Astronomers' Universe

Other titles in this series

Origins: How the Planets, Stars, Galaxies, and the Universe Began (forthcoming)
Steve Eales

The Future of the Universe (forthcoming)
A.J. Meadows

Frank Levin

Calibrating the Cosmos

How Cosmology Explains Our Big Bang Universe

Springer

Frank S. Levin, PhD
Professor Emeritus of Physics
Brown University
Providence, RI 02912
USA

Jacket illustration: WMAP image of the anisotropies in the cosmic microwave background radiation, shown on an oval projection that represents the full sky. Courtesy of NASA and the WMAP science team.

Library of Congress Control Number: 2005937514

ISBN-10: 0-387-30778-8
ISBN-13: 978-0-387-30778-7

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

Calibrating the Cosmos is based on lectures I gave in several adult-education courses. By dealing with technical details in a descriptive way, I structured the courses for people who wanted to gain some knowledge of the physical universe as it is currently understood but had neither a science nor a mathematics background. The lectures were hard science, softly presented. As with the course, so with the book: it is written for persons whose curiosity about the physical universe extends to a readiness to learn some of the relevant technical aspects, presented descriptively.

Although many primary astronomical and cosmological experiments are identified, my emphasis is on the assumptions and theoretical concepts that underlie the measurements and the efforts to interpret and understand the resulting data. Taken together, observational information and theoretical ideas about the Universe form an elegant intellectual tapestry. I have treated some of its threads only cursorily; for instance, the history of astronomy, white dwarf stars, black holes, and the theory of inflation. Some experiments on the cosmic microwave background radiation are omitted entirely, though they are indirectly referred to. Not all the sources are given for the numbers I quote. In addition, for reasons stated in Chapter 6, the question of structure (e.g., the distribution of galaxies or clusters of galaxies) is omitted entirely.

The "details" of technical details are also items that I have glossed over, usually in favor of broadly constructed descriptions. Numbers, however, are not technical details! They are indispensable elements in describing and characterizing a Universe that is known to be very BIG now but is believed to have started out very small at very tiny times. Numbers of various kinds are sprinkled throughout the book, some in the text, some in tabular form.

To help make some observational/theoretical information easier to grasp, I have followed standard practice and portrayed it graphically. In doing so, I have assumed that graphical-type

representations are no more difficult to understand than the graphs or curves that describe the behavior of the usual stock-market indicators. Furthermore, many descriptions are illustrated by simple line drawings, each designed to enhance your comprehension of a particular written analysis.

Although the book is intended for readers without a science or math background, it does contain a few equations, including Einstein's famous formula $E = Mc^2$. I have tried to explain in plain English the meaning of each of the equations and proportionalities.

Theoretical cosmology is able to generate many different "universes." To distinguish them from "our" universe, I refer to ours as the *Universe*, while those generated by theory I denote *universe* with a lowercase first letter. In other words, *a* universe as opposed to *the* Universe. In analogy to using *Universe* and *universe*, I similarly use *galaxy* to refer to a very large aggregation of stars held or bound together gravitationally, whereas *Galaxy* always signifies our own, the Milky Way Galaxy.

By the way, a long-standing question has been which, if any, of the many theoretical universes most closely corresponds to the Universe. It is likely that the answer to this question is close at hand; one of the pleasures for me in lecturing on and writing about our Universe is describing the "how" of the answer.

I am aware that emphasizing theory and theoretical concepts could pose a risk: unlike the course attendees, you cannot query me should you fail to grasp an idea or an explanation. To help minimize this risk, persons with widely varying backgrounds—the majority of whom were not technically trained—agreed to read and comment on portions of the book in draft form. Their valuable critiques have led to improvements in both the writing and the content. Naturally, any errors or infelicities that remain are my sole responsibility.

Finally, let me draw your attention to two other aspects of the book. First, it is divided into the same two portions that formed the syllabus for the adult-education courses. The first part, ending with Chapter 4, contains the background material that I thought would help the adult students in my courses understand the second portion, which concentrates on cosmology and the Universe. I hope readers of this book will find this arrangement to be beneficial as well.

The other aspect is the set of Web sites listed in the Bibliography. They not only provide ancillary material but also can function as information sources for those of you who would like to remain up to date. You might even have fun logging onto your favorite Web browser and exploring the sites produced by searching phrases such as *Big Bang*, *black holes*, *dark matter*, *gravitational lensing*, *supernovas*, *CMB*, *WMAP*, *dark energy*, *inflation*, and so forth. If reading this book encourages such activity, one of its goals will have been met. Happy reading and browsing!

[Note added in proof: In March 2006, the Wilkinson Microwave Anisotropy Probe (WMAP) team announced their first results since 2003. Included is a revised estimate of when stars were first formed (see Figure 25), and confirmation of another prediction of inflation theory, thereby adding further support for this very early Universe scenario (described in Chapters 7 and 9). Some details of the new results, which also contain an updated map of the hot and cold spots of the Universe and an extrapolation to a trillionth of a second after the Big Bang, can be found at the WMAP Web site, listed in the Bibliography.]

Contents

1. Introduction: The Splendid Science

Cosmology! The branch of knowledge concerned with the origin, evolution, and properties of the Universe, cosmology is arguably the grandest of human endeavors, for what could be grander than attempting to understand the cosmos? The quest to achieve this understanding is ancient. Its unknown origin dates back thousands of years, when people in different cultures recorded the regular motions of the planets and stars and then used their observations to create calendars, to predict celestial events, and to speculate on the origin of the cosmos.

Although the science of astronomy got its start from these venerable beginnings, cosmology itself has emerged only recently as a branch of science in the modern sense of the word. Its emergence was due in large part to the accidental discovery in 1964 of a type of radiation known as the *cosmic microwave background radiation*, now referred to by the acronym *CMB*. Once the significance of the *CMB* was understood and publicized (and I'll explain it shortly), more and more people started to do research on cosmological topics, a community was formed, textbooks were written, and the new discipline of cosmology gradually came into being. It has become one of the most bountiful of the sciences.

Cosmology's stunning revelations fall into one of two categories: theoretical or observational/experimental. Among the most important theoretical investigations is the study of model universes, especially the ones produced by the universe-generating, mathematical theory known as general relativity. Model-universe studies began soon after Albert Einstein's paper on general relativity appeared in 1916. Prior to the 1960s, however, and despite similarities between some of them and our own Universe, model-universe studies generated relatively little interest among most scientists. This was due in part to the excitement created by new research areas such as nuclear physics. Equally important, if not

more so, was the mistaken perception that experiments could not connect any of the model universes with our own Universe.

This perception was dramatically altered by the serendipitous discovery of the CMB. Curiously enough, the discovery was made by two radio astronomers working for the Bell Telephone company! (See Chapter 6 for more details.) The microwave background radiation was quickly understood to be a previously predicted type of radiation that characterizes the early history of an entire class of theoretical universes, thereby providing the previously missing connection.

The existence of the CMB implies that our Universe is a member of the class of theoretical universes described by *Big Bang cosmology*. "Big Bang" refers to a generic type of expanding universe that has evolved from an explosive event, although the phrase itself was initially meant to be disparaging. It was introduced by a proponent of a theory known as *steady state cosmology*. Rather than evolving from an explosive event, the theoretical universes of steady state cosmology exist essentially unchanged in time, having neither a beginning nor an end. However, the CMB can only be accommodated in the steady state scenario by means of *ad hoc* assumptions, whereas it is a natural ingredient of the Big Bang framework.[a] Big Bang cosmology has triumphed, becoming a new paradigm, and the phrase *Big Bang* is now well-known outside of scientific circles. The discoverers of the CMB were awarded the Nobel Prize for a discovery that proved to be one of the most consequential of the 20th century.

Suppose that the CMB had not been detected, but that the Universe was somehow known to be a member of the Big Bang class of universes. This would necessitate its containing the CMB—which it does. But because the Universe is clumpy—apart from radiation, it is mostly empty space sparsely populated by galaxies and various other objects—theory predicts that the background radiation must also be clumpy. That is, the CMB measured from one region of the sky should differ slightly from the CMB when

---

[a]An *ad hoc* assumption or theory explains one fact only. It is scientifically unsatisfactory because it has no predictive power and therefore cannot be tested.

measured from any other part of the sky. If these differences were to exist, they would mean that the CMB deviates from perfect uniformity. Were such a deviation found, it would be dramatic evidence for the existence in the early Universe of the tiny nonuniformities in the distribution of matter that eventually led to galaxy formation.

The predicted nonuniformity, known as the *anisotropy* in the CMB, aroused great interest in the cosmology/astronomy communities. It led to the launching, in the late 1980s, of a satellite bearing equipment designed to detect the anisotropy. Called the cosmic background explorer and abbreviated COBE, it obtained data in 1992 that verified the prediction. The measured anisotropy was about 1 part in 100,000, or a thousandth of a percent, very small but much larger than experimental uncertainty.

The miniscule size of the anisotropy is a feature of the utmost significance, for hidden in it are clues that, suitably interpreted, yield information about the large-scale behavior of the Universe. Such information includes the overall geometry of the Universe, the amounts of both the luminous and the nonluminous, or "dark," matter in it, and the strength of the quantity (discussed in Chapter 6) that Albert Einstein once referred to as his "greatest blunder." The anisotropy's hidden treasures have motivated a host of theoretical investigations and experimental measurements. Highly accurate data have been obtained from many experiments carried out after the COBE mission. Notable among these investigations are those carried out by the Wilkinson Microwave Anisotropy Probe (WMAP) and the Sloan Digital Sky Survey (SDSS), discussed in Chapter 7.

The WMAP and SDSS findings, first reported in 2003, have been the best sources for evaluating the quantities that I denote the *parameters of the Universe*. Defined within the context of Big Bang cosmology, these parameters uniquely specify many properties of our Universe.

That these parameters, which are derived from theory, actually *can* specify properties of our Universe is based on the widely held belief of cosmologists that our Universe is uniquely identified with a theoretical universe generated by Big Bang cosmology. Underlying this identification are the facts that both our Universe and members of a particular class of Big Bang universes are each

expanding, contain the CMB, and are *homogeneous* and *isotropic*.[b] The sharing of these common features is evidence that not only is there a unique relation between our Universe and a member of the class of Big Bang universes but also that they behave in the same way. Knowledge of one thus provides information on the other.

To learn which of the theoretical universes correlates with ours requires deducing the parameter values from measurements made, for example, on supernovas, on the CMB and on galaxies, and then inserting these values into the relevant theoretical formulas. Such an insertion will select the theoretical universe to which ours corresponds, while from it, properties and the behavior over time of our Universe can be determined.

A key aspect of the theoretical analysis, indeed, one of the most astonishing in all of modern cosmology, is that the past, present, and future size of our three-dimensional Universe is obtained from just one quantity! This single quantity is known as the *universal scale factor*, and its existence is a consequence of the homogeneity and isotropy properties that the Universe enjoys in the large. In Chapter 6, I'll explain why the scale factor exists, and in Chapter 7 I'll discuss the time evolution of the model universes generated by the scale factor.

While the size of the Universe over time is described by the scale factor, the scale factor depends on the values of the parameters. Thus there is an exquisite linkage between the CMB and the time behavior of the Universe. The parameters and scale factor play crucial roles in elucidating other aspects of the Universe, discussed in Chapters 7 and 8.

The scale factor is related to one of the most important quantities in cosmology, the Hubble constant, first identified and estimated by the astronomer Edwin Hubble. He showed that our Universe is expanding in such a way that the speeds with which galaxies are receding from the earth are proportional to their distances away from it; the proportionality constant in this relation is the one that bears his name. Although the Hubble constant itself

---

[b]I'll define and illustrate the terms *homogeneous* and *isotropic* in Chapter 6, but you can look them up now in the Glossary, which defines the other technical terms I use in this book.

is highly significant, the relation it enters is equally so, as I will show later. Known as Hubble's law, this latter relation is also a consequence of homogeneity and isotropy.

Without knowledge of the distances to galaxies beyond the Milky Way as well as their recession speeds, Hubble could not have deduced his law. Accurate measurement of astronomical distances, and later of cosmological ones, has been an essential requirement in all scientific attempts to understand the Universe, and you may well have wondered how such measurements have been accomplished. The answer is through a set of interlocking methods that form a hierarchy, one in which the easier-to-obtain shorter distances become the springboard for reaching out to longer distances. This collection of methods is known as the *cosmic distance ladder*, and although only the lower rungs were available to Hubble, they sufficed—spectacularly well—for his purposes.

Distance determination is so vital to astronomy and cosmology that *parallax*, the lowest method on the distance ladder, is the main subject of Chapter 2. When parallax fails, some of the methods that supercede it rely on the properties of certain types of stellar phenomena, for example *Cepheid variables* and *type Ia supernovas*. The role played by these exotic entities in determining distances is one of several reasons for my including the very broad discussion of stars of Chapter 4; another is the intrinsic interest that stars hold for most persons, especially stellar endstages such as white dwarfs and black holes. Furthermore, stars shine: they are the most populous of the luminous ingredients in the Universe, and gaining some understanding of them is an essential element in appreciating the cosmos.

Hubble not only needed reliable distances, he had to know the recession speeds of the galaxies. They were—and are—obtained using a mechanism that exploits the wave properties of light and radiation. An essential element in understanding the cosmos is grasping how scientists have deduced that galaxies are receding from one another, as well as how fast they are moving away. This alone is a powerful reason for my reviewing light and radiation in Chapter 3. Another is the fact that light, and radiation in general, is the sole source of observational information about the Universe (hearing, taste, and smell obviously don't work!).

The cosmic microwave background is an almost perfect example of a kind of radiation known as *blackbody*. It is crucial to the construction of a timeline for the Universe that the CMB is of this type, as I discuss in Chapter 8. And, not only is the CMB approximately blackbody in nature but so also is the radiation emitted by the sun. The interplay between these very different entities neatly illustrates both the unity of the Universe and the use of terrestrial science to explain it.

Of course, not everything one wishes to know about the cosmos has been or can be deduced by examining its various forms of radiation. Critical aspects of it remain unknown, for instance, the identity of the nonradiating dark matter as well as the nature of the *dark energy* causing the expansion of the Universe to accelerate. (This acceleration is another 20th-century discovery that has revolutionized thinking about the cosmos.) Moreover, cosmology is not yet a completely fleshed-out science, so that explanations of some observational or inferred phenomena are based on conjectures that range from the highly likely to the highly speculative (see Chapter 9).

Even though not all the answers are in, much has been ascertained. Thus, while the nature of dark matter remains a mystery, the relative amounts of the current contents of the Universe and the nature and times of occurrence of many events that took place during its evolution have been estimated. Its large-scale geometry is known. An analysis of WMAP data combined with those from other experiments leads to the time of the Big Bang as approximately 13.7 billion years ago. The diameter of the visible Universe can also be estimated. It is roughly a quarter of a million billion billion kilometers (0.25×10^{24} km),[c] or a sixth of a million billion billion miles. As you will discover in this book, these and other results, along with some of the conjectures about the cosmos, are as astonishing as any that occur in a non-cosmological context: the Universe *is* comprehensible, and cosmology explains much of it.

---

[c]The power-of-ten notation, for example, 10^{24}, is described in Appendix A.

2. Measuring Distances: On the Earth, in the Solar System, to the Nearby Stars

Distances play much the same role in astronomy and cosmology as perspective does in landscape painting: change either of them and the resulting picture changes. Accurate distances are required if you are to obtain a reliable picture of the Universe, just as they are in determining the size of the earth or the solar system or the Galaxy. The problem in each of these instances is the same: how is the requisite distance to be obtained when a direct measurement cannot be made? The solution is through the use of indirect methods, and I begin the description of them in this Chapter.

Attempts to measure distances, both successful and not, are part of the history of astronomy. Many of the successful procedures have been organized into a hierarchy known as the *cosmic distance ladder*, with each rung describing a distinct method, the lower ones typically supporting the higher ones. As one climbs the ladder, the associated distances increase; unfortunately, all of the procedures are imprecise, so that the inaccuracies of the lower-rung methods are incorporated into those of the higher rungs. Because inaccuracy is an inevitable aspect of this enterprise, great efforts have been made to ensure high precision in the shorter-distance measurements. I shall consider aspects of errors after introducing appropriate distance units.

The main distance method examined in this Chapter is denoted *parallax*. It occupies the lowest rung on the ladder and extends only to the "nearby" stars. Although such distances are small on the cosmological scale, there are two reasons for beginning with parallax: first, its assumptions and its *one-angle/one-known-distance* characteristic can be exposed in the more familiar setting of certain types of terrestrial measurements; second, it has been applied in the solar system. The former is important because

the assumptions, which are rarely identified, are not all valid in the case of cosmological measurements. The latter application is useful because I will use it as the platform for discussing some concepts and details of the solar system such as planetary orbits, as well as mass and density, quantities essential for describing the Universe both observationally and theoretically.

Parallax (also known as *trigonometric* parallax) employs a *measured angle* and a *predetermined length* to evaluate the desired stellar distance. These same two elements entered the first measurement of the earth's radius, carried out ca. 240 BCE by the Greek philosopher Eratosthenes, a one-time director of the renowned library in Alexandria, Egypt. However, the one-angle, one-known-length method is not limited to measurements of very large lengths: it can also be used to determine quite ordinary distances, such as the heights of fixed vertical objects like telephone poles, trees, or sailboat masts. Since it is simplest to explain the method for this latter class of objects, I'll introduce the discussion of parallax by describing a procedure for measuring the height of a standing telephone pole without climbing it. A key element will be identifying the relevant assumptions. After that, I'll go on to the method used by Eratosthenes.

Measuring the Height of a Standing Telephone Pole

Figure 1 is a schematic depiction of a telephone pole, whose height h is to be determined. To begin, one marks off a length D to the left of the pole; it is the predetermined-distance portion of the method. The angle-measuring device, e.g., a protractor or similar instrument, shown as the small circle with a plus sign (+) in it, is then put into the ground at this distance. By creating a line of sight from the center of the protractor to the top of the pole, indicated in the figure by the dotted line, the measurer defines an angle, labeled A, whose value (in degrees) is read off the protractor. The telephone pole, the distance D, and the dotted line form a triangle, which is a shape from plane, or flat-space, geometry—the geometry of Euclid.

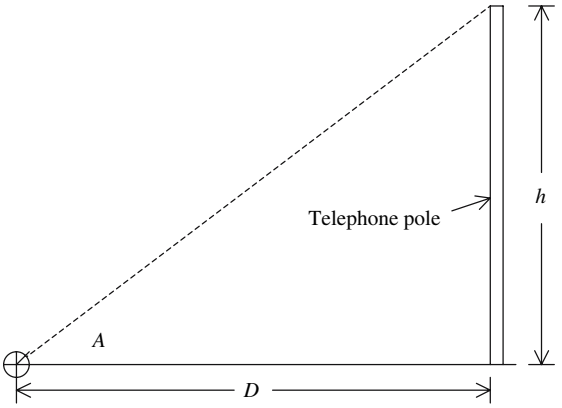

Figure 1. Illustration of the one-angle, one-measured-length method of determining a distance. To "measure" the height h of the telephone pole, one needs only to measure the preselected distance D and the angle A, where the symbol $\oplus$ represents an angle-measuring device such as a protractor.

h is uniquely determined by this construction. Of course, only D and A are measured: h itself is not. Instead, its numerical value is obtained from the other two measurements by using a mathematical formula based on plane geometry.[1] Nevertheless, the method is referred to as the *measurement of h*, just as the method of parallax is a means of "measuring" a stellar distance. Each is an example of the indirect procedure I mentioned at the beginning of the Chapter.

Although the foregoing description may seem straightforward, it contains some unspecified but crucial assumptions. First, by creating and using a triangle to define A, the geometry of flat surfaces (Euclidean geometry) is assumed to be valid. And indeed it is, as long as the distance D is not so great that the curvature of the earth's surface needs to be taken into account. This is normally the case because in any small region—that is, locally—the curvature is far too small to change the geometry from planar to spherical. But if the curvature were to become noticeable, then spherical geometry might become necessary. Spherical geometry would call for a different math formula, since the relevant distance would not be a straight line but a portion of a great circle, the type of route followed by airplanes flying long distances or ships crossing oceans.

A second assumption is that the pole and the protractor remain stationary, so that neither h nor D changes. For the type of measurement described above, this may seem like a frivolous remark, but in an expanding universe, quantities analogous to D *are changing*, and one must take care in dealing with distance. A third unstated assumption is that both the distance D and the angle A not only can be measured, but that it can be done with an accuracy sufficient for the purpose at hand. As I note later in this Chapter, for D large enough, the uncertainty in angles can become significant, whereas for most astronomic and cosmological distances it is impossible even to discern a parallax angle. When this occurs, parallax must be replaced by another method, one from a higher rung of the cosmic distance ladder.

An Aside on Angles

Since the procedure just outlined involves angles, let us take a small detour away from the next measurement—that of the earth's radius—and focus attention on the units in which angles are specified. In nontechnical applications, angles are measured in *degrees* and are indicated by the symbol ° placed as a superscript to the right of the numerical value; for instance, 30°. The degree is a concept originally associated with circles and was formulated by mathematicians of the ancient Babylonian civilization. Rather than 10, the base of the decimal system, they favored the number 60 and its multiples and divisors. In particular, they divided the circumference of a circle into 360 equal segments of arc and then defined the angle between the two radii drawn to the ends of one such segment as equal to 1°. In other words, one such arc segment *subtends* an angle of 1°. This division of a circle's circumference into 360 segments means that there are 360° in a circle. An arbitrary angle defined in this way (not equal to 1°) is shown in Figure 2.

An angle may not always be expressible as an integer number of degrees: its value may involve a fraction of a degree. The Babylonians dealt with this possibility by dividing each whole degree into 60 equal parts called *minutes*, indicated by the prime symbol, so that 1° contains 60′. And just as a degree comprises 60 minutes,

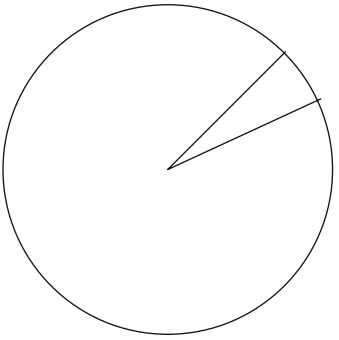

Figure 2. An angle subtended by a short arc of a circle and contained between two radii.

a minute was divided into 60 smaller portions, denoted *seconds*. The symbol for a second is the double prime, for instance 30″, which is one half of a minute. The Babylonians ended their subdivisions here; for smaller subdivisions, the modern decimal system is used, for example one tenth of a second is written 0.1″. Such small values are common in parallax measurements. Of course, one can avoid minutes altogether, replacing them with tenths of a degree (see below).

These Babylonian subdivisions into sets of 60 define time units as well: 60 minutes in an hour and 60 seconds in a minute. Despite their lacking the advantages of a decimally based set of units—the division of the day into 24 one-hour portions is an ancient Egyptian construction—the Babylonian/Egyptian system remains in effect today and is highly unlikely to be replaced: usage and tradition have trumped numerical convenience.

Back to Yesteryear: Measuring the Earth's Radius

When Eratosthenes measured the earth's radius, the cosmology he believed in was that of Aristarchus (ca. 320–250 BCE), which held (almost correctly!) that the earth was spherical,[2] that it rotated on its own axis, and that it revolved about the sun. Eratosthenes also believed, as we do not, that the earth was embedded in a spherical shell that did the actual revolving. That the earth is revolving

about the sun and not *vice versa* is the hallmark of a heliocentric cosmology.[a]

Eratosthenes had learned that at noon on the summer solstice in Syene (the Egyptian city now called Aswan), sunlight cast no shadow (it fell perpendicular to the earth's surface there). However, at noon on the same day in Alexandria, sunlight fell obliquely on the earth's surface, making an angle of about 7.2° with an upright stick. Using these facts and the then known distance of 5000 *stadia* between the two cities, he was able to calculate the value of the earth's radius (he "measured" it), expressing the result in stadia. (To convert his result to miles or kilometers, and thus determine its accuracy, one needs to know how many stadia there are in a mile or a kilometer, a point considered shortly.)

Recalling comments I made above, it might seem that a spherical earth would have required Eratosthenes to base his calculation on spherical geometry. It was not needed because the center of the earth and the two cities lie in a plane: Euclidean geometry sufficed. The formula he used relates an arc of a circle to the angle it subtends at the center and to the radius. In the case at hand, the arc length is the distance between the two cities, and the radius is that of the earth. The former is the known length of the method; the angle he needed is the one subtended at the center of the earth by the ends of the arc. Because he couldn't measure it directly, Eratosthenes replaced it by an angle of equal size that he *could* measure.

The procedure that led to the desired angle is based on Figure 3, which shows a cutaway portion of the earth, defined by the plane noted above. It passes through the earth's center and the two cities on the surface; the drawing is not to scale. Depicted in the figure are arrows denoting the parallel set of the sun's rays, which strike the earth's surface perpendicularly at Syene and obliquely at Alexandria, plus the radius from the earth's center to each city

---

[a]Opposing it was the geocentric cosmology of Aristotle and others, codified in the second century by Claudius Ptolemy through publication of his book *The Almagest*. Geocentricity held sway in Europe for well over a thousand years, until Nikolaus Copernicus challenged it in the late 1500s.[3]

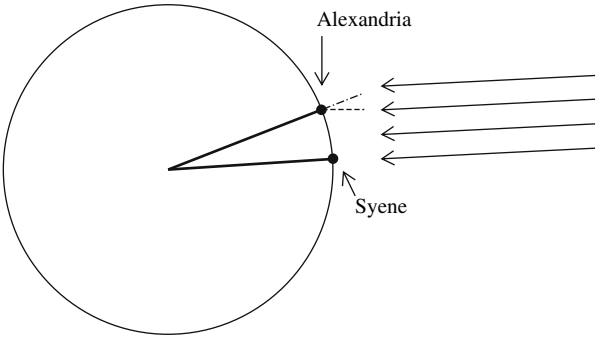

Figure 3. Illustration of the geometry used by Eratosthenes to "measure" the earth's radius. The circle denotes a plane cut through the center of the earth; the left-pointing arrows represent the parallel rays of the sun striking the earth's surface; the heavy lines are the radii to the cities of Alexandria and Syene, each symbolized by a heavy black dot; the dot-dash line is an extension of the radius to Alexandria; the dashed line is parallel to the radius to Syene.

(in bold), with the one to Syene parallel to the sun's rays. The dashed line drawn at Alexandria is parallel to both the sun's rays and the radius to Syene, and the vertical stick is represented by the dash-dot line, which is an extension of the radius to Alexandria.

There are two angles in the figure: the subtended one between the two radii—which Eratosthenes needed to know—and the one between the dot-dash and dashed lines. The key relation for him was the plane-geometry result that these two angles are equal. But, the angle between the dot-dash and dashed lines is the *same* as the angle with which the sun's rays strike the ground at Alexandria, viz., 7.2°. Hence, the subtended angle between the two radii is also 7.2°.

This conclusion, plus the fact that arc length is the product of the radius and the subtended angle (7.2°), enabled Eratosthenes to calculate the radius of the earth. He found it to be approximately 39,788 stadia, corresponding to a circumference of the earth of 250,000 stadia. His measurement of one angle and one distance plus his use of plane geometry yielded a "measurement" of the earth's circumference and radius.

The method just outlined is a triumph of the intellect, an imaginative use of reasoning based on analogy and geometry.

Although it is incisive, one must ask if the result is accurate. The only way to answer such a question is to convert stadia to contemporary units and then compare with the modern value. Assuming that the contemporary distance of 500 miles between Alexandria and Aswan is the same as the ancient Alexandria–Syene distance of 5000 stadia (however that value was obtained in the third century BCE), then one *stade* (the singular of stadia) equals one tenth of a mile. Hence, Eratosthenes's values convert to 25,000 miles for the earth's circumference and about 3979 miles for its radius.

How do these numbers compare with the contemporary values? Unfortunately, the comparison is not straightforward because the earth's circumference and radius are not uniquely defined! On the one hand, and even forgetting the existence of mountains, the earth is not spherical: its surface contains a variety of small deformations superimposed on one another, including a slight pear shape. On the other hand, even if this latter fact were to be ignored—as it can be for the present purposes since these deformations are small—there is the problem of the earth's bulge: due to its rotation, the earth is fatter at the equator than elsewhere (cf. note 2). The standard solution to this non-unique-radius problem is to use the equatorial value, in which case the earth's "radius" is found to be 3963 miles or 6378 km, and its circumference is almost identical to 25,000 miles or 40,074 km. The agreement between these values and the measurement of Eratosthenes is excellent. Not only is his result remarkably accurate, it was accepted as correct by his contemporaries, thereby demonstrating the esteem in which analytical reasoning was then regarded.

As with the measurement of the height of the telephone pole, Eratosthenes's method makes use of assumptions beyond that of a spherical earth. Three have been stated already: that measurements can be made with sufficient accuracy; that it is valid to apply Euclidean geometry to the process; and that distances and measuring devices (sticks or rods or strings of known lengths) are fixed quantities. A new one is that the radius of the earth is so much less than the sun–earth distance that rays of sunlight are parallel to one another when they hit the earth. This assumption is valid to a very high degree of accuracy. Eratosthenes undoubt-

edly accepted all of them without reservation—he may never have thought to question them. Nevertheless, they are assumptions, ones that need not, and do not, hold in all circumstances.

On the Use of Appropriate Units

The preceding assumptions are critical to the particular measurement process. In a different category are the choices of units in which to express various measured quantities, especially distances. It is standard practice to use miles or kilometers when the distances are large compared with the lengths of typical human or household objects, which are measured in inches and feet or centimeters (cm) and meters (m). Why are the latter units not used for distances on the earth's surface, or for its radius, or, especially, for astronomical distances? One answer is convenience: there is too much "bulk" when the smaller units are used, just as would be the case if you were forced to use coins rather than paper money when paying a large bill in cash.

The bulkiness of the smaller units refers to the quantity of numbers involved, as is aptly illustrated by the earth's radius. Recall that a mile is equal to 5280 feet or 63,360 inches, and a kilometer is equal to 1000 m or 100,000 cm. Using inches and centimeters as the units for the earth's radius, which from now on I will denote by D_E (the letter D stands for a distance, including that of the earth's radius, and the subscript E signifies the earth), its value in these units is equal to either 251,095,680 inches or 637,800,000 cm! The size of these numbers should make it clear that miles and km are the more appropriate units, if for no other reason than not wishing to take the time and space to write out nine digits as opposed to four. However, there is another reason for not using the preceding pairs of nine numbers, one related to the concept of *significant figures*. How many of the digits in each set of nine are needed for both accuracy and understanding, as opposed to precision? That is, how many of the nine digits are *significant*—or meaningful—in the particular context? The general answer to a significant-figures question depends entirely on the amount of inaccuracy that can be tolerated.

There have been situations in science, particularly in atomic and elementary-particle physics, where the determination of as many significant figures as possible has been the key to progress: new experimental values have led to major developments in theory, and on occasion the reverse has been true. As I will show in later Chapters, careful measurements to a sufficient number of significant figures have played essential roles in astronomy and cosmology. But, in the present context, maintaining ultimate precision is generally unnecessary: because my discussions are descriptive and not technical, no vital information will be lost by keeping only a few, rather than the entire set of nonzero digits in any large numbers.

A case in point is the value of D_E: if it is approximated either by 251,000,000 inches or 638,000,000 cm, no information vital to our purpose is omitted: the essential information resides in how many hundreds of millions there are, not in how many hundreds of thousands. The errors made by using the previous approximations are just a few hundredths of a percent, which is insignificant for our purposes. However, if one were to insist on employing the smaller units—which I do not—there is another argument behind using the approximations just cited: the mile or kilometer values of D_E are themselves approximate. Retaining the precision of all nine digits thus turns out to be an exercise in pedantry rather than in accuracy.

Even if only a few significant figures are kept, however, the total number, including the zeros, may still be bulky. The overall solution to the bulkiness problem, once only the significant figures have been retained, is to employ the power-of-ten notation. Used for both very large and very small numbers, it is described in Appendix A. In terms of this powerful notation, the value for D_E when it is expressed in the inappropriate units becomes 2.51×10^8 inches or 6.38×10^8 cm.

Copernicus, Kepler, and the Astronomical Unit

In recent years, so many new results in observational astronomy and cosmology have been publicized that it is easy to ignore how much had been learned via naked-eye astronomy. Prior to the use

of telescopes, people in many parts of the world believed the physical universe to consist of the earth, the sun, the moon, the five then-known planets (Mercury, Venus, Mars, Jupiter, and Saturn), the stars, and the lesser bodies such as comets and meteors. One application of the regularity in the motions of these bodies was to create reliable calendars for various purposes, including agriculture and religion.[4] Although there may have been calendar makers in other places and at earlier times who attempted to make reliable estimates of distances to any of the bodies listed above, the efforts of the early Greeks in this regard are the best known in the West. Their most accurately measured quantity was the earth's radius D_E, and because they were unable to determine a reliable value for the earth–sun distance, only a range of possible earth–moon distances were obtained, although the lower end of this range was remarkably good.[5]

From the time of Ptolemy until the work of Nikolaus Copernicus eventually reestablished the heliocentric solar system, the accepted cosmology of pretelescopic Europe was geocentric. And, until his analysis of Tycho Brahe's (naked-eye) data led Johannes Kepler to conclude that the planets (including the earth) moved in elliptically shaped orbits, Europeans also believed that only the circle was needed to describe planetary orbits.[6] There were, therefore, two paradigmatic shifts ushered in by Copernicus and Kepler: from "geo" to "helio," and from circles to ellipses.

In a sense, Copernicus straddled these developments: he reintroduced heliocentricity but retained the concept of circular orbits (in fact, he used coplanar circles centered on the sun). From these assumptions, plus an analysis based on plane geometry and his own observational data, Copernicus deduced the relative distances between the sun and the five non-earth planets. He expressed them in terms of the unknown earth–sun distance, publishing his results in 1543. As shown later in Table 1, these relative distances were remarkably accurate, given that his was naked-eye astronomy (even more accurate naked-eye data was obtained by Tycho Brahe about 100 years later). In view of this accuracy, determination of the size of the solar system in Copernicus's model of the cosmos needed only *one* additional measurement: that of the earth–sun distance, which I will denote by D_{ES}. The need for one additional

measurement holds true in the modern view of the solar system, due to Kepler and Isaac Newton.[6]

By expressing the five sun–planet distances in terms of the earth–sun distance, Copernicus exploited the fact that in an orbital system based on circles, any of the sun–planet distances can serve as the length unit for the other ones. However, planetary orbits are not circles but *ellipses*—as was known to astronomers in India as long ago as ca. 600 (cf. note 4) and rediscovered by Kepler about a thousand years later. In a circular orbit about the sun, the earth would always be at a constant distance from it, but for an elliptical orbit, the earth–sun distance is continuously changing. A new problem therefore arises: which of these varying distances should be taken as the unit for measuring all the other planet–sun distances? The solution to this problem resides in one of Kepler's three laws of planetary motion, which I consider after describing some features of ellipses.

Figure 4, which compares a circle and an ellipse, illustrates aspects of this new nonuniqueness problem. An ellipse is a geometric figure that is symmetric in both the up–down and the left–right directions. It looks like a squashed circle. Each of the two heavy points in Figure 4 is called a focus, and the ellipse itself is constructed such that the sum of the distances from the two foci to any point on its periphery is equal to a constant. With this construction, the longest straight line that can be drawn interior to the ellipse is the horizontal one of length $2a$. It is the *major*

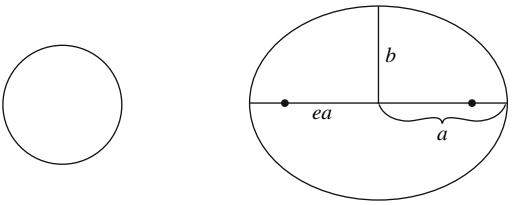

Figure 4. Comparison of (a) a circle and (b) an ellipse. The length of the semimajor axis of the ellipse is denoted a, its semiminor axis length is b, and the product of the eccentricity e and the semimajor axis a, viz., ea, is the distance from the center to either *focus*, each of which is specified by a heavy black dot.

axis of the ellipse. A vertical straight line of length $2b$ drawn through the center of the ellipse is its *minor axis* (half these distances are the semimajor and the semiminor axes, a and b; only the upper portion of the minor axis is displayed in the figure). The amount of "squashing" is characterized by the *eccentricity*, e, whose values range from 0 to 1: when e is zero, the ellipse becomes a circle, whereas for e equal to 1, the ellipse collapses to a straight line of length $2a$. In Figure 4, the eccentricity is approximately equal to 0.7.

The reason for detailing these properties is that the ellipse plays such a prominent role in the three laws of planetary motion that Kepler deduced from Tycho's wonderfully accurate, naked-eye data. These "laws" are empirical in nature, in that they were deduced from observational data rather than being theoretically based. Kepler's first law states that the orbits are ellipses with the sun located at one of the foci (not at the center). His second law is both technical in character and not relevant to my analysis, and I am therefore omitting it here; interested readers may look it up in Harrison (2000) or Webb (1999). The third law is a *universal* relation between the semimajor axis of the orbit and the corresponding period—the time it takes the planet to make a full revolution about the sun (1 year in the case of the earth).

Because of the universality of this latter relation, Kepler (and later Newton) chose D_{ES}, the semimajor axis of the earth's orbit, to be the earth–sun "distance." Since the periods of the planets were known very accurately, the third law allowed Kepler to calculate each of the five planet–sun distances with some precision; he expressed them, of course, in units of the then unknown D_{ES}. Now denoted the *astronomical unit*, abbreviated AU, D_{ES} sets the scale of the solar system. In keeping with the choice of D_{ES} as the earth–sun "distance," the other semimajor axes are each defined as the "distance" of its planet from the sun. Are the deviations from circularity of the orbits very large? No: the eccentricities of most of the planetary orbits are less than 0.1, so that the sun is much closer to the center than to the periphery of the planetary ellipses.[7]

The values of the five planet–sun distances determined by Copernicus and by Kepler are shown in Table 1. Those of Copernicus are naked-eye results, whereas those of Kepler are

Table 1. Planet–Sun Distances Expressed in AU*

Planet	Copernicus's values	Kepler's values
Mercury	0.38	0.387
Venus	0.72	0.723
Earth	1.00	1.000
Mars	1.52	1.524
Jupiter	5.22	5.200
Saturn	9.17	9.531

*Values from Webb (1999).

based on his first and third laws of planetary motion (themselves deduced from naked-eye observations). The excellent agreement between them was a strong motive for accurately measuring the astronomical unit D_{ES}, since its measurement determines all the others.

Parallax

Prior to the invention of radar, a telescope was required to obtain even an estimate of D_{ES}. It took more than 200 years after the telescope's invention—probably in the early 17th century—before D_{ES} was measured with an accuracy close to that obtained using radar. The method used was parallax.

Anyone with binocular vision should easily grasp the concept of parallax, as it is the brain's intrinsic method for estimating distance. It relies on the fact that for objects not too far away, each eye sees a different image, the images being slightly displaced from one another against the background common to both. The following simple experiment shows how this works: stretch either arm to its fullest extent in front of your face, raise your thumb, and then look at it twice, first closing one eye and then the other (it is here that binocular vision enters). By carrying out this exercise, you should find that the position of your thumb moves relative to the fixed background (from right to left or left to right, depending on which eye was closed first). A similar situation arises whenever binocular vision is used to observe a not-too-distant object. In every case, with both eyes open, the brain melds the two

images into a single one; by so doing it becomes a distance estimator (of course, this means of estimating distance becomes refined by experience).

In the case of binocular vision, parallax refers to the brain's melding of the images seen by the two eyes. In an astronomical context, parallax refers to the observation of an object from two different vantage points, typically from well-separated points on the earth or from two points on the earth's orbit separated by 6 months. Figure 5 illustrates the geometry involved: the observation points are labeled 1 and 2; the object, here taken to be a point, is labeled *O*; and the distance to *O* from the midpoint between 1 and 2 is denoted *D*. The light from *O* that reaches the observation points 1 and 2 is represented by the dashed lines in the figure (this light is either emitted, as from a star or galaxy, or scattered, as in the case of a planet). The angle *A* between the line *D* and the lines from either 1 or 2 to *O* is the *angle of parallax*.[8] Note that by identifying points 1 and 2 with a person's eyes and point *O* with his or her thumb, this construction encompasses the binocular vision example just described.

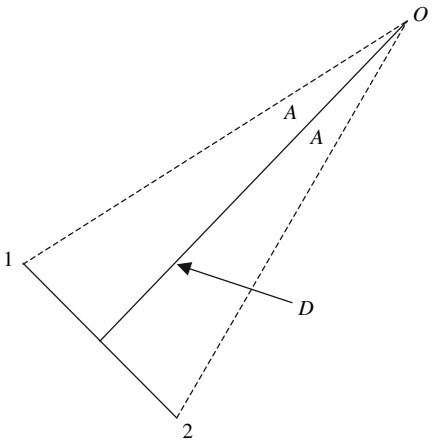

Figure 5. Illustration of the method of *parallax* (or *trigonometric parallax*) for measuring distances. The object *O* is at the distance *D*, whose length is to be measured. Points 1 and 2 are the locations of the two places where the observation of *O* occurs; the distance between points 1 and 2, indicated by the heavy line connecting them, is presumed known and, in the method of *horizontal parallax*, shown in the figure, *A* is the angle of parallax.[8] The method requires that *A* be measurable.

In the figure there are three different distances to the observation point *O*, namely, the distance *D* plus the separations between *O* and the two observation points. Although any one of them could qualify as *the* distance to *O*, the astronomical application is often to the determination of *D*, which will henceforth be designated as the desired *distance to be measured*. Because parallax is a known-distance/measured-angle procedure, each of its two elements must be determined beforehand. The known distance is the separation between the observation points 1 and 2, while *A* is the angle to be measured. In principle, the parallax angle is measured by means of observations made on the object from the two vantage points (each set of observations is made against the fixed background—the "fixed" stars[9] in the case where *O* is a nearby star).

There is a caveat associated with this procedure, as suggested by the phrase "in principle." It is based on the fact that as *D* increases, *A* decreases toward zero. Although a zero angle would occur only at an infinite distance, in practice *D* should not be so great that the angle *A* becomes too small to determine; that is, if *O* is too far away, there will be no measurable parallax. This sets a limit on the use of parallax to determine the distances to stars; correspondingly, high accuracy is required. In addition, you should bear in mind that the method of parallax involves most of the other assumptions noted previously; for example, that use of Euclidean geometry is valid.

Even taking account of the need to exercise care in measuring the angle of parallax, the procedure as just described is less straightforward than it might seem when applied to the earth–sun distance. The problem is that the object *O* in Figure 5 is a point, whereas the sun has an obvious size, in contrast with every other star seen from earth. Indeed, the angular widths of the sun and of the moon when it is closest to the earth are about the same—approximately 33′15″—thus allowing for spectacular lunar eclipses of the sun. The non-point-like character of the sun can be overcome, as was realized in the early 17th century, by measuring the parallax of either a planet or an asteroid as it transits the face of the sun. Combining geometry and Kepler's third law with the latter measurement allows D_{ES} to be determined.