a lunar eclipse is "total" (i.e., whether all of the Moon enters the earth's shadow cone)? Can observers at different locations disagree about whether a solar eclipse is total (i.e., whether the Sun is completely covered by the Moon)?

11. Argue using vector addition that the apparent motion of a planet is the same if (a) it moves on an epicycle of radius R_e and angular velocity $\omega_e = 0$, and (b) it moves around the deferent circle directly, but with the center of the deferent circle displaced by a distance R_e from the position of the earth.

12. In the Ptolemaic theory discussed in this chapter, the moon, Sun, and 5 other planets all orbit around the earth. We discussed explicitly how the distances to the Moon and Sun could be measured. What about the distances to the other planets? Could the Greeks have measured these? Why or why not? What other sorts of arguments could be used (within the Ptolemaic system) to estimate or put bounds on the size of the universe?

Projects:

1.1 Reproduce Eratosthenes' measurement of the size of the earth, perhaps by coordinating with another class in a different part of the country.

1.2 Your teacher will provide some data for the apparent angular position of the moon, along the ecliptic, over the course of several days. Do a linear curve-fit to find the average angular velocity of the moon. Is it what you expect? Look at the residuals of your fit and explain qualitatively what gives rise to the obvious feature. (The data include one unrealistic feature, which is that exact positions for the Moon are given throughout the course of several days, even when the Moon is below the horizon!) The data were taken from near the Equator when the Moon's position in the stars was close to one of the Equinox points (i.e., when the moon was at the same location relative to the stars that the Sun occupies at one of the Equinoxes). What is the distance to the Moon?

1.3 Here is another way to determine the relative distances to the Sun and moon. (Ptolemy discusses the method and credits it to Aristarchus.) During a lunar eclipse, the Moon passes through the shadow cast by the earth. Because the Sun is bigger than the Earth, the region of complete shadow "behind" the Earth is shaped not like a cylinder, but like a *cone*. From a knowledge of the size and distance of the Moon – and from observing the size of the Earth's shadow at the position of the Moon during an eclipse – the "slant" of the cone can be determined, and one can hence work out the distance to the Sun.

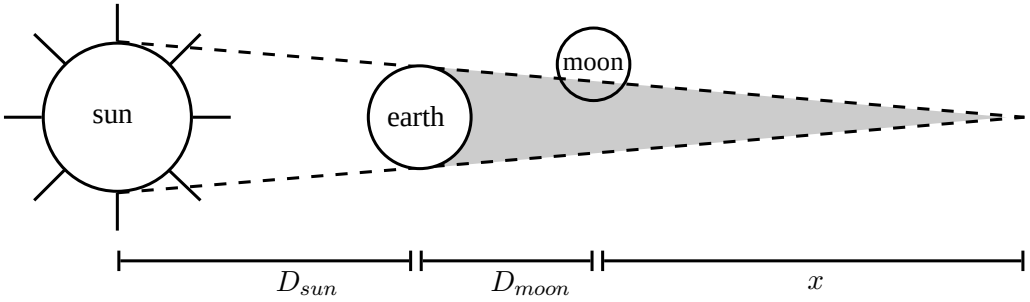

Figure 1.14: During a lunar eclipse, the moon passes into the conical shadow cast behind the earth.

Figure 1.14 shows the relevant geometry, and Figure 1.15 shows the relevant kind of observation. Use Figure 1.15 to estimate the relative size of the shadow compared to the Moon (e.g., is the shadow twice as wide as the moon, or 2.5 times as wide, or what?). Be as precise as you can. Then use the already-known relationship between the Moon's radius and the Earth's radius to determine how the size of the shadow (at the Moon's distance) compares to the size of the Earth. Knowing also already the distance to the moon allows one to then calculate the angle that the edge of the shadow slants at in the Figure. And this now known slant angle can then be related to the size and distance to the Sun. Finally, combine this relation with Equation 1.13 to yield an expression for the distance to the Sun. There are some tricky aspects to this, which should warrant further thought and discussion. In particular, it is worth considering whether this method is more or less reliable

than the method described in the main text.

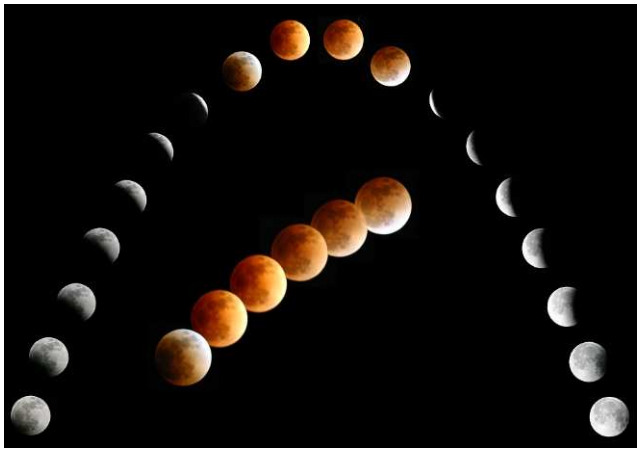

Figure 1.15: A mosaic of pictures showing the moon passing through the earth's shadow during a lunar eclipse. By looking at one of the shots of the moon where it is approximately half-shadowed, it is clear that (a) the shadow has a round edge (helping establish the round shape of the earth), and (b) that the size of the shadow is bigger than – one might guess roughly twice as big as – the size of the moon itself. Photo by Anthony Ayiomamitis, http://antwrp.gsfc.nasa.gov/apod/ap040506.html

1.4 Your teacher will provide some data for the angular position (along the ecliptic) of the Sun, over the course of several years. Use Excel to make a graph of the angle vs time. Then use Excel to compute the angular velocity vs time. What is the average value of the Sun's angular velocity? What is the period of its motion (relative to the stars)? Are these numbers related the way they should be? Does the Sun move with constant angular velocity? Describe qualitatively any deviations from constancy.

1.5 Your teacher will provide some data for the angular position (along the ecliptic) of a planet such as Mars, over the course of several years. Use Excel to make a graph of the angle vs time. Then use Excel to compute and graph the angular velocity vs time, and the angular acceleration vs time. Be able to explain in words any interesting features of the graphs, and how the graphs relate to one another.

1.6 Putting the Sun in a circular orbit around the earth with constant angular velocity accounts for its motion pretty well. But it is not exact. You probably noticed in Project 1.4 that the angular velocity is not exactly constant, but is rather sometimes a bit faster and sometimes a bit slower than usual. One way to account for this behavior in Ptolemy's system is by having the Sun move in a circular orbit with uniform angular velocity relative to the center of the circle – but displacing the center of that circle somewhat from the earth. This construction is referred to

as an "eccentric" circle, and was in fact used by Ptolemy not only for the Sun, but also for the moon and the other 5 planets. Let's see how it works for the Sun.

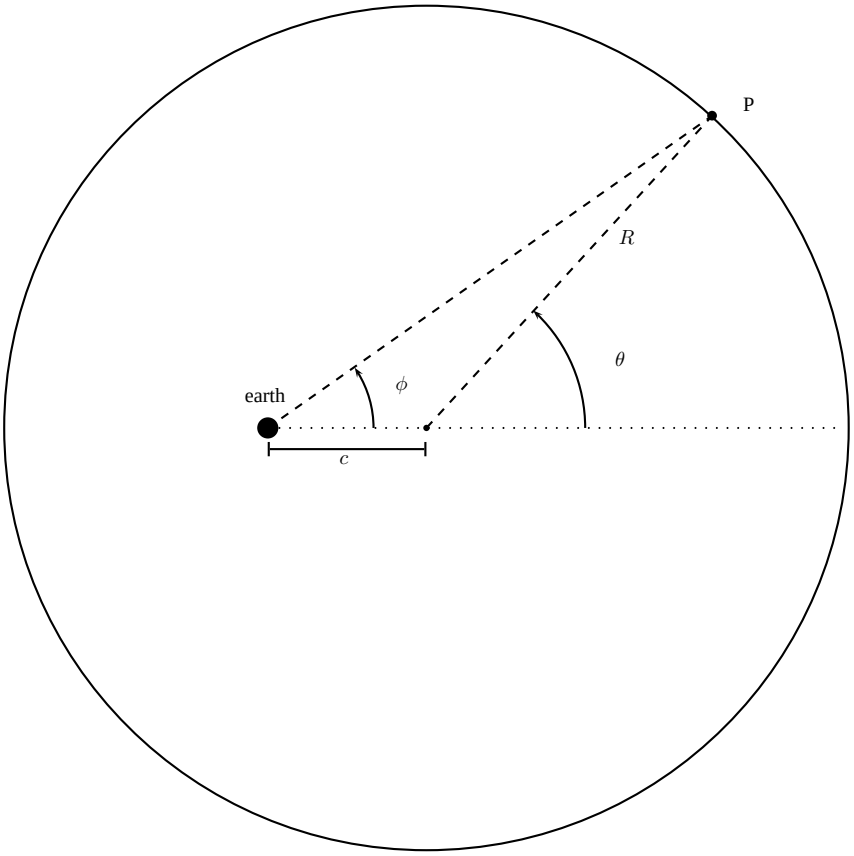

Figure 1.16: An eccentric circle. The point P (which represents a planet, here the Sun) moves around the circle with uniformly-increasing θ. But since the center of the circle doesn't coincide with the earth, the observed angle of P relative to the stars will be ϕ, which will increase in a not-quite-uniform way.

Use the triangle in the figure to write an expression for ϕ (the observable angular position of P on the ecliptic) in terms of θ. Then use the fact that θ should increase according to $\theta(t) = \theta_0 + \omega t$ (with ω a constant) to write an expression for $\phi(t)$. It should depend on c/R, and ω. Finally, do a curve-fit with the actual data for the Sun to find the best values for these parameters. How good is the fit? Examine the residuals. Are there still systematic errors?

1.7 Following up again with the Sun: hopefully you found that an eccentric orbit can reproduce quite accurately the data for the Sun's progression along the ecliptic to within an accuracy of about a tenth of a degree. Let's now explore how well you can do with the final scheme developed by Ptolemy for accounting for "anomalous" motion of the planets: the "equant." Here the motion of the Sun is around a circle

centered at the earth, but the angular velocity is constant with respect to a point called the equant which is off to the side by some distance. Draw a simple diagram to help you work out how the observed ecliptic angle of the Sun should vary in time in this scheme, and then do another curve-fit. You should find that you can do just about as well with the equant as with the eccentric (though the exact nature of the remaining small errors will be slightly different with the two devices). According to Ptolemy, the Sun has just one "anomaly" and it is hence a matter of choice whether one corrects this using an epicycle, an eccentric, or an equant. The 5 planets, however, turn out to have *three* anomalies each – so the fully detailed account of their motion required, in Ptolemy's system, all three devices for each planet! This turns out to play an important role in Copernicus' arguments for a heliocentric system.

1.8 Your teacher will provide some data for the angular position (along the ecliptic) of a planet, over the course of several years. Perform a curve-fit to find the values of ω_d, ω_e, and R_d/R_e – from Equation 1.28 – that provide the best fit to the data. (The data will be pre-arranged so that, at t=0, the planet will be in the middle of its retrograde cycle – so you can automatically set $\theta_0^d = 0$ and $\theta_0^e = \pi$.) Note also the time between subsequent retrogradings of your planet, the so-called synodic period. Be as accurate as possible and get at least two significant figures of precision for each parameter. Students should be assigned to different planets, so the class as a whole can compile and discuss the results for all the planets.

Chapter 2

The Copernican Revolution

This chapter covers the transition from the geo-centric theory of Ptolemy, to the now widely-known and -accepted *helio-centric* theory of Copernicus. We begin by simply presenting the Copernican theory of the solar system, with a focus on the numerical coincidences in the Ptolemaic system which are removed and explained by Copernicus' theory. We then briefly survey the two sets of ways that Galileo contributed to the acceptance of Copernicus' ideas – and paved the road for Newton. And finally we describe the contributions of Kepler, who clarified and supplemented Copernicus' basic scheme with some additional insights that were crucial to Newton's eventual discovery of his theory of gravity.

2.1 Copernicus

Perhaps the most surprising thing about Copernicus' revolutionary 1543 *On the Revolution of the Heavenly Spheres* is how un-revolutionary it is. Although he is arguing for a radical reconception of the place of earth (and hence mankind) in the universe, Copernicus accepts and implements almost all of the cosmological and astronomical premises of the Greeks: for example, that the universe is spherical and bounded by a sphere of fixed stars, that the proper motion of heavenly bodies is eternal circular motion, and that the small irregularities or anomalies in the motion of the heavenly bodies should be accounted for with Ptolemaic correctives such as eccentrics and epicycles. Indeed, Copernicus' own central argument for his heliocentric system was a seemingly very marginal point of detail: it allowed him to do without the particular Ptolemaic device called the equant, which Copernicus regarded as an abhorrent departure from the basic axiom of uniform circular motion. Copernicus was, in short, a surprisingly conservative revolutionary.

But nevertheless, Copernicus' work did begin a revolution that ultimately culminnated in Newton's theory of universal gravitation. So let us examine it.

Copernicus' first big claim is that the (apparent) daily westward rotation of the entire heavens is best understood (instead) as a daily *eastward* rotation of the earth:

> "Although there are so many authorities for saying that the Earth rests in the centre of the world that people think the contrary supposition inopinable

> and even ridiculous; if however we consider the thing attentively, we will see that the question has not yet been decided and accordingly is by no means to be scorned. For every apparent change of place occurs on account of the movement either of the thing seen or of the spectator, or on account of the necessarily unequal movement of both. For no movement is perceptible relatively to things moved equally in the same directions – I mean relatively to the thing seen and the spectator. Now it is from the Earth that the celestial circuit is beheld and presented to our sight. Therefore, if some movement should belong to the Earth it will appear, in the parts of the universe which are outside, as the same movement but in the opposite direction, as though the things outside were passing over. And the daily revolution in especial is such a movement. For the daily revolution appears to carry the whole universe along, with the exception of the Earth and the things around it. And if you admit that the heavens possess none of this movement, but that the Earth turns from west to east, you will find – if you make a serious examination – that as regards the apparent rising and setting of the Sun, moon, and stars the case is so."

Part of the argument for the rotating earth is that it gives a *simpler* explanation of the apparent motion of the heavens. What observation supports is merely the claim that the entire extra-terrestrial universe (the stars, the planets, the Sun, and the moon) moves in a certain way *relative to the earth*. So why have all those other objects move, when the motion of just one object – the earth – will equally well account for the observations?

Interestingly, Copernicus also rests his argument in favor of the rotating Earth on the fact that the Earth is a sphere, and circular or rotational motion is (he claims) natural and proper for a spherical object:

> "the movement of the celestial bodies is circular. For the motion of a sphere is to turn in a circle; by this very act expressing its form, in the most simple body, where beginning and end cannot be discovered or distinguished from one another, while it moves through the same parts in itself."

Thus, since "the Earth is held together between its two poles and terminates in a spherical surface"

> "Why therefore should we hesitate any longer to grant to it the movement which accords naturally with its form, rather than put the whole world in a commotion – the world whose limits we do not and cannot know? And why not admit that the appearance of daily revolution belongs to the heavens but the reality belongs to the Earth?"

If – on either of these arguments – you are willing to accept that the earth *rotates*, it is then only a little more of the same to accept that the yearly (apparent) motion of the Sun around the ecliptic, is in fact due to the yearly orbit of the earth around the (stationary) Sun. Thus, according to Copernicus, it is the Sun which is at rest at the center of the universe. The earth is then just another planet, orbiting around the Sun in a more-or-less circular trajectory. The basic scheme is sketched in Figure 2.1.

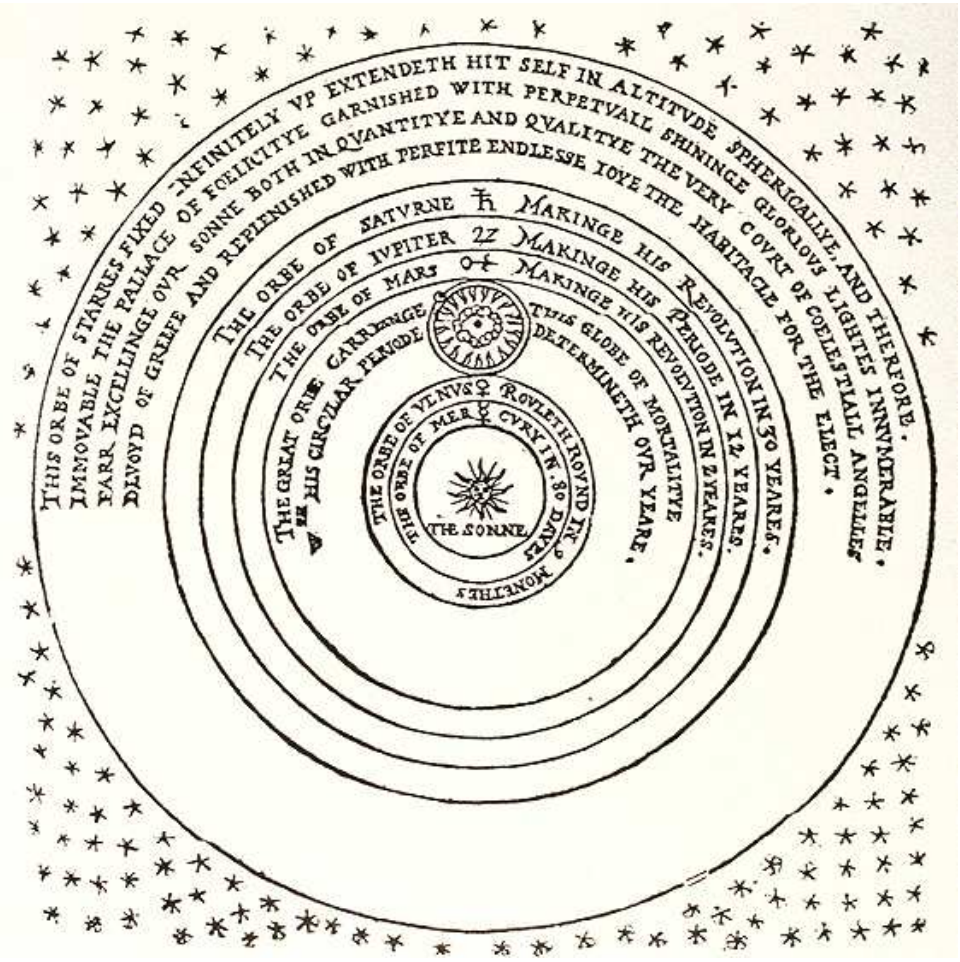

Figure 2.1: A famous depiction of Copernicus' model of the universe, from a 1576 text by Thomas Digges which included the first English translation of much of Copernicus' *De Revolutionibus*. The Sun lies at the center, surrounded in turn by the orbits of Mercury, Venus, Earth (with its moon), Mars, Jupiter, and Saturn. Note that Digges (unlike Copernicus) depicts the solar system as embedded in an expanse of stars, rather than a fixed shell. This is a natural extension from what Copernicus proposed since, unlike Ptolemy, Copernicus doesn't require the fixed stars to move. There is hence no particular reason they should be attached together (on a spherical shell) rather than spread out through space.

Let's try to understand in detail how the apparent motion of the Sun and planets is accounted for in Copernicus' helio-centric theory. Figure 2.2 shows how the annual eastward movement of the Sun around the ecliptic can be understood in terms of the posited annual orbit of the Earth around a fixed Sun. For Copernicus, the ecliptic – whose original definition is just the set of points that the Sun occupies on the map of fixed stars – can be understood in terms of the Earth's orbit. That orbit lies in a *plane* which intersects the Sun, and so the set of possible apparent positions of the Sun against the background of fixed stars is simply the circle on which the Earth's orbit plane intersects the sphere of fixed stars.

What about the progression of the seasons? The observational fact that the ecliptic is tilted with respect to the celestial equator is explained by Copernicus as follows: the axis about which the earth's daily rotation occurs, is not perpindicular to the plane of its yearly orbit around the Sun. The rotation axis does not point to the pole of the ecliptic, but is instead tilted by 23.5° and hence points toward the celestial pole (currently near the star Polaris, the "north star"). The earth's rotation axis remains (more or less) fixed in space (pointing to Polaris) as the earth orbits the Sun, as illustrated in Figure 2.3.

As already mentioned and sketched in Figure 2.1, the five planets (Mercury, Venus, Mars, Jupiter, and Saturn) will – like Earth – make roughly circular orbits around the Sun. For Mercury and Venus, whose orbits are *interior* to the Earth's orbit, it is pretty clear how their motion will be perceived, from Earth, as a kind of back-and-forth motion along the ecliptic that is always centered at the Sun. In particular, the retrograde motion of the two "inferior" planets will be explained, in Copernicus' theory, by the motion of the planets when they are on the near side of the Sun (as seen from Earth). But what about the three "superior" planets: Mars, Jupiter, and Saturn? How is their occasional retrograde motion accounted for by Copernicus?

In Ptolemy's theory, the motion of these planets was analyzed in terms of two conjoined circular motion: a "big" circular motion (of the deferent point around earth), and a "small" circular motion (of the planet on an epicycle around the deferent point). The key to understanding Copernicus' explanation for the apparent motion of these planets is to grasp that, at the level of mathematical description, the account is precisely the same! Here too the motion of (say) Mars *with respect to Earth* can be analyzed in terms of the conjunction of two circular motions – namely, the motion of Mars around the Sun (which corresponds in Ptolemy's theory to the "big" circular motion), and the motion of Earth around the Sun (which corresponds in Ptolemy's theory to the "small" circular motion, the epicycle). That is: Copernicus replaces the epicycle (which Ptolemy introduced precisely to account for the retrograde motion) with the motion of the Earth around the Sun. Figure 2.4 is an attempt to sketch the mathematical argument that the two schemes must make the same observational predictions for the motion of Mars (and Jupiter and Saturn) with respect to Earth.

Nevertheless, it is somehow harder to see intuitively how retrograde motion arises in the Copernican system, than in the Ptolemaic system. Figure 2.5 sketches the way to understand this, and the caption explains how the explanation can be understood to apply to both the inferior and superior planets.

Now we can finally see the major sense in which Copernicus' model of the solar

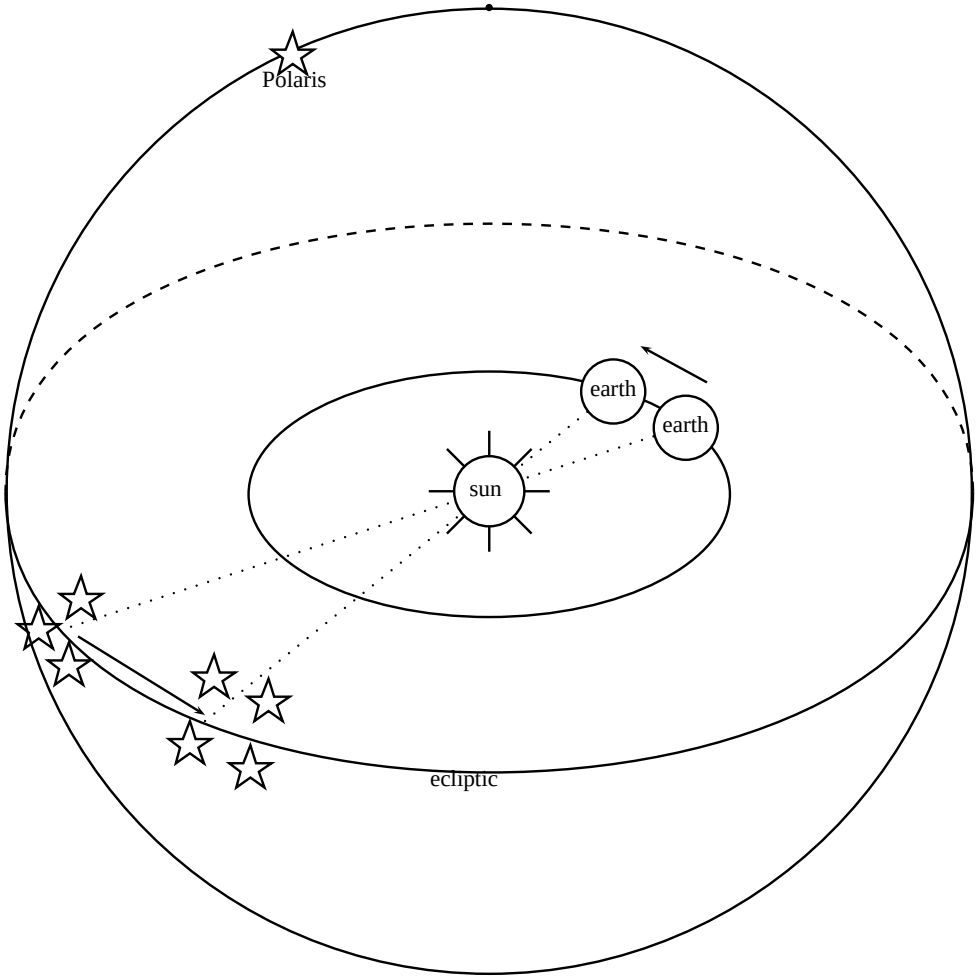

Figure 2.2: Motion of the Earth around the stationary Sun results in an apparent motion of the Sun with respect to the fixed stars. Note that in the Copernican system the ecliptic is understood in terms of the plane in which the earth's orbit lies. It is because we (on earth) always view the Sun from some point on that orbit, that the Sun always appears at some point on the ecliptic. That is, the ecliptic (thought of as a path along the sphere of fixed stars) is the intersection of the plane of the earth's orbit with the sphere of fixed stars.

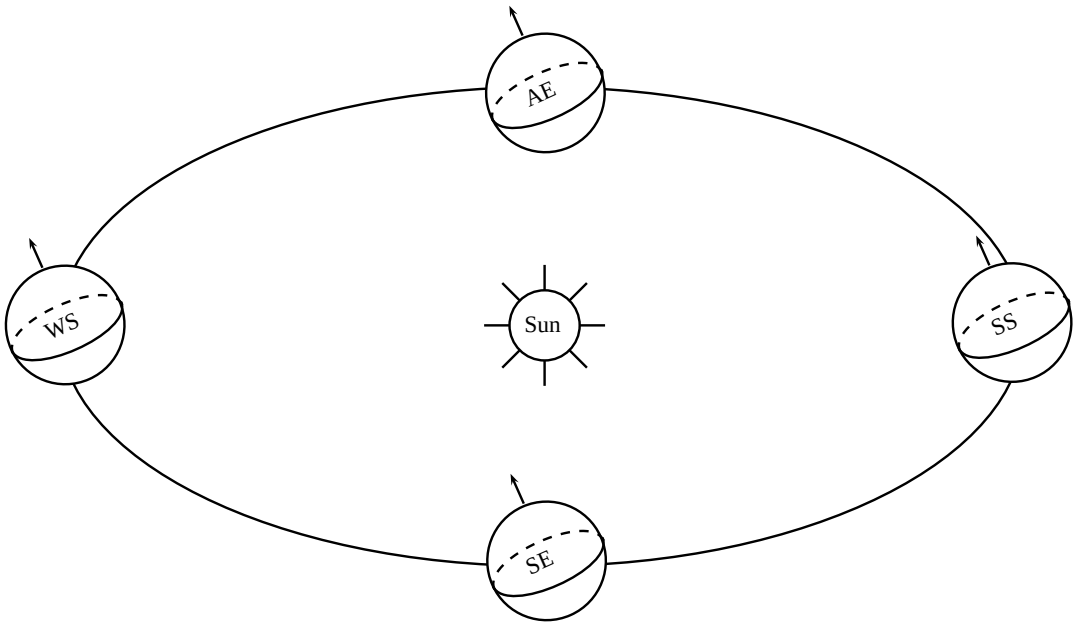

Figure 2.3: The earth is shown at four points in its yearly orbit around the Sun. At the summer solstice (SS) the 23.5° tilt of the earth's rotation axis down from the pole of the ecliptic is directly toward the Sun. On that day, the Sun will therefore be directly overhead at noon for an observer on the Tropic of Cancer (23.5° North Latitude). For observers in more northerly latitudes, the Sun will be higher in the sky at noon than at any other day during the year. Six months later, on the winter solstice (WS), the earth's rotation axis is tilted *away* from the Sun. (That, at least, is how we describe it in the Northern Hemisphere!) On that day, an observer on the Arctic Circle (66.5° North Latitude) won't see the Sun rise above the horizon at all. Observers in moderate northern latitudes will see the Sun lower in the sky at noon than on any other day during the year. Also pictured are the Autumn and Spring equinoxes (AE and SE), when the plane defined by the Earth's equator intersects the Sun. On these days all observers will see the Sun rising and setting precisely to the east and west, respectively, and the time between sunrise and sunset will be precisely half a day.

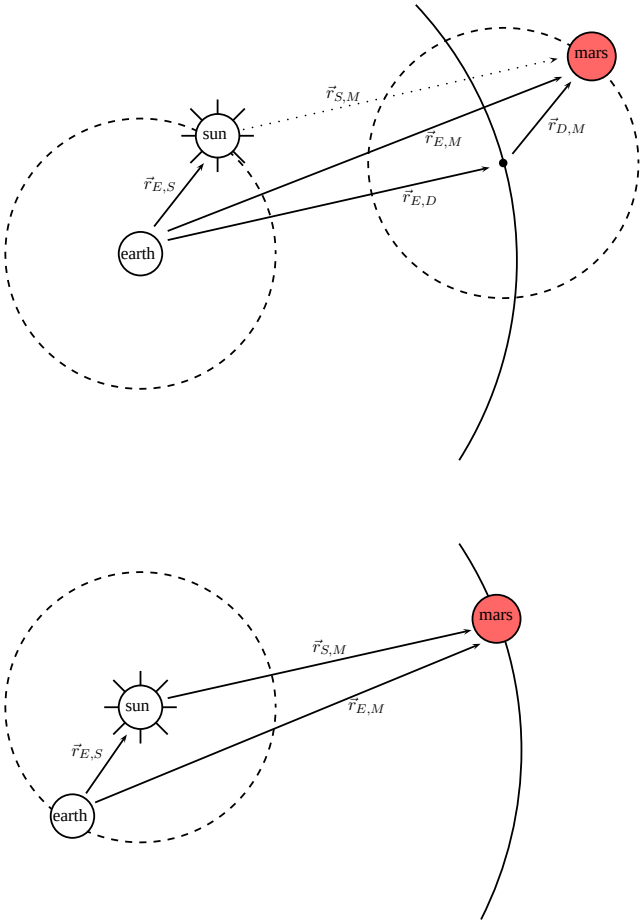

Figure 2.4: The top and bottom parts show (respectively) how the position of a planet (Mars) is analyzed in Ptolemy's theory and Copernicus' theory. For Ptolemy, the position of Mars with respect to Earth is given by the sum of two vectors (each of which maintains a constant magnitude but changes its direction uniformly in time): $\vec{r}_{E,M} = \vec{r}_{E,D} + \vec{r}_{D,M}$. It is the coincidence noted in the text that the vector representing the position of Mars on its epicycle – $\vec{r}_{D,M}$ – is just the same as the vector representing the position of the Sun relative to the Earth: $\vec{r}_{E,S}$. According to Ptolemy, the vector labeled $\vec{r}_{S,M}$ is not directly relevant to the motion of Mars – it is only by coincidence equal to his $\vec{r}_{E,D}$. By instead referencing the position of Mars directly to the Sun, Copernicus removes the coincidence and accounts for the position of Mars with respect to Earth as shown in the bottom part of the figure: $\vec{r}_{E,M} = \vec{r}_{E,S} + \vec{r}_{S,M}$. Note that, in Ptolemy's system, the absolute sizes of the deferent and epicycle are not determined – only their ratio is observationally meaningful. We have here depicted Mars' epicycle as having the same size as the Earth's orbit around the Sun to emphasize the correspondence between the two theories. It's important, though, that according to Copernicus *what Ptolemy would have called* the Sun's orbit around the Earth and Mars' epicycle *have the same size*. And since one of these sizes has already been measured, the other is determined too, which in turn fixes (what Ptolemy would have called) the size of Mars' deferent circle. But that corresponds, in Copernicus' system, to the size of Mars' orbit around the Sun. This is why one can, in Copernicus' system but not in Ptolemy's, determine the absolute size of the orbits of the planets.

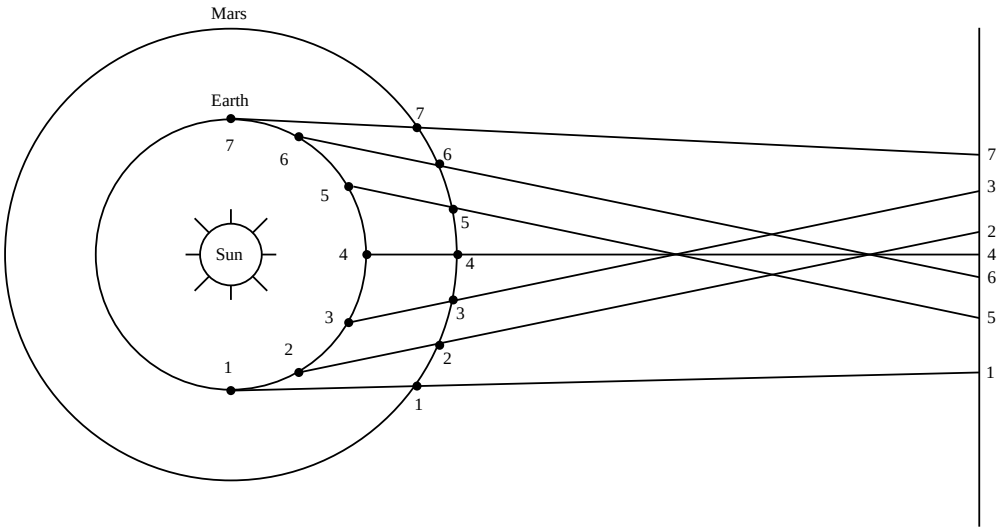

Figure 2.5: How retrograde motion arises in the Copernican system: the positions of Earth and Mars are shown at seven times over the course of half a year, as are the apparent positions of Mars against the backdrop of fixed stars. Between 1 and 2, Mars appears to move relatively quickly (to the east) relative to the stars. Between 2 and 3 it continues to move eastward, but more slowly. At 3, though, it reverses its motion and retrogresses (i.e., moves to the west) until 5, at which point it again reverses and continues its normal (easterly) motion relative to the stars. The same process also explains how the other two superior planets (Jupiter and Saturn) come to retrogress. Note that the distance to the fixed stars is (contrary to the figure) really supposed to be *much* larger than the size of Earth's orbit. So really one can perceive the retrograde motion merely from the way the angle of the "lines of sight" evolve over time: from 1 to 3 the line of sight rotates counter-clockwise, which corresponds to an apparent motion of Mars to the east. Then from 3 to 5 the line of sight rotates clockwise, corresponding to an apparent westerly Martian motion. Then from 5 to 7 the line of sight rotates again counter-clockwise. This way of understanding the origin of retrograde motion is helpful, because it allows one to immediately infer, for example, that if one lived on Mars and was charting the apparent position of the Earth relative to the (unshown) fixed stars (to the left), one would also observe retrograde motion during this same period. (Clockwise rotation of the line of sight corresponds to westerly apparent motion, whether one thinks of the line of sight extending to the right or to the left in the figure!) And one can then re-label Mars in the figure as "Earth" and re-label Earth in the figure as (say) "Venus" and hence understand (from the same figure) how the inferior planets come to retrogress. The general rule for retrograde motion can be formulated this way: a given planet will undergo retrograde motion as seen from earth when the relatively-inferior planet (which is Earth if we're talking about Mars, Jupiter, or Saturn, and is the planet itself if we're talking about Mercury or Venus) overtakes or "passes under" the relatively-superior planet.

system is *simpler* than Ptolemy's. Ptolemy's theory required a completely independent construction (deferent and epicycle) for each planet. But it turns out that the same number – the angular velocity of the Sun – shows up again and again in these separate constructions. Each planet is somehow or other "infected" with the motion of the Sun. For the two inferior planets, their *deferents* rotate at the same rate as the Sun, while for the three superior planets, their *epicycles* rotate at the same rate as the Sun. There is no reason for this in Ptolemy's system. It is just a *coincidence* that emerges when one fits the model to real data.

Copernicus removes and explains the coincidence by letting the same one circle – the Earth's orbit around the Sun – do all the jobs that were done by the coincidentally-identical circles in Ptolemy's theory. Copernicus replaces five of Ptolemy's circles (the deferents of Mercury and Venus, and the epicycles of Mars, Jupiter, and Saturn) with just one circle (the Earth's orbit around the Sun).

But we have glossed over something important. So far we have only spoken of one property of Ptolemy's circles: their speeds or angular velocities. But circles have another important property, too: *size*! And recall that, in Ptolemy's system, only the *relative* sizes of the deferent and epicycle was determined by the data – i.e., only the *ratio* R_d/R_e could be fit to observation. But in Copernicus' model, one or the other of these two radii (depending on which planet one is talking about) actually refers to the size of the Earth's orbit around the Sun – which is *known* (albeit with a large uncertainty). And so the *other* number – i.e., the absolute size of the different planet's orbit around the Sun – can be determined also, absolutely.

Copernicus quite properly advertises this aspect of his theory as one of its major virtues. By in effect reducing the number of free parameters in the theory, the whole system becomes much more tightly integrated, logically speaking, such that nothing can be adjusted without affecting the rest – or saying the same thing backwards, there is an unambiguous road from certain things that are already known (like the distance to the Sun) to new facts such as the sizes of the planetary orbits:

> "I found after much and long observation, that if the motions of the other planets were added to the motions of the earth, ... not only did the apparent behavior of the others follow from this, but the system so connects the orders and sizes of the planets and their orbits, and of the whole heaven, that no single feature can be altered without confusion among the other parts and in all the Universe. For this reason, therefore, ... have I followed this system."

As we will see later in the chapter, the fact that Copernicus' theory allows the sizes of the planetary orbits to be determined, yields great additional fruit in the hands of one of Copernicus' two great followers: Johannes Kepler.

Let us mention one more of the observational facts discovered by the Greeks and how it is explained in Copernicus' system: the precession of the equinoxes. Recall that in Ptolemy's theory, this was explained by attributing, in addition to the daily rotation, a second (more subtle) motion to the sphere of fixed stars: a sort of "wobble" by which the pole of the ecliptic moved (slowly) about the celestial pole such that, over the course of thousands of years, the celestial pole would migrate around a circle (centered at the

pole of the ecliptic). In Ptolemy's theory this motion, just like the daily rotation, was "inherited" by all of the other objects in the universe *except* the Earth which was, of course, fixed at the center. So the same argument we began with is clearly going to motivate Copernicus to instead attribute this motion to the Earth: the rotational axis of the earth does not stay *precisely* fixed in direction as the earth orbits the Sun; rather it turns by about a degree or so each century such that (refering again to Figure 2.3) after some 13,000 orbits around the Sun, the rotation axis will be tilted 23.5° to the *right*. The star Polaris will then be a whole 47° away from the north celestial pole, and it will be the middle of winter when the Earth is in the part of its orbit (to the right in the Figure) that now corresponds to summer.

We may thus summarize Copernicus' basic theory as follows. Copernicus attributes three motions to the earth: a daily rotation, a yearly orbit, and a slow precession of the rotation axis (to account for the precession of the equinoxes). This allows him to get rid of the daily rotation of the stars, planets, Sun, and moon; the motion of the Sun relative to the stars; and the long-period wobble. He also gets rid of the planetary epicycles, and, in so doing, determines the absolute sizes of all the planets' orbits. At this level of description, the theory is really elegant and extremely useful. But there are two serious caveats.

First, what about all the objections to the heliocentric solar system (which remember had been proposed in Ancient Greece by Aristarchus) that we reviewed at the end of the previous chapter? Did Copernicus have any good answers to these? Not really. For example, he recognized that his theory predicted either that the stars should display a sizable annual parallax – or that the stars were much, much further away than had been previously conceived. And, at the end of the day, he had nothing better to say than: I guess they are very far away.

> "...the dimensions of the world are so vast that though the distance from the Sun to the earth appears very large compared with the size of the orbs of some planets, yet compared with the dimensions of the sphere of fixed stars, it is as nothing."

That of course turns out to be *true*. But you can see why it didn't go very far in convincing many of his contemporaries. Copernicus also doesn't have much useful to say in response to the worry that a rotating earth would mean constant strong westerly winds, birds that can't catch worms, and rocks falling miles to the west of where they were thrown from. Essentially his response is: evidently the air and birds and rocks share in the same daily rotation that the Earth as a whole undergoes:

> "what would we say about the clouds and the other things floating in the air or falling or rising up, except that not only the Earth and the watery element with which it is conjoined are moved in this way but also no small part of the air and whatever other things have a similar kinship with the Earth? .... Hence the air which is nearest to the Earth and the things floating in it will appear tranquil..."

But as we discussed at the end of the previous chapter, this idea is really in conflict with then-widely-accepted principles of physics (rest as natural, external force needed to

explain motion, etc.). Copernicus' ideas were thus not immediately accepted by many others.

The second big problem is that the simple helio-centric theory as presented so far fails to match the observational data. To be specific: it matches the data just as well as – or probably one should say just as *poorly* as – a Ptolemaic theory which has *just* a single deferent-epicycle construction for the planets, and simple uniform circular motion for the Sun. But precisely because this theory didn't account for the details of the observed motion of the Sun, Moon, and stars, Copernicus – just like Ptolemy – had to introduce a number of suspicious devices (such as eccentric circles and additional epicycles). For Ptolemy, these devices seemed somehow marginal or minor: what's the big deal about making the orbit of Mars, for example, slightly eccentric, when it already has a big epicycle? But precisely because the basic version of Copernicus' theory is so elegant and coherent compared to Ptolemy's, the needed eccentrics and epicycles stand out rather dramatically as ugly blemishes.

Let us see in more detail how and why these blemishes were introduced.

In the previous chapter, we discussed Ptolemy's use of the eccentric in relation to the Sun. Ptolemy introduced an eccentric orbit for the Sun because, as a brute matter of observational fact, the equinoxes did not quite divide the year in half: the winter "half" of the year (from the Autumn Equinox to the Spring Equinox) was several days shorter than the summer "half" of the year, i.e., the Sun moves more rapidly along the ecliptic during the winter than during the summer. To account for this, Ptolemy displaced the center of the Sun's orbit slightly from the Earth. Similar fixes were also required for the Moon and other planets. Indeed, the planets other than the Sun required *several* such fixes each. Thus, in addition to the deferent circles being displaced somewhat from the Earth, the deferent points were made to rotate at fixed angular velocity with respect to some point (the equant) different from both the earth and the center of the deferent circle, and/or additional epicycles were added, and/or the epicycles themselves were made to be slightly eccentric with respect to the deferent point, etc.

Let's first understand how Copernicus dealt with the slightly irregular motion of the Sun, and then sketch how he treated the other planets.

Copernicus, of course, has the Earth go around the Sun, rather than the Sun go around the Earth. But Ptolemy's method of fixing up the relative motion of these two objects – making the orbit eccentric – works just as well in Copernicus' theory: one simply shifts the center of the Earth's (circular) orbit a little ways away from the Sun, as shown in Figure 2.6.

Rather curiously, Copernicus refers to the center of the Earth's circular orbit as the "Mean Sun". (This really doesn't make any sense, though there is a kind of half-sensible justification for it. Can you figure out what that is?) And he seems to regard *this* point as occupying the center of the universe – it is, for example, the "Mean Sun" (as opposed to the actual position of the actual Sun) to which Copernicus refers the orbits of all the other planets. This minor fix therefore has a rather major qualitative implication: it makes Copernicus' system not, in fact, helio-centric after all! It's of course not exactly geo-centric either, since the Earth orbits around the point Copernicus considers to be the true center. Maybe one should call it geo-centric-centric, since the true center is

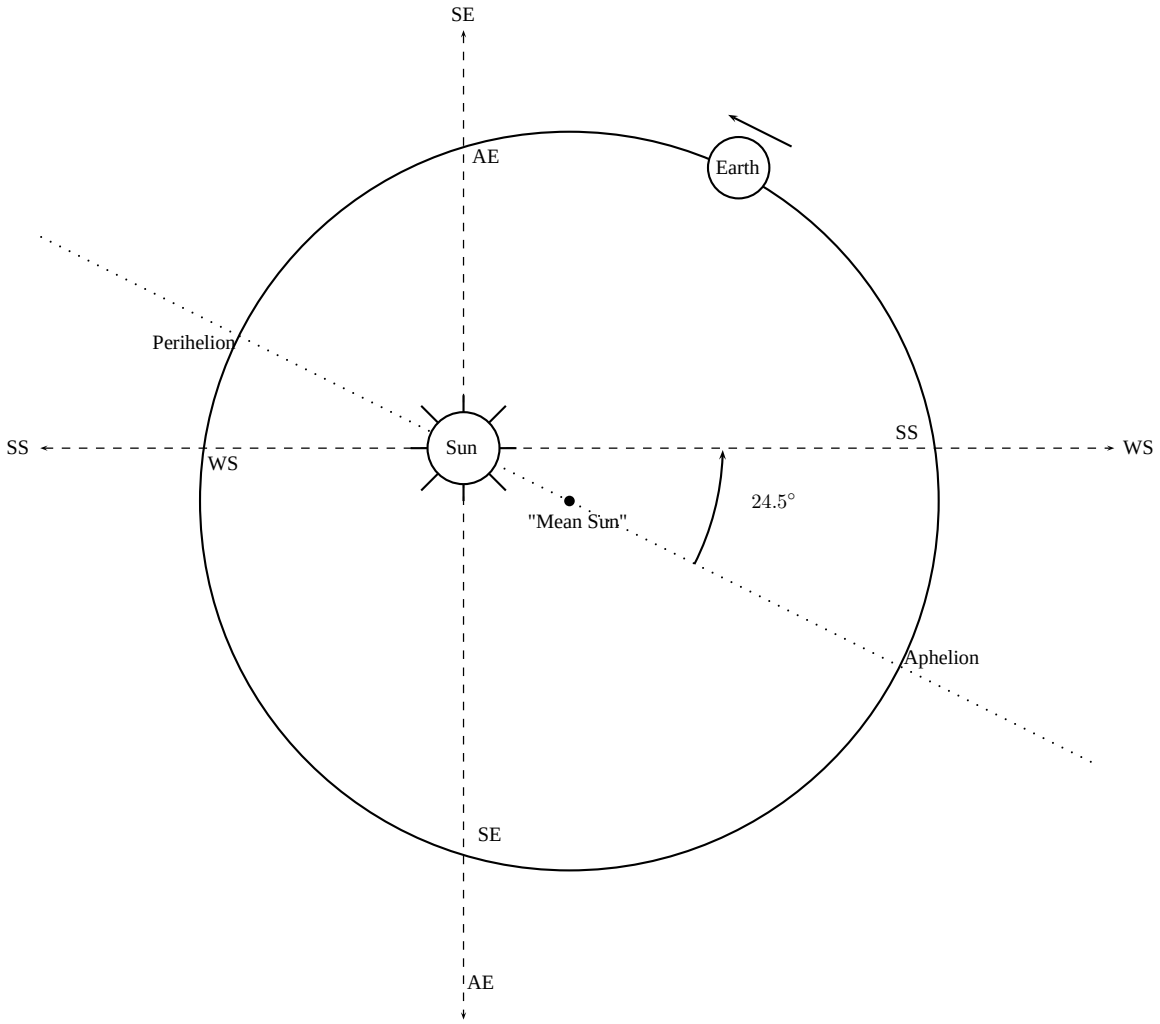

Figure 2.6: The Earth's eccentric orbit around the Sun. The horizontal and vertical lines point to the apparent position of the Sun against the distant fixed stars, as seen from earth on the solstices and equinoxes. The same symbol, SS, labels both the position of the Earth at the Summer Solstice and the apparent position of the Sun against the fixed stars at the Summer Solstice; likewise for the Autumn Equinox (AE), Winter Solstice (WS), and Spring Equinox (SE). According to Ptolemy, the Earth-Sun distance is greatest when the Sun is about 24.5° west of the Summer Solstice point. In Copernicus' model, this means the Earth reaches "aphelion" (the point in its eccentric orbit where it is farthest from the Sun) when it is 24.5° shy (as measured from the Sun) of the point on its orbit labelled SS. The center of the Earth's circular orbit is labelled "Mean Sun" (which is Copernicus' somewhat odd terminology). The observational data requires an eccentricity of about one part in twenty, meaning that the "Mean Sun" is displaced from the *Actual* Sun by about 1/20 the radius of the Earth's orbit. (Note that Copernicus refers to the actual location of the Sun not, as I just have, as the "Actual Sun" but rather as the "Apparent Sun" – because that is where it appears to be! You can see that it's slightly embarrassing for him that the Sun is displaced from the center of the Earth's orbit – i.e., that the system really isn't heliocentric!) The dotted line connecting the Mean and Actual (or "Apparent") Suns, which hence also passes through the Earth's aphelion and perihelion points, is called the "apsis."

not the Sun, but the center of the Earth's orbit. Or maybe, since Copernicus calls this point the Mean Sun, one should call the theory mean-helio-centric? The terminology, of course, doesn't matter. But it is important that Copernicus' helio-centric system really did *not* have the Sun at the center, and really did *not* refer the orbits of the planets to the actual position of the Sun. For as we will see later in the chapter, major advances were made by Kepler when he took helio-centrism more seriously than even Copernicus had!

Actually, things are even a bit worse in Copernicus' theory than I've indicated. The previous discussion and Figure 2.6 describe how Copernicus would account for the irregularity of the Sun's apparent motion *as those irregularities were known to Ptolemy.* But in the many centuries between Ptolemy and Copernicus, additional irregularities had appeared – in particular, the precise details of the eccentricity had changed. As Copernicus puts it:

> "the distance of the Sun from the centre of the orbital circle... has now become approximately 1/31st [of the radius of the Earth's orbital circle], though to Ptolemy it seemed to be 1/24th. And [the aphelion, the Earth's farthest point from the Sun], which was at that time 24.5° to the west of the summer solstice, is now 6 2/3° to the east of it."

That is, both the magnitude and direction of the eccentricity of the Earth's orbit had changed.

How did Copernicus account for this? He let the center of the Earth's orbit – the Mean Sun – move around a small circle centered at the point labeled "Mean Sun" in Figure 2.6. This is a really small irregularity which only manifests itself over long periods of time. We will basically set this aside until we are ready to apprehend its cause in Chapter 6. It is mentioned here only to provide a further sense of the ways that Copernicus needed to mar the initial elegance of his theory, with a slew of Ptolemy-style epicycles and other devices, in order to achieve the same sort of match to the observational data that Ptolemy's theory enjoyed.

But let us return to one of the virtues of Copernicus' theory, at least as claimed by Copernicus. Leaving aside the minor irregularity just discussed, the Sun's apparent motion around the ecliptic (i.e., the Earth's orbit around the Sun) is sufficiently well-described by an eccentric circle. We saw in the previous chapter that using an eccentric is mathematically equivalent to using an epicycle (with $\omega_e = 0$), and also that an equant can account for the anomalous motion basically as well as an eccentric or epicycle (though the equant is not quite mathematically equivalent to the other two). The point is, in Copernicus' way of thinking, the Earth's orbit around the Sun had a *single anomaly* – which could be fixed to the necessary precision by using an epicycle *or* an eccentric *or* an equant.

As far as the Earth-Sun system is concerned, Ptolemy and Copernicus agree. But what about the other planets? Ptolemy had found the motions of the planets to have *three* anomalies (each). Thus, in Ptolemy's system, a planet's motion would involve a deferent circle corresponding to the expected, proper, non-anomalous circular motion – with the observational data then requiring the planet to move on an epicycle, *and* for

the center of the deferent circle to be displaced somewhat from the earth (an eccentric), *and* for the angular motion of the deferent to be uniform with respect to some other point, not quite at the deferent's center (an equant).

We mentioned at the beginning of the chapter that one of Copernicus' central motivations for proposing his theory was that it allowed him to do without the equant:

> "the planetary theories of Ptolemy and most other astronomers, although consistent with the numerical data, seemed ... to present no small difficulty. For these theories were not adequate unless certain equants were also conceived; it then appeared that a planet moved with uniform velocity neither on its deferent nor about the center of its epicycle. Hence a system of this sort seemed neither sufficiently absolute nor sufficiently pleasing to the mind. Having become aware of these defects, I often considered whether there could perhaps be found a more reasonable arrangement of circles, from which every apparent inequality would be derived and in which everything would move uniformly about its proper center, as the rule of absolute motion requires."

What was Copernicus' objection to the equant? He considered it to be an abhorrent departure from the basic axiom of astronomy – that the heavenly motions should be understood in terms of uniform circular motion:

> "We must however confess that these movements are circular or are composed of many circular movements, in that they maintain these irregularities in accordance with a constant law and with fixed periodic returns: and that could not take place, if they were not circular. For it is only the circle which can bring back what is past and over with..."

We are finally in a position to see, at least in outline, how Copernicus managed without the dreaded equant.

First consider the Ptolemaic construction for some planet, say Mars. The basic scheme is as shown in the top half of Figure 2.4. The more detailed model (required to achieve adequate agreement with observation), however, added two additional anomalies (in addition to the big epicycle): the deferent circle was made to be eccentric, and the epicycle was given an equant. We have seen in Chapter 1 that, qualitatively though not precisely, the equant and eccentric are equivalent. Thus Ptolemy could have achieved equally good agreement with observation by making the epicycle eccentric, instead of using a deferent, had he so chosen. It would have been three total anomalies per planet either way, so really it's a matter of indifference. Probably Ptolemy's rationale for choosing the equant rather than a second eccentric for that third anomaly was predominantly aesthetic: it somehow seemed nicer to use each of the three devices once per planet rather than double up. In any case, Ptolemy freely used equants in his descriptions of the planets, and Copernicus made a big deal of the fact that his system could achieve the same accurate match to the observations without any equants.

To understand how Copernicus' system allowed this, one need only think about how a Ptolemaic doubly-eccentric epicycle construction (qualitatively equivalent to the

eccentric-epicycle-equant construction Ptolemy actually used) translates into Copernicus' theory. The two eccentric circles remain – only, instead of one of them being an (eccentric) epicycle, the second eccentric circle will be the Earth's (eccentric) circular motion around the Sun. That is, Copernicus achieves the same accuracy that Ptolemy achieved by introducing *three* "anomalies" per planet, with only *one* anomaly per planet – which Copernicus chose to be eccentrics. This is because one of Ptolemy's three anomalies – the epicycle – corresponded to something in Copernicus' system which isn't anomalous at all: the Earth's roughly circular orbit around the Sun.

Copernicus did have a point here, and his system is importantly better than Ptolemy's on this front. We'll discuss that point shortly. But there is also a sense in which Copernicus overstates, or maybe just misstates, the nature of this improvement. Namely: Copernicus hangs too much on the equant. He makes it sound as if using equants was a necessary feature of Ptolemy's model, and was positively prohibited in his model. But neither of these is the case. Ptolemy could have gotten along just as well by using additional eccentrics or epicycles instead of equants, and Copernicus could easily and naturally have used equants instead of the eccentrics or epicycles he did employ. So the issue of equants is a bit of a red herring – especially when one remembers that Copernicus' reasons for despising the equant don't actually hold water.

On the other hand, it is genuinely important that Copernicus' system achieved the same accuracy with one "anomaly" per planet (for which Copernicus chose eccentrics) that Ptolemy required three anomalies per planet to achieve. This indicates a real simplification (as opposed to a mere arbitrary preference for one sort of anomaly-fix over another).

And note that this implies an additional, more subtle (and hence more impressive), sense in which Copernicus' theory explains some features of the planets' motions which were, according to Ptolemy, identical by sheer coincidence. We explained previously how, despite getting its own individual deferent-epicycle construction, each planet was somehow – inexplicably – "infected" with the motion of the Sun. In particular, the angular velocity of the Sun showed up – as either the deferent or epicycle angular velocity – in the construction for each of the other 5 planets. The point here is this: *so did the slightly anomalous behavior of the Sun.* That is: not only could the same one circle (the Earth's orbit around the Sun) successfully replace *five* of Ptolemy's circles – but (much more impressively) a single "anomalous circle" (e.g., an eccentric circle) could successfully replace five of Ptolemy's anomalous circles. So it is more like trading 10 free parameters in for 2 – rather than trading 5 for 1, as we described it before. So the simplification is something like twice as good as we advertised earlier: a reduction of roughly 8 free parameters compared to Ptolemy's theory.

So where does this leave us? No doubt Copernicus' system is simpler than Ptolemy's when it comes to accounting for the observed motions of the planets. The central point in its favor is that it replaces the major epicycle used by Ptolemy for each of the five planets, with the same one circular motion: the motion of the Earth around the Sun. And Copernicus' theory is not just numerically simpler in the sense that it accounts for the same observations with four fewer circular motions (or eight overall fewer free parameters). For, as Copernicus puts it, "the mobility of the Earth binds together the

order and magnitude of their orbital circles in a wonderful harmony and sure commensurability" which allows the absolute sizes of the planetary orbits to be unambiguously determined.

On the other hand, we have also seen that the overall simplicity of Copernicus' model relative to Ptolemy's is actually quite small, since both theorists require not just one or two circles for each planet, but a variety of seemingly *ad hoc* fixes: eccentrics, minor epicycles, etc. But, on the first hand again, if one shares Copernicus' distaste for the equant, his theory gains the clear upper hand even at the level of the detailed, fully fixed models. But then – back to the other hand – we have also seen that when the need arose, Copernicus didn't hesitate to, for example, add an extra epicycle to fix some small leftover irregularity – which makes it seem like Copernicus avoided using equants only because, when anomalies proliferated, he opted to use something else (another epicycle, say) instead of an equant. And anyway, it's really not at all clear why the equant is supposed to be worse – supposed to be more of a departure from the axiom of uniform circularity – than an eccentric or epicycle.

In short: it is by no means clear that Copernicus' theory represents a significant improvement over Ptolemy's, at least in so far as a scientific understanding of the motion of the planets is concerned. It's surely a reasonable hypothesis worth pursuing, but the motion of the planets don't exactly contain any knock-down conclusive proof that Copernicus is right and the Greeks were wrong. Add to this the fact that Copernicus' theory runs up against the various objections to putting the earth in motion that we reviewed at the end of Chapter 1, and you can start to see why Copernicus' ideas took some additional time to gain wide acceptance.

2.2 Galileo

Let us then turn to the first of the two great (and roughly contemporary) followers of Copernicus: Galileo Galilei (1564-1642). Some of Galileo's important discoveries in the field of kinematics are (hopefully) familiar from your previous physics courses. Here we will just sketch those aspects of his work which helped strengthen the case for a Copernican universe.

2.2.1 Inertia

Galileo's first major contribution to the Copernican revolution had its origins in a perhaps unlikely-seeming place: his careful experiments with familiar terrestrial objects such as pendulums, balls rolling down ramps, and projectiles. His crucial discovery in this area was that free vertical motion is motion with constant downward accleration and that free horizontal motion is motion with constant speed. Furthermore, the motion of an object like a thrown rock – which moves both horizontally and vertically – can be analyzed into the separate *non-interacting* horizontal and vertical components. In particular, the downward acceleration experienced by a ball after you throw it sideways in no way influences or arises at the expense of the horizontal motion you imparted to it when you

threw it. Rather, the horizontal motion just continues uniformly and indefinitely (to the extent that air resistance can be neglected).

This is the origin of the concept "inertia" which represents a rejection of several of the Ancient Greek ideas about motion. Inertia refers to the fact that, in the absense of a resistive force like air resistance, an object will simply retain its velocity. As it would be later clarified in the first and second laws of Newton, an applied force is needed to explain *changes* in velocity, not velocity itself. That is: uniform motion at constant speed – such as the horizontal component of the motion of a projectile – is just as "natural" as rest.

Galileo's understanding of the principle of inertia was flawed in two important ways. First, Galileo's inertia really only applied to the horizontal aspect of motion, since he conceptualized "gravity" not as an external force acting on things like balls, but as a kind of inherent tendency for them to accelerate downward. (This is a last remnant of the ancient Aristotelian notion of "proper place".) Thus, for Galileo, the "natural" motion of terrestrial objects like balls is (simultaneously) constant downward acceleration and constant horizontal velocity. The second flaw in Galileo's idea of inertia was a kind of equivocation on the word "horizontal." A horizontal line extended far enough will increase in altitude and eventually stray far from the Earth. Does horizontal inertia imply that a freely moving object will follow such a trajectory? No, said Galileo. For example, if you imagine rolling a ball across the planar surface of a giant table, it seems clear that it won't just keep going forever. To do so, it would have to be going increasingly *uphill* and would hence slow down and turn around. But if the table *curved* so as to maintain a constant height with respect to the surface of the earth, then the ball *would* just keep rolling forever (assuming again an absense of friction). Hence, concludes Galileo, the "natural," constant speed, horizontal aspect of motion is really *circular*, not rectilinear. It's just that, since the relevant circle is about as big as the circumference of the Earth, we don't notice it in the flight of thrown balls.

Despite these minor flaws, Galileo's experimentally-rooted concept of inertia was a profound advance. But what does it have to do with Copernicus? Recall from the end of Chapter 1 the following argument against the possibility of a moving Earth, as phrased here by Galileo:

> "As the strongest reason of all is adduced that of heavy bodies, which, falling down from on high, go by a straight and vertical line to the surface of the earth. This is considered an irrefutable argument for the earth being motionless. For if it made the diurnal [i.e., daily] rotation, a tower from whose top a rock was let fall, being carried by the whirling of the earth, would travel many hundreds of yards to the east in the time the rock would consume in its fall, and the rock ought to strike the earth that distance away from the base of the tower. .... This argument is fortified with the experiment of a projectile sent a very great distance upward; this might be a ball shot from a cannon aimed perpindicular to the horizon. In its flight and return this consumes so much time that in our latitude the cannon and we would be carried together many miles eastward by the earth, so that the ball, falling, could never come back near the gun, but would fall as far to the west as the earth had run on ahead."

Galileo rebuts the argument by pointing out that the rocks and balls in these kinds of situations would, during their flights, *maintain* whatever horizontal speed they had *initially* by virtue of the earth's eastward rotation – and would hence land right at the base of the tower or cannon, just as is in fact observed in this kind of situation.

Of course, Copernicus had already claimed the same thing – that, somehow, projectiles (and clouds, birds, etc.) partake of the uniform circular motion that (he argued) was proper and natural for the Earth. So what has Galileo added? Only a rigorous experimental proof that this is in fact how objects really move!

It is clarifying to understand how earlier thinkers had wrongly understood the visibly curved trajectory of projectiles. They had the idea that the horizontal aspect of the motion was "violent" and "unnatural" (i.e., something artifically and externally imposed by your hand or whatever) *and therefore fleeting.* The idea was that after the ball left your hand, the vertically downward motion (toward the ball's natural or proper place) increased, while the unnatural horizontal component of the motion died out, causing the trajectory to curve downward. This had the implication that, should the ball stay in the air long enough, it should eventually be found to be moving *straight down.* And so things like birds or arrows shot from bows – things which do stay in the air for a reasonably long period of time – should therefore, at least by the end of their motions, be moving straight up and down. Which would mean they would be observed to race to the west at a thousand feet per second if the earth were rotating!

Galileo gives a number of vivid examples which help explain how his concepts of motion (to use an ironic turn of phrase) remove the ground from under this sort of objection. For instance, Galileo mentions "another experiment, which is to drop a lead ball from the top of the mast of a boat at rest, noting the place where it hits, which is close to the foot of the mast." But, according to his opponents, "if the same ball is dropped from the same place when the boat is moving, it will strike at that distance from the foot of the mast which the boat will have run during the time of fall of the lead, and for no other reason than that the natural movement of the ball when set free is in a straight line toward the center of the earth." That is, an observer watching from the shore will allegedly see the ball fall straight down, whether the boat is moving out from underneath it or not. And so, if the boat *is* moving, the ball will hit the deck some distance behind the base of the mast.

According to Galileo, however, this is just factually, observably wrong. According to an observer watching from the shore, the stone will retain its horizontal speed as it falls, and will hence trace out a curved (parabolic) trajectory. But since its horizontal speed is *maintained* and just matches that of the ship, the stone still manages to strike the ship's deck just at the base of the mast – just exactly as it would do if the ship weren't moving:

> "anyone who does it will find that the experiment shows exactly the opposite of what is written; that is, it will show that the stone always falls in the same place on the ship, whether the ship is standing still or moving with any speed you please. Therefore, the same cause holding good on the earth as on the ship, nothing can be inferred about the earth's motion or rest from the stone falling always perpindicularly to the foot of the tower."

More generally, if you are locked inside a windowless compartment (such as the hold of a ship), there is no experiment you can do that would distinguish whether you are at rest or moving with uniform velocity. In the 20th century, this principle has become a cornerstone of Einstein's Special Theory of Relativity. But it was formulated quite beautifully three centuries earlier by Galileo:

> "For a final indication of the nullity of the experiments brought forth, this seems to me the place to show you a way to test them all very easily. Shut yourself up with some friend in the main cabin below decks on some large ship, and have with you there some flies, butterflies, and other small flying animals. Have a large bowl of water with some fish in it; hang up a bottle that empties drop by drop into a narrow-mouthed vessel beneath it. With the ship standing still, observe carefully how the little animals fly with equal speed to all sides of the cabin. The fish swim indifferently in all directions; the drops fall into the vessel beneath; and, in throwing something to your friend, you need throw it no more strongly in one direction than another, the distances being equal; jumping with your feet together, you pass equal spaces in every direction. When you have observed all these things carefully (though there is no doubt that when the ship is standing still everything must happen in this way), have the ship proceed with any speed you like, so long as the motion is uniform and not fluctuating this way and that. You will discover not the least change in all the effects named, nor could you tell from any of them whether the ship was moving or standing still. In jumping, you will pass on the floor the same spaces as before, nor will you make larger jumps toward the stern than toward the prow even though the ship is moving quite rapidly, despite the fact that during the time that you are in the air the floor under you will be going in a direction opposite to your jump. In throwing something to your companion, you will need no more force to get it to him whether he is in the direction of the bow or the stern, with yourself situated opposite. The droplets will fall as before into the vessel beneath without dropping toward the stern, although while the drops are in the air the ship runs many spans. The fish in their water will swim toward the front of their bowl with no more effort than toward the back, and will go with equal ease to bait placed anywhere around the edges of the bowl. Finally the butterflies and flies will continue their flights indifferently toward every side, nor will it ever happen that they are concentrated toward the stern, as if tired out from keeping up with the course of the ship, from which they will have been separated during long intervals by keeping themselves in the air. And if smoke is made by burning some incense, it will be seen going up in the form of a little cloud, remaining still and moving no more toward one side than the other. The cause of all these correspondences of effects is the fact that the ship's motion is common to all the things contained in it, and to the air also. That is why I said you should be below decks; for if this took place above in the open air, which would not follow the course of the ship, more or less noticeable differences would be seen in some of the effects noted..."

The argument – as it bears on the Copernican revolution – is that one can replace the hold of the ship with the whole Earth, and the still water outside the ship with space surrounding the Sun, and everything would stay the same. That is, contrary to the assumption of Copernicus' (and Aristarchus') opponents, we *wouldn't* have noticed it if the Earth were rotating on its axis once per day and also orbiting the Sun once per year. Notice that this represents the complete rejection of the "cosmic graph paper" dynamics sketched at the beginning of Chapter 1.

2.2.2 Telescope

In addition to his work involving the concept of inertia, Galileo contributed significantly to the Copernican revolution with a series of observations of heavenly bodies. The first such observation dates to 1604, when Galileo observed a "nova" (meaning: new star). Such phenomena had been observed before, but were typically dismissed as atmospheric (rather than celestial), on the grounds that the heavens were perfect and unchanging. Galileo, however, collected careful observations of the apparent position of the nova over the course of time and from several different locations in Europe. The constancy of the apparent position relative to the stars – that is, the lack of parallax – proved that the nova was significantly farther away even than the moon, which remember displays roughly a full degree of parallax (due either to the rotation of the heavens or the rotation of the earth, depending on whose theory you believe). So, contrary to the reveived dogma, the heavens were not after all perfect and unchanging. And if the Greeks were wrong about that, maybe they were wrong about the Earth being at the center of the universe, too?

That's a nice argument, as far as it goes. But it doesn't go very far. And anyway, it's small potatoes compared to what Galileo did five years later, in 1609:

> "... A report reached my ears that a certain Fleming had constructed a spy-glass by means of which visible objects, though very distant from the eye of the observer, were distinctly seen as if nearby. Of this truly remarkable effect several experiences were related, to which some persons gave credence while others denied them. A few days later the report was confirmed to me .... which caused me to apply myself wholeheartedly to inquire into the means by which I might arrive at the invention of a similar instrument. This I did shortly afterwards, my basis being the theory of refraction. First I prepared a tube of lead, at the ends of which I fitted two glass lenses, both plane on one side while on the other side one was spherically convex and the other concave. Then placing my eye near the concave lens I perceived objects satisfactorily large and near, for they appeared three times closer and nine times larger than when seen with the naked eye. Next I constructed another one, more accurate, which represented objects as enlarged more than sixty times. Finally, sparing neither labor nor expense, I succeeded in constructing for myself so excellent an instrument that objects seen by means of it appeared nearly one thousand times larger and over thirty times closer than when regarded with our natural vision."

Figure 2.7: The half-moon. Notice the sunlit peaks to the left of the lunar night/day boundary (the "terminator").

What did Galileo see when he became the first person to examine the heavens through a telescope? Lots of things. And all of them supported, in one way or another, the Copernican worldview:

- *New stars.* As soon as he pointed the telescope to the sky, Galileo was "overwhelmed by the vast quantity of stars" – "...more than five hundred new stars distributed among the old ones within limits of one or two degrees of arc." The Milky Way was revealed as a vast tract of individual stars, which blended and blurred together when seen with the naked eye. And some stars were revealed by the telescope to be *double* – two stars so close together that they could not be individually discriminated with the naked eye. None of this provided any sort of direct confirmation of the Copernican theory. But it was, like the nova, indirect evidence – that the ancient dogmas were based on shamefully incomplete or downright erroneous information, and were therefore to be doubted.

- *The Moon.* When Galileo turned his telescope toward the moon, he saw

  > "that the surface of the moon is not smooth, uniform, and precisely spherical as a great number of philosophers believe it (and the other heavenly bodies) to be, but is uneven, rough, and full of cavities and prominences, being not unlike the face of the earth, relieved by chains of mountains and deep valleys."

He went on to describe the lunar sunrise, as seen from afar:

> "[N]ot only are the boundaries of shadow and light in the moon seen to be uneven and wavey, but still more astonishingly many bright points appear within the darkened portion of the moon, completely divided and separated from the illuminated part and at a considerable distance from it. After a time these gradually increase in size and brightness, and an hour or two later they become joined with the rest of the lighted part which has now increased in size. Meanwhile more and more peaks shoot up as if sprouting now here, now there, lighting up within the shadowed portion; these become larger, and finally they too are united with that same luminous surface which extends further. And on the earth, before the rising of the Sun, are not the highest peaks of the mountains illuminated by the Sun's rays while the plains remain in shadow? Does not the light go on spreading while the larger central parts of these mounts are becoming illuminated? And when the Sun has finally risen, does not the illumination of plains and hills finally become one? But on the moon the variety of elevations and depressions appears to surpass in every way the roughness of the terrestrial surface..."

See Figure 2.7 for a modern image of the beautiful lunar structure one can observe with even a cheap telescope.

- *Sunspots.* After designing a filter to reduce the intensity of the light, Galileo also used his telescope to observe the Sun. He saw that it too, like the Moon, displayed an imperfect, pock-marked surface. And the marks gradually drifted across the surface, from west to east, over the course of about two weeks – implying that the Sun *rotated*, just as Copernicus required the Earth to do. See Figure 2.8 for a modern image of Sunspots.

  Orthodox thinkers tried to insist that the Sunspots were some other celestial or atmospheric phenomenon, merely passing in front of the Sun (as opposed to being blemishes inherent to it). Galileo's careful observations of the spots' motion, however, revealed that their apparent motion slowed and increased and slowed again as they crossed the visible face of the Sun, just as they should be expected to do if they are surface features of a rotating body:

  > "I have finally concluded, and believe I can demonstrate necessarily, that [the sunspots] are contiguous to the surface of the solar body, where they are continually generated and dissolved, just like clouds around the earth, and are carried around by the Sun itself, which turns on itself in a lunar month with a revolution similar [in direction] to those others of the planets, that is, from west to east around the poles of the ecliptic; which news I think will be the funeral, or rather the extremity and Last Judgment of pseudophilosophy, of which signs were already seen in the stars, in the moon, and in the Sun."

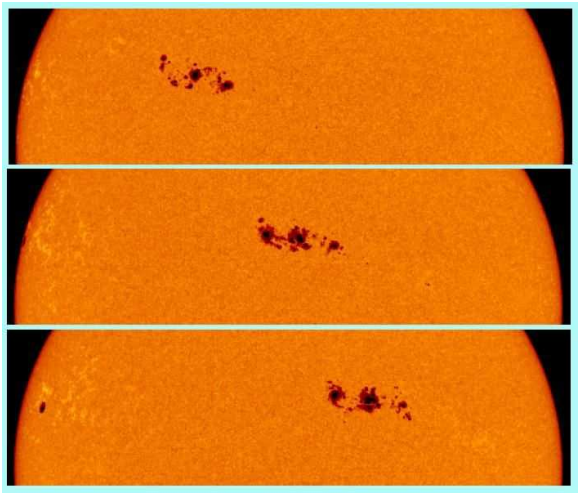

Figure 2.8: A modern image of the Sun, displaying some spots like the ones Galileo discovered.

- *Jupiter's moons.*

  > "There remains the matter which in my opinion deserves to be considered the most important of all – the disclosure of four *planets* never seen from the creation of the world up to our own time, together with the occasion of my having discovered and studied them, their arrangements, and the observations made of their movements and alterations during the past two months. I invite all astronomers to apply themselves to examine them and determine their periodic times, something which has so far been quite impossible to complete, owing to the shortness of the time."

  Subsequent observations revealed that the new planets followed Jupiter "in both its retrograde and direct movements in a constant manner" and had periods of revolution about Jupiter of, respectively: 1 day 18.5 hours; 3 days 13.3 hours; 7 days 4 hours; and 16 days 18 hours.

  Here is Galileo's description of the significance of this discovery:

  > "Here we have a fine and elegant argument for quieting the doubts of those who, while accepting with tranquil mind the revolutions of the planets about the Sun in the Copernican system, are mightily disturbed to have the moon alone revolve about the earth and accompany it in an annual rotation about the Sun. Some have believed that this structure of the universe should be rejected as impossible. But now we have not just one planet rotating about another while both run through a great orit around the Sun; our own eyes show us four stars which wander around Jupiter as does the moon around the earth, while all together trace out

Figure 2.9: A modern image of Jupiter (the big bright spot) with the four "new planets" (we now call them moons) much as Galileo would have seen them.

> a grand revolution about the Sun in the space of twelve years."

- *Phases of Venus.* Perhaps the most clear-cut and dramatic new evidence against the Ptolemaic worldview was the observation that, over the course of several months, the planet Venus displays a complete sequence of phases (just like the Earth's Moon), from "new" to slender crescent, to half, to gibbous, and finally "full." Recall that in the Ptolemaic system it was ambiguous whether Venus was closer or further to Earth than the Sun. But whichever it was, it must occupy that position for all time. Hence according to the Ptolemaic system, Venus should either display phases (roughly) on the "full" side of "half", or on the "new" side of "half" – but certainly not both. Yet it *does* display both, which proves that it must orbit

  > "around the Sun, as do also Mercury and all the other planets – something indeed believed by the Pythagoreans, Coeprnicus, Kepler, and myself, but not sensibly proved as it now is by Venus and Mercury."

It is clear that Galileo's telescopic observations shift the weight of evidence rather dramatically in favor of Copernicus.

2.2.3 Summary

The Copernican model of the solar system shatters the ancient barrier between the terrestrial and celestial realms, implying as it does that the Earth is just another planet. Galileo's application of principles extracted from terrestrial experiments to celestial phenomena represented a first step down a road that would lead Newton to a complete unification of Heaven and Earth. Galileo was also one of the first consistent champions of the idea, the basis of all modern science and everything which rests on it, that knowledge is to be gained by careful observation and experimentation, not by passive contemplation of sacred texts.

For example, in his *Dialogue Concerning the Two Chief World Systems*, in the context of the discussion (already quoted above) about the experiment with the ball falling from the mast of a moving ship, Galileo has the naive Simplicio say that, although he has never actually performed the experiment, "I certainly believe that the authorities who

Figure 2.10: A collection of modern images of Venus, taken over the course of several months in 2002.

adduced it had carefully observed it. Besides, the cause of the difference is so exactly known that there is no room for doubt." To which Salviati, Galileo's own mouthpiece in the dialogue, replies: "You yourself are sufficient evidence that those authorities may have offered it without having performed it, for you take it as certain without having done it, and commit yourself to the good faith of their dictum. Similarly it not only may be, but must be that they did the same thing too – I mean, put their faith in their predecessors, right on back without ever arriving at anyone who had performed it." (But then Salviati admits he hasn't performed it either! Nevertheless, Galileo had clearly done enough relevant experiments to generalize, and his assurance of what would happen in this particular case turns out to have been entirely warranted.)

Given all of this, it is not surprising that Galileo was considered dangerous and heretical – and hence severely restricted and persecuted – by the religious authorities of his time. For doing so much to move knowledge forward in the face of such resistance, Galileo deserves our profoundest respect and gratitude.

2.3 Kepler

Johannes Kepler (1571-1630, roughly contemporary with Galileo) made significant contributions to the growing understanding of the solar system which are in many ways complementary to Galileo's. In particular, where Galileo's discoveries are largely observational in nature, Kepler's are more technical and theoretical. But like any good theory, Kepler's ideas – at least the ones that play an important role in our story here – were grounded in observational data.

Let us thus begin by mentioning Tycho Brahe (1546-1601), a Danish astronomer who spent his life working heroically to improve the quantity and quality of observations of the positions of the stars, Sun, Moon, and planets, largely through the invention of new (and bigger and better) instruments. Characteristic of his care and ingenuity as an observer (and as manager of a large research group) is the fact that he set up two separate observatories so that independent measurements of the same positions could be carried out and compared by him. This is obviously an effective way to identify and eliminate systematic errors, and also an effective way to *measure* the uncertainty associated with a given observation. It is hard to say which is more important: the fact that Brahe's methods resulted in observations which were accurate to approximately one arc minute (1/60 of a degree, about a factor of ten better than any of his predecessors); or the fact that he genuinely *knew* that the observations were this accurate.

Near the end of his life, Brahe hired the young Johannes Kepler as an assistant, and Kepler succeeded Brahe in the post of Imperial Mathematician after Brahe's death. Kepler's major innovation was the discovery of three mathematical laws describing the precise nature of the planets' motions around the Sun. We will mainly focus on these three laws, after first discussing some preliminary points that played important roles in Kepler's thinking.

2.3.1 Preliminaries

Kepler was a supporter of the Copernican system before he discovered the three laws of planetary motion for which he is now remembered. This is partly a result of a sort of neo-Platonic mystical Sun-worship that represents another side of his personality than the side we will focus on here. For example, Kepler claims that "the Sun is the first cause of the movement of the planets and the first mover of the universe, even by reason of its own body." What is this claim for the causal primacy of the Sun based on? "[T]hese arguments are drawn from the dignity of the Sun and that of the place, and from the Sun's office of vivification and illumination in the world." Or as he put it more poetically elsewhere:

> "[The Sun] is a fountain of light, rich in fruitful heat, most fair, limpid, and pure to the sight, the source of vision, portrayer of all colors, though himself empty of color, called king of the planets for his motion, heart of the world for his power, its eye for his beauty, and which alone we should judge worthy of the Most High God, should he be pleased with a material domicile and choose a place in which to dwell with the blessed angles.... For if the Germans elect

> him as Caesar who has most power in the whole empire, who would hesitate to confer the votes of the celestial motions on him who already has been administering all other movements and changes by the benefit of the light which is entirely his possession? ... [Hence] by the highest right we return to the Sun, who alone appears, by virtue of his dignity and power, suited for this motive duty and worthy to become the home of God himself, not to say the first mover."

But he had more scientifically-respectable arguments, too. For example, early in his career, he discovered that the orbital planes of the various planets (which remember are all *near*, but not precisely *in* the plane of the ecliptic) intersect one another – not at the center of the Earth's orbit (Copernicus' "Mean Sun") but at the actual location of the Sun. This was a strong early indication that the motions of the planets were really relative to, and almost certainly caused by, the Sun:

> "Accordingly because the Sun is the node common to all the systems: therefore.... For the planets the Sun is a fixed mark, which all their revolutions regard."

As to the nature of the Sun's causal influence on the planets, Kepler convinced himself (after reading the influential 1600 book of William Gilbert) that the Sun controlled the planets through a *magnetic* influence:

> "I am much occupied with the investigation of the physical causes. My aim in this is to show that the celestial machine is to be likened not to a divine organism but rather to a clockwork..., in so far as nearly all the manifold movements are carried out by means of a single, quite simple magnetic force, as in the case of a clockwork all motions [are caused] by a simple weight."

Kepler's idea was roughly that the Sun exerted a long-range magnetic force on the planets, which pushed them around in their orbits.

This was more than a merely qualitative idea: he argued that, since the only job of this force emanating from the Sun was to move the planets in their orbits, and since the planets lie essentially in the plane of the ecliptic, the force would spread out in this two-dimensional plane – its strength therefore decreasing in proportion to the inverse of the distance from the Sun:

$$F \sim \frac{1}{r}. \tag{2.1}$$

The argument here is that, as the force spreads out from the Sun in the plane of the ecliptic, it distributes itself over circles of increasing circumference, $C = 2\pi r$. The total amount of force should be conserved, i.e., $F \times C = \text{constant}$.

Kepler also cited the rotation of the Sun (evidenced by the sunspots discovered by Galileo, though Kepler claimed to have proved, prior to 1609, and through purely theoretical arguments, that the Sun must rotate) as evidence for his idea of a magnetic force emanating from the Sun that swept the planets around in their orbits.

The crucial upshot of this (essentially wrong) idea, was an early formulation of what would later become Kepler's second law of planetary motion. Kepler argued as follows:

since the Solar force (allegedly) falls off with distance from the Sun as $1/r$, and since (according to the Ancient Greek view of Equation 1.1) the force exerted on a body is proportional to its speed, the speed of a planet in its (eccentric) orbit around the Sun should vary in inverse proportion to its distance from the Sun:

$$v \sim \frac{1}{r}. \tag{2.2}$$

Note that this applies only to the v-r relation of a given planet as it orbits the Sun. Different planets will have different sizes and weights, hence different resistances to motion, i.e., different proportionality constants in $F \sim v$. The relation between the speeds of the various planets and their respective (average) orbital radii becomes later the subject of Kepler's *third* law.

According to Equation 2.2 a planet should be moving fastest when it is closest to the Sun (perihelion) and slowest when it is farthest away (aphelion), with the above equation giving a precise description of the variation in the speed. Qualitatively, this same behavior can be produced using a circular orbit with both an eccentric and equant (and with the eccentric and equant points located on opposite sides of the circle's center). But the quantitative details of an eccentric circular orbit will be slightly different if governed by Equation 2.2 than if it were governed by the eccentric-equant construction.

In any case, it was with a commitment to the dynamical centrality of the Sun – and its alleged result, Equation 2.2 – already in hand, that Kepler began his systematic attempt to fit the data he inherited from Tycho Brahe to a theoretical system. The result of this quest was Kepler's three laws of motion, to which we now turn.

2.3.2 Kepler's Three Laws

Kepler began his careful study of the solar system by trying to understand the orbit of Mars, a problem that had been assigned to him by Brahe. Adopting first the methods of his predecessors, Kepler attempted to design an eccentric circular orbit for the planet (with, of course, the Sun at the eccentric point) and with the planet's orbital speed governed by Equation 2.2. This sounds straightforward, but is actually incredibly complex because we have no data for the position of Mars with respect to the Sun. Instead, we have data for the position of both Mars and the Sun with respect to Earth. Of course, if we knew both the angular positions *and distances* of both of these bodies, it would be a simple matter to compute their positions in space, and hence the position of one relative to the other by (vector) subtraction. But while it is easy to measure the angular position of the Sun or a planet (against the background of fixed stars), there was no means available to Kepler to determine with any accuracy or absoluteness the *distances* to these objects. Recall, for example, that the Greeks had been wrong by a factor of about 20 in determining the distance from the Earth to the Sun; this error had not yet been corrected in the 17th century!

Of course, the *relative* sizes of Earth's and Mars' orbits around the Sun were determinable by Copernicus – but only *approximately*, and Kepler's goal was no mere approximate treatment. He required extremely precise models in order to match the

extreme precision of Brahe's data. And so Kepler embarked on an almost unthinkable, decade-long "battle with Mars" in which he explored, largely through trial and error, various assumptions for the precise (relative) sizes for the circular orbits of Earth and Mars, and for their eccentricities. At some point, through a combination of ingenious methods and ferocious tenacity, he convinced himself that it could not be done: circular orbits simply could not be made consistent with Brahe's data.

It is important to appreciate that it was *close*. Kepler's best circular models reproduced the observed positions of Mars with an accuracy of better than a tenth of a degree, and would therefore have been considered perfectly adequate by any of his predecessors. But Brahe's data was accurate to a single arc minute, and so discrepancies of seven or eight arc minutes simply could not be tolerated:

> "Since the divine goodness has given to us in Tycho Brahe a most careful observer, from whose observations the error of 8' is shown in this calculation... it is right that we should with gratitude recognize and make use of this gift of God .... For if I could have treated 8' of longitude as negligible I should have already corrected sufficiently the hypothesis.... But as they could not be neglected, these 8' alone have led the way toward the complete reformation of astronomy, and have been made the subject-matter of a great part of this work."

The "complete reformation of astronomy" begins with Kepler's first law of planetary motion, stumbled upon finally toward the end of his battle with Mars:

- Planets move in elliptical orbits with the Sun at one focus.

There are two significant aspects to the first law. Most obviously, it asserts that the trajectory of a planet is not a circle at all, but is rather an elongated shape, an ellipse. With this identification, Kepler finally overthrows the ancient axiom that the motion of heavenly bodies must be circular. But equally important in the first law is that a special, defining point of the elliptical orbit – its "focus" – coincides with the location of the Sun. For this mathematically embodies Kepler's intuition that the Sun is somehow or other dynamically responsible for the motion of the planets. Let us step back and discuss the mathematics of ellipses before moving on to the other two laws.

An ellipse is most simply defined as the planar figure whose points have constant summed distances from two fixed points (the focus points). That is, for each point on the ellipse, its distance from one focus point *plus* its distance from the other focus point, is a constant – the same sum one would get for any other point on the ellipse:

$$d_1 + d_2 = \text{constant} \tag{2.3}$$

where d_1 and d_2 are as shown in Figure 2.11. Thus, one can draw an ellipse on paper by pinning the two ends of a piece of string at two points (the foci), and then moving a pencil such that the string is kept taut on both sides.

The equation satisfied by the points on the ellipse is

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{2.4}$$

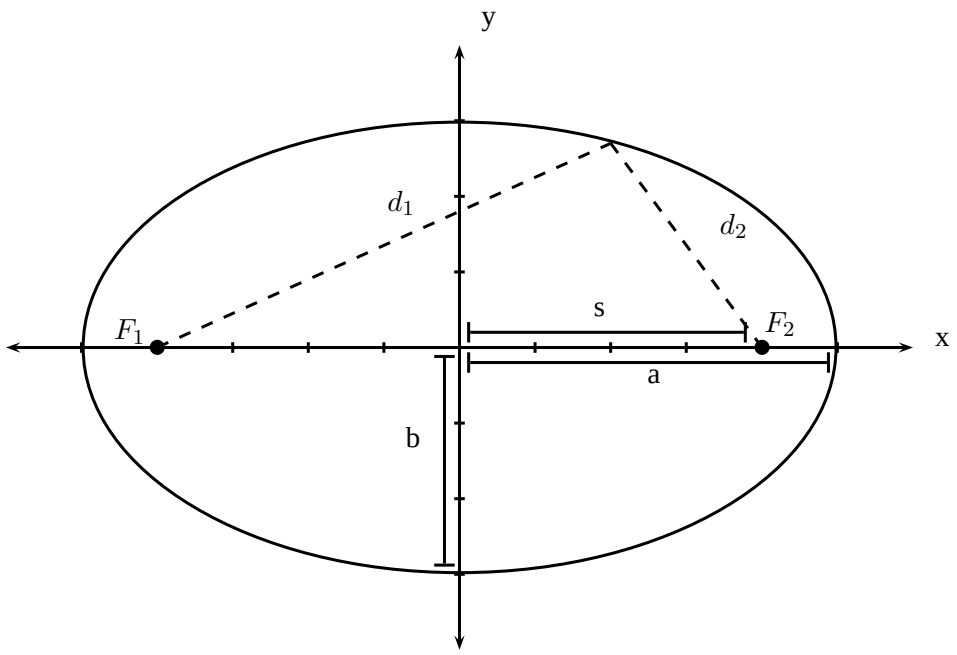

Figure 2.11: An example of an ellipse. We have here chosen the center of the ellipse as the origin of the x, y coordinate system. The two focus points are labeled F_1 and F_2. The two dotted-line distances (from a point on the ellipse to the two foci, respectively) add to the same constant value for any point on the ellipse. Here the "major axis" of the ellipse coincides with the x-axis, and the ellipse's semi-major-axis is labeled a. Its semi-minor-axis is labeled b. We have also labeled s, the distance from the center out to one of the foci along the major axis. The *eccentricity* is defined as $\epsilon = s/a = \sqrt{1 - b^2/a^2}$. Here, with $a = 5$, $b = 3$, and $s = 4$, we have $\epsilon = 4/5$ – a much larger eccentricity than the ellipses traced out by any of the planets.

which is like the equation for a circle ($x^2 + y^2 = R^2$) but with different "radii" along the x and y directions.

The eccentricity of the ellipse is a quantitative measure of its departure from circularity. The eccentricity can either be written as the ratio of the center-focus distance (s) to the semi-major axis (a), or in terms of the ratio of semi-minor-axis (b) and semi-major-axis (a):

$$\epsilon = \frac{s}{a} = \sqrt{1 - \frac{b^2}{a^2}}. \tag{2.5}$$

Finally, by placing the origin of one's coordinate system at one of the focus points instead of at the center, it can be shown that the ellipse satisfies a simple equation in polar coordinates:

$$r(\theta) = \frac{a(1 - \epsilon^2)}{1 - \epsilon \cos(\theta)} \tag{2.6}$$

where r and θ are as shown in the Figure. This expression still assumes that the major-

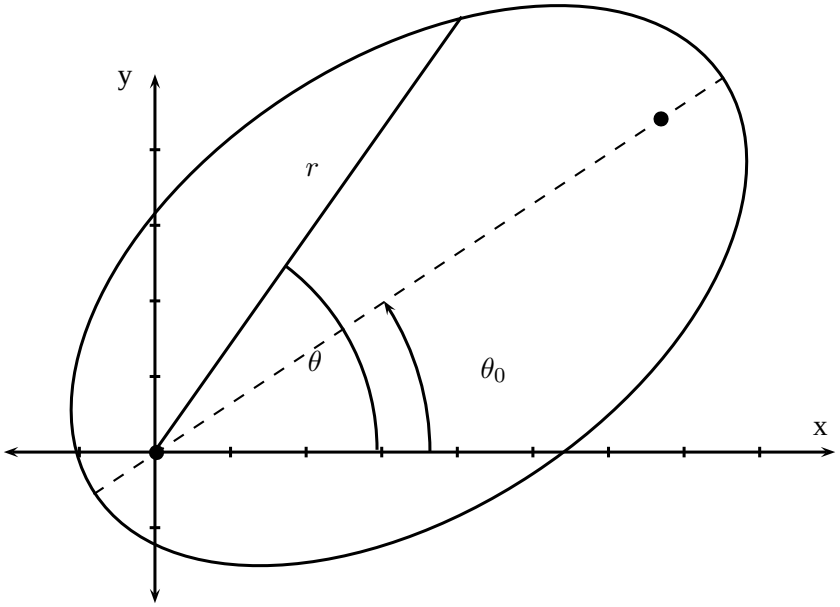

Figure 2.12: Another ellipse, now with the origin of coordinates at one of the foci, and with the major-axis rotated by θ_0 relative to the x-axis. r and θ are then related, for points on the ellipse, as in Equation 2.7

axis of the ellipse is along the x-axis, but polar coordinates make it easy to write the more general expression for an ellipse of eccentricity ϵ and semi-major-axis a whose major-axis makes an angle θ_0 with the x-axis:

$$r(\theta) = \frac{a(1-\epsilon^2)}{1-\epsilon\cos(\theta-\theta_0)}. \tag{2.7}$$

We will postpone for the Projects the proofs of all of these mathematical relations.

Kepler's second law, as it is accepted today, is a slightly modified version of his $v \sim 1/r$ rule for how a planet's speed varies as its distance from the Sun varies. The rule is formulated in a somewhat unfamiliar way: the planet is said to "sweep out equal areas in equal times." What this means is shown in Figure 2.13: as the planet moves around the Sun, the areas swept out by its coordinate vector from the Sun (e.g., the three distorted pizza-slice shapes shown) will be equal, for parts of the trajectory completed in equal times. This implies that the planet moves *faster* when it is closer to the Sun, and *slower* when it is farther from the Sun, as is evident in the Figure: the actual *distance* covered by the planet during a fixed time interval increases as the planet's distance from the Sun decreases, as is clearly necessary to keep the areas swept out in equal times equal. Qualitatively, then, Kepler's second law is the same as his original idea that the planet's speed is inversely proportional to its distance from the Sun. But as we will explore further in the Projects, the two formulations are not precisely equivalent.

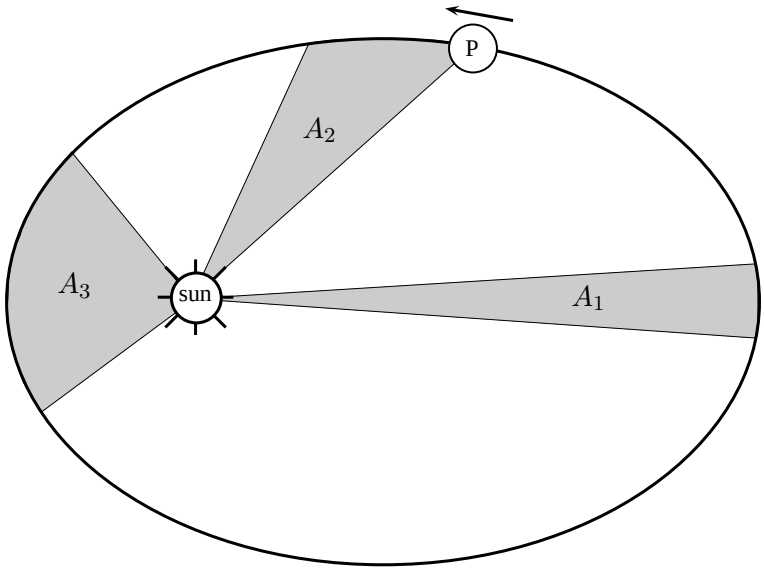

Figure 2.13: According to Kepler's second law, the areas "swept out" by a planet in equal times (e.g., the areas A_1, A_2, and A_3 shown) will be equal. This requires that the planet move a smaller distance in a given time when it is farther from the Sun, and to move a greater distance in that same time as it gets closer to the Sun – i.e., the planet's orbital speed is large when its distance from the Sun is small, and vice versa.

It is important to appreciate that Kepler's first two laws, which describe the motion of planets around the Sun, both make central reference to the position of the actual Sun: the orbits are ellipses *with the Sun at one focus*, and the speed varies such that equal areas are swept out *from the position of the Sun* in equal times. The two laws are therefore more than just descriptions of the planets' motions. They are descriptions of the planets' motions *about the Sun* – i.e., they contain strong empirical evidence in support of the Copernican concept of the Sun as the central orchestrator of the planets' motions. If the Sun weren't somehow controlling the planets from the center, but were (as in Ptolemy's theory) just another heavenly body orbiting around the Earth, it would be a bizarre and unthinkable coincidence that the motions of the planets would take on such a simple and elegant mathematical form when referenced to the position of the Sun.

Kepler's first two laws characterize the orbits of individual planets, giving precise rules for the shapes of their trajectories and the speeds with which the planets trace them out. But so far there is nothing that relates the various planets together: each planet's elliptical trajectory will have a unique size, eccentricity, and orientation; and each planet will "sweep out area" at a (uniform) rate that is different from all the other planets.

But Kepler believed deeply that there must be some hidden unity behind the apparently-unrelated orbits of the different planets. He had "discovered" early in his career that the relative sizes of the planets orbits could be "explained" by what we now consider

a bizarre numerological construction: if you nest the five perfect Platonic solids (tetrahedron, cube, octahedron, icosahedron, and dodecahedron) in a sequence of six spheres (and do it in just the right order), the relative sizes of the six spheres (approximately) match the relative sizes of the orbits of the six planets (Mercury, Venus, Earth, Mars, Jupiter, and Saturn). Of course, for a variety of reasons, scientists now consider this "explanation" to be mystical nonsense.

But much later in his career, about a decade after discovering that orbits obeying his first two laws allowed for perfect matches with Brahe's planetary data, Kepler discovered what is now known as his third law. It states that the cube of a planet's orbital radius, divided by the square of its orbital period, is a constant – the *same* constant for *all* of the planets. That is:

$$\frac{R^3}{T^2} = \text{constant}. \tag{2.8}$$

It is hard to describe the motivation that led Kepler to this discovery, since, despite arriving at a relation that would be crucial to Newton's formulation of his theory of gravitation, Kepler's quest was tinged with a sort of mysticism that is now properly regarded as unscientific. It is also difficult to overstate the kind of rapture this discovery triggered in Kepler:

> "...after I had by unceasing toil through a long period of time, using the observations of Brahe, discovered the true distances of the orbits, at last, at last, the true relation ... overcame by storm the shadows of my mind, with such fullness of agreement between my seventeen years' labor on the observations of Brahe and this present study of mine that I at first believed that I was dreaming..."

One question that may already have occured to you is: what is meant by the "orbital radius" of a planet (in Equation 2.8) since the orbits of the planets are not (according to Kepler's first law) circles? In Kepler's own formulation of the law, the R in Equation 2.8 stood for the *average* radius of the planet, though even that is not unambiguous: does this mean the average distance from the Sun of all the points on the orbit, or the distance from the Sun averaged over *time*? These will not be quite the same since (as per Kepler's second law) the planet moves *faster* through the points on its orbit that have smaller radii, and vice versa.

This is not a huge concern, since the actual orbits of the planets are not very eccentric, so the two different senses of "average" will produce almost the same R. We mention it here mostly because it is an interesting feature of the historical development of these ideas that Newton, after using Kepler's laws to arrive at his theory of gravitation, realized – by deduction from his theory – that the correct statement of Kepler's third law would have the R in Equation 2.8 being the *semi-major-axis*, i.e., the simple average of the distances of closest approach and farthest departure from the Sun.

We are finally in a position to understand how Isaac Newton arrived at his legendary theory of gravitation, the topic of Chapter 3.

Questions for Thought and Discussion:

1. Was Copernicus' theory heliocentric?

2. What are the main virtues of Copernicus' theory as against Ptolemy's? What are its weak points? If you lived in the time of Copernicus, do you think you would have been convinced by his arguments?

3. Write a short dialogue between a skeptic and a proponent of Copernicus' theory, focusing on the question of the moving Earth. Have the skeptic present the best possible arguments in favor of the idea that, if the Earth was moving, we'd have noticed it. And have the proponent give the best possible rebuttals of those arguments.

4. How might a skeptic have responded to Galileo's telescopic observations? Would reasonable disagreement with the Copernican worldview still have been possible after Galileo's work?

5. What is the difference between the "cannon ball shot straight up" type example, and the "ball dropped from the mast of a ship" type example, in the context of the debate about the moving Earth? Which type of example is more important to rebutting the claim that the Earth can't be moving because we would have noticed it? Why? What is the nature of the rebuttal?

6. Suppose Kepler had found that the orbits of the planets were, after all, eccentric circles. How might this have influenced the subsequent course of history? In particular, how would the quest to understand the causes of the planetary motions have been affected?

7. Kepler's third law states that the cube of the orbital radius is proportional to the square of the orbital period: $R^3 \sim T^2$. But the *speed* of a planet in its orbit is equal to the circumference of its orbit divided by the period: $v \sim R/T$. And so Kepler's third law can be rewritten as a statement about how the speed of the planetary orbits varies with their increasing distance from the Sun: $v^2 \sim 1/R$, i.e., $v \sim 1/\sqrt{R}$. This is different from the $v \sim 1/r$ speed-distance rule that Kepler hypothesized based on his ideas about how the Sun influenced the planets' motions (and also different from the second law as we presented it). What gives? Is there an inconsistency between the two laws?

8. Would it have been possible to identify some equivalent of Kepler's Third Law from within the Ptolemaic worldview? Why or why not?

Projects:

2.1 Use the data developed by the class last week to calculate the sizes of the orbits of all the planets (using the fact that, according to Copernicus, the radius of either Ptolemy's deferent or his epicycle should be 1 AU).

2.2 Use the Copernican model (with simple, non-eccentric circles) to derive a formula for the synodic period of a planet (i.e., the amount of time between its subsequent retrogradings) in terms of the orbital periods of the Earth and that planet. Look again at the planetary data you were given last week and see if your formula predicts the synodic period correctly.

2.3 Use the just-calculated sizes of the orbits, and also the data for the periodic times of the planets, to check Kepler's third law.

2.4 Get two lenses and try to make a telescope. Here is Galileo's description of his first attempts: "My reasoning was this. The device needs ... more than one glass.... The shape would have to be convex, ... concave, ... or bounded by parallel surfaces. But the last-named does not alter visible objects in any way; ... the concave diminishes them, and the convex, though it enlarges them, shows them indistinct and confused.... I was confined to considering what would be done by a combination of the convex and the concave. You see how this gave me what I sought."

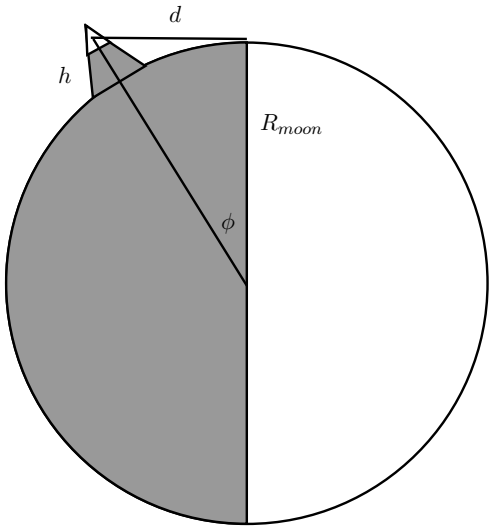

Figure 2.14: Diagram illustrating the method used by Galileo to measure the height of mountains on the moon.

2.5 Figure 2.7 shows a photograph of the moon like what Galileo would have seen through the telescope. Use it to estimate the height of the mountains on the moon, using the ideas sketched in Figure 2.14. First find a lunar mountain whose peak is lighted by the Sun, despite the mountain being on the dark side of the moon. Estimate the distance d in the Figure by measuring the mountain's apparent distance from the terminator, and comparing that to the (known) diameter of the moon. Then use the fact that $\phi \approx d/R_{moon}$. Finally, use trigonometry (or the Pythagorean theorem) to determine the height of the mountain, h.

2.6 Dot Product and Cross Product. You probably learned about the dot product of two vectors in a previous physics course: $\vec{A} \cdot \vec{B} = |\vec{A}||\vec{B}| \cos(\theta)$ where θ is the angle made by the vectors when they are drawn tail-to-tail. Argue, if you don't already know it, that the dot product can be understood as the product of the magnitude of one vector with the *component* of the second vector that is *parallel* to the first vector: $\vec{A} \cdot \vec{B} = |\vec{A}|B_{\parallel} = |\vec{B}|A_{\parallel}$. There is another way of combining two vectors called the cross product: $\vec{A} \times \vec{B}$. Its result is another vector (unlike the dot product, which gives a scalar). The *direction* of $\vec{A} \times \vec{B}$ is perpindicular to both $\vec{A}$ and $\vec{B}$ and is conventionally given by something called the right-hand-rule, which your teacher will explain in class. The *magnitude* of $\vec{A} \times \vec{B}$ is given by $|\vec{A}||\vec{B}| \sin(\theta)$ where, again, θ is the angle between $\vec{A}$ and $\vec{B}$ when they are drawn tail-to-tail. This can also be written as the product of the magnitude of one of the vectors times the component of the other vector that is *perpindicular* to the first: $|\vec{A} \times \vec{B}| = |\vec{A}|B_{\perp} = |\vec{B}|A_{\perp}$.

If the vectors $\vec{A}$ and $\vec{B}$ both lie in the x,y plane, then there are simple expressions for the dot and cross products in terms of the x and y components of the original vectors. For the dot product: $\vec{A} \cdot \vec{B} = A_x B_x + A_y B_y$. And for the cross product: $\vec{A} \times \vec{B} = A_x B_y - B_y A_x$. (Technically, this last expression gives the z-component of the *vector* $\vec{A} \times \vec{B}$, but this is the only non-zero component as long as the two vectors lie in the x,y plane.) Show that the cross product formula(s) here can all be interpreted graphically as giving the area of a parallelogram spanned by $\vec{A}$ and $\vec{B}$.

2.7 Your teacher will give you some data for the x and y coordinates of the Sun and Mars, with respect to Earth, over several years. Make a graph of the trajectory of Mars relative to Earth. Confusing and complicated, right? Now use the data to construct the position of Mars with respect to the Sun (by vector addition/subtraction) and make a new graph – of Mars' trajectory with respect to the Sun. See how the orbit simplifies! Is it an ellipse? Prove it. (This is supposed to give you an overall sense of what Kepler did – namely, by referring the motion of the planets to the Sun instead of the Earth, he found that the orbit simplified considerably and the elliptical shape of the orbit revealed itself. See the next Project to get a sense of how Kepler converted the pure angular data that was available into data for the x and y coordinates in the plane of the ecliptic.)

2.8 Kepler used an ingenious trick to convert the directly observable angular positions of the Sun and Mars into fuller knowledge of those bodies' positions in space, i.e., x and y coordinates. The trick involves recognizing that, whatever its shape, Mars' orbit around the Sun is supposed to be *periodic*, with a period of 686.9 days. Thus, every 686.9 days, Mars will be at exactly the same place. Since this is a little less than two years, Earth will be at a different point in its orbit every 686.9 days. And so one can observe, from Earth, the angular positions of the Sun and Mars every 686.9 days, and use this information to "triangulate" one's exact position with respect to each of the two bodies for each of these observations, and hence discover

the structure of Earth's orbit around the Sun by plotting its position at a number of these different times. See Figure 2.15 for the idea involved in the triangulation.

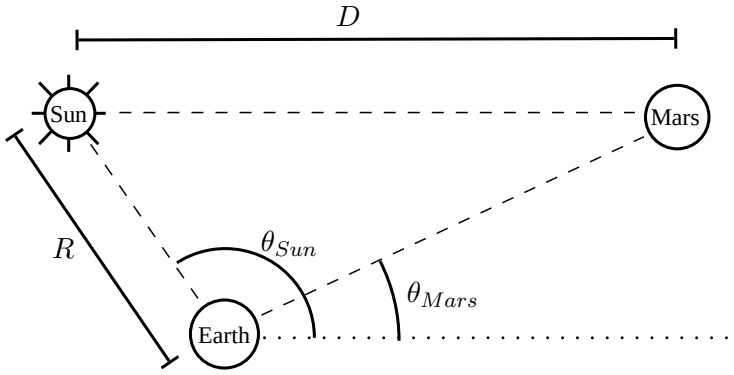

Figure 2.15: Diagram illustrating the method used by Kepler to determine the x and y coordinates of the Earth (relative to the Sun) using data for the ecliptic angles of the Sun and Mars every 686.9 days. The two ecliptic angles, θ_{Sun} and θ_{Mars}, are obtained from observation. The Earth-Sun distance R can then be inferred, in units of D, from trigonometry – as then can the x and y coordinates of the Earth's position relative to the Sun.

Your teacher will give you some data for the ecliptic angle of the Sun and Mars for a number of these points. Use the (only sort of known) distance D between the Sun and Mars as a distance unit. Then determine, for each data point, the distance R shown in the Figure. (Use the law of sines!) Then, again for each data point, use simple trigonometry to get the x and y coordinates of the Earth relative to the Sun. Make a graph of these x, y positions and you can see the trajectory of the Earth around the Sun emerge. You might try using the method you used in Project 2.7 to verify that the Earth's orbit is elliptical!

2.9 Project 2.8 showed how Kepler arrived at x, y coordinates for the Earth relative to the Sun (and so, equivalently, the Sun relative to Earth) – half of the data you were given in Project 2.7. Your assignment here is to think about how Kepler could have arrived at x, y data for the position of Mars relative to Earth – the *other* half of what you were given in Project 2.7. Here's a hint: knowing (as you now do after Project 2.8) the x and y coordinates of the Earth for all those different times, could you infer *from that* the x and y coordinates of Mars for some of those same times? What is it? How could you get positions for Mars for *other* times?

2.10 Continuing now with the data from Project 2.7, let's explore the relation between the distance of the planet from the Sun and its orbital speed. First, use Excel to compute, for all the times for which you have position data, the x and y components of Mars' velocity (v_x and v_y), its speed (v), and also its distance from the Sun (r). Is $v \sim 1/r$, as predicted by Kepler's preliminary speed rule? (The easiest way to

check this is just to multiply v and r for all the different times, and see if the value is always the same.) You should find that it is close but not exact. Maybe there is a more exact relation between (say) one of the velocity *components* and r? Check first whether the component of $\vec{v}$ that is *parallel* to $\vec{r}$ is inversely proportional to r – i.e., see if $v_\perp|\vec{r}| =$ constant. (Hint: use the expression for the dot product in terms of rectangular components given in Project 2.6.) If that doesn't work, see if there is a simple relation between r and the *perpindicular* component of $\vec{v}$ that is more exact than the preliminary $v \sim 1/r$ guess. What do you make of this? Of which law is this apparently an alternative mathematical formulation?

2.11 Consider the area "swept out" by a planet during some very small amount of time, dt. If the time is small, then the displacement of the planet, $\vec{v}\,dt$ is a good approximation to its actual trajectory (i.e., the trajectory doesn't curve much over very short times). Hence, the area swept out is a *triangle*, whose area dA is half of the parallelogram spanned by $\vec{r}$ and $\vec{v}\,dt$. Use this to help you understand what you noticed in Project 2.10.

2.12 Your teacher will give you some data for the angular coordinate of the Earth (as seen from the Sun) over time. (This is the same as the angular coordinate of the Sun as seen from Earth, plus 180° – so it's not like you have to travel to the Sun to get this data!) The idea here is to check in a relatively direct way that this data is consistent with Kepler's account of the Earth's orbit around the Sun, as specified in his first two laws of motion. To begin with, note that Kepler's second law (as re-formulated in Project 2.10) can be reformulated (again) in terms of the planet's angular velocity using $\omega = v_\perp/r$. What relationship does this imply between the planet's angular velocity and its radius r? Now combine this relation with the polar coordinates equation for an ellipse to get a relation between the angular velocity ω and the angle θ. Use the given data to calculate and then graph ω vs θ, and then do a curve-fit with the formula you developed. What is the eccentricity ϵ of the Earth's orbit? Could you achieve as good a fit to the data using an eccentric circular orbit or a circular orbit with an equant (or both at the same time)? To answer that, recall and/or reference the appropriate Projects from Chapter 1 – or do a second curve fit right now!

2.13 Figure 2.16 shows some data for the apparent positions (relative to Jupiter itself) of the four moons of Jupiter observed by Galileo, over the course of about a month. Use the graph to determine the periods of the orbits of the four moons, and also the relative *sizes* of their orbits. Can you find any mathematical relationship between the sizes and periods of the orbits, like Kepler found for the sizes and periods of the planets' orbits around the Sun?

2.14 Work out all the math connecting the various equations from the text for ellipses.

2.15 Kepler's 2nd law should apply as well to projectiles. Take the case of a ball dropped from the top of a tower (attached to the rotating earth). The ball follows a curved trajectory (as seen from an inertial frame in which the earth is rotating) to which

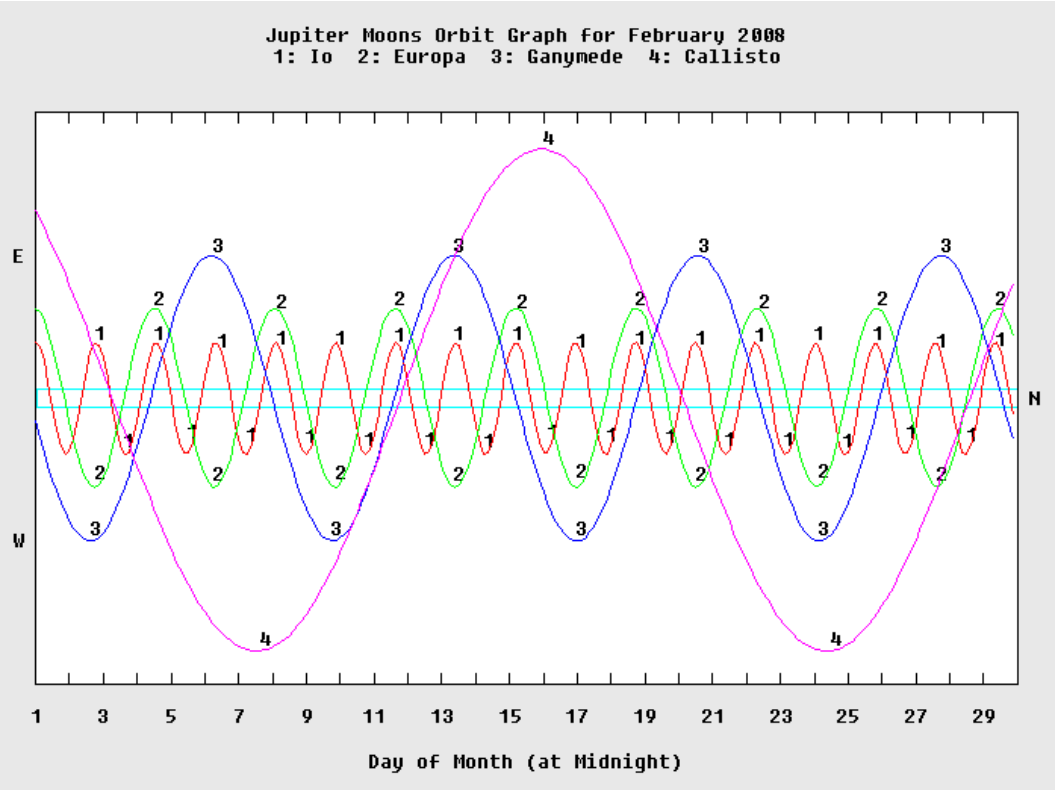

Figure 2.16: A graph showing the apparent angular (ecliptic) positions of the four moons of Jupiter (relative to Jupiter itself) over time.

one can apply Kepler's equal areas rule. Question: is the perpindicular component of the object's velocity (what an observer on the tower would refer to as the "horizontal velocity") a constant as the ball falls? By how much does it change during the fall? Approximate the distance away from the base of the tower that the ball will land due to this discrepancy.

Chapter 3

Newton's Theory of Gravitation

Isaac Newton (1643-1727) presented his theory of universal gravitation in one master stroke: his 1687 *Mathematical Principles of Natural Philosophy* or *Principia* for short. The book develops side-by-side not only his theory of gravitation, but the whole edifice of Newtonian mechanics that any reader of this book should already be reasonably familiar with. This includes the first consistent statement of rectilinear inertia (the first law of motion), $\vec{F} = m\vec{a}$ (the second law of motion), the principle of action and reaction (the third law of motion), the concept of momentum and its conservation, etc.

Newton's parallel discovery of the laws of mechanics and the theory of gravitation (not to mention calculus, which he invented along the way to solve some technical problems that arose!) is an almost unimaginable achievement, which alone ranks Newton as almost certainly the greatest scientist of all time. Because of the sheer immensity of Newton's achievement, we here (and in subsequent chapters) depart from the largely historical structure followed in the first two chapters. In particular, we are assuming that the reader has already mastered Newtonian mechanics as it is covered in standard college introductory physics courses. We will also postpone, until later chapters, discussion of several difficult and technical pieces of evidence Newton put forward as part of his initial case in support of his gravitation theory. We will cover these in time, but will treat them (somewhat a-historically) as applications of an already-established theory.

That said, we will attempt here to give a historically sensitive (if not precisely accurate) sketch of the origins, central content, and implications of Newton's theory.

The theory, of course, says that the force holding the planets (and moons) in their orbits is a *gravitational* force. Let us briefly elaborate what this means and in what ways it was novel.

First, the idea of "gravity" was not new to Newton, though he did reconceive it somewhat. Earlier thinkers had understood that many objects naturally fall when impediments are removed, and had attributed this to "gravity." But as we have seen they tended to think of this as a kind of intrinsic "heaviness" that expressed the object's "desire" to reach its "natural place" at the center of the universe or earth. Newton was the first to grasp fully that this "heaviness" is not innate to heavy objects, but is rather the expression of an *external force* exerted on them *by the earth*. This clearly required a rejection of the ancient doctrine of "natural place," a clear formulation of the principle

of rectilinear inertia, and its consistent application to both objects near the surface of the earth and heavenly bodies.

But despite this important reconception of "gravity," the main novelty of Newton's theory lies not in its advocacy of gravity-as-external-force (rather than gravity-as-innate-tendency), but rather in the theory's *extension* of the concept gravity from terrestrial objects to *all* objects, on earth and in the heavens. There are actually two aspects to this extension. First, Newton's theory involves the extension of the Earth's gravitational influence far beyond the surface of the Earth – to the Moon and beyond. And likewise, Newton attributes this same, long-range gravitational attraction to the heavenly bodies: the Moon, the Sun, the planets, comets, etc.

To further illuminate what was novel about Newton's proposal, let us see specifically how Newton's theory contrasts to the ideas of his immediate predecessors.

Copernicus for example, had speculated that gravity existed not just for the Earth, but for other heavenly bodies as well, and believed that this could account for all these bodies' apparent sphericality:

> "I myself think that gravity or heaviness is nothing except a certain natural appetancy implanted in the parts by the divine providence of the universal Artisan, in order that they should unite with one another in their oneness and wholeness and come together in the form of a globe. It is believable that this affect is present in the sun, moon, and the other bright planets and that through its efficacy they remain in the spherical figure in which they are visible, though they nevertheless accomplish their circular movements in many different ways."

But Copernicus had no concept of a "gravity" that extended over long distances and allowed separate heavenly bodies to attract one another, and hence completely missed the role of gravity in causing not only the (spherical) shape of the planets, but the (roughly circular) shape of their *orbits*. Rather, he just accepted their (he thought) circular motions as "natural" and hence in need of no causal explanation via external forces applied by other bodies.

Some important steps towards Newton's theory of gravity were taken by Kepler, who remember believed strongly (too strongly, given the evidence he possessed) that the motion of the planets was governed by forces exerted by the Sun. Kepler had specifically posited (under the title *anima motrix*) that the rotating Sun sent "spokes" out into the plane of the ecliptic, with the rotating spokes exerting the tangential force needed to keep the planets moving in roughly circular orbits. He then speculated that the would-be circular orbits were distorted into ellipses by an alternating *magnetic* attraction and repulsion, as the magnetic Earth alternately presented its North and South poles toward the Sun during its yearly orbit (and, presumably, similarly for the other planets). Despite its errors, Kepler's ideas were important because they represented the first suggestion that the planets' orbits were genuinely caused by some kind of force or influence exerted on them by the Sun.

Another great pre-Newtonian thinker, the philosopher and armchair scientist Rene Descartes, had influentially speculated that the motions of the planets could be explained

by a great cosmic whirlpool (made of some unseen fluid ether) which carried the planets in their orbits around the Sun. Newton devoted considerable space in the *Principia* (whose title is essentially the same as Descartes' own earlier book on similar topics) to proving that the whirlpool theory is untenable, since it makes predictions which contradict one another and also observational data about the motions of planets, moons, and comets. In a summary statement, Newton writes:

> "I have tried to investigate the properties of vortices in order to test whether the celestial phenomena could be explained in any way by vortices. For it is a phenomenon that the periodic times of the secondary planets that revolve about Jupiter are as the 3/2 powers of the distances from the center of Jupiter; and the same rule applies to the planets that revolve about the sun. Moreover, these rules apply to both the primary and the secondary planets [i.e., moons] very exactly, as far as astronomical observations have shown up to now. And thus if those planets are carried along by vortices revolving about Jupiter and the sun, the vortices will also have to revolve according to the same law. But the periodic times of the parts of a vortex turned out [in an earlier, fluid-mechanics analysis of vortex motion] to be in the squared ratio of the distances from the center of motion, and that ratio cannot be decreased and reduced to the 3/2 power, unless either the matter of the vortex is the more fluid the further it is form the center, or the resistance arising from a deficiency in the slipperiness of the parts of the fluid ... is increased in a greater ratio than the ratio in which the velocity is increased. Yet neither of these seems reasonable. .... It is therefore up to philosophers to see how that phenomenon of the 3/2 power can be explained by vortices."

One of the crucial similarities between Kepler's and Descartes' ideas is their failure to grasp that the curved trajectories of the planets require and imply a *centripetal* force, as opposed to a *tangential* force or an underlying tangential motion by which the planets were swept along. This key idea from Newtonian mechanics – that uniform circular motion involves a force directed not along the direction of motion, but toward the center of the circle – represents the final renunciation of the old Greek ideas about motion, e.g., "natural place" and the affiliated idea that forces produce velocity, rather than acceleration.

With all of this as historical background, let us turn to the developments that served as proximate causes for Newton's full gravitational theory, as presented in the *Principia*.

3.1 Newton's Theory of Gravitation

Kepler's accurate descriptions of the trajectories of the planets around the Sun had raised the obvious next question: what causes the planets to move this way? Kepler's own speculations about magnetic forces were arbitrary, qualitative, and unconvincing. But by the 1660s, several thinkers were hot on the trail. In that decade, Newton and Christian Huygens independently discovered the "law of circular motion" – now understood as the

claim that uniform motion at speed v around a circle of radius R involves a *centripetal acceleration* whose magnitude is given by

$$a_c = \frac{v^2}{R}. \tag{3.1}$$

In the 1660s, however, both Newton and Huygens suffered from some (perhaps familiar) confusion between centripetal and centrifugal forces. Perhaps thinking about what one experiences when one whirls a rock on a string in circles about one's head, they conceptualized v^2/R not as describing the inward acceleration or force, but rather as quantifying a circularly-moving object's "endeavor" to recede from the center – something one can feel directly in whirling a rock on a string.

Armed with this correct, but still misunderstood, formula, several thinkers hit on the idea of using it to analyze the motions of the planets – in particular, how the different planets' centrifugal "endeavors" vary with their distance from the Sun. The computation is readily made by combining the law of circular acceleration with Kepler's third law. This latter, recall, states that the planets' orbital periods vary as the 3/2 power of their orbital radii, i.e., the *squared* orbital periods of the planets are proportional to their *cubed* orbital radii:

$$T^2 \sim R^3 \tag{3.2}$$

where here we approximate the planetary orbits as *circles* and hence take R as the *radius* of the orbits.

Since the speed of a planet in its orbit is proportional to its orbital radius divided by the period

$$v \sim R/T \tag{3.3}$$

Kepler's third law can be re-written as

$$\frac{v^2}{R} \sim \frac{1}{R^2}. \tag{3.4}$$

But the left hand side is immediately recognizable as the "centrifugal force" of a given planet. So Kepler's third law implies that these forces fall off as the *inverse square* of the planets' distance from the Sun.

Some argument like this had convinced several thinkers that an "inverse square law" force might be involved in the motion of the planets around the Sun. But, lacking Newtonian mechanics (and in particular a correct concept of centripetal force/acceleration), nobody had been able to really do anything with this pregnant hypothesis.

Newton had been working on just this problem when, in 1684, his friend the astronomer Edmond Halley (of the comet) approached him with a challenge. Halley had heard that Robert Hooke (of the spring force law) and some other scientists were trying to work out the precise trajectory of a planet moving under a central inverse square force. The idea was that, if it could be proved mathematically that the resulting trajectory was an ellipse, this would provide strong support for the idea that the Sun exerted an inverse square force on the planets. Newton claimed that he had solved this problem earlier, but was (he claimed) unable to locate his notes. Whether Newton had actually solved

the problem by 1684 is, however, immaterial. In either case, the news from Halley that others were working on ideas similar to his own was perceived as a threat, and caused Newton – prodded and supported by Halley – to redouble his efforts at producing and/or writing up and publishing his work. The result, after several years of furious work, was the *Principia*.

The proof that an inverse square force produces elliptical orbits (or, really, more generally, conic sections – circles, ellipses, parabolae, or hyperbolae) is mathematically quite difficult, requiring the solution of a complicated differential equation. We will postpone the proof until Chapter 6, where we will use a computer to simulate the motion of a planet in an inverse square force field, and then prove that the resulting trajectory is elliptical.

But despite being a central piece of evidence for Newton's theory, his proof that an inverse square force produces trajectories that match those actually followed by the planets is only one part of his theory. For this proof alone would only suggest that some force with this character is operative in the solar system; it would in no way suggest that the force is *gravitational*, i.e., the same sort of force that pulls heavy objects down toward the earth. It would also not provide any direct evidence that the force in question is exerted on the planets *by and toward the Sun*, rather than being centered on some random mathematical point near the center of the solar system.

Let us thus focus the remainder of the present discussion of Newton's theory on these two points.

3.2 Newton and Kepler's Area Law

One of the crucial arguments Newton developed in support of his theory was a proof that the forces exerted on the planets are directed *toward the Sun*, and hence (like the gravitational forces exerted by the Earth) are associated with a massive body and not some mere mathematical point. Let us develop this argument using contemporary concepts and notation.

Here is a simple theorem in Newtonian mechanics. Consider a particle of mass m moving with velocity $\vec{v}$. Its momentum is then $\vec{p} = m\vec{v}$. Let us define a new quantity, its *angular momentum*, as follows:

$$\vec{L} = \vec{r} \times \vec{p} \tag{3.5}$$

where $\vec{r}$ is the position vector for the particle and the "$\times$" in the equation denotes the vector cross product. Note that just as the "angular velocity" we defined and used in previous chapters is a kind of rotational analog of (regular old) velocity, so the "angular momentum" is a kind of rotational analog of (regular old) momentum. This is something we will explore more systematically in subsequent chapters.

Let us also define the *torque* (a rotational analog of *force*) on the particle as follows:

$$\vec{\tau} = \vec{r} \times \vec{F} \tag{3.6}$$

where $\vec{F}$ is a force acting on the particle.

All this talk of "rotational analogs" suggests there might be some connection between torque and angular momentum. In particular, from this formulation of Newton's second law

$$\vec{F} = \frac{d\vec{p}}{dt} \tag{3.7}$$

(where here $\vec{F}$ represents the net force, the sum of all the individual forces that act) one might guess the "analogous" rotational formula:

$$\vec{\tau} = \frac{d\vec{L}}{dt}. \tag{3.8}$$

(where $\vec{\tau}$ is the net torque on the particle, i.e., the sum of the torques produced by all of the individual forces that act).

This turns out to be precisely right. It is not a new postulate, but a theorem based on the earlier definitions of torque and angular momentum:

$$\frac{d\vec{L}}{dt} = \frac{d}{dt}\left(\vec{r}\times\vec{p}\right) = \frac{d\vec{r}}{dt}\times\vec{p} + \vec{r}\times\frac{d\vec{p}}{dt} = \vec{r}\times\vec{F} = \vec{\tau} \tag{3.9}$$

where we have used the facts that $d\vec{r}/dt = \vec{v}$ and $\vec{v}\times\vec{p} = 0$ since the two vectors are necessarily parallel.

We will actually need only one special case of the theorem, namely, that if the angular momentum doesn't change in time, i.e., if it is a *constant* in time, the net torque must vanish.

You maybe already realized (in doing some of the Projects for Chapter 2) that Kepler's second law is mathematically equivalent to the statement that

$$\vec{r}\times\vec{v} = \text{constant}. \tag{3.10}$$

(See also the figure and adjoining caption.) Since the left hand side here differs from a planet's angular momentum only by a factor of the mass m of the planet, it should be clear that Kepler's second law is also equivalent to the statement that, for each planet,

$$\vec{L} = \text{constant}. \tag{3.11}$$

And this, according to the theorem, implies that the torque exerted on the planet (with the origin taken as the Sun) must vanish. What could ensure this? In general, both the position vector $\vec{r}$ and the force vector $\vec{F}$ will vary in some complicated way as the planet orbits the Sun. What could ensure that the torque, their cross product, always vanishes? Only that $\vec{F}$ is directed in precisely the opposite direction as $\vec{r}$ – i.e., only if the force is directed *precisely toward the Sun.*

Newton noted that not only do the planets sweep out equal areas in equal times with respect to the Sun, but the Earth's moon sweeps out equal areas in equal times with respect to the (center of the) Earth, as do Jupiter's and Saturn's several moons with respect to those planets. Thus, wherever we have one body orbiting another in the heavens, the orbiting body sweeps out area (relative to the central body) at a constant

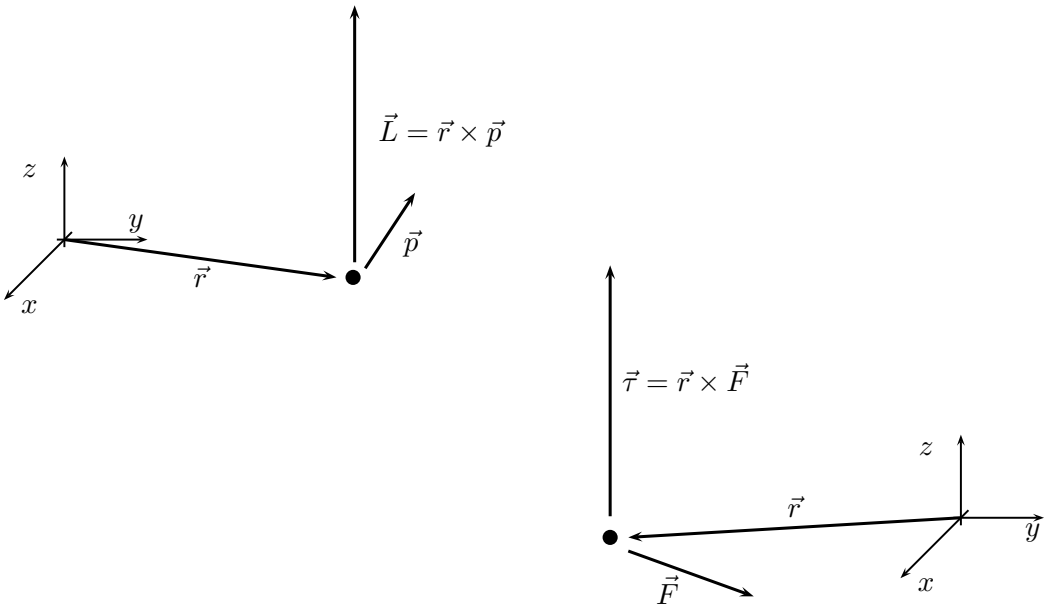

Figure 3.1: Illustrations of the vector relationships defined in Equations 3.5 and 3.6. The angular momentum $\vec{L}$ of a particle is the vector cross product of its position $\vec{r}$ and momentum $\vec{p}$. The direction of $\vec{L}$ is found by the right hand rule: orient your right hand so your (straightened) fingers point in the direction of $\vec{r}$, then rotate your hand so your fingers can (non-painfully) bend to point in the direction of $\vec{p}$. Your thumb will then point in the direction of $\vec{L}$. Note that the cross product of two vectors is orthogonal to each of the two vectors – so the right hand rule is really just a "rule of thumb" to decide between the two possible directions that are orthogonal to the plane defined by the two original vectors. In the example shown, the idea is that $\vec{r}$ and $\vec{p}$ are in the x-y plane, so their cross product is in the z-direction. The second picture just shows an example of the relationship expressed by Equation 3.6: the torque $\vec{\tau}$ produced by a force $\vec{F}$ exerted on a particle at position $\vec{r}$ is given by the cross product. The direction of the torque is orthogonal to the plane spanned by $\vec{r}$ and $\vec{F}$, as picked out by the right hand rule. It is also useful to use a right hand rule to think qualitatively about the meaning of the vectors $\vec{L}$ and $\vec{\tau}$ which can admittedly be a bit puzzling. If you point the thumb of your right hand in the direction of the vector $\vec{L}$, your fingers will naturally "curl" in a certain direction in the plane perpindicular to the vector. This direction (in which your fingers curl) is a good way to think about the meaning of angular momentum. For example, in the first figure, the angular momentum vector $\vec{L}$ is in the positive z-direction; if you point your right thumb parallel to this vector, your fingers indicate a counter-clockwise orbital motion in the x-y plane, which is precisely what the particle is doing. The meaning of the direction of the torque vector $\vec{\tau}$ can be understood in the same way. In the example shown, if you orient your right thumb parallel to $\vec{\tau}$, your fingers naturally curl to indicate (again) a counter-clockwise motion in the x-y plane. This is precisely what a torque in the z-direction means: the force that is operating is so-as to tend to turn the particle clockwise in the x-y plane.

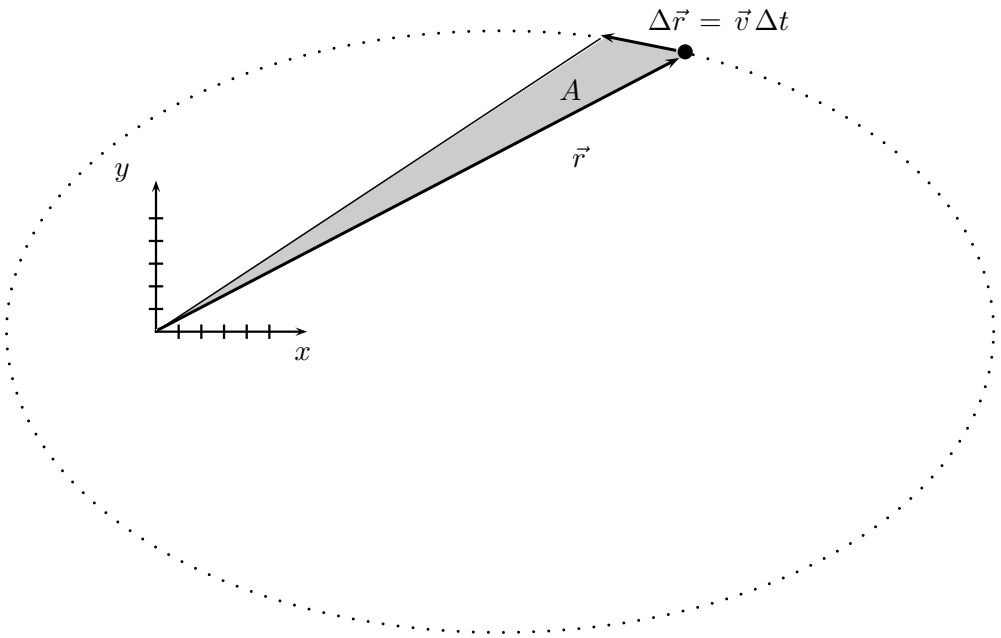

Figure 3.2: The vector $\vec{r}$ represents the position of a planet, with the Sun taken as the origin. $\Delta\vec{r}$ represents its displacement during a short time period Δt. For sufficiently small Δt, the average velocity during the time interval is close to the instantaneous velocity at the moment shown, so $\Delta\vec{r} = \vec{v}\,\Delta t$. The cross-product $\vec{r} \times \Delta\vec{r}$ is a vector pointing, by the right hand rule, out of the page. Its magnitude is the area of the parallelogram spanned by $\vec{r}$ and $\Delta\vec{r}$, whose area is just twice the shaded area A. Thus, $|\vec{r} \times \vec{v}| = 2A/\Delta t$. And so the magnitude of the planet's angular momentum is $|\vec{L}| = |\vec{r} \times \vec{p}| = m|\vec{r} \times \vec{v}| = 2mA/\Delta t$. Hence, the area swept out by the planet per unit time – $A/\Delta t$ – is just a constant ($|\vec{L}|/2m$) involving the mass of the planet and the magnitude of the planet's angular momentum. And since the mass of the planet doesn't change in time, the constancy in time of the planet's angular momentum is revealed to be mathematically equivalent to Kepler's second law, that the planet "sweeps out area" at a constant rate.

rate. And so, according to the above argument, in all of these cases, the force causing the orbit is directed precisely toward the central body that is being orbited around. It then seems irresistable to conclude that the orbits are produced by forces *exerted by* that central body.

Kepler had been right that the Sun controlled the motions of the planets (though wrong about the nature of the force). But where Kepler had argued for this claim on circumstantial evidence, Newton had produced a rigorous mathematical and mechanical proof. This convinced him he was on the right track – that the planetary motions were the result of a force exerted on the planets by the Sun.

Now: what suggested to Newton that the force exerted by the Sun was a *gravitational* force, i.e., the same sort of force that the Earth exerts on familiar, heavy terrestrial objects?

3.3 The Apple and the Moon

The story of "Newton's apple" is legendary. He was sitting under an apple tree when a falling apple (which by some accounts hit him in the head) inspired him to conceive of Universal Gravitation. What was going on in his mind that allowed him to make this connection?

First of all, he wasn't thinking about apples but about the moon. The moon orbits the earth in a roughly circular orbit whose radius is about 60 earth radii, and with a period of some 27.3 days. The centripetal acceleration of an object moving in uniform circular motion is given by

$$a_c = \frac{v^2}{R} \tag{3.12}$$

where v is its speed and R the radius of its orbit. Since the speed is related to the period through $v = 2\pi R/T$, this can be re-written

$$a_c = \frac{4\pi^2 R}{T^2}. \tag{3.13}$$

Plugging in the known values for the radius and period of the moon's orbit gives an acceleration for the moon

$$a_{moon} = 20,340\,\mathrm{km/day^2} = 0.0027\,\mathrm{m/s^2} \tag{3.14}$$

which by this time Newton clearly understood was directed toward the Earth.

The moon's acceleration is very small compared to the gravitational acceleration of heavy objects near the surface of the earth, the familiar

$$g = 9.8\ \mathrm{m/s^2}. \tag{3.15}$$

Yet still, small or not, according to Newton's second law, $\vec{F} = m\vec{a}$, this acceleration must be caused by some force exerted on the moon, and it must be a force directed toward the center of the moon's orbit, i.e., toward the Earth. But what kind of force could this be?

The falling of the apple must have triggered in Newton's mind the thought: maybe the force exerted on the Moon is the same type of force that explains the apple's acceleration toward the Earth. Maybe it's a *gravitational* force.

Since it had already been speculated by some that a force which decreased in strength as the inverse square of the distance from the Sun could account for the orbits of the planets, it immediately occured to Newton to check how the Earth's gravitational force would have to decrease with distance, if it really were the cause of both the apple's and moon's accelerations. The simplest thing is just to compute the ratio of the two accelerations:

$$\frac{a_{moon}}{g} = \frac{.0027 \text{ m/s}^2}{9.81 \text{ m/s}^2} = .000275 \approx \frac{1}{3600} = \frac{1}{(60)^2} \tag{3.16}$$

which Newton would immediately have recognized as significant, because 60 is precisely the ratio of the *distances* of the moon and apple, respectively, from the center of the earth – the moon being at a distance of 60 Earth radii, and the apple being at *one* Earth radius! So the acceleration of the moon toward Earth is less than the acceleration of the apple toward Earth by precisely the *inverse square* of their respective distances from Earth's center.

This simple numerical computation suggests several connections. First, it supports the idea that the earth's gravitational influence stretches far beyond the surface of the earth, to at least the orbit of the moon:

> "Therefore, since both forces ... are directed toward the center of the earth and are similar to each other ... they will ... have the same cause. And therefore that force by which the moon is kept in its orbit is the very one that we generally call gravity."

Accepting that involves accepting that the earth's gravitational influence falls off as the inverse square of the distance. And *that* means the Earth's gravitational influence falls off with distance in precisely the same mathematical fashion as the (as-yet-unidentified) force that the Sun exerts on the planets. It is then practically irresistable to conclude that the two forces (and in addition the forces exerted by Jupiter and Saturn on their moons) are one and the same: gravity.

Of course, all of this would require more careful and rigorous analysis. But you can see how the simple comparison of the moon's and apple's accelerations suggests the idea of a long-range, universal, inverse-square-law gravitational force.

3.4 Further Evidence for the Theory of Universal Gravitation

So far we have covered Newton's arguments that: (a) orbiting bodies are acted on by forces directed toward the central body the orbit is around, in accordance with Kepler's second law; (b) these forces (in the case of the planets and the moons of Jupiter and Saturn) must fall off in strength as the inverse square of the distance from the central

body, in accordance with Kepler's third law; (c) the Earth's gravitational influence extends (at least) to the Moon and reduces its strength in proportion to the inverse square of the distance; and therefore, probably, (d) the forces responsible for the orbits of the planets and Jupiter's and Saturn's moons are *gravitational*.

We have also mentioned, but not yet really discussed, Newton's mathematical proof that a central inverse square law force – and *only* a central inverse square law force – will produce elliptical trajectories for orbiting bodies, in accordance with Kepler's first law.

Strictly speaking, the conclusion Newton was leaning toward was not merely that the forces on the planets and moons were gravitational in nature, but that gravitation was a *universal* phenomenon: *all* bodies exert inverse-square gravitational forces on all other bodies. In order to move this claim from a probability to a certainty, Newton carefully worked out the mathematical implications of this idea and compared them to observational data. He found stunning confirmations everywhere he looked.

For example, according to the hypothesis of universal gravitation, not only should the Moon be attracted to the Earth by a gravitational force, but the Earth should be in turn gravitated toward the Moon. Indeed, according to Newtonian mechanics, the Moon does not exactly orbit around the Earth, so much as the Earth and Moon jointly orbit around their mutual center of mass. This should produce small monthly deviations of the Earth's orbit from its Keplerian ellipse. Such deviations were in fact confirmed by detailed astronomical observation.

More interestingly, the Moon's gravitational tug on the Earth falls off with distance, so the material on the side of the Earth facing the Moon is pulled toward the Moon slightly more than average, while the material on the far side of the Earth is pulled toward the Moon slightly less than average. This causes the waters to rise up in altitude on these opposite sides of the Earth. And since the Earth rotates on its axis daily, a given observer will notice the seas rising and falling twice per day. Newton's theory of universal gravitation thus produced the first correct explanation of the tides! (We'll discuss the tides in more detail in Chapter 5.)

Newton's theory also predicted that the planets should exert gravitational influences on one another. Since Jupiter and Saturn are the heaviest planets and their orbits are adjacent to one another, Newton suggested that astronomers of his time look for small perturbations in the orbits of these two planets around the time Jupiter passes "under" Saturn. (The inverse square force will be greatest around this time.) Such perturbations were eventually observed, as were similar perturbations on the other planets, including Earth. You can explore some of these effects in more detail in Chapter 6.

Newton's theory also predicted that objects could move in parabola- or hyperbola-shaped trajectories about the Sun. Careful analysis of the motion of certain comets eventually revealed that they moved in precisely these ways, with other comets moving in elliptical but extremely eccentric orbits. Halley's comet, for example, has an elliptical orbit which brings it into the inner solar system with a period of approximately 75 years. Its orbit is very eccentric, coming all the way in past the orbit of Earth to a distance from the Sun of about half an AU, and then returning again to about 35 AU, far beyond the orbit of Saturn. Notably, the comet's small deviations from a perfectly elliptical orbit

and a perfectly periodic motion can be precisely explained by the gravitational forces exerted on the comet by the planets. Careful analysis of the comet's previous trajectories including calculations of the perturbing effects of Jupiter and Saturn (based on Newton's theory) allowed scientists to predict – within a matter of weeks – the subsequent returns of Halley's comet. The accuracy of these predictions was hailed as a major piece of evidence in support of Newton's ideas.

There are several other pieces of evidence as well, some of which we will cover in subsequent chapters. To summarize, though, not only does Newton's theory account, in precise mathematical detail, for the motions of the moons and planets as these were described by Kepler; it also predicts and explains a number of small deviations from Keplerian orbits which are produced by the relatively small gravitational forces exerted by the moons and planets on each other, as well as accounting naturally for certain previously unexplained processes such as the tides. As Newton summarizes:

> "All the planets are heavy toward one another.... And hence Jupiter and Saturn near conjunction, by attracting each other, sensibly perturb each other's motions, the sun perturbs the lunar motions, and the sun and moon perturb our sea..."

Gravity, according to Newton's theory, is a universal phenomenon in which every massive body attracts – is heavy toward – every other massive body. His theory predicts and explains these effects with precise, mathematical rigor. One could not reasonably ask for a more conclusive array of evidence in support of a theory.

3.5 The precise form of the gravitational force

Having surveyed the evidence for Newton's theory and discussed one quantitative feature of the force law (that the force varies as the inverse square of the distance between the two gravitating bodies), let us here develop a more precise statement of the basic equation describing the gravitational force. We will here consider the gravitational force between two point masses, and postpone until the following chapter a discussion of the gravitational forces produced by an extended object (composed of many individually-gravitating point masses).

To begin with, recall the crucial fact identified by Galileo: in the absense of appreciable air resistance (i.e., when the gravitational force is the only one acting), projectiles move with a constant downward acceleration *independent of their mass*. For example, a dropped baseball and a dropped bowling ball will both, despite their different masses, accelerate toward the ground at

$$g = 9.8 \text{ m/s}^2. \tag{3.17}$$

This implies, according to Newton's second law $\vec{F} = m\vec{a}$, that the gravitational force exerted on an object is proportional to its mass – i.e., different objects with the same mass should also have the same *weight*.

Newton undertook his own precise experimental test of this principle, by comparing the periods of pendulums made of different substances:

> "Others have long since observed that the falling of all heavy bodies toward the earth ... takes place in equal times, and it is possible to discern that equality of the times, to a very high degree of accuracy, by using pendulums. I have tested this with gold, silver, lead, glass, sand, common salt, wood, water, and wheat. I got two wooden boxes, round and equal. I filled one of them with wood, and I suspended the same weight of gold (as exactly as I could) in the center of oscillation of the other. The boxes, hanging by equal eleven-foot cords, made pendulums exactly like each other with respect to their weight, shape, and air resistance. Then, when placed close to each other [and set into vibration], they kept swiging back and forth together with equal oscillations for a very long time. Accordingly, the amount of matter in the gold ... was to the amount of matter in the wood as the action of the motive force upon all the gold to the action of the motive force upon all the wood – that is, as the weight of one to the weight of the other."

In other words, the quantity of matter (the mass) is proportional to the weight (the gravitational force).

Newton also cites the identical rates at which Jupiter and its moons are accelerated toward the Sun by its gravitational attraction:

> "Further, that the weights of Jupiter and its satellites toward the sun are proportional to the quantities of their matter is evident from the extremely regular motion of the satellites..."

i.e., the fact that the moons (whose masses are very different from that of the planet Jupiter) manage to stay with Jupiter as they all orbit (and in so doing accelerate toward) the Sun.

So there is very strong evidence that gravitational forces are proportional to the masses (what Newton called the "quantities of matter") of the objects those forces are exerted on.

But Newton's *third* law implies that the gravitational force exerted by (say) the Earth on a ball, must be equal (in magnitude, though of course opposite in direction) to the gravitational force exerted by the ball on the Earth. So if the first force is proportional to the mass of the ball, so must be the numerically equal second force. Hence, in general, the gravitational force between two objects must *also* be proportional to the mass of the object *exerting* the force.

And, of course, we have already discussed the evidence that the gravitational force between two bodies varies with the distance between those bodies as the inverse square of their separation.

The most general expression consistent with all of this is the following:

$$F = \frac{Gm_1m_2}{r^2} \tag{3.18}$$

where F is the strength (magnitude) of the gravitational force exerted by an object of mass m_1 on an object of mass m_2 located a distance r away from it. The proportionality

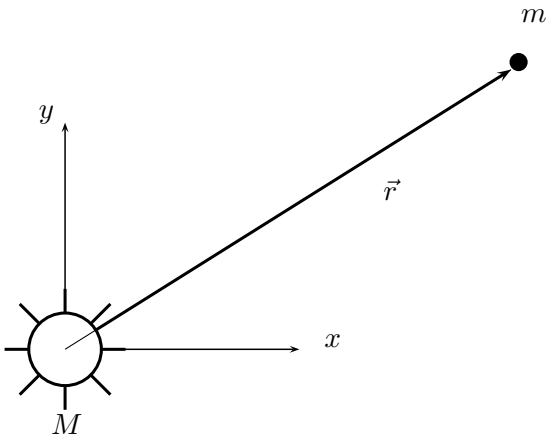

Figure 3.3: An object of mass M (shown here as the Sun) is located at the origin of a coordinate system. Another object, of mass m (here a planet), is located at position $\vec{r}$. The magnitude of the gravitational force exerted by M on m is $GMm/|\vec{r}|^2$. One can write a full vector equation for the force this way: $\vec{F} = -GMm\vec{r}/|\vec{r}|^3$.

constant G is then not dependent on the masses of the two objects involved, nor the distance between them – i.e., it is independent of all the properties the force itself depends on. In short, G is a universal constant that shall be called Newton's constant.

It is helpful also to write an expression for the force in full vector form. Since we will be so often concerned with the motion of the planets around the Sun, let us pick a coordinate system with origin at the Sun, and write an expression for the force exerted by the Sun on a planet of mass m located at position $\vec{r}$.

The magnitude of the force is just $GMm/|\vec{r}|^2$. Its direction is the direction from m back toward M, i.e., just opposite the direction of the coordinate vector $\vec{r}$. A unit vector in this direction can be constructed by dividing $\vec{r}$ by its magnitude, $|\vec{r}|$, and then putting in a minus sign. That is, the force exerted by M on m is in the direction $-\vec{r}/|\vec{r}|$. And so the (full vector) force exerted by M on m is given by

$$\vec{F} = -\frac{GMm}{|\vec{r}|^3}\vec{r}. \tag{3.19}$$

This will be a particularly useful expression for some computations we'll tackle in Chapter 6.

3.6 The Cavendish Experiment

Suppose an object of mass m_1 gravitationally orbits around another object of (much larger) mass M_1, with an approximately circular orbit of radius R_1 and period T_1. The

centripetal acceleration of the orbiting object will then be given by

$$a_1 = \frac{v_1^2}{R_1} = \frac{4\pi^2 R_1}{T_1^2}. \tag{3.20}$$

The gravitational force between the objects will, according to the results of the previous section, have magnitude

$$F = \frac{GM_1 m_1}{R_1^2}. \tag{3.21}$$

Relating the acceleration and force using Newton's second law ($F = ma$), we arrive at a relation between the orbital radius and period and the mass of the central body:

$$\frac{GM_1}{R_1^2} = \frac{4\pi^2 R_1}{T_1^2} \tag{3.22}$$

or equivalently

$$M_1 = \frac{4\pi^2}{G}\frac{R_1^3}{T_1^2} \tag{3.23}$$

which you probably recognize as a statement containing Kepler's third law.

Now suppose there is *another* such orbital system, with an object of mass m_2 orbiting around a (much heavier) object of mass M_2 with orbital radius R_2 and period T_2. Then, by the same argument as above, we should have

$$M_2 = \frac{4\pi^2}{G}\frac{R_2^3}{T_2^2}. \tag{3.24}$$

Now let's divide the last two equations. The proportionality constants (involving Newton's constant G) cancel out, leaving

$$\frac{M_1}{M_2} = \frac{R_1^3/T_1^2}{R_2^3/T_2^2}. \tag{3.25}$$

Thus, if one can empirically determine the orbital characteristics (radius and period) of two orbiting bodies, one can work out the relative masses of the two bodies the orbiting bodies are orbiting around. For example, suppose the first system is the Earth orbiting around the Sun, and the second is the Moon orbiting around the Earth. Then $R_1 = 1\,\mathrm{AU}$, $T_1 = 1\,\mathrm{year}$, $R_2 = 60\,R_{earth}$, and $T_2 = 27.3\,\mathrm{days}$. Plugging in numbers and reducing the units gives

$$\frac{M_1}{M_2} = \frac{M_{sun}}{M_{earth}} = 330,000 \tag{3.26}$$

i.e., the Sun is about three hundred and thirty thousand times more massive than the Earth.

Using the same methods, one can also relate the Earth's mass to that of Jupiter and Saturn, which also have moons that were known to Newton. Based on just this line of reasoning, Newton reports that the mass of Jupiter is about 1/1,000 that of the Sun (or about 330 times that of Earth), with Saturn being about a third that heavy.

It is remarkable that the relative masses of the planets can be so determined. But – perhaps you are already wondering – why not just compute the *masses*? If it is remarkable to know that the Sun is 330,000 times as massive as the Earth, how much more remarkable to know the mass of the Sun *in kilograms*? This would clearly be possible if the mass of the Earth (in kilograms) were known. And one can see from Equation 3.23 that (e.g.) the Sun's mass could also be calculated from the orbital characteristics of (e.g.) the Earth, if only the value of Newton's constant G were known.

But – perhaps surprisingly – neither of these quantities was known to Newton at the time the *Principia* was written. It is perhaps worth reflecting on the reasons for this. The value of G was first measured – by directly measuring the strength of the gravitational force between lead balls in a laboratory – about a *century* after the *Principia* was published, by the amateur physicist Henry Cavendish, in 1798, using an apparatus called the "torsional pendulum" which had been invented by Coulomb to measure small electrostatic forces.

Because of the equivalence, noted above, of knowing the value of G and knowing the mass of the Earth, the Cavendish experiment is sometimes referred to as "weighing the Earth." We will describe the experiment in some detail here, with the hope that you will have the opportunity to reproduce it in your class.

To begin with, here is Cavendish's description of the experiment:

> "The apparatus is very simple; it consists of a wooden arm, 6 feet long, made so as to unite great strength with little weight. This arm is suspended in an horizontal position, by a slender wire 40 inches long, and to each extremity is hung a leaden ball, about 2 inches in diameter; and the whole is inclosed in a narrow wooden case, to defend it from the wind. As no more force is required to make this arm turn round on its center, than what is necessary to twist the suspending wire, it is plain, that if the wire is sufficiently slender, the most minute force, such as the attraction of a leaden weight a few inches in diameter, will be sufficient to draw the arm sensibly aside. .... One of these [weights] was to be placed on one side of the case, opposite to one of the balls, and as near it as could conveniently be done, and the other on the other side, opposite to the other ball, so that the attraction of both these weights would conspire in drawing the arm aside; and, when its position, as affected by these weights, was ascertained, the weights were to be removed to the other side of the case, so as to draw the arm the contrary way, and the position of the arm was to be again determined; and, consequently, half the difference of these positions would show how much the arm was drawn aside by the attraction of the weights. .... I resolved to place the apparatus in a room which should remain constantly shut, and to observe the motion of the arm from without, by means of a telescope; and to suspend the leaden weights in such manner, that I could move them without entering into the room."

A sketch of the torsional pendulum apparatus is shown in Figure 3.4.

Analyzing the experiment in more detail will allow us to apply and concretize some

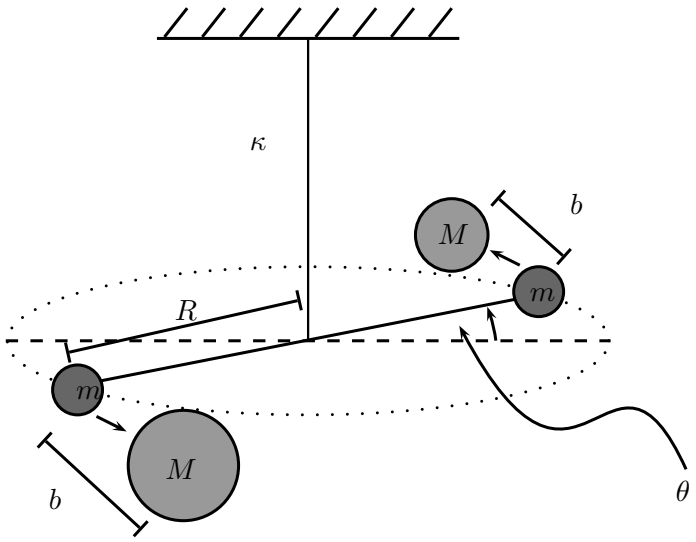

Figure 3.4: Schematic diagram of the Cavendish torsional pendulum apparatus. A "barball" (made of two masses m connected by a thin rod of length $2R$) hangs from a thin fiber. The barbell is thus free to rotate in the horizontal plane. In the absense of external forces, the barbell will orient itself along some equilibrium position (represented by the dashed line). In the Cavendish experiment, two additional balls of mass M are brought in, to a distance b from the respective ends of the barbell. The larger masses then exert (tiny!) gravitational forces on the smaller masses, which cause the barbell to rotate by a (tiny!) angle θ relative to the equilibrium position. Knowing the torsional character of the fiber (represented by the "torsion constant" κ) allows one to infer the absolute magnitude of the gravitational force between the balls, and hence, knowing their masses and separations, the value of Newton's constant G.

of the concepts – in particular "torque" and "angular momentum" – introduced earlier in this chapter. To begin with, note that the motion of the two masses constituting the hanging barbell is in a horizontal plane. Let us pick the center of the barbell (which should never move) as the origin of our coordinate system, and the $x-y$-plane to be horizontal. Then, because of the vector cross products in the definitions of these quantities, all the relevant quantities (torques and angular momenta) will be purely in the z-direction. So we can represent the vector character of these quantities with a simple plus or minus sign, indicating that the quantity is either "up" or "down" along the z-axis.

Let us first consider the hanging barbell in the absense of the larger masses. If the barbell is rotated slightly with respect to its equilibrium orientation, the twist in the fiber will produce a very tiny torque that will tend to turn the barbell back toward its equilibrium configuration. (This is really the definition of "equilibrium.") For small

angular displacements, the resulting torque will be *proportional* to the displacement, i.e.,

$$\tau = -\kappa\,\theta \tag{3.27}$$

where the angle θ is measured from the equilibrium position. Note that a positive angular displacement θ (counterclockwise as seen from above, as shown in the figure) produces a torque that tends to turn the barbell back toward $\theta = 0$ – i.e., a *negative* torque. We thus put an explicit minus sign in the above equation so that the torsional constant of the fiber, κ, is a positive quantity. One should think of κ as characterizing the "rotational stiffness" of the fiber: for example, a thicker fiber will produce a larger restoring torque given the same angular displacement, i.e., will have a larger κ.

So, if the barbell is displaced from its equilibrium orientation, the fiber twists and produces a restoring torque, whose qualitative effect is to turn the barbell back toward its equilibrium configuration. Let us now be more precise. We proved above that the net torque on an object should equal the rate of change of its angular momentum:

$$\vec{\tau} = \frac{d\vec{L}}{dt} \tag{3.28}$$

where the angular momentum of a particle is given by

$$\vec{L} = \vec{r} \times \vec{p} \tag{3.29}$$

where $\vec{p}$ is its linear momentum $\vec{p} = m\vec{v}$.

Ignoring the negligibly light rod that connects the two masses, we can think of the barbell as consisting of two particles of mass m. As long as the torsional pendulum is just rotating (and not swinging, i.e., as long as the center of the barbell remains at the origin), each mass m is always located at a distance R from the origin and its velocity $\vec{v}$ will always be perpindicular to its position vector $\vec{r}$. Each mass's angular momentum is thus given by

$$L = \pm mvR \tag{3.30}$$

where v is the speed, with the $+$ or $-$ on the right applying if the motion of the mass is (respectively) counterclockwise or clockwise (as seen from above). Note that, since the two masses are connected rigidly, they will always have the same speed and will always be moving in the same direction (clockwise or counterclockwise).

It is convenient here to introduce the angular velocity to characterize the motion of the barbell as a whole. This is just the velocity of one of the balls divided by R. The total angular momentum of the entire barbell can hence be written

$$L = 2mR^2\omega \tag{3.31}$$

where now we let the angular velocity ω carry the direction information: ω will be positive (negative) for counterclockwise (clockwise) rotation. Note also that "clockwise" and "counterclockwise" here refer to aspects of the *motion* of the barbell, not its orientation – i.e., they are directions, not places. The barbell can be, say, turned counter-clockwise ($\theta > 0$) and still be *turning* clockwise ($\omega < 0$).

Putting this expression for the barbell's angular momentum together with the basic dynamical equation for rotation, we have

$$-\kappa\,\theta = \frac{d}{dt}\left(2mR^2\omega\right) = 2mR^2\alpha \tag{3.32}$$

where $\alpha = d\omega/dt = d^2\theta/dt^2$ is the angular acceleration of the barbell. This equation can be re-written as

$$\frac{d^2\theta}{dt^2} = -\frac{\kappa}{2mR^2}\,\theta \tag{3.33}$$

which hopefully is recognized as mathematically equivalent to the equation one gets by applying $F = ma$ to a mass m on a spring of spring constant k:

$$\frac{d^2x}{dt^2} = -\frac{k}{m}\,x. \tag{3.34}$$

Such a mass on a spring will undergo simple harmonic motion (meaning that x is a sinusoidal function of time) with period

$$T = 2\pi\sqrt{\frac{m}{k}}. \tag{3.35}$$

But, to quote Feynman, "the same equations have the same solutions". So the torsional pendulum must also undergo (angularly) simple harmonic motion with period

$$T = 2\pi\sqrt{\frac{2mR^2}{\kappa}}. \tag{3.36}$$

The point of all this is that it allows a way to actually *measure* the torsional constant κ of the fiber. Rearranging, we have

$$\kappa = \frac{8\pi^2 mR^2}{T^2}. \tag{3.37}$$

So, given the mass of the balls m and the length of the barbell R, the torsional constant of the fiber can be determined by simply setting the torsional pendulum in (rotational) oscillation, and measuring the period.

And that is actually the hard part. The rest of the analysis of the Cavendish experiment will be relatively straightforward. The basic idea is now to bring in two additional balls (mass M) at distances b from the smaller balls, as shown in the Figure. Each of the large balls will exert, on the nearby smaller ball, a force F and hence a torque

$$\tau = \pm RF = \pm\frac{GMmR}{b^2} \tag{3.38}$$

with a plus or minus depending on whether the gravitational force F tends to pull the barbell counterclockwise or clockwise.

Since there are two such torques on the barbell, the total torque produced by the gravitational influence between the balls is

$$\tau = \pm \frac{2GMmR}{b^2}. \tag{3.39}$$

When the heavy balls are put in place, this gravitational torque will cause the barbell to turn slightly and subsequently oscillate about a *new equilibrium* orientation for which the gravitational torque is just cancelled by the restoring torque produced by the (now slightly twisted) fiber. Using Equation 3.27 for the latter, we get an expression for the angle θ of the new equilibrium orientation

$$\kappa\theta = \frac{2GMmR}{b^2} \tag{3.40}$$

which shows that, by *measuring* the angle θ, one can compute Newton's constant G:

$$G = \frac{\kappa\theta b^2}{2MmR} = \frac{4\pi^2\theta R b^2}{MT^2} \tag{3.41}$$

where now the final right hand side is purely in terms of directly measurable quantities.

There are a number of subtle issues that arise when one actually performs the experiment. Here we will mention only one, which is that typically one measures not the angle θ between the "original" and "new" equilibrium orientations of the barbell, but rather the angle (2θ) between the two "new" equilibrium positions one gets by placing the heavy masses in the two possible ways (where they pull the barbell counterclockwise and clockwise). It is also common in modern versions of the experiment for a small mirror to be mounted near the center of the barbell. The small angular re-orientations of the barbell can then be measured by reflecting a laser beam off the mirror onto a distant wall, and measuring the deflection of the spot on the wall. Since reflection off a mirror rotated by angle θ causes the angle of the light beam to be rotated by 2θ, this means the angle of the laser spot on the wall will move by *four* times the angle we originally defined as θ. Calling this angle ϕ, we then have that

$$G = \frac{\kappa\phi b^2}{8MmR} = \frac{\pi^2 R b \phi R b^2}{MT^2}. \tag{3.42}$$

We include this here just to make life simpler should you be asked to perform this experiment yourself in class and analyze the results.

For what it is worth, Cavendish's own 1798 implied a final result for G that was within 1% of the best contemporary value:

$$G = 6.67 \times 10^{-11}\ \frac{\mathrm{N\,m^2}}{\mathrm{kg^2}}. \tag{3.43}$$

As mentioned, though, Cavendish himself wasn't really interested in measuring what we now call Newton's constant, but was instead interested in "weighing the earth." What

he actually reports as his final conclusion is that the average density of the Earth is 5.48 times that of water, i.e.,

$$\rho = \frac{M_{\text{earth}}}{\frac{4}{3}\pi R_{\text{earth}}^3} = 5.48\,\text{g/cm}^3. \tag{3.44}$$

Plugging in the known radius of the Earth

$$R_{\text{earth}} = 6.37 \times 10^6\,\text{m} \tag{3.45}$$

and solving for the Earth's mass gives

$$M_{\text{earth}} = 6 \times 10^{24}\,\text{kg} \tag{3.46}$$

from which one can, if desired, compute the mass of the Sun (and other planets) explicitly, by using the relations that began this section.

3.7 Gravitational Energy

As is familiar to physics students, the gravitational force near the surface of the earth gives rise to an associated concept of gravitational potential energy. As always, potential energy is defined as the amount of mechanical work an external agent would have to do to arrange things a certain way (by moving them slowly into that arrangement from some initial "reference" arrangement). For objects near the surface of the earth, we took "the ground" as the reference position. In order to lift an object of mass m to a height h off the ground, an external agent has to apply a force equal to the object's weight mg through a distance h. And so the work done – and hence the gravitational potential energy – is the familiar formula: $U = mgh$.

We want to generalize this result now that we know how the gravitational force varies over large distances. We will be particularly interested in finding the gravitational potential energy between two massive objects (masses M and m, say) a distance r apart. As usual, this is simply the amount of work an external agent would have to do to bring about this arrangement. But the first question is: what should we take as the reference arrangement? For two point masses, if we pick some particular finite separation r_0 as the reference, then the gravitational potential energy will be positive or negative depending on whether r is greater or less than r_0. That's a little bit weird. And it'd also be weird to so arbitrarily pick some particular r_0.

It is thus conventional to pick $r = \infty$ as the reference arrangement. That is, we will define the potential energy between two massive objects to be zero when the two objects are infinitely far apart. This actually makes some sense, since this is the only separation for which the force between them vanishes. And anyway it is just a convention. So the question becomes: how much work does an external agent have to do to slowly "lower in" one of the masses (say m) to a distance r away from the other (fixed) mass M, starting from infinite separation? The force exerted by the external agent here will be equal and opposite to the force the masses exert on each other, i.e., it will be opposite the direction m is moving. So the work (and hence the potential energy) will come out to be negative.

And since the force varies continuously with r, we'll have to break the "lowering" process up into a sequence of short steps, and then integrate to find the work done. It is:

$$\begin{aligned} U(r) &= -\int_r^\infty \frac{GMm}{r^2} dr \\ &= -\frac{GMm}{r}. \end{aligned} \tag{3.47}$$

Since the formula for the gravitational potential energy looks alot like the formula for the gravitational *force*, I'll just call your attention explicitly to the fact that the denominator in the PE formula has r to the *first* power, while in the force formula it is of course the inverse *square* of r which appears. And there is another perhaps-misleading similarity: the minus sign that appears in the formula for the gravitational force tells us about the *direction* of the force – that it is attractive rather than repulsive. The minus sign in the potential energy formula, on the other hand, really means minus: the gravitational potential energy between two massive objects is a *scalar* and it is always *negative*.

Let's consider the case of a circular orbit (of, say, a planet of mass m around the sun, mass M). The total energy of the planet will be the sum of its kinetic energy (KE) and gravitational potential energy (PE):

$$\begin{aligned} E &= KE + PE \\ &= \frac{1}{2}mv^2 - \frac{GMm}{r} \\ &= -\frac{1}{2}\frac{GMm}{r} \end{aligned} \tag{3.48}$$

where, to get to the last line, we have used the fact that, for a circular orbit, the centripetal acceleration $a = v^2/r$ is equal to $F/m = GM/r^2$. Notice in particular that the total energy is negative (and that it is precisely half of the gravitational PE, which is of course also negative). This shouldn't bother you. The gravitational potential energy will always be negative, and for a circular orbit the kinetic energy just isn't enough to make the total energy come out positive.

Actually, a negative total E turns out to be a feature of any closed orbit, whether it is a circle or a very eccentric ellipse. An "orbit" with total energy $E = 0$ turns out to be shaped like a parabola. An object undergoing such an orbit would start infinitely far away from the sun, slowly fall in toward it (but with a little angular momentum so it doesn't just crash straight into the sun!), get slingshot around and out the other way, escaping eventually back to infinity (in a different direction) where it will be left with no kinetic energy and hence be (again) at rest. The orbits (and life histories) of some comets approximate this behavior. Orbits with $E > 0$ are also possible. They are qualitatively similar and have the shape of hyperbolae. Notice the curious fact that the shapes of the different possible orbits (in an inverse-square-law gravitational force field) are all of the different *conic sections* – the shapes (circle, ellipse, parabola, and hyperbola) that one can get by slicing a cone with a plane.

This is something we'll take up and prove in Chapter 6. For now, you should just accept it as a conjecture to be proved rigorously later.

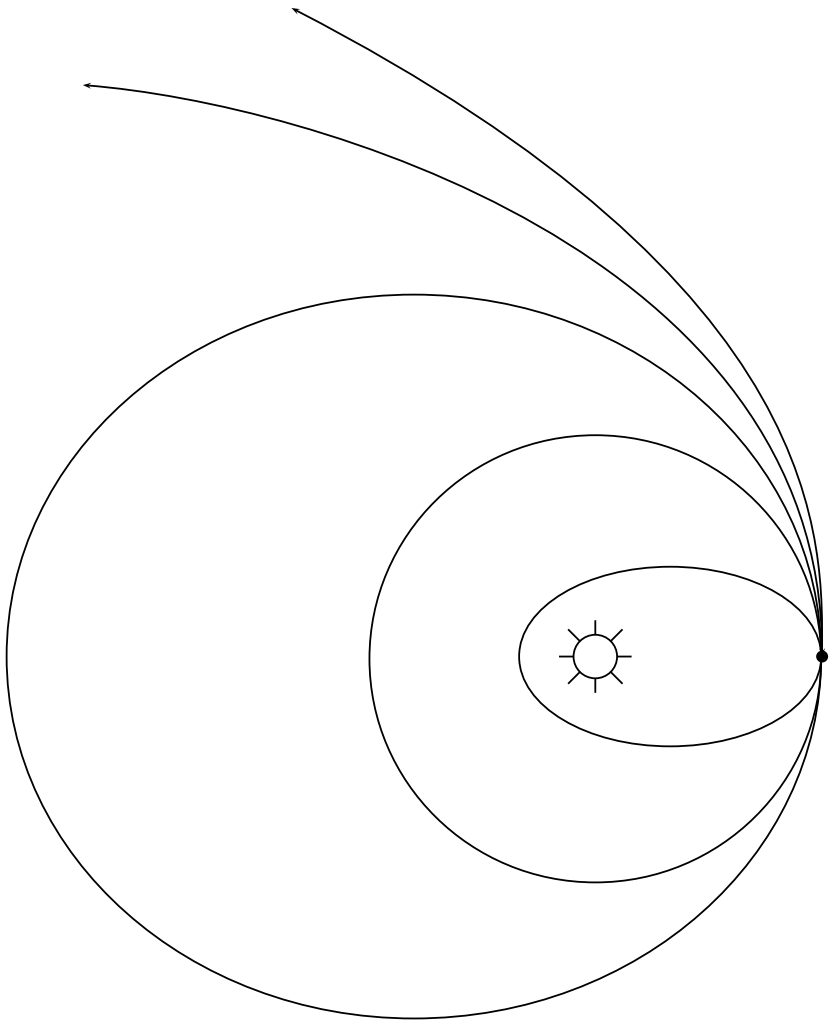

Figure 3.5: Several possible orbits for a mass at the point shown in the influence of the sun's (or whatever's) gravitational field. One may think of these as illustrating the family of trajectories possible for an object passing through the point to the right of the Sun where all the orbits cross, with a velocity directed exactly upward, but with various speeds at that point – or equivalently, with various total energies. The circular orbit corresponds to the (negative) total energy E calculated in the text. A slightly lower energy will result in the small elliptical orbit shown. A slightly higher (but still overall negative) energy will result in the larger elliptical orbit shown. A total energy of zero will produce the parabolic trajectory, while $E > 0$ corresponds to a hyperbolic trajectory.

We noted above that, for a circular orbit, the total energy E can be written in terms of the radius r as

$$E = -\frac{1}{2}\frac{GMm}{r}. \tag{3.49}$$

We just note here in passing that this same equation applies also to elliptical orbits if the "radius" r on the right hand side is interpreted as the semi-major axis of the ellipse.

Something similar is also true of the periods of the orbits. For a circular orbit, the centripetal acceleration v^2/r must equal $F/m = GM/r^2$. Re-writing the orbital speed v in terms of the period T, one gets

$$T^2 = \frac{4\pi^2}{GM}R^3 \tag{3.50}$$

which is just Kepler's third law, with the proportionality constant expanded using Newton's theory. The point is that this relation between the period and radius of a circular orbit remains true in the case of non-circular, elliptical, orbits, if one interprets R as the semi-major axis of the ellipse. This was proved rigorously by Newton and represented a minor *correction* to Kepler's third law as formulated by Kepler. We point it out here just because it is historically interesting that Newton was actually able, using his theory, to *correct* one of the things that had allowed him to formulate the theory in the first place. Of course, there is no logical circularity since what he used in arriving at the theory was the approximate truth of Kepler's third law for the approximately-circular planetary orbits; what he later corrected using the theory was a detail. In any case, the detail is interesting – implying, for example, that the orbital period of a comet with an extremely eccentric elliptical orbit and a planet with a near-circular orbit will be the same, if the semi-major axis of the former matches the radius of the latter. We will have the opportunity to check that this relation really does follow from Newton's inverse square force law in Chapter 6.

Let us mention one last important application of the gravitational potential energy implied by Newton's theory: its use in calculating *escape velocity*. This is the velocity you'd need to give to an object to get it to escape completely from the gravitational attraction that binds it – e.g., to get a rocket to escape from the gravitational pull of the earth, or to get a deep space probe to escape from the gravitational pull of the sun. The way to calculate the escape velocity is simply to ask: what velocity would we need to give the thing in order that it's total energy be *zero*, such that, instead of taking an elliptical orbit and returning back to its current location, it instead follows a parabolic trajectory and escapes to infinity?

$$E = \frac{1}{2}mv^2 - \frac{GMm}{r} = 0 \tag{3.51}$$

can be easily enough solved for v, giving

$$v_{esc} = \sqrt{\frac{2GM}{r}}. \tag{3.52}$$

For $M = M_{earth}$ and $r = R_{earth}$, we find $v_{esc} \approx 11\,km/s$. So if you could launch a projectile with this initial speed near the surface of the earth, it would escape to infinity

(though, strictly speaking, not beyond). Of course, that ignores air resistance, which would in fact be pretty important. If a projectile were actually launched with this speed, it would probably burn up before even making it out of the earth's atmosphere, much as asteroids falling into the atmosphere burn up before they hit the ground. Actually, the cases are quite parallel: an asteroid which starts out pretty far away from earth and not moving too fast, but which then gets captured by the earth's gravitational attraction, has total energy zero and hence will be (or would be, if it didn't burn up first) moving at precisely the escape velocity when it hits the earth's surface.

I'll let you work out in the Projects what the escape velocity is for some other situations.

Questions for Thought and Discussion:

1. What is "centripetal acceleration"? How is the formula $a = v^2/R$ derived?

2. If universal gravitation implies that *all* massive objects attract one another, why don't we observe pairs of familiar household objects (cats and toasters and such) attracting each other gravitationally? According to Newton's theory, how strong *is* the gravitational force between a cat and a nearby toaster?

3. If you didn't have the data from the Cavendish experiment, how might you estimate the mass of the Earth? (Such rough estimates were indeed known to Newton, but they involved extrapolation and guessing.)

4. We have discussed how Kepler's second law ("equal areas in equal times") is equivalent to the statement that each planet's angular momentum is constant in time. But suppose the angular momentum of a planet were measured relative to an origin not at the Sun, but, say, far outside the planet's orbit. Would its angular momentum still be a constant of its motion? Is there any problem here? For example, does this contradict the law of angular momentum conservation? Does it contradict the claimed connection between angular momentum conservation and Kepler's second law?

5. Imagine a flat, frictionless table top with a fixed peg sticking up in the center. One end of a spring is attached to the peg, and the other end of the spring is attached to a ball. Would it be possible to make the ball "orbit" the peg with a circular trajectory? What would be the analog of Kepler's *third* law for this system? Would Kepler's *second* law still apply?

6. Imagine a child standing near the edge of a merry-go-round, which is spinning clockwise (as seen from above). Treat the child as a particle. Using the center of the merry-go-round as an origin, what is the direction of the child's angular momentum vector? How should one think about what this direction means? If the angular velocity of the merry-go-round is decreasing (say, due to some friction), what is the direction of the torque acting on the child? What sort of force produces this torque? Are there any other forces acting that don't produce any torque?

7. How do you think the mass of the Earth's moon was determined?

8. Newton argued, more or less, as follows: the Sun exerts an inverse-square-law force on the planets which controls their orbits; the Earth's gravitational influence falls off with distance as the inverse square; therefore the force exerted by the Sun on the planets is gravitational. What kind of argument is this? Is it reasonable? It might help to compare Newton's argument to other arguments that share this one's structure, e.g.: there are two animals that are both brown, this one is a cow, therefore so is the other one.

9. In the Cavendish experiment, what is the main advantage of having all of the involved objects and forces lying in a horizontal plane? For example, why not simplify the experiment by putting one ball on a scale, then holding another ball a known distance above the first one, and seeing how much the scale reading decreases because of the additional upward inter-ball gravitational force on the first ball?

10. By what factor would the speed of the Earth (in its orbit around the Sun) have to be increased, in order for the Earth to escape the solar system?

11. When Newton claimed that the inverse square law for gravitational forces was "universal", what exactly did this mean? It turns out that certain small anomalies in the orbit of (for example) Mercury are best explained by positing that the true gravitational force law departs somewhat from Newton's law (particularly for heavy bodies and short distances). Einstein's theory of General Relativity, as far as we know today, correctly describes these deviations from Newton's theory. The question is, do these later observations and theories constitute a *refutation* of Newton's claims? Or did Newton only mean to claim, from the beginning, that the inverse square law gives a (then) adequate description of gravitational forces over a certain finite domain of situations? To answer this, it might be helpful to think of other episodes from the history of science. For example, did Copernicus refute Ptolemy, or supplement him? Did Kepler refute Copernicus, or supplement him? What other examples can you think of? The real issue here is to ponder the overall progression of scientific knowledge: is it a sequence of wrong ideas each of which is refuted by the subsequent ones, or a sequence of better and better approximations to the truth, or a sequence of true claims whose generality and scope progressively increase, or what?

Projects:

3.1 Consider a particle moving inertially (i.e., in a straight line with constant speed). Draw its trajectory and pick a (random) origin point that is not on the trajectory. Draw several $\Delta\vec{r}$ vectors representing the displacement of the particle during (equal) finite durations Δt at several different points along the trajectory. Now consider the (triangular) areas "swept out" during each of these Δt period. Does the inertially moving particle sweep out equal areas in equal times? Relate this to

the angular momentum and torque concepts introduced in the text, and explain the overall connection to Kepler's second law.

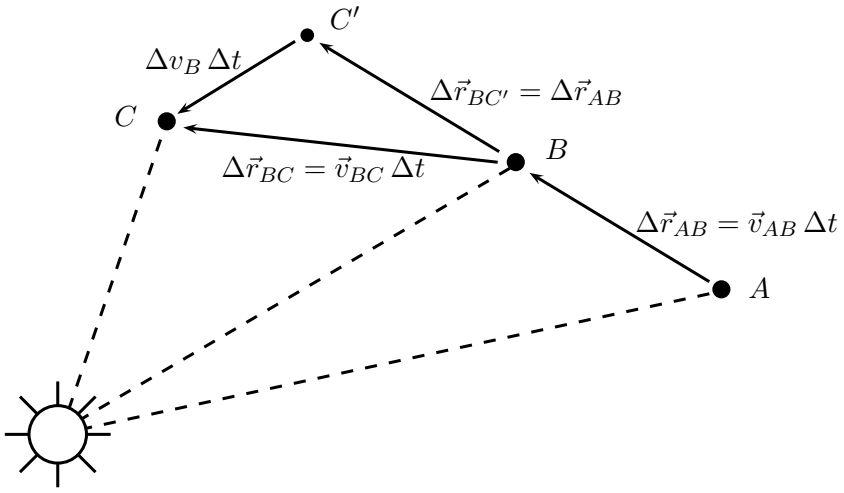

Figure 3.6: A planet receives intermittent regular impulsive forces from a central Sun. Its trajectory is thus not a smooth curve, but a polygon. The positions A, B, and C are occupied by the planet at three successive times separated by the same constant Δt. It is helpful also to consider the position C' that the planet would occupy if no force acted at point B. The displacement vector $\Delta\vec{r}_{BC'}$ is then equal to $\Delta\vec{r}_{AB}$, and so the displacement vector pointing from C' to C is proportional to the velocity change at B: $\Delta\vec{r}_{C'C} = \Delta\vec{r}_{BC} - \Delta\vec{r}_{BC'} = \Delta\vec{r}_{BC} - \Delta\vec{r}_{AB} = (\vec{v}_{BC} - \vec{v}_{AB})\,\Delta t \sim \vec{a}_B$ where a_B is the acceleration of the particle at or around point B. (We simply avoid the question of whether a_B represents an average or instantaneous acceleration by claiming only a proportionality at the end of that string of equalities.) The point is that this acceleration at point B is *toward the Sun*. Hence, the line connecting C and C' is *parallel* to the line connecting the Sun to point B. This is the key insight that then allows a relatively simple geometric argument for the claim that the triangular areas Sun-A-B and Sun-B-C are equal.

3.2 Consider a particle that is intermittently moving inertially, but which receives regular, discrete, impulsive forces directed toward a fixed center. Think, for example, of a planet orbiting the Sun, but in a universe where the Sun, instead of exerting a continuous force on the planet, delivers regular, instantaneous bursts of attractive force. The resulting trajectory of the planet would be something like a normal orbital curve, but a polygon built out of a number of discrete line segments instead of a smooth curve. Figure 3.6 shows two of the straight-line segments of such a trajectory. Use techniques similar those you used in Project 3.1 to (a) show that and (b) understand why the area swept out by the planet (in each time duration between subsequent bursts) is constant. Then do (a) and (b) again using the concepts of angular momentum and torque. Finally, think about how this relates to

the continuous-in-time version of angular momentum conservation and its relation to Kepler's 2nd law that was discussed in the text. (The geometrical part of this is essentially how Newton himself proved that Kepler's second law follows as a logical consequence of his theory of gravitation. Discussing it in terms of torque and angular momentum involves a more modern perspective.)

3.3 Here is an interesting way to derive the important formula for the centripetal acceleration of uniform circular motion: $a_c = v^2/R$. Consider a particle bouncing around on the inside of a circle, such that its trajectory is a regular polygon (an equilateral triangle, a square, a pentagon, etc.). Suppose the particle has mass m and moves with speed v and that the circle (which circumscribes the particle's trajectory) has radius R. Find an expression for the amount by which the particle's velocity changes during one of its collisions with the circle (i.e., where its trajectory makes a corner) and, dividing this by the time between such collisions, find an expression for the average centripetal acceleration of the particle. Work this out first for the case of a square orbit, then see if you can generalize to a regular polygon with N sides. You should expect that, in the limit $N \to \infty$, this average centripetal acceleration for a polygonal orbit goes into the exact expression for the instantaneous centripetal acceleration of a circular orbit. Do you indeed reproduce the familiar formula this way?

3.4 Two moons orbiting Mars – Phobos and Diemos – were discovered in 1877 by the astronomer Asaph Hall. Subsequent observations revealed Phobos to have an orbital radius of $9,380$ km and an orbital period of 7 hours, 39 minutes. Diemos was found to have an orbital radius of $23,460$ km and an orbital period of 1.26 days. What is the mass of Mars (a) in Earth masses and (b) in kilograms?

3.5 Reproduce the Cavendish experiment in your classroom and come up with a range of values for Newton's constant, G. List some of the possible systematic errors that you think influence your results.

3.6 Use your value for G from the Cavendish experiment to "weigh the earth" using the fact that, according to Newton's theory, the gravitational acceleration for an object of mass m near the surface of the Earth should be $g = GM/R^2$ where M and R are the mass and radius of Earth, respectively.

3.7 The Earth's moon has mass 7.35×10^{22} kg and radius $1,737$ km. What is the local gravitational acceleration g at the surface of the Moon? How long would it take a golf ball to drop from head height to the ground if you were on the Moon? How high could you jump?

3.8 What is the escape velocity from the surface of the Moon? Could you throw or hit a golf ball fast enough to make it escape the Moon's gravitational pull? If a meteor (which was long ago more or less at rest far from the Moon) gets captured by the Moon's gravitation and eventually strikes the surface of the Moon, roughly how fast will it be moving on impact? If the meteor's mass is, say, a thousand kilograms,

how much kinetic energy will it have? Compare this to the energy released by the explosion of a ton of TNT: 4×10^9 Joules. How do you think the Moon's craters (visible already to Galileo) were formed?

3.9 Estimate the size of the biggest asteroid (i.e., a rocky planet-like object that is smaller than our Moon) from which you *could* throw a golf ball and have it escape from the asteroid's gravity.

3.9 One theory of the cause of the extinction of the dinosaurs is that a large meteor struck the Earth and radically altered the temperature and other environmental parameters. Suppose a chunk of rocky material about 10 km across were captured by the Earth's gravitation and struck the surface. What would be its kinetic energy on impact? Compare this to the energy released by the currently most powerful nuclear weapons, which release roughly the equivalent of a million tons of exploding TNT (a "megaton"): 4×10^{15} Joules.

3.10 A geo-synchronous orbit is one in which the orbiting satellite moves with the rotating Earth such that, for example, it is always in exactly the same place on the sky as seen from Earth. What must be the radius (both in km and Earth radii) of such a satellite? Are all circular orbits with that radius geo-synchronous? Can Santa Claus get satellite TV during his less busy spring and summer months?

3.11 Compute and compare the *densities* of the Earth, Sun, Jupiter, and Saturn. When retrograding, Jupiter subtends an angular diameter of 0.00023 radians, or 0.013 degrees. Similarly, when retrograding, Saturn (not including the rings) subtends an angular diameter of 0.000093 radians, or 0.0053 degrees.

3.12 Calculate the magnitudes of the forces of gravitational attraction of the earth toward (a) the Sun, (b) the Moon, (c) Jupiter, and (d) Mars. For the planets, use the force that obtains when the earth-planet distance is at its minimum, i.e., find the maximum force.

Chapter 4

Putting it All Together

In the first three chapters, we traced the historical origins of Newton's theory of gravitation. In the final two chapters on this broad topic (Chapters 5 and 6) we trace some of the interesting and important applications of Newton's theory, giving a small taste of the role the theory has played in physics and astronomy between Newton's time and ours. The present chapter is a kind of bridge between Chapter 3 and Chapters 5 and 6. Our goal here will be to develop some important results pertaining to extended bodies (as opposed to point masses). As we'll see, these results will fill in several loopholes from the discussion in Chapter 3, and put us in a position to apply the full set of Newtonian concepts to extended objects like planets, stars, and galaxies.

The perspective that gives rise to the need for this chapter is an early "atomic" or "corpuscularian" view. (A much fuller treatment of this view is the subject of the second half of the book.) According to this view, matter basically consists of small massive particles, and the motion and dynamics of extended objects can and should ultimately be understood by *analyzing* them into their component corpuscles. Part of the idea here is that the fundamental dynamical laws (such as Newton's three laws of motion) are posited for the elementary massive corpuscles. Their applicability for extended objects (which are composed of many such corpuscles) must then be established with the appropriate mathematical theorems.

That this is so for the three laws of motion is relatively straightforward to establish, and is maybe something the reader has thought about in a previous physics course. In Section 4.1 we will reproduce the relevant derivations as a kind of warm-up to the subsequent discussions of the rotation and gravitation of extended bodies.

But one word of warning before jumping in. This chapter may feel very awkward after what has come before, since no attempt is made here to display the relevance of the ideas being discussed to astronomy or the physics of the heavens. Indeed, this chapter will feel much more like a standard physics text, with lots of boring discussion of balls rolling down inclined planes, and things like this. The reason for this is not that the ideas developed here aren't applicable to the interesting astrophysics topics we've been discussing. Instead, the point is to develop your understanding of these topics first using more mundane examples (like the balls on the ramps), just as you did last semester with $F = ma$. The many interesting astrophysical applications of all this stuff will then follow

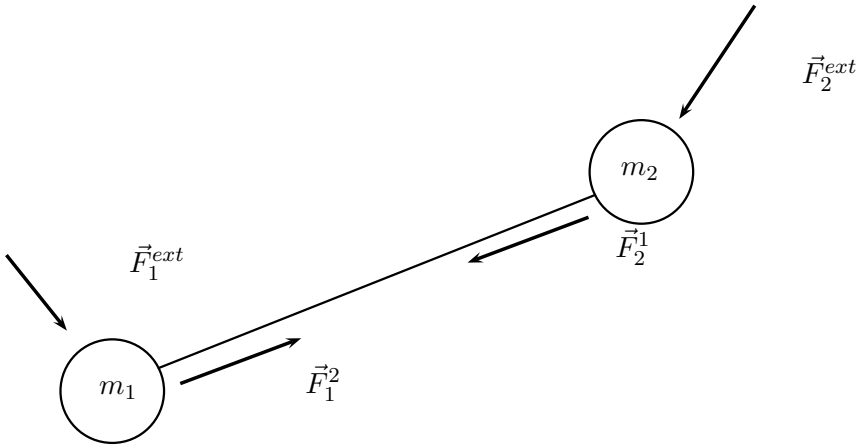

Figure 4.1: A simple example extended body, composed of two massive corpuscles labeled m_1 and m_2. (Never mind what the "rod" connecting them is made of! It is treated here as massless.) This diagram is also double-tasking as a free-body-diagram, so the forces acting on each corpuscle are also shown. These are classified as internal or external forces depending on whether or not they are exerted by other corpuscles are are part of the extended body in question. For the internal forces, our labeling convention is that $\vec{F}_1^2$ denotes the force exerted *on* mass 1 *by* mass 2.

in the next chapter. You'll just have to be patient!

4.1 Newton's Three Laws for Extended Bodies

Suppose we have an extended object composed of several elementary massive corpuscles. For simplicity, let's assume first that the object is composed of only two such particles – e.g., a barbell as shown in Figure 4.1. We'll then cycle back at the end and show how everything we've said about the barbell applies also to any arbitrary extended object.

The Figure shows the two mass points that compose the barbell, and also the forces acting on each of them. The forces can be classified as either "internal" or "external" depending on whether the force in question is exerted by an object that is part of, or outside of, the extended object in question. Specifically the Figure shows, for each particle, the net "external" force (i.e., the sum of all the forces exerted on that object by other particles in the universe, not pictured) and then also the "internal" force exerted by the *other* pictured mass.

Let's assume we're using an inertial reference frame, so that Newton's first law holds. Then, we are positing that Newton's second law applies to each of the two masses, and also that Newton's third law correctly describes the relationship between the forces the two masses exert on each other. In equations, these assumptions read

$$\vec{F}_1^{ext} + \vec{F}_1^2 = m_1\vec{a}_1 \tag{4.1}$$

and

$$\vec{F}_2^{ext} + \vec{F}_2^1 = m_2\vec{a}_2 \tag{4.2}$$

where $\vec{a}_1$ and $\vec{a}_2$ are the accelerations of the two corpuscles, and

$$\vec{F}_1^2 = -\vec{F}_2^1. \tag{4.3}$$

Let us now introduce the concept of the "center of mass" of this extended object. This is simply a point in space that in some way represents the position of the object as a whole. To be precise, it represents the *average* position of all the component corpuscles, with the masses of the corpuscles used as a weighting for the average. Mathematically, the definition is

$$\vec{R}_{CM} = \frac{m_1\vec{r}_1 + m_2\vec{r}_2}{m_1 + m_2} \tag{4.4}$$

where $\vec{r}_1$ and $\vec{r}_2$ are the position vectors for the individual corpuscles. Since the masses m_1 and m_2 are just constants independent of time, it is easy to derive (by taking derivatives with respect to time) expressions for the center of mass velocity

$$\vec{V}_{CM} = \frac{m_1\vec{v}_1 + m_2\vec{v}_2}{m_1 + m_2} \tag{4.5}$$

and acceleration

$$\vec{A}_{CM} = \frac{m_1\vec{a}_1 + m_2\vec{a}_2}{m_1 + m_2} \tag{4.6}$$

where $\vec{v}_1$ is the velocity of particle 1, $\vec{a}_2$ is the accleration of particle 2, etc.

Now let us put all of this together. The idea is to simply *add* together Equations 4.1 and 4.2:

$$m_1\vec{a}_1 + m_2\vec{a}_2 = \vec{F}_1^{ext} + \vec{F}_1^2 + \vec{F}_2^{ext} + \vec{F}_2^1. \tag{4.7}$$

But now we can use Equation 4.3 to cancel the two (equal and opposite) internal forces which appear added on the right hand side. The result is:

$$m_1\vec{a}_1 + m_2\vec{a}_2 = \vec{F}_1^{ext} + \vec{F}_2^{ext}. \tag{4.8}$$

Finally, we can recognize the left hand side as the numerator from the right hand side of Equation 4.6. This allows us to write

$$\vec{F}^{ext} = M\vec{A}_{CM} \tag{4.9}$$

where $\vec{F}^{ext}$ without any subscripts refers to the *total* external force $\vec{F}_1^{ext} + \vec{F}_2^{ext}$, and we define the total mass $M = m_1 + m_2$.

The final result here looks just like Newton's second law ($F = ma$) but is a description of the two-particle-system treated, in some sense, as a whole, as a single object. In words, what we proved is that if the constituent particles of this barbell object obey Newton's laws of motion individually, then the object as a whole will, too – with (not surprisingly) the total mass functioning as "the mass of the whole," the total external force acting as "the net force acting on the whole," and the acceleration of the center of mass point functioning as "the acceleration of the whole."

It shouldn't be surprising that this works out. After all, it was empirical observation of big extended objects that led Newton to posit his laws of motion in the first place. So there would be a pretty serious problem if positing those laws for the small constituent particles resulted in anything else for the motion of the whole. Still, it's nice to see that – and how – it works out.

Before going on, we should check explicitly that the same method works for a more general extended object, composed of some arbitrary number of elementary massive particles. The logic will be exactly the same, so let's just breeze through this quickly for the record. (If you understood the discussion of the barbell, there's really nothing new here except bigger numbers, and we won't even see those since we'll write things more abstractly, as sums.) Thus, suppose an extended object is composed of N massive particles with masses m_i and positions $\vec{r}_i$. We can then define the total mass and center of mass position by

$$M = \sum_{i=1}^{N} m_i \tag{4.10}$$

and

$$R_{CM} = \frac{\sum_{i=1}^{N} m_i \vec{r}_i}{M}. \tag{4.11}$$

The CM point's velocity and acceleration are then defined by differentiation. For example,

$$A_{CM} = \frac{\sum_{i=1}^{N} m_i \vec{a}_i}{M}. \tag{4.12}$$

We are assuming that the motion of each individual particle is governed by Newton's laws, i.e.,

$$m_i \vec{a}_i = \vec{F}_i^{net} = \vec{F}_i^{ext} + \sum_{j \neq i} \vec{F}_i^j \tag{4.13}$$

and that Newton's third law holds:

$$\vec{F}_i^j = -\vec{F}_j^i. \tag{4.14}$$

Now, just following what we did for the case $N = 2$ above, we may simply sum Equation 4.13 for all the particles:

$$\sum_{i=1}^{N} m_i \vec{a}_i = \sum_{i=1}^{N} \left(\vec{F}_i^{ext} + \sum_{j \neq i} \vec{F}_i^j \right). \tag{4.15}$$

The internal forces will cancel pairwise, leaving, on the right hand side, only the total external force:

$$\vec{F}^{ext} = \sum_{i=1}^{N} \vec{F}_i^{ext}. \tag{4.16}$$

Equation 4.15 can then be re-written, using Equation 4.12, as

$$\vec{F}^{ext} = M \vec{A}_{CM} \tag{4.17}$$

as expected. So what we found above for the barbell is really general. In the following sections, we'll follow more or less this same procedure to discuss the rotational behavior (including torque, angular momentum, and energy) and the gravitational influence of extended bodies.

4.2 Kinetic Energy of an Extended Object

Let us then turn to the following question: what is the total kinetic energy of an extended object? In one sense, the answer is trivial. The total kinetic energy is just the sum of the kinetic energies of all of the constituent particles:

$$\mathrm{KE}_{total} = \sum_{i=1}^{N} \frac{1}{2} m_i \vec{v}_i^{\,2}. \tag{4.18}$$

And for a completely general extended object, that is about all we can say. But for the special case of a *rigid* body – an extended object whose constituent particles are somehow or other "glued" together – it can be shown that the above formula simplifies in a particularly nice way, that will help clarify several concepts before they show up again later in the more complicated context of examining torque and angular momentum.

As before, let's start by exploring this for the simplest possible multi-particle rigid body, the barbell. We'll then cycle back around at the end and show that everything we found for the barbell actually applies in general.

So consider the situation pictured in Figure 4.2. The main point is that the overall motion of the barbell (assuming, for simplicity, that the motion is confined to the plane) can be characterized by specifying (i) the translational velocity of the center of mass point, and (ii) the *angular velocity* of the (remember, rigid) object as a whole *about* the center of mass point. Since the object is rigid, we can express the translational velocity of each individual particle in terms of V_{CM} and ω. Doing this, and then using the resulting expressions for the individual particle velocities to work out the total kinetic energy, allows us to rewrite the total kinetic energy in terms of V_{CM} and ω in a particularly illuminating way.

Let's go through this in detail. To begin with, suppose the center of mass velocity V_{CM} has x and y components V_{CM}^x and V_{CM}^y respectively. (We assume, only for simplicity, that all the motion is in the $x-y$ plane.) Now, what are the x and y components of the velocity of particle 1 with respect to the center of mass point, which it has by virtue of the object's rotation? Suppose the distance from the center of mass point to particle 1 is r_1. Then the *speed* of particle 1 relative to the center of mass will be $v_1 = r_1\omega$. The *direction* of its relative velocity will be perpindicular to the axis of the barbell, to the upper-left in the figure. Since the barbell is oriented at an angle θ with respect to the x-axis, we can see that the components of its relative velocity will be

$$v_1^{x,rel} = -v_1 \sin(\theta) = -r_1\omega \sin(\theta) \tag{4.19}$$

and

$$v_1^{y,rel} = v_1 \cos(\theta) = r_1\omega \cos(\theta). \tag{4.20}$$

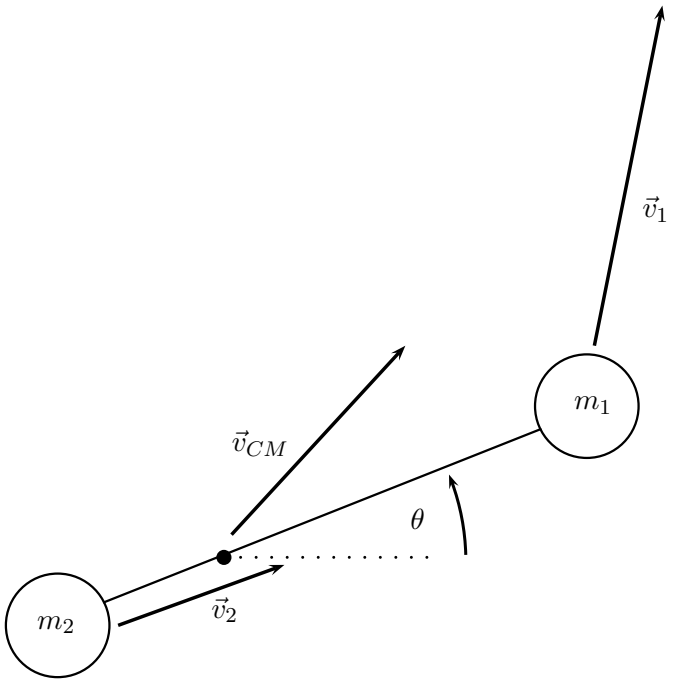

Figure 4.2: The moving barbell object. We assume $m_2 > m_1$ such that the center of mass point (represented by the solid dot in the Figure) will be nearer m_2, as shown. The object as a whole, as represented by $\vec{V}_{CM}$, moves to the upper right. The barbell is also rotating counterclockwise with angular velocity ω. The translational velocity of each particle can be understood as the velocity the particle has with respect to the center of mass point (by virtue of the rotational motion) plus the velocity of the center of mass.

That, remember, is just the velocity of particle 1 relative to the center of mass. Its actual velocity $\vec{v}_1$ will be the (vector) sum of this with the velocity of the center of mass point itself. Hence:

$$v_1^x = -r_1\omega\sin(\theta) + V_{CM}^x \tag{4.21}$$

and

$$v_1^y = r_1\omega\cos(\theta) + V_{CM}^y. \tag{4.22}$$

The story for particle 2 is exactly the same, with the exception that its velocity relative to the center of mass is in the opposite direction (down and to the right in the figure), so the signs of $v_2^{x,rel}$ and $v_2^{y,rel}$ are opposite those given above for particle 1. The results for particle 2's overall velocity components are:

$$v_2^x = r_2\omega\sin(\theta) + V_{CM}^x \tag{4.23}$$

and

$$v_2^y = -r_2\omega\cos(\theta) + V_{CM}^y \tag{4.24}$$

where r_2 is the distance between the center of mass and particle 2. Note also that, by virtue of the definition of the center of mass, the distances r_1 and r_2 are related to the masses m_1 and m_2 as follows:

$$m_1 r_1 = m_2 r_2. \tag{4.25}$$

Let's now finally put all this together. The game is to plug the expressions for the velocities of the two particles into the general expression, Equation 4.18, for the total kinetic energy. Doing this yields

$$\begin{aligned}
\mathrm{KE}_{total} &= \frac{1}{2}m_1\vec{v}_1^2 + \frac{1}{2}m_2\vec{v}_2^2 && (4.26)\\
&= \frac{1}{2}m_1\left[(v_1^x)^2 + (v_1^y)^2\right] + \frac{1}{2}m_2\left[(v_2^x)^2 + (v_2^y)^2\right] && (4.27)\\
&= \frac{1}{2}m_1\left[(-r_1\omega\sin(\theta) + V_{CM}^x)^2 + (v_1^y - r_1\omega\cos(\theta) + V_{CM}^y)^2\right] && (4.28)\\
&\quad + \frac{1}{2}m_2\left[(r_2\omega\sin(\theta) + V_{CM}^x)^2 + (-r_2\omega\cos(\theta) + V_{CM}^y)^2\right] && (4.29)
\end{aligned}$$

Expanding out all the squared binomials and organizing and simplifying, we get three groups of terms:

$$\begin{aligned}
\mathrm{KE}_{total} &= \frac{1}{2}(m_1 + m_2)\left((V_{CM}^x)^2 + (V_{CM}^y)^2\right) && (4.30)\\
&\quad + \frac{1}{2}\left(m_1 r_1^2 + m_2 r_2^2\right)\omega^2 && (4.31)\\
&\quad + (m_1 r_1 - m_2 r_2)\left[\omega\cos(\theta)V_{CM}^y - \omega\sin(\theta)V_{CM}^x\right]. && (4.32)
\end{aligned}$$

The first thing to notice is that, by virtue of the definition of the center of mass, the whole third group of terms (which, notice, were the *cross terms* in expanding out the binomials) vanishes: $m_1 r_1 - m_2 r_2 = 0$. This leaves only the first and second terms, which correspond respectively to the *purely translational* and *purely rotational* aspects of the motion. Thus, the total kinetic energy of the moving barbell can be written

$$KE_{total} = \frac{1}{2}M\vec{V}_{CM}^2 + \frac{1}{2}I\omega^2 \tag{4.33}$$

where $M = m_1 + m_2$ is just the total mass of the object, and I – the *moment of inertia* – is defined mathematically as

$$I = m_1 r_1^2 + m_2 r_2^2 \tag{4.34}$$

and should be understood as the rotational analog of mass. It is, like the actual mass, an *intrinsic* property of the object, which depends on how it is configured, but doesn't depend on how (or whether) it is moving. The purely rotational part of the total kinetic energy – $\frac{1}{2}I\omega^2$ – then depends on this intrinsic property and the angular velocity, in just the same way that the translational part of the total kinetic energy depends on the (total) mass and (overall) translational velocity.

There is one caveat that should be put on the table right away. We have spoken of the distances r_1 and r_2 as the distances of the particles from the center of mass point. That

characterization is appropriate for what we just did, but is not fully general. Really, the relevant distances are the distances between the particles and the *axis about which the rotation is occurring*. So, for example, the same physical object – the barbell – could be rotating about an axis that coincides with the line connecting the two particles. In that case, both of the particles are right on the axis, i.e., their distances from the axis are *zero*, and so the moment of inertia will also be zero. For this reason, calling the moment of inertia (as it was defined above) an intrinsic property of the object is a little misleading. It *is* an intrinsic property of the object, but only relative to a particular axis about which that object might rotate. Relative to different such axes, the same one unchanged object can have different moments of inertia. There is, not surprisingly, mathematical technology that allows one to compile the moments of inertia of a given object relative to three particularly important axes, in something like a "vector" from which one can deduce what the moment of inertia will be about *any* axis. Such information so compiled definitely captures an intrinsic property of a rigid body. We won't develop any of this further here, for the simple reason that we won't ever need to use it. But we will be talking about the moment of inertia of non-planar objects (such as spheres), so it is important to know how to properly generalize what we did above.

Back to the big picture: at least for this particular rigid object (the barbell), the total kinetic energy can be written as the sum of a "translational" piece and a "rotational" piece. This came about because, as noted, all the cross terms cancelled out above – which in turn happened because we had referred the positions of all the particles to the center of mass point. Let us now show that this central result is completely general. Hence, imagine some arbitrary body composed of N elementary particles with masses m_i and being located at position $\vec{r}_i^{\,rel}$ relative to the center of mass point. The position of the ith particle can then be written

$$\vec{r}_i = \vec{R}_{CM} + \vec{r}_i^{\,rel}. \tag{4.35}$$

Notice that, by plugging this expression into the definition of the center of mass,

$$\vec{R}_{CM} = \frac{\sum_{i=1}^{N} m_i \vec{r}_i}{M} \tag{4.36}$$

we can show that

$$\sum_{i=1}^{N} m_i \vec{r}_i^{\,rel} = 0. \tag{4.37}$$

This is the analog of the $m_1 r_1 - m_2 r_2 = 0$ condition we had for the barbell.

Let us also write the *velocities* of the particles as the velocity of the center of mass point *plus* whatever velocity a given particle has *relative* to the center of mass point. (This latter contribution will be due exclusively to the *rotational* motion.) A clever trick, explained in Figure 4.3, is then to write these relative velocities as follows:

$$\vec{v}_i^{\,rel} = \vec{\omega} \times \vec{r}_i^{\,rel} \tag{4.38}$$

where the vector $\vec{\omega}$ represents the angular velocity of the rigid body: its magnitude is just the familiar angular velocity $d\theta/dt$, and its *direction* coincides with the *axis* about

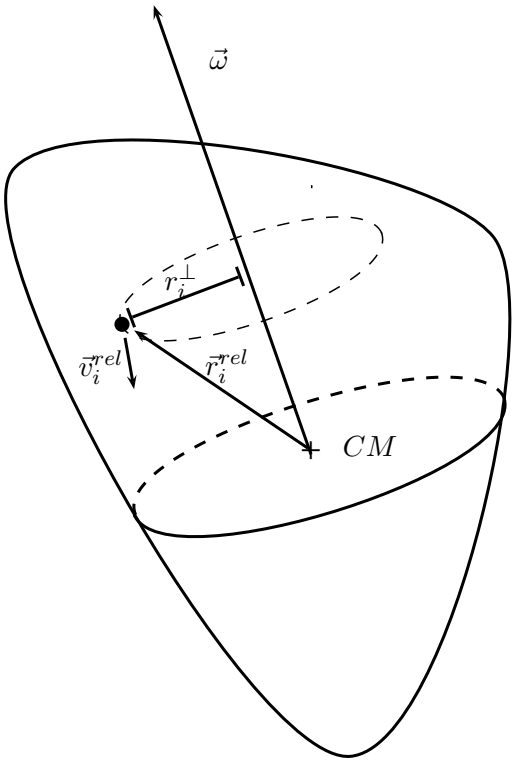

Figure 4.3: A rotating rigid body. The ith particle, marked by a black circle, moves in a circle of radius $r_i^{\perp}$ about the rotation axis. Its position relative to the center of mass point (the X marked "CM") is given by $\vec{r}_i^{rel}$. The angular velocity vector is $\vec{\omega}$. The instantaneous velocity of the particle has a direction that is tangential to the dotted-line circle (out of the page) and a magnitude $|\vec{\omega}|r_i^{\perp}$. Both of these features of $\vec{v}_i^{rel}$ are correctly captured by the formula: $\vec{v}_i^{rel} = \vec{\omega} \times \vec{r}_i^{rel}$.

which the body is rotating. (The ambiguity in sign is resolved by the right hand rule: curl the fingers of your right hand in the direction of the rotation, and your thumb points in the direction that should be assigned to $\vec{\omega}$.)

We can then write the overall velocity of the i'th particle as follows:

$$\vec{v}_i = \vec{V}_{CM} + \vec{\omega} \times \vec{r}_i^{rel} \tag{4.39}$$

and immediately plug into the beginning expression for the total kinetic energy:

$$KE_{total} = \sum_{i=1}^{N} \frac{1}{2} m_i \left(\vec{\omega} \times \vec{r}_i + \vec{V}_{CM}\right) \cdot \left(\vec{\omega} \times \vec{r}_i + \vec{V}_{CM}\right) \tag{4.40}$$

where we have written the square of the vector using the dot product. It is now relatively straightforward to get the desired result. The only tricky thing is evaluating the dot

product of a cross product with a cross product. The mathematical identity needed is:

$$\left(\vec{A}\times\vec{B}\right)\cdot\left(\vec{A}\times\vec{B}\right)=A^2B^2-\left(\vec{A}\cdot\vec{B}\right)^2. \tag{4.41}$$

We then find that the expression for the total kinetic energy can be simplified to

$$KE_{total}=\frac{1}{2}M\vec{V}_{CM}^2+\frac{1}{2}\sum_{i=1}^{N}m_i\left(\omega^2|\vec{r}_i|^2-(\vec{\omega}\cdot\vec{r}_i)^2\right) \tag{4.42}$$

plus another term, again arising from the "cross terms," which is proportional to $\sum_i m_i\vec{r}_i$ and is hence zero as shown above.

The second term simplifies, as expected, to $\frac{1}{2}I\omega^2$, where

$$I=\sum_{i=1}^{N}m_i(r_i^{\perp})^2 \tag{4.43}$$

where here $r_i^{\perp}$ is (as advertised) the perpindicular distance between particle i and the rotation axis passing through the center of mass point. One can see from the Pythagorean theorem that this distance (squared) should be given by $|\vec{r}_i|^2$ minus the square of the component of $\vec{r}_i$ that is *parallel* to that rotation axis. This component is precisely what is picked off by taking the dot product of $\vec{r}_i$ with $\vec{\omega}$, so one can see how this comes out of the complicated looking Equation 4.42. Note that, for the special case of an object that lives in the plane, rotating about an axis perpindicular to that plane, $\vec{\omega}$ and $\vec{r}_i$ are perpindicular for all i, and so their dot product vanishes, and so $|\vec{r}_i|=r_i$. That's why we could get away with ignoring this distinction at first when we were talking about the barbell moving and rotating in the plane.

The important point, of course, is just the result that for any arbitrary object composed of many point masses, the total kinetic energy (the sum of $1/2mv^2$ for all the individual particles) can indeed be written as the sum of a purely translational part and a purely rotational part:

$$\mathrm{KE}_{total}=\frac{1}{2}M\vec{V}_{CM}^2+\frac{1}{2}I\omega^2 \tag{4.44}$$

with the moment of inertia I defined as in Equation 4.43.

To see, right away, at least one important application of this idea, let's see how we can use it to analyze a standard physics textbook sort of problem: a ball rolling down a ramp.

We'll tackle this same problem later using forces and torques, but for now we can use the short-cut of using energy to calculate the acceleration. First, at the beginning of the experiment, the ball is at rest on the top of the ramp, and hence has total energy $E=Mgh$. Then it rolls down the ramp. When it reaches the bottom of the ramp, it will have some translational velocity v and also (since it is *rolling*) some angular velocity ω. Note that, if it is really rolling (and not slipping) these two quantities have to be related by $v=\omega R$. This is sometimes called the "rolling without slipping" condition. In any case, the total energy of the ball as it reaches the bottom of the ramp can be written

$$E=\frac{1}{2}Mv^2+\frac{1}{2}I\omega^2=\frac{1}{2}\left(M+\frac{I}{R^2}\right)v^2. \tag{4.45}$$

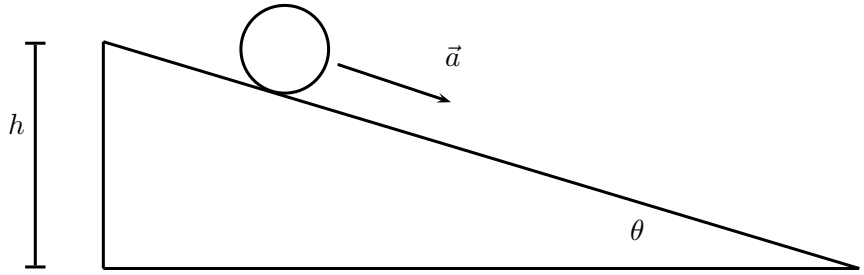

Figure 4.4: A ball of mass M and radius R rolling down an angle θ ramp.

The total energy should be conserved since (let's assume) no outside forces (such as friction) do any work. (It is worth pausing briefly to contemplate this. Friction, despite doing no mechanical work, is certainly *present* – else the ball would *slide* down the ramp, not *roll*!) We can therefore set the above expression equal to Mgh, and solve for v^2:

$$v^2 = \frac{2Mgh}{M + \frac{I}{R^2}}. \tag{4.46}$$

It is a standard one-dimensional kinematics result that, for motion with constant acceleration starting from rest, the final velocity squared is equal to $2a\Delta x$ where a is the acceleration and Δx is the distance traversed. Our Δx is just the length of the ramp: $h/\sin(\theta)$. So we may solve for the acceleration of the ball:

$$a = \frac{v^2}{2\Delta x} = \frac{g\sin(\theta)}{1 + \frac{I}{MR^2}}. \tag{4.47}$$

For an object that *slides* frictionlessly down the same ramp, the acceleration will be just $a = g\sin(\theta)$. So what we've found is that the acceleration is a little bit smaller for an object that rolls, by an amount that depends on the ratio I/MR^2. Qualitatively, the reason for this is that, for a rolling object, some of the initial gravitational kinetic energy gets "wasted" – going into the purely *rotational* kinetic energy that doesn't make the object move fast, but only turn fast.

To get a more quantitative understanding of this, we need to learn how to calculate and think about moments of inertia for some simple shapes. The simplest case is a hoop or hollow cylinder (like a piece of pipe) rotating about its symmetry axis. For such an object, *all* of the individual particles are the same distance R away from the center. And so the moment of inertia is trivial to compute:

$$I_{hoop} = \sum_i m_i r_i^2 = R^2 \sum_i m_i = MR^2 \tag{4.48}$$

where M is the total mass of the hoop or pipe.

We may plug this result into the above formula for the acceleration. We find that, for a hoop or pipe rolling down a ramp, the acceleration is precisely *half* that of an object that slides down frictionlessly, because the I/MR^2 in the denominator is just 1.

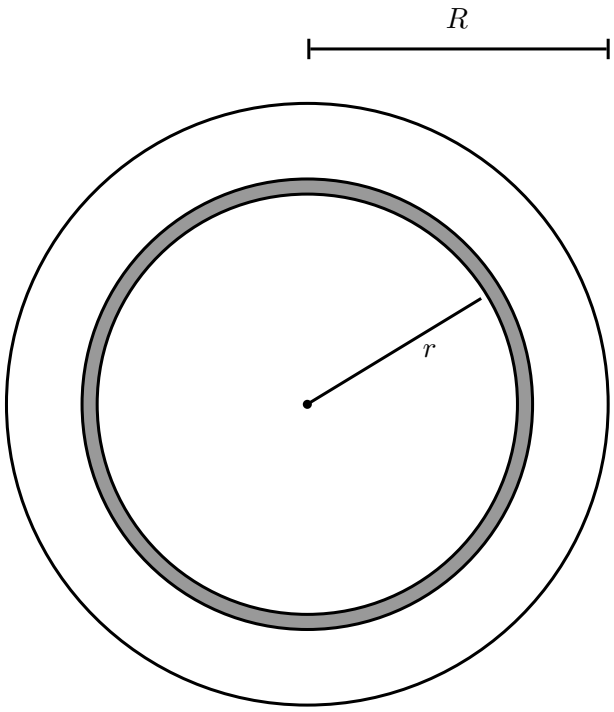

Figure 4.5: Calculating the moment of inertia of a disc of radius R and total mass M. The shaded hoop has radius r and thickness dr, hence total area $dA = 2\pi r dr$. The whole disc has area $A = \pi R^2$, so the shaded ring is a fraction $dA/A = 2rdr/R^2$ of the total area. Hence the mass of the shaded ring is given by $dm = 2Mrdr/R^2$.

One can then hopefully begin to see that, for some other object like a sphere, since *some* of the mass is *closer* to the rotation axis than the object's radius R, the quantity I/MR^2 will be somewhat *less* than 1, and the acceleration will be accordingly *greater*. So, for example, in a downhill rolling race, a solid sphere should beat a hoop. It might be worth pausing at this point and actually performing some races with household objects, and thinking about the results!

Let's calculate the moment of inertia for one more object. (We'll then just quote the result for a ball, and leave the proof for the Projects.) Consider a solid disc, like a pancake, rolling down the ramp. (Nevermind that a pancake wouldn't roll too well. A solid cylindrical candle is maybe a more realistic equivalent example.) The idea is to imagine chopping the object up into a whole bunch of small-ish "pieces" each of whose moment of inertia is known. The moment of inertia for the whole object can then be calculated by adding (or more precisely, integrating) up the contributions from all the pieces.

It is convenient to divide the disc up into a number of small hoops, since we just calculated the moment of inertia for a hoop. Taking the whole disc to have mass M and radius R, we focus our attention on a small hoop of radius r and thickness dr. Its

area is then $dA = 2\pi r dr$, which is $2rdr/R^2$ of the total area $A = \pi R^2$. Hence, assuming the mass of the disc is uniformly distributed over its area, the shaded hoop should have mass $dm = 2Mrdr/R^2$. The idea is to treat each of these small hoops as a "particle" composing the full disc. We can thus calculate the disc's moment of inertia by using

$$I = \sum_i m_i r_i^2 = \int dm r^2 = \int_0^R \frac{2Mrdr}{R^2} r^2. \tag{4.49}$$

Performing the integral gives

$$I_{disc} = \frac{1}{2} MR^2 \tag{4.50}$$

which certainly seems plausible given the qualitative discussion above. Hence, a solid disc or cylinder (e.g., a candle or a can of tomato paste) should roll down a ramp at 2/3 the rate of the intrepid frictionlessly sliding object.

What about a solid sphere (e.g., an orange)? Imagine what you'd have to do to "morph" a solid cylinder like a candle into the shape of a sphere. Basically, you have to move some of the mass from the edges of the cylinder, down toward the axis. Thus, on average, the mass of a solid sphere is a little closer to the axis than for a cylinder, so we'd expect, for a solid sphere, the moment of inertia to be a little less than for a disc of the same mass and radius. It turns out the exact formula is

$$I_{sphere} = \frac{2}{5} MR^2 \tag{4.51}$$

which seems reasonable and can be proven with just a little work. (See the Projects.)

4.3 Rotational Dynamics of a Rigid Body

Let's now follow this same pattern of discussion, and work out some important results for the rotational dynamics of rigid bodies. The starting point here will be the concepts of torque and angular momentum which we introduced, in Chapter 3, for point particles. Recall that the angular momentum $\vec{L}$ of a particle located at position $\vec{r}$ and moving with momentum $\vec{p}$ was defined as

$$\vec{L} = \vec{r} \times \vec{p} \tag{4.52}$$

with the multiplication on the right being the vector cross product. Also, the torque $\vec{\tau}$ produced by a force $\vec{F}$ acting on a particle at position $\vec{r}$ is

$$\vec{\tau} = \vec{r} \times \vec{F}. \tag{4.53}$$

We showed in the last chapter that, with these definitions and Newton's second law, the net torque on a particle (i.e., the sum of the torques produced by all the forces acting on it) was equal to the rate of change of its angular momentum:

$$\vec{\tau}_{net} = \frac{d\vec{L}}{dt}. \tag{4.54}$$

Our goal here will be to see what these basic facts about individual point particles imply about the behavior of an extended, rigid object composed of many such particles "glued" together.

We'll first show that the total angular momentum of a rigid object can be written as the sum of two terms – just as with the total kinetic energy – one of which pertains to the motion of the object as a whole (as captured by the velocity of the center of mass point), with the other pertaining to the rotational motion of the object *about* the center of mass. We'll then work out how the net torque on a rigid body relates to its total angular momentum. And finally we'll show how to apply all of this to a couple of examples.

Let's begin by working out the simple case of the barbell. Refer back again to Figure 4.2. We want to calculate the total angular momentum of the barbell by adding up the angular momentum of the two constituent particles:

$$\vec{L} = \vec{L}_1 + \vec{L}_2 = m_1\vec{r}_1 \times \vec{v}_2 + m_2\vec{r}_2 \times \vec{v}_2. \tag{4.55}$$

The "trick" will be to write the positions $\vec{r}$ and velocities $\vec{v}$ of the two particles in terms of the position $\vec{R}_{CM}$ and velocity $\vec{V}_{CM}$ of the center of mass point. From the Figure, it is apparent that

$$r_1^x = R_{CM}^x + r_1\cos(\theta) \tag{4.56}$$

and

$$r_1^y = R_{CM}^y + r_1\sin(\theta) \tag{4.57}$$

and likewise for particle 2:

$$r_2^x = R_{CM}^x - r_2\cos(\theta) \tag{4.58}$$

and

$$r_2^y = R_{CM}^y - r_2\sin(\theta). \tag{4.59}$$

We already worked out how the x and y components of the velocities of particles 1 and 2 relate to V_{CM} in Equations 4.21 - 4.24. We may use the fact that $\vec{r} \times \vec{v}$ will be in the z direction if $\vec{r}$ and $\vec{v}$ both lie in the $x-y$ plane, and will have magnitude

$$|\vec{r} \times \vec{v}| = r_x v_y - r_y v_x \tag{4.60}$$

and then plug all of the above expressions for the vector components into Equation 4.55. When the dust settles (using the fact that $m_1 r_1 - m_2 r_2 = 0$ to eliminate several terms) the result is

$$|\vec{L}| = (m_1 + m_2)\left(R_{CM}^x V_{CM}^y - R_{CM}^y V_{CM}^x\right) + (m_1 r_1^2 + m_2 r_2^2)\omega \tag{4.61}$$

which can be understood as a special case of

$$\vec{L} = \vec{L}_{orbital} + \vec{L}_{spin} \tag{4.62}$$

where the "orbital" and "spin" contributions to the total angular momentum are, as advertised, associated with the motion of the object as a whole and the purely rotational motion respectively:

$$\vec{L}_{orbital} = \vec{R}_{CM} \times \vec{P}_{CM} = M\vec{R}_{CM} \times \vec{V}_{CM} \tag{4.63}$$

where M is the total mass, and

$$\vec{L}_{spin} = I\vec{\omega} \tag{4.64}$$

where I is the moment of inertia as defined previously.

We can prove that this is indeed the correct general formula by working through a parallel calculation for an arbitrary rigid body composed of N point particles. Using again the trick from the previous section of writing

$$\vec{r}_i = \vec{R}_{CM} + \vec{r}_i^{\,rel} \tag{4.65}$$

and

$$\vec{v}_i = \vec{V}_{CM} + \vec{\omega} \times \vec{r}_i^{\,rel} \tag{4.66}$$

we can plug into

$$\begin{aligned} \vec{L} &= \sum_{i=1}^{N} m_i \vec{r}_i \times \vec{v}_i & (4.67)\\ &= \sum_{i=1}^{N} m_i \left(\vec{R}_{CM} + \vec{r}_i^{\,rel}\right) \times \left(\vec{V}_{CM} + \vec{\omega} \times \vec{r}_i^{\,rel}\right) & (4.68)\\ &= \sum_{i=1}^{N} m_i \vec{R}_{CM} \times \vec{V}_{CM} + \sum_{i=1}^{N} m_i \vec{r}_i^{\,rel} \times \left(\vec{\omega} \times \vec{r}_i^{\,rel}\right) & (4.69) \end{aligned}$$

plus two other terms which vanish because $\sum_i m_i \vec{r}_i^{\,rel} = 0$. The first non-vanishing term in the last line is just the "orbital" angular momentum defined above. It is the angular momentum that a *point particle* would have if it had the same mass, position, and velocity as our rigid body. The second term is the "spin" angular momentum, which is nonzero when the body as a whole is rotating about its center of mass.

Note the analogy here between angular momentum and kinetic energy. Of course the former is a vector and the latter is a scalar. But the two quantities are similar in that the total amount of both, for a rigid body, can be written as a sum of two terms, one depending on the translational motion of the object as a whole (as described by the velocity of the center of mass point) and the other depending on the rotational motion (as described by the angular velocity).

Now let's figure out how forces affect the angular momentum of a rigid body. The intermediate concept here is "torque." For a single point particle at position $\vec{r}$, we defined the torque exerted by a force $\vec{F}$ as

$$\vec{\tau} = \vec{r} \times \vec{F}. \tag{4.70}$$

It was then simple to prove that

$$\frac{d\vec{L}}{dt} = \frac{d}{dt}\left(\vec{r} \times \vec{p}\right) = \frac{d\vec{r}}{dt} \times \vec{p} + \vec{r} \times \frac{d\vec{p}}{dt} = \vec{r} \times \vec{F} = \vec{\tau} \tag{4.71}$$

where we have used the fact that $d\vec{r}/dt = \vec{v}$ is parallel to $\vec{p} = m\vec{v}$, and so their cross product vanishes. Note that the $\vec{F}$ here represents the net force acting on the particle, and so the $\vec{\tau}$ at the end represents the net torque.

We can follow a similar derivation to get the corresponding law for a rigid body composed of many elementary particles. The total angular momentum is just

$$\vec{L} = \sum_{i=1}^{N} \vec{r}_i \times \vec{p}_i. \tag{4.72}$$

Taking the derivative with respect to time and following the steps taken for the single particle, we arrive at

$$\frac{d\vec{L}}{dt} = \frac{d}{dt} \sum_{i=1}^{N} \vec{r}_i \times \vec{p}_i = \sum_{i=1}^{N} \vec{r}_i \times \vec{F}_i = \sum_{i=1}^{N} \vec{\tau}_i = \vec{\tau} \tag{4.73}$$

where now the $\vec{\tau}$ at the end means the (doubly) net torque – i.e., the net torque on each particle (produced by the sum of all the individual forces acting on that particle), added up for all the particles.

We may simplify the expression for the net torque $\vec{\tau}$ somewhat by doing what we did in the very first section above – distinguishing between "external" and "internal" forces, and showing that the effects of the internal forces cancel out so that only the external force needs to be considered. To see how this goes here, we can write the net force acting on a given particle as

$$\vec{F}_i = \vec{F}_i^{ext} + \sum_{j \neq i} \vec{F}_{i,j} \tag{4.74}$$

where the first term on the right represents the total external force acting on particle i (that is, the sum of any forces acting on that particle which aren't exerted by other particles in the body), and the second term adds up the forces exerted on particle i by other particles in the body.

When we plug this into the expression for the total torque on the body, we get two terms

$$\tau = \sum_{i=1}^{N} \vec{r}_i \times \vec{F}_i^{ext} + \sum_{i=1}^{N} \sum_{j \neq i} \vec{r}_i \times \vec{F}_{i,j} \tag{4.75}$$

representing, respectively, the total torque produced by external and internal forces.

Now let us show that (or really, under what conditions) the second term vanishes. Note that it is a double sum – first we are summing over each of the particles, but then, for each particle, we sum over all of the other particles in the body. So for any given *pair* of particles (say, the 19th and 47th particles) there are two terms in the double sum (one when $i = 19$ and $j = 47$ and then another when $i = 47$ and $j = 19$). We can thus rewrite the double sum as a sum over all the possible *pairs* of particles, with two terms for each pair:

$$\tau^{int} = \sum_{pairs} \left(\vec{r}_i \times \vec{F}_{i,j} + \vec{r}_j \times \vec{F}_{j,i} \right). \tag{4.76}$$

That was the hard part. Now we can appeal to Newton's third law – $\vec{F}_{i,j} = -\vec{F}_{j,i}$ – to rewrite this as

$$\tau^{int} = \sum_{pairs} (\vec{r}_i - \vec{r}_j) \times \vec{F}_{i,j}. \tag{4.77}$$

It is then easy to see that – *if the forces exerted by a pair of particles on one another are along the line connecting the two particles* – the thing being summed is a cross product of two vectors that are *parallel*, so that each term in the sum (and hence the sum) is *zero*, and the internal forces won't contribute anything to the net torque on the body.

What is the status of this extra requirement? It is satisfied by the Newtonian gravitational force as discussed in the last chapter, and also by the electrostatic forces which play a central role in holding together the actual atomic and sub-atomic constituents of matter. And the upshot of this assumption about the microscopic forces – that internal forces won't contribute any net torque to a body – is certainly consistent with our macroscopic experience with everyday objects. For example, consider our friend the barbell sitting on a (frictionless) table. Suppose the two masses exerted (gravitational or electrical or whatever) forces on each other that were consistent with Newton's third law, but *not* along the line joining the two particles. Then there would be a net torque on the barbell, and it would "spontaneously" (without any intervention from the outside) start rotating! Since we never see something like this happen, it is good evidence that such internal forces don't in fact exist. Nevertheless, it is worth appreciating that this is an extra assumption and that exceptions can and do exist.

Let us summarize and put together the main points developed so far in this section. First, as with kinetic energy, the total angular momentum of a rigid body can be written as the sum of two terms – one pertaining to its translational motion (the "orbital" angular momentum) and one pertaining to the purely rotational motion (the "spin" angular momentum). Second, the net torque on an object – which under typical conditions will only get contributions from external forces – is equal to the rate of change of its total angular momentum. We may summarize all of this with the following equation:

$$\vec{\tau}^{ext} = \frac{d}{dt}\left(\vec{L}^{orbital} + \vec{L}^{spin}\right) \tag{4.78}$$

where $\vec{L}^{orbital}$ and $\vec{L}^{spin}$ were defined in Equations 4.63 and 4.64.

Let us now see how we can apply all of this to gain further insight into the example considered before: a ball (or other round object) rolling down a ramp. One of the important points is that a lot of the quantities defined above, for example the torque and the orbital angular momentum, are relative to the chosen coordinate system. For example a given force might exert no torque relative to one coordinate system, but nevertheless exert a large torque relative to some other coordinate system. So we will have to be very careful – very explicit – about how we are choosing our coordinates. This is not a bad thing, though; it's a good thing. For we can take advantage of the fact that all of the above is true for *any* coordinate system, to choose coordinate systems that make certain calculations easier.

Thus, consider first the situation shown in Figure 4.6. Notice that this Figure is doing two separate jobs: first, it's a picture of the situation (the ball on the ramp), and second, it's a free body diagram cataloging the forces which act on the ball (which is of course our object of interest). There are three relevant forces: the weight force $\vec{W}$ whose magnitude is Mg and which can be treated as acting at the center of mass of the ball, a normal force $\vec{N}$ which is directed perpindicular to the ramp and acts on the point of the

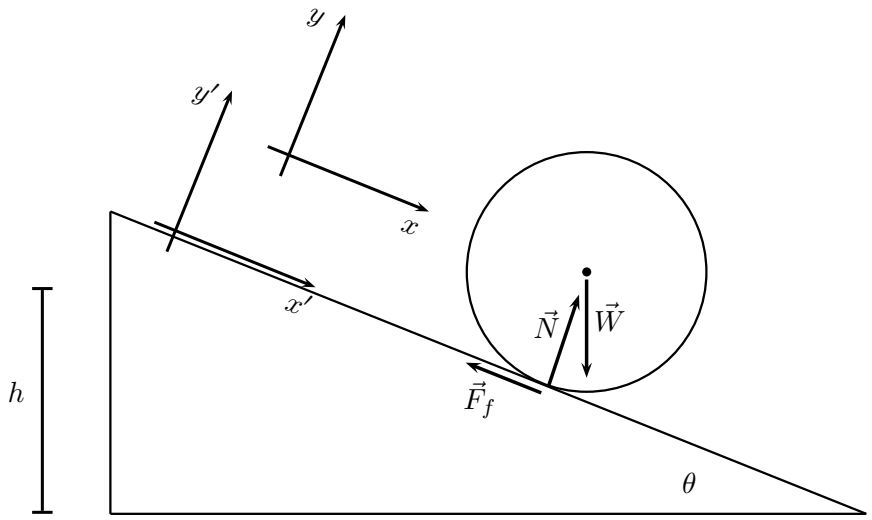

Figure 4.6: A ball rolling down a ramp, with forces shown as in a free body diagram.

ball which touches the ramp, and finally a friction force $\vec{F}_f$ whose direction is parallel to the ramp and which, like the normal force, acts on the point of the ball that contacts the ramp. Why should we include the friction force? Because we are interested in the case of a ball *rolling* down a ramp. If there were no friction, the ball wouldn't roll; it would simply *slide*. (You might be worried that this contradicts what we said above when we did this using energy conservation. The point is that the friction force exists, but, since the ball is rolling but not sliding, does no work. So the friction force must be included to understand the dynamics of the ball, yet it can be ignored when treating the same system using energy conservation.)

To begin with, let's use the coordinate system indicated by the x and y axes shown in the Figure. (We'll use the ones labeled x' and y' shortly.) The x axis is parallel to the ramp, so we can already say that the translational acceleration of the ball should be in the positive x-direction.

Let's first apply Newton's second law and see what we can say about the translational motion of the ball. We can read off from the free body diagram that the net force in the x-direction is the x-component of the weight force minus the (magnitude of the) friction force:

$$F_x^{net} = Mg\sin\theta - F_f. \tag{4.79}$$

According to Newton's second law, this should equal the ball's mass times the x-component of its acceleration. Thus

$$Mg\sin\theta - F_f = Ma. \tag{4.80}$$

Note that, since we don't actually know anything about the size of the friction force, we can't yet solve for the acceleration a, which is our ultimate goal. And bringing in the y-components of the forces won't help either. This will only tell us that the normal force

must just cancel the y-component of the weight force – interesting enough if we want to know how big the normal force is, but not helpful for our present goal.

Of course, the resolution is going to be that we need to consider the *rotational* motion of the ball. So let's do that. To begin with, notice that, *for our choice of coordinate system*, as the ball rolls down the ramp its center of mass coordinate $\vec{R}_{CM}$ and its overall velocity $\vec{V}_{CM}$ will be parallel. Hence, during the entire process, the ball has no orbital angular momentum:

$$\vec{L}^{orbital} = M\vec{R}_{CM} \times \vec{V}_{CM} = 0. \tag{4.81}$$

Its total angular momentum will therefore be equal to its spin angular momentum:

$$\vec{L} = I\vec{\omega}. \tag{4.82}$$

It is also worth noting that, since all of the motion is in the $x - y$ plane, the rotational quantities such as the angular momentum, angular velocity, and torque will be purely in the z-direction – which we can handle by treating them as scalars and using the sign (positive or negative) to indicate (respectively) a clockwise or counter-clockwise "sense." We may thus re-write the previous vector equation more simply as

$$L = I\omega \tag{4.83}$$

where now a positive L and ω will indicate what we expect on physical grounds: the ball will be rotating clockwise as it rolls.

Let's finally analyze the torque. To find the total torque on the ball, we need to add up the torques produced by all three of the forces. Actually, despite initial appearances, it will be simpler to break the weight force up into its x and y components, and hence calculate the torques due to *four* forces. To begin with, the x-component of the weight force produces no torque, because it acts at the center of mass point, which is itself purely in the x-direction relative to our chosen origin. The y-component of the weight force does exert a (positive, clockwise) torque. But the normal force (whose magnitude remember is just equal to the y-component of the weight force) exerts a precisely-cancelling (negative, counter-clockwise) torque. The easiest way to see this is to remember that the magnitude of the cross product $\vec{r} \times \vec{F}$ can be thought of as the magnitude of $\vec{F}$ times the component of $\vec{r}$ that is perpindicular to $\vec{F}$ – here this component is just the x-coordinate of the point on the ball where the normal force acts, which is of course just equal to the lever arm for the y-component of the weight force.

So at the end of the day, the net torque on the ball is just the torque produced by the friction force:

$$\tau = RF_f \tag{4.84}$$

where R is the radius of the ball (i.e., the component of $\vec{r}$ for the contact point that is perpindicular to the friction force). Notice that the torque is positive because it tends to rotate the ball in the clockwise direction.

We can finally put all of this together. The net torque should equal the time rate of change of the total angular momentum:

$$RF_f = \frac{d}{dt} I\omega = I\alpha \tag{4.85}$$

where α is the angular acceleration of the ball. If the ball is really rolling (without slipping), we must have $\alpha = a/R$. This is just the "rolling without slipping" condition that we discussed when we did this same problem using energy in the previous section – or technically, it is the time derivative of that earlier condition (which related the translational and angular velocities). So we have that $RF_f = Ia/R$, or

$$F_f = \frac{Ia}{R^2}. \tag{4.86}$$

This is precisely the extra information about the friction force that we needed, in order to solve Equation 4.80 for the acceleration. Plugging this last expression into that earlier equation, we get

$$Mg\sin(\theta) - \frac{Ia}{R^2} = Ma \tag{4.87}$$

which can be solved for the acceleration:

$$a = \frac{Mg\sin(\theta)}{M + \frac{I}{R^2}} = \frac{g\sin(\theta)}{1 + \frac{I}{MR^2}} \tag{4.88}$$

which is, happily, just what we found before using energy.

It is worth analyzing the same problem once again, in essentially the same way, but using a different coordinate system. This will show how, although lots of things change in the intermediate analysis, the final result turns out the same. So, consider now the same situation but using the x' and y' axes shown on the earlier Figure. The ball will now have *both* orbital and spin angular momentum:

$$L = MRV_{CM} + I\omega. \tag{4.89}$$

And we'll have to rethink the torques produced by the "4" forces, too. With the original coordinate system, only the friction force produced torque. But now the friction force is directed right back toward the origin and hence produces no torque! The torques produced by the normal force and the y'-component of the weight force still cancel. So the net torque is just that produced by the y'-component of the weight force:

$$\tau = RMg\sin(\theta). \tag{4.90}$$

Note that this torque is definitely *bigger* than what we said the net torque was (RF_f) when we were using the other coordinate system, because $Mg\sin(\theta)$ is definitely bigger than F_f – if it weren't, the ball certainly couldn't be acclerating down the ramp! So the different coordinate system really does result in all these quantities being different.

Yet still, as if by magic, everything still works out. For, even with this new coordinate system, we should have that the net torque equals the time derivative of the angular momentum:

$$RMg\sin(\theta) = \frac{d}{dt}\left(MRV_{CM} + I\omega\right) = MRa + I\alpha \tag{4.91}$$

which becomes

$$RMg\sin(\theta) = MRa + I\frac{a}{R} \tag{4.92}$$

when we bring in the "rolling without slipping" condition. This can be easily solved for a with the by-now expected result:

$$a = \frac{g\sin(\theta)}{1 + \frac{I}{MR^2}}. \tag{4.93}$$

It works out the same way despite the torque being bigger because that bigger torque now has to produce not only an increasing L_{spin}, but also an increasing $L_{orbital}$ (which, remember, was zero for the earlier coordinate system).

4.4 The Top

So far we have only considered examples of rotating bodies for which the rotation axis is constant. But the general formalism we have developed applies in the more general case, too. Let us therefore briefly explore the motion of the familiar toy, the top, to illustrate the power and generality of the results developed above. The amazing thing about a top, of course, is that once it is set spinning, it – as if by magic – refuses to tip over, but instead rotates around with a motion called "precession." To understand how this comes about, let's start by thinking about what happens to a non-spinning object (a ruler, say) if it is placed at an angle and then released.

Figure 4.7 shows the initial situation (in solid lines) and then the situation at some later moment (in dotted lines). We can understand the down-and-to-the-left acceleration of the center of mass point just by considering the down-and-to-the-left net force which is produced jointly by the weight, normal, and friction forces (shown in the figure). But in addition to *translating*, the ruler also *rotates*. This can be understood by considering the torque acting on the ruler, and how this causes its angular momentum to change. If we pick an origin at the point of contact between the ruler and the table, the two contact forces which act at that point will produce no torque. The net torque is then just the torque produced by the weight force. This torque is nonzero in magnitude, and out-of-the-page (counter-clockwise) in direction.

Equation 4.3 tells us that the net torque equals the rate of change of the total angular momentum. Hence, during some short period of time Δt, the angular momentum $\vec{L}$ will change by $\vec{\tau}\Delta t$. Since the ruler is initially at rest, it has zero initial angular momentum, and so its total angular momentum after Δt will be just $\vec{\tau}\Delta t$ – which of course points out-of-the-page. Since the ruler is pivoting about a fixed point, it will have both some orbital and spin angular momentum. The pivoting ruler is similar to the rolling ball, and there is some equation similar to the earlier "rolling without slipping" condition that relates the orbital and spin motions. All that matters here, though, is that the orbital and spin contributions to the total angular momentum will both be in the same direction, hence both of them are out-of-the-page. This means a counter-clockwise orbital motion of the center of mass about the pivot point, and also a counter-clockwise rotational motion of the ruler about its center of mass. Which is just a fancy way of describing the motion we already know will happen as the ruler begins to fall.

The point of discussing the ruler in that way is just to set up a contrast for the following discussion of a top. What is the difference between a (spinning) top and a

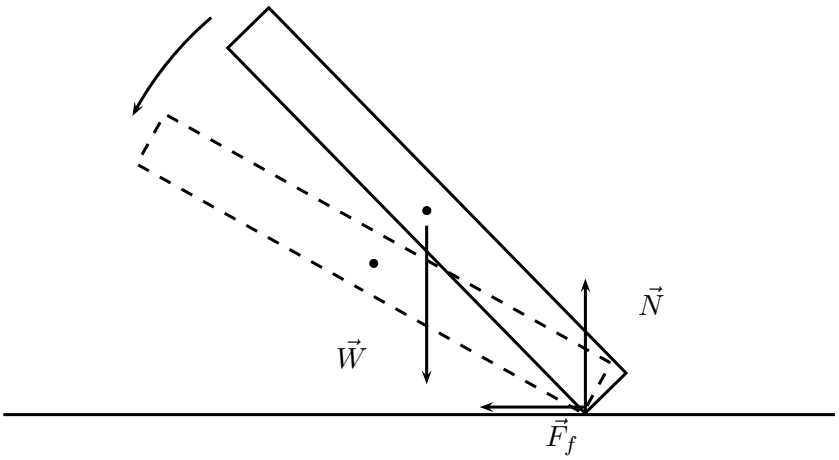

Figure 4.7: A ruler resting on a table, released from rest at an angle, begins to fall. The motion of the center of mass point can be understood on the basis of the three forces shown: a weight force, a normal force, and a friction force. These produce a net force in the "bottom-left" direction. But we can understand this – plus the purely rotational aspect of the falling motion – on the basis of torque and angular momentum, too. Taking the contact point between the ruler and the table as the origin, the normal and friction forces will produce no torque. There is therefore a net torque equal to the torque produced by the weight force. This torque is counter-clockwise or, as a vector, out-of-the-page. The initially stationary ruler has zero angular momentum, and so, at the end of some short period of time Δt, Equation 4.3 requires that the ruler have new total angular momentum $\vec{L} = \vec{\tau}\Delta t$, which will (like $\vec{\tau}$) point out-of-the-page. Since the ruler is pivoting about the fixed point of contact with the table, its total angular momentum will have both an orbital and a spin contribution, and these will be related by something like the "rolling without slipping" condition we've used in previous examples. The details aren't important. What matters is just that the basic dynamical equation for rotational motion requires that the ruler have, after Δt, both some out-of-the-page (counter-clockwise) orbital motion and some out-of-the-page (counter-clockwise) rotational motion – precisely as shown in the Figure and expected from common experience.

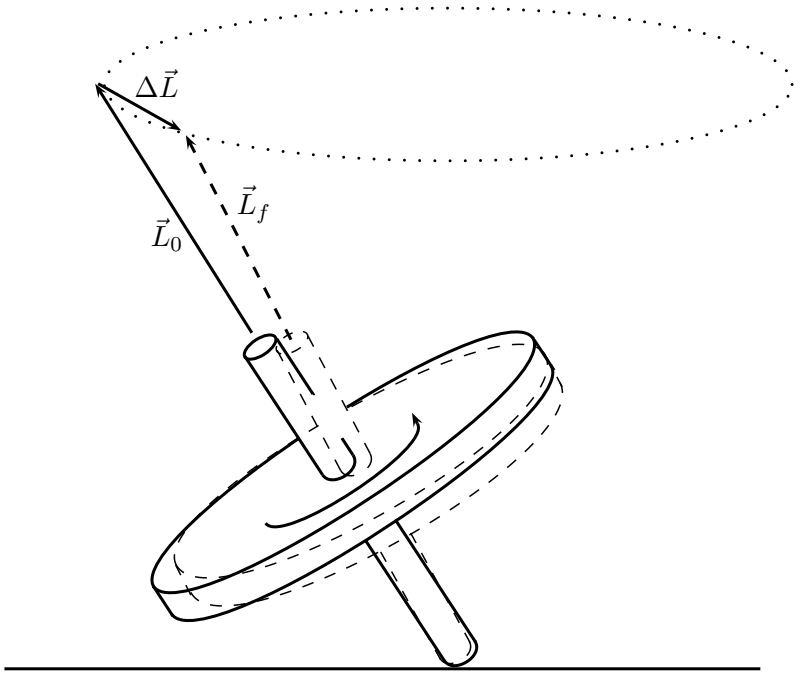

Figure 4.8: A spinning top, released in the same way as the ruler considered previously. Unlike the ruler, the top has some initial (spin) angular momentum. So although the forces and torques acting on the top are the same as the corresponding ones acting on the ruler, the new angular momentum after some short time Δt is quite different. It is, instead of *zero* plus $\Delta\vec{L} = \vec{\tau}\Delta t$, *the large spin angular momentum vector* – $\vec{L}_0$ in the Figure – plus $\Delta\vec{L}$. The new angular momentum vector – $\vec{L}_f$ in the Figure – is therefore (remember, Δt is small) of the same magnitude as the original one; it just points in a slightly different direction. The angular momentum vector will be again rotated in the next short time interval Δt, and the next, and the next – the net result being that the angular momentum vector (hence also the spin axis of the top) sweeps out a cone as suggested by the dotted-line circle in the figure. This motion – which you should demand to see in real life in class – is called "precession."

ruler? Only that the top is spinning – and so *already has considerable angular momentum* when it is let go. Let us run through the same sort of analysis, and see how this changes things.

To begin with, the forces and torques acting on the top are the same as they were with the ruler. With the origin of the coordinate system taken at the contact point between the top and the floor, only the weight force contributes to the net torque, and this is out-of-the-page. The key thing is that, where the ruler had *zero* angular momentum initially, the (spinning) top has a *lot* of (spin) angular momentum – as shown by the vector $\vec{L}_0$ in Figure 4.8. It is then required by the basic dynamical equation for rotation,

that the angular momentum vector after some short period of time Δt be

$$\vec{L}_f = \vec{L}_0 + \Delta\vec{L} \tag{4.94}$$

where, just as for the ruler, $\Delta\vec{L} = \vec{\tau}\Delta t$ is small and in the out-of-the-page direction.

It is then clear from the properties of vector addition that the new angular momentum vector $\vec{L}_f$ will have (in the limit of small Δt) the same magnitude as $\vec{L}_0$, but will point in a slightly different direction (a little bit toward us, out of the page). How is this possible, i.e., what does such a final angular momentum vector tell us about the new state of motion of the top?

Here is the first answer that should occur to you: the top should start tipping over, just like the ruler did. But this isn't right. Here's why. Suppose it did tip over like the top. The falling motion would have associated with it some little bit of angular momentum out-of-the-page, just as it did for the ruler. But in order to have fallen, the spin axis of the spinning top would have to have changed, i.e., the spin angular momentum vector would have to now point slightly more toward the horizontal than it did initially. This would imply a (contribution to) $\Delta\vec{L}$ in the down-and-to-the-left direction. But such a change in the angular momentum would require a *torque* in that same direction. And, simply put, there *is* no torque in that direction! So (in a perhaps not fully satisfying way!) that is the proof that the top, unlike the ruler, cannot tip over.

We can now start to see what it must actually do. The final angular momentum vector has the same magnitude as the initial one, and is just turned a little bit in direction. This will be achieved if the top doesn't fall at all, but instead just turns its orientation, as shown in the Figure. And then the situation is the same as it was initially (but just now oriented in a slightly different direction) so the same thing happens again in the next short period of time Δt. And so on. The result is a continuous re-orientation of the spin axis around a cone, as suggested by the dotted line path in the figure. This motion is called *precession*.

As already mentioned, the fact that a spinning top doesn't fall, but instead precesses, seems somehow magical (or at least very counterintuitive), and the above sort of formal analysis somehow leaves one not fully satisfied. There should be no shame in acknowledging this. Actually, what the above "formal analysis" shows is really just that the precessional motion described is a consistent steady state solution. The non-obvious aspect that is so counter-intuitive is: how does the top get into this precessional steady state in the first place?

The answer to this is subtle and complicated, but gesturing vaguely in the direction of some of the subtleties can at least give one a sense that there is nothing magical happening. To begin with, when the top is first released, it *does* actually fall just a tiny bit, at least for a split second. We argued above that this is impossible, but actually it isn't. We neglected to mention it before, but the precessional motion of the top actually implies that the top posseses not only *spin* angular momentum, but also some *orbital* angular momentum. For the orientation of the top shown in the figure, the center of mass velocity associated with the precessional motion will be out-of-the-page, which implies (by the right hand rule) an orbital angular momentum that is up-and-to-the-right. So

actually there is no question of finding some torque to account for the down-and-to-the-left $\Delta\vec{L}$ associated with a tiny bit of initial falling. What actually happens is that, while falling just a bit, the top *converts* a little bit of its *spin* angular momentum into *orbital* angular momentum. The *total* angular momentum is actually *constant*, so no external torques are required to explain a change in $\vec{L}$. There is only some relatively complicated story about internal forces (exerted on and by, for example, the ball bearings that connect the spinning part of the top to its axis) rearranging the overall distribution of total angular momentum.

But that isn't precisely true either! It's true that, when first released, the top falls just a bit, trading some of its spin angular momentum for orbital angular momentum associated with the precessional motion. But the top actually falls a little bit *too far* and overshoots the orientation at which it could stably precess. Something like the same story then happens in reverse. The result is that the spin axis of the top exhibits *another*, secondary sort of precession – it precesses in a (smaller) cone centered on the (moving!) precession axis already discussed. This secondary wobbling motion is called "nutation" and can easily be observed with a real top. Of course, for a real top there will always be a little bit of friction associated with the contact "point" – this works to damp out the nutational motion, helping the top achieve the smooth, steady state precessional motion we discussed at the beginning.

Of course, in time, the same sorts of frictional effects will reduce the magnitude of the spin angular momentum and change the character of the orbital (precessional) angular momentum – until, for example, the spinning part of the top hits the table, substantial new forces are introduced, and the magical motion turns more mundane. It is definitely worth spending some time with a top or gyroscope to observe – and contemplate – some of these effects.

Let us finally calculate the *rate* of the precessional motion (assuming it has reached a constant, steady-state rate). We have already argued that in a short time period Δt, the angular momentum of the top will change by an amount

$$\Delta\vec{L} = \vec{\tau}\,\Delta t. \tag{4.95}$$

The component of its (spin) angular momentum that is in the horizontal plane will be

$$L_{horizontal} = I\omega\sin\phi \tag{4.96}$$

where ϕ is the angle the top's spin axis makes with the vertical. The precessing top thus sweeps through an angle

$$\Delta\theta = \frac{\Delta L}{L_{horizontal}} = \frac{\tau\Delta t}{I\omega\sin\phi} \tag{4.97}$$

in time Δt. Which means that its precessional angular velocity $\Omega = \Delta\theta/\Delta t$ is

$$\Omega = \frac{\tau}{I\omega\sin\phi}. \tag{4.98}$$

This implies a precessional *period*

$$T_{prec} = \frac{2\pi}{\Omega} = \frac{2\pi I\omega\sin(\phi)}{\tau}. \tag{4.99}$$

For a top, the magnitude of the torque τ is given by

$$\tau = MgR\sin(\phi) \tag{4.100}$$

where R is the distance from the pivot point to the center of mass. This gives

$$\Omega = \frac{MgR}{I\omega}. \tag{4.101}$$

or a precessional period of

$$T_{prec} = \frac{2\pi I\omega}{MgR} \tag{4.102}$$

Note that for a given ω, the precession rate (or period) is independent of the tilt angle ϕ. And – more interestingly – the precession rate is inversely proportional to the spin rate ω. So, for example, as friction works to steadily decrease the spin rate ω, the precession rate Ω will *increase*. Or in terms of the period, as ω decreases (due to friction, say), the top will take less and less time to precess. This feature too is readily observable with a real top.

4.5 Newton's Spherical Shell Theorem

So far in this chapter we've been exploring the implications of some basic postulates about point particles – Newton's laws of motion, and definitions of kinetic energy, torque, and angular momentum – when we consider extended (especially rigid) bodies that are composed of many such point particles.

In this section, we address a similar question about Newton's other law – the law of universal gravitation that we discussed in the previous chapter.

What exactly is the issue or question here? We can begin by formulating it this way: if we take the gravitational inverse-square force law – Equation 3.18 or if you prefer Equation 3.19 – as a statement of the gravitational force between two *point particles*, what will be the gravitational force exerted by (and/or on) an extended *collection* of point particles, such as a planet?

You should perhaps be a little puzzled about how this question could be coming up at this point. Didn't Newton get the inverse-square-law in the first place by examining (among other things) the behavior of the planets? So how could there be any question about whether it should apply to them?

It's a good question! In a way, there should be no question about whether an inverse-square-law force toward the Sun is responsible for the motions of the planets. But remember that a crucial part of Newton's theory was that this force was *gravitational* – i.e., the same sort of force that the Earth exerts on apples and (he argued) the Moon. And a big part of the argument for *that* aspect of the theory was the claim that the Earth's gravitational influence falls off as the inverse-square of the distance from its center – a point he argued for by, you'll recall, comparing the accelerations of the apple and the Moon, and noticing that they are in the ratio 1:3600, i.e., the inverse square of their respective distances from the center of the Earth.

But in what sense exactly is the apple one Earth radius away from the Earth? The Earth is composed (presumably) of many tiny massive particles, distributed (presumably) more or less uniformly throughout its whole spherical volume. And the whole idea of the gravitational force is supposed to be that each one of these constituent particles exerts its own gravitational force on the apple – the *net* gravitational force on the apple then being the vector sum of all of these little forces, which have *all sorts of different magnitudes and directions*. So it is *not at all obvious* that the net gravitational force on the apple (assuming gravitation works the way Newton hypothesized, for the Earth's and apple's constituent particles) can be written as

$$F = \frac{GM_{earth}m_{apple}}{R_{earth}^2} \tag{4.103}$$

where M_{earth} is the total mass of the Earth (i.e., the sum of the masses of all of the Earth's little constituent particles) and R_{earth} is the Earth's radius (i.e., the distance between the apple and the *center* of the Earth).

This requires proof. And that proof requires calculus. Interestingly, there is some historical evidence that Newton had hit on the basic idea of his theory of gravitation one or even two decades prior to the publication of the *Principia*, but delayed the publication of his ideas precisely because he lacked the mathematical tools needed to complete this proof, i.e., to fill in this one logical gap in the argument as it was presented in the previous chapter. (So kudos to you if you noticed the gap last week!) Of course, during that time (and among other things!) Newton *invented integral calculus* to solve precisely this problem. This should no doubt add to your assessment of the grandeur of Newton's achievements as a scientist. He set the bar for himself extremely high when it came to producing really rigorous, conclusive arguments, with all the i's dotted and t's crossed, for his proposed conclusions. And then he worked tremendously hard to live up to his own high expectations.

Back to the actual proof. What we want to establish is that a spherical blob of total mass M (such as the Earth) acts, gravitationally, the same as a *point mass* of mass M, located at the center of the sphere, would. We can make this slightly simpler by noting that a spherical blob of mass can be thought of as a collection of a bunch of concentric *spherical shells*. So if we can prove that a single spherical shell of mass M exerts the same gravitational force (at least on objects outside the shell) as would a point mass M located at the center of the shell, the corresponding point for spherical blobs will follow. So this is what we will prove.

Let's warm up to it, though, by considering first a simpler problem, for the sake of laying out the overall approach: what is the nature of the gravitational force exerted by a *line segment* of length L and mass M, on a mass m located some distance d away from the segment's center (and, for simplicity, along a line perpindicular to the segment itself)? See Figure 4.9.

As shown in the Figure, we'll pick a coordinate system with the x-axis along the line segment and with its origin at the center of the segment. Our goal is to then calculate the force exerted by the line segment on a point of mass m located some distance d away along the y-axis. Pay careful attention to the way this is set up, because that is the real

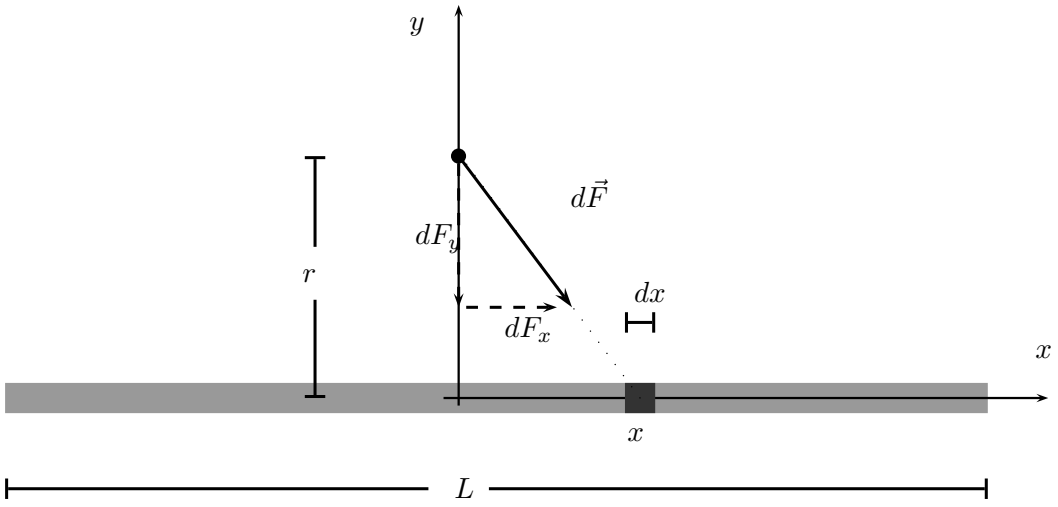

Figure 4.9: How to calculate the gravitational force exerted on a point mass m by a line segment of length L and total mass M.

point of going through this (intrinsically uninteresting) example first. Now let's focus our attention on one little (and non-special) piece of the line segment – say, the piece of width dx located at position x. Note that, since we are treating dx as infinitessimally small, we can treat this piece as a point mass. Assuming the total mass M is uniformly distributed between $x = -L/2$ and $x = L/2$, the mass of this little piece will be the same fraction of M as dx is of L: $dM = dx\, M/L$. Since this little piece can be treated as a point, the gravitational force $d\vec{F}$ it exerts on the mass m will have magnitude

$$dF = \frac{G\, dM\, m}{r^2} \tag{4.104}$$

where r, the distance between the two point masses, is given by $r = \sqrt{d^2 + x^2}$. and direction as shown in the figure, along the line from m to dM.

A little thought should convince you that, when we add up the forces on m due to all of the different pieces of the line segment, the x-components will add up to zero and only a y-component will remain. (For each piece on the right side that gives a positive dF_x, there is a corresponding piece on the left that contributes an exactly-cancelling negative dF_x.) So to find the net force on m, we need only consider the y-component of the force $d\vec{F}$. This is given by

$$dF_y = dF \frac{d}{r} \tag{4.105}$$

which, plugging everything in, reduces to

$$dF_y = \frac{GdMm}{r^2}\frac{d}{r} = \frac{GMmd}{L}\frac{dx}{(x^2 + d^2)^{3/2}}. \tag{4.106}$$

Arguably there should be a minus sign on the right hand side since the y-component of the force is in the negative y direction. But it is already obvious from the *attractive* character of the gravitational force that the net force is going to be *toward the line segment*. We should be able to remember that without needing to confuse the equations with explicit minus signs.

What remains is now only to add up – by *integrating* – all of the little contributions dF_y from all of the little pieces of the line segment, to find the total gravitational force exerted on the mass m by the segment. Note, though, that what we've done so far is the hard part, and the part that actually takes some understanding of and careful application of ideas from physics. This is not the kind of problem that we'll focus on much in this course, but calculations very similar to this will show up *all the time* if you continue taking more advanced courses in physics. So it is worth formulating a very important principle here: once you get the idea of this sort of calculation, it becomes second-nature to recognize, right away, that you are going to have to perform an integral to get the answer. The crucial lesson is: *don't try to write down an integral right away.* Instead, do what we did here: focus on *one little piece* of whatever is (eventually) going to be integrated over, and develop *slowly and carefully* an expression for its contribution to the thing you (eventually) want to know the total of. *Only after getting this expression, and being absolutely certain that you've gotten it correctly, should you perform the now purely mathematical and (assuming you can do calculus) relatively trivial step of adding up – integrating – all of the tiny contributions to find the total.*

Since we've done the hard part, we can follow our own advice and now, finally, integrate to find the total force on m:

$$F_y = \int dF_y = \int_{-L/2}^{L/2} \frac{GMmd}{L} \frac{dx}{(x^2+d^2)^{3/2}}. \tag{4.107}$$

The constants can be taken outside the integral, and we are left with

$$F_y = \frac{GMmd}{L} \int_{-L/2}^{L/2} \frac{dx}{(x^2+d^2)^{3/2}} = \frac{GMm}{d\sqrt{d^2+L^2/4}} \tag{4.108}$$

where the integral can be done by looking it up in a table or making a trig substitution.

It is always worth pausing at the end of a calculation like this and asking: does the result make sense? First of all, does it have the right units? We know that Newton's constant G times a mass times another mass, divided by a distance squared, will give a force – because that is the form of the basic gravitational force law postulated for point particles. And that is indeed what we have in the above expression.

So far so good. What about various physical limits in which we can intuitively reason out what the answer should be? For example, suppose the line segment is really really long: $L \to \infty$. Thinking about that physically, the same total mass M is being distributed ever more "thinly" across a longer and longer line segment. Since the force between m and one of the tiny pieces of the segment falls off as $1/r^2$, only the part of the segment near $x = 0$ should contribute appreciably. But this region contains a vanishing fraction of the total mass M as we let $L \to \infty$. Plus, the parts of the

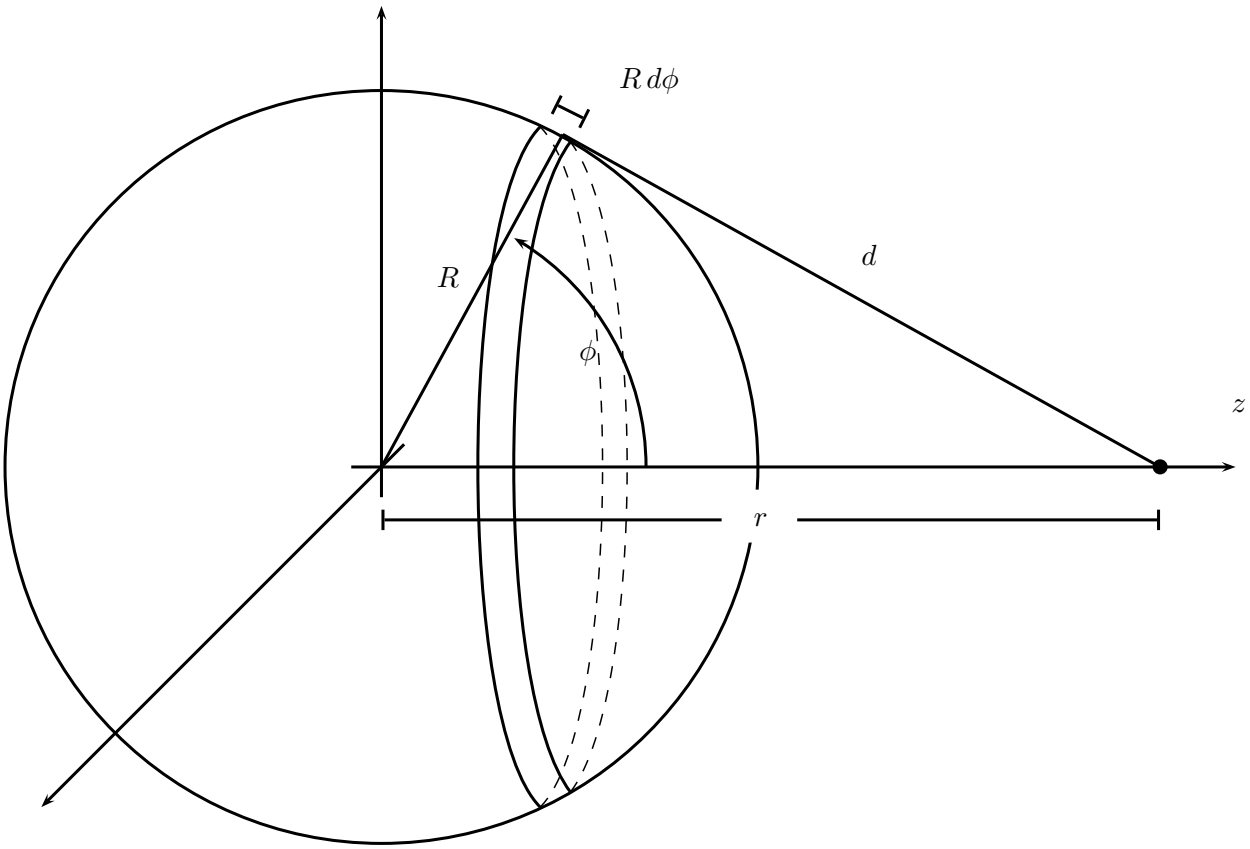

Figure 4.10: How to calculate the gravitational force exerted on a point mass m by a spherical shell of total mass M and radius R.

segment off on the two sides tend increasingly to pull m in opposite directions, resulting in increasingly cancelling contributions to the total force. So it seems that the total force should probably go to zero in this limit. And that is precisely what happens if we take the $L \to \infty$ limit of Equation 4.108.

What if we consider being very very far away from the line segment, i.e., $d \to \infty$? Intuitively, if we are very very far from the line segment (much farther than it is long), the fact that it is a line segment instead of a point should stop mattering, and we should get back the inverse-square-law expression for the force between two points. And again, that is precisely what we get: for $d \gg L$, we can neglect the $L^2/4$ inside the square root sign compared to the d^2, and we get back that the force is GMm/d^2.

So it seems reasonable to believe that we calculated the force correctly.

Of course, we didn't actually care about the force due to a line segment. We want to know the force produced by a uniform spherical shell. So let's set that up and work through it, following much the same procedure.

Consider the setup as shown in Figure 4.10. The spherical shell has total mass M

and radius R, and we assume it is symmetrical – i.e., the mass is distributed uniformly over the surface. We want to calculate the total gravitational force exerted by the shell on a point mass m located a distance r away from the shell's center. We'll start by picking out a small piece of the shell to focus our attention on. To begin with, we might pick any old point on the surface, a distance d from the point mass m. If the mass of that piece is dM, then the magnitude of the (attractive) force between the piece and the point mass will of course just be

$$|d\vec{F}| = \frac{GdMm}{d^2} \tag{4.109}$$

the z-component of which is

$$dF_z = \frac{GdMm}{d^2}\frac{r - R\cos(\phi)}{d}. \tag{4.110}$$

One recognizes, however, that there will be a whole *ring* of similar pieces with the same distance d from the point mass. Each small part of the ring will contribute the same amount to the z-component of the force on m. And all of the *other* components will cancel out – for each little piece of the ring that contributes some force in the x-direction, for example, there is some corresponding piece on the opposite side of the ring that contributes a precisely-cancelling component. So, conveniently, the total force exerted *by the ring* on the point mass m will be given by Equation 4.110 with dM now the mass of the entire ring.

That's a good start. Let's now work out an expression for dM in terms of the angle ϕ. The ring is supposed to have angular width $d\phi$, which means its actual width (in inches) is $R\,d\phi$. The ring is essentially a long skinny rectangle (wrapped around the surface of the sphere), so its total area dA will be its width times the "circumference" $2\pi R\sin(\phi)$:

$$dA = 2\pi R^2 \sin(\phi)d\phi. \tag{4.111}$$

The mass dM should then be the same fraction of the total mass M, as the area dA is a fraction of the sphere's total surface area $A = 4\pi R^2$:

$$dM = M\frac{dA}{A} = M\frac{2\pi R^2\sin(\phi)d\phi}{4\pi R^2} = \frac{M}{2}\sin(\phi)d\phi. \tag{4.112}$$

We can then plug this expression for dM into Equation 4.110:

$$dF_z = \frac{GMm}{2}\sin(\phi)d\phi\frac{r - R\cos(\phi)}{d^3}. \tag{4.113}$$

Finally, we can use the law of cosines to write the distance d in terms of R, r, and ϕ:

$$d^2 = R^2 + r^2 - 2Rr\cos(\phi). \tag{4.114}$$

The resulting expression for the force is:

$$dF_z = \frac{GMm}{2}\frac{r - R\cos(\phi)}{\left(R^2 + r^2 - 2Rr\cos(\phi)\right)^{3/2}}\sin(\phi)d\phi. \tag{4.115}$$

That's the hard part. Now we just remember that we were interested in the force produced by the whole spherical shell, not just this ring. So we want to now add up the contributions to the force produced by all the rings between $\phi = 0$ and $\phi = \pi$ which compose the shell. That is easily accomplished by integrating:

$$F_z = \int dF_z = \int_0^\pi \frac{GMm}{2} \frac{r - R\cos(\phi)}{\left(R^2 + r^2 - 2Rr\cos(\phi)\right)^3/2} \sin(\phi) d\phi. \tag{4.116}$$

The integral is tricky, so it's worth spending a moment talking about how to do it. The easiest way is probably to first make a substitution: $u = r - R\cos(\phi)$. Then $du = R\sin(\phi)d\phi$ and the expression from the denominator can be rewritten

$$R^2 + r^2 - 2Rr\cos(\phi) = R^2 - r^2 + 2ru. \tag{4.117}$$

Finally, when $\phi = 0$, $u = r - R$ and when $\phi = \pi$, $u = r + R$. So the force produced by the shell can be re-written

$$F_z = \frac{GMm}{2R} \int_{r-R}^{r+R} \frac{u\, du}{\left(R^2 - r^2 - 2ru\right)^{3/2}}. \tag{4.118}$$

This integral is now a little more manageable. It can be found in standard integral tables (or done using more clever substitutions). The result is:

$$F_z = \frac{GMm}{r^2} \times \frac{1}{2} \left(\frac{r + R}{\sqrt{(r+R)^2}} + \frac{r - R}{\sqrt{(r-R)^2}} \right). \tag{4.119}$$

It may look like the mess in parentheses is just 2, which would cancel the 1/2 and give us exactly the result we were hoping for. But it's a little more subtle – and a little more interesting – than that. Remember that the square root function always gives back a positive answer. Since r and R are both positive, when we add them, we necessarily get something positive. And the square root of the square of this positive number will therefore be the same as the number itself. So the first term in parentheses is just 1. But, for the second term, $r - R$ can be *either* positive or negative: it is positive if $r > R$ (i.e., if the point mass m is outside the spherical shell, as it was originally shown in the figure) – but negative if $r < R$ (i.e., if the point mass m is inside the spherical shell). In the first case, the same argument as before tells us that the second term in parentheses is also 1, and so the stuff in parentheses will total 2. But in the latter case, $r < R$, the second term in parentheses will be -1, and the stuff in parentheses will total *zero*.

So we may write the final result, for the force exerted by a spherical shell of mass M and radius R on a point mass m located a distance r away from the center of the shell, this way:

$$F = \begin{cases} 0 & r < R \\ -\frac{GMm}{r^2} & r > R \end{cases} \tag{4.120}$$

where the minus sign reminds us that the force is attractive, i.e., back toward the sphere.

This is a really remarkable result. The first wonderful feature of this result is that, for points *inside* a symmetric spherical shell of mass, the net gravitational force is *zero*. This

isn't as important as the other half of the result, just because we are usually interested in the gravitational effect of (say) planets on things outside, rather than inside, the planet. But still, this half of the result is interesting if only because it is surprising. One probably would have guessed that, since the gravitational force gets weak fast as one gets farther from the mass exerting it, one would be attracted to whichever part of the shell one was closest to. That reasoning is right, but leaves out something important: when one is close to a given part of the shell, there is only a little bit of it that is pulling one toward it. On the other hand, practically the whole other side is pulling one toward it. Granted, it's further away, so each little piece of it isn't pulling as hard – but there are, evidently, just enough more such little pieces that their net effect is precisely as big as the effect of the (fewer but stronger-pulling) nearby pieces. Thought of this way, the result has a certain intuitive plausibility to it. But still, what we learn from doing the rigorous calculation – that the cancellation is exact and the force exactly zero – is amazing.

Second, and more importantly: for points outside the spherical shell, the spherical shell acts just like a point mass with the same total mass as the shell and located at the center of the shell. And, as discussed a while back, this implies that a symmetrical spherical blob of mass (such as the Earth) will do the same. We will close this chapter by quoting Newton's summary of this point:

> "After I had found that the gravity toward a whole planet arises from and is compounded of the gravities toward the parts and that toward each of the individual parts it is inverse[ly] proportional to the squares of the distances from the parts, I was still not certain whether that proportion of the inverse square obtained exactly in a total force compounded of a number of forces, or only nearly so. For it could happen that a proportion which holds exactly enough at very great distances might be markedly in error near the surface of the planet, because there the distances of the particles may be unequal and their situations dissimilar. But at length ... I discerned the truth of the proposition dealt with here."

To summarize and close this section and this chapter, we've shown for the gravitational inverse-square law just what we started out by showing for the laws of motion (in particular $F = ma$): it is possible and consistent to postulate these laws as applying primarily to the ultimate, "atomic" particles of which ordinary macroscopic objects (like apples and planets) are composed. The applicability of the laws to these larger objects can then be established by means of mathematical theorems.

But there is one difference between the two cases which is worth mentioning. The laws of mechanics (such as $F = ma$) will apply to any aggregate body. There were essentially no extra assumptions in the derivation which would qualify the result. On the other hand, the proof that extended objects obey a gravitational inverse-square force law (if their constituent particles do) relied on the assumption of *spherical symmetry*. What we technically proved is that a spherically symmetric body acts (on other bodies outside it) just as would a point mass (with the same total mass and located at the real body's center). But – as we saw explicitly with the "warm-up" example of the massive line segment – a non-spherically-symmetric body will *not* produce simple inverse-square

forces on objects outside it, nor will it respond exactly as an equivalent point mass would to gravitational forces exerted on it. This turns out to have interesting implications and applications in the context of astrophysics, some of which we will take up in the following chapter.

Questions for Thought and Discussion:

1. Why is the "orbital angular momentum" called "orbital"? Consider the motion of the Earth using the Sun as the origin of a coordinate system. Does the Earth have any orbital angular momentum? Does it have any spin angular momentum? Are these in the same direction? How do their magnitudes compare?

2. Can you think of a situation in which an object's spin angular momentum is (at least partially) converted into orbital angular momentum? How about vice versa?

3. A man on a motorcycle is riding a "wheelie." Can the wheelie be maintained if he moves with constant velocity? Explain why or why not.

4. When you slam on the brakes in your car, the front end dips down a bit toward the ground and the back end rises up. (A more extreme example of the same phenomenon is braking while riding a bike – if you brake too fast, you can spill forward right over the handlebars.) Relatedly, when you step hard on the gas, the opposite happens: the front end lifts up a little and the back end dips down. Explain this effect using the concepts of torque and angular momentum.

5. Why do many sports cars (and all serious drag racers) have their engines mounted in the back of the car rather than the front?

6. A rocket ship in outer space accelerates from rest along a straight line. Take a coordinate system whose origin is not along the ship's trajectory. Does the ship develop any angular momentum as it accelerates? What force produces the necessary torque?

7. A car accelerates from rest. Take a coordinate system with its origin on the ground. Does the car develop any angular momentum as it accelerates? What force or forces produce the necessary torque? (This is subtle and probably surprising.)

8. Are the angular momentum and the angular velocity of a rigid body necessarily parallel? Give an example where they are not.

9. Are the orbital and spin angular momenta of a rigid body necessarily parallel? Give an example where they point in precisely opposite directions.

10. Consider a single planet orbiting a star. Technically, unless the star is infinitely more massive than the planet, the star is not quite stationary, but itself orbits around the (stationary) center of mass point of the star-planet system. Now suppose that the gravitational influence of the star on the planet propagates through

the intervening space at some finite speed, such that the gravitational force on the planet *now* is not toward the position of the star *now*, but rather toward where the star was a little while ago. And vice versa for the force exerted on the star by the planet. What is the qualitative effect on the motion of the two bodies? Is the total angular momentum of the two-body system constant in time? (Actually, there exist gravitational analogues of magnetic forces which almost exactly cancel out this anomalous, total-angular-momentum-changing component of the gravitational force. So don't take this question too seriously as a hint at what actually happens. It is rather only an opportunity to think about a plausible situation where forces exerted by two objets on each other may not be exactly co-linear.)

11. Consider two tops that are identical in shape, but different in size. They are set spinning with the same spin angular velocity ω. Which will have the higher precessional angular velocity Ω, the bigger one or the smaller one?

12. We proved that the gravitational force exerted by a spherical shell of mass, on a point mass in its interior, is zero. Does this mean you can build an anti-gravity shield by crawling inside a large ball or balloon?

13. We proved that the gravitational force exerted *on* a point mass *by* a uniform spherical shell of mass, is the same as would be exerted by a point mass, located at the center of the shell, and with the same total mass of the shell. In short, we showed that spherical shells of mass act just like point masses in so far as their *production* of gravitation is concerned. But what about their *response* to gravitation? Can anything be said about the force exerted, say, *by* a point mass *on* a spherical shell?

Projects:

4.1 Consider an extended body made of many point masses glued together. Show that, in a *uniform* gravitational field (like that near the surface of the earth), the total gravitational force on the body can be treated as acting at the body's center of mass even though, in fact, there is some small gravitational force on each of the body's constituent particles. Use this to explain why dropped objects do not spontaneously start rotating as they fall.

4.2 Could an extended body experience a net torque (about its center of mass) from a *non-uniform* gravitational field? Sketch and discuss an example.

4.3 Calculate the moment of inertia of a spherical shell of mass M and radius R.

4.4 Calculate the moment of inertia of a solid sphere of mass M and radius R.

4.5 A painter is standing on a piece of scaffolding near the side of a building. The platform is a long board, attached to the scaffolding at each end. Suppose the man weighs 200 pounds and that the weight of the board is negligible compared to this.

How big are the forces exerted on the board by the scaffolding (on both ends) if the man stands right in the center of the board? What if he stands 3/4 of the way toward one side?

4.6 A ladder is propped up against the side of a house. It makes a 30° angle with the vertical. The top end of the ladder is slippery, so the side of the house exerts a (horizontal) normal force on it, but no (vertical) friction force. The ground exerts both a (vertical) normal and a (horizontal) friction force on the bottom end of the ladder. What must be the coefficient of friction between the ladder and the ground in order for this to be a stable situation? Now qualitatively, what happens when somebody gets on the ladder? Is the ladder more likely to slip out with a person near the bottom or near the top?

4.7 A big spool of wire turns about a frictionless axle. Its moment of inertia about this axis is I, and its radius is R. A mass M is hung from the wire, which hangs down and slowly uncoils the spool. What is the angular acceleration of the spool? (If your answer is that $\alpha = RMg/I$ that's a good start: it's approximately correct if M is really small. But try to find the exact answer that is correct even if M is big. Also, for fun, you could try doing this at least a couple of different ways – e.g., using energy and using torque and angular momentum.)

4.8 An ice-skater is twirling with her arms out. She then pulls her arms in. Assume that the contact between her skates and the ice is frictionless. (Or if you don't buy that, have it be a twirling, spacewalking astronaut.) What is the net torque on the ice-skater as she pulls her arms in? By what amount should her angular momentum change? Make some reasonable assumptions about her size, and estimate the fraction by which her angular velocity will increase when she pulls her arms in. Does her rotational kinetic energy go up, go down, or stay the same? If it changes, where does the extra/missing energy come from / go?

4.9 Re-do the calculation of the gravitational force of a line segment, using the angle ϕ (between the one tiny piece of the segment and the center of the segment, as seen from the point where the mass m is located) as the main variable rather than x.

4.10 Measure/estimate the actual precession rate of an actual top, and use estimates/calculations of its moment of inertia to estimate its spin rate ω. Or devise some way of measuring/estimating ω in order to *predict* the precession rate Ω, and then check the prediction.

4.11 Use the fact that the gravitational potential energy between two point masses is given by

$$U = -\frac{GMm}{r} \tag{4.121}$$

to calculate the potential energy of a mass m near a line segment of total mass M, as in the example in the text. After doing the integral to get the potential energy, take the derivative with respect to d and see if you get what you should.

4.12 Calculate the gravitational potential energy between a point mass m and a symmetric spherical shell of radius R and total mass M. Sketch a graph of the result. Take the derivative to calculate the force as a way of verifying that your calculation was correct.

4.13 Suppose (only mildly contrary to fact) that the Earth is a solid sphere of uniform density. And suppose that a big hole were drilled straight through it, passing through the center. Now consider the force exerted by the whole Earth on some "test particle" (like an apple, say) at some point in this hole, say a distance $r < R_{earth}$ from the center of the planet. Work out an expression for the force's dependence on r. What sort of motion would the apple undergo if you dropped it in the hole?

Chapter 5

Astrophysical Applications

The last two chapters have explored two rather different topics: (i) Newton's discovery of the law of universal gravitation, according to which massive particles exert an inverse-square-law gravitational force on one another and (ii) the rotational dynamics of extended (and in particular rigid) bodies. Our goal in this present chapter is to develop some applications of and connections between these ideas, by surveying a number of interesting discoveries from the period between Newton and today in which gravitation and/or rotation play some interesting role. We focus on applications from astronomy and astrophysics, considering (in proper Newtonian spirit) the Earth as part of the heavens. Especially the latter parts of this chapter depart somewhat from the earlier practice of explaining always not just "what is true" but also "how it was figured out." We instead just survey some interesting conclusions that have emerged from more recent research, without giving all of the historical background that would make every detail clear. One goal is to sketch some of the ways the two foundational topics of the previous chapters play an important role in these more recent discoveries. Another goal is simply to tempt you to want to learn more about these things in more advanced coursework.

5.1 The Shape of the Earth

As discussed in Chapter 1, the Ancient Greeks knew, some 2000 years prior to Newton's theory of gravitation, that the Earth was a sphere. And it is – *approximately*. Of course the surface of the Earth is marked with hills, valleys, and mountains. Such features are produced by terrestrial causes such as volcanoes and erosion; they are not the departures from perfect sphericality that will concern us here, as they have nothing (or at any rate, less) to do with rotation and gravitation.

Instead we will focus on an interesting systematic departure of the Earth from perfect sphericality: it "bulges" slightly at the Equator and is in fact a slighly *oblate spheroid*. The oblateness can be quantified this way: the "radius" of the Earth at the poles is a bit less than the "radius" at the equator. The difference is small compared to the radius itself, but surprisingly big on human scales. It is about 21.3 kilometers, or about 13

miles, or about 0.335% of the Earth's average radius:

$$f = \frac{R_E - R_P}{R} = .00335. \tag{5.1}$$

The earliest observational evidence pertaining to the Earth's oblateness was acquired in the 1600's and was noted, by Newton, in the Principia:

> "some astronomers, sent to distant regions to make astronomical observations, have observed that their pendulum clocks went more slowly near the equator than in our regions. And indeed M. Richer first observed this in the year 1672 on the island of Cayenne. For while he was observing the transit of the fixed stars across the merdian in the month of August, he found that his clock was going more slowly than in its proper proportion to the mean motion of the sun, the difference being [2 minutes and 28 seconds] every day. Then by constructing a simple pendulum that would oscillate in seconds as measured by the best clock, he noted the length of the simple pendulum, and he did this frequently, every week for ten months. Then, when he had returned to France, he compared the length of this pendulum with the length of a seconds pendulum at Paris (which was 3 Paris feet and 8 3/5 lines long) and found that it was shorter than the Paris pendulum, the difference being 1 1/4 lines."

The idea here is that the period of a pendulum depends on its length and the local acceleration of gravity, g, according to

$$T = 2\pi\sqrt{\frac{L}{g}}. \tag{5.2}$$

Astronomers had constructed very precise pendulum clocks, whose lengths were carefully "tuned" to tick precisely once per second. What was found, however, was that such clocks failed to keep accurate time if they were transported too far to the north or south, i.e., to a different latitude. The obvious explanation for this would be that different weather (e.g., changes in temperature or humidity) caused the *length* of a given pendulum to change a little bit when it was transported to a new latitude. But even when such effects were corrected for, the inconsistency persisted. So the only possible conclusion was that the Earth's gravitational acceleration, g, was not actually a constant – as it would have to be for a spherically symmetric Earth – but instead varied slightly with latitude.

As Newton reports, in order to tick with the same period, a pendulum at the Equator must be a little *shorter* than one in "our regions." It is clear from the above formula for the period that this implies that g is a little *smaller* near the Equator than it is closer to the Poles. This can be understood as the result of two related factors: the rotation of the Earth, and the Equatorial bulge which is caused by the rotation.

To begin with, think of the Earth as a perfect sphere with an additional layer of matter piled up near the Equator, as in Figure 5.1. An observer at point C in the Figure is a distance h *farther away* from the dominant, spherical part of the Earth's mass, than

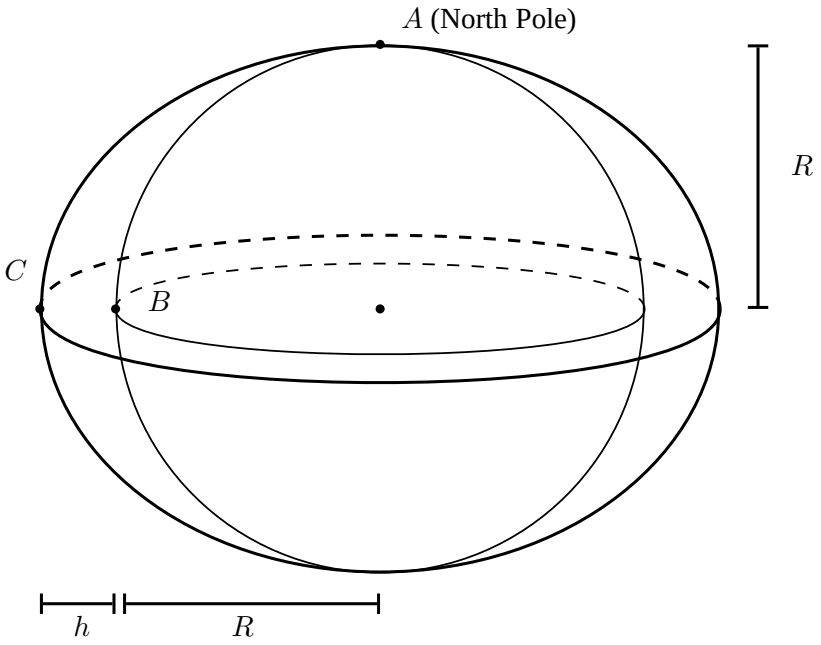

Figure 5.1: The Earth's Equatorial bulge: the Earth can be thought of as a perfect sphere of radius R, plus an additional layer of matter, thickest around the Equator, where its thickness (i.e., the difference between the "radius" at the Equator and the "radius" at the Pole) is h.

an observer at point A. This tends to decrease g near the Equator since the strength of the gravitational effect produced by the Earth's spherical core, falls off with distance. On the other hand, an observer at C has an extra layer of matter (of thickness h) right below him, and this tends to *increase* his g compared to an observer at A. It turns out (but is certainly not obvious) that the former effect is bigger in magnitude than the latter, i.e., the overall effect of the Equatorial bulge is to decrease g at the Equator relative to the Pole.

Actually, the fact that the Earth is *rotating* also contributes to the variation of g with latitude. This is because we usually define g as the acceleration that we would observe for a freely-falling object *from a reference frame that is attached to the Earth*. But since the Earth rotates, such a reference frame is not inertial – and so we cannot expect Newton's second law to apply! As we will discuss in detail shortly, we can still use Newton's laws in non-inertial reference frames if we introduce certain fudge factors called "inertial forces" – the most important and familiar of which is the so-called "centrifugal force" which tends to pull objects away from the axis of rotation. The magnitude of this (fictitious) centrifugal force turns out to be proportional to the mass m of the object it acts on, just like the gravitational force. So in practice the centrifugal force cannot be distinguished from a true gravitational force – which is why the two are usually lumped together and jointly described as an "effective gravitational force."

Of course, no such centrifugal force really exists. The point is, if one finds oneself in a non-inertial reference frame, it will *feel like* they do. And it is often convenient to indulge those feelings and use a non-inertial reference frame for the analysis of certain physical phenomena, even though, in principle, things could always be analyzed using an inertial reference frame (and only real forces!) too.

In any case, the immediate point is that, according to observers on the Earth, there is a centrifugal force which opposes and partly counteracts the gravitational force. The "effective" gravitational force on an object of mass m – which is equal to m times the effective gravitational accleration $\vec{g}_{eff}$, this being a definition of $\vec{g}_{eff}$ – is the vector sum of these two. Since the centrifugal force is strongest (and also most directly opposite the true gravitational force) for observers at the Equator, the rotation of the Earth also contributes to the systematic decrease of g_{eff} near the Equator.

Note also that these two causes of the systematic variation of g with latitude – the oblateness of the Earth and its rotation – are not unrelated. The earth is oblate *because* it rotates! It is precisely the centrifugal force which causes the (only semi-rigid) Earth to bulge out around its waist.

To briefly mention some of the interesting history: in the mid 1700's, scientists undertook a new, more direct method of measuring the shape of the Earth. They measured the actual distance, in miles, along the surface of the Earth that corresponded to moving North or South by one degree of Latitude, for different Latitudes. As expected on the basis of Newton's theory, the distance was a bit less near the Poles than near the Equator – i.e., the Earth really did bulge around the Equator. Today, the amount of oblateness or flattening can be measured very precisely from space using satellite images. And ground-based techniques of measuring g_{eff} are so precise that tiny local variations can be used to locate valuable underground deposits of natural resources!

Our goal in the rest of this section will be to understand, in a little more detail, how the rotation of the Earth, coupled with its self-gravitation, accounts for the observed amount of the Earth's oblateness. We'll then discuss the use of non-inertial (rotating) reference frames and the associated centrifugal forces, and apply these ideas to analyze again the relationship between the Earth's rotation rate, its oblateness, and the variation in the effective gravitational acceleration g_{eff} with latitude.

5.1.1 The Earth's Oblateness

If the Earth were a perfectly rigid sphere, and it were set rotating, nothing would happen. It would retain its spherical shape. But a somewhat elastic or liquid Earth will be flung outward, away from the axis of its rotation – just as pizza dough is stretched outward when it is tossed, spinning, into the air. Although this does not match the actual chronological process by which the Earth achieved its present shape, it is clarifying, in trying to derive a quantitative relationship between the Earth's rotation rate and its oblateness, to have in mind the following story: suppose the Earth used to be a perfectly rigid sphere rotating at a certain rate ω, but then "softened" and hence relaxed into its present oblate shape. For example, suppose it used to be a big, perfectly-spherical ice cube which then *melted*, allowing the water to flow into a new, energy minimizing,

equilibrium configuration.

A crucial point is that the rearrangement of matter that occurs when the Earth "melts" will be produced exclusively by internal forces. Indeed, for the moment, we may simply ignore the fact that the Earth orbits the Sun, and instead imagine it to be rotating on its axis at some fixed point in otherwise-empty space. There simply aren't any relevant external forces at all, so clearly whatever rearrangement occurs must be the result of purely internal forces. And since, as shown in the previous Chapter, such internal forces will not produce any net torque, *the angular momentum of the Earth will have to remain constant even as it melts and adjusts its shape.*

Here's why this is so crucial. Suppose we ignored it and made the following argument. The rotating earth has some rotational kinetic energy

$$KE = \sum_i \frac{1}{2} m_i v_i^2 = \frac{1}{2} I \omega^2 \tag{5.3}$$

and also some gravitational potential energy

$$PE \approx -\frac{3}{5}\frac{GM^2}{R}. \tag{5.4}$$

Now we contemplate the possibility that some of the matter from near one of the poles (point A in Figure 5.1) should move toward the Equator (which is clearly how we're going to get from a sphere to an oblate spheroid). Suppose it moves first along the (quarter-circle) path from point A to point B, and then along the straight line from B to C. During the first part of the path, it is always moving precisely horizontally, maintaining a fixed "altitude". So there is no change in its gravitational potential energy. But then, in moving from B to C, it has to move "uphill", which *increases* the overall gravitational potential energy. And since, by virtue of the rotation, the matter at the Equator has to be moving *faster* than the matter at the Pole, moving some matter from the Pole to the Equator entails an increase in the kinetic energy, too. And so, apparently, *the total energy must increase if any matter is moved from the Pole toward the Equator.* And so an initially-spherical Earth that "melts" certainly should *not* spontaneously acquire an oblate shape!

Of course, that argument is wrong, for the reason we've already hinted at. The problem is that it assumes that the Earth's overall rotation rate ω is the same before and after the contemplated movement of some matter from the Pole to the Equator. But, as we have argued, it isn't the angular velocity ω that would be constant if the initially spherical Earth melted and reconfigured itself – rather, its (spin) angular momentum $L = I\omega$ would be constant.

Moving some matter from the Pole (where $r_\perp = 0$) to the Equator (where $r_\perp = R$) of course *increases* the Earth's moment of inertia, I. So the conservation of angular momentum implies that the angular velocity ω must actually *decrease.* And since the kinetic energy is proportional to I to the first power, but ω to the *second* power, this means that the contemplated re-organization of matter will actually *decrease* the overall kinetic energy of the Earth: the kinetic energy of the one little blob of mass that moved

will indeed increase, but the rest of the Earth will slightly slow its rotation and hence decrease its overall kinetic energy – resulting in a net decrease.

It turns out that, at least for a while, this decrease in the net kinetic energy is bigger than the associated increase in the potential energy. So matter will spontaneously "flow" from the region near the Poles to the region near the Equator. At some point, though, an equilibrium is reached, beyond which further transfer of material from the Pole to the Equator would decrease the kinetic energy less than it increased the potential energy – i.e., such further transfer of material would increase, rather than decrease, the total energy. We can find a quantitative expression for the equilibrium shape of the Earth by noting that, in equilibrium, the total energy change produced by moving a tiny piece of matter from the Pole to the Equator should *vanish*.

To proceed with the calculation note that the total kinetic energy can be written in terms of the spin angular momentum $L = I\omega$ as follows:

$$KE = \frac{1}{2}I\omega^2 = \frac{L^2}{2I}. \tag{5.5}$$

Now, by taking differentials on both sides, we can write the following simple expression for the change ΔKE in kinetic energy that is produced by a small change ΔI in the moment of inertia:

$$\Delta KE = -\frac{1}{2}\frac{L^2}{I^2}\Delta I = -\frac{1}{2}\omega^2\Delta I. \tag{5.6}$$

Note that for positive ΔI, the change in kinetic energy is *negative*. That's just what we reasoned out in words above. And note that, if we had forgotten about the conservation of angular momentum and just naively taken differentials of $KE = \frac{1}{2}I\omega^2$, we'd have gotten the same final expression but with the opposite sign: $\Delta KE = \frac{1}{2}\omega^2\Delta I$. And then we'd never be able to understand why the Earth bulges at its Equator.

Let's now try to develop an actual formula for the height of the bulge. Consider the particular sort of change contemplated above – a small chunk of matter, say of mass m, being moved from the Pole to the Equator. Since, as mentioned before, its $r_\perp$ (the quantity that enters into this chunk's contribution to the moment of inertia) increases from zero to R (the radius of the Earth), we have

$$\Delta I = mR^2 \tag{5.7}$$

and hence

$$\Delta KE = -\frac{1}{2}mR^2\omega^2. \tag{5.8}$$

Of course, technically speaking the R here should be the Earth's *equatorial* radius, not its polar radius. But we'll ignore this difference here, because it turns out not to make a significant difference. (Taking it into account would only introduce a small correction to the already-small thing we are here calculating: the difference between the equatorial and polar radii!)

Now what about the change in the gravitational potential energy associated with moving this chunk of matter from the Pole to the Equator? The idea is to first move the chunk "horizontally" along the (initially spherical) surface of the Earth, from A to B in

the Figure. Since the Earth is spherical, the potential energy of a mass m hunk should be the same at B as it was at A, and so

$$\Delta PE_{A\to B} = 0. \tag{5.9}$$

We then have to move the hunk *up* a little bit, from point B to point C. Let's call this extra vertical distance h – it is just the difference between the Polar and Equatorial radii that we are trying to calculate. Then the gravitational potential energy change, when this one chunk of matter moves, is

$$\Delta PE_{B\to C} = mgh. \tag{5.10}$$

Of course, the relevant acceleration of gravity g will vary a little bit between B and C. But it only varies a little bit, and we can ignore this for the purposes of the present calculation.

We now need only plug Equations 5.8 and Equations 5.10 into the equilibrium condition

$$\Delta KE + \Delta PE = 0. \tag{5.11}$$

The result is

$$mgh - \frac{1}{2}mR^2\omega^2 = 0 \tag{5.12}$$

or, solving for h and expressing the gravitational acceleration g in terms of Newton's constant and the physical properties of the Earth ($g = GM/R^2$),

$$h = \frac{R^4\omega^2}{2GM}. \tag{5.13}$$

This is the amount by which the Equatorial radius of an approximately spherical body will exceed its Polar radius (assuming it's rigid enough to rotate as a whole, but also fluid enough to relax into this equilibrium configuration).

A nice dimensionless measure of the oblateness is the so-called "flattening parameter" f – the difference in the Equatorial and Polar radii, divided by the (say, average) radius

$$f = \frac{h}{R} = \frac{R^3\omega^2}{2GM}. \tag{5.14}$$

What does this formula predict for the oblateness of the Earth? It is easy enough to plug in numbers: $R = 6.37 \times 10^6 m$, $M = 5.97 \times 10^{24} kg$, $\omega = 2\pi$ radians/day $= 7.27 \times 10^{-5}$ rad/sec, and $G = 6.67 \times 10^{-11} m^3/kg\, s^2$. The resulting prediction is

$$h = 11\, km \tag{5.15}$$

or

$$f = 0.0017 \tag{5.16}$$

which is about a factor of two shy of the actual observed numbers. As we'll see in the rest of this chapter, it's pretty good for these kinds of problems even to get the order of

magnitude right. Often, and certainly here, there are a lot of really complicated details that we just ignore or approximate over. So getting within a factor of two definitely counts as achieving a decent quantitative understanding of the observed facts – and also leaves plenty of room for more sophisticated work in the future!

But actually here the factor-of-two discrepancy (between Equation 5.13 and the true value $h \approx 21.3$ kilometers) is a result of a pretty bad flaw in the above argument. (Did you notice it?!) We assumed that the potential energy of a hunk of matter at points A and B was the same, such that $\Delta PE_{A\to B} = 0$. That would indeed be true, as we said above, if the Earth were perfectly spherical. But of course the whole point of this discussion is that it *isn't*! And indeed, thinking about it qualitatively, it's pretty clear that in moving from point A to point B, we are moving *closer and closer* to the extra "belt" of matter surrounding the Equator – i.e., as far as *gravitation* is concerned, the path from A to B is going to be decidedly *downhill*. And so in fact $\Delta PE_{A\to B}$ will not be zero, but will be negative. It stands to reason that it, like $\Delta PE_{B\to C}$, should be roughly proportional to h – and indeed it turns out that these two contributions to ΔPE, one positive and the other negative, are of roughly the same order of magnitude:

$$\Delta PE_{A\to B} \approx -\frac{1}{2}\Delta PE_{B\to C} \approx -\frac{1}{2}mgh. \tag{5.17}$$

And so the *total* change in potential energy associated with the contemplated transfer of a small chunk of matter from the Pole to the Equator turns out to be more like

$$\Delta PE_{A\to C} = \Delta PE_{A\to B} + \Delta PE_{B\to C} \approx \frac{1}{2}mgh \tag{5.18}$$

which has the effect of *doubling* our earlier estimate for h, bringing the prediction much better in line with the actual observations. The rather subtle and difficult task of calculating $\Delta PE_{A\to B}$ will be further explored in the Projects.

Newton's theory of gravitation allows us to understand how primordial clouds of gas and dust could clump up under the mutual gravitational attraction of their parts, and form spherical blobs – the sphere being the natural result when lots of individual particles of matter try to get as close as they can to one another. The upshot of the above calculations is that Newton's theory allows us also to understand not only why the Earth and other heavenly bodies are more or less spherical, but also why and by how much they deviate from perfect sphericality due to rotation.

5.1.2 Rotating Reference Frames

We have just analyzed the oblateness of the Earth in terms of a certain trade-off in energies: if an initially-rigid and perfectly spherical rotating Earth were to melt, the gravitational potential energy would be increased by having some of the matter flow from the Poles to the Equator; but the overall kinetic energy would be decreased. Initially, the decrease would be greater than the increase, so matter *would* spontaneously flow toward the Equator – until an equilibrium is reached for which further such transfer of matter is energetically indifferent.

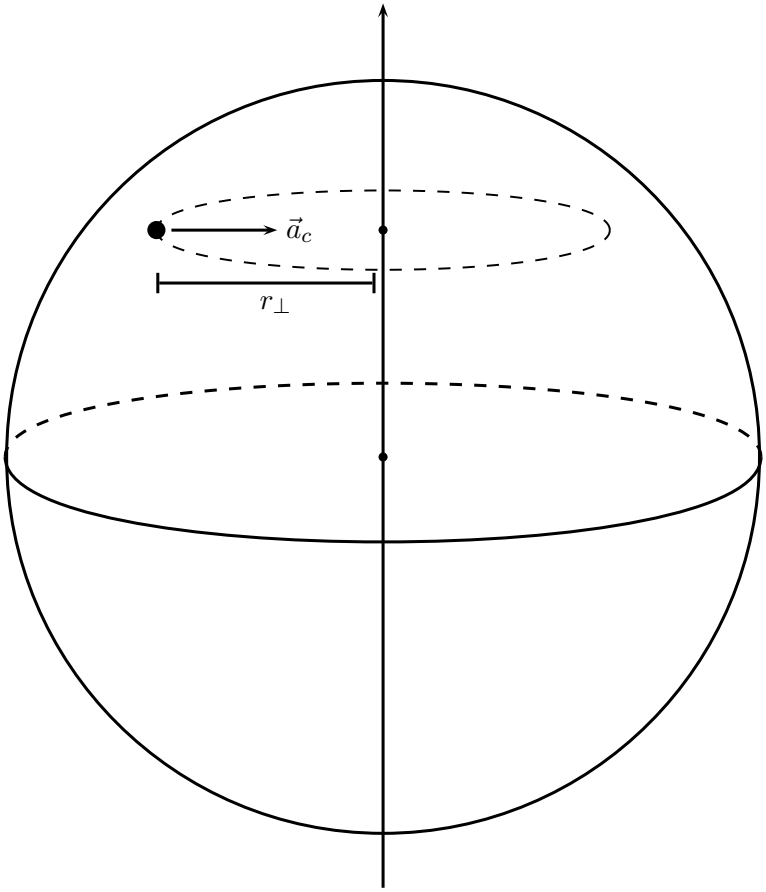

Figure 5.2: Some random particle in the Earth, undergoing uniform circular motion with radius $r_\perp$ and centripetal acceleration $\vec{a}_c$. The centripetal acceleration's magnitude will be $a_c = v^2/r_\perp = r_\perp\omega^2$, where ω is the Earth's angular velocity.

It is also possible to understand the oblateness by using a (non-inertial!) reference frame that co-rotates with the Earth. To see how to do this, let's first think about how Newton's second law, $\vec{F} = m\vec{a}$, applies to some particle of Earth if we use an inertial reference frame. Let's assume the particle is stationary with respect to the turning Earth, i.e., undergoing uniform circular motion with speed $v = r_\perp\omega$ and centripetal acceleration, directed perpindicularly in toward the Earth's rotation axis, of magnitude $a_c = v^2/r_\perp = r_\perp\omega^2$. See Figure 5.2.

The point is just that, according to Newton's second law, the net force acting on the object – the vector sum of whatever gravitational, frictional, normal, electric, magnetic, etc. forces are acting on it – will add up to its mass m times the centripetal acceleration $\vec{a}_c$. That is:

$$\vec{F}_{net} = m\vec{a}_c. \tag{5.19}$$

Since the centripetal acceleration is, well, toward the center, let us write this a little

more explicitly as

$$\vec{F}_{net} = -m\omega^2 r_\perp \hat{r} \tag{5.20}$$

where $\hat{r}$ is a unit vector pointing radially outward (in the cylindrical rather than the spherical sense), i.e., perpindicularly outward from the central rotation axis.

Now, what if we contemplate the motion of this same particle from the point of view of a coordinate system that rotates around with the rotating Earth? The main point is just this: relative to such a coordinate system, the particle isn't ever moving! And so, in particular, its acceleration is zero. Since the question of what forces act is not in any way dependent on our (subjective, arbitrary) choice of reference frame, note that this makes Newton's second law *false*. The net force is *not zero*. Yet, as reckoned in this co-rotating reference frame, the acceleration *is zero*.

None of that should be too interesting or surprising, but is maybe clarifying about why the concept of inertial reference frames is so important for Newtonian dynamics (in particular, why the first law of motion is more than a mere special case of the second). What's interesting and surprising is that we can *make* Newton's second law hold, even in the non-inertial frame, by cooking the books a little bit.

Here is the trick: whatever reference frame we choose to use, Equation 5.20 remains true. How can we reconcile this with the fact that, in the co-rotating frame, the acceleration is zero? We may simply rewrite Equation 5.20 this way:

$$\vec{F}_{net} + m\omega^2 r_\perp \hat{r} = 0 \tag{5.21}$$

and interpret the right hand side as the mass m times the acceleration $\vec{a} = 0$ in this non-inertial frame! We can then interpret the left hand side as some kind of modified or "effective" net force: it is the sum of all the *real* forces and a fictitious *centrifugal force* of magnitude $m\omega^2 r_\perp$.

Let's try to come to grips with this by considering the simplest possible example: a rock sitting on the ground somewhere at the Equator. Suppose there are just two forces acting on the rock: a weight force of magnitude W and a normal force of magnitude N. (Of course, the weight force is down, toward the center of the Earth, and the normal force is up, away from the center of the Earth.) Since the rock is rotating around with the Earth it has centripetal acceleration of magnitude

$$a_c = \omega^2 R \tag{5.22}$$

where R is the radius of the Earth. So evidently it must be that the weight force is just a little larger in magnitude than the normal force: $W > N$. In particular, we must have that

$$W - N = m\omega^2 R \tag{5.23}$$

in accordance with Newton's second law.

Now what if we consider this same situation using a non-inertial reference frame that co-rotates with the Earth? It may seem at first that there is a contradiction: The weight force is bigger than the normal force, yet the rock doesn't accelerate! Ah, but there is also the centrifugal force which, despite not really existing, must be treated as real *if*

we insist on using this non-inertial frame of reference. And then, of course, there is no problem: the weight force pulls the rock in one direction with a certain force, and the normal force and the centrifugal force *together* pull the rock equally hard in the opposite direction, resulting in zero acceleration:

$$W - N - m\omega^2 R = 0. \tag{5.24}$$

Note also that the centrifugal force is proportional to the mass m of the rock, just like the weight force $W = mg$. So it is conventional to group these two forces (one real, one fictitious) together into a single so-called "effective" gravitational force:

$$W_{eff} = W - m\omega^2 R = m(g - \omega^2 R) \tag{5.25}$$

where the quantity in parentheses is then defined as the "effective gravitational acceleration":

$$g_{eff} = g - \omega^2 R. \tag{5.26}$$

Of course, in the general case (of an object not necessarily at the Equator) we'd have to recognize the vector character of all these quantities. So the general formula for the effective gravitational acceleration is:

$$\vec{g}_{eff} = \vec{g} + \omega^2 r_\perp \hat{r}. \tag{5.27}$$

Now let's see how this relates to the oblateness of the Earth. Consider some random hunk of (say) water at the surface of the Earth at some latitude ϕ. See Figure 5.3. The important point is that the effective gravitational accleration (which determines the local meaning of "up" and "down" in the rotating coordinate system) will be tilted slightly away from its expected direction of "true down", i.e., toward the center of the Earth.

Since the surface of the Earth is largely liquid (and even the solid parts are relatively plastic on long, geological time-scales), its surface will everywhere be approximately perpindicular to the local $\vec{g}_{eff}$. And so if we can just calculate how $\vec{g}_{eff}$ varies with latitude, we can determine exactly the angle that "effective up" makes with "true up" at different latitudes, and from that understand the shape of the Earth.

Let's begin by breaking the centrifugal force up into "true horizontal" and "true vertical" components. The horizontal piece is

$$F_c^{horiz} = F_c \sin(\phi) = m\omega^2 R \cos(\phi) \sin(\phi). \tag{5.28}$$

The "true vertical" component is

$$F_c^{vert} = F_c \cos(\phi) = m\omega^2 R \cos^2(\phi). \tag{5.29}$$

Assuming that the true gravitational acceleration $\vec{g}$ is directed toward the center of the Earth, we then have that

$$g_{eff}^{vert} = g - \omega^2 R \cos^2(\phi) \tag{5.30}$$

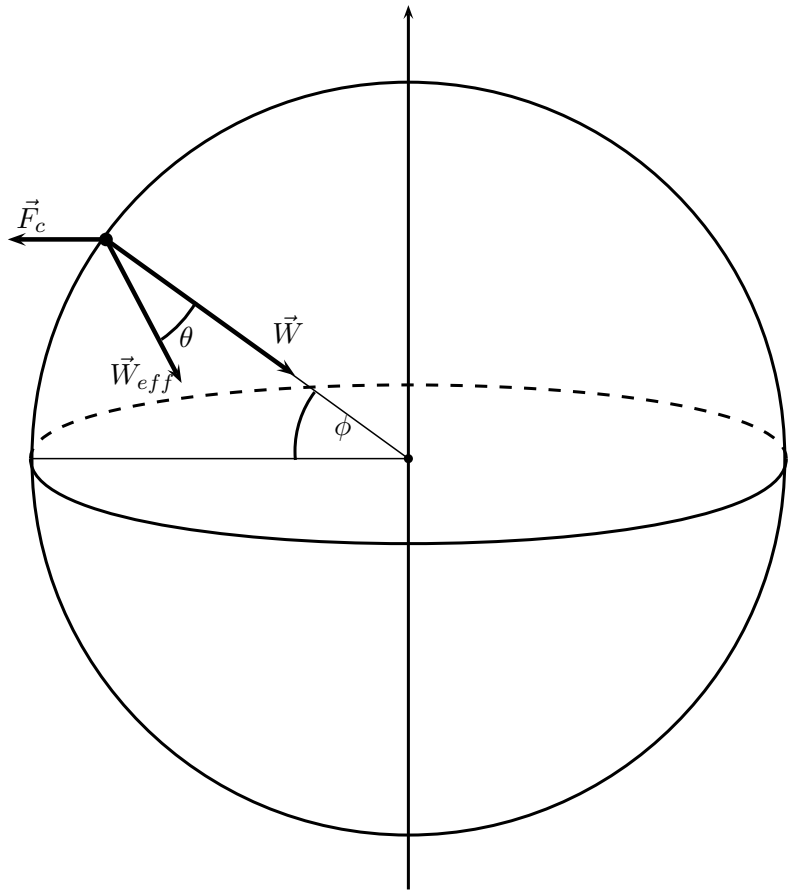

Figure 5.3: The gravitational ($\vec{W}$) and centrifugal ($\vec{F}_c$) forces acting on a hunk of, say, water at the surface of the ocean at latitude angle ϕ. The centrifugal force has magnitude $F_c = m\omega^2 r_\perp = m\omega^2 R|\cos(\phi)|$. The vector sum of these two forces (one real and one fictitious!), the effective gravitational force $\vec{W}_{eff}$, is also shown. The important point is that, because of the centrifugal force's contribution, $\vec{W}_{eff}$ does not point toward the center of the Earth, i.e., does not point in the direction we've been calling "true down." In equilibrium, the water's surface will be perpindicular to $\vec{W}_{eff}$, and so the surface will not be a perfect sphere.

which expresses the (small) reduction in g_{eff} with latitude that was primarily responsible for the effects noted first by Richer and discussed by Newton.

The horizontal component of $\vec{g}_{eff}$ will then be just the relevant component of the centrifugal force (divided by m):

$$g_{eff}^{horiz} = \omega^2 R \cos(\phi) \sin(\phi). \tag{5.31}$$

The angle θ that $\vec{g}_{eff}$ makes with "true vertical" at latitude ϕ is therefore given by $\tan(\theta) = g_{eff}^{horiz}/g_{eff}^{vert}$. But since the right hand side is very small, we might as well use the small angle approximation: $\tan(\theta) \approx \theta$. Moreover, since the second term in Equation 5.30 is small compared to the first term, we can here get away with approximating the angle as

$$\theta = \frac{\omega^2 R \cos(\phi) \sin(\phi)}{g} = \frac{\omega^2 R^3 \cos(\phi) \sin(\phi)}{GM} \tag{5.32}$$

where we have used the fact that $g = GM/R^2$.

Now imagine traveling along the surface of the Earth from the North Pole down to the Equator, and keeping track of the change in "true altitude" (distance from the center of the Earth) as one moves. A decrease in latitude by $d\phi$ corresponds to a linear distance $ds = R\, d\phi$ along a meridian of the Earth. Over this distance, the "true altitude" will increase by

$$dh = \theta\, ds = \frac{\omega^2 R^4 \cos(\phi) \sin(\phi)}{GM}\, d\phi. \tag{5.33}$$

And so the total increase in "true altitude" between the Pole and the Equator can be found by integrating:

$$h = \int dh = \frac{\omega^2 R^4}{GM} \int_0^{\pi/2} \cos(\phi) \sin(\phi) d\phi = \frac{\omega^2 R^4}{2GM} \tag{5.34}$$

which is the same result we got before in a different way. Or more precisely: this is the same *wrong* result we got before in a different way. And the reason we got the wrong result again is that we let the same wrong assumption creep in here! Before we wrongly, at first, assumed that there was no potential energy change associated with moving a hunk of matter from point A to point B of Figure 5.1. That is equivalent to assuming that the gravitational force does no work on a particle moving along the (quarter circle) path from A to B, which would be true precisely if $\vec{g}$ had no "horizontal" component, which is what we assumed here.

Of course, what we eventually realized before – that the journey from A to B is gravitationally "downhill" – implies here that $\vec{g}$ *does* have a "horizontal" component. Why? Because there is this extra belt of matter around the Equator which attracts our test particle and tilts the true gravitational acceleration g a little bit toward the Equator. And that means we *underestimated* the amount by which the surface of the Earth at latitude ϕ tilts relative to "true horizontal". Evidently this extra tilt that results from the not-quite-radial character of $\vec{g}$ contributes approximately as much to h as the (direct) centrifugal force contribution we already calculated.

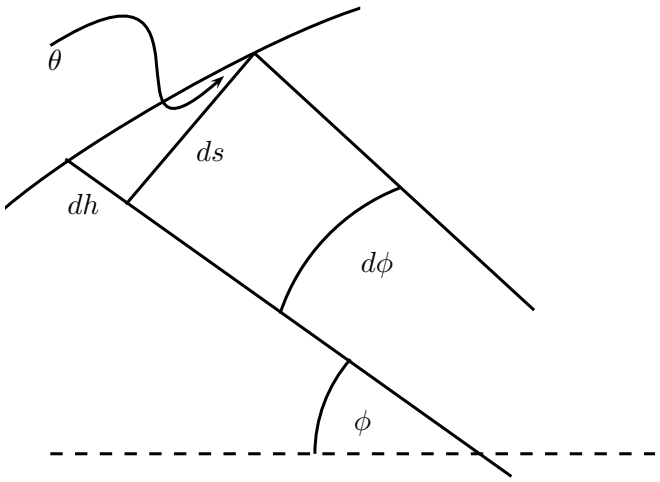

Figure 5.4: The surface of a bit of water at latitude ϕ. Over a small horizontal distance ds which spans a small latitude range $d\phi$, the height of the water increases (from the Polar to the Equatorial side) by an amount $dh = \theta\, ds$.

Let's finally return to the observations which began this whole discussion – the fact that the period of the same identical fixed-length pendulum varies somewhat with latitude, indicating that the local effective acceleration of gravity, g_{eff}, also varies somewhat with latitude. We've already written down equations for the horizontal and vertical components of $\vec{g}_{eff}$. In principle, the magnitude g_{eff} can be found from the Pythagorean Theorem, but since the horizontal component (despite its important role in determining the shape of the Earth!) is always small compared to the vertical component of g, it's a very good approximation to take

$$|\vec{g}_{eff}| = g_{eff}^{vert} = g^{vert} - \omega^2 R \cos^2(\phi). \tag{5.35}$$

Between the poles and the Equator, $\cos^2(\phi)$ varies between zero and one. So the part of the difference in g_{eff} between the Poles and the Equator that is attributable directly to the Earth's rotation is just

$$\Delta g_{eff} = R\omega^2 = 0.034\,\mathrm{m/s^2} \tag{5.36}$$

with the g_{eff} at the Equator of course being this much *smaller* than at the Poles.

Actual empirical measurement reveals that g_{eff} varies by just a little more than this:

$$\Delta g_{eff} = 0.052\,\mathrm{m/s^2}. \tag{5.37}$$

The extra discrepancy is of course due to the fact that g^{vert} itself varies a little bit with latitude, it being, evidently, $0.018\,\mathrm{m/s^2}$ smaller at the Equator than at the Poles.

The main reason for this difference was mentioned earlier: at sea level at the Equator, one is further from the center of the Earth, by a height h, than at the Poles. If one were on a *ladder* of this height (rather than a several-miles-thick slab of solid, gravitating earth!) above a perfectly-spherical Earth, the difference in g at the two locations would be

$$\Delta g = \frac{GM}{R^2} - \frac{GM}{(R+h)^2}. \tag{5.38}$$

To simplify this, it is useful to write the second term as

$$\frac{GM}{R^2}\left(1+\frac{h}{R}\right)^{-2} \tag{5.39}$$

and then use the purely mathematical fact that

$$(1+x)^n \approx 1 + nx + \frac{1}{2}n(n-1)x^2 + \cdots \tag{5.40}$$

for small x. (This can be derived, for example, by Taylor expanding the left hand side about $x = 0$.)

Since h/R is small, we may use this approximation and keep only the first-order term. The result is that

$$\frac{GM}{(R+h)^2} \approx \frac{GM}{R^2} - \frac{2GMh}{R^3} \tag{5.41}$$

or

$$\Delta g \approx \frac{2GMh}{R^3} = 2g\frac{h}{R}. \tag{5.42}$$

Plugging in numbers (in particular, the true value for h) gives

$$\Delta g = 0.066\,\mathrm{m/s^2}. \tag{5.43}$$

Of course, that's not quite right, because a mass at the Equator is not on a ladder of height h above a spherical Earth, but is rather supported by an enormous slab of gravitating material. This turns out, not surprisingly, to increase g at the Equator more than it does at the Poles, i.e., to contribute negatively to what we've been calling Δg. Evidently this extra negative contribution is just what brings our previously-calculated $\Delta g = .066\,\mathrm{m/s^2}$ in line with the empirically correct $\Delta g = .018\,\mathrm{m/s^2}$.

Notice that we have again here skirted the question of how to actually *calculate* the contributions to $\vec{g}$ that arise from the gravitational effect of the Earth's Equatorial bulge. This is not in principle all that difficult to treat exactly, but requires some rather sophisticated math. We'll take it up in the Projects and, at least, work up some order-of-magnitude estimates to convince us that everything makes sense.

There is one last thing we'll need in the Projects. We mentioned above that the actual, empirically-measured difference between the gravitational acceleration g (not g_{eff}, but the genuine gravitational field g) at the Pole and Equator is $\Delta g = 0.018\,\mathrm{m/s^2}$, where we are talking about sea level at both locations. We also just calculated that climbing to the top of a height-h ladder near the surface of the Earth has the effect of

reducing g by $0.066\,\mathrm{m/s^2}$. It is then straightforward to calculate that the Δg – *between points at the same "true altitude" at the Pole and Equator* – is going to be:

$$\Delta g = (0.018 - 0.066)\,\mathrm{m/s^2} = -0.048\,\mathrm{m/s^2}. \tag{5.44}$$

Just for clarity, what this means is that the (genuine) gravitational acceleration at sea level at the Equator is $0.048\,\mathrm{m/s^2}$ *greater* than the (genuine) gravitational acceleration at a point just *above* the Pole such that the two points are equidistant from the Earth's center. Qualitatively, it of course makes perfect sense that, for two points equidistant from the Earth's center, the gravitational acceleration would be stronger at the point that is nearer to the bulging part of the mass distribution.

5.2 Tides

There is another way in which the Earth's surface bulges away from perfect sphericality, familiar to anyone who has ever visited the ocean: the tides. Let's try to understand the physical origin of the tides, first qualitatively and then with some mathematical analysis.

First some basic qualitative facts about ocean tides. At least at most locations on the Earth, there are roughly two high and two low tides per day – "roughly" because, strictly speaking, the average time between two subsequent high tides is not precisely 12 hours, but rather about 12 hours and 25 minutes. This is just half of 24 hours and 50 minutes, which happens to be the amount of time it takes a given point on the Earth to rotate all the way around and arrive at the same place – not the same place with respect to the Sun (24 hours) or the stars (23 hours 56 minutes), but *the same place with respect to the Moon*. So that is the first and most obvious piece of evidence that the tides are controlled, somehow, by the Moon.

Actually, even this was a controversial claim for a surprisingly long period in history. Many commentators had speculated that the Moon is somehow or other controlling the tides, but nobody understood *how* and nobody was able to explain satisfactorily why there were *two* high tides per day. A naive explanation involving the Moon would have, say, the Moon pulling the Earth's water toward it a bit, causing an extra-high pile-up of water on the side of the Earth facing toward the Moon, and an extra-low deficit of water on the side of the Earth facing away from the Moon. Then, as the Earth rotated (all the way around every 24 hours 50 minutes!) underneath the moon, a given point on the Earth's surface would pass alternately through the high- and low-water regions, resulting in one high and one low-tide per day. It's a nice story, but, unfortunately, it is contradicted by the observations.

Galileo also came up with a speculative theory in which the twice-per-day rising and falling of the tide was explained (in some way that is a little obscure, and not too important because it is definitely wrong) by some sort of interaction between the two primary motions of the Earth: its daily rotation and its yearly orbit around the Sun.

The point is just to acknowledge that the tides are confusing and complicated. They were only first properly understood by Newton, using (what else?) his theory of universal gravitation.

The basic idea of Newton's gravitational explanation of the tides is this. Since (in accordance with Newton's third law) not only does the Earth exert an attractive gravitational force on the Moon, but also vice versa, the Earth itself undergoes uniform circular motion (centered on the Earth-Moon center of mass point) and is thus constantly accelerating toward the Moon. But the parts of the Earth that are closest to the Moon will experience – because the gravitational force decreases with distance – a *stronger than average* gravitational attraction toward the Moon, while parts of the Earth that are farthest from the Moon will experience – for the same reason – a *weaker than average* gravitational attraction toward the Moon.

The point is that – *relative to this average attraction toward the Moon* (as embodied, say, by the gravitational acceleration of the point at the Earth's geometrical center) – the stuff on the side of the Earth nearest the Moon will be attracted (just a little bit) toward the Moon, while stuff on the side of the Earth farthest from the Moon will be (just a little bit) *repelled*, *away from the Moon*. And so stuff – like the water in the oceans – that is more or less free to flow around and re-position itself will tend to pile up at these two opposite positions on the Earth's surface. And that, obviously, is where it'll be high tide. And so a typical point on the Earth's surface will pass through *both* high-tide regions per day.

This is illustrated in Figure 5.5. The normal arrows represent the strength of the gravitational force exerted on that part of the Earth by the Moon. As discussed above, the points closer to the Moon experience a greater than average attraction to the Moon and the point farthest from the Moon experiences a smaller than average attraction to the Moon. In addition, points like those at the top and bottom of the Earth in the Figure experience an attraction that is approximately the same magnitude as average, but tilted at a slight angle. The double arrows represent the *difference* between the actual attraction at a point, and the average attraction. The upshot is clear: relative to the average motion of the Earth as a whole, the surfaces on the top and bottom (of the Figure) are pushed in/down, while the surfaces on the sides are pushed out/up. The result is something like the elliptical shape (technically a prolate spheroid) indicated in the Figure – though the extent of the tidal bulges is significantly exaggerated there.

One should think of this ellipsoid as an equilibrium shape that the surface of the oceans would make if this were the only relevant effect. But of course, the Earth itself is spinning around once per day (or once every 24 hours 50 minutes relative to the Moon). So, as a kind of first approximation, one should think of the oceans as always making roughly this equilibrium ellipsoid, with the tidal bulges essentially fixed in space relative to the Moon – but with the solid parts of the Earth rotating around, underneath and through the tidal bulges. In particular, since there are two tidal bulges, a given point on the Earth's surface will revolve around through this relatively fixed pattern of high- and low- water, passing alternately through high- and low-tide regions.

Of course, how one describes this is going to be reference-frame dependent. From the point of view of the Earth itself, there are two big tidal bulges which race around and around and around, trying to keep up with the Moon (and the point in the sky opposite the Moon, respectively) as it rises and sets each day.

Let's see if we can now calculate the actual *amplitude* of the tides, i.e., the difference

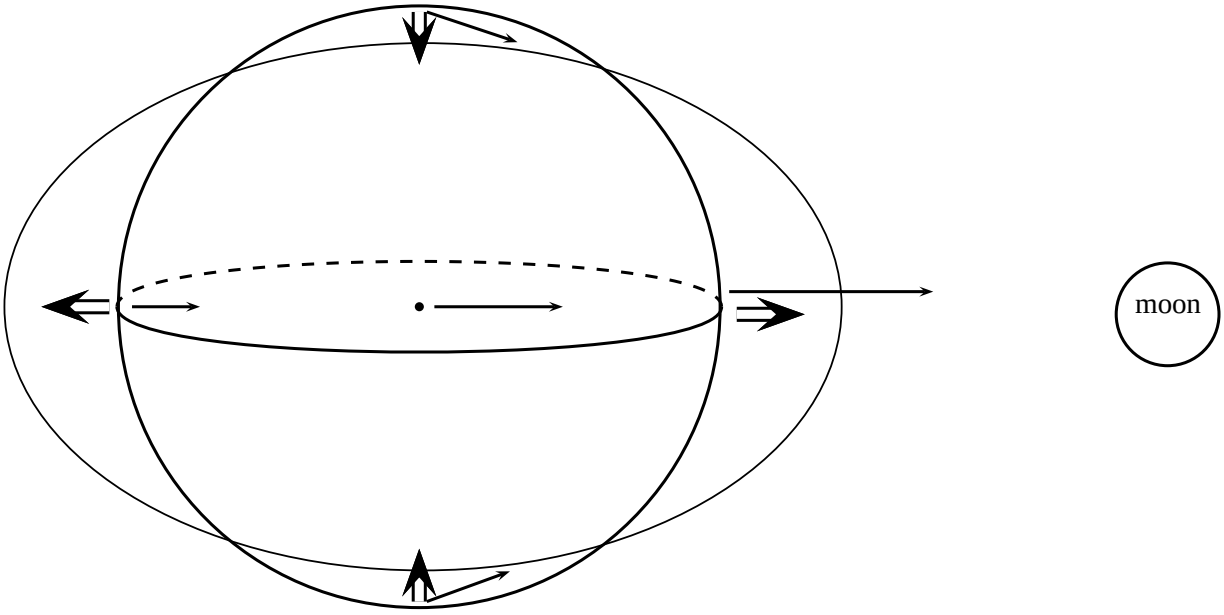

Figure 5.5: The tidal forces produced by the Moon on the Earth. The single arrows represent the gravitational force (per unit mass) exerted on a given part of the Earth. The double arrows represent the *difference* between the actual force acting at a point and the average force. This difference is called the "tidal force." The decrease of the Moon's gravitational influence with distance explains why the tidal force is toward the Moon on the right and away from the Moon on the left. The fact that the force is always directed straight toward the Moon explains the slight "tilt" of the forces at the top and bottom of the figure, which in turn results in a tidal force that points back in toward the center of the Earth. The net result of these differential forces is that water flows toward the two high-tide points on the right and left of the diagram, as indicated by the (much exaggerated) elliptical surface shown. Note finally that the three dimensional shape generated by these tidal forces will have rotational symmetry about the Earth-Moon axis. So it is low-tide not only at the top and bottom of the Earth (as shown in the figure) but also on the parts of the Earth that come out of the page, and the parts that go into the page.

h in height between the high- and low-tide points shown in the previous Figure. As with the calculations of the amplitude of the Equatorial bulge in the previous section, there are several ways to do this. The simplest is probably to use the equilibrium argument which says that the total energy change associated with moving (say) some mass m blob of water from the high-tide surface to the low-tide surface should be zero. You can work it out this way in the Projects. Here we'll adopt the slightly less straightforward, but in some ways more revealing, method of calculating first the effective gravitational accleration at different points on the Earth's surface, and then using this to calculate the "slant" of the equilibrium surface relative to "true horizontal" – just as we did in the previous section as a final way to analyze the oblateness of the Earth.

Figure 5.6 shows a cross-section of the Earth. We assume, to begin with, that the Earth is spherical. This seems like a dubious assumption given the previous section, but all we are going to calculate here is the *extra* deviation to the Earth's shape produced by the tidal interaction with the Moon. The idea is that the Earth's Equatorial oblateness is caused by its (daily) rotation – so we can ignore both the rotation and the oblateness in order to isolate the effect of the tides.

The Figure shows the two relevant contributions to the effective gravitational accleration, g_{eff}, at a point near the Earth's surface that is an angle θ down from the Earth-Moon axis. One of the contributions, g_{moon}, is of course the result of the Moon's gravitational influence. The other, g_c, is equal to the "average" (centripetal) acceleration of the Earth (toward the Moon, or, equivalently, about the Earth-Moon center of mass point).

The clearest way to think of this is to assume that we are using a non-inertial reference frame that is attached to the Earth. Since the Earth as a whole accelerates to the right in the Figure (if one uses an inertial frame), there will exist (in this Earth-attached frame) a "fictitious" gravitation-like force that pushes everything back to the left with a force proportional to its mass: $F_c = mg_c$. Note that the magnitude of g_c is just the accleration of the Earth (as a whole, on average) toward the Moon: $g_c = GM_{moon}/r^2$.

It is important here to appreciate that this is a non-inertial *but also non-rotating* reference frame. Using a rotating reference frame (e.g., centered at the Earth-Moon center of mass point and rotating in tandem with those bodies' mutual orbits) to analyze this problem is certainly possible. You can work it out in the Projects. But it can be a little confusing because, really by definition, the effect we are here trying to isolate and understand – the effect of the Moon's tidal forces on the shape of the Earth – has absolutely nothing to do with rotation. As we saw in the last section, rotation produces relatively *large* deviations from perfect sphericality, on the order of tens of kilmeters. The tides, of course, are nowhere near that high! (Luckily!) So we need to be careful to isolate the purely tidal effects we are interested in, by systematically avoiding any assumption (which may creep into the analysis if we're not careful to avoid it) that the Earth is rotating. So for now we forget about the rotational/spin motion of the Earth, and treat it as having a fixed orientation with respect to the fixed stars. Then, a reference frame that is rigidly attached to the Earth will be accelerating (because the Earth accelerates toward the Moon) *but not rotating*. And so the fictitious forces needed to use this non-inertial reference frame will be as described in the previous paragraph.

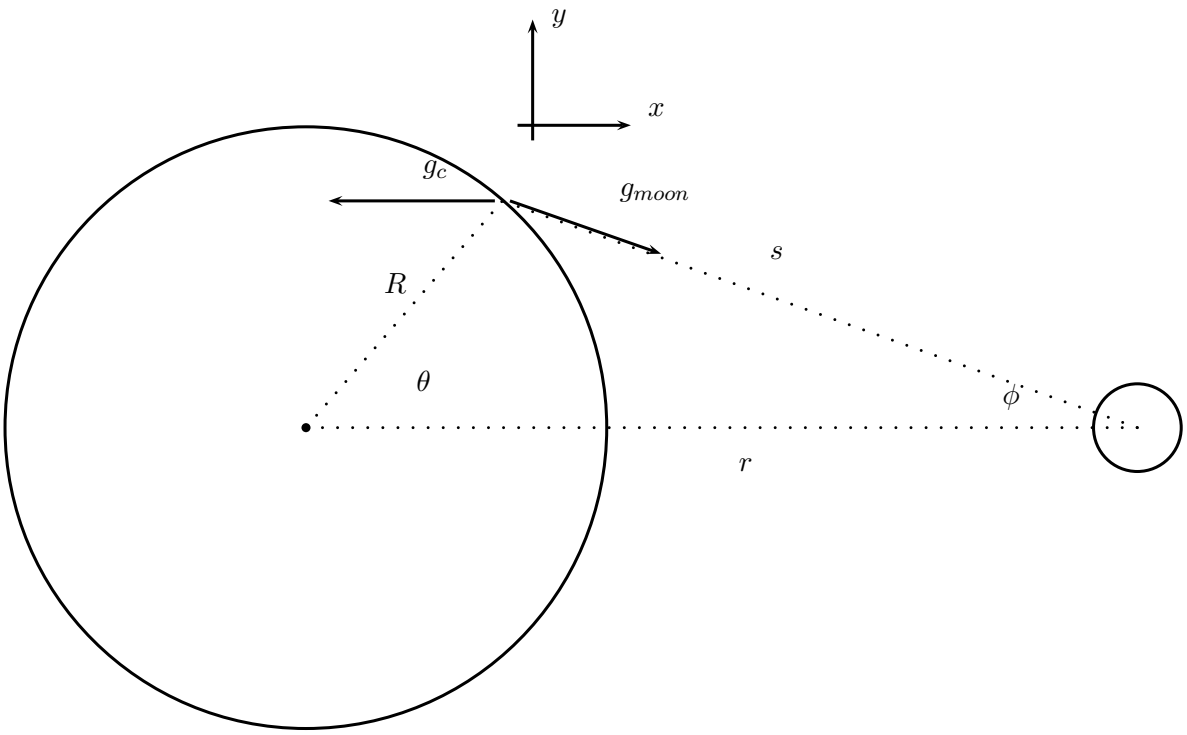

Figure 5.6: Diagram for calculation of tidal contributions to g_{eff} near the surface of the Earth.

The Figure also indicates an $x-y$ coordinate system, which will help us in writing down expressions for the x and y components of these two contributions to the effective gravitational acceleration for a point a distance R from the origin:

$$g^x_{eff} = \frac{GM_{moon}}{s^2}\cos(\phi) - g_c \tag{5.45}$$

$$= \frac{GM_{moon}}{r^2 + R^2 - 2Rr\cos(\theta)} - \frac{GM_{moon}}{r^2} \tag{5.46}$$

and

$$g^y_{eff} = -\frac{GM_{moon}}{s^2}\sin(\phi) \tag{5.47}$$

where we have used the law of cosines: $s^2 = r^2 + R^2 - 2Rr\cos(\theta)$.

Now we are going to make some simplifying approximations. The idea is essentially to expand things in the small parameter $R/r \approx 1/60$ and keep only the leading non-vanishing contributions in each term. As suggested (but understated) in the Figure, the angle ϕ is already small: $\sin(\phi) \approx R\sin(\theta)/r$. So we can then make the crudest possible approximation for the denominator in the expression for the y-component: $s^2 \approx r^2$. This gives

$$g^y_{eff} \approx -\frac{GM_{moon}R\sin(\theta)}{r^3}. \tag{5.48}$$

On the other hand, $\cos(\phi) \approx 1$. So we need to be more careful to pick off the similarly-sized contribution in the expression for the x-component. In particular, we'll ignore the R^2 – but not the $2Rr\cos(\theta)$ term – in the denominator. This gives

$$\begin{aligned} g^x_{eff} &\approx \frac{GM_{moon}}{r^2 - 2Rr\cos(\theta)} - \frac{GM_{moon}}{r^2} && (5.49)\\ &= \frac{GM_{moon}}{r^2}\left[\left(1 - \frac{2R}{r}\cos(\theta)\right)^{-1} - 1\right] && (5.50)\\ &\approx \frac{2GM_{moon}R\cos(\theta)}{r^3} && (5.51)\end{aligned}$$

which is indeed the same order of magnitude as the y component.

It will eventually be useful to have an expression for the Moon's tidal force not just on the surface of the Earth, but at an arbitrary location. The previous expressions can be easily converted by replacing $R\cos(\theta)$ with x, and $R\sin(\theta)$ with y:

$$g_e ff^x(x,y) \approx \frac{GM_{moon}}{r^3}2x \tag{5.52}$$

and

$$g_e ff^y(x,y) \approx \frac{GM_{moon}}{r^3}y. \tag{5.53}$$

So those are the x and y components of the tidal force (per unit mass... probably we should say the tidal acceleration). As with the calculation of the size of the Equatorial bulge, however, it's really the *horizontal* component that directly affects the "slant" of the equilibrium ocean surface at angle θ relative to "true horizontal". It is easy enough to work out that

$$\begin{aligned} g^{horiz}_{eff} &= g^x_{eff}\sin(\theta) - g^y_{eff}\cos(\theta) && (5.54)\\ &= \frac{3GM_{moon}R}{r^3}\cos(\theta)\sin(\theta) && (5.55)\end{aligned}$$

and hence that the angle made by the water surface at angle θ relative to "true horizontal" will be

$$\alpha = \frac{g^{horiz}_{eff}}{g} = \frac{3M_{moon}R^3}{M_{earth}r^3}\cos(\theta)\sin(\theta). \tag{5.56}$$

And, still just following the earlier calculation, this means that, over a small angle $d\theta$ at angle θ, the height of the water (relative to the initial spherical shape, i.e., constant height) will decrease by

$$dh = \alpha R d\theta = \frac{3M_{moon}R^4}{M_{earth}r^3}\cos(\theta)\sin(\theta)d\theta. \tag{5.57}$$

We need only finally integrate this from $\theta = 0$ to $\theta = \pi/2$ to find the total difference h between the heights of the low- and high-tide points:

$$h = \int dh = \frac{3}{2}\frac{M_{moon}}{M_{earth}}\frac{R^4}{r^3}. \tag{5.58}$$

If we plug in the actual values for the Moon's and Earth's masses and the relevant distances, we find

$$h = 54\,\mathrm{cm} \tag{5.59}$$

or about two feet – certainly in the ballpark of the actual variations observed.

Actually, though, it is not at all uncommon for high- and low-tides to differ by two or three times this estimate, or more. The reason for this can be qualitatively understood by thinking again about what's happening using a reference frame that co-rotates with the Earth. Then the story one tells is that there are these two giant tidal waves which are constantly propagating around the Earth to the west. If the whole surface of the Earth were covered with water, the tidal bulges would more or less just flow around, and the above calculation would be pretty accurate. But, of course, they can't – there's *land* in the way! So, for example, the tidal bulge in the Atlantic Ocean runs up pretty hard against the whole east coast of the Americas, and has to somehow go *around* that land mass to get, just a few hours later, into the Pacific. So there is a tendency for the water to pile up more along the east coast than it would if there were no land there, much as a small ripple in the bathtub can make the water level go up and down with significantly greater amplitude when the ripple sloshes against the edge of the tub. And then, after all that water races around the continents into the Pacific, the two giant streams (from the north and south) meet in the middle and again create a bulge with an even-greater-than-equilibrium height. Of course, the details of this are *extremely* complicated and vary significantly between different geographical locations, even locations that are relatively close together. So if you want to know exactly when it will be high- or low-tide at a given location on a given day, consult a tide table! These are based on empirical fits to historical data, and so are much more reliable than any possible calculation a physicist could make. On the other hand, if you want to really understand what produces the tides and how to think about them, well, now you do!

There are several additional points that should be mentioned. First, although we've talked as if the tides are produced exclusively by the gravitational influence of the Moon, everything we've said applies equally much to the Sun. What low-high-tide difference h would be produced by the Sun (if we could isolate its tidal effect)? We can immediately co-opt our previous result, just changing everywhere the word "Moon" to "Sun" and re-interpreting the r to mean now the distance between the Earth and the Sun:

$$h_{Sun} = \frac{3}{2}\frac{M_{sun}}{M_{earth}}\frac{R^4}{r^3} = 25\,\mathrm{cm} \tag{5.60}$$

which turns out, by sheer coincidence, to be of the same order of magnitude as the h produced by the Moon.

Of course, the Sun and Moon are both always present and always influencing the Earth's waters. The interesting point is that, depending on their relative alignment, the Sun and Moon can produce particularly *strong* tides, or particularly weak tides. Consider for example New Moon – when the Moon and the Sun are both in (roughly) the same direction relative to Earth. Then the tidal bulges produced by the two bodies are right on top of each other (one bulge on each side of the planet), and their amplitudes

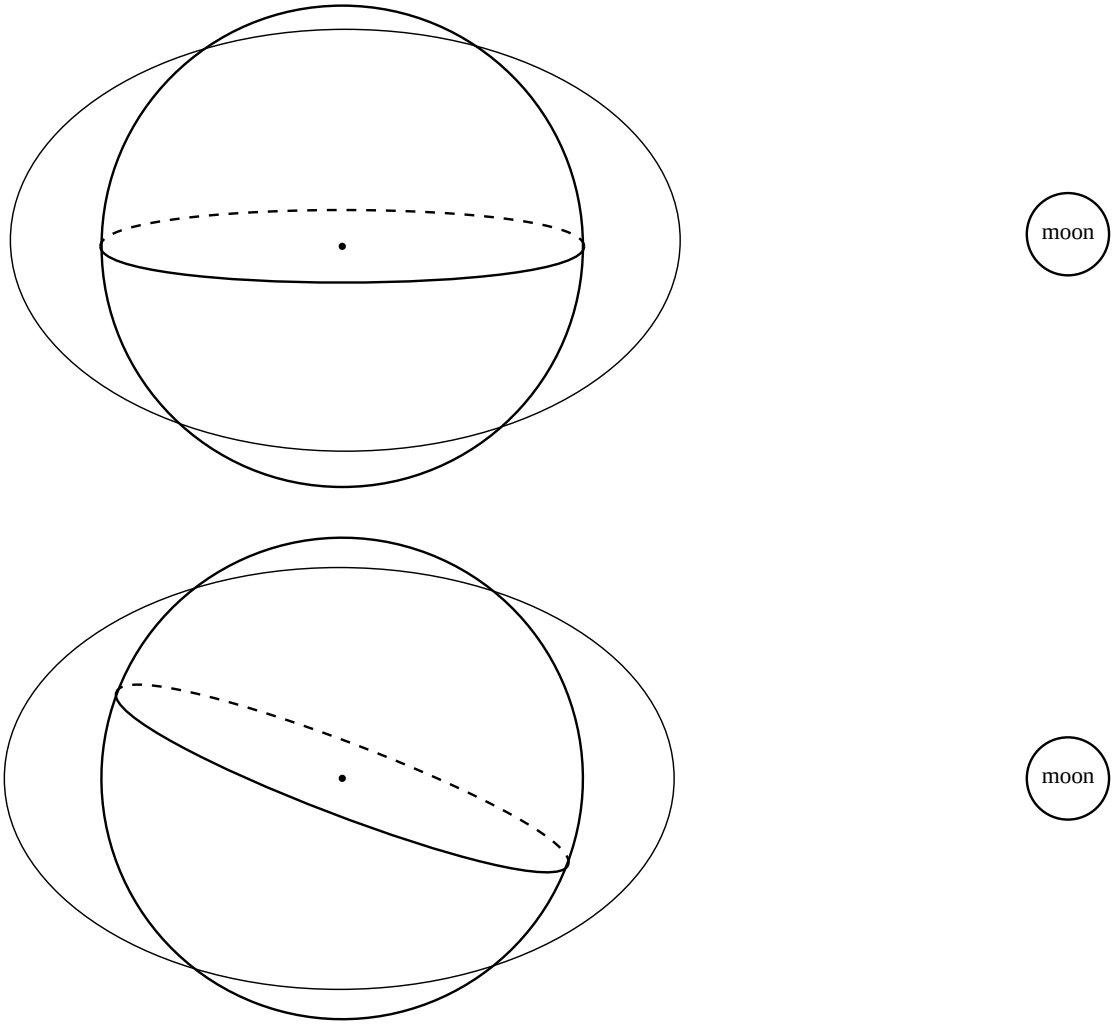

Figure 5.7: Two possible orientations of the Earth's tidal bulge relative to its geography. In the top panel, the Moon lies in the plane of the Earth's equator. An observer on the Equator will experience two equally high and two equally low tides on this day. Observers at other location will also experience two equally high and two equally low tides, but they won't be as high and low (respectively) as at the Equator. In the bottom panel, the Earth's spin axis is tilted down toward the Moon, so the two tidal bulges are somewhat north and south of the Equator, respectively. On this day, an observer at the Equator will observe two equally high high tides and two equally low low tides (but they won't be as high and low, respectively, as they were on the day pictured in the top panel). An observer at moderate latitude will observe two high tides, but one of them will be considerably higher than the other. The two low tides will be about the same. Observers at very extreme latitudes (near the north or south Poles) may experience just one high tide and just one low tide on this day! That should give you a sense of the monthly variations that are possible in the tides, and how those variations vary by latitude. The *seasonal* variations mentioned in the text arise the same way, but with the Sun replacing the Moon in the Figure.

add. Hence, one expects particularly strong tides (meaning particularly high high-tides and particularly low low-tides) around New Moon – and, as you can see with a little thought, also around Full Moon. By contrast, when the Moon is about half full, so the Sun, Moon, and Earth make a right triangle, the Moon tries to create a high tide at the same place (on Earth) that the Sun tries to create a low tide, and vice versa. That is, their effects tend to *cancel* resulting in particularly weak tides (not too high high-tides and not too low low-tides). The weak tides that occur when the moon is half full (either waxing or waning) are called "neap tides". The strong tides that occur at Full and New Moon are called "spring tides" – not because they happen in the spring, but, evidently, because the waters spring up particularly high then.

There are, however, some seasonal (and monthly) variations in the tides as well, which have to do with the fact that the Earth's spin axis is not perpindicular to, but tilted relative to, the plane of the ecliptic. See Figure 5.7. Additional seasonal and monthly variations are produced by the fact that neither the Moon's orbit around the Earth nor the Earth's around the Sun are circular. Instead, as discovered first by Kepler, the orbits are slightly eccentric ellipses. The distance r between the Earth and the Moon, for example, varies up and then down away from its average by about 5 % each month. And since the strength of the tides depends on this distance to the *third power*, the relatively small changes in the distance to the Moon can produce relatively large changes (15-20% variations away from average) in the strength of the tides. (The same is also true for the Earth's orbit around the Sun, but since this is eccentric only by one or two percent, the corresponding seasonal variations in the Sun's tidal influence are smaller.)

So next time you visit the ocean, pay attention to the tides. In particular, notice how the rising and falling of the tide correlates with the location and phase of the moon.

5.3 The Non-Spherical Earth and Associated Torques

As we mentioned in passing in Chapter 4, a *uniform* gravitational field – like that near the surface of the Earth – will exert a net force *but no net torque* on an object, no matter how complicated its shape. On the other hand, a *non-uniform* gravitational field – like the spherically-symmetric radially-inward field produced by a moon or planet or star – can exert not only a net force but also a net torque on an arbitrarily shaped object. Consider, for example, the situation depicted in Figure 5.8.

It can be shown (we won't bother here) that a spherically symmetric object, however, cannot have such a gravitational torque exerted on it. (Actually, it's sort of the converse of the earlier proof that a spherically symmetric body acts, gravitationally, just like a point mass – the point here is that such a body also *re*-acts, gravitationally, just like a point mass.) In order for a gravitational torque to be produced on an object, the object must lie in a non-uniform gravitational field and must itself be non-spherically-symmetric. Of course, just like the imaginary giant barbell in the Figure, the Earth sits in the not-quite-uniform gravitational field of the Moon. (The tidal forces we analyzed in the previous section can be thought of as nothing but the departures of the Moon's gravitational field from uniformity in the vicinity of the Earth.) Moreover, both of the last two sections have concerned themselves with respects in which the Earth fails to

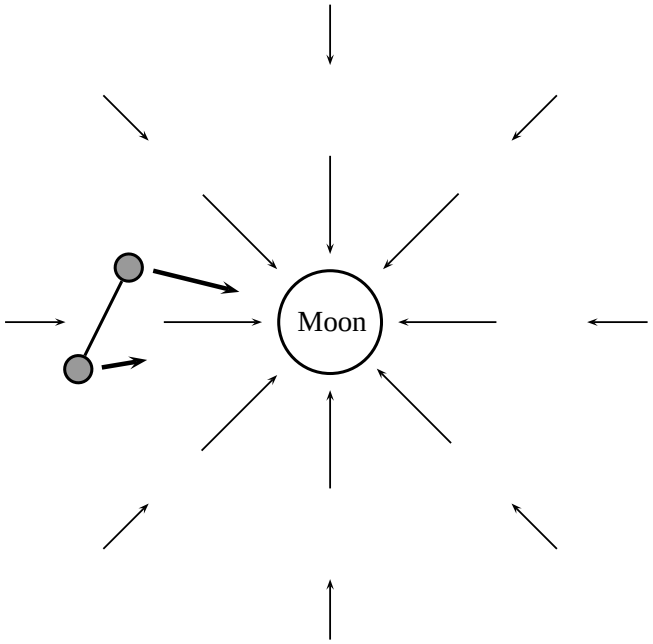

Figure 5.8: A giant barbell in space near (say) the Moon. The net gravitational force on the barbell (the sum of the two forces exerted on the two masses) pulls it toward the Moon. But because the gravitational field produced by the Moon is not uniform, the two masses composing the barbell have different forces exerted on them – which produces, in addition to the overall tendency to accelerate toward the moon, a *torque* which tends to rotate the barbell (clockwise in the Figure).

be perfectly symmetric. So one should expect that the Earth's various bulges result in *torques*, which – in some ways we'll now explore – affect the rotational state of the Earth.

5.3.1 The Tidal Torque

In the previous section, we discussed how the non-uniformity of the Moon's gravitational field near the Earth (the tidal forces) produces two tidal bulges, one on the side facing the Moon and the other on the opposite side. We calculated the equilibrium height of the bulges and discussed the simple equilibrium model in which a given point on the Earth's surface just rotates around, moving alternately through the high- and low-tide regions, and thus experiences two high- and two low-tides per day.

There is also, however, an important dynamical coupling between the rotation of the Earth and the tidal bulges. As viewed from (say) an inertial reference frame above the Earth-Moon system, the tidal bulges have to move (pretty fast!) relative to the rotating Earth, just to stay in their equilibrium positions. And because there is some friction between the solid rotating Earth under the oceans, and the waters themselves, the tidal bulges don't *quite* keep up. Put another way, the rotation of the Earth is

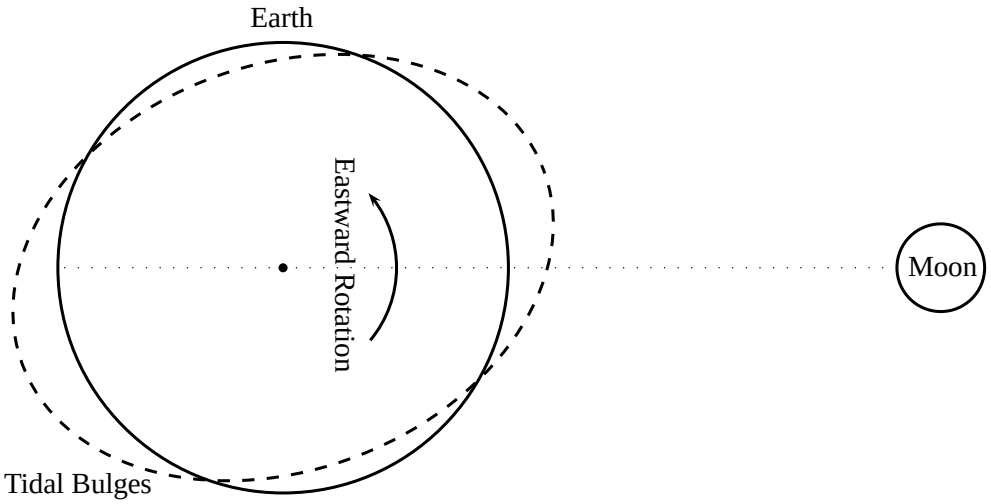

Figure 5.9: The perspective here is looking down from far above the North Pole. The Earth's daily Eastward Rotation drags the tidal bulges a little bit to the east relative to their equilibrium positions (which would be along the line connecting the Earth's and Moon's centers). (Note that the picture exaggerates this. In fact, the tidal bulges are only a few degrees to the east of the Earth-Moon line.) The tidally bulging Earth therefore acts just like the giant barbell from the previous Figure: the bulge that is closer to the Moon is attracted toward the Moon more strongly than, and in a slightly different direction from, the bulge on the other side, producing a torque that tends to turn the Earth to the west. Or here is a slightly better way to think about the same thing. If one imagines the slightly-rotated tidally bugling Earth superimposed on the tidal forces as shown in Figure 5.5, the result is clearly a net torque: both bulges are pulled, by those tidal forces, in a way that tends to produce a clockwise (i.e., westerly) rotation of the bulging Earth. Of course, since the Earth is already rotating to the East, the result of this torque is a (very gradual) decrease of the easterly angular velocity. That is, because of the torque exerted by the Moon on the tidally bulging Earth, the length of the day is very gradually increasing!

constantly pulling the bulges away from their equilibrium positions (just under the Moon and opposite it). The result is that the tidal bulges are not precisely in their equilibrium positions, but are instead pulled a few degrees to the east by the rotating Earth. See Figure 5.9.

As explained in the Figure's caption, the Moon's tidal forces produce a net torque on the Earth because of this slight departure of the tidal bulges from their equilibrium orientation. This torque acts to slow down the rate at which the Earth rotates, i.e., to increase the length of the day. Indeed, scientists have measured that the length of the day is increasing by about 1.6 milli-seconds per century.

It is very interesting to consider this process from the point of view of the Earth-

Moon system. For that system, all of the complicated frictional and tidal/gravitational forces that are involved in the slowing down of the Earth's rotation are *internal forces*, which can therefore produce no *net* torque. The total angular momentum of the Earth-Moon system must therefore be a constant – which means that, since the (eastward) spin angular momentum of the Earth is *decreasing*, the (eastward) orbital angular momentum of the Moon must be *increasing*. It can be shown that the orbital angular momentum for a roughly circular orbit is proportional to the square root of the radius of the orbit – so increasing orbital angular momentum implies increasing radius. Thus, it follows from the conservation of angular momentum that the size of the Moon's orbit should be slowly increasing.

Amazingly, this too has been directly measured in recent decades. When the Apollo astronauts landed on the Moon in the early 1970s, they left some mirrors (technically "corner reflectors") from which Earth-based scientists can reflect light. Measuring the amount of time it takes for a pulse of light shot toward the Moon to be reflected and subsequently detected (and knowing the speed of light) allows for extremely precise measurements of the distance to the Moon. And indeed, this distance has been *measured* to be increasing at a rate of approximately 3.5 cm per year.

By the way, the reason for this gradual (but measurable) change in the Moon's orbit can be understood without mentioning angular momentum conservation. As we have seen, the torque on the tidally bulging Earth can be understood as a result of the bulge closest to the Moon being attracted to the Moon more strongly than the bulge on the far side. But also vice versa: the tidal bulge closest to the Moon *attracts* the Moon more strongly than does the bulge on the far side. So the net force exerted by the Earth on the Moon is not quite toward the center of the Earth, but rather ever-so-slightly tilted toward the Moon's direction of motion. This component does positive work on the Moon, increasing its energy and allowing it to "climb" into ever-higher orbits.

The upshot of all this is that the Earth-Moon system is not in equilibrium. The Earth's "daily" rotation rate is decreasing, and the Moon's orbital radius is increasing. When and how will these gradual changes cease? A little thought reveals the answer: when the Earth's (spin) angular velocity matches the Moon's (orbital) angular velocity. In other words: when the Earth daily rotation slows so much that it always presents the same face to the Moon. Then the Moon (which *already* always presents the same face to the Earth) and the Earth will be "tidally locked" in a face-to-face dance.

So – as a result of the subtle interplay of tidal forces, friction between the oceans and the sea floor, and the laws of rotational dynamics – your distant ancentors may someday be able to see the Moon in the sky all the time (or never, depending on where they live).

5.3.2 Torque on the Equatorial Bulge

We previously treated the Earth's Equatorial Bulge as an intrinsically interesting feature that can be understood and explained using Newton's theory of gravitation and some concepts of rotational dynamics and kinematics. But the Earth's Equatorial bulge is interesting for another reason, too: just as with the tidal bulges, the tidal forces exerted by the Moon (and Sun) interact with the Equatorial bulge to produce a torque. And

this torque, like the one on the tidal bulges, results in some interesting gradual changes in the Earth's state of rotation.

Actually, the effect of the torque (exerted jointly by the Sun and Moon) on the Equatorial bulge is something we've already discussed, way back in Chapter 1: the precession of the equinoxes – that subtle long-period turning of the Earth's rotation axis that was (despite its roughly 26,000 year period) noticed already by the Ancient Greeks.

In principle, the mechanism here is simple. The Earth spins, like a top. And – *because the Earth is not quite perfectly spherical* – the Sun and Moon exert a gravitational torque on the Earth. The torque is produced by the Sun's and Moon's tidal forces, which would tend to align the Earth's Equatorial plane with the plane of the Ecliptic (the plane of the Earth's orbit around the Sun, roughly also the Moon's orbit around the Earth). But just as the gravitational torque on the spinning top causes it to precess rather than tip over, so with the Earth: the gravitational torques exerted on its bulging Equator by the Sun and Moon cause its spin (or spin angular momentum) axis to sweep out a cone, always staying roughly the same 23.5° away from the fixed point among the stars called the Pole of the Ecliptic.

Really, the story here is precisely like the story from the previous chapter for the top. So there are only two things to fill in. First: why and how, exactly, does the Moon or Sun exert a net torque on the Earth? And second: how big is that torque, and does it – in accordance with Equation 4.102 – account for the observed rate of one revolution per 26,000 years?

We'll discuss the first point here and then leave the second part (a hard but very cool calculation) for the Projects.

Actually, there's not that much to say since the effect is the same as that for the tidal bulges. See Figure 5.10 for a sketch of the Equatorially bulging Earth sitting in the tidal force field produced by (say) the Moon.

The only subtlety is that, since the orientation of the Earth's spin axis is (approximately!) the same throughout the month or year, the tidal forces and bulge will not always be exactly as depicted in the Figure. It is probably easiest here to think first about the tidal forces exerted by the Sun. Then, Figure 5.10 will depict the situation correctly at the Summer Solstice (with the Sun to the left along the negative x-axis) and also the Winter Solstice (with the Sun to the right along the positve x-axis). These two times turn out to correspond to the torque being a *maximum*. And it is important that at these two times the torque is in the same *sense*, the same direction.

Around the equinoxes, however, the situation is rather different. The relevant tidal forces are as shown in Figure 5.11. As should be clear qualitatively from the Figure, the torque now *vanishes*. Hence, over the course of the year, the torque exerted by the Sun on the Equatorially bulging Earth varies back and forth (twice) between some maximum value and zero.

Since this back and forth variation in the torque turns out to be extremely fast compared to the main effect produced by the torque (the 26,000 year period precession of the equinoxes), it is reasonable to calculate an average torque, and then treat the phenomenon as if that average torque were exerted steadily in time. We may guess that the average torque produced by the Sun's tidal forces will be about half of the maximum

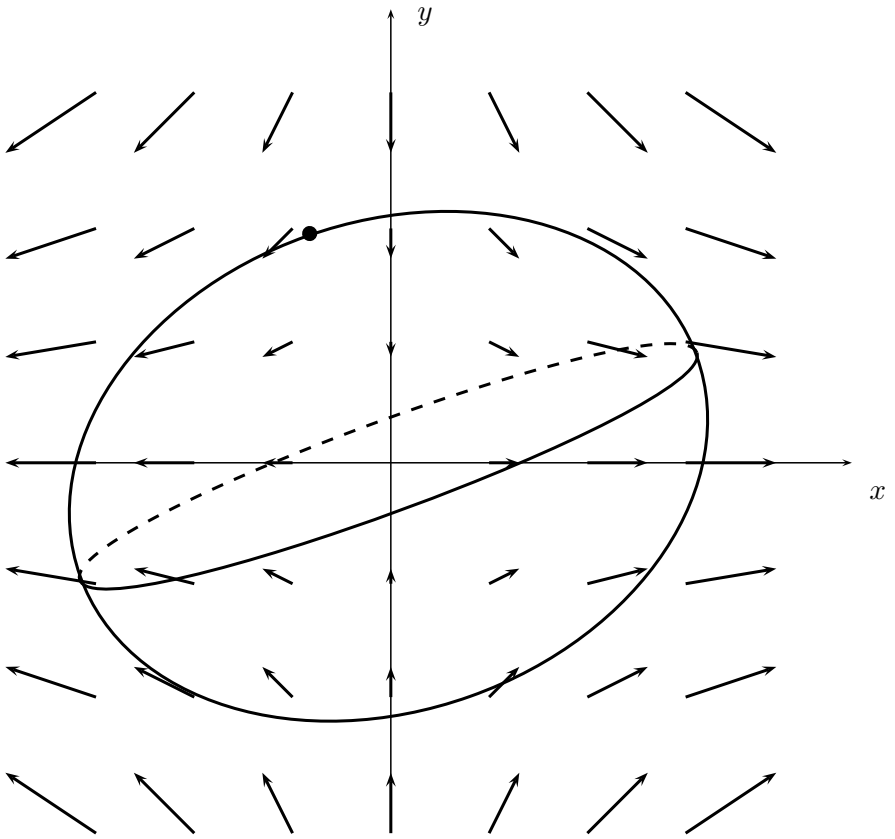

Figure 5.10: The Equatorially bulging Earth lies in the tidal force field produced by (say) the Moon. The picture will be accurate if the Moon is to the right along the x-axis, or to the left along the negative x-axis. Clearly the tidal force on the bulge on the right tends to turn it clockwise, as does the tidal force on the bulge on the left. There is therefore a net torque exerted on the Earth by these tidal forces.

torque (exerted at the Solstices):

$$\tau_{Sun}^{avg} \approx \frac{1}{2}\tau_{Sun}^{max}. \tag{5.61}$$

Now finally note that everything we've just said about the tidal forces and torques produced by the Sun, applies in just the same way to the Moon. The only difference really is that we don't have terms for the points in the Moon's orbit around the Earth which correspond to the Solstices and Equinoxes – i.e., the times when the Earth's spin axis is tilted maximally toward or away from the Moon (the "Lunar Solstices") or tilted down perpindicularly from the Moon-Earth line (the "Lunar Equinoxes"). So it would have been a little harder to describe and understand. But if you followed the discussion for the Sun, everything is truly the same for the Moon – except that the relevant torque (produced by the Moon's tidal forces on the Earth's Equatorial bulge) oscillates back

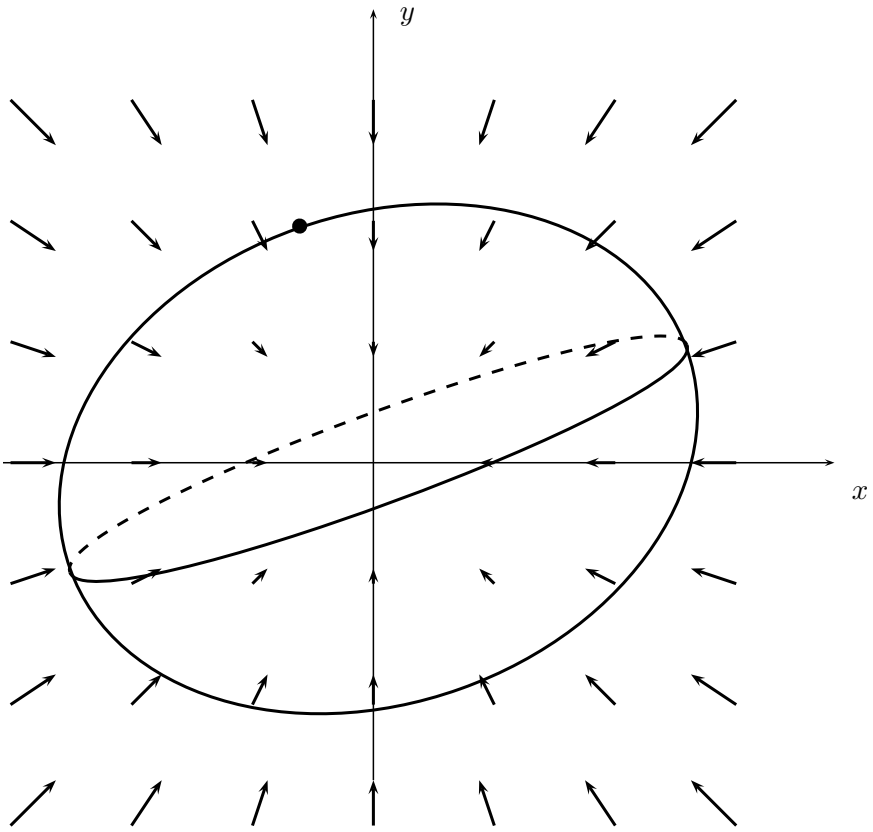

Figure 5.11: The Earth sitting in the tidal force field produced by the Sun around the Spring or Autumn Equinox. The perspective is (say) from the Sun. The tidal forces are radially symmetric in a plane perpindicular to the Earth-Sun line, and so produce no torque on the Earth – or at least, it is clear from the picture that there is no torque on the particular planar slice of Earth shown. But all of the other slices will have the same symmetry pattern, and so, indeed, the total torque will vanish.

and forth between its maximum value and zero twice per month, rather than twice per year. But we still have, in analogy with the previous equation, that

$$\tau_{Moon}^{avg} \approx \frac{1}{2}\tau_{Moon}^{max}. \tag{5.62}$$

The time-averaged *total* torque exerted on the Earth (in virtue of its Equatorial bulge) is therefore

$$\tau_{total}^{avg} = \tau_{Sun}^{avg} + \tau_{Moon}^{avg} = \frac{1}{2}\left(\tau_{Sun}^{max} + \tau_{Moon}^{max}\right). \tag{5.63}$$

So if we can calculate (or approximate) the "max" torque produced on the Equatorial bulge by the Sun at the Solstices – and by the Moon at the "Lunar Solstices" – we'll be able to plug the resulting total torque into Equation 4.102 and see if, indeed, this

process accounts quantitatively for the observed precession rate. But we'll leave that fun project for the Projects.

5.4 Measuring Masses

Back in Chapter 4, we discussed the Cavendish experiment in which Newton's gravitational constant G was first measured. Because the gravitational acceleration $g = 9.8\,m/s^2$ of objects near the Earth's surface is readily measureable, and because this acceleration is given, according to Newton's theory, by

$$g = \frac{GM_{earth}}{R_{earth}^2} \tag{5.64}$$

– and because the radius of the Earth is also known – the measurement of Newton's constant G allows the mass of the Earth to be computed. This is why, as we discussed, this laboratory measurement was and is often referred to as a means of "weighing the Earth."

We also discussed, in that earlier chapter, how a similar approach could be used to determine the mass of the Sun. Since, for example, the (centripetal) acceleration of the Earth toward the Sun is known

$$a_{earth} = \omega^2 R = \frac{2\pi^2 R}{T^2} = \frac{4\pi^2 \times 1\,AU}{(1\,year)^2} = .0059\,m/s^2 \tag{5.65}$$

and given, according to Newton's theory, by

$$a_{earth} = \frac{GM_{sun}}{R^2} \tag{5.66}$$

where $R = 1\,AU$ is the Earth's orbital radius, the mass of the Sun can be worked out:

$$M_{sun} = \frac{R^2 a_{earth}}{G} = \frac{4\pi^2 R^3}{T^2 G} = 2 \times 10^{30}\,kg \tag{5.67}$$

or about 300,000 times the mass of the Earth.

Here is the principle involved: whenever a relatively light body moves under the gravitational influence of a relatively heavy body, and the relevant kinematical properties of the light body (its acceleration toward and distance from the heavy body) can be measured directly, the mass of the heavy body can be inferred. This is a relatively simple point, but an extremely important and fruitful one for modern astronomy and astrophysics. For example, it is by this same method that the masses of other planets can be determined – but only if those planets have *moons*!

Moons orbiting Mars, Jupiter, and Saturn were discovered when (or shortly after) Galileo first pointed a telescope to the heavens. Thus Newton, in the Principia, was already able to estimate the masses of these planets. A more recent and particularly interesting instance is the planet (recently demoted to "dwarf planet" status) Pluto, which was discovered in 1930. Pluto's mass, however, remained unknown until 1978, when a moon ("Charon") orbiting Pluto was discovered.

The apparent (angular) diameter of Charon's orbit, combined with knowledge of the distance to the planet-moon system, allowed the absolute size (R) of its orbit to be determined. Observations over time also allow the period of Charon's orbit to be determined. By plugging this information into Equation 5.67, Pluto's mass can be calculated. You can work through that calculation in the Projects.

We should note, though, that this is a bit over-simplified. The mass of Pluto turns out to be pretty low – so low that its moon, Charon, really is not "relatively light" compared to it. Indeed, it turns out that the center of mass of the Pluto-Charon system is not within Pluto's body at all, but is rather in the empty space between them. (By comparison, the center of mass of the whole solar system is somewhere inside the Sun – slightly toward Jupiter from its center, typically; similarly, the center of mass of the Earth-Moon system is within the Earth, some 1700 km below the Earth's surface, on the side facing the Moon, obviously.) The Pluto-Charon system is therefore sometimes classified as a "dwarf double planet" system rather than a (dwarf) planet plus a moon.

Another interesting (and only recently-discovered) fact about the Pluto-Charon system is that both bodies are "tidally locked" to one another. This is also a result of the fact that the two bodies are of comparable mass.

Anyway, the fact that the two bodies are of comparable mass – and hence must really be described as each orbiting around their mutual center of mass – requires a somewhat more careful analysis to convert the observed kinematical information into a determination of their masses. Let's work this out in general for two objects of mass m_1 and m_2, orbiting around their mutual center of mass with (circular) orbits of radii R_1 and R_2, as shown in Figure 5.12.

For the Pluto-Charon system, we would observe the system "edge-on" rather than the "face-on" perspective shown in the Figure. The latter, however, is a little simpler for analyzing the physics. In any case, no matter what perspective we have on the system, as long as we can observe it over time (and as long as the absolute distance to the system is known, so the apparent angular separations can be converted into absolute distances) we can determine the radii of the two orbits, R_1 and R_2. These are distances measured from the (empty) center of mass point, so one might wonder how this point can be located. The answer is simple: it is the center of the two observable orbits.

It follows from the definition of the center of mass that the product m_1R_1 should equal m_2R_2. This can be converted into an expression for the mass ratio:

$$\frac{m_1}{m_2} = \frac{R_2}{R_1}. \tag{5.68}$$

An additional algebraic constraint on the two masses can then be inferred from orbital dynamics. According to Newton's theory, the mass m_2 exerts on m_1 a force of magnitude $F = Gm_1m_2/(R_1 + R_2)^2$ which produces acceleration $a_1 = Gm_2/(R_1 + R_2)^2$. But this is just the observed centripetal acceleration of m_1, so we may write

$$\frac{Gm_2}{(R_1 + R_2)^2} = \frac{v_1^2}{R_1} = \frac{4\pi^2 R_1}{T^2} \tag{5.69}$$

where T is the period of the orbit. The same reasoning leads to a corresponding condition

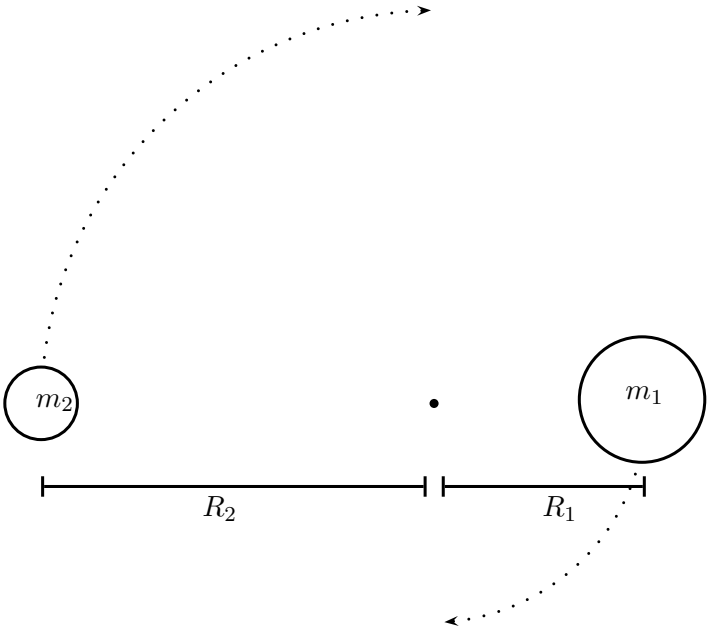

Figure 5.12: A binary system: two bodies, of mass m_1 and m_2 respectively, orbit about their mutual center of mass (the black dot in the Figure) with circular orbits of radii R_1 and R_2.

for the other mass:

$$\frac{Gm_1}{(R_1+R_2)^2} = \frac{4\pi^2 R_2}{T^2}. \tag{5.70}$$

The previous two equations can then be added, and the result simplified, to give an expression for the sum of the masses:

$$m_1 + m_2 = \frac{4\pi^2(R_1+R_2)^3}{GT^2}. \tag{5.71}$$

With the ratio and the sum both determined by observable quantities, it is then clear that the two masses – m_1 and m_2 – can each be uniquely determined.

This would all probably be analytical overkill if it were useful only as a way to figure out the mass of Pluto and its moon Charon. But in fact this same technique can be used to determine the masses of *stars*, many of which, perhaps surprisingly, are found to be trapped in a gravitational orbit with another star – a so-called binary star system. For example, two of the best-known stars – Polaris (the north star) and Sirius (the bright star near Orion) – happen actually to be members of binary star systems.

The simplest kind of case is a binary star system in which the two stars are individually observable, such that their individual orbits can (as in the case of Pluto and Charon) be tracked over time. If the absolute distance to the binary star system can also be determined, it is then straightforward to measure R_1, R_2, and T from observation, and hence to infer (just as sketched above) the masses of the two stars.

Actually, it is more commonly successful to measure the masses of stars in binary systems in a slightly more subtle way. This takes advantage of the so-called Doppler effect, which is probably familiar in the case of sound: the paradigm example is the ambulance siren that sounds higher in pitch as the ambulance approaches you, but then appears to drop in pitch as the ambulance passes you and starts to recede. The physics involved here is that the *observed* frequency f of a wave (such as the sound wave emitted by the ambulance siren) depends not only on the *intrinsic* frequency of the source, f_0, but also on the radial velocity, v_r, of the source – i.e., the rate at which its distance from the observer is decreasing. For a sound wave, the relevant formula for the Doppler shift is

$$\Delta f = f - f_0 = f_0 \frac{v_r}{c} \tag{5.72}$$

where c is the speed of sound. For a *light* wave, the formula is the same (at least as long as v_r is small compared to c), but with c now the speed of light: $c = 3 \times 10^8\, m/s$.

The upshot is that, by carefully monitoring the *frequency* of light emitted by stars, one can learn something about the speed with which they move, toward or away from the observer. For stars in a binary system as discussed above – but viewed "edge on" – the radial velocity will oscillate back and forth (say, around zero) with a maximum absolute value

$$v^{max} = \omega R = \frac{2\pi R}{T} \tag{5.73}$$

where T is the period of the orbit and R is its radius.

The point is then that, for a so-called "spectroscopic binary" in which this Doppler wobble can be detected for both stars, we can rewrite the above mass-determination equations purely in terms of the radial velocity amplitudes, $v_1^{max} = \omega R_1$ and $v_2^{max} = \omega R_2$, instead of the radii R_1 and R_2 which are, as a matter of observational fact, much harder to measure than the velocities.

The only problem is that, if we just determine the v^{max} values from spectroscopic data without actually resolving the precise motion of the two stars, there is no way to know whether the binary system is being viewed precisely "edge-on." To be general, we should assume that the system is inclined at some angle i, in which case the maximum observed radial velocities are given by

$$v^{max} = \omega R \sin(i) = \frac{2\pi R}{T} \sin(i) \tag{5.74}$$

where $i = 0$ corresponds to the "face-on" perspective shown in the Figure and $i = \pi/2$ corresponds to the "edge-on" perspective. With this more general relationship, the relevant formulas for the masses of the two stars in the binary become

$$\frac{m_1}{m_2} = \frac{v_2^{max}}{v_1^{max}} \tag{5.75}$$

and

$$m_1 + m_2 = \frac{T}{2\pi G} \frac{(v_1^{max} + v_2^{max})^3}{\sin^3(i)}. \tag{5.76}$$

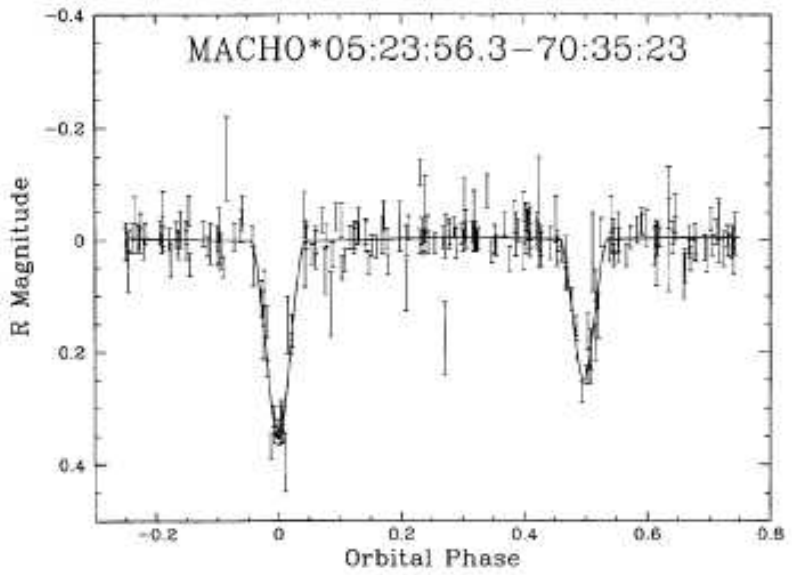

Figure 5.13: The "light curve" for an eclipsing binary star system. (A light curve is a plot of the intensity of light coming from a certain source, plotted against time. Since the light curve for this system is *periodic*, the intensity has been plotted against the phase of the period. This makes the structure of the periodic curve much clearer to the eye.)

In some cases, the two stars in a binary system can be observed to *eclipse* one another as they orbit. The light curve for one such eclipsing binary is shown in Figure 5.13. The eclipsing implies that the system is being observed edge-on, such that $\sin(i) \approx 1$. In such cases, careful observations of T, v_1^{max}, and v_2^{max} allow very accurate determinations of the masses of the two stars. In other cases, there is no way of determining i from the observations, and the most one can do is put a lower limit on the masses.

At this point a fair question would be: who cares about the masses of stars? Part of the answer would surely be that, as we now know, gravity plays a crucial role in the formation and evolution of stars. So if you want to understand stars – which means, if you want to understand the universe and our place in it – you better know something about the source of gravitation, which is mass. As just one concretization of this (perhaps otherwise unsatisfying) answer, note that empirical studies of relatively nearby binary star systems reveal an amazing correlation between stars' mass and luminosity. See Figure 5.14. "Luminosity" refers to a star's intrinsic brightness – the total amount of energy radiated, as light, per unit time. This can be determined by measuring the *intensity* of the star's light – that is, the energy per unit time passing through a unit area (e.g., a detector) here on Earth – and then multiplying by the area of a sphere

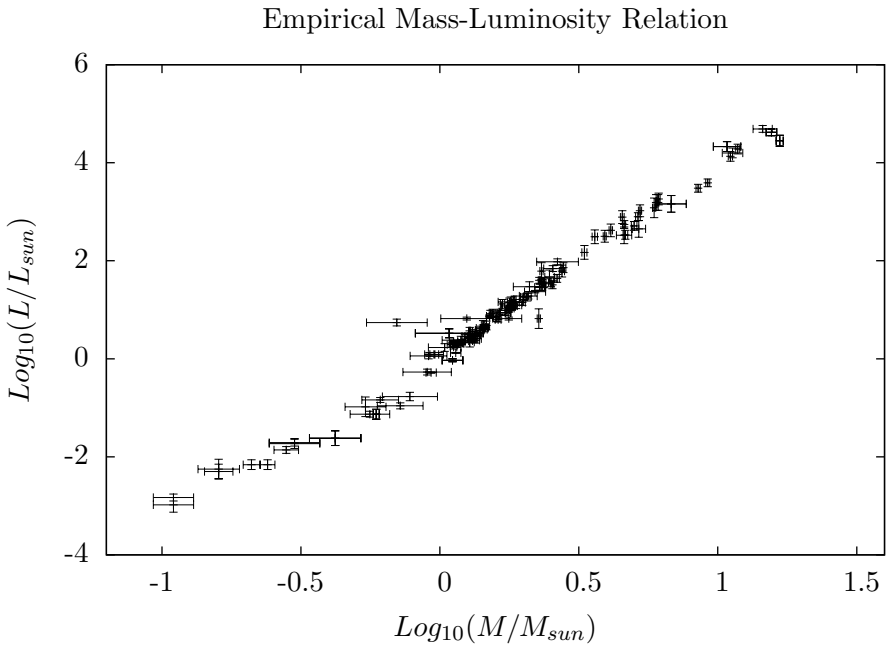

Figure 5.14: Graph of the mass-luminosity relation. Data are from Popper 1980 and are for "main sequence" stars only. See C/O problem 7.9 on page 200.

whose radius is the distance from that star to the Earth, the idea being that the star radiates its light uniformly in all directions, so the total amount of light (the luminosity) should equal the amount of light per unit area (as sampled over some very small area here on Earth) times the total area through which starlight of that intensity passes. As an equation,

$$L = I \times 4\pi D^2 \tag{5.77}$$

where L is the star's intrinsic luminosity, I is the measured intensity of its light here on Earth, and D is the distance to the star.

Anyway, the mass-luminosity correlation indicates that the directly observable features of stars (such as their brightness, but including also such features as color and radius) are intimately related to their intrinsic internal structures. Better understanding the details of this connection between the hidden internal structure of stars and their outward appearances is a major part of astrophysics. As you can imagine, this relies not only on the theory of gravitation, but also thermodynamics, hydrodynamics, optics, and even nuclear physics – because it is the nuclear process of *fusion*, occuring in the cores of stars, which fuels them.

One particularly interesting implication of the empirical mass-luminosity relation is that massive stars live much shorter lives. All other things being equal, one might have thought that a more massive star would burn longer than a less massive star, since it has more internal fuel. (The fusion reaction that powers stars is the nuclear "burning"

of Hydrogen into Helium. More massive stars, however, will continue to burn after the Hydrogen is used up – by fusing Helium into Carbon, for example.) But the empirical relationship between mass and luminosity shows that all other things are not equal. A star that is, say, twice the mass of the Sun will be about 10 times brighter, and will therefore burn through its fuel in roughly a fifth the time. (The Sun will run out of Hydrogen fuel – and puff up into a red giant, before eventually settling back down to become a "white dwarf" – in about 5 Billion years.)

A star that is a hundred times the mass of the Sun will use up its fuel millions of times faster, and hence have a lifetime that is thousands of times shorter than the Sun. Such massive stars not only end their lives sooner than the Sun – they also end it much more dramatically. We will discuss in the next section.

5.5 Cataclysms

In our discussion of the Earth's tides above, we noted that, because the overall tidal effect depends on the third power of the distance between the central and orbiting bodies, relatively small variations in this distance (as result from the Moon's not-quite-circular orbit) can produce relatively large fluctuations in the strength of tidal effects. This is of course a general fact about tidal forces, which applies not just to the Moon's tidal influence on the Earth, but also the Earth's on the Moon, Pluto's on Charon, and so forth. Let us think about the following thought experiment in terms of a generic planet-moon system.

Imagine that some planet's moon was somehow brought into progressively smaller and smaller circular orbits around the planet. The planet's tidal effect on its moon would grow and grow, in accordance with the inverse-cube law just mentioned, and so – at least to the extent that the moon is deformable over the relevant timescales – its departure from sphericality would increase. But at some point (i.e., at some particular distance from the planet) a dramatic transition will have to occur: the tidal forces acting on the planet would become comparable in size to the (largely gravitational) forces by which the moon holds itself together as an integrated body. At this point, the moon would be unable to hold itself together, and would be literally torn apart by the tidal forces.

To estimate when this should happen, we may calculate the distance at which, say, a rock on the side of the moon facing the planet is pulled just as hard toward the planet (by the tidal force) as it is pulled toward the moon (by the moon's own gravitational force). The strength of the gravitational field produced by the moon's own gravity is

$$g_{Moon} = \frac{GM_{moon}}{R_{moon}^2} \tag{5.78}$$

while the tidal field exerted by the planet is (for a rock on the near side)

$$g_{tidal} = \frac{2GM_{planet}R_{moon}}{r^3} \tag{5.79}$$

where R is the radius of the moon and r is the critical planet-moon separation.

If the moon were just at this critical distance where the two previous expressions are equal, a rock released from just above the Moon's surface would be at a kind of unstable equilibrium point, and might either fall back to the moon's surface or be pulled by the tidal forces toward the planet. Solving for the critical value of r by equating the previous two expressions gives

$$r_{crit} = R_{moon} \left(\frac{2M_{planet}}{M_{moon}} \right)^{1/3}. \tag{5.80}$$

This critical distance is usually referred to as the Roche limit, after the physicist Edouard Roche who first discovered it in 1850.

It is clarifying to re-write this expression in terms of the mass *densities* of the moon and planet, given by

$$\rho = \frac{M}{\frac{4}{3}\pi R^3} \tag{5.81}$$

for each of the two bodies. The result is that the critical radius – the Roche limit – is proportional to the radius of the planet:

$$r_{crit} = R_{planet} \left(\frac{2\rho_{planet}}{\rho_{moon}} \right)^{1/3}. \tag{5.82}$$

For the Earth-Moon system, a quick calculation reveals that the Roche limit occurs at about five and a half times the radius of the Earth – well inside the actual distance to the Moon, about sixty Earth radii. And since the tidal interaction discussed earlier is causing the Moon to slowly *increase* its distance from the Earth, we needn't worry that our (distant) ancestors will someday see the Moon ripped apart.

Interestingly, though, such a fate *does* lie in the future for several other moons in the solar system. Phobos (one of Mars' moons) and Triton (one of Neptune's moons) both orbit their planets in a way that is, in a sense, opposite to the Moon's orbit around Earth. Phobos orbits Mars *faster* than Mars rotates: the Phobosian month on Mars is shorter than a Martian day! In the case of Triton, the moon actually orbits the planet in a retrograde fashion, i.e., opposite the direction of the planet's "daily" rotation. (As seen from way to the north of the solar system, Neptune – like all the other planets – orbits the Sun counter-clockwise, and Neptune – also like all the other planets – *rotates* counter-clockwise. But Triton's orbit around Neptune is clockwise.) In both cases, the effect is to reverse the sense of the slow tidal evolution discussed above for the Earth-Moon system: the two moons in question are (unlike the Earth's moon) getting ever closer to their planets. And so at some point (millions of years in the future) they will reach the relevant Roche limits for their respective planets and be shredded.

Perhaps it has occured to you that the famous rings of Saturn could be the dusty remnants of a tidally shredded moon. Indeed, Saturn does have a number of moons, all of which are farther away than its famous rings. And, indeed, it turns out that (making some reasonable estimates for the density of the moons) the current moons are outside the Roche limit, while the rings are inside. So it is entirely possible that the rings formed, at some point in the past, when tidal (or other) interactions pulled a previously-coherent moon inside the critical radius. Another possibility is that the rings

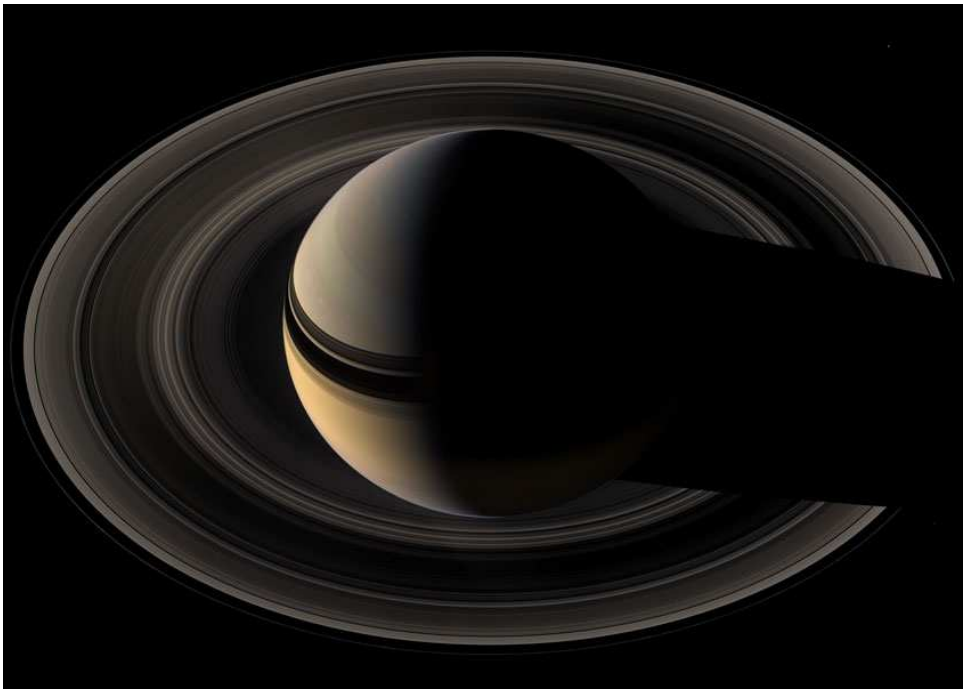

Figure 5.15: A picture of Saturn and its beautiful rings. Several of Saturn's moons are also present, though it is hard to tell that they are all more distant from the planet than the rings. This photo was taken in 2007 by the Cassini spacecraft. What aspect of the picture proves immediately that it wasn't taken from Earth?

are a well-preserved remnant of a primordial swirl of dust that clumped up billions of years ago to form Saturn and its moons. Under this hypothesis, the rings are not the debris of an ex-moon, but rather the ingredients that would have formed a moon had they not found themselves at a distance from the central planet for which the tidal forces prevented the usual moon-formation process of gravitational clumping.

The "tidal cataclysms" we've been discussing here can in principle occur not only for moons which get too close to their planets, but also for planets which get too close to their stars, or even stars in binary systems which get too close to their partners. We will explore this a bit in the Projects. But at least for the case of the moons in the solar system, although the effect is interesting to understand and contemplate, it is a bit moot. For another type of cataclysm will eventually befall many of these systems.

At some point, several billion years in the future, the Sun will start to exhaust the Hydrogen that fuels the internal Hydrogen-to-Helium fusion reactions which power it. As this happens, Helium – the inert by-product of this fusion reaction – will tend to pile up in the core. The inert core will cool somewhat and contract, allowing the still-Hydrogen-rich material above to fall in somewhat, a process which heats the Sun's outer layers. This, in turn, will dramatically increase the rate at which Hydrogen-to-Helium fusion is occuring there, heating the outer layers even further. All this newly generated

heat will cause the outer layers of the star to puff up – the net result being that the Sun will become a so-called "red giant" star.

The radius of the newly-formed red giant will exceed the (current) Sun's radius by a factor of about 100 – which means, among other things, that the Sun will then occupy what used to be the orbits of several of the inner planets, probably including Earth. (And even if Earth is spared in this process, the dramatic increase in the Sun's total luminosity will increase the Earth's average temperature far beyond the boiling point of water, making Earth in any case not exactly hospitable.)

Eventually – i.e., after wiping out much of the solar system – the Sun will really run out of Hydrogen fuel, and settle back down to a smaller size. At this point its "life" is essentially over. All that will remain is an inert core of mostly Helium, which will simply sit there and slowly cool off over the subsequent billions of years.

A more dramatic death awaits stars which are significantly more massive than the Sun. For the Sun, the Helium by-product of the primary Hydrogen-to-Helium fusion process is inert – it doesn't participate in any further nuclear processes. But for stars which are ten to a hundred times heavier than the Sun, the temperature and pressure in the star's core greatly exceed those in the core of the Sun. And, it turns out, under such conditions further energy-producing nuclear reactions are possible. For example, three Heliums can fuse together to form Carbon in the so-called triple-alpha process. (The name is because Helium nuclei, which fuse in this process, are also known as "alpha particles.") And likewise, Carbon can fuse with Helium to form Oxygen – which can in turn fuse with another Helium to form Neon – and so on, to heavier and heavier elements.

It is now understood that virtually all of the elements heavier than Helium were created, in stars, in precisely this process. So, for example, the trace amounts of Carbon, Nitrogen, and Oxygen in our Sun (which incidentally act as catalysts in a special Hydrogen-to-Helium fusion reaction called the CNO cycle) signify that the Sun is not a first-generation star, but was rather formed from the remnants of an earlier cycle of stellar evolution. And of course it also means that *we* – who are made of lots of Carbon and Oxygen and Nitrogen – are, in the memorable phrase of Carl Sagan, "starstuff pondering the stars."

Anyway, there is a definite end to this process of fusion reactions leading to heavier and heavier elements in massive stars: Iron. It turns out that fusion reactions from Iron to anything heavier than Iron are *endothermic* – you don't get energy out, but rather have to put it in. That's why nuclear reactors here on Earth – which proceed by *fission*, or the breaking apart of nuclei into smaller components – always begin with elements (such as Uranium) which are heavier than Iron. In terms of nuclear energy, Iron is the bottom of the barrel. You can get energy out by fusing smaller nuclei together or by breaking larger nuclei apart, but once you have Iron you are truly stuck.

So what we said above about Helium in the case of the Sun, applies in a more fundamental, non-negotiable kind of way to Iron for stars massive enough to produce it. That is, the Iron eventually produced by such stars forms a truly inert core, which just grows and grows as the fusion of still-remaining lighter elements continue above it. Since the core is inert, however, it doesn't produce any heat and hence doesn't contribute

much to the ability of the star to hold itself up against its own tremendous weight – i.e., against the inherent gravitational attraction of all its parts.

Eventually, the ultimate cataclysm occurs. The inert Iron core is simply unable to support the weight of the material above, and gives way: the entire star implodes, along the way crushing the electrons from the Iron atoms in its core right into their nuclei, where they are literally forced to react with protons. The result is that the core is converted into a uniform and immensely hot soup of neutrons. And now things get really interesting. Eventually, after shrinking in linear size by about 5 orders of magnitude – i.e., after being crushed to something like 10^{-15} times its original volume – the core finally again becomes very stiff, very difficult to compress further. All those neutrons, by virtue of a phenomenon that can only be understood using quantum mechanics, really don't like to get too close together. Over-simplifying only a little, the result is a very big "bounce": the ten or more solar masses worth of material that is racing in toward the collapsing core at tremendous speed suddenly encounter something akin to a brick wall. So all that material bounces off the suddenly solid ball of neutrons that used to be the core, and flies now *outward* at tremendous speed.

The implosion has been converted into an *explosion*. This process is called a core-collapse *supernova*. Most of the material of the star is blown out into surrounding space, often leaving behind an observable remnant called a planetary nebula such as that shown in Figure 5.16.

Also left behind by the supernova explosion is the solid ball of neutrons that used to be the star's inert Iron core. Such an object is called a "neutron star." We've already mentioned that, in the collapse, the volume of the core gets compressed by some 15 orders of magnitude. This is, not surprisingly, about the same as the ratio between the volume of a normal atom and the volume occupied by the atom's nucleus. Thus, a neutron star has roughly the same total mass as the Sun, but an incredibly large density comparable to that (or actually several times bigger than that) of atomic nuclei. A single teaspoon of neutron star material would weigh as much as a billion cars! Perhaps more dramatically, this means that the neutron star has a radius of only about 10 kilometers. A neutron star thus has as much matter as the Sun, compressed into a ball no larger than a small town!

Neutron stars don't shine in the visible part of the spectrum the way normal stars do, but they can be detected and observed by astronomers nevertheless. The first observation of (what was only later identified as) a neutron star occured in 1967 when two radio astronomers, Jocelyn Bell and Antony Hewish noticed a curious and extremely regular pulsation in the radio signal coming from a certain direction in the sky. They initially thought the signal must be some kind of noise in the apparatus, or of some other terrestrial origin, because the precisely-regular beep-beep-beeping seemed too strange to admit a heavenly origin. But that conclusion eventually became inescapable, and the mysterious astronomical source was dubbed a "pulsar."

The discoverers briefly considered the possibility that the beeping was being emitted by extra-terrestrials! But cooler heads prevailed, and in time the consensus developed that pulsars were rotating neutron stars, emitting a burst of radio-wave radiation toward us each time a certain part of their magnetized bodies passed by.

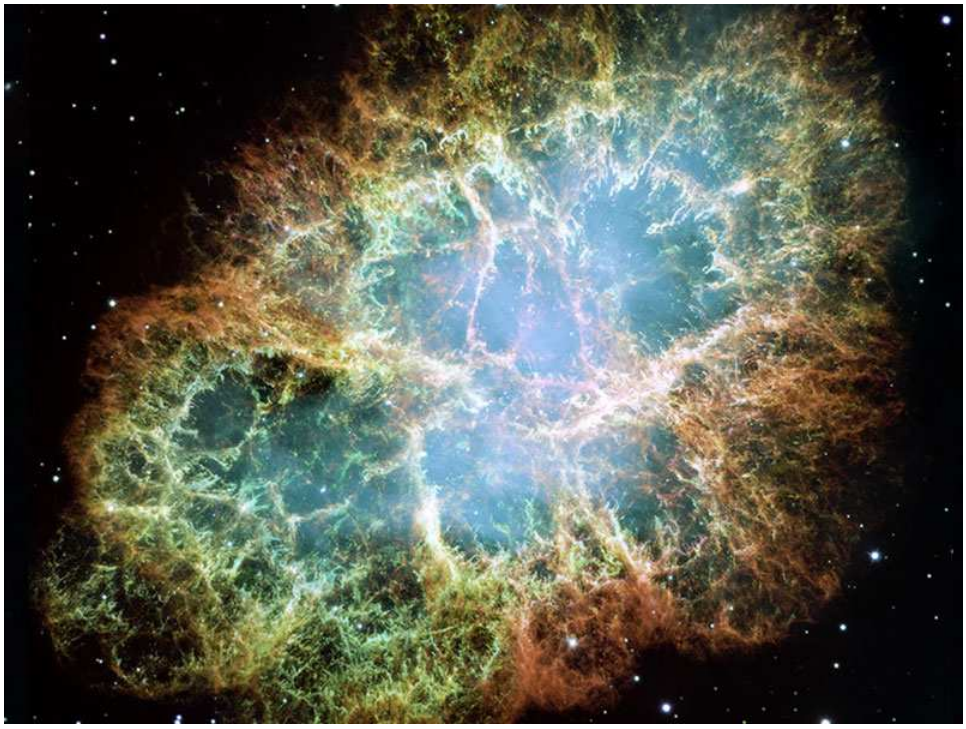

Figure 5.16: The Crab nebula. The supernova which produced it was visible to the naked eye from Earth and was actually observed and recorded by Chinese astronomers in 1054. Buried in the rubble is a rotating neutron star – the Crab pulsar – which was first identified by radio astronomers (but not yet understood to be a rotating neutron star) in 1968.

Many hundreds of pulsars were subsequently discovered, most with rotational periods of around one second. Note that this is one of the key pieces of evidence in favor of interpreting puslars as rotating *neutron* stars: if an ordinary Sun-like star were rotating with a period of one second, the result would be not just a small Equatorial bulge, but complete centrifugal annilihation (like the batter on the electric mixer pulled too soon out of the bowl).

Actually, even some neutron stars are fairly close to this limit. The Crab pulsar – the rotating neutron star in the Crab nebula shown in Figure 5.16 – has a period of only 0.033 seconds. And other pulsars have been discovered whose periods are as short as a few milliseconds. But no pulsars have been observed with periods less than a millisecond. This is additional evidence for the rotating neutron star model of pulsars. Following the calculation above of the critical distance for tidal disintegration, we may estimate the critical period for centrifugal disintegration as follows.

Assuming a roughly spherical body of mass M and radius R, the gravitational field

at the surface has magnitude

$$g = \frac{GM}{R^2} \tag{5.83}$$

while the centripetal acceleration of a point on the Equator (i.e., the centrifugal contribution to the effective gravitational field there, if we use a co-rotating reference frame) is

$$g_c = \omega^2 R = \frac{4\pi^2 R}{T^2}. \tag{5.84}$$

If these are equal, it means the disintegrative centrifugal force (on, say, some random neutron near the Equator) is comparable to the gravitational force keeping it together with the rest of the star. We may thus set them equal and solve for the critical rotation period

$$T_{crit} = 2\pi\sqrt{\frac{R^3}{GM}}. \tag{5.85}$$

If the rotation period is shorter than this (i.e., if the rotation is faster), the body will be torn apart by centrifugal forces.

What is the critical rotation period for a neutron star? Plugging in the rough numbers $M \approx M_{sun}$ and $R \approx 10\,km$ gives

$$T_{crit}^{NS} \approx 0.5\,ms. \tag{5.86}$$

So, on the premise that pulsars are rotating neutron stars, we can understand why no sub-millisecond pulsars have been observed.

The intriguing question of how a neutron star could get to be rotating up to a thousand times per second will be left for the Projects.

5.6 New Discoveries

Not only can gravitation be used to indirectly measure masses of previously known objects like moons and stars – it can also be used to discover entirely new objects! A contemporary terrestrial example of this was noted earlier in the chapter: sensitive modern instruments can measure the gravitational field $\vec{g}$ near the surface of the Earth with such great precision, that the tiny local fluctuations produced by, for example, underground mineral deposits can be detected. Such measurements have thus allowed scientists to know, beforehand, where to dig or drill to tap into valuable natural resources. This is a remarkable and beautiful example of the long-term practical benefits of progress in basic science.

5.6.1 New Planets

A less immediately practical but even more dramatic example of using gravitation to make new discoveries occurred in the 19th century. Recall that, according to the Ancient Greeks, there were (in addition to the Sun and Moon) *five* planets: Mercury, Venus, Mars, Jupiter, and Saturn. Of course, with the Copernican revolution, it was realized that the

Earth too was a planet, bringing the number to six. In the last section we noted the 20th Century discovery of the ninth planet (or, at any rate, what was formerly considered the ninth planet), Pluto. But when and how were the seventh and eighth planets – Uranus and Neptune – discovered?

Uranus was first recognized as something other than an ordinary star by the great English astronomer William Herschel in 1781. He stumbled on it essentially at random, in the course of his ongoing systematic surveys of the heavens. Herschel originally suspected that the newly discovered object was a previously unobserved comet, but subsequent observations revealed a more-or-less circular orbit around the Sun with a period of about 84 years. This object, subsequently named Uranus, was therefore a new planet. Its orbital radius was about 19 AU, or roughly twice that of Saturn, which previously marked the outer fringe of the known solar system.

Over the subsequent decades, though, the increasingly detailed observations of Uranus' actual motion increasingly failed to match up with theoretical expectations. This is not to say, for example, that Uranus orbited the Sun in a square rather than an ellipse, in gross violation of Kepler's laws. Actually, by this time it was known that all of the planets violated Kepler's laws to some small extent, because their orbits are influenced not only by the gravitational force of the Sun, but also by small gravitational forces exerted by the other planets. The point here is that the observed motion of Uranus seemed anomalous *even when these tiny inter-planetary perturbations were taken into account.*

Two quite reasonable hypotheses arose to explain the discrepancy. One possibility was that Newton's theory of gravitation simply didn't apply for an object at such a tremendous distance from the Sun. After all, Kepler's laws – taken here as summaries of the motion of the planets known about by Kepler – were the central pieces of evidence for Newton's theory, and that evidence pertained only to objects whose separation was at most the distance between Saturn and the Sun. There simply was no direct empirical data to support the extrapolation of Newton's inverse square law to longer distance scales. And clearly, by *some* appropriate modification to Newton's formula (i.e., by inventing the right fudge factor) the anomalous behavior of Uranus could be accounted for.

The alternative hypothesis was the existence of *another* previously-unknown object, whose gravitational influence on Uranus could in principle account for the small anomalies in its observed motion. This idea remained just another speculative gesture toward an appropriate fudge factor until two scientists, John Adams from England and Urban Leverrier from France, undertook to calculate the precise position and orbit of the hypothesized object. Adams and Leverrier worked independently and didn't know of each other's work, and the successful outcome led to a great international controversy. Adams probably finished the relevant calculations first, but his request to astronomers at the English Royal Observatory went unheeded for some time, since Adams was "merely" an unknown mathematician.

Leverrier, on the other hand, sent his predictions to a colleague at an observatory in Berlin. The eighth planet, Neptune, was discovered right away, in 1846, in just the region of sky that Leverrier (and Adams) had predicted. Neptune had an orbital radius of about 30 AU, and an orbital period of 165 years.

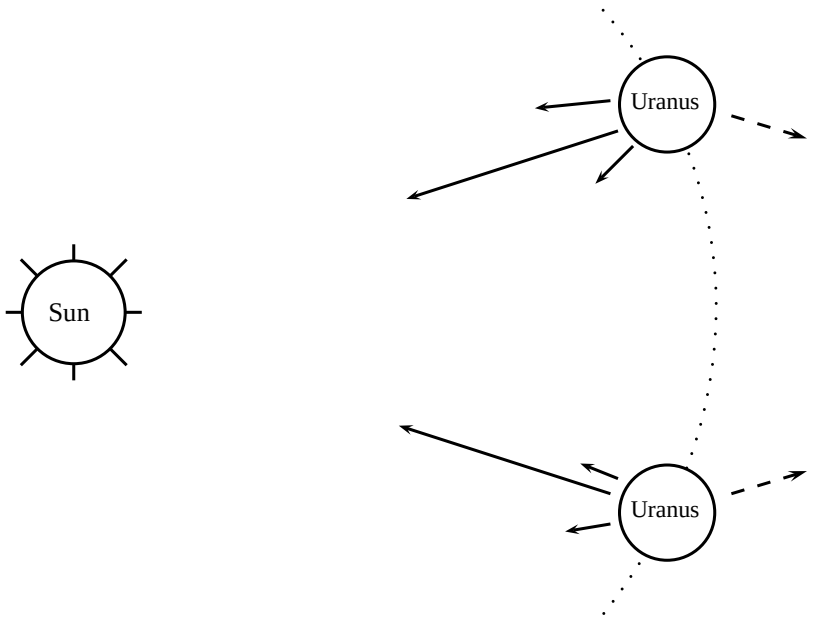

Figure 5.17: Schematic description of the calculations made by Adams and Leverrier. The acceleration of Uranus is produced by the joint effect of the gravitational forces exerted by the other bodies in the solar system. The forces exerted by the Sun, Jupiter, and Saturn are indicated by the three solid arrows, at two different times. Since the actual acceleration can be inferred from observation, the "anomalous forces" – the dotted-arrows in the Figure – can be computed. These are the gravitational forces exerted by the hypothetical new object, which of course turns out to be the new planet Neptune. Note that if the new planet were stationary, its position could be calculated by "triangulation." But since it, too, is expected to be in orbit around the Sun, the calculation is a little more subtle. If its own orbital radius were known, Kepler's third law would tell us the rate of its motion around the Sun, and its motion could be explicitly corrected for in the triangulation. In fact, both Adams and Leverrier made what turns out to have been a rather bogus assumption about the orbital radius of the hypothetical new object, based on a curious (one might say numerological) regularity in the orbital radii of the then-known planets called Bode's Law. As it turns out, their assumption about the orbital radius of Neptune was off by about 20%. It was a matter of sheer dumb luck that this bogus assumption didn't significantly affect Adams' and Leverrier's predictions!

Figure 5.17 gives a somewhat schematic indication of the kinds of calculations Adams and Leverrier made to predict the existence of Neptune, whose successful discovery was regarded as a major triumph for Newton's theory of gravitation. One scientist later described it this way:

> "The explanation by Newton of the observed facts of the motions of the moon, the way he accounted for precession and nutation and for the tides, the way in which Laplace [using Newton's theory] explained every detail of the planetary motions – these achievements may seem to the professional astronomer equally, if not more, striking and wonderful.... But to predict in the solitude of the study, with no weapons other than pen, ink, and paper, an unknown and enormously distant world, to calculate its orbit when as yet it had never been seen, and to be able to say to a practical astronomer, 'Point your telescope in such a direction at such a time, and you will see a new planet hitherto unknown to man' – this must always appeal to the imagination with dramatic intensity."

Actually, the same story more or less repeated itself, in a slightly less dramatic way, with the discovery of Pluto. In the decades after Neptune's discovery, *its* orbit was observed to deviate slightly from theoretical predictions, just as had that of Uranus decades earlier. This time, however, the deviations were much smaller. And so, although people this time guessed right away that the deviations were probably caused by yet another previously unknown planet, it was much harder to get a reliable estimate of that undiscovered planet's location. Pluto was finally discovered in 1930 as a result of these calculations, but this was after several decades of failed searches. And the specific calculations which led directly to Pluto's discovery were subsequently shown to be erroneous (in a more significant way than were those of Adams and Leverrier). Persistance and dumb luck thus played a great enough role in Pluto's discovery that it usually isn't considered any great triumph of Newton's theory of gravity. Nevertheless, it was ultimately Newton's theory which made that discovery possible, if only in an indirect sense.

We now understand better why the search for Pluto was so fraught with difficulty. Pluto is significantly less massive than any of the other planets – the next smallest, Mercury, is 20 times heavier! Thus, Pluto's gravitational perturbation on Neptune is very small. Moreover, Pluto turns out to be just one of a larger group of small, planet-ish objects occupying the outer fringes of the Solar System and marginally perturbing the orbit of (especially) Neptune. As Pluto was the first of these so-called trans-Neptunian objects (TNOs) to be discovered, it was naturally treated initially as another planet. But as more and more TNOs were discovered in the 1990s and 2000s, it became increasingly clear that Pluto had more in common (including its size, composition, and orbital character) with these other objects than it did with the eight planets. Pluto turns out not to even be the biggest of the TNOs. So you can see why Pluto was recently demoted from full planetary status – i.e., using a neologism inspired by this controversial episode, why Pluto was "plutoed."

5.6.2 Exo-planets

In the 1990s, the discovery of new planets extended beyond our own solar system for the first time. Of course, once it was understood that the stars were more or less like the Sun, only farther away, it became natural to speculate that other stars, like the Sun, would be centers of planetary systems. Newton even mentions this possibility in the Principia. But the first genuine empirical discovery of a planet orbiting a Sun-like star was made only very recently, in the 1990s. The star in question is 51 Pegasi, and its planet – *51 Pegasi b* – was detected indirectly, via its gravitational influence on the star.

The physics here is very similar to that presented already in the discussion of binary systems such as Pluto and Charon or binary stars. The idea is that, strictly speaking, the planet doesn't orbit a central, fixed, star. Rather, the star and planet both orbit around their mutual center of mass. Since only the star is directly observable (at least with current technology, and even this is starting to have exceptions), the planet manifests itself in the tiny back-and-forth periodic wiggle of the position of the star. Observations of the amplitude and period of this wiggle then allow some inferences about the properties of (or, less specifically but more profoundly, the *existence* of) the planet.

Actually, just as with the mass determinations of binary stars, it is more common (i.e., currently possible/easier!) to observe not the back-and-forth wiggle in space, but, instead, the back-and-forth fluctuations in *frequency* from which the periodic oscillations in the radial velocity can be inferred. Then the same formalism we developed before – Equations 5.75 and 5.76 – can be used to infer the mass and orbital radius of the invisible planet.

This is precisely the method that astronomers used to discover *51 Pegasi b*. A graph of the radial velocity of the parent star, 51 Pegasi, as a function of time is shown in Figure 5.18, and shown again as a function of the phase of the inferred periodic cycle, in Figure 5.19. The star moves in and out relative to us with a period $T = 4.23\, days$ and an amplitude of about $56\, m/s$. Unfortunately, the planet does not appear to pass in front of the star during its orbit, so the inclination i of its orbit remains unknown. Nevertheless, it is possible to put a lower limit on the planet's mass. This turns out to be

$$m_2 \geq .45\, M_{Jup} \tag{5.87}$$

where M_{Jup} is the mass of (our own) Jupiter. It is also possible to infer from the data that 51 Pegasi b has an orbital radius of $R_2 = .05\, AU$. So the planet is (probably) roughly as big as Jupiter, but – compared to the real Jupiter – very close to its parent star. It, and the many other extra-solar planets like it which have been subsequently discovered, are therefore sometimes called "hot Jupiters."

You should probably be wondering: how exactly did the scientists determine the *mass* of this extra-solar planet? In our discussion of measuring masses in binary systems (such as the Pluto-Charon system or a double star system), we found that one must determine empirically not only the period of the orbit(s), but also the radii or maximum radial velocities of *each* of the two bodies, in order to determine either of the masses. Recall, for example, Equations 5.75 and 5.76. But the extra-solar planet discussed here remained *invisible*: so while the period and v^{max} of the *star's* wobble could be observed,

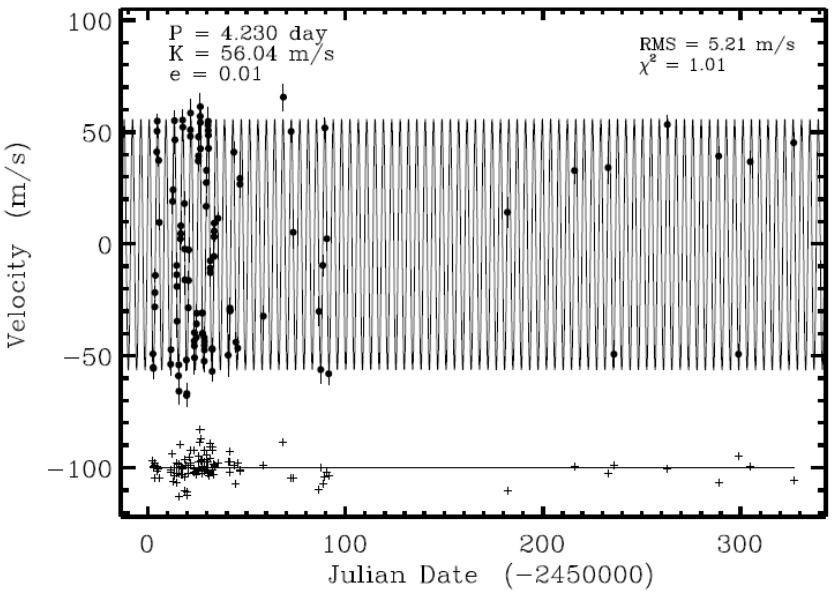

Figure 5.18: Data for the radial velocity (inferred from Doppler effect observations) of the star 51 Pegasi over the course of about a year. The wiggly line is a sinusoidal fit to the data, which maybe looks a little suspicious given the seemingly random character of the data. The residuals for the fit, however, are shown below and indicate that the fit is quite good. (Observations of nearby, non-wiggling stars indicate that there is about a 5 m/s uncertainty on any of the velocity measurements – so the residuals are just the size one would expect given the inherent accuracy of the data.) Note also how the data were taken over the course of the year: lots and lots of observations over a month or two to get an accurate guess of the periodicity, and then just a few measurements, almost randomly spaced over the subsequent months, to test whether the guessed periodicity continues to fit the data over a longer timescale. As the points to the right in the plot of the residuals shows, it does. This provides much more confidence that the fit is correct, than would (say) the same total number of data points crammed into just a month of observation, or the same number of data points uniformly spaced over an entire year.

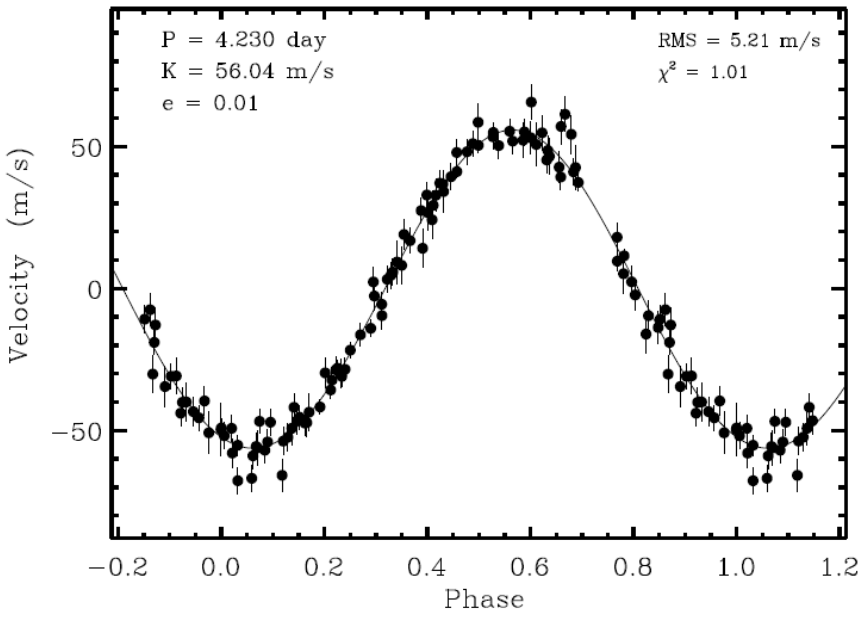

Figure 5.19: The same data as the previous figure, but plotted as a function of the phase of the inferred 4.231 day cycle. This makes the sinusoidal character of the in-and-out motion of the star particularly clear.

the v^{max} for the actual planet could *not* be. One of the three crucial pieces of evidence seems to be missing!

Here is the resolution to this puzzle: we can make an educated guess about the mass of the star by measuring its luminosity and then using the mass-luminosity relation discussed above and shown in Figure 5.14. This, of course, requires an assumption that the star in question is relevantly like the stars whose masses and luminosities were shown to be so correlated. But there is abundant evidence for this hidden away in the light emitted by the stars – in particular, in their *spectra*, i.e., in the distribution of their emitted light across the frequency spectrum.

Most of the subsequently discovered extra-solar planets were discovered by more or less the same method. And so that's, in a nutshell, how scientists in recent decades have established that, as had long been suspected, there do exist planets orbiting stars other than our own (the Sun) – and how they measure the mass and orbit of the planets to boot. At this writing, hundreds of extra-solar planets have been positively detected, and the rate of their discovery is continuing to accelerate.

5.6.3 Dark Matter

Let us close with one more example of a recent discovery made using Newton's theory of gravitation: the discovery of so-called "dark matter." This follows roughly the same pattern discussed above under the heading of measuring masses. Moons orbit planets and planets orbit stars, and the orbital character of the orbiting body (in particular its period and radius) can be used to infer the mass of the central gravitating body. Similarly, it turns out that stars arrange themselves in enormous clusters called galaxies, with the individual stars all (more or less, on average) orbiting around the galactic center.

A particularly beautiful type of galaxy – see for example Figure 5.20 – has most of the stars clumped up into a spiralling disc.

Using the same Doppler-effect-related spectroscopic techniques described before, astronomers can measure the speed with which individual stars (or groups of them) orbit around the center of their galaxies. For a star on the outer fringes of its galaxy, the orbital speed should be given approximately by the familiar Newtonian calculation which sets the centripetal gravitational force (produced collectively by all the stars in the galaxy) equal to the mass m of the star in question times its centripetal acceleration, $a_c = v^2/R$. Thus we expect

$$\frac{GMm}{R^2} = m\frac{v^2}{R} \tag{5.88}$$

where M is the total mass of the galaxy and R is the galactic radius of the star in question. This reduces to

$$v = \sqrt{\frac{GM}{R}}. \tag{5.89}$$

Of course, we've assumed here that the rest of the galaxy can be treated as if it were a single point of mass M located at the galactic center. For a star anywhere near the middle of the galaxy, this is a terrible approximation – for such a star, "the rest of the galaxy" will be pulling it in several directions at once and from several different distances. But for stars out on the outer fringes of the galaxy, "the rest of the galaxy" *is* all pulling it in the same direction. Of course, the mass is distributed in something like a disc shape (not a perfect sphere) so we might worry that there are corrections to the simple point mass formula like those we dealt with in discussing the Earth's Equatorial bulge. And, indeed, such corrections should exist. Nevertheless, they will be increasingly small corrections for stars that are truly on the fringes, very far from the galactic center.

All of this is just meant to underscore that, although Equation 5.89 is derived with the crudest possible approximations, we have good reason to think it should apply to stars on the outer fringes of galaxies. Yet, when the orbital velocities of such stars are actually *measured*, they do not appear to vary with R in the way that Equation 5.89 suggests they should – i.e., decrease with R as $1/\sqrt{R}$. Instead, what is observed is that the orbital velocities of stars on the outer fringes of galaxies tend to be quite *constant* – independent of R. See Figure 5.21.

What does this mean? Obviously it means that one of the assumptions we've made in generating the wrong expectation, is itself wrong. One possibility (again, just like in the discussion of Neptune's discovery) is that Newton's formula for the gravitational

Figure 5.20: The galaxy M51, also known as the Whirlpool Galaxy. It is located about 30 million light years away, and has a radius of (roughly, since there's no well-defined edge) about 30 thousand light years or about 9 kiloparsecs. (Recall that a parsec is the distance a star would have to be from the Sun in order to exhibit a parallax of one second of arc, i.e., 1/3600 of a degree. The closest stars to the Sun are about a parsec away, which is about 200,000 AU. It's nice to have a sense of the relative order of magnitude of these things. To summarize: the nearest stars are hundreds of thousands of times (5 orders of magnitude) farther away from us than the Sun. And the galaxy – ours turns out to be roughly the same size as the Whirlpool – is another 10,000 times (four orders of magnitude) bigger than that. The galaxy, then is some 9 orders of magnitude – a billion times – bigger than the Earth's orbit around the Sun. The distance between Galaxies is then another factor of a thousand – three more orders of magnitude – bigger than that. And it turns out galaxies themselves form clusters, with relative gaps between *them*. And believe it or not, even the galaxy clusters form clusters – "superclusters" they're called. So there is important and interesting structure in the universe across an incredibly broad spectrum of length scales. And we haven't yet even begun to discuss the *small* end of the spectrum!

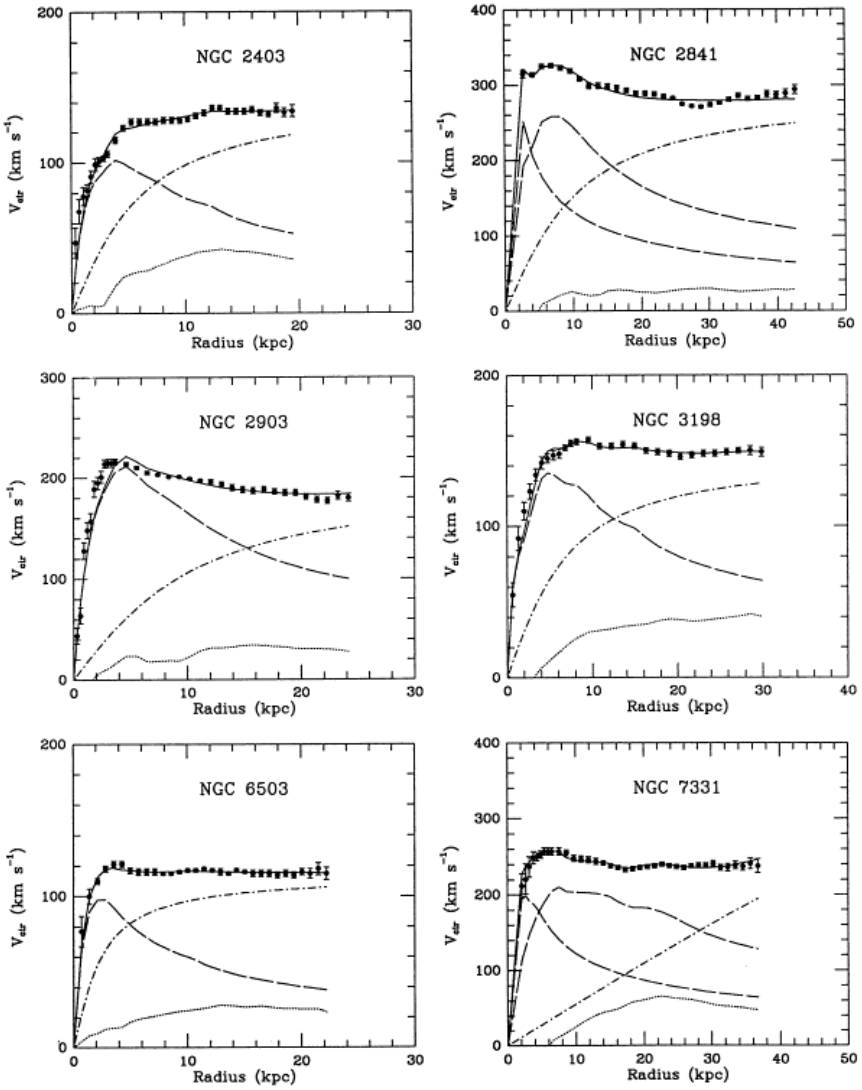

Figure 5.21: Observed rotation curves for six "typical" galaxies. Dots are data points for the rotational velocity (as measured via the Doppler effect). The three curves below are the components of a three-parameter fit to the rotation curve data: "the dashed curves are for the visible components, the dotted curves for the gas, and the dash-dot curves for the dark [matter] halo. The fitting parameters are the mass-to-light ratio of the disc (M/L), the halo core radius (r_c), and the halo asymptotic circular velocity (V_h)." Image and parts of caption from "Extended rotation curves of spiral galaxies – Dark haloes and modified dynamics" by K. G. Begeman, A. H. Broeils, and R.H. Sanders, Royal Astronomical Society, Monthly Notices, vol. 249, April 1 1991, pages 523-537.

force simply doesn't apply at these (now *really*) large distance scales. This is considered a going hypothesis in current research – the idea generally goes by the name Modified Newtonian Dynamics, or MOND for short.

But by far the more popular interpretation of the surprising data is the hypothesis of so-called "dark matter." The idea is that, although the stars we have been talking about *appear* to be on the fringes of the galaxy, in the sense that virtually all of the *observable* matter (i.e., the other stars) are much nearer the galaxy's center, in fact those stars are not near the galaxy's "edge" because the galaxy consists not just of the visible stars but also of some mysterious non-visible ("dark") matter which, nevertheless, gravitates.

Another way to put the problem and the (hypothesized) solution is this: if you just calculate, using Equation 5.89 and the actual velocities and radii for some stars near the (apparent) fringes of a galaxy, the mass M of the galaxy – that calculated mass is substantially *bigger* than the mass you would have guessed by counting up all the stars in the galaxy and multiplying by the average mass of a star. So there must exist, in addition to the stars (which both gravitate and produce light), some "dark matter" (which gravitates but does not produce light).

Note that the dark matter is no mere marginal correction. Current estimates (based not only on the velocities of stars in galaxies, but several other methods as well) suggest that there is something like five or ten times more mass in dark matter than in ordinary "light matter" (mostly stars).

All of this obviously raises the question: what *is* this dark matter? No answer can be given, because it is simply not yet known. Some have speculated that the dark matter is ordinary matter that does not produce light – e.g., billions of Jupiter-sized "planets" roaming around the universe. This is an intriguing possibility if only because it doesn't require the postulation of any wholly new type of matter. It is, however, very difficult to understand where all these Jupiters would have come from. (Suffice it to say that otherwise strongly-confirmed theories of the evolution of stars and planets do not suggest that such Jupiters could be produced in the needed numbers and with the needed spatial distribution.) Other proposed dark matter candidates include exotic new sub-atomic particles (beyond the electrons and quarks of which ordinary matter is made). Such models are, in a way, more consistent with the overall astronomical evidence. But they suffer from the fact that none of the candidate particles have ever been observed in particle physics experiments.

The identity of dark matter thus remains a profound mystery.

And while it may perhaps feel a bit anti-climactic, that is a fitting way to close our survey of the astrophysical applications and implications of Newton's theory of gravitation. As we have seen, Newton's theory forms the crucial support for virtually everything we have discovered about our world and our universe. But it also continues to provide the basic context for the questions and puzzles at the current frontier of our knowledge. Surely there could be no stronger testament to the theory, especially considering that we are now well into the fourth century after its publication!

Questions for Thought and Discussion:

1. Suppose a piece of pizza dough were thrown up, spinning, repeatedly. (Or equivalently: suppose it were set spinning in outer space.) Would it keep getting flatter and flatter indefinitely? Or would it, like the Earth, reach some kind of equilibrium beyond which further flattening would increase the total energy? What's the relevant difference, if any, between the pizza dough and the Earth?

2. Everyone knows that the highest point on Earth is the top of Mt. Everest (on the border between Nepal and Tibet). But actually, this depends on what you mean by "highest." The point on Earth *whose distance from the Earth's center* is greatest, is the top of Mt. Chimborazo in Ecuador. What is going on here? How can the highest point (by the usual meaning of "highest") not be furthest from the center? What exactly is the "usual meaning" of "highest"?

3. Because of the bulging of the earth near the equator, the source of the Mississippi River, although high above sea level, is nearer to the center of the Earth than is its mouth. How can the river flow 'uphill'?

4. If you turn your car to the right, you experience being pulled to the left, e.g., pushed up against the left side of the interior of the car. Is there really a force pushing you left?

5. Suppose the Earth were perfectly spherical. Would your weight change as a result of moving North or South, i.e., changing your latitude? Would the reading of your bathroom scale change? Explain. What does a bathroom scale actually measure?

6. Near the equator, during what part (or parts) of the year are high tides highest and low tides lowest? How about near the poles? How about in middle latitudes?

7. The caption to Figure 5.7 perhaps gave the impression that – wherever there *are* two low tides per day – the two low tides should be equally low. As you may have noticed on ocean visits, this is not true. The two low tides each day are not necessarily equally low. The question is: can this be understood from the "equilibrium" model of tides that most of the text's discussion (and Figure 5.7 in particular) is based on? Or must we resort to the complicated sloshing of tidal waves to understand this? To make the assignment a little more concrete: can you come up with a scenario (i.e., a relative arrangement of the Earth, Moon, and Sun) in which (say) observers at middle latitudes will experience two low tides that are not equally low?

8. Would you expect the amplitude of tides to be higher in Hawaii, or in Florida (at about the same latitude)? Why?

9. How, if at all, would the tides be different if, instead of orbiting one another, the Earth was rigidly stapled to the cosmic graph paper, with the Moon orbiting around it?

10. When we calculated the height h of the (rotation-produced) Equatorial bulge, our formula was too small by roughly a factor of two because we ignored the gravitational effect of the Equatorial bulge itself. (Thinking in terms of $\vec{g}_{eff}$ near the surface of the Earth, the point is that the Equatorial bulge makes the true gravitational acceleration $\vec{g}$ itself have a "true horizontal" component. Or, thinking in terms of the energy argument, the point is that a mass of material would actually be moving gravitationally "downhill" in going from the pole to the Equator, i.e., $\Delta PE^{A\to B}$ is not zero, but negative and about half as big as $\Delta PE^{B\to C} = mgh$.) Was there a parallel error in our calculation of the height h of the tides? That is, is the true equilibrium tide height h (the altitude difference between the high and low tide points) produced by the Moon about twice what we said? About a meter, rather than $54\,cm$?

11. The text discussed how, in millions of years, the Earth and Moon will become "tidally locked" in a face-to-face dance in which the same face of the rotating Earth is always pointing toward the Moon. It was mentioned in passing that the Moon *already* orbits in such a way that it presents the same constant face to the Earth. Why do you think it does this? Is this just a coincidence? (You better not say yes – it is extremely common for moons orbiting other planets in the solar system to orbit this way!) If your answer has something to do with tides, does this square with the fact that the Moon is dry (no oceans)?

12. Following up on the previous question, can you explain why it is so common for moons in the solar system to have very circular (as opposed to highly elliptical) orbits?

13. The text encouraged you to think about the gradual increase in the orbital radius of the Moon in terms of the Work-Energy theorem. But what was said there was actually a little sketchy. It's true that the net gravitational force exerted by the Earth on the Moon has (because of especially the near-side tidal bulge) a small "easterly" component. We hinted (too quickly) above that this was a component of the net force that was parallel to the direction of motion. Hence, by the Work-Energy theorem, we said, positive work is being done on the Moon and so its total energy should increase – which we then interpreted as meaning that its orbital radius should increase. But that's not what the Work-Energy theorem says! The theorem says that the net work done on an object should equal the change in its *kinetic* energy. But when the orbital radius of an orbiting body (in a roughly-circular orbit) increases, its kinetic energy does *down*, not up! (You should prove this to yourself.) Does this mean the Work-Energy Theorem is actually contradicted by the behavior of the Earth-Moon system? How can you resolve this paradox?

14. Consider the distant future in which the Earth-Moon system has become tidally locked. Now step back and think about the system comprising the Sun and the Earth-Moon. What should happen in the even more distant future?

15. Do you think Pluto should be classified as a planet? Why or why not? What, if anything, hinges on this question? Is it a pointless discussion?

16. Our discussion of determining the masses of the stars in binary star systems assumed that the stars' orbits were *circular*. Is this necessary? Is there any reason that highly-elliptical orbits for stars in binary star systems should be rare?

17. We all live well inside the Earth's Roche limit. How come we aren't ripped apart by tidal forces?

18. Think about the calculations that Adams and Leverrier made to predict the location of Neptune, as sketched in Figure 5.17. Can you understand why they needed to make some assumption about Neptune's oribtal radius? Strictly speaking, given Kepler's thrid law, wouldn't only one orbital radius be consistent with the two "anomalous forces" shown in the Figure? So why was this assumption necessary? Think about what other factors were glossed over in the text, including the precision with which these anomalous forces could be calculated. Also, roughly what period of time must elapse between the two times when Uranus' position (and the forces acting on it) are shown in that figure?

19. The extra-solar planet discussed in detail in the text, 51 Pegasi b, was described as a "hot Jupiter." Though not all of the other currently known extra-solar planets are hot Jupiters, *most of them are*. Do you think this means that most planets outside our own solar system are hot Jupiters? Why or why not?

20. The text explained how, by measuring the masses of stars in binary systems close enough to the earth that their intrinsic luminosities can also be calculated, the empirical Mass-Luminosity relation (plotted in Figure 5.14) was worked out. Explain qualitatively how, once this Mass-Luminosity relationship is known, one could use it to determine the distance to another binary system which, say, is sufficiently far away that its distance cannot be determined by parallax.

21. An ordinary star is in a close binary orbit with a neutron star. Suppose now that the ordinary star becomes a red giant, such that its outer surface gets inside the neutron star companion's Roche lobe. What will start happening, and what do you think will happen to the neutron star eventually?

Projects:

5.1 In the text, we derived Equation 5.13 for the oblateness of a rotating sphere like the Earth in two different ways. There is a third way, which is probably easier than the other two (especially once you understand the other two!). It involves using energy considerations as in the first method, but using a non-inertial co-rotating frame of reference as in the second method. The crucial point is that the centrifugal force which appears in the co-rotating frame implies an additional contribution to

the potential energy. Work out an expression for this, and use it to re-derive (yet again) Equation 5.13.

5.2 Let's try to estimate the quantity $\Delta PE^{A\to B}$ that plays an important role in the calculation of how the Earth's oblateness depends on its rotation rate ω. The simplest model is probably to think of the Earth as a perfect sphere plus a "hula hoop" near the Equator. The spherical part is, of course, spherically symmetric and so won't contribute anything to $\Delta PE^{A\to B}$. We need then only try to estimate the contribution from the "hula hoop." To begin with, write down some approximate expressions for the mass and radius of the hula hoop, in terms of the total mass of the Earth M, its radius R, and the height of the Equatorial bulge h. (The hoop's mass should probably be something like the total mass of the Earth times the fraction of the Earth that is in the Equatorial bulge as opposed to the underlying spherical core, which fraction will have to be estimated. The hoop's radius should probably be R_{earth} or $R_{earth}/2$ or something like that, to take account of the fact that not all of the Equatorial bulge is right at the Equator, i.e., much of the mass of the bulge is closer to the spin axis than R_{earth}.) Now use calculus to develop expressions for the potential energy of a point mass m a distance r from the center of a hula hoop (radius R_{hoop} and mass M_{hoop}) (a) along the symmetry axis and (b) in the plane of the hoop. (For this problem it will be sufficient to expand these expressions in powers of R_{hoop}/r and drop terms smaller than R_{hoop}^2/r^3, if you want.) Now compare the potential energy at the same distance, $r = R_{earth}$, along the two different directions – i.e., at the Pole vs. at the Equator. You should find that the difference in potential energy is

$$\Delta PE = -\frac{3}{4}\frac{GM_{hoop}mR_{hoop}^2}{R_{earth}^3} \tag{5.90}$$

which should reduce to something in the neighborhood of

$$\Delta PE \sim -\frac{1}{2}mgh \tag{5.91}$$

but with probably, some other dimensionless fraction (like 2/3 or 9/32 or something) out front, depending on exactly what you said when you estimated M_{hoop} and R_{hoop}. So the point of this calculation is only to show that you can get in the general ballpark of the result claimed in the text – namely, that $\Delta PE^{A\to B}$ is *in the neighborhood of* $-1/2$ times $\Delta PE^{B\to C} = mgh$, which effectively *doubles* the prediction for the Earth's Equatorial flattening, bringing that prediction very very close to the actual, empirically measured value.

5.3 Here's another nice model for the not-quite-spherical Earth. This has the advantage of being simpler than the sphere-plus-hula-hoop model considered previously, but the disadvantage of failing to possess the same rotational symmetry as the actual Equatorially bulging Earth. That can cause problems, but is actually OK so long as we restrict ourselves to discussing features of the Earth's gravitational field that

are confined to some cross-sectional plane like that shown in Figure 5.1. Here is the model: pretend that the Earth is a dumbbell, i.e., two point masses separated by some distance and both located near (but not quite at) the real Earth's center. Suppose that the two masses each have one half of the Earth's total mass so that, together, they are (in terms of total mass) equivalent to the actual Earth. Then: what should their separation be in order to reproduce the empirical fact reported at the end of section 5.1 – namely, that at equal distances R_{earth} from the center, the gravitational acceleration near the Pole is $0.048\,\mathrm{m/s^2}$ smaller than the gravitational acceleration near the Equator? The idea here is to take this empirical fact as fixing the (otherwise free) separation parameter in the model. We can then test the accuracy of the model, for example, as follows: what does it predict for the quantity $\Delta PE_{A\to B}$ that plays some role in the energy-based calculation of the size h of the Equatorial bulge? (Note: later, in Project 5.8, we will use this same model to calculate the period of the Earth's precessional motion!)

5.4 Jupiter is an oblate spheroid just like Earth, but with an observationally measured flattening parameter of $f \approx .065$. This flattening is so large that Jupiter's oblate shape is noticable just looking through a telescope! Given values for Jupiter's mass ($M_{jup} = 1.9 \times 10^{27} kg$) and radius ($R_{jup} = 70,000\,km$), what do you think its spin angular velocity should be? What is the corresponding period of revolution (i.e., the duration of a Jovian day)? This last can be estimated by watching observable surface features (such as the famous "great red spot") move across the surface. Your teacher will tell you the rotational period that comes from such observations so you can check the accuracy of your prediction. (By the way: do you understand how Jupiter's mass and radius can be known?)

5.5 The Sun's mass is $M_{sun} = 2 \times 10^{30} kg$ and its radius is $R_{sun} = 7 \times 10^8\, m$. Observation of Sun Spots progressing slowly and systematically across the visible face of the Sun suggest that the Sun rotates with a period of about 25 days. What do you predict should be height of the Sun's Equatorial bulge and/or its flattening parameter? Should the Sun's oblateness be obvious through a telescope the way Jupiter's is?

5.6 In the discussion of the Earth-Moon tidal interaction, we mentioned that the Moon's orbital angular momentum is proportional to the square root of its orbital radius. Show that this is right.

5.7 Using angular momentum conservation, find the angular velocity at which the Earth and Moon will both move, long in the future when they are finally tidally-locked, face to face. How long will the Earth day be then? How long with the "month" (the period of the Moon's orbit) be?

5.8 Approximate the torque exerted on the Earth by the Moon due to the Earth's tidal bulge. Roughly how long will it take for the Earth and Moon to become tidally locked?

5.9 Use the model developed in Project 5.3 and Equations 5.52-5.53 to calculate the *torque* exerted on the (bulging) Earth by the Moon, say during the part of the monthly cycle when the Earth's spin axis is tilted maximally toward the Moon – i.e., the Lunar equivalent of the Summer Solstice. You should be able to get out an actual honest-to-god number (of Newton-meters or whatever your favorite unit of torque is). Now think about how this torque varies during the monthly Lunar cycle. What do you think the *average* torque is? Now do all of this again for the torque exerted by the Sun (to whatever extent, that is, doing anything again is required). Add the two results together to find the total time-average torque exerted on the Earth. And finally plug the result into Equation 4.99 from Chapter 4 to *predict* the Earth's precessional period. (Recall from Chapter 1 that the actual period is about $26,000$ years. You should get something in this ballpark, which is pretty cool given the crudeness of this model for the Earth. We'll count it as definitely understanding the cause of the observed rate for the "precession of the equinoxes.")

5.10 According to a speculative theory going back to George Howard Darwin (son of the biologist Charles Darwin), our moon might have been formed from material of the Earth's crust flung off by the rotating Earth. How fast would the Earth have had to rotate at that time to make the latter picture plausible?

5.11 Here's a cute little model that will help you understand the tides: consider a little "barbell" type thing made of two masses m connected by a spring of spring constant k and rest length L. Suppose this object is orbiting another object (a star or planet or whatever) of mass M, with an orbital radius R. Consider the various ways it could orbit (axis-on, side-on, no spin angular momentum, spinning fast, etc.) and address, for the different types of orbit (or different moments during the orbit as appropriate): what is the separation between the two masses?

5.12 The moon, like the Earth, is not a perfect sphere. Its biggest "radius" exceeds its smallest "radius" by about 2.2 kilometers. Can you understand this number based on the physics in this chapter? In particular: is the Moon's 2.2 kilometer bulginess a result of rotation or tidal forces or what? What about the fact that the Moon doesn't rotate – i.e., that it always presents the same face to the Earth?

5.13 In the discussion of the torque exerted by the Moon on the Earth's tidal bulges (and its effects) it was mentioned that the length of the day is increasing by 1.6 milli-seconds per century, and that the radius of the Moon's orbit is increasing by 3.5 cm per year. From each of these numbers, calculate the rate of change of the associated angular momentum. They should be about the same (with one positive and one negative), in light of angular momentum conservation for the combined Earth-Moon system. Are they?

5.14 Not long after Pluto's discovery in 1930, its distance from the Sun was measured to be about 39.5 AU. (Actually, Pluto's orbit is highly elliptical, so that's just an average. You should be able to explain how this distance could be measured!) As

seen from Earth, around the time Pluto's distance from the Earth is 38.5 AU, its moon Charon appears to oscillate back and forth about Pluto with a period $T = 6.39$ days and with an amplitude of $3.4 \times 10^{-}6$ radians. (This is equivalent to the angular diameter of Charon's orbit being 1.4 arc-seconds.) What is Pluto's mass? What not-quite-true assumption explains why your answer is slightly different than the currently accepted value of 1.52×10^{21} kg?

5.15 Our derivation of the "Roche limit" for tidal disruption of a moon left something important out. (Actually it left several important things out, but this is the biggest and easiest to address.) For any moon which is in danger of approaching the Roche limit, it is likely that the tidal forces are already strong enough to have gotten the moon into a tidally locked synchronous rotation in which it always presents the same face to the planet. This means, as viewed from an inertial frame, that the moon will be *rotating*, which means that there will be a "centrifugal" tendency for the moon to come apart, in addition to the tidal effect noted in the earlier derivation. It turns out that the centrifugal effect is just about as big as the tidal effect, so it really should be included. So include it! For definiteness: calculate the centrifugal component to the effective gravity at the surface of the moon. This is given by

$$g_c = \omega^2 R_{moon} \tag{5.92}$$

so your only task here is to find an expression for the rotational angular velocity ω in terms of the mass of (M_{planet}) and distance to (r) the planet. Hint: for a tidally locked synchronous orbit the *rotational* and *orbital* angular velocities are the same. Now that you've got that worked out, develop a new-and-improved formula for the Roche limit.

5.16 Using the new and improved formula for the Roche limit that you got from the previous Project, calculate the Roche limit for Saturn. Of course, if you were to look up the fact that Saturn's radius is 6×10^7 meters, you could calculate the Roche limit in meters. But it is more revealing to just calculate the dimensionless multiplier by which the critical distance exceeds the planet's radius. Saturn's average mass density is about $0.7\,\mathrm{g/cm^3}$. What's a reasonable value to use for the mass density of the (perhaps shredded) moon? Is the result more or less consistent with the picture of Saturn in Figure 5.15 and the hypothesis that the rings exist as rings because they are inside the Roche limit?

5.17 You might think some special mechanism is needed to explain how a neutron star could come to be rotating up to a thousand times per second. But in fact, the conservation of angular momentum is sufficient. First, explain qualitatively why the angular momentum of the progenitor star's core should be conserved during the core-collapse supernova which produces the neutron star, and why this collapse would magnify any small initial angular velocity into a much larger angular velocity. Write an expression for the core's moment of inertia in terms of its mass and radius, and then use conservation of angular momentum to derive an expression for the final rotation period as a function of the initial rotation period and the initial and

final radii. What initial rotation period is needed to produce a millisecond pulsar? Is this reasonable? (Hint: recall Galileo's sunspot observations from Chapter 2.)

5.18 Estimate the amount by which the gravitational binding self-energy of the core changes when a star undergoes a core-collapse supernova. Work out the actual number in Joules. Is this an increase or a decrease in its energy? Where do you think the missing (or extra?) energy comes from (or goes?)?

5.19 It turns out that only about a hundredth of a percent (0.0001) of the energy difference calculated in the previous Project is converted into visible radiation. (Most of the energy escapes in the form of neutrinos, a type of particle that is copiously produced as a by-product of the electron + proton → neutron reaction which occurs during the collapse.) But still, this is a tiny fraction of a huge amount of energy. Calculate the luminosity of a supernova if this entire energy is given off over a period of about a month (which is about the period during which a typical supernova is at its brightest). For comparison, the Sun's luminosity is about

$$L_{sun} = 4 \times 10^{26}\, W. \tag{5.93}$$

5.20 The text discusses how a neutron star is formed during – and then left behind by – a core-collapse supernova explosion. It is possible, however, for the core to turn not into a neutron star but something else instead: a black hole. As a preliminary definition, we would say that a black hole is any object for which the escape velocity from the surface exceeds the speed of light ($c = 3 \times 10^8\, m/s$). The idea is then that even light cannot escape from the surface, and so the object will appear black. It turns out, however, that according to general relativity, one cannot have such a small, rigid black hole with a well-defined (albeit unobservable) surface. Rather, such an object would necessarily be unstable and collapse indefinitely, forming a point or "singularity." (Actually, this shouldn't be taken too seriously either, since at some point such a high density will be reached that even general relativity doesn't apply – but then nobody has any way to guess what might happen.) In any case, although it doesn't exactly have a "surface," even a point mass will have some specific distance away from it – the so-called "Event Horizon" radius – at which the escape velocity equals the speed of light. The idea is then that light (or anything else, since it is another principle of relativity theory that nothing can go faster than light!) which finds itself inside the Event Horizon can never escape. Find an expression for the Event Horizon radius in terms of the mass of the central body. How big is it (in kilometers) for the Earth? For the Sun? By what factor would you have to compress a neutron star (whose mass is the same as the Sun's mass and whose radius is 10 km) to convert it into a black hole?

5.21 Your teacher will give you some data for the radial velocities of the two stars in an eclipsing binary star system, over time. Determine the masses of the two stars.

5.22 Use the graph and associated data discussed in the text, to work through the calculation of 51 Pegasi b's orbital radius and mass. (Hint: this is mostly an

exercise in appropriately simplifying Equations 5.75 and 5.76 for the special case that one mass is much greater than the other. If you do that first, it should then be relatively straightforward to plug in the numbers given in the text for the period and amplitude of 51 Pegasi's radial velocity oscillation.

5.23 Your teacher will give you a file containing data for the radial velocity of a star at a number of times over the course of several months. The intrinsic luminosity of this star is about 10 times the luminosity of the Sun. Try to find a good sinusoidal fit to the data, and use the results to argue for the existence of (and calculate a lower limit on the mass of) an associated extra-solar-planet.

5.24 Stars on the (apparent) fringes of our own galaxy, the Milky Way, appear to orbit the center with a speed of about 225 km/sec. For a star whose galactic radius is 17 kiloparsecs (about twice as far from the galactic center as the Sun), what does this imply about the total mass of the Milky Way galaxy? How does this compare to the results of statistical studies which show that the Milky Way contains about 50 Billion ($10^1 1$) stars comparable to the Sun? Based on the numbers given here, what fraction of the Milky Way's total mass is dark matter?

Chapter 6

Numerical Solutions of Differential Equations

As a final sub-topic under this first half of the course (on astronomy, gravitation, and astrophysics), we are going to discuss the use of a computer to solve differential equations. We'll begin with some general discussion to make clear (a) what a differential equation is, (b) what it means to solve it numerically using a computer, and (c) how to go about actually doing this. Then we'll develop and practice our new skills by taking one final look at several lingering loose ends from earlier chapters. For example, you'll finally be able to demonstrate something that we have claimed and used but not yet actually proved – that the proposed inverse-square-law gravitational force field produced by the Sun gives rise to *elliptical* orbits for the planets, with the Sun at one focus. You will also be able to try out *other* force laws (e.g., an inverse-cube-law, or whatever else you want to make up) and see what sorts of planetary trajectories are produced. This fills an important gap in the logic from Kepler's laws to Newton's theory: an inverse-square-law force is not just *a* way to account for the Keplerian orbits of the planets; it's the *only* way to account for them. As another example, you'll also be able to explore the departures from perfect Keplerian orbits that are produced by the gravitational forces exerted by the planets on each other.

The structure of this chapter will be slightly different from the previous ones. The main text here will hardly concern physics at all, but will instead help you get up to speed on some mathematical and computational ideas that you'll then use – to explore some interesting physics – in the Projects.

6.1 Overview of Differential Equations

Let's begin with some basic background about differential equations as such. What is a differential equation (DE)? Basically, it's any equation involving the *derivatives* of an unknown function. To *solve* a differential equation means to find the function whose derivatives have the properties stated in the DE. This probably sounds more complicated and unfamiliar than it is – you've definitely seen DEs before, even if you didn't realize

it.

For example, remember back to the very beginning of your physics course, and the topic of motion with constant acceleration. One of our early formulations of this was that the velocity $v(t) = dx(t)/dt$ should change uniformly, i.e., linearly, in time. Mathematically, that means

$$\frac{dx(t)}{dt} = v_0 + at \tag{6.1}$$

where v_0 (the "initial velocity") is the value of dx/dt at $t = 0$, and a is the rate at which the velocity, dx/dt, increases. The point is that this is an equation which tells us something about the *derivative* of some (unknown) function, $x(t)$ – i.e., it is a differential equation. It specifies that the derivative of the unknown function $x(t)$ changes in a certain way with time.

The *solution* of this DE (as one can check by explicit differentiation) is the familiar expression:

$$x(t) = x_0 + v_0 t + \frac{1}{2}at^2. \tag{6.2}$$

Although Equation 6.1 is indeed a differential equation, it is a sort of special case because here the derivative of the unknown function is specified in terms of the independent variable (t) alone. In a more typical DE, the derivative of the unknown function depends not just on the independent variable (i.e., the variable that the unknown function is a function *of*), but on the function itself, too.

Here's a familiar physics example of that type: recall that the force exerted by an ideal spring with spring constant k is given by the formula

$$F = -kx \tag{6.3}$$

where x is the displacement of the spring from its equilibrium length. (The minus sign reminds us that the spring exerts a *restoring* force – i.e., the force is always in the opposite direction of the displacement.) So if one has a particle of mass m attached to a spring of spring constant k (and no other forces act) then Newton's second law takes the form:

$$-kx = ma. \tag{6.4}$$

But the acceleration a and the position x are not unrelated quantities: a is the *second derivative* of x. And so the equation of motion can be written

$$\frac{d^2x(t)}{dt^2} = -\frac{k}{m}x(t) \tag{6.5}$$

which makes it clear that this is a differential equation. One should of course think of this as an equation which one is supposed to solve for $x(t)$. But one cannot simply solve for $x(t)$ algebraically, because the information the equation specifies about $x(t)$ is more abstract: it tells us only that $x(t)$ is a function whose second derivative is proportional to the function itself (the proportionality constant being $-k/m$).

You probably already realized (or remembered or can show by substitution and explicit differentiation) that the solution here looks like this:

$$x(t) = A\sin(\omega t + \phi) \tag{6.6}$$

where A and ϕ are constants (the amplitude and phase, respectively) and – in order that this actually *be* a solution to Equation 6.5 – the angular velocity must be given by $\omega = \sqrt{k/m}$. All of this of course makes sense, as one expects a mass on a spring to *oscillate*, and that is precisely what Equation 6.6 describes.

Let's introduce a bit of terminology that will help draw our attention to several important features of differential equations, on the basis of which they can be classified.

- The *order* of a DE is defined as the number of the highest derivative appearing in it. So our first example above – Equation 6.1 – is a *first order* DE, while the second example – Equation 6.5 – is a *second order* DE.

- A *linear* DE is one in which each term contains the unknown function (or one of its derivatives) to *the first power*. Thus, both of the example DEs mentioned above are linear. In equation 6.1, the unknown function $x(t)$ only appears once, in dx/dt, and that is to the first power. Similarly, in Equation 6.5, the two terms contain, respectively, the second derivative of the unknown function *to the first power*, and the unknown function itself *to the first power*. Here's a random example of a DE that is *non-linear*:

$$\frac{dx(t)}{dt} = \gamma \left[x(t)\right]^2 \tag{6.7}$$

  where γ is a constant. And here's another less obvious example of a non-linear DE:

$$\frac{dx(t)}{dt} = \gamma \cos\left[\alpha x(t)\right]. \tag{6.8}$$

  where γ and α are constants. The way to understand this last one is to think of expanding the cosine as its Taylor series: $\cos(z) = 1 - z^2/2 + z^4/24 - \cdots$. This makes it clear that terms involving $x(t)$ to higher-than-the-first-power are going to be present, even though they are sort of hidden inside the cosine function. But clearly pretty much any time the unknown function (or one of its derivatives!) appears in the argument of some other function in the differential equation, the DE in question will fail to be linear.

- A final distinction one sometimes encounters in the taxonomy of DEs is *homogenous* vs *non-homogenous* equations. These terms primarily apply to DEs that are linear – the issue is then whether *all* of the terms contain the unknown function (or its derivatives) to the first power (in which case the DE is homogenous) or whether there is also a term in which the unknown function doesn't appear at all (in which case the DE is in-homogenous). Thus, the first example given earlier – Equation 6.1 – is *in-homogenous* because there is a term (namely the whole right hand side) which doesn't contain $x(t)$ at all. By contrast, the second example – Equation 6.5 – is *homogenous*.

The point of introducing this terminology is mostly just to give you a better sense of some of the different kinds of equations one sometimes encounters. And having seen the terminology probably gives you a sense that some equations are going to be easier/harder

than others to solve. For example, it probably isn't shocking or incomprehensible to hear that, all other things being equal: first-order DEs are easier to solve than second-order DEs (and so on); linear DEs are easier to solve than non-linear DEs; and homogenous DEs are easier to solve than in-homogenous DEs.

Let's now turn (briefly) to the problem we eventually want to solve, just to get it on the table: $\vec{F} = m\vec{a}$ for a planet orbiting the sun. We might as well write this equation out more explicitly just so we can have it in mind as an example for what follows. Assuming we place the Sun (mass M) at the origin, and let the planet orbit in the $x - y$ plane, then, according to Newton's theory of gravitation, the force felt by an orbiting planet (mass m) will be

$$\vec{F} = -\frac{GMm}{r^3}\vec{r} \tag{6.9}$$

where G is Newton's constant and $\vec{r} = x\hat{i} + y\hat{j}$ is the position of the planet at a given moment. Of course, the whole point is that these are functions of time: $x = x(t)$ and $y = y(t)$. And since the force is going to be set equal to m times the planet's acceleration (which in turn can be written as the second derivative of the position) it's pretty clear that we're going to have a differential equation.

Actually we have *two*, one for each of the components of the position:

$$\frac{d^2x(t)}{dt^2} = -GM\frac{x}{(x^2+y^2)^{3/2}} \tag{6.10}$$

and

$$\frac{d^2y(t)}{dt^2} = -GM\frac{y}{(x^2+y^2)^{3/2}} \tag{6.11}$$

where we have used the fact that $r = |\vec{r}| = \sqrt{x^2+y^2}$. Note that each equation is *second-order* and *non-linear*! Note also that the two equations are *coupled* – meaning that the unknown function $y(t)$ appears in the equation for the (second derivative of) $x(t)$, and vice versa.

So it looks like these are going to be hard equations to solve! That is precisely why we're going to approach them here by using a computer to solve them numerically. As it turns out, though, these equations can be solved *exactly*. That is, by a sequence of clever changes-of-variables, analytical tricks, and magic incantations, one can actually generate explicit expressions for $x(t)$ and $y(t)$. And despite being way more *elegant* than using a computer to generate approximate, numerical solutions, doing it this way is *hard*. Students with adequate math backgrounds can work though this hard calculation in the Projects. But even if you are able to do that, you will still gain important insight and develop some really important skills by approaching it the way we're going to do in the rest of this chapter.

6.2 Euler's Method

In order to explain the basic approach to solving a DE numerically using a computer, we'll first spend some time discussing an algorithm called "Euler's Method" which turns

out to be really inaccurate for most problems. So you probably won't (and certainly shouldn't) use Euler's method if you ever actually care about solving a DE with some modicum of accuracy. Nevertheless, it is simple to understand, so I'll try to convey the basic idea using Euler's method and then show you some more sophisticated algorithms that are more useful in practice.

One more preliminary before we jump into Euler's method. Any higher-order DE can always be converted into a collection of several first-order DEs. For example, take the DE

$$\frac{d^2x(t)}{dt^2} = -\frac{k}{m}x(t) \tag{6.12}$$

that was mentioned above. This is second-order. But we can introduce an additional unknown function $v(t)$, which is just the first derivative of $x(t)$:

$$\frac{dx(t)}{dt} = v(t). \tag{6.13}$$

And so then the original equation can be re-written in terms of v this way:

$$\frac{dv(t)}{dt} = -\frac{k}{m}x(t). \tag{6.14}$$

The point is, the last two equations *together* are exactly equivalent to the original DE – Equation 6.12. We've replaced a single second-order DE with a pair of first-order DEs. This is good, because (as we'll see momentarily) first-order DEs are easy to solve numerically using a computer. Indeed, solving lots of them at the same time is really no harder, since the computer is doing all the work. By the way, note that the two first-order DEs are *coupled*: the equation specifying the derivative of v depends on x, and vice versa. But, again, as we'll see shortly, this doesn't introduce any serious trouble.

OK, so finally let's talk about Euler's method. As we've just discussed, there is no loss of generality in talking about a first-order DE, so we'll do that. We'll first discuss this in a somewhat abstract way, and then work through a particular concrete example in detail.

So: suppose we have a first-order DE for some unknown function $x(t)$:

$$\frac{dx(t)}{dt} = f[x,t] \tag{6.15}$$

where the f on the right hand side represents some arbitrary expression involving $x(t)$ and t. This will be known for a given problem. And suppose we are given an *initial condition* – i.e., the value of $x(t)$ at $t = 0$:

$$x(0) = x_0. \tag{6.16}$$

Here's the idea. Since we know $x(0)$, we can evaluate f (i.e., the right hand side of our DE) for $t = 0$. This tells us the rate at which x is changing, dx/dt, at $t = 0$. But from *that* we can estimate the value of x some short time Δt (think "a millionth of a second") later:

$$x(\Delta t) \approx x_0 + \Delta t\, f(x_0, 0). \tag{6.17}$$

You should pause here and make absolutely sure you fully understand why that last equation makes sense. If you miss this, you'll have missed the whole thing!

Once you understand why Equation 6.17 makes sense, we can raise an important question: why is it only *approximately* true? As you hopefully already realized, Equation 6.17 is based on the approximate equality of the instantaneous velocity dx/dt at $t = 0$, and the *average velocity* $\Delta x/\Delta t$ for a duration Δt beginning at $t = 0$. What's *exactly* true is this:

$$x(\Delta t) = x_0 + \Delta t\, v_{avg} \tag{6.18}$$

where v_{avg} is the average velocity:

$$v_{avg} = \frac{x(\Delta t) - x_0}{\Delta t}. \tag{6.19}$$

The point is that, according to Equation 6.15, $f(x_0, 0)$ is equal to the instantaneous velocity right at $t = 0$. This will not in general be equal to v_{avg}, but – the idea is – if Δt isn't too big, it should be *close*. And so the two sides of Equation 6.17 should be, while not exactly the same, *close*. (You should pause here and make sure that you can now *derive* Equation 6.17.)

Let's recap. From the given DE and the given initial value, we've figured out a "pretty good" approximation to the value of the unknown function at a slightly later time, $t = \Delta t$. Now the whole point is that we can do that same thing again and again and again. *Now* we know – or at least have a good approximation to – $x(\Delta t)$. So we can evaluate $f[x(\Delta t), \Delta t]$. This tells us the rate of change of $x(t)$ for $t = \Delta t$. And so, knowing both x and dx/dt at $t = \Delta t$, we can approximate the value of x at the "next" moment, $t = 2\Delta t$:

$$x(2\Delta t) \approx x(\Delta t) + \Delta t\, f[x(\Delta t), \Delta t]. \tag{6.20}$$

And so on.

In case the notation is confusing you, just think of it this way. We're going to get an approximate value for x at a bunch of different times: $t = 0$, $t = \Delta t$, $t = 2\Delta t$, etc. So let's just call these times t_0, t_1, t_2, etc., and the corresponding x values x_0, x_1, x_2, etc. Then the Euler's method algorithm can be written:

$$x_{n+1} = x_n + \Delta t\, f(x_n, t_n) \tag{6.21}$$

which amounts to: use the "current" value of x to figure out the "current" rate at which x is changing, and then use *that* to approximate the "next" value of x. And then keep going as long as you want.

That concludes the abstract overview. Now let's see how this works for a simple concrete example:

$$\frac{dx(t)}{dt} = x(t) \tag{6.22}$$

with, say, the initial condition $x_0 = 1$. This is a nice example because (a) it is simple and (b) we don't actually need any numerical algorithm to solve it. The solution will be a function whose derivative is equal to the original function (that's just what the DE says)

and which has the value 1 at $t = 0$ (that's what the initial condition says). But anybody who has taken calculus already knows that the solution is the exponential function

$$x(t) = e^t \tag{6.23}$$

whose derivative is

$$\frac{dx(t)}{dt} = e^t \tag{6.24}$$

which, as required, equals the function itself! The fact that we know the exact solution will help us see that and why the approximation produced by Euler's method is just that – an approximation.

Let's think about using Euler's method to fill in the following table, which initially contains just the initial value, $x(0) = 1$. We'll take the "step size" Δt to be 0.3. You might want to open Excel (or equivalent) and work through this yourself as you read.

t	x(t)	$\frac{dx(t)}{dt}$
0.0	1.0	
0.3		
0.6		
0.9		
1.2		

The first thing we can do is to use the DE, Equation 6.22, to fill in the $dx(t)/dt$ column in the first row. The DE tells us that $dx/dt = x$, so we fill the 1.0 we read off from the x column in the same row. That gives the following updated table.

t	x(t)	$\frac{dx(t)}{dt}$
0.0	1.0	1.0
0.3		
0.6		
0.9		
1.2		

Now we apply Euler's method, according to which the "next" value of $x(t)$ is approximated by the "previous" value (1.0) plus Δt (0.3) times the "previous" right hand side (1.0). That gives $x(0.3) = 1.3$, which we can fill into the table.

t	x(t)	$\frac{dx(t)}{dt}$
0.0	1.0	1.0
0.3	1.3	
0.6		
0.9		
1.2		

If you are doing this yourself, e.g., in Excel – and you should be! – you'll want to let the computer to the work. That is, don't try to figure out in your head what "the previous value plus Δt times the previous slope" equals. Just type it in as a *formula* which refers to the contents of the appropriate cells in the previous row.

Now we are in a position to just repeat the same steps over and over again. We can fill in the dx/dt column in the second row by again appealing to the fact that, according to Equation 6.22, dx/dt just equals x. So we put a 1.3 there. And then we can fill in the $x(t)$ cell in the $t = 0.6$ row: it's just the previous x value (1.3) plus Δt (0.3) times the previous dx/dt value (1.3). This works out to $x(0.6) = 1.69$. We'll put this in the table, along with the next few values that fill it out.

t	x(t)	$\frac{dx(t)}{dt}$
0.0	1.0	1.0
0.3	1.3	1.3
0.6	1.69	1.69
0.9	2.197	2.197
1.2	2.8561	2.8561

And of course we could keep going if we wanted to. The point is, by following this method, we're able to generate a list of the values of the function $x(t)$ at a bunch of equally spaced times, which is just what we want.

It is illuminating to make a graph of the $x(t)$ values we just calculated, and compare them to the exact solution, $x(t) = e^t$. The two are shown in Figure 6.1.

The first point, of course, lies exactly on the curve. But the next point misses by a little bit – then the next misses by a little bit *more* – and so on. Let's try to understand graphically why the second point is a little bit off, and why the error grows with each subsequent step.

Figure 6.2 shows a blow-up of the first two points and the exact curve. Also shown is a line connecting the two points on the exact curve which correspond to $t = 0$ and $t = 0.3$. The slope of this line is, of course, just the *average* slope of the exact curve in this interval. And the point is that this is close to – but not exactly equal to – the slope of the exact curve *at* $t = 0$ which, by construction, matches the slope of the line that

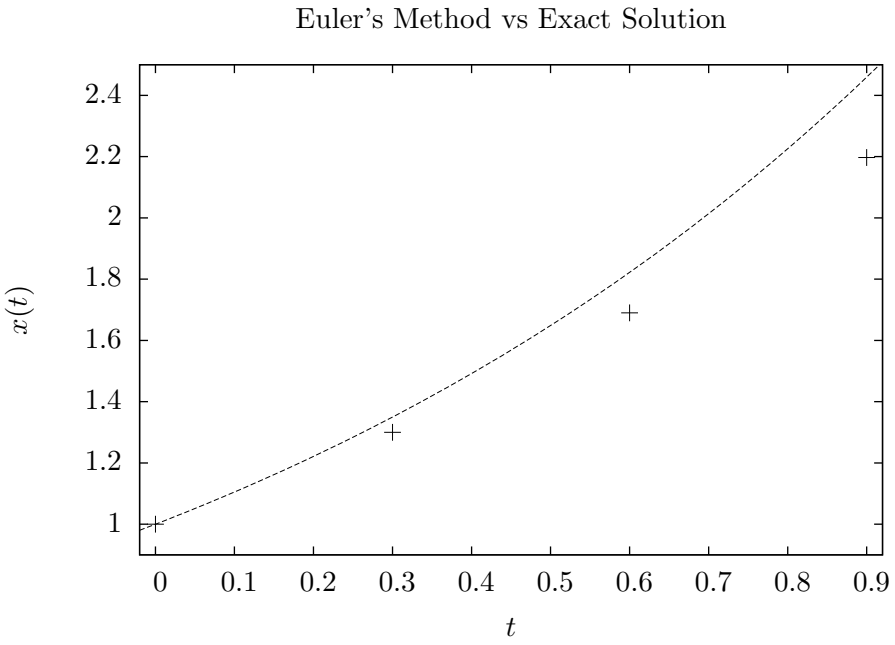

Figure 6.1: Comparison between the first few points generated by the Euler's method solution to $dx(t)/dt = x(t)$ with $x(0) = 1$, and the exact solution $x(t) = e^t$. (The exact solution is the smooth curve, and the Euler's method points are the crosses.)

connects the first two Euler method points. Remember, the whole *content* of Euler's method is to approximate the average slope over some interval, with the slope at the beginning of the interval. As discussed already, we can minimize the error by using a smaller Δt; but – for any nonzero Δt – there will be *some* error.

Think too about the way the error accumulates. The story over the second interval (i.e., between $t = 0.3$ and $t = 0.6$) is much the same as the story over the first interval: the *exact* increase of the function over this interval is just the width of the interval multiplied by the *average slope* over this interval; but (just as before) the average slope isn't known, so we have to use, as a proxy, the instantaneous slope at the beginning of the interval. But here's a further problem with this second interval which we didn't encounter with the first interval: we don't actually *know* the instantaneous slope at the beginning of the interval, if that means the instantaneous slope of the exact solution at $t = 0.3$. Rather, what we know is the slope that that curve would have, if $x(0.3)$ were the value we got from the first Euler step (1.3) – which, as we've already discussed, it isn't! So instead of even using the (already wrong) slope at the beginning of the interval to proxy for the average slope across the interval, we actually end up using the slope of some point a little to the *left* of the beginning of the interval – the point where the exact curve has the value 1.3 – to proxy for the average.

The point is that there is a double sense in which the error is cumulative as we

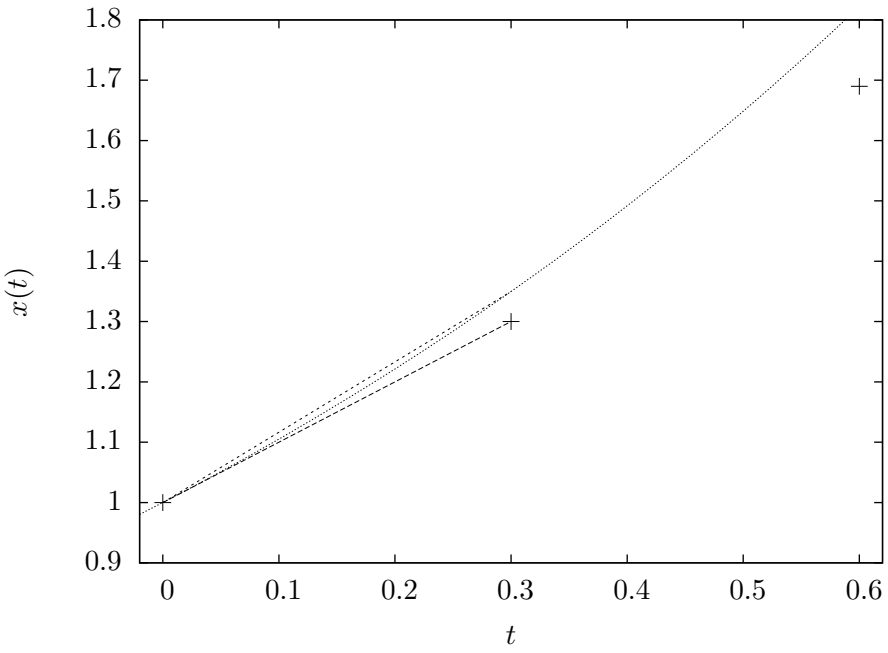

Figure 6.2: The first step is not quite right because it is based on the (not quite true) assumption that the average slope over the interval is equal to the slope at the beginning of the interval. Both slopes are shown.

proceed to higher and higher times. First, each additional step contributes some of the same sort of error we had during the first step, the error associated with equating the average slope across the interval with the instantaneous slope at the beginning of the interval. And second, because our best approximation to dx/dt at a given point depends on the value of x there – and because (for all but the initial points) we don't know the exact value of x – the slope used in Euler's method doesn't even match the real slope at the beginning of the interval.

We happen to have chosen an "extreme" example in which these two sorts of error leapfrog off of one another and cause the Euler method approximation to diverge badly from the exact solution. The point is, things aren't always quite as bad as we're seeing here. On the other hand, it's good to know how bad things can be when you are trying to decide whether or not to believe the results of a computation – the point being, of course, that in actual use we won't have "the exact solution" in hand to compare to!

6.3 Better Algorithms

It is traditional to summarize Euler's method by writing

$$x_{n+1} = x_n + k1 \tag{6.25}$$

where

$$k1 = \Delta t\, f(x_n, t_n) \tag{6.26}$$

is the approximation to the change in the function x between the "current" and "next" times that you get by following the line whose slope at the "current" time matches the slope there. As discussed before, this provides a pretty good approximation to x_{n+1} so long as Δt is small. But it isn't perfect, and it's clear that it's going to be off *systematically* since it equates the average slope over some finite duration Δt with the instantaneous slope at the *beginning* of that duration.

A little thought suggests a better approach. Instead of approximating the average slope over Δt by the instanteous slope at the beginning of that interval, we could approximate the average slope by the instantaneous slope at the *middle* of the interval. This is likely to be a much better approximation. The problem is, we don't know dx/dt at the middle of the interval – we only know it at the beginning! But what we can do is *use Euler's method* to take a "trial step" over to the midpoint of the interval. This will give us at least an *approximate* value for x there. And then we can plug this approximate value back into f to get the slope at the midpoint. It's plausible (though not completely obvious) that this will result in a more accurate algorithm overall, and indeed it does. It's called the *midpoint method* and can be summarized like this:

$$x_{n+1} = x_n + k2 \tag{6.27}$$

where

$$k1 = \Delta t\, f(x_n, t_n) \tag{6.28}$$

$$k2 = \Delta t\, f(x_n + k1/2, t_n + \Delta t/2). \tag{6.29}$$

Don't let this one be a total black box. Pause and follow the logic of this: $k1$ is just the Euler's method approximation for $x_{n+1} - x_n$. So $x_n + k1/2$ will be a naive guess for the value of x at the midpoint of the time interval. And so $f(x_n + k1/2, t_n + \Delta t/2)$ – the slope at the midpoint of the interval *as approximated by Euler's method* – will be the more accurate guess for the average slope of $x(t)$ across the whole interval. And so this times Δt should be a very good approximation to the amount by which x should change during Δt.

Let's walk through how to use the midpoint method to solve the same toy DE we used above

$$\frac{dx(t)}{dt} = x(t) \tag{6.30}$$

using the same initial condition $x(0) = 1$ and stepsize $\Delta t = 0.3$ we used before.

To begin with, let's set up a table which contains the initial condition and columns for calculating the quantities $k1$ and $k2$ introduced above.

t	x(t)	k1	k2
0.0	1.0		
0.3			
0.6			
0.9			
1.2			

The first thing we can do is calculate the value of $k1$ in the first row. This is just Δt (0.3) times "f" evaluated at $t = 0$. Here f is just x, so we have $k1 = 0.3$. Then $k2$ is Δt times f evaluated at $x(0) + k1/2$. Here that works out to be $k2 = 0.3(1 + .15) = .345$. Note that if you are following along with Excel – and again you should be! – you shouldn't be typing "0.3" (etc.) but should instead be writing a *formula* which refers to (e.g.) a cell where you declare a value for Δt. That way, once you get it set up, if you want to *change* Δt – for example, to find out how the accuracy of the algorithm scales with the stepsize – then you only have to change one thing. Anyway, at this point your table should look like this

t	x(t)	k1	k2
0.0	1.0	.3	.345
0.3			
0.6			
0.9			
1.2			

where, just to be clear, the value for $k1$ is computed according to a formula which refers to the x value (from that same line) and the Δt value (from somewhere else in the sheet). Then the $k2$ value is computed according to a formula which refers to x, Δt, and the already-computed $k1$.

That was the hard part. Now we can compute the $x(t)$ value in the $t = 0.3$ row. It's $x(0.3) = x(0.0) + k2$, where, of course, here $k2$ means $k2(0.0)$, i.e., the $k2$ value from the previous row of the table. That gives $x(0.3) = 1.345$. Then we can compute the new $k1$ and $k2$ values (by just copying and pasting the formulas we already typed in just above), and... wash, rinse, repeat. The table ends up looking like this:

t	x(t)	k1	k2
0.0	1.0	.3	.345
0.3	1.345	.4035	.464025
0.6	1.809025	.542708	.624114
0.9	2.433139	.729942	.839433
1.2	3.272571	.981771	1.12904

Figure 6.3 shows a graph of the midpoint method's results here, compared both to the exact solution ($x(t) = e^t$) and the Euler's method approximation using the same stepsize. It is immediately clear that (all other things, in particular the step-size, being equal) the midpoint method is more accurate than Euler's method (but still not perfect).

Of course, one might object that really the midpoint method cheats. What we *call* the "midpoint method" with a stepsize $\Delta t = 0.3$ actually involves taking these trial steps of size $\Delta t/2 = 0.15$. So maybe the only reason the midpoint method is more accurate, is because we are really in effect using half the reported stepsize. This is a good objection to think about carefully – in the Questions and Projects!

In any case, the immediate upshot is clear: with only a marginal increase in the difficulty of setting things up, the midpoint method provides a more accurate solution of a given DE than does Euler's method.

And it turns out you can keep going this way. For example: instead of just taking one trial step to get a better approximation for the change in x over the time interval Δt, one can take two trial steps – with the first trial step being used to then take a more accurate second trial step, which then one uses finally to calculate the change in x. And so on. A robust and widely used method along these lines (which involves taking a kind of weighted average of four different trial steps) is called the *4th Order Runge-Kutta* algorithm. It is defined this way:

$$x_{n+1} = x_n + \left(\frac{k1}{6} + \frac{k2}{3} + \frac{k3}{3} + \frac{k4}{6} \right) \tag{6.31}$$

where

$$k1 = \Delta t\, f(x_n, t_n) \tag{6.32}$$
$$k2 = \Delta t\, f(x_n + k1/2, t_n + \Delta t/2) \tag{6.33}$$
$$k3 = \Delta t\, f(x_n + k2/2, t_n + \Delta t/2) \tag{6.34}$$
$$k4 = \Delta t\, f(x_n + k3, t_n + \Delta t) \tag{6.35}$$

It's probably OK to just accept this as a black-box, although it's easy enough to get the gist of it: $k1$ is just the Euler's method approximation to Δx; $k2$ is then the better guess based on the midpoint recalculation; $k3$ is then an *even better* approximation to Δx based on *using* $k2$, in the same way that $k1$ is used to recalculate Δx when computing $k2$; and finally $k4$ uses $k3$ to extrapolate over to the *end* of the time interval Δt and represents the Δx you'd get by using the slope *there* as a stand-in for the average across the whole

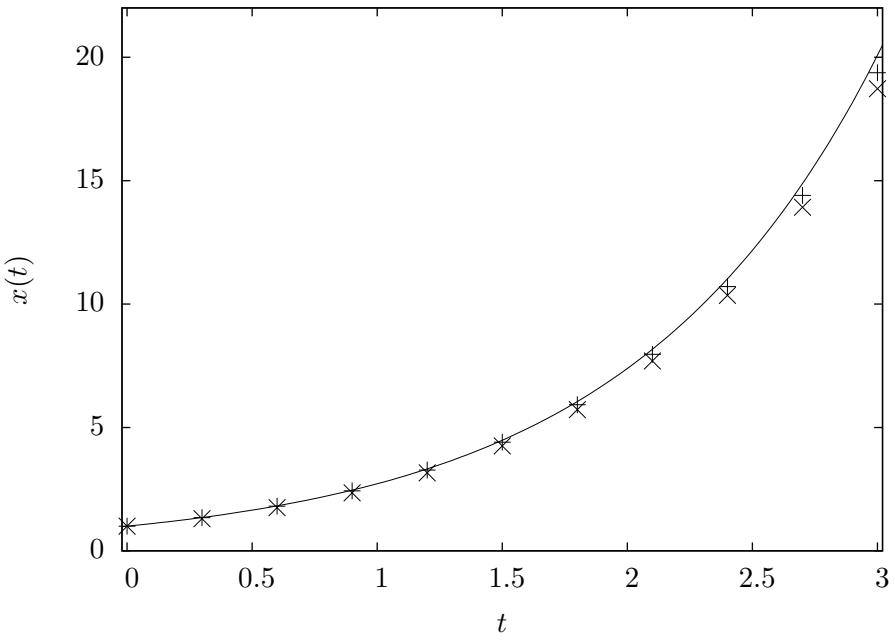

Figure 6.3: Midpoint method (the crosses) and Euler's method (the Xs) with the same stepsize. The curve is the exact solution. Clearly the midpoint method is more accurate.

Δt. And then some cleverly-chosen weighted average of all these approximations turns out, quite plausibly, to be a *very very* good approximation to the real Δx.

To solve a real problem (like you'll do for the planet orbiting the sun) it's definitely worth investing some extra time at the beginning to use one of the more complicated but more accurate algorithms. It takes more time to set up, but once it's set up, you just copy and paste and the computer does all the really hard work.

Again at this point it would be a good idea to sit down in front of a computer and use Excel or some equivalent to play with the same simple toy example

$$\frac{dx(t)}{dt} = x \tag{6.36}$$

using the more complicated but more accurate algorithms.

We reproduce here, in the following table, the first several values one gets using the 4th order Runge-Kutta algorithm, as well as the various intermediate numbers ($k1$ through $k4$) that are involved. Use this as a check to make sure that your own calculations are giving the right numbers. That is, you should stop reading right now, sit down in front of a computer, open Excel, and make sure you understand how to reproduce the numbers in this table!

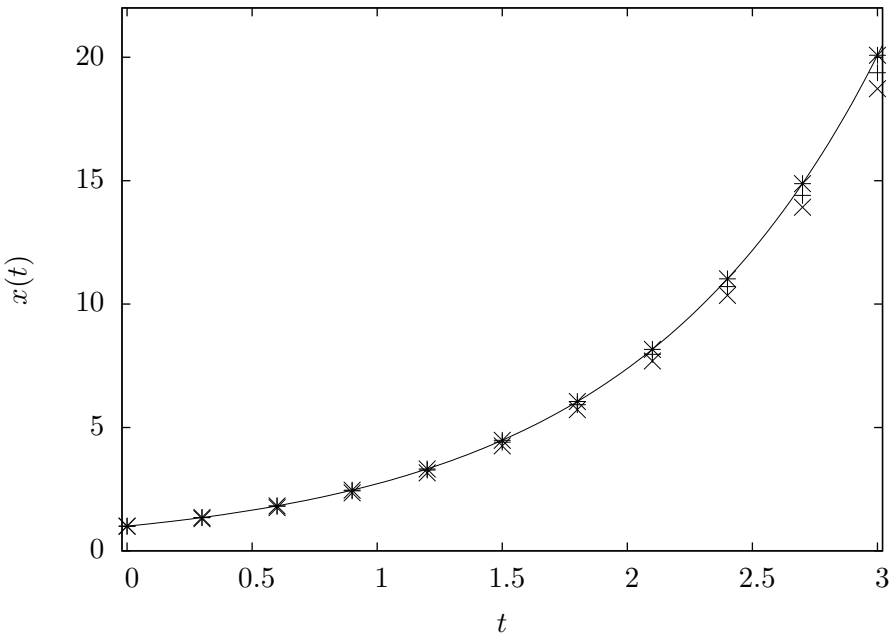

Figure 6.4: 4th order Runge-Kutta algorithm, vs Midpoint and Euler.

t	x(t)	k1	k2	k3	k4
0.0	1.0	0.3	0.345	0.35175	0.405525
0.3	1.349838	0.404951	0.465694	0.474805	0.547393
0.6	1.822061	0.546618	0.628611	0.64091	0.738891

The numbers in the t column are of course just given, as is the first entry in the $x(t)$ column. The four k columns in the $t = 0$ row are then filled in in order: $k1$ at $t = 0$ is just the step size (0.3) times the $x(t)$ value in that row. Then $k2$ is the step size (0.3) times $x(t) + k1/2$. Then $k3$ is the step size (0.3) times $x(t) + k2/2$. And finally $k4$ is the step size (0.3) times $x(t) + k3$. Finally, with all four of the k's computed, the value of $x(t)$ at the "next" time – $t = 0.3$ – is computed: $x(0.3) = x(0)+k1(0)/6+k2(0)/3+k3(0)/3+k4(0)/6$. The result happens to be $x(0.3) = 1.349838$. And with that new value of x computed, the cycle begins again – computing the four k's across the $t = 0.3$ row, on the basis of which $x(0.6)$ can be computed, and so on.

Figure 6.4 shows the results of this computation for times up to three seconds, comparing the 4th order R-K results with the midpoint and Euler methods (shown previously) and the exact formula. As far as one can tell from the graph, the 4th order R-K algorithm reproduces the exact formula exactly. As the logirithmic graph of the errors shown in Figure 6.5 shows, it is not perfect. But it is *substantially* better than the other

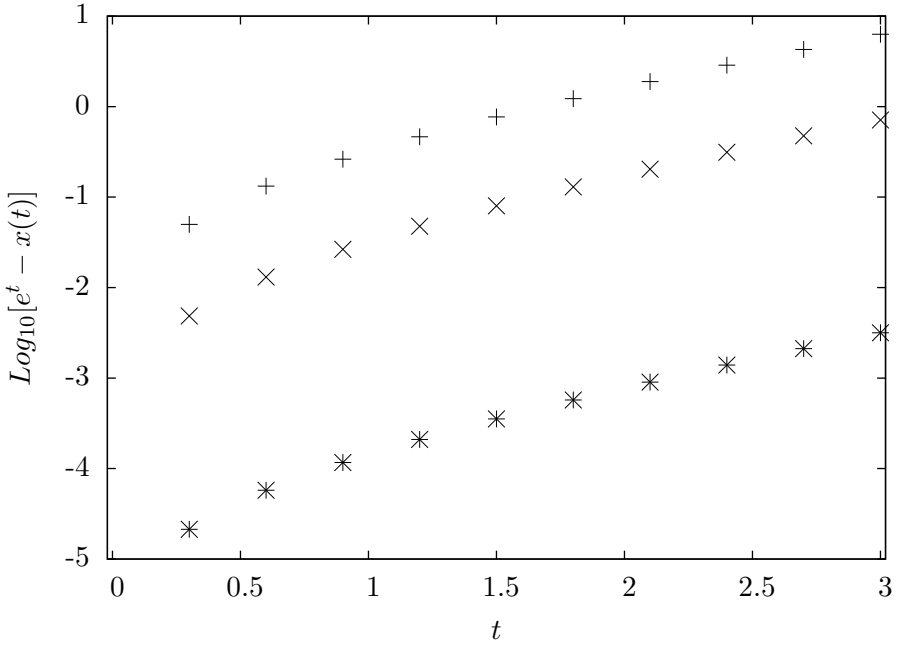

Figure 6.5: The (base 10) logarithm of the error (i.e., the exact value minus the value given by the computational algorithm) for each step. Results for the Euler method (crosses), the Midpoint method (X's), and the 4th order R-K method (stars) are all shown. Notice that, for the reasons discussed earlier, the total error (for a given method) increases with time. It doesn't typically increase with time quite as fast as it does in this particular problem, but it's good to see how quickly the results *can* become completely unreliable. So one has to be increasingly careful the longer one runs a simulation. Note also that, at a given moment, the Midpoint method is about one order of magnitude more accurate than the Euler method, and that the R-K method is about three orders of magnitude more accurate than Euler. (This last statement depends on the stepsize we've chosen, so don't take it as any kind of dogmatic, out-of-context statement about the accuracies.)

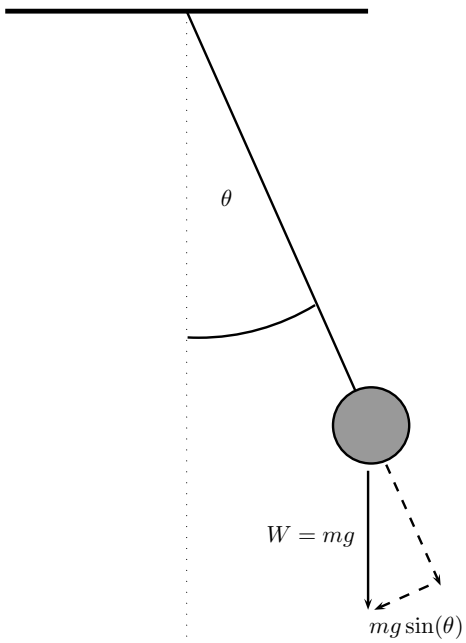

Figure 6.6: A simple pendulum constructed from a mass m and a (massless) string of length L. When displaced at angle θ from the vertical, the tangential component of the weight force produces a restoring torque (about the point where the string attaches to the ceiling) of magnitude $\tau = mgL\sin(\theta)$.

two methods.

6.4 A more involved example: the pendulum

Now that we understand the basic idea of using a computer to numerically solve a differential equation, let's work through a slightly less artificial example: the physical motion of a simple pendulum.

Consider a pendulum made of a point mass m hanging from a massless string of fixed length L. See Figure 6.6. When the pendulum is displaced to angle θ, there is a net torque (about the point where the string attaches to the ceiling). This is produced by the tangential component of the gravitational force acting on the mass. The torque is:

$$\tau = -mgL\sin(\theta) \tag{6.37}$$

where the minus sign indicates that the direction of the torque (e.g., clockwise in the figure) is opposite the (e.g., counter-clockwise) sense of the angular displacement itself.

Now we can use the results from Chapter 4 to write down the differential equation governing the pendulum. The moment of inertia of the pendulum about the pivot point is just $I = mL^2$, independent of the angle θ. The basic law of rotational dynamics is

that the net torque should equal the time rate of change of the angular momentum. Here that comes down to

$$\tau = \frac{dL}{dt} = I\alpha \tag{6.38}$$

since the angular momentum $L = I\omega$ and I is independent of time.

Plugging in our expression for the torque and remembering that $\alpha = d\omega/dt = d^2\theta/dt^2$, we have:

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\sin(\theta) \tag{6.39}$$

which is a second-order, *non-linear* differential equation for $\theta(t)$.

Recall that, if the pendulum swings with small amplitude, then $\sin(\theta) \approx \theta$, and the exact non-linear DE turns into the following linear equation

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\theta \tag{6.40}$$

which has exact solution

$$\theta(t) = A\,\sin(\omega t + \phi) \tag{6.41}$$

with $\omega = \sqrt{g/L}$. This corresponds to a back-and-forth motion with period $T = 2\pi/\omega = 2\pi\sqrt{L/g}$. This is a well-known formula for the period of a pendulum; the interesting thing here is to see not only how it is derived, but also that the derivation applies only for small amplitudes. For larger amplitudes, the period will deviate significantly from the above formula.

Unfortunately, there's no simple explicit way to write down exactly how the period varies with amplitude. That's essentially because no simple explicit solution to the full non-linear differential equation exists. So this is an excellent candidate for solving numerically using a computer.

Since it is a second-order DE, let's first convert it to a pair of first-order DEs (each of which can be solved using any of the methods discussed previously). Thus we introduce the angular velocity ω and write:

$$\frac{d\theta}{dt} = \omega \tag{6.42}$$

and

$$\frac{d\omega}{dt} = -\frac{g}{L}\sin(\theta). \tag{6.43}$$

And let's take as initial conditions $\theta(0) = \theta_0$, $\omega(0) = 0$ – i.e., the pendulum is pulled over by angle θ_0 and let go.

Let's get really concrete and sketch how we could work through this computation in Excel or equivalent. Suppose the length of the pendulum is $L = 1$ meter, suppose $\theta_0 = 1$ radian, and for convenience we'll take $g = 10m/s^2$. A pendulum of this length moving with small amplitude would have a period of about two seconds. Our amplitude, one radian, isn't exactly small. But still we expect a period of not more than a few seconds. So let's use a step size of $\Delta t = 0.05$ seconds, at least at first. (A smaller step size would of course produce more accurate results, but this is an OK starting point.)

We'll start by using Euler's method, just to get clear on the approach. Set up a table with a column for the time t, one for the variable θ, one for the variable ω, and then *two more columns* – one for the "$k1$" associated with the variable θ, and another for the "$k1$" associated with the variable ω. It looks like this:

t	$\theta(t)$	$\omega(t)$	$k1_\theta$	$k1_\omega$
0.00	1.0	0.0		
0.05				
0.10				

Now we can start filling in the table. Note that each of the $k1$'s – $k1_\theta$ and $k1_\omega$ – is a function of, in principle, *both* of the dependent variables: θ and ω. Though in our case here, $k1_\theta$ depends only on ω, and vice versa. In particular: $k1_\theta$ is given by

$$k1_\theta = \Delta t\, f_\theta(\theta, \omega, t) \tag{6.44}$$

where the function f_θ is just the right hand side of the first-order differential equation for θ, Equation 6.42. Thus, concretely, we have

$$k1_\theta = \Delta t\, \omega \tag{6.45}$$

and similarly for the $k1$ associated with the variable ω:

$$k1_\omega = \Delta t\, f_\omega(\theta, \omega, t) \tag{6.46}$$

where f_ω is the right hand side of Equation 6.43. Thus,

$$k1_\omega = -\Delta t\, \frac{g}{L} \sin(\theta). \tag{6.47}$$

Equations 6.45 and 6.47 are what we need to continue filling in the table.

The two $k1$'s in the $t = 0$ row can be computed on the basis of the θ and ω values in that row (and, of course, Δt). Then, "new" values for θ and ω can be computed: $\theta(0.05) = \theta(0.00) + k1_\theta(0.00)$ and $\omega(0.05) = \omega(0.00) + k1_\omega(0.00)$. And so on. The first few rows of the table look like this:

t	$\theta(t)$	$\omega(t)$	$k1_\theta$	$k1_\omega$
0.0	1.0	0.0	0.0	-0.42074
0.1	1.0	-0.42074	-0.02104	-0.42074
0.2	0.97896	-0.84147	-0.04207	-0.41496

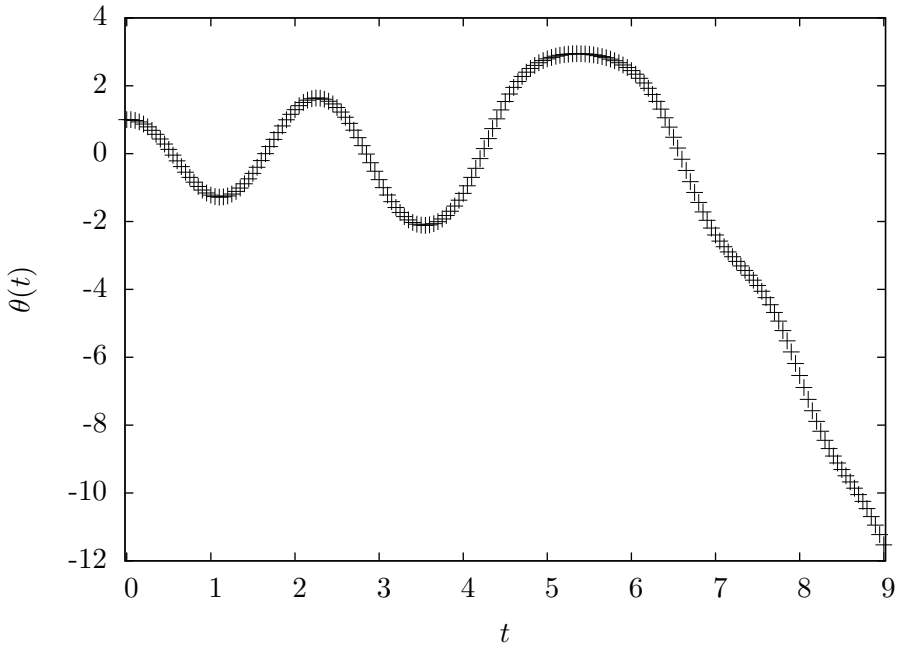

Figure 6.7: Numerical solution of the pendulum equation using Euler's method and a step size of $\Delta t = 0.05$ seconds. The qualitative behavior is roughly correct, at least for a while, but there are some signs that the solution is not very accurate: the amplitude increases with time (which is unphysical since it implies that the total energy is not conserved), and after just a few oscillations, the solution goes completely crazy. This suggests using a smaller step size and/or one of the more accurate algorithms.

Once you get this set up, it's trivial to have the computer compute as many rows as you want. Figure 6.7 shows a graph of the resulting $\theta(t)$ vs t for the first ten seconds. We clearly are reproducing the qualitatively correct oscillatory motion of the pendulum – at least at first. But after only a couple of swings, the motion is obviously unphysical. For one thing, the amplitude is increasing with time. (We know this is unphysical, for it implies a failure of energy conservation. Indeed, you can see already from the numbers in the previous table that energy isn't being conserved exactly. In the first row, the pendulum is at $\theta = 0$ at rest, i.e., with $\omega = 0$. In the second row, the pendulum is *still* at $\theta = 0$, and hence has the same *potential* energy, but now it's *moving* with $\omega = -0.42074$ and so has some *kinetic* energy, *too*!) And something completely crazy starts happening between about 6 and 10 seconds. All of these issues are of course a result of the inherent numerical inaccuracy of this method. We could improve things by sticking with the Euler method and using a smaller stepsize, or by sticking with $\Delta t = 0.05$ seconds and using one of the more accurate algorithms – or both.

You are encouraged to work through this same problem (the pendulum) using the midpoint or 4th order Runge-Kutta algorithm. See the Projects.

And then it is not too much harder to finally tackle the problem that motivated all of this: solving for the motion of a planet in the inverse-square-law gravitational force field produced by the Sun. Remember that the defining equations here are Equations 6.10 and 6.11, which we may rewrite as a set of *four* coupled *first-order* equations by introducing the x- and y-components of the velocity:

$$\frac{dx}{dt} = v_x \tag{6.48}$$

$$\frac{dv_x}{dt} = -GM\frac{x}{(x^2+y^2)^{3/2}} \tag{6.49}$$

$$\frac{dy}{dt} = v_y \tag{6.50}$$

$$\frac{dv_y}{dt} = -GM\frac{y}{(x^2+y^2)^{3/2}}. \tag{6.51}$$

We then have four "f" functions

$$f_x(x, v_x, y, v_y) = v_x \tag{6.52}$$

$$f_{v_x}(x, v_x, y, v_y) = -GMx/(x^2+y^2)^{3/2} \tag{6.53}$$

$$f_y(x, v_x, y, v_y) = v_y \tag{6.54}$$

$$f_{v_y}(x, v_x, y, v_y) = -GMy/(x^2+y^2)^{3/2} \tag{6.55}$$

which allow us, for example, to calculate the four associated $k1$'s in order to calculate the trajectory using Euler's method.

Some additional Projects will step you through this, and then a number of interesting and illuminating applications.

Questions for Thought and Discussion:

1. In what sense is, say, the midpoint method better than Euler's method? Sure, it's more accurate for a given stepsize, but you have to (or, really, the computer has to) do twice as much work per step. So why not just stick with Euler's method and reduce the step size by a factor of two, instead of switching to the more complicated algorithm? Indeed, aren't the midpoint method (with a given step size) and Euler's method (with half that step size) equally accurate? If you're not sure how to answer this, do Project 6.1 and then come back!

Projects:

6.1 Set up an Excel worksheet to solve the toy DE discussed in the text ($dx/dt = x$, $x(0) = 1$) using both the Euler and midpoint methods. Calculate $x(3)$ using the Euler method with a stepsize $\Delta t = 0.3$, and subtract this from e^3 to get the total error. Now calculate the same thing using some different stepsizes (say, $\Delta t = .1$, .03, .01, .003, and .001), and make a graph showing the total error vs. the stepsize.

You should notice that, for the pretty small stepsizes, the total error is roughly proportional to the stepsize – so that, for example, to increase the accuracy of the final result by a factor of 2, you need to reduce the stepsize by a factor of 2. Now do all these calculations again with the midpoint method. If, using the midpoint method, you reduce the stepsize by a factor of 2, by what factor is the total error reduced?

6.2 Implement the 4th order Runge-Kutta algorithm for the toy example ($dx/dt = x$, $x(0) = 1$). For reasonably small-ish stepsizes, how does the total error scale with the stepsize? That is, if you reduce the stepsize by a factor of 2, by what factor is the total error reduced?

6.3 Finish up the pendulum example we began in the text by implementing the 4th order Runge-Kutta algorithm. Note that there will be *eight* k's – a $k1_\theta$, a $k1_\omega$, a $k2_\theta$, ..., and a $k4_\omega$. Use the parameters mentioned in the text, and a step-size of $\Delta t = 0.1$ seconds. By what factor is the period of the pendulum (moving with amplitude one radian) greater than the period of the same pendulum at small amplitude?

6.4 Explore the accuracy of your R-K pendulum program. Think of some way to quantify the accuracy – e.g., calculate the total energy at each step, or monitor the maximum θ the pendulum gets to on each cycle, or something else clever that you think of. Now systematically vary the step size Δt and explore how the accuracy changes. Make an appropriate graph to display your findings.

6.5 Using an algorithm and a step-size whose accuracy you trust, find the period of a pendulum for the following amplitudes: 0.01 radians, 0.1 radians, 0.3 radians, 0.5 radians, 1 radian, 1.5 radians, 2.0 radians, 2.5 radians, and 3.0 radians. Make a nice graph to display the results. (Note that you can physically interpret the meaning of amplitudes greater than $\pi/2$ by imagining that the string is replaced with a massless rigid rod.)

6.6 Write a program to solve for the trajectory of a planet in the gravitational field of the sun, using Euler's method. Hint: use one year as your time unit, and one AU as your distance unit. Then, from the hopefully familiar formula

$$\frac{4\pi^2 R}{T^2} = \frac{GM}{R^2} \tag{6.56}$$

(where M is the mass of the Sun) you can infer that

$$GM = 4\pi^2 \mathrm{AU}^3/\mathrm{year}^2 \tag{6.57}$$

which comes in handy when you have to type in a formula for the gravitational force – i.e., the right hand sides of Equations 6.10 and 6.11.

Note that, if you approximate the Earth's orbit as a circle, its orbital speed is $v = 2\pi$ AU/year. So, to test your program, maybe start with the following initial

conditions: $x(0) = 1$, $y(0) = 0$, $v_x(0) = 0$, $v_y(0) = 2\pi$. If everything is set up correctly, this should launch your planet on a roughly circular orbit with a radius of about 1 (meaning 1 AU) and a period of about 1 (meaning 1 year). Use a stepsize of about .01. Of course, you won't get an exact circle, because Euler's method isn't all that accurate. But you should at least get something recognizably orbit-ish. (Make a scatterplot of the x and y coordinates to see the trajectory. What happens over the course of several years?)

6.7 Soup up your Euler's method program by adding columns for the $k2$'s and thus implement the midpoint method (or both Euler and midpoint simultaneously). Test the program to make sure it's working correctly.

6.8 Soup up your earlier program even further by adding columns as appropriate to implement the 4th order R-K algorithm. (Note that there will be 16 total columns of k's!) Test it.

6.9 Assuming now you've got a worksheet that can calculate the trajectory of the planet according to all three methods, compare them using again the Earth-like circular orbit. With the right initial conditions, the period of the orbit should be exactly one year. But of course none of these algorithms are exact, so you should find that at $t = 1$ the planet is not quite back to its starting point.

The point here is to see how these inaccuracies scale with the stepsize. So, for example, start by finding the position error after one year if you use $N = 100$ steps, i.e., a step size of 0.01 year. (You'll have to think about how best to characterize "the positional error".) Then use $N = 200$ (i.e., $\Delta t = .005$) and so on for several additional values. Make a plot (a log-log plot is particularly good here) of the error vs N for each of the three methods. Its slope (which characterizes the quality of the algorithm) should be different for each of the three algorithms.

6.10 Once you have something whose accuracy you trust, use your program to make some generic (moderately eccentric) orbit for a planet. Prove that the resulting points in fact lie on an ellipse with the Sun at one focus. (You'll have to go back to Chapter 2 and think about which mathematical definition of an ellipse is the best one to use to make this test.)

6.11 It is easy to show that, for a circular orbit, the total energy E (the kinetic energy plus the potential energy of the planet) depends on the radius R of the orbit as follows:

$$E = -\frac{1}{2}\frac{GMm}{R} \tag{6.58}$$

where M is the mass of the Sun and m is the mass of the planet. Show that this same formula is true for the more general elliptical orbits, if R is interpreted as the semi-major axis. That is, show that a bunch of different orbits (e.g., a circle, and then several ellipses with differing eccentricities) which all have the same R, also have the same total energy E. Since you can easily control the energy E by setting initial conditions (e.g., by doing a little algebra to work out what initial velocity

needs to be for a given initial radius, in order for the total energy to be some fixed value), it is probably easiest to do this by creating several different orbits that you know have the same energy, and then verifying that they also have the same R.

6.12 Similar to the previous one, but with the formula for how the period T of the orbit depends on the radius R:

$$T^2 = \frac{4\pi^2}{GM} R^3. \tag{6.59}$$

Show that this is true (in general, for elliptical orbits, with R the semi-major axis) by producing a number of orbits with the same R and showing that they have the same period T.

6.13 Write a program that has, as "sources", both the Sun and another planet, which we will think of here as Jupiter. Put in a fixed circular orbit for Jupiter at the right radius and speed. Then calculate the force on the earth at each moment from the combined forces exerted by the (stationary) sun and (moving) Jupiter. What is the effect of Jupiter on the orbit of the earth? (You won't see any, unless you track the trajectory for many years! But you can, for example, make Jupiter *heavier* and/or *closer to the Earth* and see what happens, in order to get a qualitative sense of the effect. You should be able to connect this to one of the astronomical effects mentioned way back in Chapter 2, but not yet explained!

6.14 Use the code from Project 6.13 to find the trajectory of a moon orbiting a planet ("Jupiter") as the planet orbits the Sun.

6.15 Play around some more with the program that models the combined gravitational effects (on Earth) of the Sun and (something like) Jupiter. Make Jupiter pretty heavy and pretty close to Earth, and give Earth's orbit enough eccentricity that it reaches almost as far as Jupiter. What can happen? What are the implications? Is there some reason other than just tidal forces which causes planets to tend to have circular orbits?

6.16 Newton proved in the Principia that the inverse square force law is the *only* one that gives elliptical orbits with the Sun at a focus. Try to produce some supporting evidence for this by changing the force law and seeing what happens to the resulting trajectories. Try, for example, an inverse *cube* law. What shape would the planet's trajectories have if they were connected to the Sun by springs (instead of gravitational forces)? What if the gravitational force varied as $1/r^{2+\epsilon}$ where ϵ is some small number like .01?

6.16 Work through the polar-coordinates calculus to prove (without resorting to numerical solutions) that the inverse-square force field produces elliptical orbits, with the general energy and period equations discussed in the earlier projects.