CELESTIAL MECHANICS

Jeremy Tatum
University of Victoria

LibreTexts™

University of Victoria

Celestial Mechanics

Jeremy Tatum

The LibreTexts libraries are Powered by MindTouch® and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

This text was compiled on 09/17/2020

TABLE OF CONTENTS

Celestial mechanics is the branch of astronomy that deals with the motions of celestial objects. Historically, celestial mechanics applies principles of physics (classical mechanics) to astronomical objects, such as stars and planets, to produce ephemeris data.

1: NUMERICAL METHODS

This chapter is not intended as a comprehensive course in numerical methods. Rather it deals, and only in a rather basic way, with the very common problems of numerical integration and the solution of simple (and not so simple!) equations. Specialist astronomers today can generate most of the planetary tables for themselves; but those who are not so specialized still have a need to look up data in tables.

2: CONIC SECTIONS

A particle moving under the influence of an inverse square force moves in an orbit that is a conic section; that is to say an ellipse, a parabola or a hyperbola. We shall prove this from dynamical principles in a later chapter. In this chapter we review the geometry of the conic sections. We start off, however, with a brief review (eight equation-packed pages) of the geometry of the straight line.

3: PLANE AND SPHERICAL TRIGONOMETRY

4: COORDINATE GEOMETRY IN THREE DIMENSIONS

5: GRAVITATIONAL FIELD AND POTENTIAL

This chapter deals with the calculation of gravitational fields and potentials in the vicinity of various shapes and sizes of massive bodies. The reader who has studied electrostatics will recognize that this is all just a repeat of what he or she already knows.

6: THE CELESTIAL SPHERE

If you look up in the sky, it appears as if you are at the centre of a vast crystal sphere with the stars fixed on its surface. This sphere is the celestial sphere. It has no particular radius; we record positions of the stars merely by specifying angles. We see only half of the sphere; the remaining half is hidden below the horizon. In this section we describe the several coordinate systems that are used to describe the positions of stars and other bodies on the celestial sphere.

7: TIME

In this chapter we briefly discuss the several time scales that are in use in astronomy, such as Universal Time, Mean Solar Time, Ephemeris Time, Terrestrial Dynamical Time, and the several types of second, hour, day and year that are or have been in use. Some of the items in this chapter will be given only in short note form or single sentence definitions, particularly where they have already been discussed. Others will require a bit more discussion.

8: PLANETARY MOTIONS

In this chapter, I do not attempt to calculate planetary ephemerides, which will come in a later chapter. Rather, I discuss in an idealistic and qualitative manner how it is that a planet sometimes moves in one direction and sometimes in another. That the treatment in this chapter is both idealistic and qualitative by no means implies that it will be devoid of Equations or of quantitative results, or that the matter discussed in this chapter will have no real practical or observational value.

9: THE TWO BODY PROBLEM IN TWO DIMENSIONS

In this chapter we show how Kepler’s laws can be derived from Newton’s laws of motion and gravitation, and conservation of angular momentum, and we derive formulas for the energy and angular momentum in an orbit. We show also how to calculate the position of a planet in its orbit as a function of time. The discussion here is limited to two dimensions. The corresponding problem in three dimensions, and how to calculate an ephemeris of a planet or comet in the sky, is discussed elsewhere.

10: COMPUTATION OF AN EPHEMERIS

11: PHOTOGRAPHIC ASTROMETRY

Astrometry is the art and science of measuring positions of celestial objects, and indeed the first step in determining the orbit of a new asteroid or comet is to obtain a set of good astrometric positions. For much of the twentieth century, most astrometric positions were determined photographically.

A visual binary is a gravitationally bound system that can be resolved into two stars. These stars are estimated, via Kepler's 3rd law, to have periods ranging from a number of years to thousands of years. A visual binary consists of two stars, usually of a different brightness.

18: SPECTROSCOPIC BINARY STARS

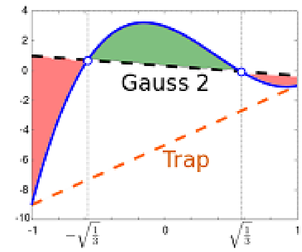
There are many binary stars whose angular separation is so small that we cannot distinguish the two components even with a large telescope – but we can detect the fact that there are two stars from their spectra. In favorable circumstances, two distinct spectra can be seen. It might be that the spectral types of the two components are very different – perhaps a hot A-type star and a cool K-type star, and it is easy to recognize that there must be two stars there.

BACK MATTER

CHAPTER OVERVIEW

1: NUMERICAL METHODS

This chapter is not intended as a comprehensive course in numerical methods. Rather it deals, and only in a rather basic way, with the very common problems of numerical integration and the solution of simple (and not so simple!) equations. Specialist astronomers today can generate most of the planetary tables for themselves; but those who are not so specialized still have a need to look up data in tables.



There are many occasions when one may wish to integrate an expression numerically rather than analytically. Sometimes one cannot find an analytical expression for an integral, or, if one can, it is so complicated that it is just as quick to integrate numerically as it is to tabulate the analytical expression. Or one may have a table of numbers to integrate rather than an analytical equation.

The Newton-Raphson method is very suitable for the solution of polynomial equations.

In nearly all cases encountered in practice Newton-Raphson method is very rapid and does not require a particularly good first guess. Nevertheless for completeness it should be pointed out that there are rare occasions when the method either fails or converges rather slowly.

1.1: Introduction to Numerical Methods

I believe that, when I was a young student, I had some vague naive belief that every Equation had as its solution an explicit algebraic formula, and every function to be integrated had a similar explicit analytical function for the answer. It came as quite an eye-opener to me when I began to realize that this was far from the case. There are many mathematical operations for which there is no explicit formula, and yet more for which numerical solutions are either easier, or quicker or more convenient than algebraic solutions. I also remember being impressed as a student with the seemingly endless number of "special functions" whose properties were listed in textbooks and which I feared I would have to memorize and master. Of course, we now have computers, and over the years I have come to realize that it is often easier to generate numerical solutions to problems rather than try to express them in terms of obscure special functions with which few people are honestly familiar. Now, far from believing that every problem has an explicit algebraic solution, I suspect that algebraic solutions to problems may be a minority, and numerical solutions to many problems are the norm.

This chapter is not intended as a comprehensive course in numerical methods. Rather it deals, and only in a rather basic way, with the very common problems of numerical integration and the solution of simple (and not so simple!) Equations. Specialist astronomers today can generate most of the planetary tables for themselves; but those who are not so specialized still have a need to look up data in tables such as The Astronomical Almanac, and I have therefore added a brief section on interpolation, which I hope may be useful. While any of these topics could be greatly expanded, this section should be useful for many everyday computational purposes.

I do not deal in this introductory chapter with the huge subject of differential Equations. These need a book in themselves. Nevertheless, there is an example I remember from student days that has stuck in my mind ever since. In those days, calculations were done by hand-operated mechanical calculators, one of which I still fondly possess, and speed and efficiency, as well as accuracy, were a prime concern - as indeed they still are today in an era of electronic computers of astonishing speed. The problem was this: Given the differential Equation

$$\frac{dy}{dx} = \frac{x+y}{x-y} \tag{1.1.1}$$

with initial conditions $y = 0$ when $x = 1$, tabulate y as a function of x. It happens that the differential Equation can readily be solved analytically:

$$\ln(x^2+y^2) = 2\tan^{-1}(y/x) \tag{1.1.2}$$

Yet it is far quicker and easier to tabulate y as a function of x using numerical techniques directly from the original differential Equation 1.1.1 than from its analytical solution 1.1.2.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.2: Numerical Integration

There are many occasions when one may wish to integrate an expression numerically rather than analytically. Sometimes one cannot find an analytical expression for an integral, or, if one can, it is so complicated that it is just as quick to integrate numerically as it is to tabulate the analytical expression. Or one may have a table of numbers to integrate rather than an analytical Equation. Many computers and programmable calculators have internal routines for integration, which one can call upon (at risk) without having any idea how they work. It is assumed that the reader of this chapter, however, wants to be able to carry out a numerical integration without calling upon an existing routine that has been written by somebody else.

There are many different methods of numerical integration, but the one known as Simpson's Rule is easy to program, rapid to perform and usually very accurate. (Thomas Simpson, 1710 - 1761, was an English mathematician, author of *A New Treatise on Fluxions*.)

Suppose we have a function $y(x)$ that we wish to integrate between two limits. We calculate the value of the function at the two limits and halfway between, so we now know three points on the curve. We then fit a parabola to these three points and find the area under that.

In the figure I.1, $y(x)$ is the function we wish to integrate between the limits $x_2 - \delta x$ and $x_2 + \delta x$. In other words, we wish to calculate the area under the curve. y_1, y_2 and y_3 are the values of

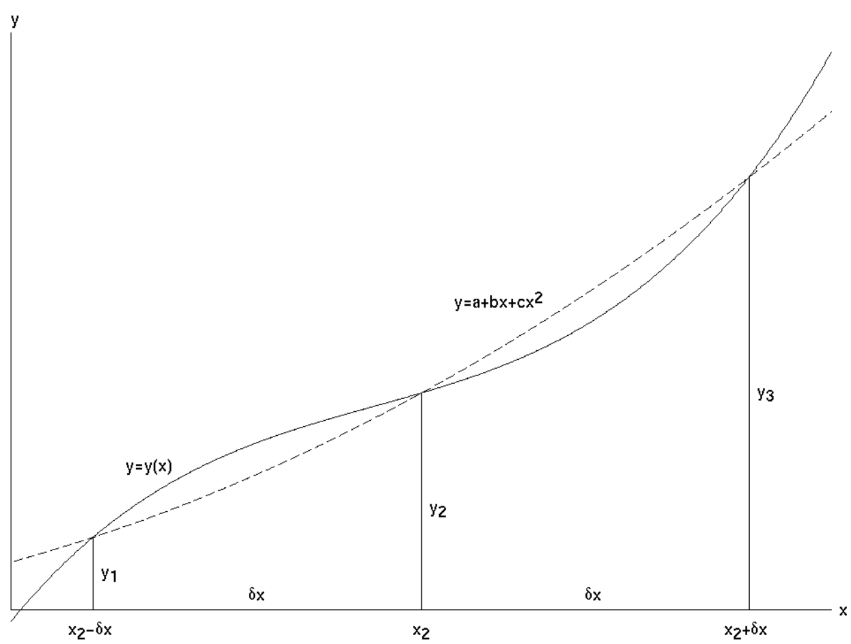


FIGURE I.1 Simpson's Rule gives us the area under the parabola (dashed curve) that passes through three points on the curve $y = y(x)$. This is approximately equal to the area under $y = y(x)$.

the function at $x_2 - \delta_x$, x_2 and $x_2 + \delta_x$, and $y = a + bx + cx^2$ is the parabola passing through the points $(x_2 - \delta x, y_1)$, (x_2, y_2) and $(x_2 + \delta x, y_3)$.

If the parabola is to pass through these three points, we must have

$$y_1 = a + b(x_2 - \delta x) + c(x_2 - \delta x)^2 \tag{1.2.1}$$

$$y_2 = a + bx + cx^2 \tag{1.2.2}$$

$$y_3 = a + b(x_2 + \delta x) + c(x_2 + \delta x)^2 \tag{1.2.3}$$

We can solve these Equations to find the values of a, b and c. These are

$$a=y_2-\frac{x_2(y_3-y_1)}{2\delta x}+\frac{x_2^2(y_3-2y_2+y_1)}{2(\delta x)^2} \tag{1.2.4}$$

$$b=\frac{y_3-y_1}{2\delta x}-\frac{x_2(y_3-2y_2+y_1)}{(\delta x)^2} \tag{1.2.5}$$

$$c=\frac{y_3-2y_2+y_1}{2(\delta x)^2} \tag{1.2.6}$$

Now the area under the parabola (which is taken to be approximately the area under $y(x)$) is

$$\int_{x_2-\delta x}^{x_2+\delta x}\left(a+bx+cx^2\right)dx=2\left[a+bx_2+cx_2^2+\frac{1}{3}c(\delta x)^2\right]\delta x \tag{1.2.7}$$

On substituting the values of a, b and c, we obtain for the area under the parabola

$$\frac{1}{3}(y_1+4y_2+y_3)\delta x \tag{1.2.8}$$

and this is the formula known as Simpson's Rule.

For an example, let us evaluate $\int_0^{\pi/2}\sin x dx$.

We shall evaluate the function at the lower and upper limits and halfway between. Thus

$$\begin{aligned} x&=0, & y&=0 \\ x&=\pi/4, & y&=1/\sqrt{2} \\ x&=\pi/2, & y&=1 \end{aligned}$$

The interval between consecutive values of x is $\delta x=\pi/4$.

Hence Simpson's Rule gives for the area

$$\frac{1}{3}\left(0+\frac{4}{\sqrt{2}}+1\right)\frac{\pi}{4}$$

which, to three significant figures, is 1.00. Graphs of $\sin x$ and $a+bx+cx^2$ are shown in figure $\mathrm{I.2a}$. The values of a, b and c, obtained from the formulas above, are

$$a=0, \quad b=\frac{\sqrt{32}-2}{\pi}, \quad c=\frac{8-\sqrt{128}}{\pi^2}$$

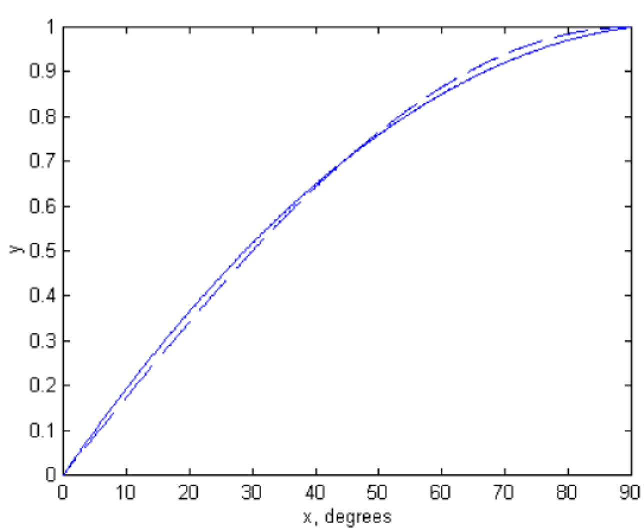


FIGURE I.2a

The result we have just obtained is quite spectacular, and we are not always so lucky. Not all functions can be approximated so well by a parabola. But of course the interval $\delta x = \pi/4$ was ridiculously coarse. In practice we subdivide the interval into numerous very small intervals. For example, consider the integral

$$\int_0^{\pi/4} \cos^{\frac{3}{2}} 2x \sin x \, dx.$$

Let us subdivide the interval 0 to $\pi/4$ into ten intervals of width $\pi/40$ each. We shall evaluate the function at the end points and the nine points between, thus:

x	$\cos^{\frac{3}{2}} x \sin x dx$
0	$y_1 = 0.000\,000\,000$
$\pi/40$	$y_2 = 0.077\,014\,622$
$2\pi/40$	$y_3 = 0.145\,091\,486$
$3\pi/40$	$y_4 = 0.196\,339\,002$
$4\pi/40$	$y_5 = 0.224\,863\,430$
$5\pi/40$	$y_6 = 0.227\,544\,930$
$6\pi/40$	$y_7 = 0.204\,585\,473$
$7\pi/40$	$y_8 = 0.159\,828\,877$
$8\pi/40$	$y_9 = 0.100\,969\,971$
$9\pi/40$	$y_{10} = 0.040\,183\,066$
$10\pi/40$	$y_{11} = 0.000\,000\,000$

The integral from 0 to $2\pi/40$ is $\frac{1}{3}(y_1 + 4y_2 + y_3)\delta x$, δx being the interval $\pi/40$. The integral from $3\pi/40$ to $4\pi/40$ is $\frac{1}{3}(y_3 + 4y_4 + y_5)\delta x$. And so on, until we reach the integral from $8\pi/40$ to $10\pi/40$. When we add all of these up, we obtain for the integral from 0 to $\pi/4$,

$$\frac{1}{3}(y_1 + 4y_2 + 2y_3 + 4y_4 + 2y_5 + \ldots\ldots + 4y_{10} + y_{11})\,\delta x$$

$$= \frac{1}{3}[y_1 + y_{11} + (y_2 + y_4 + y_6 + y_8 + y_{10}) + 2(y_3 + y_5 + y_7 + y_9)]\delta x,$$

which comes to $0.108\,768\,816$

We see that the calculation is rather quick, and it is easily programmable (try it!). But how good is the answer? Is it good to three significant figures? Four? Five?

Since it is fairly easy to program the procedure for a computer, my practice is to subdivide the interval successively into 10, 100, 1000 subintervals, and see whether the result converges. In the present example, with N subintervals, I found the following results:

N	integral
10	$0.108\,768\,816$
100	$0.108\,709\,621$
1000	$0.108\,709\,466$
10000	$0.108\,709\,465$

This shows that, even with a course division into ten intervals, a fairly good result is obtained, but you do have to work for more significant figures. I was using a mainframe computer when I did the calculation with 10000 intervals, and the answer was displayed on my screen in what I would estimate was about one fifth of a second.

There are two more lessons to be learned from this example. One is that sometimes a change of variable will make things very much faster. For example, if one makes one of the (fairly obvious?) trial substitutions $y = \cos x$, $y = \cos 2x$ or $y^2 = \cos 2x$, the integral becomes

$$\int_{1/\sqrt{2}}^{1}(2y^2-1)^{3/2}dy, \quad \int_0^1\sqrt{\frac{y^3}{8(1+y)}}dy \quad \text{or} \quad \int_0^1\frac{y^4}{\sqrt{2(1+y^2)}}dy.$$

Not only is it very much faster to calculate any of these integrands than the original trigonometric expression, but I found the answer $0.108\,709\,465$ by Simpson's rule on the third of these with only 100 intervals rather than $10,000$, the answer appearing on the screen apparently instantaneously. (The first two required a few more intervals.)

To gain about one fifth of a second may appear to be of small moment, but in truth the computation went faster by a factor of several hundred. One sometimes hears of very large computations involving massive amounts of data requiring overnight computer runs of eight hours or so. If the programming speed and efficiency could be increased by a factor of a few hundred, as in this example, the entire computation could be completed in less than a minute.

The other lesson to be learned is that the integral does, after all, have an explicit algebraic form. You should try to find it, not only for integration practice, but to convince yourself that there are indeed occasions when a numerical solution can be found faster than an analytic one! The answer, by the way, is $\frac{\sqrt{18}\ln(1+\sqrt{2})-2}{16}$.

You might now like to perform the following integration numerically, either by hand calculator or by computer.

$$\int_0^2\frac{x^2dx}{\sqrt{2-x}}$$

At first glance, this may look like just another routine exercise, but you will very soon find a small difficulty and wonder what to do about it. The difficulty is that, at the upper limit of integration, the integrand becomes infinite. This sort of difficulty, which is not infrequent, can often be overcome by means of a change of variable. For example, let $x=2\sin^2\theta$, and the integral becomes

$$8\sqrt{2}\int_0^{\pi/2}\sin^5\theta d\theta$$

and the difficulty has gone. The reader should try to integrate this numerically by Simpson's rule, though it may also be noted that it has an exact analytic answer, namely $\sqrt{8192}/15$.

Here is another example. It can be shown that the period of oscillation of a simple pendulum of length l swinging through 90° on either side of the vertical is

$$P=\sqrt{\frac{8l}{g}}\int_0^{\pi/2}\sqrt{\sec\theta}d\theta.$$

As in the previous example, the integrand becomes infinite at the upper limit. I leave it to the reader to find a suitable change of variable such that the integrand is finite at both limits, and then to integrate it numerically. (If you give up, see Section 1.13.) Unlike the last example, this one has no simple analytic solution in terms of elementary functions. It can be written in terms of special functions (elliptic integrals) but they have to be evaluated numerically in any case, so that is of little help. I make the answer

$$P=2.3607\pi\sqrt{\frac{l}{g}}.$$

For another example, consider

$$\int_0^\infty\frac{dx}{x^5\left(e^{1/x}-1\right)}$$

This integral occurs in the theory of blackbody radiation. To help you to visualize the integrand, it and its first derivative are zero at $x=0$ and $x=\infty$ and it reaches a maximum value of 21.201435 at $x=0.201405$. The difficulty this time is the infinite upper limit. But, as in the previous two examples, we can overcome the difficulty by making a change of variable. For example, if we let $x=\tan\theta$, the integral becomes

$$\int_0^{\pi/2} \frac{c^3\left(c^2+1\right) d\theta}{e^c-1}, \text{ where } c = \cot\theta = 1/x.$$

The integrand is zero at both limits and is easily calculable between, and the value of the integral can now be calculated by Simpson's rule in a straightforward way. It also has an exact analytic solution, namely $\pi^4/15$, though it is hard to say whether it is easier to arrive at this by analysis or by numerical integration.

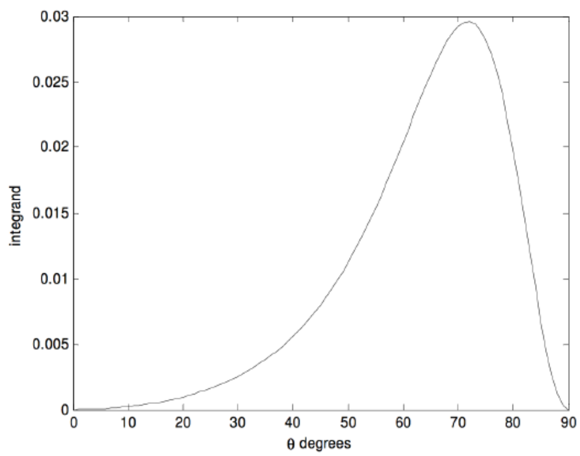
Here’s another:

$$\int_0^\infty \frac{x^2 dx}{(x^2+9)(x^2+4)^2}$$

The immediate difficulty is the infinite upper limit, but that is easily dealt with by making a change of variable: $x = \tan\theta$. The integral then becomes

$$\int_{\theta=0}^{\pi/2} \frac{t(t+1)d\theta}{(t+9)(t+4)^2}$$

in which $t = \tan^2\theta$. The upper limit is now finite, and the integrand is easy to compute - except, perhaps, at the upper limit. However, after some initial hesitation the reader will probably agree that the integrand is zero at the upper limit. The integrand looks like this:



It reaches a maximum of $0.029\,5917$ at $\theta = 71^\circ.789\,962$. Simpson’s rule easily gave me an answer of $0.015\,708$. The integral has an analytic solution (try it) of $\pi/200$.

There are, of course, methods of numerical integration other than Simpson’s rule. I describe one here without proof. I call it “seven-point integration”. It may seem complicated, but once you have successfully programmed it for a computer, you can forget the details, and it is often even faster and more accurate than Simpson’s rule. You evaluate the function at $6n+1$ points, where n is an integer, so that there are $6n$ intervals. If, for example, $n = 4$, you evaluate the function at 25 points, including the lower and upper limits of integration. The integral is then:

$$\int_a^b f(x)dx = 0.3 \times (\Sigma_1 + 2\Sigma_2 + 5\Sigma_3 + 6\Sigma_4)\delta x, \tag{1.2.9}$$

where δx is the size of the interval, and

$$\Sigma_1 = f_1 + f_3 + f_5 + f_9 + f_{11} + f_{15} + f_{17} + f_{21} + f_{23} + f_{25}, \tag{1.2.10}$$

$$\Sigma_2 = f_7 + f_{13} + f_{19}, \tag{1.2.11}$$

$$\Sigma_3 = f_2 + f_6 + f_8 + f_{12} + f_{14} + f_{18} + f_{20} + f_{24} \tag{1.2.12}$$

and

$$\Sigma_4 = f_4 + f_{10} + f_{16} + f_{22}. \tag{1.2.13}$$

Here, of course, $f_1 = f(a)$ and $f_{25} = f(b)$. You can try this on the functions we have already integrated by Simpson's rule, and see whether it is faster.

Let us try one last integration before moving to the next section. Let us try

$$\int_0^{10} \frac{1}{1+8x^3} dx.$$

This can easily (!) be integrated analytically, and you might like to show that it is

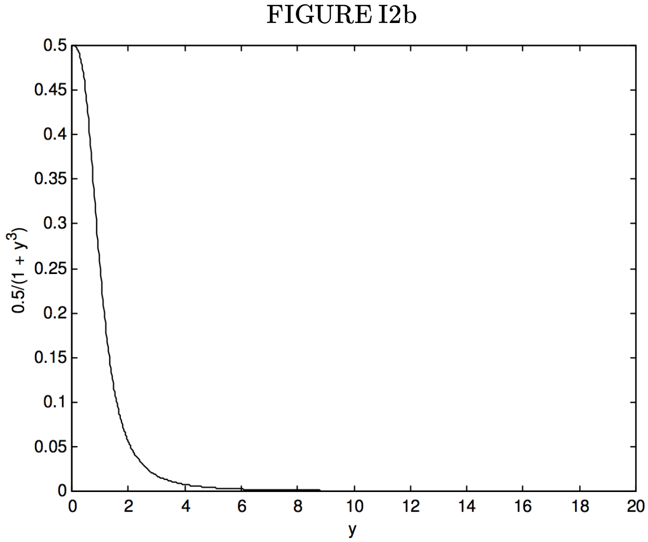
$$\frac{1}{12}\ln\frac{147}{127} + \frac{1}{\sqrt{12}}\tan^{-1}\sqrt{507} + \frac{\pi}{\sqrt{432}} = 0.6039748.$$

However, our purpose in this section is to learn some skills of numerical integration. Using Simpson's rule, I obtained the above answer to seven decimal places with 544 intervals. With seven-point integration, however, I used only 162 intervals to achieve the same precision, a reduction of 70. Either way, the calculation on a fast computer was almost instantaneous. However, had it been a really lengthy integration, the greater efficiency of the seven point integration might have saved hours. It is also worth noting that $x \times x \times x$ is faster to compute than x^3. Also, if we make the substitution y = 2x, the integral becomes

$$\int_0^{20} \frac{0.5}{1+y^3} dy.$$

This reduces the number of multiplications to be done from 489 to 326 – i.e. a further reduction of one third. But we have still not done the best we could do. Let us look at the function $\frac{0.5}{1+y^3}$, in figure I.2b:

FIGURE I2b



We see that beyond $y = 6$, our efforts have been largely wasted. We don't need such fine intervals of integration. I find that I can obtain the same level of precision − i.e. an answer of 0.6039748− using 48 intervals from $y = 0$ to 6 and 24 intervals from $y = 6$ to 20. Thus, by various means we have reduced the number of times that the function had to be evaluated from our original 545 to 72, as well as reducing the number of multiplications each time by a third, a reduction of computing time by 91. This last example shows that it often advantageous to use fine intervals of integration only when the function is rapidly changing (i.e. has a large slope), and to revert to coarser intervals where the function is changing only slowly.

The *Gaussian quadrature* method of numerical integration is described in Sections 1.15 and 1.16.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.3: Quadratic Equations

Any reader of this book will know that the solutions to the quadratic Equation

$$ax^2+bx+c=0 \tag{1.3.1}$$

are

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a} \tag{1.3.2}$$

and will have no difficulty in finding that the solutions to

$$2.9x^2-4.7x+1.7=0$$

are

$$x=1.0758 \text{ or } 0.5449.$$

We are now going to look, largely for fun, at two alternative iterative numerical methods of solving a quadratic Equation. One of them will turn out not to be very good, but the second will turn out to be sufficiently good to merit our serious attention.

In the first method, we re-write the quadratic Equation in the form

$$x=\frac{-\left(ax^2+c\right)}{b}$$

We guess a value for one of the solutions, put the guess in the right hand side, and hence calculate a new value for x. We continue iterating like this until the solution converges.

For example, let us guess that a solution to the Equation $2.9x^2-4.7x+1.7=0$ is $x=0.55$. Successive iterations produce the values

$$\begin{array}{ll}
0.54835 & 0.54501\\
0.54723 & 0.54498\\
0.54648 & 0.54496\\
0.54597 & 0.54495\\
0.54562 & 0.54494\\
0.54539 & 0.54493\\
0.54524 & 0.54493\\
0.54513 & 0.54494\\
0.54506 & 0.54492
\end{array}$$

We did eventually arrive at the correct answer, but it was very slow indeed even though our first guess was so close to the correct answer that we would not have been likely to make such a good first guess accidentally.

Let us try to obtain the second solution, and we shall try a first guess of 1.10, which again is such a good first guess that we would not be likely to arrive at it accidentally. Successive iterations result in

$$\begin{array}{l}
1.10830\\
1.11960\\
1.13515
\end{array}$$

and we are getting further and further from the correct answer!

Let us try a better first guess of 1.05. This time, successive iterations result in

$$\begin{array}{l}
1.04197\\
1.03160\\
1.01834
\end{array}$$

Again, we are getting further and further from the solution.

No more need be said to convince the reader that this is not a good method, so let us try something a little different.

We start with

$$ax^2 + bx = -c \tag{1.3.3}$$

Add ax^2 to each side:

$$2ax^2 + bx = ax^2 - c \tag{1.3.4}$$

or

$$(2ax + b)x = ax^2 - c \tag{1.3.5}$$

Solve for x:

$$x = \frac{ax^2 - c}{2ax + b} \tag{1.3.6}$$

This is just the original Equation written in a slightly rearranged form. Now let us make a guess for x, and iterate as before. This time, however, instead of making a guess so good that we are unlikely to have stumbled upon it, let us make a very stupid first guess, for example $x = 0$. Successive iterations then proceed as follows.

$$\begin{array}{c} 0.00000 \\ 0.36170 \\ 0.51751 \\ 0.54261 \\ 0.54491 \\ 0.54492 \end{array}$$

and the solution converged rapidly in spite of the exceptional stupidity of our first guess. The reader should now try another very stupid first guess to try to arrive at the second solution. I tried $x = 100$, which is very stupid indeed, but I found convergence to the solution 1.0758after just a few iterations.

Even although we already know how to solve a quadratic Equation, there is something intriguing about this. What was the motivation for adding ax^2 to each side of the Equation, and why did the resulting minor rearrangement lead to rapid convergence from a stupid first guess, whereas a simple direct iteration either converged extremely slowly from an impossibly good first guess or did not converge at all?

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.4: The Solution of f(x) = 0

The title of this section is intended to be eye-catching. Some Equations are easy to solve; others seem to be more difficult. In this section, we are going to try to solve any Equation at all of the form $f(x) = 0$ (which covers just about everything!) and we shall in most cases succeed with ease.

Figure I.3 shows a graph of the Equation $y = f(x)$. We have to find the value (or perhaps values) of x such that $f(x) = 0$.

We guess that the answer might be x_g, for example. We calculate $f(x_g)$. It won't be zero, because our guess is wrong. The figure shows our guess x_g, the correct value x, and $f(x_g)$. The tangent of the angle θ is the derivative $f'(x)$, but we cannot calculate the derivative there because we do not yet know x. However, we can calculate $f'(x_g)$, which is close. In any case $\tan\theta$, or $f'(x_g)$, is approximately equal to $f(x_g)/(x_g - x)$, so that

$$x \approx x_g - \frac{f(x_g)}{f'(x_g)} \tag{1.4.1}$$

will be much closer to the true value than our original guess was. We use the new value as our next guess, and keep on iterating until

$$\left|\frac{x_g - x}{x_g}\right|$$

is less than whatever precision we desire. The method is usually extraordinarily fast, even for a wildly inaccurate first guess. The method is known as **Newton-Raphson iteration**. There are some cases where the method will not converge, and stress is often placed on these exceptional cases in mathematical courses, giving the impression that the Newton-Raphson process is of limited applicability. These exceptional cases are, however, often artificially concocted in order to illustrate the exceptions (we do indeed cite some below), and in practice Newton-Raphson is usually the method of choice.

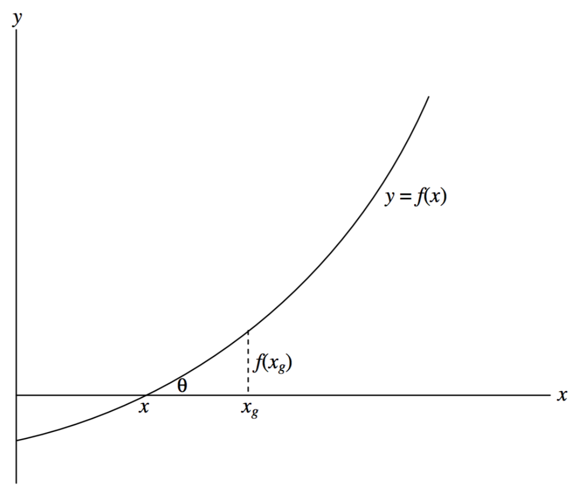


FIGURE I.3

I shall often drop the clumsy subscript g, and shall write the Newton-Raphson scheme as

$$x = x - f(x)/f'(x), \tag{1.4.2}$$

meaning "start with some value of x, calculate the right hand side, and use the result as a new value of x". It may be objected that this is a misuse of the $=$ symbol, and that the above is not really an "Equation", since x cannot equal x minus something. However, when the correct solution for x has been found, it will satisfy $f(x) = 0$, and the above is indeed a perfectly good Equation and a valid use of the $=$ symbol.

Example 1.4.1

Solve the Equation $1/x = \ln x$

We have $f = 1/x - \ln x = 0$

And $f' = -(1+x)/x^2$,

from which $x - f/f'$ becomes, after some simplification,

$$\frac{x[2+x(1-\ln x)]}{1+x},$$

so that the Newton-Raphson iteration is

$$x = \frac{x[2+x(1-\ln x)]}{1+x}.$$

There remains the question as to what should be the first guess. We know (or should know!) that $\ln 1 = 0$ and $\ln 2 = 0.6931$, so the answer must be somewhere between 1 and 2. If we try $x = 1.5$, successive iterations are

$$\begin{gathered} 1.735\,081\,403 \\ 1.762\,915\,391 \\ 1.763\,222\,798 \\ 1.763\,222\,834 \\ 1.763\,222\,835 \end{gathered}$$

This converged quickly from a fairly good first guess of 1.5. Very often the Newton-Raphson iteration will converge, even rapidly, from a very stupid first guess, but in this particular example there are limits to stupidity, and the reader might like to prove that, in order to achieve convergence, the first guess must be in the range

$$0 < x < 4.319\,136\,566$$

Example 1.4.2

Solve the unlikely Equation $\sin x = \ln x$

We have $f = \sin x - \ln x$ and $f' = \cos x - 1/x$,

and after some simplification the Newton-Raphson iteration becomes

$$x = x\left[1 + \frac{\ln x - \sin x}{x\cos x - 1}\right].$$

Graphs of $\sin x$ and $\ln x$ will provide a first guess, but in lieu of that and without having much idea of what the answer might be, we could try a fairly stupid $x = 1$. Subsequent iterations produce

$$\begin{gathered} 2.830\,487\,722 \\ 2.267\,902\,211 \\ 2.219\,744\,452 \\ 2.219\,107\,263 \\ 2.219\,107\,149 \\ 2.219\,107\,149 \end{gathered}$$

Example 1.4.3

Solve the Equation $x^2 = a$ (A new way of finding square roots!)

$$f = x^2 - a, \quad f' = 2x.$$

After a little simplification, the Newton-Raphson process becomes

$$x = \frac{x^2 + a}{2x}.$$

For example, what is the square root of 10? Guess 3. Subsequent iterations are

$$3.166\,666\,667$$
$$3.162\,280\,702$$
$$3.162\,277\,661$$
$$3.162\,277\,661$$

Example 1.4.4

Solve the Equation $ax^2+bx+c=0$ (A new way of solving quadratic Equations!)

$$f=ax^2+bx+c=0,$$

$$f'=2ax+b.$$

Newton-Raphson:

$$x=x-\frac{ax^2+bx+c}{2ax+b},$$

which becomes, after simplification,

$$x=\frac{ax^2-c}{2ax+b}.$$

This is just the iteration given in the previous section, on the solution of quadratic Equations, and it shows why the previous method converged so rapidly and also how I really arrived at the Equation (which was via the Newton-Raphson process, and not by arbitrarily adding ax^2 to both sides!)

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.5: The Solution of Polynomial Equations

The Newton-Raphson method is very suitable for the solution of polynomial Equations, for example for the solution of a quintic Equation:

$$a_0+a_1x+a_2x^2+a_3x^3+a_4x^4+a_5x^5=0. \tag{1.5.1}$$

Before illustrating the method, it should be pointed out that, even though it may look inelegant in print, in order to evaluate a polynomial expression numerically it is far easier and quicker to nest the parentheses and write the polynomial in the form

$$a_0+x(a_1+x(a_2+x(a_3+x(a_4+xa_5)))). \tag{1.5.2}$$

Working from the inside out, we see that the process is a multiplication followed by an addition, repeated over and over. This is very easy whether the calculation is done by computer, by calculator, or in one's head.

For example, evaluate the following expression in your head, for $x=4$:

$$2-7x+2x^2-8x^3-2x^4+3x^5.$$

You couldn't? But now evaluate the following expression in your head for $x=4$ and see how (relatively) easy it is:

$$2+x(-7+x(2+x(-8+x(-2+3x)))).$$

Fortran

As an example of how efficient the nested parentheses are in a computer program, here is a FORTRAN program for evaluating a fifth degree polynomial. It is assumed that the value of x has been defined in a FORTRAN variable called X, and that the six coefficients $a_0, a_1, \ldots a_5$ have been stored in a vector as $\mathrm{A}(1),\ \mathrm{A}(2), \ldots\ \mathrm{A}(6)$.

```
      Y = 0.
      DO1I = 1, 5
1     Y = (Y + A(7-I))*X
      Y = Y + A(1)
```

The calculation is finished!

We return now to the solution of

$$f(x)=a_0+a_1x+a_2x^2+a_3x^3+a_4x^4+a_5x^5=0. \tag{1.5.3}$$

We have

$$f'(x)=a_1+2a_2x+3a_3x^2+4a_4x^3+5a_5x^4. \tag{1.5.4}$$

Now

$$x=x-f/f', \tag{1.5.5}$$

and after simplification,

$$x=\frac{-a_0+x^2(a_2+x(2a_3+x(3a_4+4a_5x)))}{a_1+x(2a_2+x(3a_3+x(4a_4+5a_5x)))}, \tag{1.5.6}$$

which is now ready for numerical iteration.

For example, let us solve

$$205+111x+4x^2-31x^3-10x^4+3x^5=0 \tag{1.5.7}$$

A reasonable first guess could be obtained by drawing a graph of this function to see where it crosses the x-axis, but, truth to tell, the Newton-Raphson process usually works so well that one need spend little time on a first guess; just use the first number that comes into your head, for example, $x=0$. Subsequent iterations then go

$$\begin{array}{l} 1.846\,847 \\ -1.983\,713 \\ -1.967\,392 \\ -1.967\,111 \\ -1.967\,110 \end{array}$$

A question that remains is: How many solutions are there? The general answer is that an nth degree polynomial Equation has n solutions. This statement needs to be qualified a little. For example, the solutions need not be real. The solutions may be imaginary, as they are, for example, in the Equation

$$1 + x^2 = 0 \tag{1.5.8}$$

or complex, as they are, for example, in the Equation

$$1 + x + x^2 = 0. \tag{1.5.9}$$

If the solutions are real they may not be distinct. For example, the Equation

$$1 - 2x + x^2 = 0 \tag{1.5.10}$$

has two solutions at $x = 1$, and the reader may be forgiven for thinking that this somewhat stretches the meaning of "two solutions". However, if one includes complex roots and repeated real roots, it is then always true that an nth degree polynomial has n solutions. The five solutions of the quintic Equation we solved above, for example, are

$$\begin{array}{lll} 4.947\,845 & & \\ 2.340\,216 & & \\ -1.967\,110 & & \\ -0.993\,808 & + & 1.418\,597i \\ -0.993\,808 & - & 1.418\,597i \end{array}$$

Can one tell in advance how many real roots a polynomial Equation has? The most certain way to tell is to plot a graph of the polynomial function and see how many times it crosses the x-axis. However, it is possible to a limited extent to determine in advance how many real roots there are. The following "rules" may help. Some will be fairly obvious; others require proof.

The number of real roots of a polynomial of odd degree is odd. Thus a quintic Equation can have one, three or five real roots. Not all of these roots need be distinct, however, so this is of limited help. Nevertheless a polynomial of odd degree always has at least one real root. The number of real roots of an Equation of even degree is even - but the roots need not all be distinct, and the number of real roots could be zero.

An upper limit to the number of real roots can be determined by examining the signs of the coefficients. For example, consider again the Equation

$$205 + 111x + 4x^2 - 31x^3 - 10x^4 + 3x^5 = 0. \tag{1.5.11}$$

The signs of the coefficients, written in order starting with a_0, are

$$+++--+$$

Run your eye along this list, and count the number of times there is a change of sign. The sign changes twice. This tells us that there are not more than two positive real roots. (If one of the coefficients in a polynomial Equation is zero, i.e. if one of the terms is "missing", this does not count as a change of sign.)

Now change the signs of all coefficients of odd powers of x:

$$+-++--$$

This time there are three changes of sign. This tells us that there are not more than three negative real roots.

In other words, the number of changes of sign in $f(x)$ gives us an upper limit to the number of positive real roots, and the number of changes of sign in $f(-x)$ gives us an upper limit to the number of negative real roots.

One last "rule" is that complex roots occur in conjugate pairs. In our particular example, these rules tell us that there are not more than two positive real roots, and not more than three negative real roots. Since the degree of the polynomial is odd, there is at least one real root, though we cannot tell whether it is positive or negative.

In fact the particular Equation, as we have seen, has two positive real roots, one negative real root, and two conjugate complex roots.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.6: Failure of the Newton-Raphson Method

This section is written reluctantly, for fear it may give the impression that the Newton-Raphson method frequently fails and is of limited usefulness. This is not the case; in nearly all cases encountered in practice it is very rapid and does not require a particularly good first guess. Nevertheless for completeness it should be pointed out that there are rare occasions when the method either fails or converges rather slowly.

One example is the quintic Equation that we have just encountered:

$$205+111x+4x^2-31x^3-10x^4+5x^5=0 \tag{1.6.1}$$

When we chose $x=0$ as our first guess, we reached a solution fairly quickly. If we had chosen $x=1$, we would not have been so lucky, for the first iteration would have taken us to -281, a very long way from any of the real solutions. Repeated iteration will eventually take us to the correct solution, but only after many iterations. This is not a typical situation, and usually almost any guess will do.

Another example of an Equation that gives some difficulty is

$$x=\tan x, \tag{1.6.2}$$

an Equation that occurs in the theory of single-slit diffraction.

We have

$$f(x)=x-\tan x=0 \tag{1.6.3}$$

and

$$f'(x)=1-\sec^2 x=-\tan^2 x. \tag{1.6.4}$$

The Newton-Raphson process takes the form

$$x=x+\frac{x-\tan x}{\tan^2 x}. \tag{1.6.5}$$

The solution is $x=4.493\,409$, but in order to achieve this the first guess must be between 4.3 and 4.7. This again is unusual, and in most cases almost any reasonable first guess results in rapid convergence.

The Equation

$$1-4x+6x^2-4x^3+x^4=0 \tag{1.6.6}$$

is an obvious candidate for difficulties. The four identical solutions are $x=1$, but at $x=1$ not only is $f(x)$ zero, but so is $f'(x)$. As the solution $x=1$ is approached, convergence becomes very slow, but eventually the computer or calculator will record an error message as it attempts to divide by the nearly zero $f'(x)$.

I mention just one last example very briefly. When discussing orbits, we shall encounter an Equation known as Kepler's Equation. The Newton-Raphson process almost always solves Kepler's Equation with spectacular speed, even with a very poor first guess. However, there are some very rare occasions. almost never encountered in practice, where the method fails. We shall discuss this Equation in Chapter 9.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.7: Simultaneous Linear Equations, N = n

Consider the Equations

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + a_{15}x_5 = b_1 \tag{1.7.1}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 + a_{25}x_5 = b_2 \tag{1.7.2}$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 + a_{35}x_5 = b_3 \tag{1.7.3}$$

$$a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + a_{44}x_4 + a_{45}x_5 = b_4 \tag{1.7.4}$$

$$a_{51}x_1 + a_{52}x_2 + a_{53}x_3 + a_{54}x_4 + a_{55}x_5 = b_5 \tag{1.7.5}$$

There are two well-known methods of solving these Equations. One of these is called Cramer's Rule. Let D be the determinant of the coefficients. Let D_i be the determinant obtained by substituting the column vector of the constants $b_1,\ b_2,\ b_3,\ b_4,\ b_5$ for the ith column in D. Then the solutions are

$$x_i = D_i/D \tag{1.7.6}$$

This is an interesting theorem in the theory of determinants. It should be made clear, however, that, when it comes to the practical numerical solution of a set of linear Equations that may be encountered in practice, this is probably the most laborious and longest method ever devised in the history of mathematics.

The second well-known method is to write the Equations in matrix form:

$$\mathbb{A}\mathbf{x} = \mathbf{b} \tag{1.7.7}$$

Here $\mathbb{A}$ is the matrix of the coefficients, $\mathbf{x}$ is the column vector of unknowns, and $\mathbf{b}$ is the column vector of the constants. The solutions are then given by

$$\mathbf{x} = \mathbb{A}^{-1}\mathbf{b}, \tag{1.7.8}$$

where $\mathbb{A}^{-1}$ is the inverse or reciprocal of $\mathbb{A}$. Thus the problem reduces to inverting a matrix. Now inverting a matrix is notoriously labour-intensive, and, while the method is not quite so long as Cramer's Rule, it is still far too long for practical purposes.

How, then, should a system of linear Equations be solved?

Consider the Equations

$$7x - 2y = 24 \tag{1.7.8}$$

$$3x + 9y = 30 \tag{1.7.9}$$

Few would have any hesitation in multiplying the first Equation by 3, the second Equation by 7, and subtracting. This is what we were all taught in our younger days, but few realize that this remains, in spite of knowledge of determinants and matrices, the fastest and most efficient method of solving simultaneous linear Equations. Let us see how it works with a system of several Equations in several unknowns.

Consider the Equations

$$9x_1 - 9x_2 + 8x_3 - 6x_4 + 4x_5 = -9 \tag{1.7.10}$$

$$5x_1 - x_2 + 6x_3 + x_4 + 5x_5 = 58 \tag{1.7.11}$$

$$2x_1 + 4x_2 - 5x_3 - 6x_4 + 7x_5 = -1 \tag{1.7.12}$$

$$2x_1 + 3x_2 - 8x_3 - 5x_4 - 2x_5 = -49 \tag{1.7.13}$$

$$8x_1 - 5x_2 + 7x_3 + x_4 + 5x_5 = 42 \tag{1.7.14}$$

We first eliminate x_1 from the Equations, leaving four Equations in four unknowns. Then we eliminate x_2, leaving three Equations in three unknowns. Then x_3, and then x_4, finally leaving a single Equation in one unknown. The following table shows how it is done.

In columns 2 to 5 are listed the coefficients of x_1, x_2, x_3, x_4 and x_5, and in column 6 are the constant terms on the right hand side of the Equations. Thus columns 2 to 6 of the first five rows are just the original Equations. Column 7 is the sum of the numbers in columns 2 to 6, and this is a most important column. The boldface numbers in column 1 are merely labels.

Lines 6 to 9 show the elimination of x_1. Line 6 shows the elimination of x_1 from lines 1 and 2 by multiplying line 2 by 9 and line 1 by 5 and subtracting. The operation performed is recorded in column 1. In line 7, x_1 is eliminated from Equations 1 and 3 and so on.

	x_1	x_2	x_3	x_4	x_5	b	$\sum$
1	9	−9	8	−6	4	−9	−3
2	5	−1	6	1	5	58	74
3	2	4	−5	−6	7	−1	1
4	2	3	−8	−5	−2	−49	−59
5	8	−5	7	1	5	42	58
$\mathbf{6} = 9\times\mathbf{2} - 5\times\mathbf{1}$		36	14	39	25	567	681
$\mathbf{7} = 2\times\mathbf{1} - 9\times\mathbf{3}$		−54	61	42	−55	−9	−15
$\mathbf{8} = \mathbf{3} - \mathbf{4}$		1	3	−1	9	48	60
$\mathbf{9} = 4\times\mathbf{3} - \mathbf{5}$		21	−27	−25	23	−46	−54
$\mathbf{10} = 3\times\mathbf{6} + 2\times\mathbf{7}$			164	201	−35	1 683	2 013
$\mathbf{11} = \mathbf{6} - 36\times\mathbf{8}$			−94	75	−299	−1 161	−1 479
$\mathbf{12} = 7\times\mathbf{6} - 12\times\mathbf{9}$			422	573	−101	4 521	5 415
$\mathbf{13} = 47\times\mathbf{10} + 82\times\mathbf{11}$				15 597	−26 163	−16 101	−26 667
$\mathbf{14} = 211\times\mathbf{11} + 47\times\mathbf{12}$				42 756	−67 836	−32 484	−57 654
$\mathbf{15} = 5199\times\mathbf{14} - 14252\times\mathbf{13}$					20 195 712	60 587 136	80 782 848

(1.7.15)

The purpose of Σ ? This column is of great importance. Whatever operation is performed on the previous columns is also performed on Σ, and Σ must remain the sum of the previous columns. If it does not, then an arithmetic mistake has been made, and it is immediately detected. There is nothing more disheartening to discover at the very end of a calculation that a mistake has been made and that one has no idea where the mistake occurred. Searching for mistakes takes far longer than the original calculation. The Σ-column enables one to detect and correct a mistake as soon as it has been made.

We eventually reach line 15, which is

$$20\,195\,712 x_5 = 60\,587\,136, \tag{1.7.16}$$

from which

$$x_5 = 3. \tag{1.7.17}$$

x_4 can now easily be found from either or both of lines 13 and 14, x_3 can be found from any or all of lines 10, 11 and 12, and so on. When the calculation is complete, the answers should be checked by substitution in the original Equations (or in the sum of the five Equations). For the record, the solutions are $x_1 = 2,\ x_2 = 7,\ x_3 = 6,\ x_4 = 4$ and $x_5 = 3$.

Of course, if you have only two simultaneous Equations to solve, it is easy to write down explicit algebraic expressions for the solutions, and that may be the fastest and most efficient way of doing it. Thus, if

$$a_{11}x + a_{12}y = b_1 \tag{1.7.9}$$

and

$$a_{21}x + a_{22}y = b_2, \tag{1.7.10}$$

the solutions are

$$x = c(b_1 a_{22} - b_2 a_{12}) \tag{1.7.11}$$

and

$$y = c(b_2 a_{11} - b_1 a_{21}), \tag{1.7.12}$$

where

$$c = 1/(a_{11}a_{22} - a_{12}a_{21}). \tag{1.7.18}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.8: Simultaneous Linear Equations, N > n

Consider the following Equations

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + b_1 = 0 \tag{1.8.1}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + b_2 = 0 \tag{1.8.2}$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + b_3 = 0 \tag{1.8.3}$$

$$a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + b_4 = 0 \tag{1.8.4}$$

$$a_{51}x_1 + a_{52}x_2 + a_{53}x_3 + b_5 = 0 \tag{1.8.5}$$

Here we have five Equations in only three unknowns, and there is no solution that will satisfy all five Equations exactly. We refer to these Equations as the *Equations of condition*. The problem is to find the set of values of x_1, x_2 and x_3 that, while not satisfying any one of the Equations exactly, will come closest to satisfying all of them with as small an error as possible. The problem was well stated by Carl Friedrich Gauss in his famous *Theoria Motus*. In 1801 Gauss was faced with the problem of calculating the orbit of the newly discovered minor planet Ceres. The problem was to calculate the six elements of the planetary orbit, and he was faced with solving more than six Equations for six unknowns. In the course of this, he invented the method of least squares. It is hardly possible to describe the nature of the problem more clearly than did Gauss himself:

"...as all our observations, on account of the imperfection of the instruments and the senses, are only approximations to the truth, an orbit based only on the six absolutely necessary data may still be liable to considerable errors. In order to diminish these as much as possible, and thus to reach the greatest precision attainable, no other method will be given except to accumulate the greatest number of the most perfect observations, and to adjust the elements, not so as to satisfy this or that set of observations with absolute exactness, but so as to agree with all in the best possible manner."

If we can find some set of values of x_1, x_2 and x_3 that satisfy our five Equations fairly closely, but without necessarily satisfying any one of them exactly, we shall find that, when these values are substituted into the left hand sides of the Equations, the right hand sides will not be exactly zero, but will be a small number known as the residual, R.

Thus:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + b_1 = R_1 \tag{1.8.6}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + b_2 = R_2 \tag{1.8.7}$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + b_3 = R_3 \tag{1.8.8}$$

$$a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + b_4 = R_4 \tag{1.8.9}$$

$$a_{51}x_1 + a_{52}x_2 + a_{53}x_3 + b_5 = R_5 \tag{1.8.10}$$

Gauss proposed a "best" set of values such that, when substituted in the Equations, gives rise to a set of residuals such that the sum of the squares of the residuals is least. (It would in principle be possible to find a set of solutions that minimized the sum of the absolute values of the residuals, rather than their squares. It turns out that the analysis and the calculation involved is a good deal more difficult than minimizing the sum of the squares, with no very obvious advantage.) Let S be the sum of the squares of the residuals for a given set of values of x_1, x_2 and x_3. If any one of the x-values is changed, S will change - unless S is a minimum, in which case the derivative of S with respect to each variable is zero. The three Equations

$$\frac{\partial S}{\partial x_1} = 0, \quad \frac{\partial S}{\partial x_2} = 0, \quad \frac{\partial S}{\partial x_3} = 0 \tag{1.8.11}$$

express the conditions that the sum of the squares of the residuals is least with respect to each of the variables, and these three Equations are called the *normal Equations*. If the reader will write out the value of S in full in terms of the variables x_1, x_2 and x_3, he or she will find, by differentiation of S with respect to x_1, x_3 and x_3 in turn, that the three normal Equations are

$$A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + B_1 = 0 \tag{1.8.12}$$

$$A_{12}x_1 + A_{22}x_2 + A_{23}x_3 + B_2 = 0 \tag{1.8.13}$$

$$A_{13}x_1 + A_{23}x_2 + A_{33}x_3 + B_3 = 0 \tag{1.8.14}$$

where

$$A_{11} = \sum a_{i1}^2, \quad A_{12} = \sum a_{i1}a_{i2}, \quad A_{13} = \sum a_{i1}a_{i3}, \quad B_1 = \sum a_{i1}b_i, \tag{1.8.15}$$

$$A_{22} = \sum a_{i2}^2, \quad A_{23} = \sum a_{i2}a_{i3}, \quad B_2 = \sum a_{i2}b_i, \tag{1.8.16}$$

$$A_{33} = \sum a_{i3}^2, \quad B_3 = \sum a_{i3}b_i, \tag{1.8.17}$$

and where each sum is from $i = 1$ to $i = 5$.

These three normal Equations, when solved for the three unknowns x_1, x_2 and x_3, will give the three values that will result in the lowest sum of the squares of the residuals of the original five Equations of condition.

Let us look at a numerical example, in which we show the running checks that are made in order to detect mistakes as they are made. Suppose the Equations of condition to be solved are

$$7x_1 - 6x_2 + 8x_3 - 15 = 0 \quad -6$$

$$3x_1 + 5x_2 - 2x_3 - 27 = 0 \quad -21$$

$$2x_1 - 2x_2 + 7x_3 - 20 = 0 \quad -13$$

$$4x_1 + 2x_2 - 5x_3 - 2 = 0 \quad -1$$

$$9x_1 - 8x_2 + 7x_3 - 5 = 0 \quad 3$$

$$-108 \quad -69 \quad -71$$

The column of numbers to the right of the Equations is the sum of the coefficients (including the constant term). Let us call these numbers s_1, s_2, s_3, s_4, s_5.

The three numbers below the Equations are $\sum a_{i1}s_i, \quad \sum a_{i2}s_i, \quad \sum a_{i3}s_i$

Set up the normal Equations:

$$159x_1 - 95x_2 + 107x_3 - 279 = 0 \quad -108$$

$$-95x_1 + 133x_2 - 138x_3 + 31 = 0 \quad -69$$

$$107x_1 - 138x_2 + 191x_3 - 231 = 0 \quad -71$$

The column of numbers to the right of the normal Equations is the sum of the coefficients (including the constant term). These numbers are equal to the row of numbers below the Equations of condition, and serve as a check that we have correctly set up the normal Equations. The solutions to the normal Equations are

$$x_1 = 2.474 \quad x_2 = 5.397 \quad x_3 = 3.723$$

and these are the numbers that satisfy the Equations of condition such that the sum of the squares of the residuals is a minimum.

I am going to suggest here that you write a computer program, in the language of your choice, for finding the least squares solutions for N Equations in n unknowns. You are going to need such a program over and over again in the future – not least when you come to Section 1.12 of this chapter!.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.9: Nonlinear Simultaneous Equations

We consider two simultaneous Equations of the form

$$f(x,\ y)=0, \tag{1.9.1}$$

$$g(x,\ y)=0 \tag{1.9.2}$$

in which the Equations are not linear.

As an example, let us solve the Equations

$$x^2=\frac{a}{b-\cos y} \tag{1.9.3}$$

$$x^3-x^2=\frac{a(y-\sin y\cos y)}{\sin^3 y}, \tag{1.9.4}$$

in which a and b are constants whose values are assumed given in any particular case.

This may seem like an artificially contrived pair of Equations, but in fact a pair of Equations like this does appear in orbital theory.

We suggest here two methods of solving the Equations.

In the first, we note that in fact x can be eliminated from the two Equations to yield a single Equation in y:

$$F(y)=aR^3-R^2-2SR-S^2=0, \tag{1.9.5}$$

where

$$R=1/(b-\cos y) \tag{1.9.5a}$$

and

$$S=(y-\sin y\cos y)/\sin^3 y. \tag{1.9.5b}$$

This can be solved by the usual Newton-Raphson method, which is repeated application of $y=y-F/F'$. The derivative of F with respect to y is

$$F'=3aR^2R'-2RR'-2(S'R+SR')-2SS' \tag{1.9.6}$$

where

$$R'=-\frac{\sin y}{(b-\cos y)^2} \tag{1.9.6a}$$

and

$$S'=\frac{\sin y(1-\cos 2y)-3\cos y(y-\frac{1}{2}\sin 2y)}{\sin^4 y} \tag{1.9.6b}$$

In spite of what might appear at first glance to be some quite complicated Equations, it will be found that the Newton-Raphson process, $y=y-F/F'$, is quite straightforward to program, although, for computational purposes, F and F' are better written as

$$F=-S^2+R(-2S+R(-1+aR)), \tag{1.9.7a}$$

and

$$F'=3aR^2R'-2(R+S)(R'+S') \tag{1.9.7b}$$

Let us look at a particular example, say with $a=36$ and $b=4$. We must of course, make a first guess. In the orbital application, described in Chapter 13, we suggest a first guess. In the present case, with $a=36$ and $b=4$, one way would be to plot graphs of Equations 1.9.3 and 1.9.4 and see where they intersect. We have done this in Figure 1.4, from which we see that y must be close to 0.6.

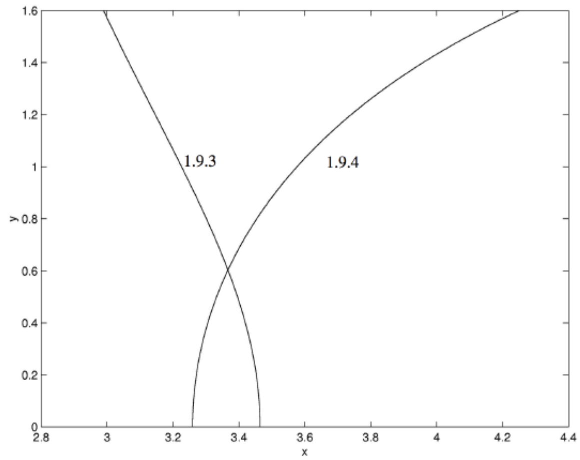


FIGURE 1.4
Equations 1.9.3 and 1.9.4 with $a = 36$ and $b = 4$.

With a first guess of $y = 0.6$, convergence to $y = 0.60292$ is reached in two iterations, and either of the two original Equations then gives $x = 3.3666$.

We were lucky in this case in that we found we were able to eliminate one of the variables and so reduce the problem to a single Equation on one unknown. However, there will be occasions where elimination of one of the unknowns may be considerably more difficult or, in the case of two simultaneous transcendental Equations, impossible by algebraic means. The following iterative method, an extension of the Newton-Raphson technique, can nearly always be used. We describe it for two Equations in two unknowns, but it can easily be extended to n Equations in n unknowns.

The Equations to be solved are

$$f(x,\ y) = 0 \tag{1.9.8}$$

$$g(x,\ y) = 0. \tag{1.9.9}$$

As with the solution of a single Equation, it is first necessary to guess at the solutions. This might be done in some cases by graphical methods. However, very often, as is common with the Newton-Raphson method, convergence is rapid even when the first guess is very wrong.

Suppose the initial guesses are $x + h$, $y + k$, where x, y are the correct solutions, and h and k are the errors of our guess. From a first-order Taylor expansion (or from common sense, if the Taylor expansion is forgotten),

$$f(x+h, y+k) \approx f(x,y) + hf_x + kf_y. \tag{1.9.10}$$

Here f_x and f_y are the partial derivatives and of course $f(x,\ y) = 0$. The same considerations apply to the second Equation, so we arrive at the two linear Equations in the errors h, k:

$$f_x h + f_y k = f, \tag{1.9.11}$$

$$g_x h + g_y k = g. \tag{1.9.12}$$

These can be solved for h and k:

$$h = \frac{g_y f - f_y g}{f_x g_y = f_y g_x}, \tag{1.9.13}$$

$$k = \frac{f_x g - g_x f}{f_x g_y - f_y g_x}. \tag{1.9.14}$$

These values of h and k are then subtracted from the first guess to obtain a better guess. The process is repeated until the changes in x and y are as small as desired for the particular application. It is easy to set up a computer program for solving any two Equations; all that will change from one pair of Equations to another are the definitions of the functions f and g and their partial derivatives.

In the case of our example, we have

$$f=x^2-\frac{a}{b-\cos y} \tag{1.9.15}$$

$$g=x^3-x^2-\frac{a(y-\sin y\cos y)}{\sin^3 y} \tag{1.9.16}$$

$$f_x=2x \tag{1.9.17}$$

$$f_y=\frac{a\sin y}{(b-\cos y)^2} \tag{1.9.18}$$

$$g_x=x(3x-2) \tag{1.9.19}$$

$$g_y=\frac{a[3(y-\sin y\cos y)\cos y-2\sin^3 y]}{\sin^4 y} \tag{1.9.20}$$

In the particular case where $a=36$ and $b=4$, we can start with a first guess (from the graph - Figure I.4) of $y=0.6$ and hence $x=3.3$. Convergence to one part in a million is reached in three iterations, the solutions being $x=3.3666$, $y=0.60292$.

A simple application of these considerations arises if you have to solve a polynomial Equation $f(z)=0$, where there are no real roots, and all solutions for z are complex. You then merely write $z=x+iy$ and substitute this in the polynomial Equation. Then equate the real and imaginary parts separately, to obtain two Equations of the form

$$R(x,\ y)=0 \tag{1.9.21}$$

$$I(x,\ y)=0 \tag{1.9.22}$$

and solve them for x and y. For example, find the roots of the Equation

$$z^4-5z+6=0. \tag{1.9.23}$$

It will soon be found that we have to solve

$$R(x,\ y)=x^4-6x^2y^2+y^4-5x+6=0 \tag{1.9.24}$$

$$I(x,\ y)=4x^3-4xy^2-5=0 \tag{1.9.25}$$

It will have been observed that, in order to obtain the last Equation, we have divided through by y, which is permissible, since we know z to be complex. We also note that y now occurs only as y^2, so it will simplify things if we let $y^2=Y$, and then solve the Equations

$$f(x,Y)=x^4-6x^2Y+Y^2-5x+6=0 \tag{1.9.26}$$

$$g(x,Y)=4x^3-4xY-5=0 \tag{1.9.27}$$

It is then easy to solve either of these for Y as a function of x and hence to graph the two functions (figure I.5):

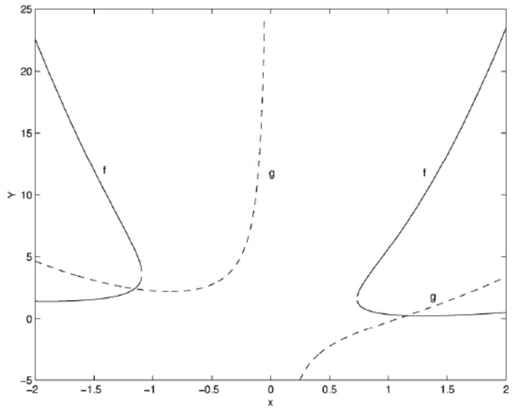


FIGURE I.5

This enables us to make a first guess for the solutions, namely

$$x = -1.2, \quad Y = 2.4$$

and

$$x = +1.2, \quad Y = 0.25818$$

We can then refine the solutions by the extended Newton-Raphson technique to obtain

$$x = -1.15697, \quad Y = 2.41899$$

$$x = +1.15697, \quad Y = 0.25818$$

so the four solutions to the original Equation are

$$z = -1.15697 \pm 1.55531i$$

$$z = 1.15697 \pm 0.50812i$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.10: 1.10- Besselian Interpolation

In the days before the widespread use of high-speed computers, extensive use was commonly made of printed tables of the common mathematical functions. For example, a table of the Bessel function $J_0(x)$ would indicate

$$J_0(1.7) = 0.397\,984\,859$$

$$J_0(1.8) = 0.339\,986\,411$$

If one wanted the Bessel function for $x = 1.762$, one would have to interpolate between the tabulated values.

Today it would be easier simply to calculate the Bessel function for any particular desired value of the argument x, and there is less need today for printed tables or to know how to interpolate. Indeed, most computing systems today have internal routines that will enable one to calculate the commoner functions such as Bessel functions even if one has only a hazy notion of what a Bessel function is.

The need has not entirely passed, however. For example, in orbital calculations, we often need the geocentric coordinates of the Sun. These are not trivial for the nonspecialist to compute, and it may be easier to look them up in *The Astronomical Almanac*, where it is tabulated for every day of the year, such as, for example, July 14 and July 15. But, if one needs y for July 14.395, how does one interpolate?

In an ideal world, a tabulated function would be tabulated at sufficiently fine intervals so that linear interpolation between two tabulated values would be adequate to return the function to the same number of significant figures as the tabulated points. The world is not perfect, however, and to achieve such perfection, the tabulation interval would have to change as the function changed more or less rapidly. We need to know, therefore, how to do nonlinear interpolation.

Suppose a function $y(x)$ is tabulated at $x = x_1$ and $x = x_2$, the interval $x_2 - x_1$ being δx. If one wishes to find the value of y at $x + \theta\delta x$, linear interpolation gives

$$y(x_1 + \theta\Delta x) = y_1 + \theta(y_2 - y_1) = \theta y_2 + (1-\theta)y_1, \tag{1.10.1}$$

where $y_1 = y(x_1)$ and $y_2 = y(x_2)$. Here it is assumed that is a fraction between 0 and 1; if θ is outside this range (that is negative, or greater than 1) we are extrapolating, not interpolating, and that is always a dangerous thing to do.

Let us now look at the situation where linear interpolation is not good enough. Suppose that a function $y(x)$ is tabulated for four points x_1, x_2, x_3, x_4 of the argument x, the corresponding values of the function being y_1, y_2, y_3, y_4. We wish to evaluate y for $x = x_2 + \theta\delta x$, where δx is the interval $x_2 - x_1$ or $x_3 - x_2$ or $x_4 - x_3$. The situation is illustrated in figure I.6A.

A possible approach would be to fit a polynomial to the four adjacent points:

$$y = a + bx + cx^2 + dx^3. \tag{1.10.1}$$

We write down this Equation for the four adjacent tabulated points and solve for the coefficients, and hence we can evaluate the function for any value of x that we like in the interval between x_1 and x_4. Unfortunately, this might well involve more computational effort than evaluating the original function itself.

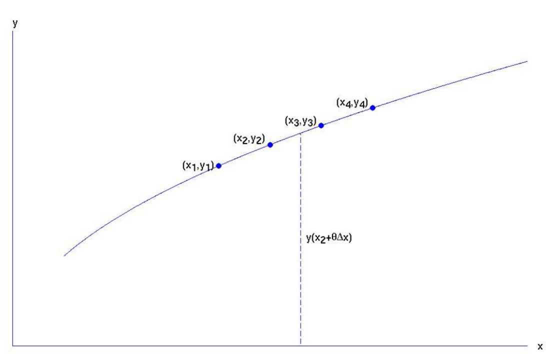
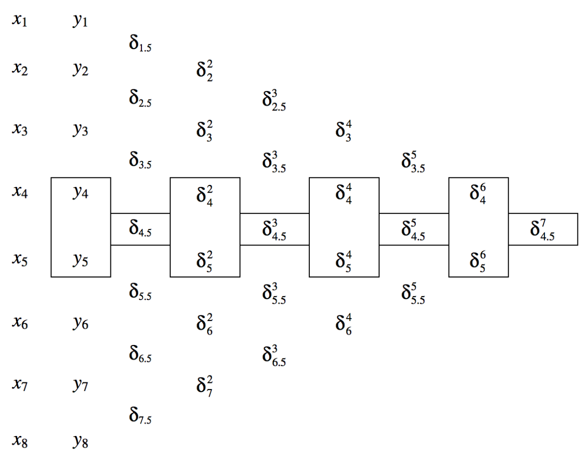


FIGURE I.6A

The problem has been solved in a convenient fashion in terms of finite difference calculus, the logical development of which would involve an additional substantial chapter beyond the intended scope of this book. I therefore just provide the method only, without proof.

The essence of the method is to set up a table of differences as illustrated below. The first two columns are x and y. The entries in the remaining columns are the differences between the two entries in the column immediately to the left. Thus, for example, $\delta_{4.5} = y_5 - y_4$, $\delta_4^2 = \delta_{4.5} - \delta_{3.5}$, etc.



Let us suppose that we want to find y for a value of x that is a fraction θ of the way from x_4 to x_5. Bessel's interpolation formula is then

$$y(x) = \frac{1}{2}(y_4 + y_5) + B_1\delta_{4.5} + B_2(\delta_4^2 + \delta_5^2) + B_3\delta_{4.5}^3 + B_4(\delta_4^4 + \delta_5^4) + \ldots \tag{1.10.3}$$

Here the B_n are the Besselian interpolation coefficients, and the successive terms in parentheses in the expansion are the sums of the numbers in the boxes in the table.

The Besselian coefficients are

$$B_n(\theta) = \frac{1}{2}\begin{pmatrix} \theta + \frac{1}{2}n - 1 \\ n \end{pmatrix} \quad \text{if n is even,} \tag{1.10.4}$$

and

$$B_n(\theta) = \frac{\theta - \frac{1}{2}}{n}\begin{pmatrix} \theta + \frac{1}{2}n - \frac{3}{2} \\ n - 1 \end{pmatrix} \quad \text{if n is odd.} \tag{1.10.5}$$

The notation $\begin{pmatrix} m \\ n \end{pmatrix}$ means the coefficient of x m in the binomial expansion of $(1 + x)^n$.

Explicitly,

$$B_1 = \theta - \frac{1}{2} \tag{1.10.6}$$

$$B_2 = \frac{1}{2}\theta(\theta - 1)/2! = \theta(\theta - 1)/4 \tag{1.10.7}$$

$$B_3 = (\theta - \frac{1}{2})\theta(\theta - 1)/3! = \theta(0.5 + \theta(-1.5 + \theta))/6 \tag{1.10.8}$$

$$B_4 = \frac{1}{2}(\theta + 1)\theta(\theta - 1)(\theta - 2)/4! = \theta(2 + \theta(-1 + \theta(-2 + \theta)))/48 \tag{1.10.9}$$

$$B_5 = (\theta - \frac{1}{2})(\theta + 1)\theta(\theta - 1)(\theta - 2)/5! = \theta(-1 + \theta(2.5 + \theta^2(-2.5 + \theta)))/120 \tag{1.10.10}$$

The reader should convince him- or herself that the interpolation formula taken as far as B_1 is merely linear interpolation. Addition of successively higher terms effectively fits a curve to more and more points around the desired value and more and more accurately reflects the actual change of y with x.

t	*y*			
1	0.920 6928			
		−26037		
2	0.918 0891		−2600	
		−28637		9
3	0.915 2254		−2591	
		−31228		8
4	0.912 1026		−2583	
		−33811		10
5	0.908 7215		−2573	
		−36384		13
6	0.905 0831		−2560	
		−38944		11
7	0.901 1887		−2549	
		−41493		
8	0.897 0394			

The above table is taken from *The Astronomical Almanac* for 1997, and it shows the y-coordinate of the Sun for eight consecutive days in July. The first three difference columns are tabulated, and it is clear that further difference columns are unwarranted.

If we want to find the value of y, for example, for July 4.746, we have $\theta = 0.746$ and the first three Bessel coefficients are

$$\begin{aligned} B_1 &= +0.246 \\ B_2 &= -0.047\,371 \\ B_3 &= -0.007\,768\,844 \end{aligned}$$

The reader can verify the following calculations for y from the sum of the first 2, 3 and 4 terms of the Besselian interpolation series formula. The sum of the first two terms is the result of linear interpolation.

Sum of the first 2 terms, $y = 0.909\,580\,299$
Sum of the first 3 terms, $y = 0.909\,604\,723$
Sum of the first 4 terms, $y = 0.909\,604\,715$

Provided the table is not tabulated at inappropriately coarse intervals, one need rarely go past the third Bessel coefficient. In that case an alternative and equivalent interpolation formula (for $t = t_4 + \theta\Delta t)$, which avoids having to construct a difference table, is

$$\begin{aligned} y(t_4 + \theta\Delta t) \quad = \quad & -\tfrac{1}{6}\theta[(2-\theta(3-\theta))y_3 + (1-\theta)y_6] \\ & + \tfrac{1}{2}[(2+\theta(-1+\theta(-2+\theta)))y_4 + \theta(2+\theta(1-\theta))y_5]. \end{aligned} \tag{1.10.2}$$

Readers should check that this gives the same answer, at the same time noting that the nested parentheses make the calculation very rapid and they are easy to program on either a calculator or a computer.

Exercise 1.10.1: Bessel Coefficients

From the following table, construct a difference table up to fourth differences. Calculate the first four Bessel coefficients for $\theta = 0.73$. Hence calculate the value of y for $x = 0.273$.

x	y
0.0	+0.381300
0.1	+0.285603
0.2	+0.190092
0.3	+0.096327
0.4	+0.008268
0.5	−0.067725

(1.10.3)

Answers

- $B_1 = +0.23$
- $B_2 - -0.049275$
- $B_3 = -7.5555 \times 10^{-3}$
- $B_4 = +9.021841875 \times 10^{-3}$
- $y = 0.121289738$

Note: the table was calculated from a formula, and the interpolated answer is correct to nine significant figures.

Exercise 1.10.2: Linear Interpolation vs. Besselian Interpolation

From the following table of $\sin x$, use linear interpolation and Besselian interpolation to estimate $\sin 51^\circ$ to three significant figures.

$$\begin{array}{cc} x^\circ & \sin x \\ \\ 0 & 0.0 \\ 30 & 0.5 \\ 60 & \sqrt{3}/2 - 0.86603 \\ 90 & 1.0 \end{array} \tag{1.10.4}$$

Answers

- By linear interpolation, $\sin 51^\circ = 0.634$.
- By Besselian interpolation, $\sin 51^\circ = 0.776$.

The correct value is 0.777. You should be impressed – but there is more on interpolation to come, in Section 1.11.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.11: Fitting a Polynomial to a Set of Points - Lagrange Polynomials and Lagrange Interpolation

Given a set of n points on a graph, there any many possible polynomials of sufficiently high degree that go through all n of the points. There is, however, just one polynomial of degree less than n that will go through them all. Most readers will find no difficulty in determining the polynomial. For example, consider the three points (1 , 1), (2 , 2) , (3 , 2). To find the polynomial $y = a_0 + a_1x + a_2x^2$ that goes through them, we simply substitute the three points in turn and hence set up the three simultaneous Equations

$$\begin{aligned} 1 &= a_0 + a_1 + a_2 \\ 2 &= a_0 + 2a_1 + 4a_2 \\ 2 &= a_0 + 3a_1 + 9a_2 \end{aligned} \tag{1.11.1}$$

and solve them for the coefficients. Thus $a_0 = -1$, $a_1 = 2.5$ and $a_3 = -0.5$. In a similar manner we can fit a polynomial of degree $n-1$ to go exactly through n points. If there are more than n points, we may wish to fit a least squares polynomial of degree $n-1$ to go close to the points, and we show how to do this in sections 1.12 and 1.13. For the purpose of this section (1.11), however, we are interested in fitting a polynomial of degree $n-1$ exactly through n points, and we are going to show how to do this by means of Lagrange polynomials as an alternative to the method described above.

While the smallest-degree polynomial that goes through n points is usually of degree $n-1$, it could be less than this. For example, we might have four points, all of which fit exactly on a parabola (degree two). However, in general one would expect the polynomial to be of degree $n-1$, and, if this is not the case, all that will happen is that we shall find that the coefficients of the highest powers of x are zero.

That was straightforward. However, what we are going to do in this section is to fit a polynomial to a set of points by using some functions called *Lagrange polynomials*. These are functions that are described by Max Fairbairn as "cunningly engineered" to aid with this task.

Let us suppose that we have a set of n points:

$$(x_1, y_1), (x_1, y_1), (x_2, y_2), \ldots\ \ldots (x_i, y_i), \ldots\ \ldots (x_n, y_n), \tag{1.11.1}$$

and we wish to fit a polynomial of degree $n-1$ to them.

I assert that the function

$$y = \sum_{i=1}^{n} y_i L_i(x) \tag{1.11.2}$$

is the required polynomial, where the n functions , $L_i(x)$, $i = 1, n$, are n *Lagrange polynomials*, which are polynomials of degree $n-1$ defined by

$$L_i(x) = \prod_{j=1,\, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{1.11.3}$$

Written more explicitly, the first three Lagrange polynomials are

$$L_1(x) = \frac{(x-x_2)(x-x_3)(x-x_4)\ldots\ \ldots(x-x_n)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)\ldots\ \ldots(x_1-x_n)}, \tag{1.11.4}$$

and

$$L_2(x) = \frac{(x-x_1)(x-x_3)(x-x_4)\ldots\ \ldots(x-x_n)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)\ldots\ \ldots(x_2-x_n)} \tag{1.11.5}$$

and

$$L_3(x) = \frac{(x-x_1)(x-x_2)(x-x_4)\ldots\ldots(x-x_n)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)\ldots\ldots(x_3-x_n)} \tag{1.11.6}$$

At first encounter, this will appear meaningless, but with a simple numerical example it will become clear what it means and also that it has indeed been cunningly engineered for the task.

Consider the same points as before, namely (1 , 1), (2 , 2) , (3 , 2). The three Lagrange polynomials are

$$L_1(x) = \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{1}{2}(x^2-5x+6), \tag{1.11.7}$$

$$L_2(x) = \frac{(x-1)(x-3)}{(2-1)(2-3)} = -x^2+4x-3, \tag{1.11.8}$$

$$L_3(x) = \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{1}{2}(x^2-3x+2). \tag{1.11.9}$$

Equation 1.11.2 for the polynomial of degree $n-1$ that goes through the three points is, then,

$$y = 1\times\frac{1}{2}(x^2-5x+6)+2\times(-x^2+4x-3)+2\times\frac{1}{2}(x^2-3x+2); \tag{1.11.10}$$

that is

$$y = -\frac{1}{2}x^2+\frac{5}{2}x-1, \tag{1.11.11}$$

which agrees with what we got before.

One way or another, if we have found the polynomial that goes through the n points, we can then use the polynomial to interpolate between nontabulated points. Thus we can either determine the coefficients in $y = a_0+a_1x^2+a_2x^2\ldots$ by solving n simultaneous Equations, or we can use Equation 1.11.2 directly for our interpolation (without the need to calculate the coefficients a_0, a_1, etc.), in which case the technique is known as *Lagrangian interpolation*. If the tabulated function for which we need an interpolated value is a polynomial of degree less than n, the interpolated value will be exact. Otherwise it will be approximate. An advantage of this over Besselian interpolation is that it is not necessary that the function to be interpolated be tabulated at equal intervals in x. Most mathematical functions and astronomical tables, however, are tabulated at equal intervals, and in that case either method can be used.

Let us recall the example that we had in Section 1.10 on Besselian interpolation, in which we were asked to estimate the value of $\sin 51^\circ$ from the table

x°	$\sin x$
0	0.0
30	0.5
60	$\sqrt{3}/2 = 0.86603$
90	1.0

(1.11.2)

The four Lagrangian polynomials, evaluated at $x = 51$, are

$$L_1(51) = \frac{(51-30)(51-60)(51-90)}{(0-30)(0-60)(0-90)} = -0.0455, \tag{1.11.3}$$

$$L_2(51) = \frac{(51-0)(51-60)(51-90)}{(30-0)(30-60)(30-90)} = +0.3315, \tag{1.11.4}$$

$$L_3(51) = \frac{(51-0)(51-30)(51-90)}{(60-0)(60-30)(60-90)} = +0.7735, \tag{1.11.5}$$

$$L_4(51) = \frac{(51-0)(51-30)(51-60)}{(90-0)(90-30)(90-60)} = -0.0595. \tag{1.11.6}$$

Equation 1.11.2 then gives us

$$\sin 51^\circ = 0 \times (-0.0455) + 0.5 \times 0.3315 + 0.86603 \times 0.7735 + 1 \times (-0.0595)$$
$$= 0.776.$$

This is the same as we obtained with Besselian interpolation, and compares well with the correct value of 0.777. I point out again, however, that the Lagrangian method can be used if the function is not tabulated at equal intervals, whereas the Besselian method requires tabulation at equal intervals.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.12: Fitting a Least Squares Straight Line to a Set of Observational Points

Very often we have a set of observational points (x_i, y_i), $i = 1$ to N, that seem to fall roughly but not quite on a straight line, and we wish to draw the "best" straight line that passes as close as possible to all the points. Even the smallest of scientific hand calculators these days have programs for doing this – but it is well to understand precisely what it is that is being calculated.

Very often the values of x_i are known "exactly" (or at least to a high degree of precision) but there are appreciable errors in the values of y_i. In figure I.6B I show a set of points and a plausible straight line that passes close to the points.

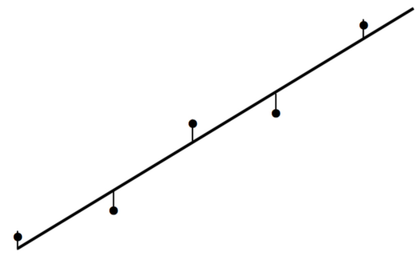


FIGURE I.6B

Also drawn are the vertical distances from each point from the straight line; these distances are the *residuals* of each point.

It is usual to choose as the "best" straight line that line such that the sum of the squares of these residuals is least. You may well ask whether it might make at least equal sense to choose as the "best" straight line that line such that the sum of the absolute values of the residuals is least. That certainly does make good sense, and in some circumstances it may even be the appropriate line to choose. However, the "least squares" straight line is rather easier to calculate and is readily amenable to statistical analysis. Note also that using the *vertical* distances between the points and the straight line is appropriate only if the values of x_i are known to much higher precision than the values of y_i. In practice, this is often the case – but it is not always so, in which case this would not be the appropriate "best" line to choose.

The line so described – i.e. the line such that the sum of the squares of the vertical residuals is least is often called loosely the "least squares straight line". Technically, it is the least squares linear regression of y *upon* x. It might be under some circumstances that it is the values of y_i that are known with great precision, whereas there may be appreciable errors in the x_i. In that case we want to minimize the sum of the squares of the *horizontal* residuals, and we then calculate the *least squares linear regression of x upon y*. Yet again, we may have a situation in which the errors in x and y are comparable (not necessarily exactly equal). In that case we may want to minimize the sum of the squares of the *perpendicular* residuals of the points from the line. But then there is a difficulty of drawing the x- and y-axes to equal scales, which would be problematic if, for example, x were a time and y a distance.

To start with, however, we shall assume that the errors in x are negligible and we want to calculate the least squares regression of y upon x. We shall also make the assumption that all points have *equal weight*. If they do not, this is easily dealt with in an obvious manner; thus, if a point has twice the weight of other points, just count that point twice.

So, let us suppose that we have N points, (x_i, y_i), $i = 1$ to N, and we wish to fit a straight line that goes as close as possible to all the points. Let the line be $y = a_1x + a_0$. The *residual* R_i of the ith point is

$$R_i = y_i - (a_1x_i + a_0). \tag{1.12.1}$$

We have N simultaneous linear Equations of this sort for the two unknowns a_1 and a_0, and, for the least squares regression of y upon x, we have to find the values of a_1 and a_0 such that the sum of the squares of the residuals is least. *We already know how to do this* from Section 1.8, so the problem is solved. (Just make sure that you understand that, in Section 1.8 we were using x for the unknowns and a for the coefficients; here we are doing the opposite!)

Now for an *Exercise*. Suppose our points are as follows:

$$\begin{array}{cc} x & y \\ \\ 1 & 1.00 \\ 2 & 2.50 \\ 3 & 2.75 \\ 4 & 3.00 \\ 5 & 3.50 \end{array} \tag{1.12.1}$$

i.) Draw these points on a sheet of graph paper and, using your eye and a ruler, draw what you think is the best straight line passing close to these points.

ii.) Write a computer program for calculating the least squares regression of y upon x. You've got to do this sooner or later, so you might as well do it now. In fact you should already (after you read Section 1.8) have written a program for solving N Equations in n unknowns, so you just incorporate that program into this.

iii.) Now calculate the least squares regression of y upon x. I make it $y = 0.55x + 0.90$. Draw this on your graph paper and see how close your eye-and-ruler estimate was!

iv.) How are you going to calculate the least squares regression of x upon y? Easy! Just use the same program, but read the x-values for y and the y-values for x! No need to write a second program! I make it $y = 0.645x + 0.613$. Draw that on your graph paper and see how it compares with the regression of y upon x.

The two regression lines intersect at the centroid of the points, which in this case is at (3.00, 2.55). If the errors in x and y are comparable, a reasonable best line might be one that passes through the centroid, and whose slope is the mean (arithmetic? geometric?) of the regressions of y upon x and x upon y. However, in Section 1.12 I shall give a reference to where this question is treated more thoroughly.

If the regressions of y upon x and x upon y are respectively $y = a_1x + a_0$ and $y = b_1x + b_0$, the quantity $\sqrt{a_1/b_1}$ is called the correlation coefficient r between the variates x and y. If the points are exactly on a straight line, the correlation coefficient is 1. The correlation coefficient is often used to show how well, or how badly, two variates are correlated, and it is often averred that they are highly correlated if r is close to 1 and only weakly correlated if r is close to zero. I am not intending to get bogged down in formal statistics in this chapter, but a word of warning here is in order. If you have just two points, they are necessarily on a straight line, and the correlation coefficient is necessarily 1 – but there is no evidence whatever that the variates are in any way correlated. The correlation coefficient by itself does not tell how closely correlated two variates are. The *significance* of the correlation coefficient depends on the number of points, and the significance is something that can be calculated numerically by precise statistical tests.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.13: Fitting a Least Squares Polynomial to a Set of Observational Points

I shall start by assuming that the values of x are known to a high degree of precision, and all the errors are in the values of y. In other words, I shall calculate a least squares polynomial regression of y upon x. In fact I shall show how to calculate a least squares quadratic regression of y upon x, a quadratic polynomial representing, of course, a parabola. What we want to do is to calculate the coefficients $a_0,\ a_1,\ a_2$ such that the sum of the squares of the residual is least, the residual of the ith point being

$$R_i = y_i - (a_0 + a_1 x_i + a_2 x_1^2). \tag{1.13.1}$$

You have N simultaneous linear Equations of this sort for the three unknowns $a_0,\ a_1$ and a_2. You already know how to find the least squares solution for these, and indeed, after having read Section 1.8, you already have a program for solving the Equations. (Remember that here the unknowns are $a_0,\ a1$ and a_2 – not x! You just have to adjust your notation a bit.) Thus there is no difficulty in finding the least squares quadratic regression of y upon x, and indeed the extension to polynomials of higher degree will now be obvious.

As an *Exercise*, here are some points that I recently had in a real application:

$$\begin{array}{cc} x & y \\ 395.1 & 171.0 \\ 448.1 & 289.0 \\ 517.7 & 399.0 \\ 583.3 & 464.0 \\ 790.2 & 620.0 \end{array} \tag{1.13.1}$$

Draw these on a sheet of graph paper and draw by hand a nice smooth curve passing as close as possible to the point. Now calculate the least squares parabola (quadratic regression of y upon x) and see how close you were. I make it $y = -961.34 + 3.7748x - 2.247 \times 10^{-3} x^2$. It is shown in Figure I.6C.

I now leave you to work out how to fit a least squares cubic (or indeed any polynomial) regression of y upon x to a set of data points. For the above data, I make the cubic fit to be

$$y = -2537.605 + 12.4902x - 0.017777x^2 + 8.89 \times 10^{-6} x^3. \tag{1.13.2}$$

This is shown in Figure I.6D, and, on the scale of this drawing it cannot be distinguished (within the range covered by x in the figure) from the quartic Equation that would go exactly through all five points.

The cubic curve is a “better” fit than either the quadratic curve or a straight line in the sense that, the higher the degree of polynomial, the closer the fit and the less the residuals. But higher degree polynomials have more “wiggles”, and you have to ask yourself whether a high-degree polynomial with lots of “wiggles” is really a realistic fit, and maybe you should be satisfied with a quadratic fit. Above all, it is important to understand that it is very dangerous to use the curve that you have calculated to *extrapolate* beyond the range of x for which you have data – and this is especially true of higher-degree polynomials.

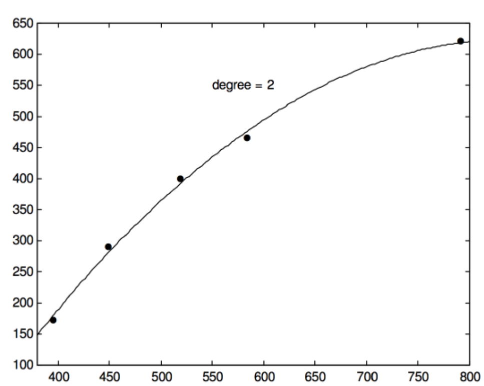


FIGURE I.6C

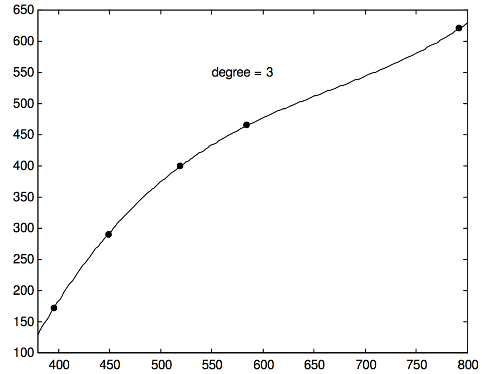


FIGURE I.6D

What happens if the errors in x are not negligible, and the errors in x and y are comparable in size? In that case you want to plot a graph of y against x on a scale such that the unit for x is equal to the standard deviation of the x-residuals from the chosen polynomial and the unit for y is equal to the standard deviation of the y-residuals from the chosen polynomial. For a detailed and thorough account of how to do this, I refer you to a paper by D. York in Canadian Journal of Physics, **44**, 1079 (1966).

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.14: Legendre Polynomials

Consider the expression

$$(1-2rx+r^2)^{-1/2}, \tag{1.14.1}$$

in which $|x|$ and $|r|$ are both less than or equal to one. Expressions similar to this occur quite often in theoretical physics - for example in calculating the gravitational or electrostatic potentials of bodies of arbitrary shape. See, for example, Chapter 5, Sections 5.11 and 5.12.

Expand the expression 1.14.1 by the binomial theorem as a series of powers of r. This is straightforward, though not particularly easy, and you might expect to spend several minutes in obtaining the coefficients of the first few powers of r. You will find that the coefficient of r^l is a polynomial expression in x of degree l. Indeed the expansion takes the form

$$(1-2rx+r^2)^{-1/2} = P_0(x)+P_1(x)r+P_2(x)r^2+P_3(x)r^3\ldots \tag{1.14.2}$$

The coefficients of the successive power of r are the *Legendre polynomials*; the coefficient of r^l, which is $P_l(x)$, is the Legendre polynomial of order l, and it is a polynomial in x including terms as high as x^l. We introduce these polynomials in this section because we shall need them in Section 1.15 on the derivation of Gaussian Quadrature.

If you have conscientiously tried to expand expression 1.14.1, you will have found that

$$P_0(x)=1, \quad P_1(x)=x, \quad P_2(x)=\frac{1}{2}(3x^2-1), \tag{1.14.3}$$

though you probably gave up with exhaustion after that and didn't take it any further. If you look carefully at how you derived the first few polynomials, you may have discovered for yourself that you can obtain the next polynomial as a function of two earlier polynomials. You may even have discovered for yourself the following *recursion relation*:

$$P_{l+1} = \frac{(2l+1)xP_l - lP_{l-1}}{l+1}. \tag{1.14.4}$$

This enables us very rapidly to obtain higher order Legendre polynomials, whether numerically or in algebraic form. For example, put $l=1$ and show that Equation 1.12.4 results in $P_2=\frac{1}{2}(3x^2-1)$. You will then want to calculate P_3, and then P_4, and more and more and more. Another way to generate them is form the Equation

$$P_{l+1} = \frac{1}{2^l l!}\frac{d^l}{dx^l}(x^2-1)^l. \tag{1.14.5}$$

Here are the first eleven Legendre polynomials:

$$\begin{aligned}
&P_0 = 1\\
&P_1 = x\\
&P_2 = \tfrac{1}{2}(3x^2-1)\\
&P_3 = \tfrac{1}{2}(5x^3-3x)\\
&P_4 = \tfrac{1}{8}(35x^4-30x^2+3)\\
&P_5 = \tfrac{1}{16}(63x^5-70x^3+15x)\\
&P_6 = \tfrac{1}{16}(231x^6-315x^4+105x^2-5)\\
&P_7 = \tfrac{1}{16}(429x^7-693x^5+315x^3-35x)\\
&P_8 = \tfrac{1}{128}(6435x^8-12012x^6+6930x^4-1260x^2+35\\
&P_9 = \tfrac{1}{128}(12155x^9-25740x^7+18018x^5-4620x^3+315x)\\
&P_{10} = \tfrac{1}{256}(46189x^{10}-109395x^8+90090x^6-30030x^4+3465x^2-63)
\end{aligned} \tag{1.14.6}$$

The polynomials with argument $\cos\theta$ are given in Section 5.11 of Chapter 5.

In what follows in the next section, we shall also want to know the roots of the Equations $P_l = 0$ for $l > 1$. Inspection of the forms of these polynomials will quickly show that all odd polynomials have a root of zero, and all nonzero roots occur in positive/negative pairs. Having read Sections 1.4 and 1.5, we shall have no difficulty in finding the roots of these Equations. The roots $x_{l,i}$ are in the following table, which also lists certain coefficients $c_{l,i}$, that will be explained in Section 1.15.

Roots of $P_l = 0$

l	$x_{l,i}$	$c_{l,i}$
2	±0.577 350 269 190	1.000 000 000 00
3	±0.774 596 669 241	0.555 555 555 56
	0.000 000 000 000	0.888 888 888 89
4	±0.861 136 311 594	0.347 854 845 14
	±0.339 981 043 585	0.652 145 154 86
5	±0.906 179 845 939	0.236 926 885 06
	±0.538 469 310 106	0.478 628 670 50
	0.000 000 000 000	0.568 888 888 89
6	±0.932 469 514 203	0.171 324 492 38
	±0.661 209 386 466	0.360 761 573 05
	±0.238 619 186 083	0.467 913 934 57
7	±0.949 107 912 343	0.129 484 966 17
	±0.741 531 185 599	0.279 705 391 49
	±0.405 845 151 377	0.381 830 050 50
	0.000 000 000 000	0.417 959 183 68
8	±0.960 289 856 498	0.101 228 536 29
	±0.796 666 477 414	0.222 381 034 45
	±0.525 532 409 916	0.313 706 645 88
	±0.183 434 642 496	0.362 683 783 38
9	±0.968 160 239 508	0.081 274 388 36
	±0.836 031 107 327	0.180 648 160 69
	±0.613 371 432 701	0.260 610 696 40
	±0.324 253 423 404	0.312 347 077 04
	0.000 000 000 000	0.330 239 355 00
10	±0.973 906 528 517	0.066 671 343 99
	±0.865 063 366 689	0.149 451 349 64
	±0.679 409 568 299	0.219 086 362 24
	±0.433 395 394 129	0.269 266 719 47
	±0.148 874 338 982	0.295 524 224 66
11	±0.978 228 658 146	0.055 668 567 12
	±0.887 062 599 768	0.125 580 369 46

	±0.730 152 005 574	0.186 290 210 93
	±0.519 096 129 207	0.233 193 764 59
	±0.269 543 155 952	0.262 804 544 51
	0.000 000 000 000	0.272 925 086 78
12	±0.981 560 634 247	0.047 175 336 39
	±0.904 117 256 370	0.106 939 325 99
	±0.769 902 674 194	0.160 078 328 54
	±0.587 317 954 287	0.203 167 426 72
	±0.367 831 498 998	0.233 492 536 54
	±0.125 233 408 511	0.249 147 045 81
13	±0.984 183 054 719	0.040 484 004 77
	±0.917 598 399 223	0.092 121 499 84
	±0.801 578 090 733	0.138 873 510 22
	±0.642 349 339 440	0.178 145 980 76
	±0.448 492 751 036	0.207 816 047 54
	±0.230 458 315 955	0.226 283 180 26
	0.000 000 000 000	0.232 551 553 23
14	±0.986 283 808 697	0.035 119 460 33
	±0.928 434 883 664	0.080 158 087 16
	±0.827 201 315 070	0.121 518 570 69
	±0.687 292 904 812	0.157 203 167 16
	±0.515 248 636 358	0.185 538 397 48
	±0.319 112 368 928	0.205 198 463 72
	±0.108 054 948 707	0.215 263 853 46
15	±0.987 992 518 020	0.030 753 242 00
	±0.937 273 392 401	0.070 366 047 49
	±0.848 206 583 410	0.107 159 220 47
	±0.724 417 731 360	0.139 570 677 93
	±0.570 972 172 609	0.166 269 205 82
	±0.394 151 347 078	0.186 161 000 02
	±0.201 194 093 997	0.198 431 485 33
	0.000 000 000 000	0.202 578 241 92
16	±0.989 400 934 992	0.027 152 459 41
	±0.944 575 023 073	0.062 253 523 94
	±0.865 631 202 388	0.095 158 511 68
	±0.755 404 408 355	0.124 628 971 26
	±0.617 876 244 403	0.149 595 988 82
	±0.458 016 777 657	0.169 156 519 39
	±0.281 603 550 779	0.182 603 415 04
	±0.095 012 509 838	0.189 450 610 46
17	±0.990 575 475 315	0.024 148 302 87
	±0.950 675 521 769	0.055 459 529 38

(1.14.1)

±0.880 239 153 727	0.085 036 148 32
±0.781 514 003 897	0.111 883 847 19
±0.657 671 159 217	0.135 136 368 47
±0.512 690 537 086	0.154 045 761 08
±0.351 231 763 454	0.168 004 102 16
±0.178 484 181 496	0.176 562 705 37
0.000 000 000 000	0.179 446 470 35

For interest, I draw graphs of the Legendre polynomials in figures I.7 and I.8.

Figure I.7. Legendre polynomials for even I

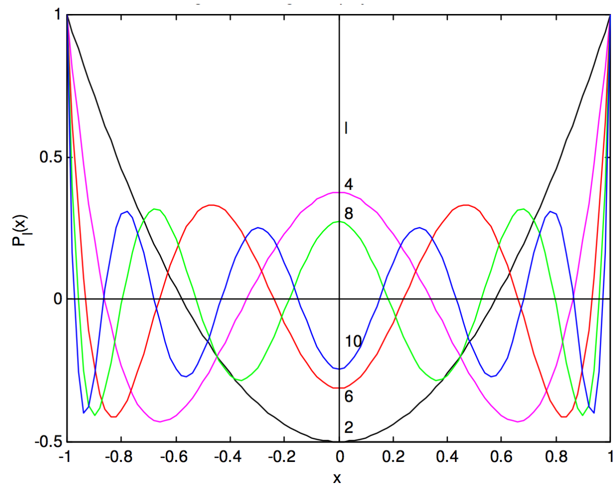
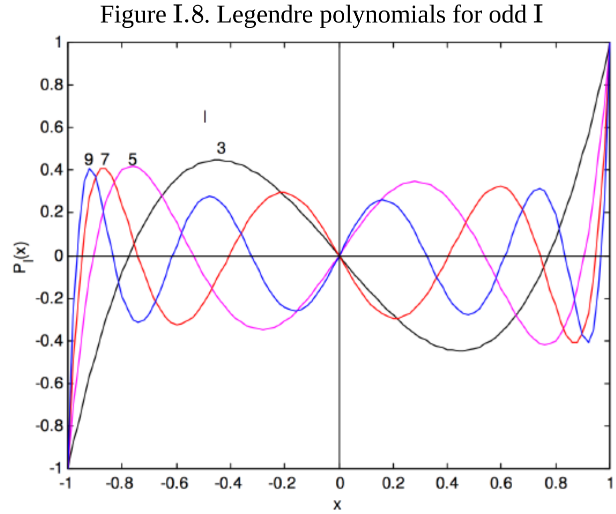


Figure I.8. Legendre polynomials for odd I



For further interest, it should be easy to verify, by substitution, that the Legendre polynomials are solutions of the differential Equation

$$(1-x^2)y'' - 2xy' + l(l+1)y = 0. \tag{1.14.7}$$

The Legendre polynomials are solutions of this and related Equations that appear in the study of the vibrations of a solid sphere (spherical harmonics) and in the solution of the Schrödinger Equation for hydrogen-like atoms, and they play a large role in quantum mechanics.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.15: Gaussian Quadrature - the Algorithm

Gaussian quadrature is an alternative method of numerical integration which is often much faster and more spectacular than Simpson's rule. Gaussian quadrature allows you to carry out the integration

$$\int_{-1}^{1} f(x) dx. \tag{1.15.1}$$

But what happens if your limits of integration are not ± 1? What if you want to integrate

$$\int_{a}^{b} F(t) dt? \tag{1.15.2}$$

That is no problem at all – you just make a change of variable. Thus, let

$$x = \frac{2t-a-b}{b-a}, \quad t = \frac{1}{2}[(b-a)x+a+b], \tag{1.15.3}$$

and the new limits are then $x = \pm 1$.

At the risk of being pedagogically unsound I'll describe first, without any theoretical development, just what you do, with an example – as long as you promise to look at the derivation afterwards, in Section 1.16.

For our example, let's try to evaluate

$$I = \int_{0}^{\pi/2} \sin\theta d\theta. \tag{1.15.4}$$

Let us make the change of variable given by Equation 1.15.3(with $t = \theta$, $a = 0$, $b = \pi/2$), and we now have to evaluate

$$I = \int_{-1}^{1} \frac{\pi}{4} \sin\frac{\pi}{4}(x+1) dx. \tag{1.15.5}$$

For a 5-point Gaussian quadrature, you evaluate the integrand at five values of x, where these five values of x are the solutions of $P_5(x) = 0$ given in Section 1.14, P_5 being the Legendre polynomial. That is, we evaluate the integrand at $x = \pm 0.906\,469\,514\,203, \ \pm 0.538\,469\,310\,106$and 0.

I now assert, without derivation (until later), that

$$I = \sum_{i=1}^{5} c_{5,i} f(x_{5,i}), \tag{1.15.6}$$

where the coefficients $c_{l,i}$ (all positive) are listed with the roots of the Legendre polynomials in Section 1.14.

Let's try it.

$x_{5,i}$	$f(x_{5,i})$	$c_{5,i}$
$+0.906\,179\,845\,939$	$0.783\,266\,908\,39$	$0.236\,926\,885\,06$
$+0.538\,469\,310\,106$	$0.734\,361\,739\,69$	$0.478\,628\,670\,50$
$0.000\,000\,000\,000$	$0.555\,360\,367\,27$	$0.568\,888\,888\,89$
$-0.538\,469\,310\,006$	$0.278\,501\,544\,60$	$0.478\,628\,670\,50$
$-0.906\,179\,845\,939$	$0.057\,820\,630\,35$	$0.236\,926\,885\,06$

(1.15.1)

and the expression 1.15.6 comes to $1.000\,000\,000\,04$ and might presumably have come even closer to 1 had we given $x_{l,i}$, and $c_{l,i}$, to more significant figures.

You should now write a computer program for Gaussian quadrature – you will have to store the $x_{l,i}$ and $c_{l,i}$ of course. You have presumably already written a program for Simpson's rule.

In a text on integration, the author invited the reader to evaluate the following integrals by Gaussian quadrature:

$$\begin{array}{llll} (a) & \int_1^{1.5} x^2 \ln x dx & (e) & \int_0^{\pi/4} e^{3x} \sin 2x dx \\ (b) & \int_0^1 x^2 x^{-x} dx & (f) & \int_1^{1.6} \frac{2x}{x^2-4} dx \\ (c) & \int_0^{0.35} \frac{2}{x^2-4} dx & (g) & \int_3^{3.5} \frac{x}{\sqrt{x^2-4}} dx \\ (d) & \int_0^{\pi/4} x^2 \sin x dx & (h) & \int_0^{\pi/4} \cos^2 x dx \end{array} \tag{1.15.2}$$

All of these can be integrated analytically, so I am going to invite the reader to evaluate them first analytically, and then numerically by Simpson's rule and again by Gaussian quadrature, and to see at how many points the integrand has to be evaluated by each method to achieve nine or ten figure precision. I tried, and the results are as follows. The first column is the answer, the second column is the number of points required by Simpson's rule, and the third column is the number of points required by Gaussian quadrature.

$$\begin{array}{llll} (a) & 0.192\,259\,358 & 33 & 4 \\ (b) & 0.160\,602\,794 & 99 & 5 \\ (c) & -0.176\,820\,020 & 19 & 4 \\ (d) & 0.088\,755\,284\,4 & 111 & 5 \\ (e) & 2.588\,628\,633 & 453 & 7 \\ (f) & -0.733\,969\,175 & 143 & 8 \\ (g) & 0.636\,213\,346 & 31 & 5 \\ (h) & 0.642\,699\,082 & 59 & 5 \end{array} \tag{1.15.3}$$

Let us now have a look at four of the integrals that we met in Section 1.2.

1. $\int_0^1 \frac{x^4 dx}{\sqrt{2(1+x^2)}}$. This was straightforward. It has an analytic solution of $\frac{\sqrt{18}\ln(1+\sqrt{2})-2}{16} = 0.108\,709\,465$. I needed to evaluate the integral at 89 points in order to get this answer to nine significant figures using Simpson's rule. To use Gaussian quadrature, we note that integrand contains only even powers of x and so it is symmetric about $x=0$, and therefore the integral is equal to $\frac{1}{2}\int_{-1}^1 \frac{x^4 dx}{\sqrt{2(1+x^2)}}$, which makes it immediately convenient for Gaussian quadrature! I give below the answers I obtained for 3- to 7-point Gaussian quadrature.

$$\begin{array}{rl} 3 & 0.108\,667\,036 \\ 4 & 0.108\,711\,215 \\ 5 & 0.108\,709\,441 \\ 6 & 0.108\,709\,463 \\ 7 & 0.108\,709\,465 \\ \text{Correct answer} & 0.108\,709\,465 \end{array} \tag{1.15.4}$$

2. $\int_0^2 \frac{y^2 dy}{\sqrt{2-y}}$. This had the difficulty that the integrand is infinite at the upper limit. We got round this by means of the substitution $y = 2\sin^2\theta$, and the integral becomes $\sqrt{128}\int_0^{\pi/2} \sin^5\theta d\theta$. This has an analytic solution of $\sqrt{8192}/15 = .6033977866$ I needed 59 points to get this answer to ten significant figures using Simpson's rule. To use Gaussian quadrature we can let $y = 1+x$, so that the integral becomes , $\int_{-1}^1 \frac{(1+x)^2 dy}{\sqrt{1-x}}$, which seems to be immediately suitable for Gaussian quadrature. Before we proceed, we recall that the integrand becomes infinite at the upper limit, and it still does so after our change of variable. We note, however, that with Gaussian quadrature, *we do not evaluate the integrand at the upper limit*, so that this would appear to be a great advantage of the method over Simpson's method. Alas! – this turns out not to be the case. If, for example, we use a 17-point quadrature, the largest value of x for which we evaluate the integrand is equal to the largest solution of $P_{17}(x) = 0$, which is 0.9906. We just cannot ignore the fact that the integrand shoots up to infinity beyond this, so we have left behind a large part of the integral. Indeed, with a 17-point Gaussian quadrature, I obtained an answer of 5.75, which is a long way from the correct answer of 6.03.

Therefore we have to make a change of variable, as we did for Simpson's method, so that the upper limit is finite. We chose $y=2\sin^2\theta$ which changed the integral to $\sqrt{128}\int_0^{\pi/2}\sin^5\theta d\theta$. To make this suitable for Gaussian quadrature, we must now make the further substitution (see Equation 1.15.3) $x=4\theta/\pi-1$, $\theta=\frac{\pi}{4}(x+1)$. If we wish to impress, we can make the two substitutions in one step, thus: Let $y=2\sin^2\frac{\pi}{4}(1+x)$, $x=\frac{4}{\pi}\sin^{-1}\sqrt{\frac{y}{2}}-1$. The integral becomes $\sqrt{8}\pi\int_{-1}^{1}\sin^5\frac{\pi}{4}(1+x)dx$, and there are no further difficulties. With a 9-point integration, I obtained the answer, correct to ten significant figures, $6.033\,977\,866$ Simpson's rule required 59 points.

3. $\int_0^{\pi/2}\sqrt{\sec\theta}d\theta$. This integral occurs in the theory of a simple pendulum swinging through 90°. As far as I can tell it has no simple analytical solution unless we have recourse to unfamiliar elliptic integrals, which we would have to evaluate numerically in any case. The integral has the difficulty that the integrand is infinite at the upper limit. We get round this by means of a substitution. Thus let $\sin\phi=\sqrt{2}\sin\frac{1}{2}\theta$. (Did you not think of this?) The integral becomes $\sqrt{2}\int_0^{\pi/2}\frac{d\phi}{\sqrt{1-\frac{1}{2}\sin^2\phi}}$. I needed 13 points by Simpson's rule to get the answer to ten significant figures, $2.622\,057\,554$ In order to make the limits ±1, suitable for Gaussian quadrature, we can make the second substitution (as in example 2), $\phi=\frac{\pi}{4}(x+1)$. If we wish truly to impress our friends, we can make the two substitutions in one step, thus: Let $\sin\frac{\pi}{4}(1+x)=\sqrt{2}\sin\frac{1}{2}\theta$. (No one will ever guess how we thought of that!) The integral becomes $\frac{\pi}{2}\int_{-1}^{1}\frac{dx}{\sqrt{2-\sin^2\frac{\pi}{4}(x+1)}}$, which is now ready for Gaussian quadrature. I obtained the answer $2.622\,057\,554$in a 10-point Gaussian quadrature, which is only a little faster than the 13 points required by Simpson's rule.

4. $\int_0^{\infty}\frac{dy}{y^5\left(e^{1/y}-1\right)}$. This integral occurs in the theory of blackbody radiation. It has the difficulty of an infinite upper limit. We get round this by means of a substitution. Thus let $y=\tan\theta$. The integral becomes $\int_0^{\pi/2}\frac{c^3(c^2+1)}{e^c-1}d\theta$, where $c=\cot\theta$. It has an analytic solution of $\pi^4/15=6.493\,939\,402$ I needed 261 points by Simpson's rule to get the answer to ten significant figures. To prepare it for Gaussian quadrature, we can let $\theta=\frac{\pi}{4}(x+1)$, as we did in example 2, so that the integral becomes ,where $\frac{\pi}{4}\int_{-1}^{1}\frac{c^3(c^2+1)}{e^c-1}dx$, where $c=\cot\frac{\pi}{4}(x+1)$. Using 16- point Gaussian quadrature, I got 6.48. Thus we would need to extend our table of constants for the Gaussian method to much higher order in order to use the method successfully. Doubtless the Gaussian method would then be faster than the Simpson method – but we do not need an extensive (and difficult-to-calculate) set of constants for the latter. A further small point: You may have noticed that it is not immediately obvious that the integrand is zero at the end points, and that some work is needed to prove it. But with the Gaussian method you don't evaluate the integrand at the end points, so that is one less thing to worry about!

Thus we have found that in most cases the Gaussian method is far faster than the Simpson method. In some cases it is only marginally faster. In yet others it probably would be faster than Simpson's rule, but higher-order constants are needed to apply it. Whether we use Simpson's rule or Gaussian quadrature, we have to carry out the integration with successively higher orders until going to higher orders results in no further change to the number of significant figures desired.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.16: Gaussian Quadrature - Derivation

In order to understand why Gaussian quadrature works so well, we first need to understand some properties of polynomials in general, and of Legendre polynomials in particular. We also need to remind ourselves of the use of Lagrange polynomials for approximating an arbitrary function.

First, a statement concerning polynomials in general: Let P be a polynomial of degree n, and let S be a polynomial of degree less than $2n$. Then, if we divide S by P, we obtain a quotient Q and a remainder R, each of which is a polynomial of degree less than n.

That is to say:

$$\frac{S}{P}=Q+\frac{R}{P}. \tag{1.16.1}$$

What this means is best understood by looking at an example, with $n=3$. For example,

let

$$P=5x^3-2x^2+3x+7 \tag{1.16.2}$$

and

$$S=9x^5+4x^4-5x^3+6x^2+2x-3. \tag{1.16.3}$$

If we carry out the division $S \div P$ by the ordinary process of long division, we obtain

$$\frac{9x^5+4x^4-5x^3+6x^2+2x-3}{5x^3-2x^2+3x+7}=1.8x^2+1.52x-1.472-\frac{14.104x^2+4.224x-7.304}{5x^3-2x^2+3x+7}. \tag{1.16.4}$$

For example, if $x=3$, this becomes

$$\frac{2433}{133}=19.288-\frac{132.304}{133}. \tag{1.16.1}$$

The theorem given by Equation 1.16.1 is true for any polynomial P of degree l. In particular, it is true if P is the Legendre polynomial of degree l.

Next an important property of the Legendre polynomials, namely, if P_n and P_m are Legendre polynomials of degree n and m respectively, then

$$\int_{-1}^{1} P_n P_m dx=0 \quad \text{unless } m=n. \tag{1.16.5}$$

This property is called the *orthogonal* property of the Legendre polynomials.

I give here a proof. Although it is straightforward, it may look formidable at first, so, on first reading, you might want to skip the proof and go on the next part (after the next short horizontal dividing line).

From the symmetry of the Legendre polynomials (see figure $\text{I}.7$), the following are obvious:

$$\int_{-1}^{1} P_n P_m dx \neq 0 \quad \text{if } m=n \tag{1.16.2}$$

and

$$\int_{-1}^{1} P_n P_m=0 \quad \text{if one (but not both) of } m \text{ or } n \text{ is odd.} \tag{1.16.3}$$

In fact we can go further, and, as we shall show,

$$\int_{-1}^{1} P_n P_m dx = 0 \quad \text{unless } m = n, \text{ whether } m \text{ and } n \text{ are even or odd.} \tag{1.16.4}$$

Thus P_m satisfies the differential Equation (see Equation 1.14.7)

$$(1-x^2)\frac{d^2P_m}{dx^2} - 2x\frac{dP_m}{dx} + m(m+1)P_m = 0, \tag{1.16.6}$$

which can also be written

$$\frac{d}{dx}\left[(1-x^2)\frac{dP_m}{dx}\right] + m(m+1)P_m = 0. \tag{1.16.7}$$

Multiply by P_n:

$$P_n\frac{d}{dx}\left[(1-x^2)\frac{dP_m}{dx}\right] + m(m+1)P_mP_n = 0, \tag{1.16.8}$$

which can also be written

$$\frac{d}{dx}\left[(1-x^2)P_n\frac{dP_m}{dx}\right] - (1-x^2)\frac{dP_n}{dx}\frac{dP_m}{dx} + m(m+1)P_mP_n = 0. \tag{1.16.9}$$

In a similar manner, we have

$$\frac{d}{dx}\left[(1-x^2)P_m\frac{dP_n}{dx}\right] - (1-x^2)\frac{dP_n}{dx}\frac{dP_m}{dx} + n(n+1)P_mP_n = 0. \tag{1.16.10}$$

Subtract one from the other:

$$\frac{d}{dx}\left[(1-x^2)\left(P_n\frac{dP_m}{dx} - P_m\frac{dP_n}{dx}\right)\right] + [m(m+1)-n(n+1)]P_mP_n = 0. \tag{1.16.11}$$

Integrate from -1 to $+1$:

$$\left[(1-x^2)\left(P_n\frac{dP_m}{dx} - P_m\frac{dP_n}{dx}\right)\right]_{-1}^{1} = [n(n+1)-m(m+1)]\int_{-1}^{1} P_mP_n dx. \tag{1.16.12}$$

The left hand side is zero because $1-x^2$ is zero at both limits.

Therefore, unless $m = n$,

$$\int_{-1}^{1} P_mP_n dx = 0. \qquad \text{Q.E.D.} \tag{1.16.13}$$

I now assert that, if P_l is the Legendre polynomial of degree l, and if Q is any polynomial of degree less than l, then

$$\int_{-1}^{1} P_l Q dx = 0. \tag{1.16.14}$$

I shall first prove this, and then give an example, to see what it means.

To start the proof, we recall the recursion relation (see Equation 1.14.4 – though here I am substituting $l-1$ for l) for the Legendre polynomials:

$$lP_l = (2l-1)xP_{l-1} - (l-1)P_{l-2}. \tag{1.16.15}$$

The proof will be by induction.

Let Q be any polynomial of degree less than l. Multiply the above relation by Qdx and integrate from -1 to $+1$:

$$l\int_{-1}^{1} P_lQdx = (2l-1)\int_{-1}^{1} xP_{l-1}Qdx - (l-1)\int_{-1}^{1} P_{l-2}Qdx. \tag{1.16.16}$$

If the right hand side is zero, then the left hand side is also zero.

A correspondent has suggested to me a much simpler proof. He points out that you could in principle expand Q in Equation 1.16.14 as a sum of Legendre polynomials for which the highest degree is $l-1$. Then, by virtue of Equation 1.16.13, every term is zero.

For example, let $l=4$, so that

$$P_{l-2}=P_2=\frac{1}{2}(3x^2-1) \tag{1.16.17}$$

and

$$xP_{l-1}=xP_3=\frac{1}{2}(5x^4-3x^2), \tag{1.16.18}$$

and let

$$Q=2(a_3x^3+a_2x^2+a_1x+a_0). \tag{1.16.19}$$

It is then straightforward (and only slightly tedious) to show that

$$\int_{-1}^{1} P_{l-2}Q\,dx=\left(\frac{6}{5}-\frac{2}{3}\right)a_2 \tag{1.16.20}$$

and that

$$\int_{-1}^{1} xP_{l-1}Q\,dx=\left(\frac{10}{7}-\frac{6}{5}\right)a_2. \tag{1.16.21}$$

But

$$7\left(\frac{10}{7}-\frac{6}{5}\right)a_2-3\left(\frac{6}{5}-\frac{2}{3}\right)a_2=0, \tag{1.16.22}$$

and therefore

$$\int_{-1}^{1} P_4Q\,dx=0. \tag{1.16.23}$$

We have shown that

$$l\int_{-1}^{1} P_lQ\,dx=(2l-1)\int_{-1}^{1} xP_{l-1}Q\,dx-(l-1)\int_{-1}^{1} P_{l-2}Q\,dx=0 \tag{1.16.24}$$

for $l=4$, and therefore it is true for all positive integral l.

You can use this property for a parlour trick. For example, you can say: "Think of any polynomial. Don't tell me what it is – just tell me its degree. Then multiply it by (here give a Legendre polynomial of degree more than this). Now integrate it from -1 to $+1$. The answer is zero, right?" (Applause.)

Thus: Think of any polynomial. $3x^2-5x+7$. Now multiply it by $5x^3-3x$. OK, that's $15x^5-25x^4-2x^3+15x^2-21x$. Now integrate it from -1 to $+1$. The answer is zero.

Now, let S be any polynomial of degree less than $2l$. Let us divide it by the Legendre polynomial of degree l, P_l, to obtain the quotient Q and a remainder R, both of degree less than l. Then I assert that

$$\int_{-1}^{1} S\,dx=\int_{-1}^{1} R\,dx. \tag{1.16.25}$$

This follows trivially from Equations 1.16.1 and 1.16.14. Thus

$$\int_{-1}^{1} S dx = \int_{-1}^{1} (Q P_l + R) dx = \int_{-1}^{1} R dx. \tag{1.16.26}$$

Example: Let $S = 6x^5 - 12x^4 + 4x^3 + 7x^2 - 5x + 7$. The integral of this from -1 to $+1$ is 13.86. If we divide S by $\frac{1}{2}(5x^3 - 3x)$, we obtain a quotient of $2.4x^2 - 4.8x + 3.04$ and a remainder of $-0.2x^2 - 0.44x + 7$. The integral of the latter from -1 to $+1$ is also 13.86.

I have just described some properties of Legendre polynomials. Before getting on to the rationale behind Gaussian quadrature, let us remind ourselves from Section 1.11 about Lagrange polynomials. We recall from that section that, if we have a set of n points, the following function:

$$y = \sum_{i=1}^{n} y_i L_i(x) \tag{1.16.27}$$

(in which the n functions $L_i(x)$, $i = 1, n$, are Lagrange polynomials of degree $n-1$) is the polynomial of degree $n-1$ that passes exactly through the n points. Also, if we have some function $f(x)$ which we evaluate at n points, then the polynomial

$$y = \sum_{i=1}^{n} f(x_i) L_i(x) \tag{1.16.28}$$

is a jolly good approximation to $f(x)$ and indeed may be used to interpolate between nontabulated points, even if the function is tabulated at irregular intervals. In particular, if $f(x)$ is a polynomial of degree $n-1$, then the expression 1.16.28is an exact representation of $f(x)$.

We are now ready to start talking about quadrature. We wish to approximate $\int_{-1}^{1} f(x) dx$ by an n-term finite series

$$\int_{-1}^{1} f(x) dx \approx \sum_{i=1}^{n} c_i f(x_i), \tag{1.16.29}$$

where $-1 < x_i < 1$. To this end, we can approximate $f(x)$ by the right hand side of Equation 1.16.28, so that

$$\int_{-1}^{1} f(x) dx \approx \int_{-1}^{1} \sum_{i=1}^{n} f(x_i) L_i(x) dx = f(x_i) \int_{-1}^{1} \sum_{i=1}^{n} L_i(x) dx. \tag{1.16.30}$$

Recall that the Lagrange polynomials in this expression are of degree $n-1$.

The required coefficients for Equation 1.16.29are therefore

$$c_i = \int_{-1}^{1} L_i(x) dx. \tag{1.16.31}$$

Note that at this stage the values of the x_i have not yet been chosen; they are merely restricted to the interval [−1 , 1].

Now let's consider $\int_{-1}^{1} S(x) dx$, where S is a polynomial of degree less than $2n$, such as, for example, the polynomial of Equation 1.16.3. We can write

$$\int_{-1}^{1} S(x) dx = \int_{-1}^{1} \sum_{i=1}^{n} S(x_i) L_i(x) dx = \int_{-1}^{1} \sum_{i=1}^{n} L_i(x) [Q(x_i) P(x_i) + R(x_i)] dx. \tag{1.16.32}$$

Here, as before, P is a polynomial of degree n, and Q and R are of degree less than n.

If we now choose the x_i to be the roots of the Legendre polynomials, then

$$\int_{-1}^{1} S(x) dx = \int_{-1}^{1} \sum_{i=1}^{n} L_i(x) R(x_i) dx. \tag{1.16.33}$$

Note that the integrand on the right hand side of Equation 1.16.33 is an *exact representation of* $R(x)$. But we have already shown (Equation 1.16.26) that $\int_{-1}^{1} S(x)dx = \int_{-1}^{1} R(x)dx$, and therefore

$$\int_{-1}^{1} S(x)dx = \int_{-1}^{1} R(x)dx = \sum_{i=1}^{n} c_i R(x_i) = \sum_{i=1}^{n} c_i S(x_i). \tag{1.16.34}$$

It follows that the Gaussian quadrature method, if we choose the roots of the Legendre polynomials for the n abscissas, will yield exact results for any polynomial of degree less than $2n$, and will yield a good approximation to the integral if $S(x)$ is a polynomial representation of a general function $f(x)$ obtained by fitting a polynomial to several points on the function.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

1.17: Frequently-needed Numerical Procedures

Many years ago I gradually became aware that there were certain mathematical Equations and procedures that I found myself using over and over again. I therefore set aside some time to write short computer programs for dealing with each of them, so that whenever in the future I needed, for example, to evaluate a determinant, I had a program already written to do it. I show here a partial list of the programs I have for instant use by myself whenever needed. I would suggest that the reader might consider compiling for him- or herself a similar collection of small programs. I have found over the years that they have saved me an immense amount of time and effort. Most programs are very short and required only a few minutes to write (although this depends, of course, on how much programming experience one has), though a few required a bit more effort. Some programs are so short – consisting of a few lines only - that they might be thought to be too trivial to be worth writing. These include, for example, programs for solving a quadratic Equation or for solving two simultaneous linear Equations. Yet I have perhaps used these particularly simple ones more than any others, and they have been of use out of all proportion to the almost negligible effort required to write them. Here, then, is a partial list, and I do suggest that the reader will be repaid enormously over the years if he takes a short time to write similar programs. Of course many or even most of them are readily available in prepackaged programs. But there are enormous advantages in writing your own programs. Quite apart from the extra programming practice that they provide, you know exactly what your own programs do, you can tailor them exactly to your own requirements, you know their strengths and their weaknesses or limitations, and you don't have to struggle for hours over an instruction manual trying to understand how to use them, only to find in the end that they don't do exactly what you want.

Solve quadratic Equation
Solve cubic Equation
Solve quintic Equation
Solve $f(x) = 0$ by Newton-Raphson
Solve $f(x, y) = 0$, $g(x, y) = 0$ by Newton-Raphson
Tabulate $y = f(x)$
Tabulate $y = f(x, a)$
Fit least-squares straight line to data
Fit least-squares cubic Equation to data
Solve two simultaneous linear Equations
Solve three simultaneous linear Equations
Solve four simultaneous linear Equations
Solve $N(> 4)$ simultaneous linear Equations in two, three or four unknowns by least squares
Multiply column vector by square matrix
Invert matrix
Diagonalize matrix
Find eigenvectors and eigenvalues of matrix
Test matrix for orthogonality
Evaluate determinant
Convert between rectangular and polar coordinates
Convert between rectangular and spherical coordinates
Convert between direction cosines and Euler angles
Fit a conic section to five points
Numerical integration by Simpson's rule
Gaussian quadrature
Given any three elements of a plane triangle, calculate the remaining elements
Given any three elements of a spherical triangle, calculate the remaining elements

In addition to these common procedures, there are many others that I have written and have readily to hand that are of more specialized use tailored to my own particular interests, such as

Solve Kepler's Equation
Convert between wavelength and wavenumber

Calculate LS-coupling line strengths
Convert between relativity factors such as $\gamma = 1/\sqrt{1-\beta^2}$

Likewise, you will be able to think of many formulas special to your own interests that you use over and over again, and it would be worth your while to write short little programs for them.

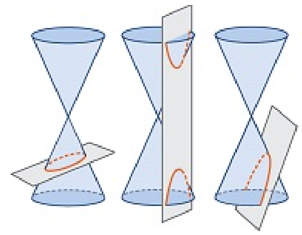
Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

2: CONIC SECTIONS

A particle moving under the influence of an inverse square force moves in an orbit that is a conic section; that is to say an ellipse, a parabola or a hyperbola. We shall prove this from dynamical principles in a later chapter. In this chapter we review the geometry of the conic sections. We start off, however, with a brief review (eight equation-packed pages) of the geometry of the straight line.



2.1: The Straight Line

It might be thought that there is rather a limited amount that could be written about the geometry of a straight line. We can manage a few Equations here, however, (there are 35 in this section on the Straight Line) and we shall return for more on the subject in Chapter 4.

Most readers will be familiar with the Equation for a straight line:

$$y = mx + c \tag{2.2.1}$$

The slope (or gradient) of the line, which is the tangent of the angle that it makes with the x-axis, is m, and the intercept on the y-axis is c. There are various other forms that may be of use, such as

$$\frac{x}{x_0} + \frac{y}{y_0} = 1 \tag{2.2.2}$$

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1} \tag{2.2.3}$$

which can also be written

$$\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0 \tag{2.2.4}$$

$$x\cos\theta + y\sin\theta = p \tag{2.2.5}$$

The four forms are illustrated in figure II.1.

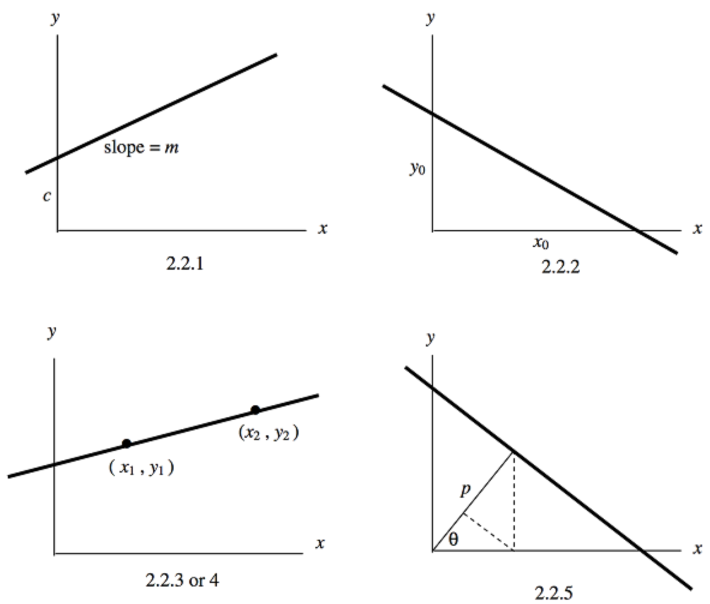


FIGURE II.1

A straight line can also be written in the form

$$Ax + By + C = 0. \tag{2.2.6}$$

If $C = 0$, the line passes through the origin. If $C \neq 0$, no information is lost, and some arithmetic and algebra are saved, if we divide Equation 2.2.6 by C and re-write it in the form

$$ax + by = 1. \tag{2.2.7}$$

Let $P(x, y)$ be a point on the line and let $P_0(x_0, y_0)$ be a point in the plane not necessarily on the line. It is of interest to find the perpendicular distance between P_0 and the line. Let S be the square of the distance between P_0 and P. Then

$$S=(x-x_0)^2+(y-y_0)^2 \tag{2.2.8}$$

We can express this in terms of the single variable x by substitution for y from Equation 2.2.7. Differentiation of S with respect to x will then show that S is least for

$$x=\frac{a+b(bx_0-ay_0)}{a^2+b^2} \tag{2.2.9}$$

The corresponding value for y, found from Equations 2.2.7 and 2.2.9, is

$$y=\frac{b+a(ay_0-bx_0)}{a^2+b^2}. \tag{2.2.10}$$

The point P described by Equations 2.2.9 and 2.2.10 is the closest point to P_0 on the line. The perpendicular distance of P from the line is $p=\sqrt{S}$ or

$$p=\frac{1-ax_0-by_0}{\sqrt{a^2+b^2}}. \tag{2.2.11}$$

This is positive if P_0 is on the same side of the line as the origin, and negative if it is on the opposite side. If the perpendicular distances of two points from the line, as calculated from Equation 2.2.11, are of opposite signs, they are on opposite sides of the line. If $p=0$, or indeed if the numerator of Equation 2.2.11 is zero, the point $P_0(x_0, y_0)$ is, of course, on the line.

Let $A(x_1, y_1)$, $B(x_2, y_2)$ and $C(x_3, y_3)$ be three points in the plane. What is the area of the triangle ABC? One way to answer this is suggested by figure II.2.

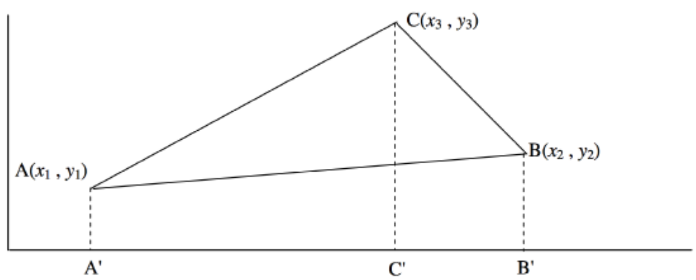


FIGURE II.2

We see that

area of triangle ABC = area of trapezium A′ACC′ (see comment*)
+ area of trapezium C′CBB′
− area of trapezium A′ABB′.

$$=\frac{1}{2}(x_3-x_1)(y_3+y_1)+\frac{1}{2}(x_2-x_3)(y_2+y_3)-\frac{1}{2}(x_2-x_1)(y_2+y_1) \tag{2.1.1}$$

$$=\frac{1}{2}[x_1(y_2-y_3)+x_2(y_3-y_1)+x_3(y_1-y_2)] \tag{2.1.2}$$

$$=\frac{1}{2}\begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix} \tag{2.2.12}$$

* Since writing this section I have become aware of a difference in U.S./British usages of the word "trapezium". Apparently in British usage, "trapezium" means a quadrilateral with two parallel sides. In U.S. usage, a trapezium means a quadrilateral with no parallel sides, while a quadrilateral with two parallel sides is a "trapezoid". As with many words, either British or U.S. usages may be heard in Canada. In the above

derivation, I intended the British usage. What is to be learned from this is that we must always take care to make ourselves clearly understood when using such ambiguous words, and not to assume that the reader will interpret them the way we intend.

The reader might like to work through an alternative method, using results that we have obtained earlier. The same result will be obtained. In case the algebra proves a little tedious, it may be found easier to work through a numerical example, such as: calculate the area of the triangle ABC, where A, B, C are the points (2,3), (7,4), (5,6) respectively. In the second method, we note that the area of a triangle is $\frac{1}{2}\times\text{base}\times\text{height}$. Thus, if we can find the length of the side BC, and the perpendicular distance of A from BC, we can do it. The first is easy:

$$(\mathrm{BC})^2=(x_3-x_2)^2+(y_3-y_2)^2. \tag{2.2.13}$$

To find the second, we can easily write down the Equation to the line BC from Equation 2.2.3, and then re-write it in the form 2.2.7. Then Equation 2.2.11 enables us to find the perpendicular distance of A from BC, and the rest is easy.

If the determinant in Equation 2.2.12 is zero, the area of the triangle is zero. This means that the three points are collinear.

The angle between two lines

$$y=m_1x+c_1 \tag{2.2.14}$$

and

$$y=m_2x+c_2 \tag{2.2.15}$$

is easily found by recalling that the angles that they make with the x-axis are $\tan^{-1}\ m_1$ and $\tan^{-1}\ m_2$ together with the elementary trigonometry formula $\tan(A-B)=(\tan A-\tan B)/(1+\tan A\tan B)$. It is then clear that the tangent of the angle between the two lines is

$$\frac{m_2-m_1}{1+m_1m_2}. \tag{2.2.16}$$

The two lines are at right angles to each other if

$$m_1m_2=-1 \tag{2.2.17}$$

The line that bisects the angle between the lines is the locus of points that are equidistant from the two lines. For example, consider the two lines

$$-2x+5y=1 \tag{2.2.18}$$

$$30x-10y=1 \tag{2.2.19}$$

Making use of Equation 2.2.11, we see that a point (x,y) is equidistant from these two lines if

$$\frac{1+2x-5y}{\sqrt{29}}=\pm\frac{1-30x+10y}{\sqrt{1000}}. \tag{2.2.20}$$

The significance of the $\pm$ will become apparent shortly. The + and − choices result, respectively, in

$$-8.568x+8.079y=1 \tag{2.2.21}$$

and

$$2.656x+2.817y=1. \tag{2.2.22}$$

The two continuous lines in figure $\mathrm{II}.3$ are the lines 2.2.18 and 2.2.19. There are two bisectors, represented by Equations 2.2.21 and 2.2.22, shown as dotted lines in the figure, and they are at right angles to each other. The choice of the + sign in Equation 2.2.20 (which in this case results in Equation 2.2.21, the bisector in figure $\mathrm{II}.3$ with the positive slope) gives the bisector of the sector that contains the origin.

An Equation of the form

$$ax^2+2hxy+by^2=0 \tag{2.2.23}$$

can be factored into two linear factors with no constant term, and it therefore represents two lines intersecting at the origin. It is left as an exercise to determine the angles that the two lines make with each other and with the x axis, and to show that the lines

$$x^2+\left(\frac{a-b}{h}\right)xy-y^2=0 \tag{2.2.24}$$

are the bisectors of 2.2.23 and are perpendicular to each other.

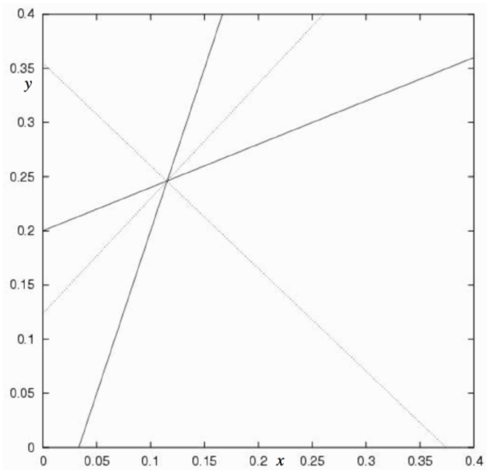


FIGURE II.3

Given the Equations to three straight lines, can we find the area of the triangle bounded by them? To find a general algebraic expression might be a bit tedious, though the reader might like to try it, but a numerical example is straightforward. For example, consider the lines

$$x-5y+12=0, \tag{2.2.25}$$

$$3x+4y-9=0, \tag{2.2.26}$$

$$3x-y-3=0. \tag{2.2.27}$$

By solving the Equations in pairs, it is soon found that they intersect at the points (−0.15789, 2.36842), (1.4, 1.2) and (1.92857, 2.78571). Application of Equation 2.2.12 then gives the area as 1.544. The triangle is drawn in figure II.4. Measure any side and the corresponding height with a ruler and see if the area is indeed about 1.54.

But now consider the three lines

$$x-5y+12=0, \tag{2.2.28}$$

$$3x+4y-9=0, \tag{2.2.29}$$

$$3x+23y-54=0. \tag{2.2.30}$$

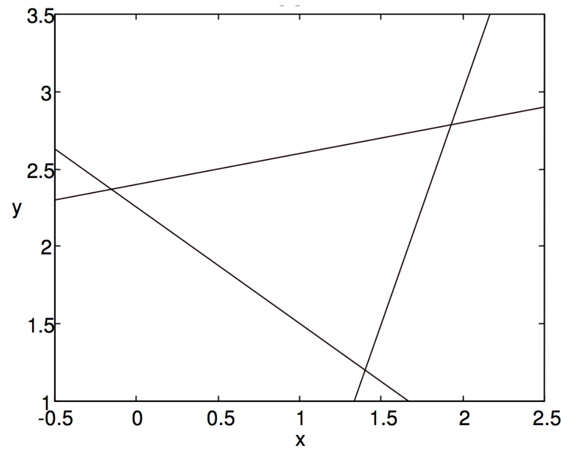


FIGURE II.4

By solving the Equations in pairs, it will be found that all three lines intersect at the same point (please do this), and the area of the triangle is, of course, zero. Any one of these Equations is, in fact, a linear combination of the other two. You should draw these three lines accurately on graph paper (or by computer). In general, if three lines are

$$A_1x + B_1y + C_1 = 0 \tag{2.2.31}$$

$$A_2x + B_2y + C_2 = 0 \tag{2.2.32}$$

$$A_3x + B_3y + C_3 = 0 \tag{2.2.33}$$

they will be concurrent at a single point if

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} = 0. \tag{2.2.34}$$

Thus the determinant in Equation 2.2.12 provides a test of whether three points are collinear, and the determinant in Equation 2.2.34 provides a test of whether three lines are concurrent.

Finally - at least for the present chapter - there may be rare occasion to write the Equation of a straight line in polar coordinates. It should be evident from figure II.5 that the Equations

$$r = p\csc(\theta - \alpha) \text{ or } r = p\csc(\alpha - \theta) \tag{2.2.35}$$

describe a straight line passing at a distance p from the pole and making an angle α with the initial line. If $p = 0$, the polar Equation is merely $\theta = \alpha$.

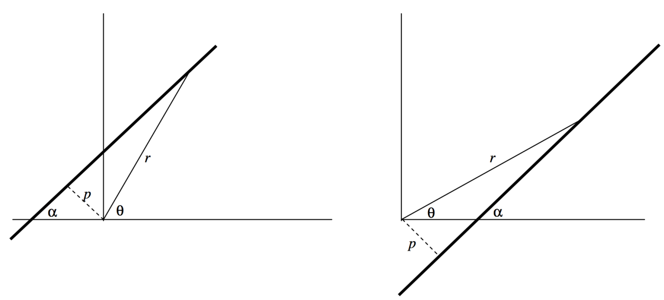


FIGURE II.5

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

2.2: The Ellipse

An ellipse is a figure that can be drawn by sticking two pins in a sheet of paper, tying a length of string to the pins, stretching the string taut with a pencil, and drawing the figure that results. During this process, the sum of the two distances from pencil to one pin and from pencil to the other pin remains constant and equal to the length of the string. This method of drawing an ellipse provides us with a formal definition, which we shall adopt in this chapter, of an ellipse, namely:

An ellipse is the locus of a point that moves such that the sum of its distances from two fixed points called the foci is constant (see figure II.6).

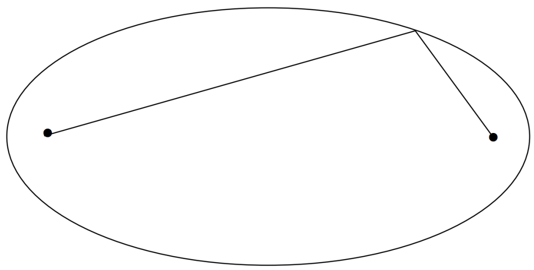


FIGURE II.6

We shall call the sum of these two distances (i.e the length of the string) $2a$. The ratio of the distance between the foci to length of the string is called the *eccentricity* e of the ellipse, so that the distance between the foci is $2ae$, and e is a number between 0 and 1.

The longest axis of the ellipse is its major axis, and a little bit of thought will show that its length is equal to the length of the string; that is, $2a$. The shortest axis is the minor axis, and its length is usually denoted by $2b$. The eccentricity is related to the ratio b/a in a manner that we shall shortly discuss.

The ratio

$$\eta = (a-b)/a \tag{2.2.1}$$

is called the *ellipticity* if the ellipse. It is merely an alternative measure of the noncircularity. It is related to the eccentricity, and we shall obtain that relation shortly, too. Until then, Figure $\text{II}.7$ shows pictorially the relation between the two.

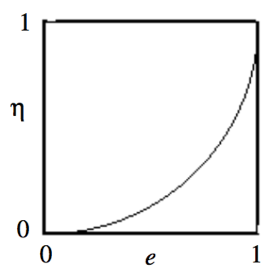


FIGURE II.7

We shall use our definition of an ellipse to obtain its Equation in rectangular coordinates. We shall place the two foci on the x-axis at coordinates $(-ae, 0)$ and $(ae, 0)$ (see figure II.8).

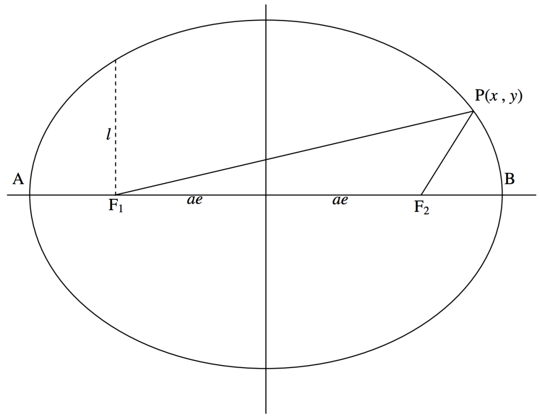


FIGURE II.8

The definition requires that $\mathbf{PF}_1 + \mathbf{PF}_2 = 2a$. That is:

$$\left[(x+ae)^2+y^2\right]^{\frac{1}{2}}+\left[(x-ae)^2+y^2\right]^{\frac{1}{2}}=2a, \tag{2.3.1}$$

and this is the Equation to the ellipse. The reader should be able, after a little bit of slightly awkward algebra, to show that this can be written more conveniently as

$$\frac{x^2}{a^2}+\frac{y^2}{a^2(1-e^2)}=1. \tag{2.3.2}$$

By putting $x=0$, it is seen that the ellipse intersects the y-axis at $\pm a\sqrt{1-e^2}$ and therefore that $a\sqrt{1-e^2}$ is equal to the semi minor axis b. Thus we have the familiar Equation to the ellipse

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{2.3.3}$$

as well as the important relation between a, b and e:

$$b^2=a^2(1-e^2) \tag{2.3.4}$$

The reader can also now derive the relation between ellipticity η and eccentricity e:

$$\eta=1-\sqrt{(1-e^2)}. \tag{2.3.5}$$

This can also be written

$$e^2=\sqrt{\eta(2-\eta)} \tag{2.3.6}$$

or

$$e^2+(\eta-1)^2=1. \tag{2.3.7}$$

This shows, incidentally, that the graph of η versus e, which we have drawn in figure II.7, is part of a circle of radius 1 centred at $e=0,\ \eta=1$.

In figures II.9 I have drawn ellipses of eccentricities 0.1 to 0.9 in steps of 0.1, and in figure II.10 I have drawn ellipses of ellipticities 0.1 to 0.9 in steps of 0.1. You may find that ellipticity gives you a better idea than eccentricity of the noncircularity of an ellipse. For an exercise, you should draw in the positions of the foci of each of these ellipses, and decide whether eccentricity or ellipticity gives you a better idea of the "ex-centricity" of the foci. Note that the eccentricities of the orbits of Mars and Mercury are, respectively, about 0.1 and 0.2 (these are the most eccentric of the planetary orbits except for comet-like Pluto), and it is difficult for the eye to see that they depart at all from circles - although, when the foci are drawn, it is obvious that the foci are "ex-centric".

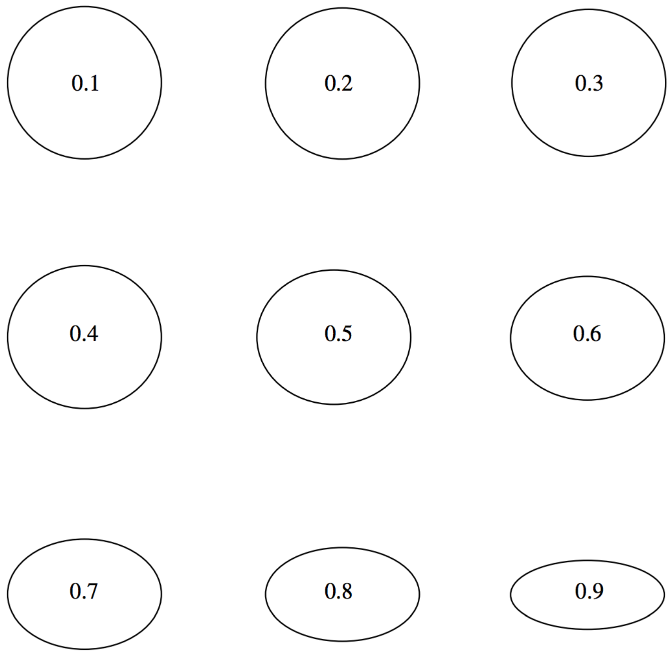


FIGURE II.9: The number inside each ellipse is its eccentricity.

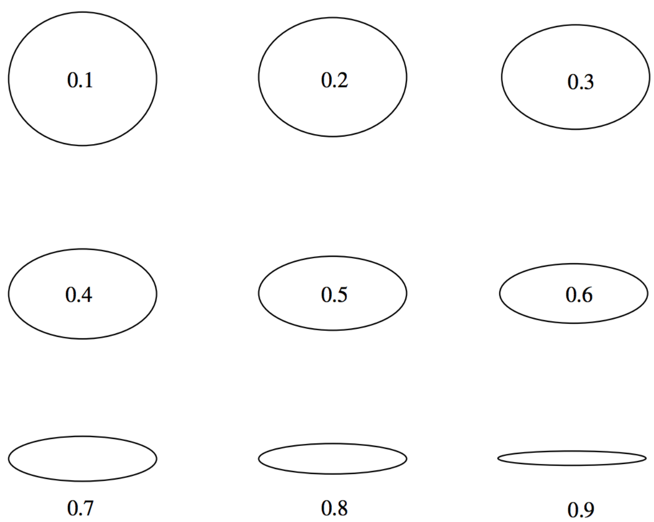


FIGURE II.10: The figure inside or below each ellipse is its ellipticity.

In the theory of planetary orbits, the Sun will be at one focus. Let us suppose it to be at F_2 (see figure II.8). In that case the distance $\mathrm{F}_2\mathrm{B}$ is the perihelion distance q, and is equal to

$$q = a(1-e). \tag{2.3.8}$$

The distance $\mathrm{F}_2\mathrm{A}$ is the aphelion distance Q (pronounced ap-helion by some and affelion by others − and both have defensible positions), and it is equal to

$$Q = a(1+e). \tag{2.3.9}$$

A line parallel to the minor axis and passing through a focus is called a *latus rectum* (plural: *latera recta*). The length of a semi latus rectum is commonly denoted by l(sometimes by p). Its length is obtained by putting $x = ae$ in the Equation to the ellipse, and it will be readily found that

$$l = a(1-e^2). \tag{2.3.10}$$

The length of the semi latus rectum is an important quantity in orbit theory. It will be found, for example, that the energy of a planet is closely related to the semi major axis a of its orbit, while its angular momentum is closely related to the semi latus rectum.

The circle whose diameter is the major axis of the ellipse is called the *eccentric circle* or, preferably, the *auxiliary circle* (figure $\text{II}.11$). Its Equation is

$$x^2 + y^2 = a^2. \tag{2.3.11}$$

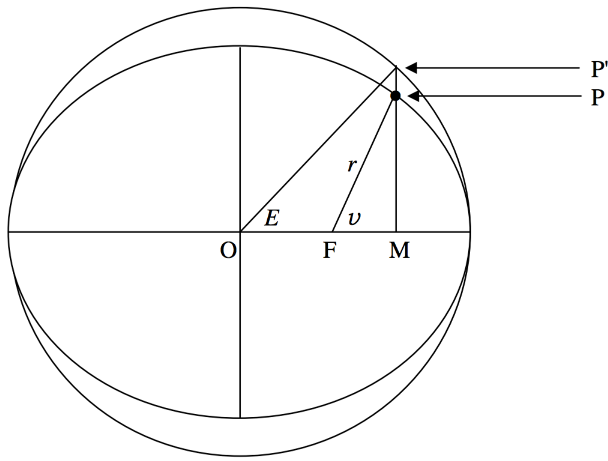


FIGURE II.11

In orbit theory the angle v (denoted by f by some authors) is called the true anomaly of a planet in its orbit. The angle E is called the *eccentric anomaly*, and it is important to find a relation between them.

We first note that, if the eccentric anomaly is E, the abscissas of P' and of P are each $a\cos E$. The ordinate of P' is $a\sin E$. By putting $x = a\cos E$ in the Equation to the ellipse, we immediately find that the ordinate of P is $b\sin E$. Several deductions follow. One is that any point whose abscissa and ordinate are of the form

$$x = a\cos E, \quad y = b\sin E \tag{2.3.12}$$

is on an ellipse of semi major axis a and semi minor axis b. These two Equations can be regarded as parametric Equations to the ellipse. They can be used to describe an ellipse just as readily as

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{2.3.13}$$

and indeed this Equation is the E-eliminant of the parametric Equations.

The ratio $\text{PM}/\text{P}'\text{M}$ for any line perpendicular to the major axis is b/a. Consequently the area of the ellipse is b/a times the area of the auxiliary circle; and since the area of the auxiliary circle is πa^2, it follows that the area of the ellipse is πab.

In figure $\text{II}.11$, the distance r is called the *radius vector* (plural *radii vectores*), and from the theorem of Pythagoras its length is given by

$$r^2 = b^2\sin^2 E + a^2(\cos E - e)^2. \tag{2.3.14}$$

On substituting $1 - \cos^2 E$ for $\sin^2 E$ and $a^2(1 - e^2)$ for b^2, we soon find that

$$r = a(1 - e\cos E) \tag{2.3.15}$$

It then follows immediately that the desired relation between v and E is

$$\cos v = \frac{\cos E - e}{1 - e\cos E}. \tag{2.3.16}$$

From trigonometric identities, this can also be written

$$\sin v = \frac{\sqrt{1-e^2}\sin E}{1-e\cos E} \tag{2.3.17a}$$

or

$$\tan v = \frac{\sqrt{1-e^2}\sin E}{\cos E - e} \tag{2.3.17b}$$

or

$$\tan\frac{1}{2}v = \sqrt{\frac{1+e}{1-e}}\tan\frac{1}{2}E. \tag{2.3.17c}$$

The inverse formulas may also be useful:

$$\cos E = \frac{e+\cos v}{1+e\cos v} \tag{2.3.17d}$$

$$\sin E = \frac{\sin v\sqrt{1-e^2}}{e+\cos v} \tag{2.3.17e}$$

$$\tan E = \frac{\sin v\sqrt{1-e^2}}{e+\cos v} \tag{2.3.17f}$$

$$\tan\frac{1}{2}E = \sqrt{\frac{1-e}{1+e}}\tan\frac{1}{2}v. \tag{2.3.17g}$$

There are a number of miscellaneous geometric properties of an ellipse, some, but not necessarily all, of which may prove to be of use in orbital calculations. We describe some of them in what follows.

Tangents to an Ellipse

Find where the straight line $y = mx + c$ intersects the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \tag{2.3.18}$$

The answer to this question is to be found by substituting $mx + c$ for y in the Equation to the ellipse. After some rearrangement, a quadratic Equation in x results:

$$\left(a^2m^2 + b^2\right)x^2 + 2a^2cmx + a^2\left(c^2 - b^2\right) = 0. \tag{2.3.19}$$

If this Equation has two real roots, the roots are the x-coordinates of the two points where the line intersects the ellipse. If it has no real roots, the line misses the ellipse. If it has two coincident real roots, the line is tangent to the ellipse. The condition for this is found by setting the discriminant of the quadratic Equation to zero, from which it is found that

$$c^2 = a^2m^2 + b^2. \tag{2.3.20}$$

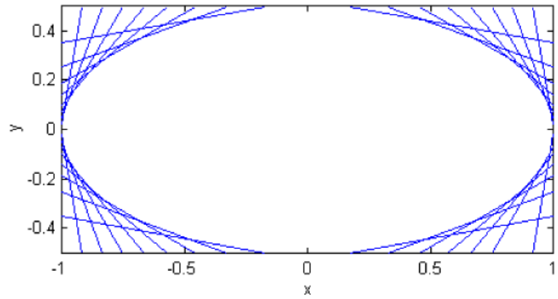
Thus a straight line of the form

$$y = mx \pm \sqrt{a^2m^2 + b^2} \tag{2.3.21}$$

is tangent to the ellipse.

Figure II.12 shows several such lines, for $a = 2b$ and slopes $(\tan^{-1} m)$ of 0° to 180° in steps of 10°

FIGURE II.12



Director Circle

The Equation we have just derived for a tangent to the ellipse can be rearranged to read

$$m^2\left(a^2-x^2\right)+2mx+b^2-y^2=0. \tag{2.3.22}$$

Now the product of the slopes of two lines that are at right angles to each other is -1 (Equation 2.2.17). Therefore, if we replace m in the above Equation by $-1/m$ we shall obtain another tangent to the ellipse, at right angles to the first one. The Equation to this second tangent becomes (after multiplication throughout by m)

$$m^2\left(b^2-y^2\right)-2mx+a^2-x^2=0. \tag{2.3.23}$$

If we eliminate m from these two Equations, we shall obtain an Equation in x and y that describes the point where the two perpendicular tangents meet; that is, the Equation that will describe a curve that is the locus of the point of intersection of two perpendicular tangents. It turns out that this curve is a circle of radius $\sqrt{a^2+b^2}$, and it is called the *director circle*.

It is easier than it might first appear to eliminate m from the Equations. We merely have to add the Equations 2.3.22 and 2.3.23:

$$m^2\left(a^2+b^2-x^2-y^2\right)+a^2+b^2-x^2-y^2=0. \tag{2.3.24}$$

For real m, this can only be if

$$x^2+y^2=a^2+b^2, \tag{2.3.25}$$

which is the required locus of the director circle of radius $\sqrt{a^2+b^2}$. It is illustrated in figure II.13.

We shall now derive an Equation to the line that is tangent to the ellipse at the point $(x_1,\ y_1)$.

Let $(x_1,\ y_1)=(a\cos E_1, b\sin E_1)$ and $(x_2,\ y_2)=(a\cos E_2, b\sin E_2)$ be two points on the ellipse.

The line joining these two points is

$$\frac{y-b\sin E_1}{x-a\cos E_1}=\frac{b(\sin E_2-\sin E_1)}{a(\cos E_2-\cos E_1)}=\frac{2b\cos\frac{1}{2}(E_2+E_1)\sin\frac{1}{2}(E_2-E_1)}{-2a\sin\frac{1}{2}(E_2+E_1)\sin\frac{1}{2}(E_2-E_1)}=-\frac{b\cos\frac{1}{2}(E_2+E_1)}{a\sin\frac{1}{2}(E_2+E_1)}. \tag{2.3.26}$$

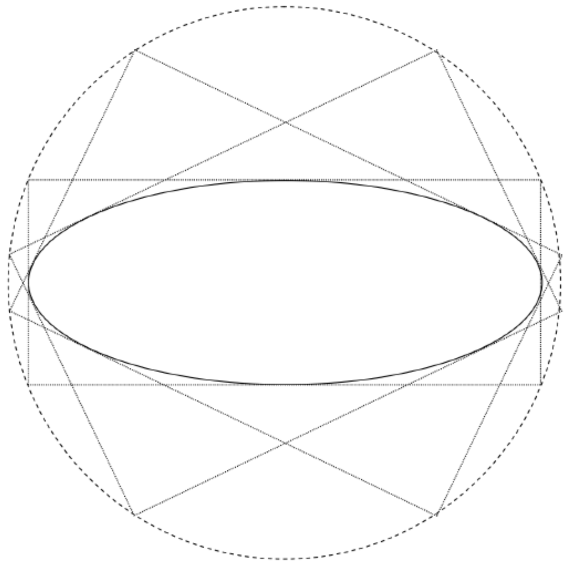


FIGURE II.13

Now let E_2 approach E_1, eventually coinciding with it. The resulting Equation

$$\frac{y-b\sin E}{x-a\cos E}=-\frac{b\cos E}{a\sin E}, \tag{2.3.27}$$

in which we no longer distinguish between E_1 and E_2, is the Equation of the straight line that is tangent to the ellipse at $(a\cos E\,, b\sin E)$. This can be written

$$\frac{x\cos E}{a}+\frac{y\sin E}{b}=1 \tag{2.3.28}$$

or, in terms of $(x_1,\ y_1)$,

$$\frac{x_1 x}{a^2}+\frac{y_1 y}{b^2}=1, \tag{2.3.29}$$

which is the tangent to the ellipse at $(x_1,\ y_1)$.

An interesting property of a tangent to an ellipse, the proof of which I leave to the reader, is that $\mathrm{F_1P}$ and $\mathrm{F_2P}$ make equal angles with the tangent at P. If the inside of the ellipse were a reflecting mirror and a point source of light were to be placed at $\mathrm{F_1}$, it would be imaged at $\mathrm{F_2}$. (Have a look at figure $\mathrm{II}.6$ or $\mathrm{II}.8$.) This has had an interesting medical application. A patient has a kidney stone. The patient is asked to lie in an elliptical bath, with the kidney stone at $\mathrm{F_2}$. A small explosion is detonated at $\mathrm{F_1}$; the explosive sound wave emanating from $\mathrm{F_1}$ is focused as an implosion at $\mathrm{F_2}$ and the kidney stone at $\mathrm{F_2}$ is shattered. Don't try this at home.

Directrices

The two lines $x=\pm a/e$ are called the directrices (singular *directrix*) of the ellipse (figure $\mathrm{II}.14$).

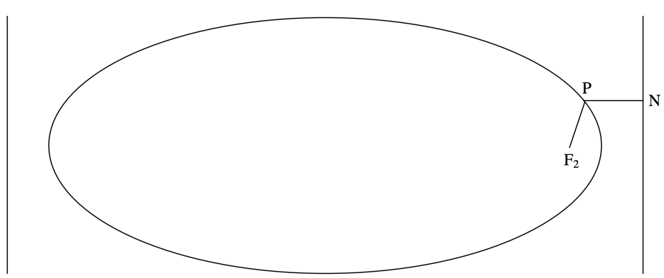


FIGURE II.14

The ellipse has the property that, for any point P on the ellipse, the ratio of the distance PF_2 to a focus to the distance PN to a directrix is constant and is equal to the eccentricity of the ellipse. Indeed, this property is sometimes used as the definition of an ellipse, and all the Equations and properties that we have so far derived can be deduced from such a definition. We, however, adopted a different definition, and the focus-directrix property must be derived. This is straightforward, for, (recalling that the abcissa of F_2 is ae) we see from figure II.14 that the square of the desired ratio is

$$\frac{(x-ae)^2+y^2}{(a/e-x)^2}. \tag{2.3.30}$$

On substitution of

$$b^2\left(1-\left(\frac{x}{a}\right)^2\right)=a^2\left(1-e^2\right)\left(1-\left(\frac{x}{a}\right)^2\right)=\left(1-e^2\right)\left(a^2-x^2\right) \tag{2.3.31}$$

for y^2, the above expression is seen to reduce to e^2.

Another interesting property of the focus and directrix, although a property probably with not much application to orbit theory, is that if the tangent to an ellipse at a point P intersects the directrix at Q, then P and Q subtend a right angle at the focus. (See figure II.15).

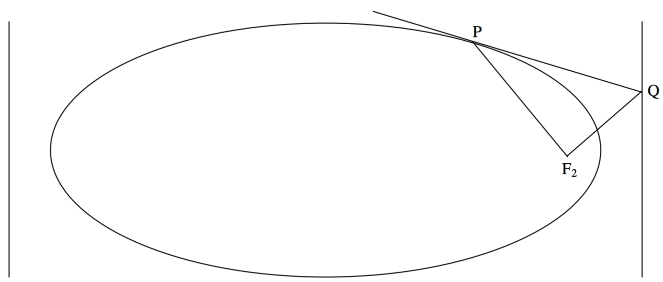


FIGURE II.15

Thus the tangent at $P=(x_1,\ y_1)$ is

$$\frac{x_1x}{a^2}+\frac{y_1y}{b^2}=1 \tag{2.3.32}$$

and it is straightforward to show that it intersects the directrix $x=a/e$ at the point

$$\left(\frac{a}{e},\frac{b^2}{y_1}\left(1-\frac{x_1}{ae}\right)\right). \tag{2.2.2}$$

The coordinates of the focus F_2 are $(ae,0)$. The slope of the line PF_2 is $(x_1-ae)/y_1$ and the slope of the line QF_2 is

$$\frac{\frac{b^2}{y_1}\left(1-\frac{x_1}{ae}\right)}{\frac{a}{e}-ae}. \tag{2.2.3}$$

It is easy to show that the product of these two slopes is -1, and hence that PF_2 and QF_2 are at right angles.

Conjugate Diameters

The left hand of figure II.16 shows a circle and two perpendicular diameters. The right hand figure shows what the circle would look like when viewed at some oblique angle. The circle has become an ellipse, and the diameters are no longer perpendicular. The diameters are called *conjugate diameters* of the ellipse. One is conjugate to the other, and the other is conjugate to the one. They have the property - or the definition - that each bisects all chords parallel to the other, because this property of bisection, which is obviously held by the perpendicular diameters of the circle, is unaltered in projection.

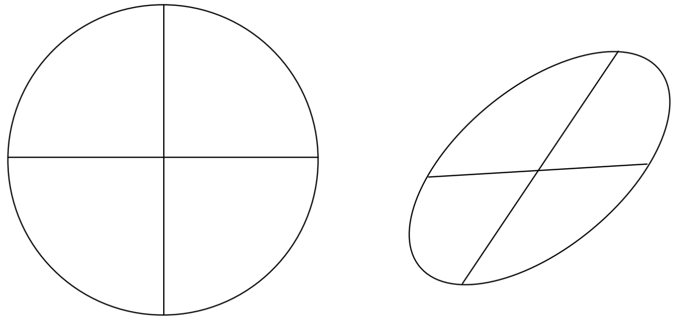


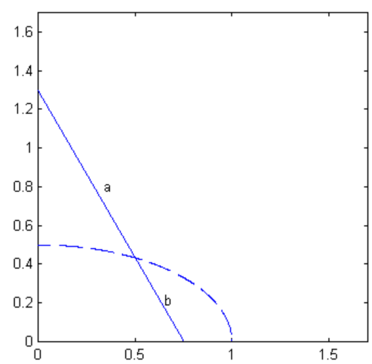
FIGURE II.16

It is easy to draw two conjugate diameters of an ellipse of eccentricity e either by making use of this last-mentioned property or by noting that that the product of the slopes of two conjugate diameters is e^2-1. The proof of this is left for the enjoyment of the reader.

A Ladder Problem.

No book on elementary applied mathematics is complete without a ladder problem. A ladder of length $a+b$ leans against a smooth vertical wall and a smooth horizontal floor. A particular rung is at a distance a from the top of the ladder and b from the bottom of the ladder. Show that, when the ladder slips, the rung describes an ellipse. (This result will suggest another way of drawing an ellipse.) See figure $\text{II}.17$.

FIGURE II.17



If you have not done this problem after one minute, here is a hint. Let the angle that the ladder makes with the floor at any time be E. That is the end of the hint.

The reader may be aware that some of the geometrical properties that we have discussed in the last few paragraphs are more of recreational interest and may not have much direct application in the theory of orbits. In the next subsection we return to properties and Equations that are very relevant to orbital theory - perhaps the most important of all for the orbit computer to understand.

Polar Equation to the Ellipse

We shall obtain the Equation in polar coordinates to an ellipse whose focus is the pole of the polar coordinates and whose major axis is the initial line $(\theta=0^\circ)$ of the polar coordinates. In figure $\text{II}.18$ we have indicated the angle θ of polar coordinates, and it may occur to the reader that we have previously used the symbol v for this angle and called it the true anomaly. Indeed at present, v and θ are identical, but a little later we shall distinguish between them.

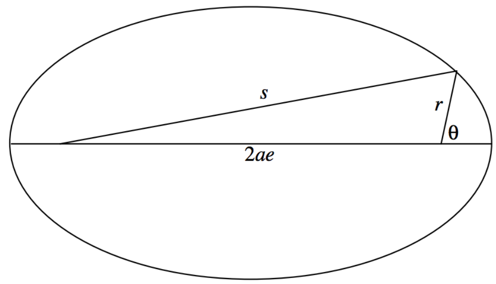


FIGURE II.18

From our definition of the ellipse, $s=2a-r$, and so

$$s^2=4a^2-4ar+r^2. \tag{2.3.33}$$

From the cosine formula for a plane triangle,

$$s^2=4a^2e^2+r^2+4aer\cos\theta. \tag{2.3.34}$$

On equating these expressions we soon obtain

$$a\left(1-e^2\right)=r\left(1+e\cos\theta\right). \tag{2.3.35}$$

The left hand side is equal to the semi latus rectum l, and so we arrive at the polar Equation to the ellipse, focus as pole, major axis as initial line:

$$r=\frac{1}{1+e\cos\theta}. \tag{2.3.36}$$

If the major axis is inclined at an angle ω to the initial line (figure II.19), the Equation becomes

$$r=\frac{l}{1+e\cos(\theta-\omega)}=\frac{l}{1+e\cos v}. \tag{2.3.37}$$

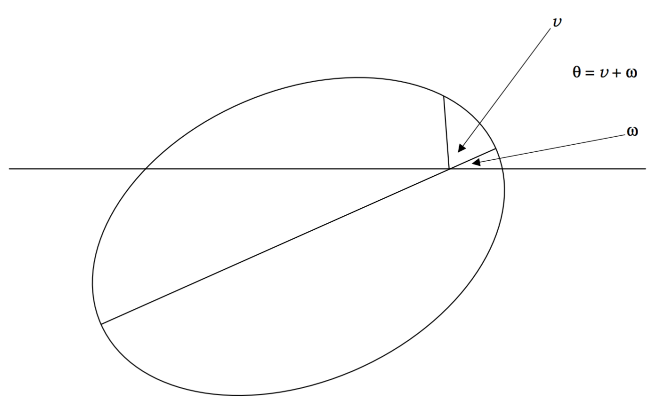


FIGURE II.19

The distinction between θ and v is now evident. θ is the angle of polar coordinates, ω is the angle between the major axis and the initial line (ω will be referred to in orbital theory as the "argument of perihelion"), and v, the true anomaly, is the angle between the radius vector and the initial line.

Contributors and Attributions

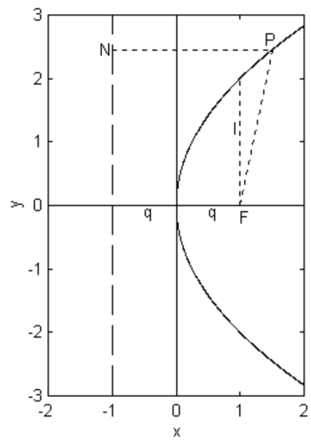
- Jeremy Tatum (University of Victoria, Canada)

2.3: The Parabola

We define a parabola as the locus of a point that moves such that its distance from a fixed straight line called the *directrix* is equal to its distance from a fixed point called the *focus*. Unlike the ellipse, a parabola has only one focus and one directrix. However, comparison of this definition with the focus - directrix property of the ellipse (which can also be used to define the ellipse) shows that the parabola can be regarded as a limiting form of an ellipse with eccentricity equal to unity.

We shall find the Equation to a parabola whose directrix is the line $y=-q$ and whose focus is the point $(q, 0)$. Figure $\mathrm{II}.20$ shows the parabola. F is the focus and O is the origin of the coordinate system. The vertex of the parabola is at the origin. In an orbital context, for example, the orbit of a comet moving around the Sun in parabolic orbit, the Sun would be at the focus F, and the distance between vertex and focus would be the perihelion distance, for which the symbol q is traditionally used in orbit theory.

FIGURE II.20



From figure $\mathrm{II}.20$, it is evident that the definition of the parabola $(\mathrm{PF}=\mathrm{PN})$ requires that

$$(x-q)^{2}+y^{2}=(x+q)^{2}, \tag{2.3.1}$$

from which

$$y^{2}=4 q x, \tag{2.3.2}$$

which is the Equation to the parabola.

Exercise 2.3.1

Sketch the following parabolas:

a. $y^{2}=-4 q x$
b. $x^{2}=4 q y$
c. $x^{2}=-4 q y,$
d. $(y-2)^{2}=4 q(x-3)$.

The line parallel to the y-axis and passing through the focus is the *latus rectum*. Substitution of $x=q$ into $y^{2}=4 a x$ shows that the latus rectum intersects the parabola at the two points $(q, \pm 2 q)$, and that the length l of the semi latus rectum is $2 q$.

The Equations

$$x=q t^{2}, \quad y=2 q t \tag{2.3.3}$$

are the parametric Equations to the parabola, for $y^{2}=4 q x$ results from the elimination of t between them. In other words, if t is any variable, then any point that satisfies these two Equations lies on the parabola.

Most readers will know that if a particle is moving with constant speed in one direction and constant acceleration at right angles to that direction, as with a ball projected in a uniform gravitational field or an electron moving in a uniform electric field, the path is a parabola. In the constant speed direction the distance is proportional to the time, and in the constant acceleration direction, the distance is proportional to the square of the time, and hence the path is a parabola.

Tangents to a Parabola.

Where does the straight line $y = mx + c$ intersect the parabola $y^2 = 4qx$? The answer is found by substituting $mx + c$ for y to obtain, after rearrangement,

$$m^2x^2 + 2(mc - 2q)x + c^2 = 0. \tag{2.3.4}$$

The line is tangent if the discriminant is zero, which leads to

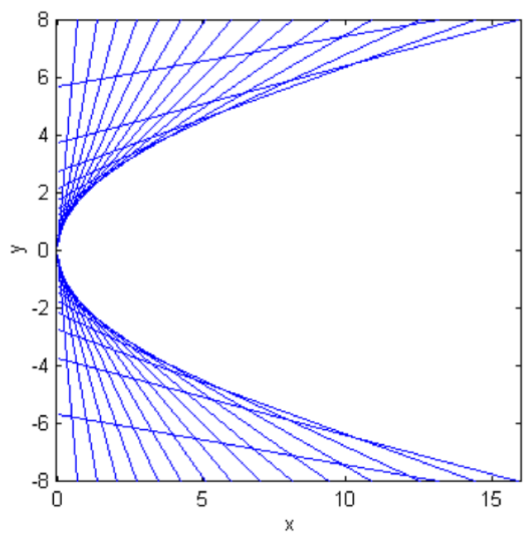
$$c = q/m. \tag{2.3.5}$$

Thus a straight line of the form

$$y = mx + q/m \tag{2.3.6}$$

is tangent to the parabola. Figure $\text{II}.22$ illustrates this for several lines, the slopes of each differing by 5° from the next.

FIGURE II.22



We shall now derive an Equation to the line that is tangent to the parabola at the point (x_1, y_1).

Let $(x_1, y_1) = (qt_1^2, 2qt_1)$ be a point on the parabola, and
Let $(x_2, y_2) = (qt_2^2, 2qt_2)$ be another point on the parabola.

The line joining these two points is

$$\frac{y - 2qt_1}{x - qt_1^2} = \frac{2q(t_2 - t_1)}{q(t_2^2 - t_1^2)} = \frac{2}{t_2 + t_1}. \tag{2.3.7}$$

Now let t_2 approach t_1, eventually coinciding with it. Putting $t_1 = t_2 = t$ in the last Equation results, after simplification, in

$$ty = x + qt^2, \tag{2.3.8}$$

being the Equation to the tangent at $(qt^2, 2qt)$.

Multiply by $2q$:

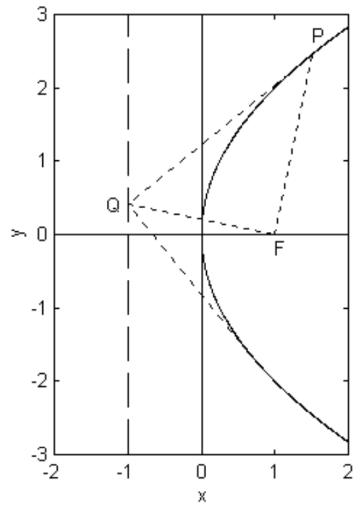
$$2qty = 2q(x + qt^2) \tag{2.3.9}$$

and it is seen that the Equation to the tangent at (x_1, y_1) is

$$y_1 y = 2q(x_1 + x). \tag{2.3.10}$$

There are a number of interesting geometric properties, some of which are given here. For example, if a tangent to the parabola at a point P meets the directrix at Q, then, just as for the ellipse, P and Q subtend a right angle at the focus (figure $\text{II}.23$). The proof is similar to that given for the ellipse, and is left for the reader.

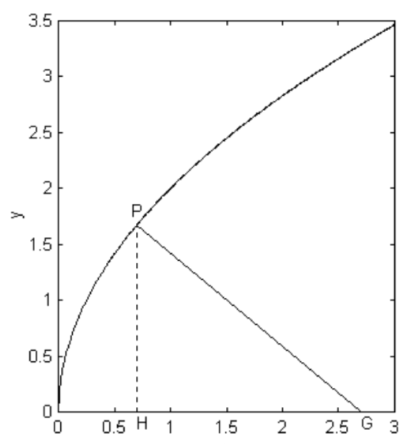
FIGURE II.23



The reader will recall that perpendicular tangents to an ellipse meet on the director circle. The analogous theorem *vis-à-vis* the parabola is that perpendicular tangents meet on the directrix. This is also illustrated in figure $\text{II}.23$. The theorem is not specially important in orbit theory, and the proof is also left to the reader.

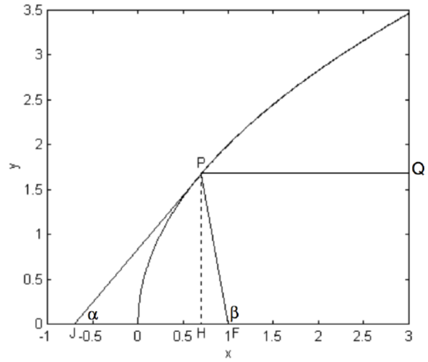
Let PG be the normal to the parabola at point P, meeting the axis at G (figure $\text{II}.24$). We shall call the length GH the subnormal. A curious property is that the length of GH is always equal to l, the length of the semi latus rectum (which in figure $\text{II}.24$ is of length 2 − i.e. the ordinate where $x = 1$), irrespective of the position of P. This proof again is left to the reader.

FIGURE II.24



The following two geometrical properties, while not having immediate applications to orbit theory, certainly have applications to astronomy.

FIGURE II.25



The tangent at P makes an angle α with the x-axis, and PF makes an angle β with the x-axis (figure $\mathrm{II}.25$). We shall show that $\beta=2\alpha$ and deduce an interesting consequence.

The Equation to the tangent (see Equation 2.3.8) is $ty=x+qt^2$, which shows that

$$\tan\alpha=1/t. \tag{2.3.11}$$

The coordinates of P and F are, respectively, $\left(qt^2,2qt\right)$ and $(q,0)$, and so, from the triangle PFH, we find.

$$\tan\beta=\frac{2t}{t^2-1}. \tag{2.3.12}$$

Let $\tau=1/t$, then $\tan\alpha=\tau$ and $\tan\beta=2\tau/(1-\tau^2)$, which shows that $\beta=2\alpha$.

This also shows that triangle JFP is isosceles, with the angles at J and P each being α. This can also be shown as follows.

From the Equation $ty=x+qt^2$, we see that J is the point $(-qt^2,0)$, so that $\mathrm{JF}=q(t^2+1)$.

From triangle PFH, we see that

$$(\mathrm{PF})^2=4q^2t^2+q^2\left(t^2-1\right)^2-q^2\left(t^2+1\right)^2. \tag{2.3.13}$$

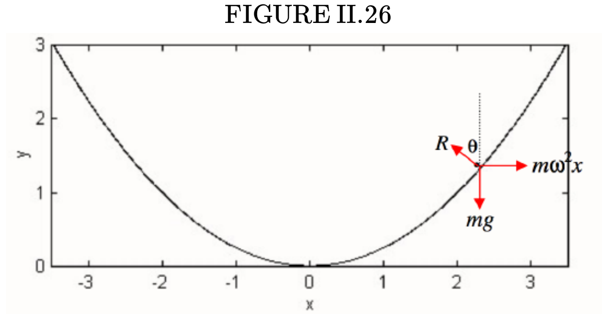
Therefore

$$\mathrm{PF}=\mathrm{JF}. \tag{2.3.14}$$

Either way, since the triangle JPF is isosceles, it follows that QP and PF make the same angle α to the tangent. If the parabola is a cross section of a telescopic mirror, any ray of light coming in parallel to the axis will be focussed at F, so that a paraboloidal mirror, used on-axis, does not suffer from spherical aberration. (This property holds, of course, only for light parallel to the axis of the paraboloid, so that a paraboloidal mirror, without some sort of correction, gives good images over only a narrow field of view.)

Now consider what happens when you stir a cup of tea. The surface takes up a shape that looks as though it might resemble the parabola $y=x^2/(4q)$ - see figure $\mathrm{II}.26$:

FIGURE II.26



Suppose the liquid is circulating at angular speed ω. A tea leaf floating on the surface is in equilibrium (in the rotating reference frame) under three forces: its weight mg, the centrifugal force $m\omega^2x$ and the normal reaction R. The normal to the surface makes an angle θ with the vertical (and the tangent makes an angle θ with the horizontal) given by

$$\tan\theta = \frac{\omega^2 x}{g}. \tag{2.3.15}$$

But the slope of the parabola $y = x^2/(4q)$ is $x/(2q)$, so that the surface is indeed a parabola with semi latus rectum $2q = g/\omega^2$.

This phenomenon has been used in Canada to make a successful large telescope (diameter $6\ \mathrm{m}$) in which the mirror is a spinning disc of mercury that takes up a perfectly paraboloidal shape. Another example is the spin casting method that has been successfully used for the production of large, solid glass paraboloidal telescope mirrors. In this process, the furnace is rotated about a vertical axis while the molten glass cools and eventually solidifies into the required paraboloidal effect.

Exercise 2.3.1

The 6.5 metre diameter mirrors for the twin Magellan telescopes at Las Campañas, Chile, have a focal ratio $f/1.25$. They were made by the technique of spin casting at The University of Arizona's Mirror Laboratory. At what speed would the furnace have had to be rotated in order to achieve the desired focal ratio? (Answer $= 7.4$ rpm.) Notice that $f/1.25$ is quite a deep paraboloid. If this mirror had been made by traditional grinding from a solid disc, what volume of material would have had to be removed to make the desired paraboloid? (Answer - a whopping 5.4 cubic metres, or about 12 tons!)

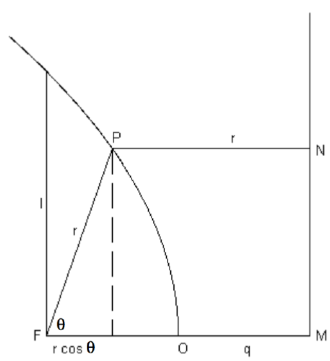
Polar Equation to the Parabola

As with the ellipse, we choose the focus as pole and the axis of the parabola as initial line. We shall orient the parabola so that the vertex is toward the right, as in figure $\mathrm{II}.27$.

We recall the focus-directrix property, $\mathrm{FP} = \mathrm{PN}$. Also, from the definition of the directrix, $\mathrm{FO} = \mathrm{OM} = q$, so that $\mathrm{FM} = 2q = l$, the length of the semi latus rectum. It is therefore immediately evident from figure $\mathrm{II}.27$ that $r\cos\theta + r = 2q = l$, so that the polar Equation to the parabola is

$$r = \frac{l}{1+\cos\theta}. \tag{2.3.16}$$

FIGURE II.27



This is the same as the polar Equation to the ellipse (Equation 2.3.36), with $e = 1$ for the parabola. I have given different derivations for the ellipse and for the parabola; the reader might like to interchange the two approaches and develop Equation 2.3.36 in the same manner as we have developed Equation 2.3.16.

When we discuss the hyperbola, I shall ask you to show that its polar Equation is also the same as 2.3.36. In other words, Equation 2.3.36 is the Equation to a *conic section*, and it represents an ellipse, parabola or hyperbola according to whether $e < 1$, $e = 1$ or $e > 1$.

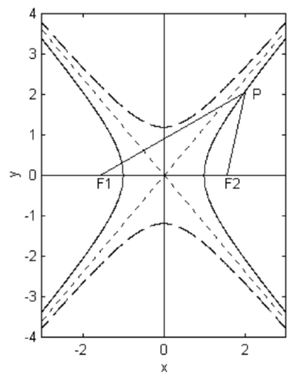
Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

2.4: The Hyperbola

A hyperbola is the locus of a point that moves such that the difference between its distances from two fixed points called the foci is constant. We shall call the difference between these two distances $2a$ and the distance between the foci $2ae$, where e is the eccentricity of the hyperbola, and is a number greater than 1. See figure $\text{II}.28$.

FIGURE $\text{II}.28$



For example, in a Young's double-slit interference experiment, the mth bright fringe is located at a point on the screen such that the path difference for the rays from the two slits is m wavelengths. As the screen is moved forward or backwards, this relation continues to hold for the mth bright fringe, whose locus between the slits and the screen is therefore a hyperbola. The "Decca" system of radar navigation, first used at the D-Day landings in the Second World War and decommissioned only as late as 2000 on account of being rendered obsolete by the "GPS" (Global Positioning Satellite) system, depended on this property of the hyperbola. (Since writing this, part of the Decca system has been re-commissioned as a back-up in case of problems with GPS.) Two radar transmitters some distance apart would simultaneously transmit radar pulses. A ship would receive the two signals separated by a short time interval, depending on the difference between the distances from the ship to the two transmitters. This placed the ship on a particular hyperbola. The ship would also listen in to another pair of transmitters, and this would place the ship on a second hyperbola. This then placed the ship at one of the four points where the two hyperbolas intersected. It would usually be obvious which of the four points was the correct one, but any ambiguity could be resolved by the signals from a third pair of transmitters.

In figure $\text{II}.28$, the coordinates of F_1 and F_2 are, respectively, $(-ae, 0)$ and $(ae, 0)$.

The condition $\text{PF}_1 - \text{PF}_2 = 2a$ requires that

$$\left[(x+ae)^2+y^2\right]^{\frac{1}{2}} - \left[(x-ae)^2+y^2\right]^{\frac{1}{2}} = 2a, \tag{2.5.1}$$

and this is the Equation to the hyperbola. After some arrangement, this can be written

$$\frac{x^2}{a^2} - \frac{y^2}{a^2(e^2-1)} = 1, \tag{2.5.2}$$

which is a more familiar form for the Equation to the hyperbola. Let us define a length b by

$$b^2 = a^2(e^2-1). \tag{2.5.3}$$

The Equation then becomes

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1, \tag{2.5.4}$$

which is the most familiar form for the Equation to a hyperbola.

Example 1

When a meteor streaks across the sky, it can be tracked by radar. The radar instrumentation can determine the range (distance) of the meteoroid as a function of time. Show that, if the meteoroid is moving at constant speed (a questionable assumption, because it must be decelerating, but perhaps we can assume the decrease in speed is negligible during the course of the observation), and if the range r is plotted against the time, the graph will be a hyperbola. Show also that, if r^2 is plotted against t, the graph will be a parabola of the form

$$r^2 = at^2 + bt + c, \tag{2.4.1}$$

where $a = V^2,\ b = -2V^2 t_0,\ c = V^2 t_0^2 + r_0^2,\ V = \text{speed of the meteoroid},\ t_0 = \text{time of closest approach},\ r_0 = \text{distance of closest approach}$.

Radar observation of a meteor yields the following range-time data:

t(s)	r(km)
0.0	101.4 *
0.1	103.0
0.2	105.8
0.3	107.8
0.4	111.1
0.5	112.6
0.6	116.7
0.7	119.3
0.8	123.8 *
0.9	126.4
1.0	130.6
1.1	133.3
1.2	138.1
1.3	141.3 *

(2.4.2)

Assume that the velocity of the meteor is constant.

Determine i. The time of closest approach (to 0.01 s)
ii. The distance of closest approach (to 0.1 km)
iii. The speed (to 1.0 km s^{-1})

If you wish, just use the three asterisked data to determine a, b and c. If you are more energetic, use all the data, and determine a, b and c by least squares, and the probable errors of V, t_0 and r_0.

The distance between the two vertices of the hyperbola is its transverse axis, and the length of the semi transverse axis is a − but what is the geometric meaning of the length b? This is discussed below in the next subsection (on the conjugate hyperbola) and again in a later section on the impact parameter.

The lines perpendicular to the x-axis and passing through the foci are the two *latera recta*. Since the foci are at $(\pm ae,\ 0)$, the points where the latera recta intersect the hyperbola can be found by putting $x = ae$ into the Equation to the hyperbola, and it is then found that the length l of a semi latus rectum is

$$l = a(e^2 - 1). \tag{2.5.5}$$

Definition: The Conjugate Hyperbola

The Equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = -1 \tag{2.5.6}$$

is the Equation to the *conjugate hyperbola*.

The conjugate hyperbola is drawn dashed in figure $\text{II}.28$, and it is seen that the geometric meaning of b is that it is the length of the semi transverse axis of the conjugate hyperbola. It is a simple matter to show that the eccentricity of the conjugate hyperbola is $e/\sqrt{e^2-1}$.

Definition: The Asymptotes

The lines

$$y=\pm\frac{bx}{a} \tag{2.5.7}$$

are the *asymptotes* of the hyperbola.

Equation 2.5.7 can also be written

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=0. \tag{2.5.8}$$

Thus

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}=c \tag{2.5.9}$$

is the hyperbola, the asymptotes, or the conjugate hyperbola, if $c=+1$, 0 or -1 respectively. The asymptotes are drawn as dotted lines in figure $\text{II}.28$.

The semi angle ψ between the asymptotes is given by

$$\tan\psi=b/a. \tag{2.5.10}$$

Exercise 1

If the eccentricity of a hyperbola is e, show that the eccentricity of its conjugate is $\dfrac{e}{\sqrt{e^2-1}}$.

Exercise 1: Corollary

No one will be surprised to note that this implies that, if the eccentricities of a hyperbola and its conjugate are equal, then each is equal to $\sqrt{2}$.

The Directrices

The lines $y=\pm a/e$ are the directrices, and, as with the ellipse (and with a similar proof), the hyperbola has the property that the ratio of the distance PF_2 to a focus to the distance PN to the directrix is constant and is equal to the eccentricity of the hyperbola. This ratio (i.e. the eccentricity) is less than one for the ellipse, equal to one for the parabola, and greater than one for the hyperbola. It is not a property that will be of great importance for our purposes, but is worth mentioning because it is a property that is sometimes used to define a hyperbola. I leave it to the reader to draw the directrices in their correct positions in figure $\text{II}.28$.

Parametric Equations to the Hyperbola.

The reader will recall that the point $(a\cos E, b\sin E)$ is on the ellipse $(x^2/a^2)+\left(y^2/b^2\right)=1$ and that this is evident because this Equation is the E-eliminant of $x=a\cos E$ and $y=b\sin E$. The angle E has a geometric interpretation as the eccentric anomaly. Likewise, recalling the relation $\cosh^2\phi-\sinh^2\phi=1$, it will be evident that $(x^2/a^2)-\left(y^2/b^2\right)=1$ can also be obtained as the $\phi-$eliminant of the Equations

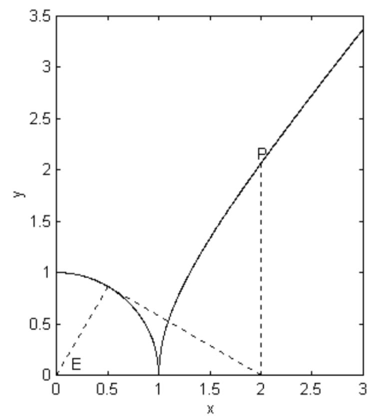
$$x=a\cosh\phi,\quad y=b\sinh\phi \tag{2.5.11}$$

These two Equations are therefore the parametric Equations to the hyperbola, and any point satisfying these two Equations lies on the hyperbola. The variable ϕ is not an angle, and has no geometric interpretation analogous to the eccentric anomaly of an ellipse. The Equations

$$x = a \sec E, \quad y = b \tan E \tag{2.5.12}$$

can also be used as parametric Equations to the hyperbola, on account of the trigonometric identity $1 + \tan^2 E = \sec^2 E$. In that case, the angle E does have a geometric interpretation (albeit not a particularly interesting one) in relation to the auxiliary circle, which is the circle of radius a centred at the origin. The meaning of the angle should be evident from figure $\text{II}.29$, in which E is the eccentric angle corresponding to the point P.

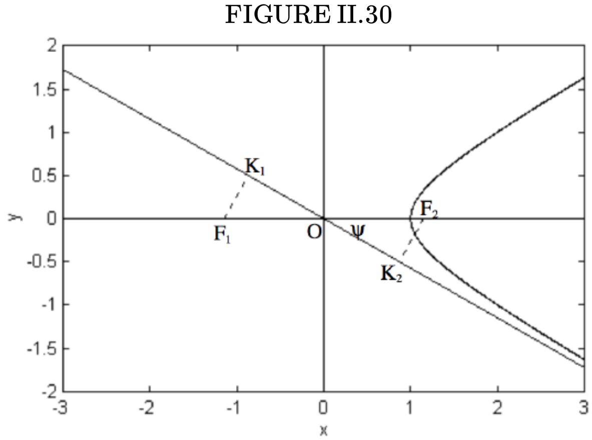
FIGURE II.29



Impact Parameter

A particle travelling very fast under the action of an inverse square attractive force (such as an interstellar meteoroid or comet - if there are such things - passing by the Sun, or an electron in the vicinity of a positively charged atomic nucleus) will move in a hyperbolic path. We prove this in a later chapter, as well as discussing the necessary speed. We may imagine the particle initially approaching from a great distance along the asymptote at the bottom right hand corner of figure $\text{II}.30$. As it approaches the focus, it no longer moves along the asymptote but along an arm of the hyperbola.

FIGURE II.30



The distance $\text{K}_2\,\text{F}_2$, which is the distance by which the particle would have missed F_2 in the absence of an attractive force, is commonly called the impact parameter. Likewise, if the force had been a repulsive force (for example, suppose the moving particle were a positively charged particle and there were a centre of repulsion at F_1, F_1K_1 would be the impact parameter. Clearly, F_1K_1 and F_2K_2 are equal in length. The symbol that is often used in scattering theory, whether in celestial mechanics or in particle physics, is b - but is this b the same b that goes into the Equation to the hyperbola and which is equal to the semi major axis of the conjugate hyperbola?

$\mathrm{OF}_2 = ae$, and therefore $\mathrm{K}_2\mathrm{F}_2 = ae\sin\psi$. This, in conjunction with $\tan\psi = b/a$ and $b^2 = a^2(e^2-1)$, will soon show that the impact parameter is indeed the same b that we are familiar with, and that b is therefore a very suitable symbol to use for impact parameter.

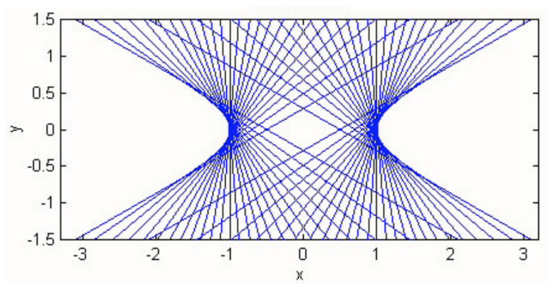
Tangents to the Hyperbola

Using the same arguments as for the ellipse, the reader should easily find that lines of the form

$$y = mx \pm \sqrt{a^2m^2 - b^2} \tag{2.5.13}$$

are tangent to the hyperbola. This is illustrated in figure $\mathrm{II}.33$ for a hyperbola with $b = a/2$, with tangents drawn with slopes 30° to 150° in steps of 5°. (The asymptotes have $\psi = 26^\circ 34'$.) (Sorry, but there are no figures $\mathrm{II}.31$ or $\mathrm{II}.32$ - computer problems!)

FIGURE II.31



Likewise, from similar arguments used for the ellipse, the tangent to the hyperbola at the point (x, y) is found to be

$$\frac{x_1x}{a^2} - \frac{y_1y}{b^2} = 1. \tag{2.5.14}$$

Director Circle

As for the ellipse, and with a similar derivation, the locus of the points of intersection of perpendicular tangents is a circle, called the director circle, which is of radius $\sqrt{(a^2-b^2)}$. This is not of particular importance for our purposes, but the reader who is interested might like to prove this by the same method as was done for the director circle of the ellipse, and might like to try drawing the circle and some tangents. If $b > a$, that is to say if $\psi > 45^\circ$ and the angle between the asymptotes is greater than 90°, the director circle is not real and it is of course not possible to draw perpendicular tangents.

Rectangular Hyperbola

If the angle between the asymptotes is 90°, the hyperbola is called a rectangular hyperbola. For such a hyperbola, $b = a$, the eccentricity is $\sqrt{2}$, the director circle is a point, namely the origin, and perpendicular tangents can be drawn only from the asymptotes.

The Equation to a rectangular hyperbola is

$$x^2 - y^2 = a^2 \tag{2.5.15}$$

and the asymptotes are at 45° to the x axis.

Let $\mathrm{O}x'$, $\mathrm{O}y'$ be a set of axes at 45° to the x axis. (That is to say, they are the asymptotes of the rectangular hyperbola.) Then the Equation to the rectangular hyperbola referred to its asymptotes as coordinate axes is found by the substitutions

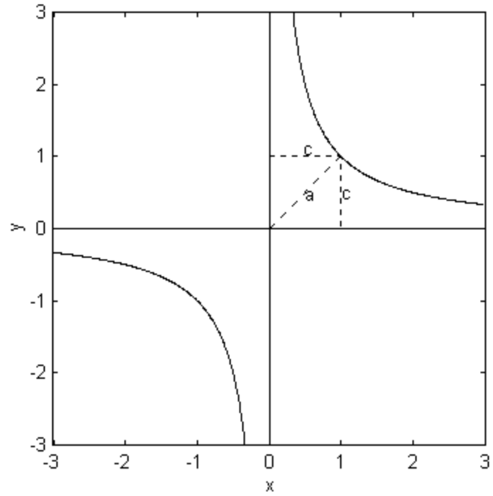
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos 45^\circ & \sin 45^\circ \\ -\sin 45^\circ & \cos 45^\circ \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} \tag{2.5.16}$$

into $x^2 - y^2 = a^2$. This results in the Equation

$$x'y' = \frac{1}{2}a^2 = c^2, \quad \text{where} \quad c = a/\sqrt{2}, \tag{2.5.17}$$

for the Equation to the rectangular hyperbola referred to its asymptotes as coordinate axes. The geometric interpretation of c is shown in figure $\mathrm{II}.32$, which is drawn for $c = 1$, and we have called the coordinate axes x and y. The length of the semi transverse axis is $c\sqrt{2}$.

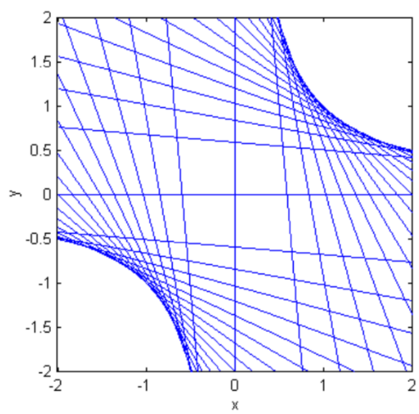
FIGURE II.32



The simple Equation $y = 1/x$ is a rectangular hyperbola and indeed it is this Equation that is shown in figure $\text{II}.32$.

It is left to the reader to show that the parametric Equations to the rectangular hyperbola $xy = c^2$ (we have dropped the primes) are $x = ct$, $y = c/t$, that lines of the form $y = mx \pm 2c\sqrt{-m}$ are tangent to $xy = c^2$ (figure II.35, drawn with slopes from 90° to 180° in steps of 5°), and that the tangent at (x_1, y_1) is $x_1 y + y_1 x = 2c$.

FIGURE II.33



Equation of a Hyperbola Referred to its Asymptotes as Axes of Coordinates

We have shown that the Equation to a *rectangular* hyperbola referred to its asymptotes as axes of coordinates is $x'y' = \frac{1}{2}a^2 = c^2$. In fact the Equation $x'y' = c^2$ is the Equation to any hyperbola (centred at $(0,\ 0)$), not necessarily rectangular, when referred to its asymptotes as axes of coordinates, where $c^2 = \frac{1}{4}(a^2 + b^2)$ In the figure below I have drawn a hyperbola and a point on the hyperbola whose coordinates with respect to the horizontal and vertical axes are $(x,\ y)$, and whose coordinates with respect to the asymptotes are (x', y'). I have shown the distances x and y with blue dashed lines, and the distances x' and y' with red dashed lines. The semiangle between the asymptotes is ψ.

The Equation to the hyperbola referred to the horizontal and vertical axes is

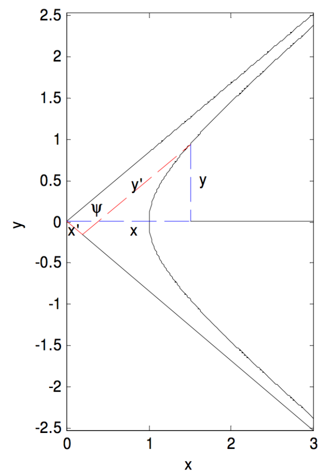
$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{2.5.18}$$

From the drawing, we see that

$$x = (x' + y')\cos\psi, \quad y = (y' - x')\sin\psi. \tag{2.5.19a,b}$$

If we substitute these into Equation 2.5.18, and also make use of the relation $\tan\psi = b/a$ (Equation 2.5.10), we arrive at the Equation to the hyperbola referred to the asymptotes as axes of coordinates:

$$x'y' = \frac{1}{4}(a^2 + b^2) = c^2. \tag{2.5.20}$$



Polar Equation to the Hyperbola

We found the polar Equations to the ellipse and the parabola in different ways. Now go back and look at both methods and use either (or both) to show that the polar Equation to the hyperbola (focus as pole) is

$$r = \frac{l}{1 + e\cos\theta}. \tag{2.5.21}$$

This is the polar Equation to any conic section - which one being determined solely by the value of e. You should also ask yourself what is represented by the Equation

$$r = \frac{l}{1 - e\cos\theta}. \tag{2.5.22}$$

Try sketching it for different values of e.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

2.5: Conic Sections

We have so far defined an ellipse, a parabola and a hyperbola without any reference to a cone. Many readers will know that a plane section of a cone is either an ellipse, a parabola or a hyperbola, depending on whether the angle that the plane makes with the base of the cone is less than, equal to or greater than the angle that the generator of the cone makes with its base. However, given the definitions of the ellipse, parabola and hyperbola that we have given, proof is required that they are in fact conic sections. It might also be mentioned at this point that a plane section of a circular cylinder is also an ellipse. Also, of course, if the plane is parallel with the base of the cone, or perpendicular to the axis of the cylinder, the ellipse reduces to a circle.

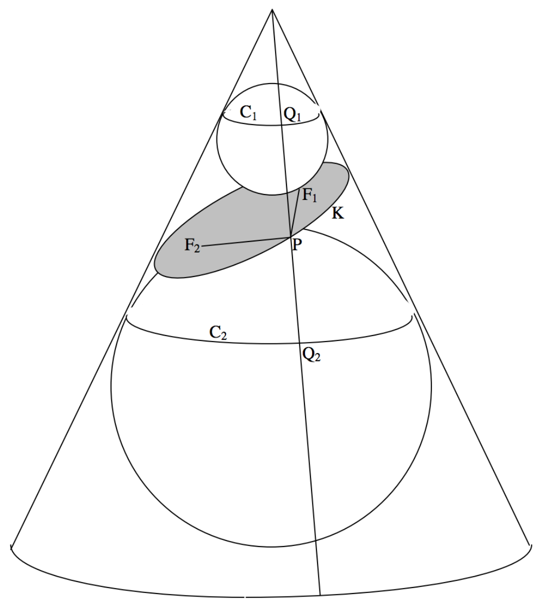


FIGURE II.36

A simple and remarkable proof can be given in the classical Euclidean "Given. Required. Construction. Proof. Q.E.D." style.

Proof

Given: A cone and a plane such that the angle that the plane makes with the base of the cone is less than the angle that the generator of the cone makes with its base, and the plane cuts the cone in a closed curve K. Figure $\mathrm{II}.36$.

Required: To prove that K is an ellipse.

Construction: Construct a sphere above the plane, which touches the cone internally in the circle C_1 and the plane at the point F_1. Construct also a sphere below the plane, which touches the cone internally in the circle C_2 and the plane at the point F_2.

Join a point P on the curve K to F_1 and to F_2.
Draw the generator that passes through the point P and which intersects C_1 at Q_1 and C_2 at Q_2.

Proof: $\mathrm{PF}_1 = \mathrm{PQ}_1$ (Tangents to a sphere from an external point.)

$\mathrm{PF}_2 = \mathrm{PQ}_2$ (Tangents to a sphere from an external point.)

$\therefore \mathrm{PF}_1 + \mathrm{PF}_2 = \mathrm{PQ}_1 + \mathrm{PQ}_2 = \mathrm{Q}_1\mathrm{Q}_2$

and $\mathrm{Q}_1\mathrm{Q}_2$ is independent of the position of P, since it is the distance between the circles C_1 and C_2 measured along a generator.

$$\therefore \ \mathrm{K} \text{ is an ellipse and } \mathrm{F}_1 \text{ and } \mathrm{F}_2 \text{ are its foci.} \qquad \text{(Q.E.D.)}$$

A similar argument will show that a plane section of a cylinder is also an ellipse.

The reader can also devise drawings that will show that a plane section of a cone parallel to a generator is a parabola, and that a plane steeper than a generator cuts the cone in a hyperbola. The drawings are easiest to do with paper, pencil, compass and ruler, and will require some ingenuity. While I have seen the above proof for an ellipse in several books, I have not seen the corresponding proofs for a parabola and a hyperbola, but they can indeed be done, and the reader should find it an interesting challenge. If the reader can use a computer to make the drawings and can do better than my poor effort with figure $\text{II}.36$, s/he is pretty good with a computer, which is a sign of a misspent youth.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

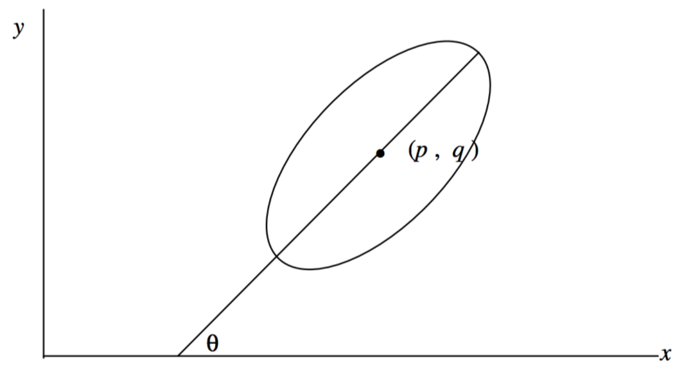
2.6: The General Conic Section

The Equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{2.7.1}$$

represents an ellipse whose major axis is along the x axis and whose centre is at the origin of coordinates. But what if its centre is not at the origin, and if the major axis is at some skew angle to the x axis? What will be the Equation that represents such an ellipse? Figure II.37.

FIGURE II.37



If the centre is translated from the origin to the point (p, q), the Equation that represents the ellipse will be found by replacing x by $x-p$ and y by $y-q$. If the major axis is inclined at an angle θ to the x axis, the Equation that represents the ellipse will be found by replacing x by $x\cos\theta+y\sin\theta$ and y by $-x\sin\theta+y\cos\theta$. In any case, if the ellipse is translated or rotated or both, x and y will each be replaced by linear expressions in x and y, and the resulting Equation will have at most terms in x^2, y^2, xy, x, y and a constant. The same is true of a parabola or a hyperbola. Thus, any of these three curves will be represented by an Equation of the form

$$ax^2+2hxy+by^2+2gx+2fy+c=0. \tag{2.7.2}$$

(The coefficients a and b are not the semi major and semi minor axes.) The apparently random notation for the coefficients arises because these figures are plane sections of three-dimensional surfaces (the ellipsoid, paraboloid and hyperboloid) which are described by terms involving the coordinate z as well as x and y. The customary notation for these three-dimensional surfaces is very systematic, but when the terms in z are removed for the two- dimensional case, the apparently random notation a, b, c, f, g, h remains. In any case, the above Equation can be divided through by the constant term without loss of generality, so that the Equation to an ellipse, parabola or hyperbola can be written, if preferred, as

$$ax^2+2hxy+by^2+2gx+2fy+1=0. \tag{2.7.3}$$

Is the converse true? That is, does an Equation of this form always necessarily represent an ellipse, parabola or hyperbola?

Not quite. For example,

$$6x^2+xy-y^2-17x-y+12=0 \tag{2.7.4}$$

represents two straight lines (it can be factored into two linear terms - try it), while

$$2x^2-4xy+4y^2-4x+4=0 \tag{2.7.5}$$

is satisfied only by a single point. (Find it.)

However, a plane section of a cone can be two lines or a single point, so perhaps we can now ask whether the general second degree Equation must always represent a conic section. The answer is: close, but not quite.

For example,

$$4x^2+12xy+9y^2+14x+21y+6=0 \tag{2.7.6}$$

represents two parallel straight lines, while

$$x^2+y^2+3x+4y+15=0 \tag{2.7.7}$$

cannot be satisfied by any real (x, y).

However, a plane can intersect a *cylinder* in two parallel straight lines, or a single straight line, or not at all. Therefore, if we stretch the definition of a cone somewhat to include a cylinder as a special limiting case, then we can say that the general second degree Equation in x and y does indeed always represent a conic section.

Is there any means by which one can tell by a glance at a particular second degree Equation, for example

$$8x^2+10xy-3y^2-2x-4y-2=0, \tag{2.7.8}$$

what type of conic section is represented? The answer is yes, and this one happens to be a hyperbola. The discrimination is made by examining the properties of the determinant

$$\Delta=\begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix} \tag{2.7.9}$$

I have devised a table after the design of the dichotomous tables commonly used by taxonomists in biology, in which the user is confronted by a couplet (or sometimes triplet) of alternatives, and is then directed to the next part of the table. I shall spare the reader the derivation of the table; instead, I shall describe its use.

In the table, I have used the symbol $\bar{a}$ to mean the cofactor of a in the determinant, $\bar{h}$ the cofactor of h, $\bar{g}$ the cofactor of g, etc. Explicitly, these are

$$\bar{a}=bc-f^2, \tag{2.7.10}$$

$$\bar{b}=ca-g^2, \tag{2.7.11}$$

$$\bar{c}=ab-h^2, \tag{2.7.12}$$

$$\bar{f}=gh-af, \tag{2.7.13}$$

$$\bar{g}=hf-bg \tag{2.7.14}$$

and

$$\bar{h}=fg-ch. \tag{2.7.15}$$

The first column labels the choices that the user is asked to make. At the beginning, there are two choices to make, 1 and $1'$ The second column says what these choices are, and the fourth column says where to go next. Thus, if the determinant is zero, go to 2; otherwise, go to 5. If there is an asterisk in column 4, you are finished. Column 3 says what sort of a conic section you have arrived at, and column 5 gives an example.

No matter what type the conic section is, the coordinates of its centre are $(\bar{g}/\bar{c},\ \bar{f}/\bar{c})$ and the angle θ that its major or transverse axis makes with the x axis is given by

$$\tan 2\theta=\frac{2h}{a-b}. \tag{2.7.16}$$

Thus if x is first replaced with $x+\bar{g}/\bar{c}$ and y with $y+\bar{f}/\bar{c}$, and then the new x is replaced with $x\cos\theta - y\sin\theta$ and the new y with $x\sin\theta + y\cos\theta$, the Equation will take the familiar form of a conic section with its major or transverse axis coincident with the x axis and its centre at the origin. Any of its properties, such as the eccentricity, can then be deduced from the familiar Equations. You should try this with Equation 2.7.8.

Key to the Conic Sections

1	$\Delta=0$		2	
1'	$\Delta\neq0$		5	
2	$\bar{c}>0$	Point	*	$x^2-2xy+2y^2-2x+2=0$
2'	$\bar{c}=0$		3	
2"	$\bar{c}<0$	Two nonparallel straight lines	4	
3	$\bar{h}=0$	Straight line	*	$4x^2+4xy+y^2+12x+6y+9=0$
3'	$\bar{h}\neq0$	Two parallel straight lines	*	$4x^2+12xy+9y^2+14x+21y+6=0$
4	$a+b=0$	Two perpendicular straight lines	*	$6x^2+5xy-6y^2+x+8y-2=0$
4'	$a+b\neq0$	Two straight lines, neither parallel nor perpendicular	*	$6x^2-xy-y^2+34x+13y-12=0$
5	$\bar{c}>0$		6	
5'	$\bar{c}=0$	Parabola	*	$9x^2-12xy+4y^2-18x-101y+19=0$
5"	$\bar{c}<0$		8	
6	$a\Delta>0$	Nothing	*	$x^2+y^2+3x+4y+15=0$
6'	$a\Delta<0$		7	
7	$a=b$ and $h=0$	Circle	*	$x^2+y^2-6x-8y+9=0$
7'	Not so	Ellipse	*	$14x^2-4xy+11y^2-44x-58y+71=0$
8	$a+b=0$	Rectangular hyperbola	*	$7x^2-48xy-7y^2+10x-28y+100=0$
8'	$a+b\neq0$	Hyperbola (not Rectangular)	*	$8x^2+10xy-3y^2-2x-4y-2=0$

When faced with a general second degree Equation in x and y, I often find it convenient right at the start to calculate the values of the cofactors from Equations 2.7.10 − 2.7.15.

Here is an exercise that you might like to try. Show that the ellipse $ax^2+2hxy+by^2+2gx+2fy+1=0$ is contained within the rectangle whose sides are

$$x=\frac{\bar{g}\pm\sqrt{\bar{g}^2-\bar{a}\bar{c}}}{\bar{c}} \tag{2.7.18}$$

$$y=\frac{\bar{f}\pm\sqrt{\bar{f}^2-\bar{b}\bar{c}}}{\bar{c}} \tag{2.7.19}$$

In other words, these four lines are the vertical and horizontal tangents to the ellipse.

This is probably not of much use in celestial mechanics, but it will probably be useful in studying Lissajous ellipses, or the Stokes parameters of polarized light. It is also useful in programming a computer to draw, for example, the ellipse $14x^2-4xy+11y^2-44x-58y+71=0$. To do this, you will probably want to start at some value of x and calculate the two corresponding values of y, and then move to another value of x. But at which value of x should you start? Equation 2.7.18 will tell you.

But what do Equations 2.7.18 and 2.7.19 mean if the conic section Equation $ax^2+2hxy+by^2+2gx+2fy+1=0$ is not an ellipse? They are still useful if the conic section is a hyperbola. Equations 2.7.18 and 2.7.19 are still vertical and horizontal tangents - but in this case the hyperbola is entirely outside the limits imposed by these tangents. If the axes of the hyperbola are horizontal and vertical, one or other of Equations 2.7.18 and 2.7.19 will fail.

If the conic section is a parabola, Equations 2.7.18 and 2.7.19 are not useful, because $c = 0$. There is only one horizontal tangent and only one vertical tangent. They are given by

$$x = \frac{\bar{a}}{2\bar{g}} \tag{2.7.20}$$

and

$$y = \frac{\bar{b}}{2\bar{f}} \tag{2.7.21}$$

If the axis of the parabola is horizontal or vertical, one or other of Equations 2.7.20 and 2.7.21 will fail.

If the second degree Equation represents one or two straight lines, or a point, or nothing, I imagine that all of Equations 2.7.18 – 2.7.21 will fail - unless perhaps the Equation represents horizontal or vertical lines. I haven't looked into this; perhaps the reader would like to do so.

Here is a problem that you might like to try. The Equation $8x^2 + 10xy - 3y^2 - 2x - 4y - 2 = 0$ represents a hyperbola. What are the Equations to its axes, to its asymptotes, and to its conjugate hyperbola? Or, more generally, if $ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0$ represents a hyperbola, what are the Equations to its axes, to its asymptotes, and to its conjugate hyperbola?

Before starting, one point worth noting is that the original hyperbola, its asymptotes, and the conjugate hyperbola) have the same centre, which means that g and f are the same for each, and they have the same axes, which means that a, h, and b are the same for each. They differ only in the constant term.

If you do the first problem, $8x^2 + 10xy - 3y^2 - 2x - 4y - 2 = 0$, there will be a fair amount of numerical work to do. When I did it I didn't use either pencil and paper or a hand calculator. Rather I sat in front of a computer doing the numerical calculations with a Fortran statement for every stage of the calculation. I don't think I could have done it otherwise without making lots of mistakes. *The very first thing I did* was to work out the cofactors $\bar{a}$, $\bar{h}$, $\bar{b}$, $\bar{g}$, $\bar{f}$, $\bar{c}$ and store them in the computer, and also the coordinates of the centre (x_0, y_0) of the hyperbola, which are given by $x_0 = \bar{g}/\bar{c}$, $y_0 = \bar{f}/\bar{c}$.

Whether you do the particular numerical problem, or the more general algebraic one, I suggest that you proceed as follows. First, refer the hyperbola to a set of coordinates x', y' whose origin coincides with the axes of the hyperbola. This is done by replacing x with $x' + x_0$ and y with $y' + y_0$. This will result in an Equation of the form $ax'^2 + 2hx'y' + by'^2 + c' = 0$. The coefficients of the quadratic terms will be unchanged, the linear terms will have vanished, and the constant term will have changed. At this stage I got, for the numerical example, $8x'^2 + 10x'y' - 3y'^2 - 1.8163 = 0$.

Now refer the hyperbola to a set of coordinates x'', y'' whose axes are parallel to the axes of the hyperbola. This is achieved by replacing x' with $x''\cos\theta - y''\sin\theta$ and y' with $x''\sin\theta + y''\cos\theta$, where $\tan 2\theta = 2h/(a-b)$. There will be a small problem here, because this gives two values of θ differing by 90°, and you'll want to decide which one you want. In any case, the result will be an Equation of the form $a''x''^2 + b''y''^2 + c' = 0$, in which a'' and b'' are of opposite sign. Furthermore, if you happen to understand the meaning of the noise "The trace of a matrix is invariant under an orthogonal transformation", you'll be able to check for arithmetic mistakes by noting that $a'' + b'' = a + b$. If this is not so, you have made a mistake. Also, the constant term should be unaltered by the rotation (note the single prime on the c). At this stage, I got $9.933x''^2 - 4.933y''^2 - 1.8163 = 0$. (All of this was done with Fortran statements on the computer - no actual calculation or writing done by me - and the numbers were stored in the computer to many significant figures).

In any case this Equation can be written in the familiar form $\frac{x''^2}{A^2} - \frac{y''^2}{B^2} = 1$, which in this case I made to be $\frac{x''^2}{0.4283^2} - \frac{y''^2}{0.6088^2} = 1$. We are now on familiar ground. The axes of the hyperbola are $x'' = 0$ and $y'' = 0$, the asymptotes are $\frac{x''^2}{A^2} - \frac{y''^2}{B^2} = 0$ and the conjugate hyperbola is $\frac{x''^2}{A^2} - \frac{y''^2}{B^2} = -1$.

Now, starting from $\frac{x''^2}{A^2} - \frac{y''^2}{B^2} = -1$ for the asymptotes, or from $\frac{x''^2}{A^2} - \frac{y''^2}{B^2} = -1$ for the conjugate hyperbola, we reverse the process. We go to the single-primed coordinates by replacing x'' with $x'\cos\theta + y'\sin\theta$ and y'' with $-x'\sin\theta + y'\cos\theta$, and then to the original coordinates by replacing x' with $x - x_0$ and y' with $y - y_0$.

This is what I find:

Original hyperbola: $ax^2+2hxy+by^2+2gx+2fy+c=0$

Conjugate hyperbola: $ax^2+2hxy+by^2+2gx+2fy+c_{\text{conj}}=0$,
where $c_{\text{conj}}=-(2g\bar{g}+2f\bar{f}+c\bar{c})/\bar{c}=-(2gx_0+2fy_0+c)$.

Asymptotes: $ax^2+2hxy+by^2+2gx+2fy+c_{\text{asymp}}=0$,
where c_{asymp} can be written in any of the following equivalent forms:

$$c_{\text{asymp}}=+(a\bar{g}^2+2h\bar{g}\bar{f}+b\bar{f}^2)/\bar{c}^2=ax_0^2+2hx_0y_0+by_0^2=-(g\bar{g}+f\bar{f})/\bar{c}. \tag{2.6.1}$$

[The last of these three forms can be derived very quickly by recalling that a condition for a general second degree Equation in x and y to represent two straight lines is that the determinant Δ should be zero. A glance at this determinant will show that this implies that $g\bar{g}+f\bar{f}+c\bar{c}=0$.]

Axes of hyperbolas: $(y-x\tan\theta-y_0+x_0\tan\theta)(y+x\cot\theta-y_0-x_0\cot\theta)=0$,
where $\tan 2\theta=2h/(a-b)$.

Example:

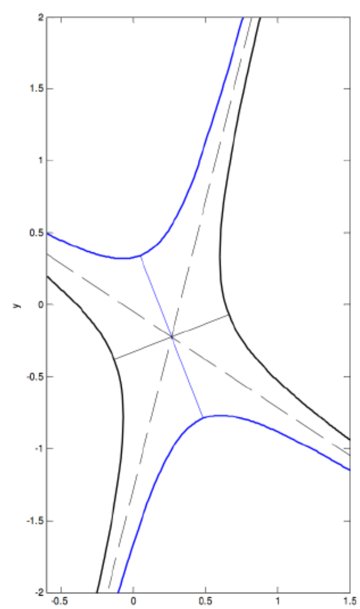
Original hyperbola: $8x^2+10xy-3y^2-2x-4y-2=0$

Conjugate hyperbola: $8x^2+10xy-3y^2-2x-4y+\frac{80}{49}=0$

Asymptotes: $8x^2+10xy-3y^2-2x-4y-\frac{9}{49}=0$,
which can also be written $(4x-y-\frac{9}{7})(2x+3y+\frac{1}{7})=0$

Axes of hyperbolas: $(y-0.3866x+0.3275)(y+2.5866x-0.4613)$

These are shown in the figure below - the original hyperbola in black, the conjugate in blue.



The centre is at (0.26531, −0.22449).

The slopes of the two asymptotes are 4 and $-\frac{2}{3}$. From Equation 2.2.16 we find that the tangent of the angle between the asymptotes is $\tan 2\psi=\frac{14}{5}$, so that $2\psi=70^\circ.3$, and the angle between the asymptote and the major axis of the original hyperbola is $54^\circ.8$, or $\tan\psi=1.419$. This is equal (see Equations 2.5.3 and 2.5.10) to $\sqrt{e^2-1}$, so the eccentricity of the original hyperbola is 1.735. From Section 2.2, shortly Equation 2,5,6, we soon find that the eccentricity of the conjugate hyperbola is $\csc\psi=1.223$.

An interesting question occurs to me. We have found that, if $ax^2+2hxy+by^2+2gx+2fy+c=0$ is a hyperbola, then the Equations to the conjugate hyperbola and the asymptotes are of a similar form, namely $ax^2+2hxy+by^2+2gx+2fy+c_{\text{conj}}=0$ and $ax^2+2hxy+by^2+2gx+2fy+c_{\text{asymp}}=0$, and we found expressions for c_{conj} and c_{asymp}. But what if $ax^2+2hxy+by^2+2gx+2fy+c=0$ is not a hyperbola? What if it is an ellipse? What do the other Equations represent, given that an ellipse has neither a conjugate nor asymptotes?

For example, $14x^2-4xy+11y^2-44x-58y+71=0$ is an ellipse. What are $14x^2-4xy+11y^2-44x-58y+191=0$ and $14x^2-4xy+11y^2-44x-58y+131=0$? I used the key on page 47, and it told me that the first of these Equations is satisfied by no real points, which I suppose is the Equation's way of telling me that there is no such thing as the conjugate to an ellipse. The second Equation was supposed to be the "asymptotes", but the key shows me that the Equation is satisfied by just one real point, namely (2 , 3), which coincides with the centre of the original ellipse. I didn't expect that. Should I have done so?

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

2.7: Fitting a Conic Section Through Five Points

FIGURE II.38

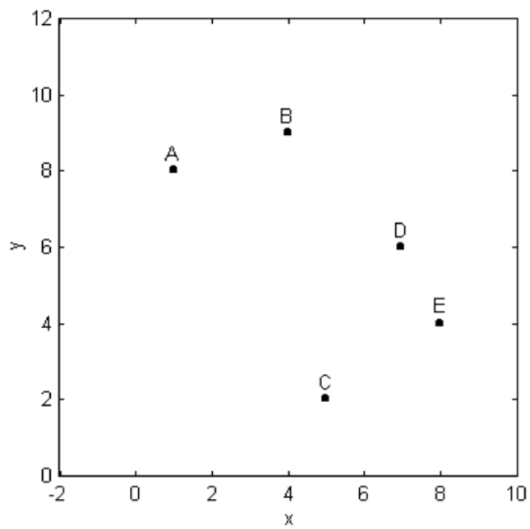


Figure $\text{II}.38$ shows the five points $\text{A}(1,8)$, $\text{B}(4,9)$, $\text{C}(5,2)$, $\text{D}(7,6)$, $\text{E}(8,4)$. Problem: Draw a conic section through the five points.

The first thing to notice is that, since a conic section is of the form

$$ax^2+2hxy+by^2+2gx+2fy+1=0 \tag{2.8.1}$$

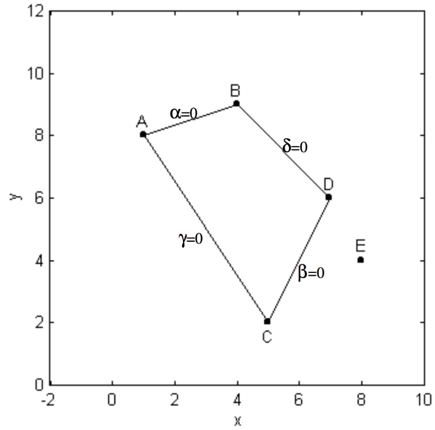
(remember that there is no loss of generality by taking the constant term to be 1), five points are necessary and sufficient to define a conic section uniquely. One and only one conic section can be drawn through these five points. We merely have to determine the five coefficients. The most direct (but not the fastest or most efficient) way to do this is to substitute each of the (x , y) pairs into the Equation in turn, thus obtaining five linear Equations in the five coefficients.

There is a better way.

We write down the Equations for the straight lines AB, CD, AC and BD. Let us call these Equations $\alpha=0$, $\beta=0$, $\gamma=0$ and $\delta=0$ respectively (figure $\text{II}.39$).

Then $\alpha\beta=0$ is the Equation that represents the two straight lines AB and CD, and $\gamma\delta=0$ is the Equation that represents the two straight lines AC and BD. The Equation $\alpha\beta+\lambda\gamma\delta=0$, where λ is an arbitrary constant, is a second degree Equation that represents any conic section that passes through the points A, B, C and D. By inserting the coordinates of E in this Equation, we can find the value of λ that forces the Equation to go through all five points. This model of unclarity will become clear on following an actual calculation for the five points of the present example.

FIGURE II.39



The four straight lines are

$$\alpha=0: \quad x-3y+23=0$$

$$\beta=0: \quad 2x-y-8=0$$

$$\gamma=0: \quad 3x+2y-19=0$$

$$\delta=0: \quad x+y-13=0$$

The two pair of lines are

$$\alpha\beta=0: \quad 2x^2-7xy+3y^2+38x+y-184=0$$

$$\gamma\delta=0: \quad 3x^2+5xy+2y^2-58x-45y+247=0$$

and the family of conic sections that passes through A, B, C and D is

$\alpha\beta+\lambda\gamma\delta=0:$

$$(2+3\lambda)x^2-(7-5\lambda)xy+(3+2\lambda)y^2+(38-58\lambda)x+(1-45\lambda)y-184+247\lambda=0.$$

Now substitute $x=8$, $y=4$ to force the conic section to pass through the point E. This results in the value

$$\lambda=\frac{76}{13}.$$

The Equation to the conic section passing through all five points is therefore

$$508x^2+578xy+382y^2-7828x-6814y+32760=0$$

We can, if desired, divide this Equation by 2 (since all coefficients are even), or by 32760 (to make the constant term equal to 1) but, to make the analysis that is to follow easier, I choose to leave the Equation in the above form, so that the constants f, g and h remain integers.

The constants have the values

$$a=508, b=382, c=32760, f=-3407, g=-3914, h=289$$

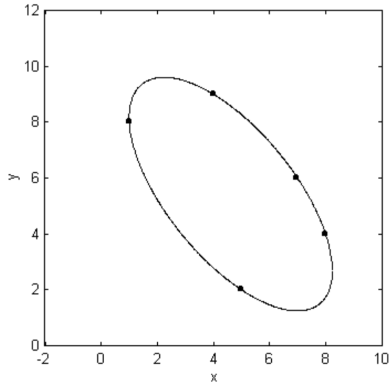
and the cofactors have the values

$$\bar{a}=906\,671, \quad \bar{b}=1\,322\,684, \quad \bar{c}=110\,535$$

$$\bar{f}=599\,610, \quad \bar{g}=510\,525, \quad \bar{h}=3\,867\,358$$

Let us consult the dichotomous table. The value of the determinant is $\Delta=a\bar{a}+h\bar{h}+g\bar{g}$ (or $h\bar{h}+b\bar{b}+f\bar{f}$, or $g\bar{g}+f\bar{f}+c\bar{c}$; try all three sums to check for arithmetic mistakes). It comes to $\Delta=-419\,939\,520$, so we proceed to option 5. $\bar{c}>0$, so we proceed to option 6. a and Δ have opposite signs, so we proceed to 7. a does not equal b, nor is h equal to zero. Therefore we have an ellipse. It is drawn in figure II.40.

FIGURE II.40



The centre of the ellipse is at $(4.619,\ 5.425)$ and its major axis is inclined at an angle $128^\circ\ 51'$ to the x-axis. If we now substitute $x+4.619$ for x and $y+5.425$ for y, and then substitute $x\cos 128^\circ 51' + y\sin 128^\circ 51'$ for the new value of x and $-x\sin 128^\circ 51' + y\cos 128^\circ 51'$ for the new value of y, the Equation will assume its the familiar form for an ellipse referred to its axes as coordinate axes and its centre as origin.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

2.8: Fitting a Conic Section Through n Points

What is the best ellipse passing near to the following 16 points?

$$\begin{array}{llll}(\ 1,50) & (11,58) & (20,63) & (30,60) \\ (42,59) & (48,52) & (54,46) & (61,42) \\ (61,19) & (45,12) & (35,10) & (25,13) \\ (17,17) & (14,22) & (\ 5,29) & (\ 3,43)\end{array}$$

This is answered by substituting each point (x, y) in turn in the Equation

$$ax^2 + 2hxy + by^2 + 2gx + 2fy + 1 = 0, \tag{2.9.1}$$

thus obtaining 16 Equations in the coefficients $a,\ h,\ b,\ g,\ f$. (The constant term can be taken to be unity.) These are the Equations of condition. The five normal Equations can then be set up and solved to give those values for the coefficients that will result in the sum of the squares of the residuals being least, and it is in that sense that the "best" ellipse results. The details of the method are given in the chapter on numerical methods. The actual solution for the points given above is left as an exercise for the energetic.

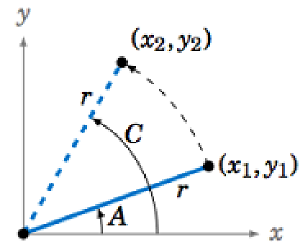
It might be thought that we are now well on the way to doing some real orbital theory. After all, suppose that we have several positions of a planet in orbit around the Sun, or several positions of the secondary component of a visual binary star with respect to its primary component; we can now fit an ellipse through these positions. However, in a real orbital situation we have some additional information as well as an additional constraint. The additional information is that, for each position, we also have a time. The constraint is that the orbit that we deduce must obey Kepler's second law of planetary motion - namely, that the radius vector sweeps out equal areas in equal times. We shall have to await Part II before we get around actually to computing orbits.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

3: PLANE AND SPHERICAL TRIGONOMETRY



3.1: INTRODUCTION

It is assumed in this chapter that readers are familiar with the usual elementary formulas encountered in introductory trigonometry. We start the chapter with a brief review of the solution of a plane triangle. While most of this will be familiar to readers, it is suggested that it be not skipped over entirely, because the examples in it contain some cautionary notes concerning hidden pitfalls.

3.2: PLANE TRIANGLES

This section is to serve as a brief reminder of how to solve a plane triangle. While there may be a temptation to pass rapidly over this section, it does contain a warning that will become even more pertinent in the section on spherical triangles.

3.3: CYLINDRICAL AND SPHERICAL COORDINATES

3.4: VELOCITY AND ACCELERATION COMPONENTS

3.5: SPHERICAL TRIANGLES

We are fortunate in that we have four formulas at our disposal for the solution of a spherical triangle, and, as with plane triangles, the art of solving a spherical triangle entails understanding which formula is appropriate under given circumstances. Each formula contains four elements (sides and angles), three of which, in a given problem, are assumed to be known, and the fourth is to be determined.

3.6: ROTATION OF AXES, TWO DIMENSIONS

3.7: ROTATION OF AXES, THREE DIMENSIONS. EULERIAN ANGLES

3.8: TRIGONOMETRICAL FORMULAS

A reference a set of commonly-used trigonometric formulas i sprovide. Anyone who is regularly engaged in problems in celestial mechanics or related disciplines will be familiar with most of them.

3.1: Introduction

It is assumed in this chapter that readers are familiar with the usual elementary formulas encountered in introductory trigonometry. We start the chapter with a brief review of the solution of a plane triangle. While most of this will be familiar to readers, it is suggested that it be not skipped over entirely, because the examples in it contain some cautionary notes concerning hidden pitfalls.

This is followed by a quick review of spherical coordinates and direction cosines in three-dimensional geometry. The formulas for the velocity and acceleration components in two dimensional polar coordinates and three-dimensional spherical coordinates are developed in section 3.4.

Section 3.5 deals with the trigonometric formulas for solving spherical triangles. This is a fairly long section, and it will be essential reading for those who are contemplating making a start on celestial mechanics.

Sections 3.6 and 3.7 deal with the rotation of axes in two and three dimensions, including Eulerian angles and the rotation matrix of direction cosines.

Finally, in section 3.8, a number of commonly encountered trigonometric formulas are gathered for reference.

Contributor

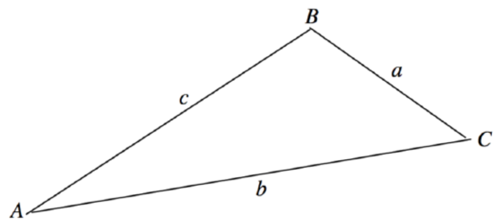
- Jeremy Tatum (University of Victoria, Canada)

3.2: Plane Triangles

This section is to serve as a brief reminder of how to solve a plane triangle. While there may be a temptation to pass rapidly over this section, it does contain a warning that will become even more pertinent in the section on spherical triangles.

Conventionally, a plane triangle is described by its three angles A, B, C and three sides a, b, c, with a being opposite to A, b opposite to B, and c opposite to C. See figure III.1.

FIGURE III.1



It is assumed that the reader is familiar with the sine and cosine formulas for the solution of the triangle:

$$\frac{a}{\sin A}=\frac{b}{\sin B}=\frac{c}{\sin C} \tag{3.2.1}$$

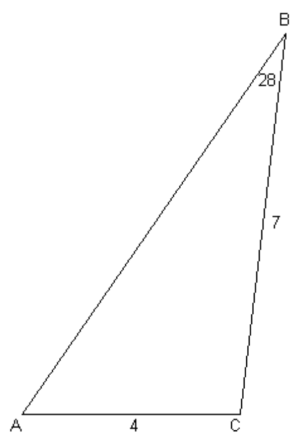
and

$$a^2=b^2+c^2-2bc\cos A, \tag{3.2.2}$$

and understands that the art of solving a triangle involves recognition as to which formula is appropriate under which circumstances. Two quick examples - each with a warning - will suffice.

Example: A plane triangle has sides $a=7$ inches, $b=4$ inches and angle $B=28^\circ$. Find the angle A.

FIGURE III.2



See figure III.2

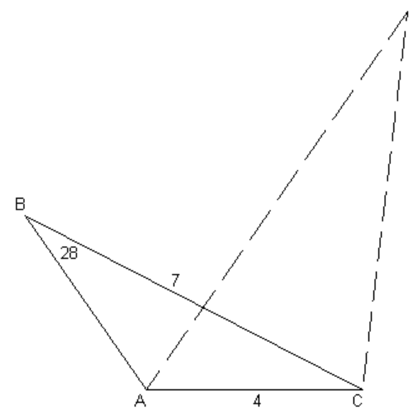
We use the sine formula, to obtain

$$\sin A=\frac{7\sin 28^\circ}{4}=0.821575$$

$$A=55^\circ 14'.6$$

The pitfall is that there are two values of A between 0° and 180° that satisfy $\sin A=0.821575$, namely $55^\circ\ 14'.6$ and $124^\circ\ 45'.4$. Figure III.3 shows that, given the original data, either of these is a valid solution.

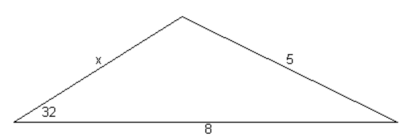
FIGURE III.3



The lesson to be learned from this is that all inverse trigonometric functions $(\sin^{-1}, \cos^{-1}, \tan^{-1})$ have two solutions between 0° and 360°. The function $\sin^{-1}$ is particularly troublesome since, for positive arguments, it has two solutions between 0° and 180°. The reader must always be on guard for "quadrant problems" (i.e. determining which quadrant the desired solution belongs to) and is warned that, unless particular care is taken in programming calculators or computers, quadrant problems are among the most frequent problems in trigonometry, and especially in spherical astronomy.

Example: Find x in the triangle illustrated in figure $\text{III}.4$.

FIGURE III.4



Application of the cosine rule results in

$$25 = x^2 + 64 - 16x\cos 32^\circ$$

Solution of the quadratic Equation yields

$$x = 4.133 \text{ or } 9.435$$

This illustrates that the problem of "two solutions" is not confined to angles alone. Figure $\text{III}.4$ is drawn to scale for one of the solutions; the reader should draw the second solution to see how it is that two solutions are possible.

The reader is now invited to try the following "guaranteed all different" problems by hand calculator. Some may have two real solutions. Some may have none. The reader should draw the triangles accurately, especially those that have two solutions or no solutions. It is important to develop a clear geometric understanding of trigonometric problems, and not merely to rely on the automatic calculations of a machine. Developing these critical skills now will pay dividends in the more complex real problems encountered in celestial mechanics and orbital computation.

PROBLEMS

1. $a = 6 \;\; b = 4 \;\; c = 7 \; C = ?$
2. $a = 5 \;\; b = 3 \;\; C = 43^\circ \; c = ?$
3. $a = 7 \;\; b = 9 \;\; C = 110^\circ \; B = ?$
4. $a = 4 \;\; b = 5 \;\; A = 29^\circ \; c = ?$
5. $a = 5 \;\; b = 7 \;\; A = 37^\circ \; B = ?$
6. $a = 8 \;\; b = 5 \;\; A = 54^\circ \; C = ?$
7. $A = 64^\circ \; B = 37^\circ \; a/c = ? \; b/c = ?$
8. $a = 3 \;\; b = 8 \;\; c = 4 \; C = ?$

9. $a=4$ $b=11$ $A=26^\circ$ $c=?$

The reader is now further invited to write a computer program (in whatever language is most familiar) for solving each of the above problems for arbitrary values of the data. Lengths should be read in input and printed in output to four significant figures. Angles should be read in input and printed in output in degrees, minutes and tenths of a minute (e.g. $47^\circ\ 12'.9$). Output should show two solutions if there are two, and should print "NO Solution" if there are none. This exercise will familiarize the reader with the manipulation of angles, especially inverse trigonometric functions in whatever computing language is used, and will be rewarded in future more advanced applications.

Solutions to problems.

1. $C=86^\circ 25'.0$
2. $c=3.473$
3. $B=40^\circ 00'.1$
4. $c=7.555$ or 1.191
5. $B=57^\circ 24'.6$ or $122^\circ 35'.4$
6. $C=95^\circ 37'.6$ or $23^\circ 37'.6$
7. $a/c=0.9165$ $b/c=0.6131$
8. No real solution
9. No real solution

The area of a plane triangle is $\frac{1}{2}\times\text{base}\times\text{height}$, and it is easy to see from this that

$$\text{Area}=\frac{1}{2}bc\sin A=\frac{1}{2}ca\sin B=\frac{1}{2}ab\sin C \tag{3.2.3}$$

By making use $\sin^2 A=1-\cos^2 A$ and $\cos A=\left(b^2+c^2-a^2\right)/(2bc)$, we can express this entirely in terms of the lengths of the sides:

$$\text{Area}=\sqrt{s(s-a)(s-b)(s-c)}, \tag{3.2.4}$$

where s is the semi-perimeter $\frac{1}{2}(a+b+c)$.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

3.3: Cylindrical and Spherical Coordinates

It is assumed that the reader is at least somewhat familiar with cylindrical coordinates (ρ, ϕ, z) and spherical coordinates (r, θ, ϕ) in three dimensions, and I offer only a brief summary here. Figure III.5 illustrates the following relations between them and the rectangular coordinates (x, y, z).

$$x = \rho\cos\phi = r\sin\theta\cos\phi \tag{3.3.1}$$

$$y = \rho\sin\phi = r\sin\theta\sin\phi \tag{3.3.2}$$

$$z = r\cos\theta \tag{3.3.3}$$

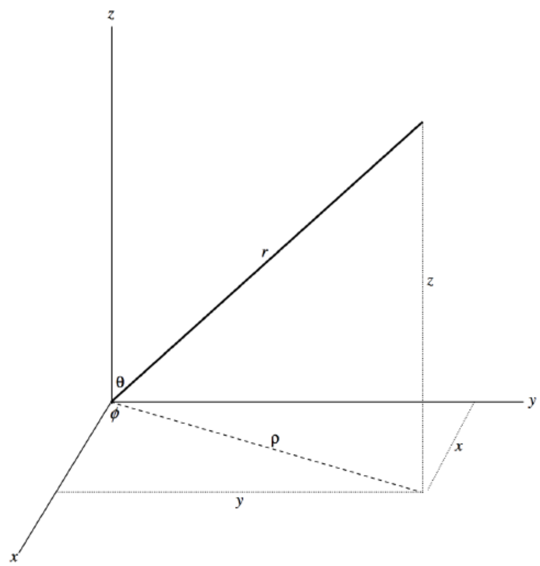


FIGURE III.5

The inverse relations between spherical and rectangular coordinates are

$$r = \sqrt{x^2+y^2+z^2} \tag{3.3.4}$$

$$\theta = \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}} \tag{3.3.5}$$

$$\phi = \cos^{-1}\frac{x}{\sqrt{x^2+y^2}} = \sin^{-1}\frac{y}{\sqrt{x^2+y^2}} \tag{3.3.6}$$

The coordinates r, θ and ϕ are called, respectively, the "radial", "polar" or "meridional", and "azimuthal" coordinates respectively.

Note that r is essentially positive (the symbol √ denotes the positive or absolute value of the square root). The angle θ is necessarily between 0° and 180° and therefore there is no quadrant ambiguity in the evaluation of θ. The angle ϕ, however, can be between 0° and 360°. Therefore, in order to determine ϕ uniquely, both of the above formulas for ϕ must be evaluated, or the signs of x and y must be inspected. It does not suffice to calculate ϕ from $\phi = \tan^{-1}(y/x)$ alone. The reader, however, should be aware that some computer languages and some hand calculator functions will inspect the signs of x and y for you and will return ϕ in its correct quadrant. For example, in FORTRAN, the function ATAN2(X,Y) (or DATAN2(X,Y) in double precision) will return ϕ uniquely in its correct quadrant (though perhaps as a negative angle, in which case 360° should be added to the outputted angle) provided the arguments X and Y are inputted with their correct signs. This can save an immense amount of trouble in programming, and the reader should become familiar with this function.

Direction Cosines

The direction to a point in three dimensional space relative to the origin can be described, as we have seen, by the two angles θ and ϕ. Another way of describing the direction to a point, or the orientation of a vector, is to give the angles α, β, γ that the vector makes with the x-, y- and z-axes, respectively (see figure $\text{III}.5$). The angle γ is the same as the angle θ.

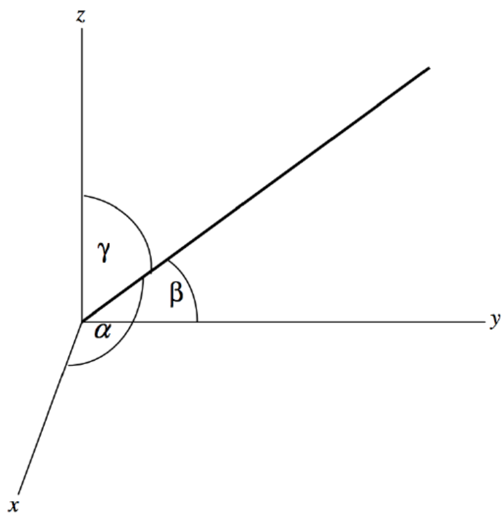


FIGURE III.6

More commonly one quotes the cosines of these three angles. These are called the *direction cosines*, and are often denoted by (l, m, n). It should not take long for the reader to be convinced that the relation between the direction cosines and the angles θ and ϕ are

$$l = \cos\alpha = \sin\theta\cos\phi \tag{3.3.7}$$

$$m = \cos\beta = \sin\theta\sin\phi \tag{3.3.8}$$

$$n = \cos\gamma = \cos\theta \tag{3.3.9}$$

These are not independent, and are related by

$$l^2 + m^2 + n^2 = 1. \tag{3.3.10}$$

A set of numbers that are multiples of the direction cosines - i.e. are proportional to them - are called *direction ratios*.

Latitude and Longitude

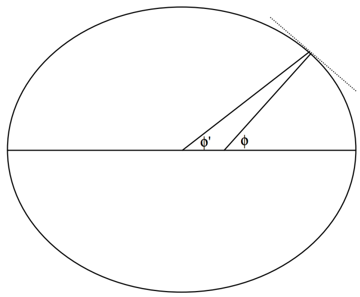
The figure of the Earth is not perfectly spherical, for it is slightly flattened at the poles. For the present, however, our aim is to become familiar with spherical coordinates and with the geometry of the sphere, so we shall suppose the Earth to be spherical. In that case, the position of any town on Earth can be expressed by two coordinates, the latitude ϕ, measured north or south of the equator, and the longitude λ, measured eastwards or westwards from the meridian through Greenwich. These symbols, ϕ for latitude and λ for longitude, are unfortunate, but are often used in this context. In terms of the symbols θ, ϕ for spherical coordinates that we have used hitherto, the east longitude would correspond to ϕ and the latitude to $90^\circ - \theta$.

A plane that intersects a sphere does so in a circle. If that plane passes through the centre of the sphere (so that the centre of the circle is also the centre of the sphere), the circle is called a *great circle*. All the meridians (the circles of fixed longitude that pass through the north and south poles) including the one that passes through Greenwich, are great circles, and so is the equator. Planes that do not pass through the centre of the sphere (such as parallels of latitude) are *small circles*. The radius of a parallel of latitude is equal to the radius of the sphere times the cosine of the latitude.

We have used the example of latitude and longitude on a spherical Earth in order to illustrate the concepts of great and small circles. Although it is not essential to pursue it in the present context, we mention in passing that the true figure of the Earth at mean sea level is a *geoid* - which merely means the shape of the Earth. To a good approximation, the geoid is an oblate spheroid (i.e. an ellipse rotated about its minor axis) with semi major axis $a = 6378.140 \text{ km}$ and semi minor axis $c = 6356.755 \text{ km}$. The ratio $(a - c)/a$ is called the *geometric ellipticity* of the Earth and it has the value $1/298.3$ The mean radius of the Earth, in the sense of the radius of a sphere having the same volume as the actual geoid, is a $\sqrt[3]{a^2c} = 6371.00$ km.

It is necessary in precise geodesy to distinguish between the geographic or geodetic latitude ϕ of a point on the Earth's surface and its geocentric latitude ϕ'. Their definitions evident from figure $\text{III}.7$. In this figure, the ellipticity of the Earth is greatly exaggerated; in reality it would scarcely be discernible. The angle ϕ is the angle between a plumb-bob and the equator. This differs from ϕ' partly because the gravitational field of a spheroid is not the same as that of an equal point mass at the centre, and partly because the plumb bob is pulled away from the Earth's rotation axis by centrifugal force.

FIGURE III.7



The relationship between ϕ and ϕ' is

$$\phi - \phi' = 692'.74 \sin 2\phi - 1'.16 \sin 4\phi. \tag{3.3.4}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

3.4: Velocity and Acceleration Components

Two-dimensional polar coordinates

Sometimes the symbols r and θ are used for two-dimensional polar coordinates, but in this section I use (ρ, ϕ) for consistency with the (r, θ, ϕ) of three-dimensional spherical coordinates. In what follows I am setting vectors in **boldface**. If you make a print-out, you should be aware that some printers apparently do not print Greek letter symbols in boldface, even though they appear in boldface on screen. You should be on the look-out for this. Symbols with ^ above them are intended as unit vectors, so you will know that they should be in boldface even if your printer does not recognize this. If in doubt, look at what appears on the screen.

FIGURE III.8

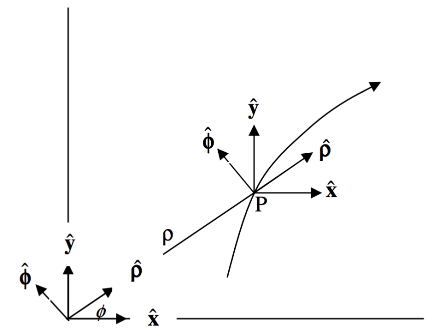


Figure III.8 shows a point P moving along a curve such that its polar coordinates are changing at rates $\dot{\rho}$ and $\dot{\phi}$. The drawing also shows fixed unit vectors $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ parallel to the x- and y-axes, as well as unit vectors $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\phi}}$ in the radial and transverse directions. We shall find expressions for the rate at which the unit radial and transverse vectors are changing with time. (Being unit vectors, their magnitudes do not change, but their directions do.)

We have

$$\hat{\boldsymbol{\rho}} = \cos\phi\hat{\boldsymbol{x}} + \sin\phi\hat{\boldsymbol{y}} \tag{3.4.1}$$

and

$$\hat{\boldsymbol{\phi}} = -\sin\phi\hat{\boldsymbol{x}} + \cos\phi\hat{\boldsymbol{y}}. \tag{3.4.2}$$

$$\therefore \quad \dot{\hat{\boldsymbol{\rho}}} = -\sin\phi\dot{\phi}\hat{\boldsymbol{x}} + \cos\phi\dot{\phi}\hat{\boldsymbol{y}} = \dot{\phi}(-\sin\phi\hat{\boldsymbol{x}} + \cos\phi\hat{\boldsymbol{y}}) \tag{3.4.3}$$

$$\therefore \quad \dot{\hat{\boldsymbol{\rho}}} = \dot{\phi}\hat{\boldsymbol{\phi}} \tag{3.4.4}$$

In a similar manner, by differentiating Equation 3.4.2. with respect to time and then making use of Equation 3.4.1, we find

$$\dot{\hat{\boldsymbol{\phi}}} = -\dot{\phi}\hat{\boldsymbol{\rho}} \tag{3.4.5}$$

Equations 3.4.4 and 3.4.5 give the rate of change of the radial and transverse unit vectors. It is worthwhile to think carefully about what these two Equations mean.

The position vector of the point P can be represented by the expression $\boldsymbol{\rho} = \rho\hat{\boldsymbol{\rho}}$. The velocity of P is found by differentiating this with respect to time:

$$\mathbf{v} = \dot{\boldsymbol{\rho}} = \dot{\rho}\hat{\boldsymbol{\rho}} + \rho\dot{\hat{\boldsymbol{\rho}}} = \dot{\rho}\hat{\boldsymbol{\rho}} + \rho\dot{\phi}\hat{\boldsymbol{\phi}}. \tag{3.4.6}$$

The radial and transverse components of velocity are therefore $\dot{\phi}$ and $\rho\dot{\phi}$ respectively. The acceleration is found by differentiation of Equation 3.4.6, and we have to differentiate the products of two and of three quantities that vary with time:

$$\begin{aligned} \mathbf{a} = \dot{\mathbf{v}} &= \ddot{\rho}\hat{\boldsymbol{\rho}} + \dot{\rho}\dot{\hat{\boldsymbol{\rho}}} + \dot{\rho}\dot{\phi}\hat{\boldsymbol{\phi}} + \rho\ddot{\phi}\hat{\boldsymbol{\phi}} + \rho\dot{\phi}\dot{\hat{\boldsymbol{\phi}}} \\ &= \ddot{\rho}\hat{\boldsymbol{\rho}} + \dot{\rho}\dot{\phi}\hat{\boldsymbol{\phi}} + \dot{\rho}\dot{\phi}\hat{\boldsymbol{\phi}} + \rho\ddot{\phi}\hat{\boldsymbol{\phi}} - \rho\dot{\phi}^2\hat{\boldsymbol{\rho}} \\ &= \left(\ddot{\rho} - \rho\dot{\phi}^2\right)\hat{\boldsymbol{\rho}} + \left(\rho\ddot{\phi} + 2\dot{\rho}\dot{\phi}\right)\hat{\boldsymbol{\phi}}. \end{aligned} \tag{3.4.7}$$

The radial and transverse components of acceleration are therefore $(\ddot{\rho} - \rho\dot{\phi}^2)$ and $(\rho\ddot{\phi} + 2\dot{\rho}\dot{\phi})$ respectively.

Three-Dimensional Spherical Coordinates

In figure III.9, P is a point moving along a curve such that its spherical coordinates are changing at rates $\dot{r}, \dot{\theta}, \dot{\phi}$. We want to find out how fast the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$ in the radial, meridional and azimuthal directions are changing.

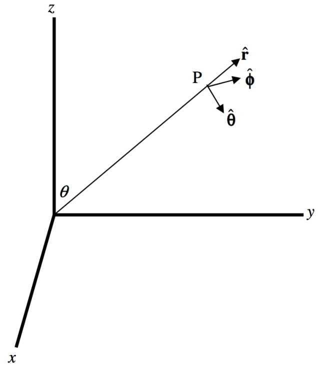


FIGURE III.9

We have

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\hat{\mathbf{x}} + \sin\theta\sin\phi\hat{\mathbf{y}} + \cos\theta\hat{\mathbf{z}} \tag{3.4.8}$$

$$\hat{\boldsymbol{\theta}} = \cos\theta\cos\phi\hat{\mathbf{x}} + \cos\theta\sin\phi\hat{\mathbf{y}} - \sin\theta\hat{\mathbf{z}} \tag{3.4.9}$$

$$\hat{\boldsymbol{\phi}} = -\sin\phi\hat{\mathbf{x}} + \cos\phi\hat{\mathbf{y}} \tag{3.4.10}$$

$$\therefore \quad \dot{\hat{\mathbf{r}}} = (\cos\theta\dot{\theta}\cos\phi - \sin\theta\sin\phi\dot{\phi})\hat{\mathbf{x}} + (\cos\theta\dot{\theta}\sin\phi + \sin\theta\cos\phi\dot{\phi})\hat{\mathbf{y}} - \sin\theta\dot{\theta}\hat{\mathbf{z}}. \tag{3.4.11}$$

We see, by comparing this with Equations 3.4.9 and 3.4.10 that

$$\dot{\hat{\mathbf{r}}} = \dot{\theta}\hat{\boldsymbol{\theta}} + \sin\theta\dot{\phi}\hat{\boldsymbol{\phi}} \tag{3.4.12}$$

By similar arguments we find that

$$\dot{\hat{\boldsymbol{\theta}}} = \cos\theta\dot{\phi}\hat{\boldsymbol{\phi}} - \dot{\theta}\hat{\mathbf{r}} \tag{3.4.13}$$

and

$$\dot{\hat{\boldsymbol{\phi}}} = -\sin\theta\dot{\phi}\hat{\mathbf{r}} - \cos\theta\dot{\phi}\hat{\boldsymbol{\theta}} \tag{3.4.14}$$

These are the rates of change of the unit radial, meridional and azimuthal vectors. The position vector of the point P can be represented by the expression $\mathbf{r} = r\,\hat{\mathbf{r}}$. The velocity of P is found by differentiating this with respect to time:

$$\begin{aligned} \mathbf{v} &= \dot{\mathbf{r}} = \dot{r}\hat{\mathbf{r}} + r\,\dot{\hat{\mathbf{r}}} = \dot{r}\hat{r} + r(\dot{\theta}\hat{\boldsymbol{\theta}} + \sin\theta\dot{\phi}\hat{\boldsymbol{\phi}}) \\ &= \dot{r}\hat{\mathbf{r}} + r\,\dot{\theta}\hat{\boldsymbol{\theta}} + r\sin\theta\dot{\phi}\hat{\boldsymbol{\phi}} \end{aligned} \tag{3.4.15}$$

The radial, meridional and azimuthal components of velocity are therefore $\dot{r}$, $r\dot{\theta}$ and $r\sin\theta\dot{\phi}$ respectively.

The acceleration is found by differentiation of Equation 3.4.15.

It might not be out of place here for a quick hint about differentiation. Most readers will know how to differentiate a product of two functions. If you want to differentiate a product of several functions, for example four functions, a, b, c and d, the procedure is $(abcd)' = a'bcd + ab'cd + abc'd + abcd'$.

In the last term of Equation 3.4.15, all four quantities vary with time, and we are about to differentiate the product.

$$\mathbf{a} = \dot{\mathbf{v}} = \ddot{r}\hat{\mathbf{r}} + \dot{r}(\dot{\theta}\hat{\boldsymbol{\theta}} + \sin\theta\dot{\phi}\hat{\boldsymbol{\phi}}) + \dot{r}\dot{\theta}\hat{\boldsymbol{\theta}} + r\ddot{\theta}\hat{\boldsymbol{\theta}} + r\dot{\theta}(\cos\theta\dot{\phi}\hat{\boldsymbol{\phi}} - \dot{\theta}\hat{\mathbf{r}}) + \dot{r}\sin\theta\dot{\phi}\hat{\boldsymbol{\phi}} + r\cos\theta\dot{\theta}\dot{\phi}\hat{\boldsymbol{\phi}} + r\sin\theta\ddot{\phi}\hat{\boldsymbol{\phi}} + r\sin\theta\dot{\phi}(-\sin\theta\dot{\phi}\hat{\mathbf{r}} - \cos\theta\dot{\phi}\hat{\boldsymbol{\theta}}) \tag{3.4.16}$$

On gathering together the coefficients of $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\phi}}$, we find that the components of acceleration are:

- Radial: $\ddot{r} - r\dot{\theta}^2 - r\sin^2\theta\dot{\phi}^2$
- Meridional: $r\ddot{\theta} + 2\dot{r}\dot{\theta} - r\sin\theta\cos\theta\dot{\phi}^2$
- Azimuthal: $2\dot{r}\dot{\phi}\sin\theta + 2r\dot{\theta}\dot{\phi}\cos\theta + r\sin\theta\ddot{\phi}$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

3.5: Spherical Triangles

As with plane triangles, we denote the three angles by A, B, C and the sides opposite to them by a, b, c. We are fortunate in that we have four formulas at our disposal for the solution of a spherical triangle, and, as with plane triangles, the art of solving a spherical triangle entails understanding which formula is appropriate under given circumstances. Each formula contains four elements (sides and angles), three of which, in a given problem, are assumed to be known, and the fourth is to be determined.

Three important points are to be noted before we write down the formulas.

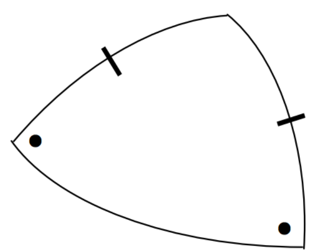
1. The formulas are valid only for triangles in which the three sides are arcs of great circles. They will not do, for example, for a triangle in which one side is a parallel of latitude.
2. The sides of a spherical triangle, as well as the angles, are all expressed in angular measure (degrees and minutes) and not in linear measure (metres or kilometres). A side of 50° means that the side is an arc of a great circle subtending an angle of 50° at the centre of the sphere.
3. The sum of the three angles of a spherical triangle add up to more than 180°.

In this section are now given the four formulas without proof, the derivations being given in a later section. The four formulas may be referred to as the sine formula, the cosine formula, the polar cosine formula, and the cotangent formula. Beneath each formula is shown a spherical triangle in which the four elements contained in the formula are highlighted.

The sine formula:

$$\frac{\sin a}{\sin A} = \frac{\sin b}{\sin B} \left(= \frac{\sin c}{\sin C} \right) \tag{3.5.1}$$

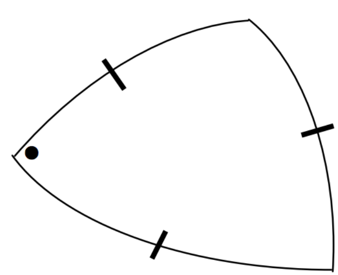
FIGURE III.10



The cosine formula:

$$\cos a = \cos b \cos c + \sin b \sin c \cos A \tag{3.5.2}$$

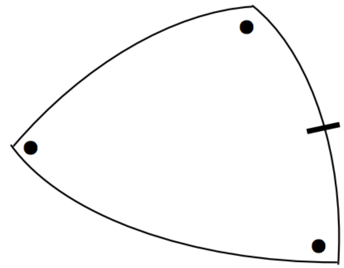
FIGURE III.11



The polar cosine formula:

$$\cos A = -\cos B \cos C + \sin B \sin C \cos a \tag{3.5.3}$$

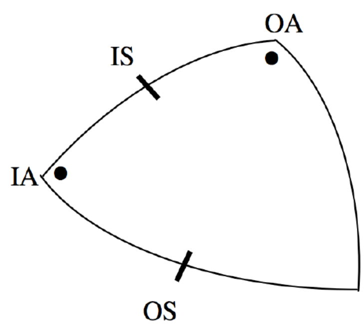
FIGURE III.12



The cotangent formula:

$$\cos b \cos A = \sin b \cot c - \sin A \cot C \tag{3.5.4}$$

FIGURE III.13



The cotangent formula is a particularly useful and frequently needed formula, and it is unfortunate that it is not only difficult to commit to memory but, even with the formula written out in front of one, it is often difficult to decide which is b, which is A and so on. However, it should be noted from the drawing that the four elements, side-angle-side-angle, lie adjacent to each other in the triangle, and they may be referred to as outer side (OS), inner angle (IA), inner side (IS) and outer angle (OA) respectively. Many people find that the formula is much easier to use when written in the form

$$\cos(\mathrm{IS})\cos(\mathrm{IA}) = \sin(\mathrm{IS})\cot(\mathrm{OS}) - \sin(\mathrm{IA})\cot(\mathrm{OA}) \tag{3.5.5}$$

The reader will shortly be offered a goodly number of examples in the use of these formulas. However, during the course of using the formulas, it will be found that there is frequent need to solve deceptively simple trigonometric Equations of the type

$$4.737 \sin\theta + 3.286 \cos\theta = 5.296 \tag{3.5.6}$$

After perhaps a brief pause, one of several methods may present themselves to the reader - but not all methods are equally satisfactory. I am going to suggest four possible ways of solving this Equation. The first method is one that may occur very quickly to the reader as being perhaps rather obvious - but there is a cautionary tale attached to it. While the method may seem very obvious, a difficulty does arise, and the reader would be advised to prefer one of the less obvious methods. There are, incidentally, two solutions to the Equation between 0° and 360°. They are $31^\circ 58'.6$ and $78^\circ 31'.5$.

Method i

The obvious method is to isolate $\cos\theta$:

$$\cos\theta = 1.611\ 686 - 1.441\ 570\ \sin\theta. \tag{3.5.1}$$

Although the constants in the problem were given to four significant figures, do not be tempted to round off intermediate calculations to four. It is a common fault to round off intermediate calculations prematurely. The rounding-off can be done at the end.

Square both sides, and write the left hand side, $\cos^2\theta$, as $1 - \sin^2\theta$. We now have a quadratic Equation in $\sin\theta$:

$$3.078\ 125 \sin^2\theta - 4.646\ 717 \sin\theta + 1.597\ 532 = 0. \tag{3.5.2}$$

The two solutions for $\sin\theta$ are $0.529\ 579$ and $0.908\ 014$ and the four values of θ that satisfy these values of $\sin\theta$ are $31^\circ 58'.6,\ 148^\circ\ 01'.4,\ 78^\circ\ 31'.5$ and $101^\circ 28'.5$.

Only two of these angles are solutions of the original Equation. The fatal move was to square both sides of the original Equation, so that we have found solutions not only to

$$\cos\theta = 1.611\ 686 - 1.441\ 570 \sin\theta \tag{3.5.3}$$

but also to the different Equation

$$-\cos\theta = 1.611\ 686 - 1.441\ 570\ \sin\theta. \tag{3.5.4}$$

This generation of extra solutions always occurs whenever we square an Equation. For this reason, method (i), however tempting, should be avoided, particularly when programming a computer to carry out a computation automatically and uncritically.

If in doubt whether you have obtained a correct solution, substitute your solution in the original Equation. You should always do this with any Equation of any sort, anyway.

Method ii

This method makes use of the identities

$$\sin\theta = \frac{2t}{1+t^2}, \quad \cos\theta = \frac{1-t^2}{1+t^2}, \tag{3.5.5}$$

where $t = \tan\frac{1}{2}\theta$.

When applied to the original Equation, this results in the quadratic Equation in t:

$$8.582t^2 - 9.474t + 2.010 = 0 \tag{3.5.6}$$

with solutions

$$t = 0.286528 \quad \text{and} \quad t = 0.817410 \tag{3.5.7}$$

The only values of θ between 0° and 360° that satisfy these are the two correct solutions $31^\circ\ 58'.6$ and $78^\circ\ 31'.5$.

It is left as an exercise to show, using this method algebraically, that the solutions to the Equation

$$a\sin\theta + b\cos\theta = c \tag{3.5.8}$$

are given by

$$\tan\frac{1}{2}\theta = \frac{a \pm \sqrt{a^2+b^2-c^2}}{b+c}. \tag{3.5.9}$$

This shows that there are no real solutions if $a^2+b^2<c^2$, one real solution if $a^2+b^2=c^2$, and two real solutions if $a^2+b^2>c^2$.

Method iii

We divide the original Equation

$$4.737\sin\theta + 3.286\cos\theta = 5.296 \tag{3.5.10}$$

by the "hypotenuse" of 4.737 and 3.286; that is, by $\sqrt{(4.737^2+3.286^2)} = 5.765151$.

Thus

$$0.821\ 661\ \sin\theta + 0.569\ 976\cos\theta = 0.918\ 623 \tag{3.5.11}$$

Now let $0.821\ 661 = \cos\alpha$ and $0.569976 = \sin\alpha$ (which we can, since these numbers now satisfy $\sin^2\alpha\cos^2\alpha = 1$) so that $\alpha = 34^\circ\ 44'.91$.

We have

$$\cos \alpha \sin \theta+\sin \alpha \cos \theta=0.918\ 623 \tag{3.5.12}$$

or

$$\sin (\theta+\alpha)=0.918623 \tag{3.5.13}$$

from which

$$\theta+\alpha=66^{\circ}\ 43^{\prime}.54 \text{ or } 113^{\circ}\ 16^{\prime}.46 \tag{3.5.14}$$

Therefore

$$\theta=31^{\circ}\ 58^{\prime}.6 \text{ or } 78^{\circ}\ 31^{\prime}.5 \tag{3.5.15}$$

Method iv

Methods ii and iii give explicit solutions, so there is perhaps no need to use numerical methods. Nevertheless, the reader might like to solve, by Newton-Raphson iteration, the Equation

$$f(\theta)=a \sin \theta+b \cos \theta-c=0, \tag{3.5.16}$$

for which

$$f^{\prime}(\theta)=a \cos \theta-b \sin \theta. \tag{3.5.17}$$

Using the values of a, b and c from the example above and using the Newton-Raphson algorithm, we find with a first guess of 45° the following iterations, working in radians:

$$\begin{array}{lll} 0.785\ 398 & & \\ 0.417\ 841 & & \\ 0.541\ 499 & & \\ 0.557\ 797 & & \\ 0.558\ 104 & & \\ 0.558\ 104 & = & 31^{\circ} 58^{\prime}.6 \end{array} \tag{3.5.18}$$

The reader should verify this calculation, and, using a different first guess, show that NewtonRaphson iteration quickly leads to $78^{\circ}\ 31^{\prime}.5$.

Having now cleared that small hurdle, the reader is invited to solve the spherical triangle problems below. Although these twelve problems look like pointless repetitive work, they are in fact all different. Some have two solutions between 0° and 360° ; others have just one. After solving each problem, the reader should sketch each triangle - especially those that have two solutions - in order to see how the two-fold ambiguities arise. The reader should also write a computer program that will solve all twelve types of problem at the bidding of the user. Answers should be given in degrees, minutes and tenths of a minute, and should be correct to that precision. For example, the answer to one of the problems is $47^{\circ}\ 37^{\prime}.3$. An answer of $47^{\circ}\ 37^{\prime}.2$ or $47^{\circ}\ 37^{\prime}.4$ should be regarded as wrong. In celestial mechanics, there is no place for answers that are "nearly right". An answer is either right or it is wrong. (This does not mean, of course, that an angle can be measured with no error at all; but the answer to a calculation given to a tenth of an arcminute should be correct to a tenth of an arcminute.)

Exercise 1

All angles and sides in degrees

$$\begin{array}{lllll}
10. & a=64 & b=33 & c=37 & C=? \\
11. & a=39 & b=48 & C=74 & c=? \\
12. & a=16 & b=37 & C=42 & B=? \\
13. & a=21 & b=43 & A=29 & c=? \\
14. & a=67 & b=54 & A=39 & B=? \\
15. & a=49 & b=59 & A=14 & C=? \\
16. & A=24 & B=72 & c=19 & a=? \\
17. & A=79 & B=84 & c=12 & C=? \\
18. & A=62 & B=49 & a=44 & b=? \\
19. & A=59 & B=32 & a=62 & c=? \\
20. & A=47 & B=57 & a=22 & C=? \\
21. & A=79 & B=62 & C=48 & c=?
\end{array} \tag{3.5.19}$$

Solutions

10. $28^\circ 18'.2$
11. $49^\circ 32'.4$
12. $117^\circ 31'.0$
13. $30^\circ 46'.7$ or $47^\circ 37'.3$
14. $33^\circ 34'.8$
15. $3^\circ 18'.1$ or $162^\circ 03'.9$
16. $7^\circ 38'.2$
17. $20^\circ 46'.6$
18. $36^\circ 25'.5$
19. $76^\circ 27'.7$
20. $80^\circ 55'.7$ or $169^\circ 05'.2$
21. $28^\circ 54'.6$

Derivation of the formulas

Before moving on to further problems and applications of the formulas, it is time to derive the four formulas which, until now, have just been given without proof. We start with the cosine formula. There is no loss of generality in choosing rectangular axes such that the point A of the spherical triangle ABC is on the z-axis and the point B and hence the side c are in the zx-plane. The sphere is assumed to be of unit radius.

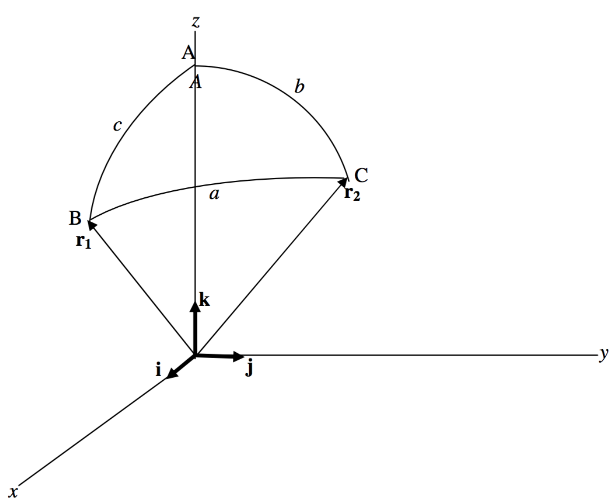


FIGURE III.14

If $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are unit vectors directed along the $x-$, $y-$ and $z-$axes respectively, inspection of the figure will show that the position vectors of the points B and C with respect to the centre of the sphere are

$$\mathbf{r}_1 = \mathbf{i}\sin c + \mathbf{k}\cos c \tag{3.5.7}$$

and

$$\mathbf{r}_2 = \mathbf{i}\sin b\cos A + \mathbf{j}\sin b\sin A + \mathbf{k}\cos b \tag{3.5.8}$$

respectively.

The scalar product of these vectors (each of magnitude unity) is just the cosine of the angle between them, namely $\cos a$, from which we obtain immediately

$$\cos a = \cos b\cos c + \sin b\sin c\cos A. \tag{3.5.9}$$

To obtain the sine formula, we isolate $\cos A$ from this Equation, square both sides, and write $1 - \sin^2 A$ for $\cos^2 A$. Thus,

$$(\sin b\sin c\cos A)^2 = (\cos a - \cos b\cos c)^2, \tag{3.5.10}$$

and when we have carried out these operations we obtain

$$\sin^2 A = \frac{\sin^2 b\sin^2 c - \cos^2 a - \cos^2 b\cos^2 c + 2\cos a\cos b\cos c}{\sin^2 b\sin^2 c}. \tag{3.5.11}$$

In the numerator, write $1 - \cos^2 b$ for $\sin^2 b$ and $1 - \cos^2 c$ for $\sin^2 c$, and divide both sides by $\sin^2 a$. This results in

$$\frac{\sin^2 A}{\sin^2 a} = \frac{1 - \cos^2 a - \cos^2 b - \cos^2 c + 2\cos a\cos b\cos c}{\sin^2 a\sin^2 b\sin^2 c}. \tag{3.5.12}$$

At this stage the reader may feel that we are becoming bogged down in heavier and heavier algebra and getting nowhere. But, after a careful look at Equation 3.5.12, it may be noted with some delight that the next line is:

Therefore

$$\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}. \tag{3.5.13}$$

The derivation of the polar cosine formula may also bring a small moment of delight. In figure III.15, $\mathrm{A'B'C'}$ is a spherical triangle. ABC is also a spherical triangle, called the polar triangle to $\mathrm{A'B'C'}$. It is formed in the following way. The side BC is an arc of a great circle 90° from $\mathrm{A'}$; that is, BC is part of the equator of which $\mathrm{A'}$ is pole. Likewise CA is 90° from $\mathrm{B'}$ and AB is 90° from $\mathrm{C'}$. In the drawing, the side $\mathrm{B'C'}$ of the small triangle has been extended to meet the sides AB and CA of the large triangle. It will be evident from the drawing that the angle A of the large

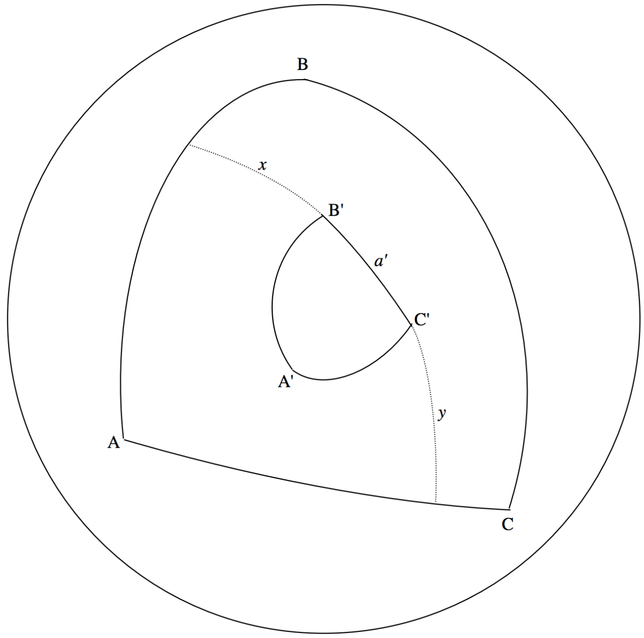


FIGURE III.15

triangle is equal to $x+a'+y$. Further, from the way in which the triangle ABC was formed, $x+a'$ and $a'+y$ are each equal to 90°. From these relations, we see that

$$A+A=[(x+a')+y]+[x+(a'+y)] \tag{3.5.20}$$

or

$$2A=180^\circ+x+y=180^\circ+A-a' \tag{3.5.21}$$

Therefore

$$A=180^\circ-a' \tag{3.5.22}$$

In a similar manner,

$$B=180^\circ-b' \text{ and } C=180^\circ-c' \tag{3.5.23}$$

Now, suppose $f(A',B',C',a',b',c')=0$ is any relation between the sides and angles of the triangle $\mathrm{A'B'C'}$. We may replace a' by $180^\circ-A$, b' by $180^\circ-B$, and so on, and this will result in a relation between A, B, C, a, b and c ; that is, it will result in a relation between the sides and angles of the triangle ABC.

For example, the Equation

$$\cos a'=\cos b'\cos c'+\sin b'\sin c'\cos A' \tag{3.5.14}$$

is valid for the triangle $\mathrm{A'B'C'}$. By making these substitutions, we find the following formula valid for triangle ABC:

$$-\cos A=\cos B\cos C-\sin B\sin C\cos a, \tag{3.5.15}$$

which is the polar cosine formula.

The reader will doubtless like to try starting from the sine and cotangent formulas for the triangle $\mathrm{A'B'C'}$ and deduce corresponding polar formulas for the triangle ABC, though this, unfortunately, may give rise to some anticlimactic disappointment.

I know of no particularly interesting derivation of the cotangent formula, and I leave it to the reader to work through the rather pedestrian algebra. Start from

$$\cos a=\cos b\cos c+\sin b\sin c\cos A \tag{3.5.24}$$

and

$$\cos c = \cos a \cos b + \sin a \sin b \cos C. \tag{3.5.25}$$

Eliminate $\cos c$ (but retain $\sin c$) from these Equations, and write $1-\sin^2 b$ for $\cos^2 b$. Finally substitute $\frac{\sin c \sin A}{\sin C}$ for $\sin a$, and, after some tidying up, the cotangent formula should result.

At this stage, we have had some practice in solving the four spherical triangle formulas, and we have derived them. In this section we encounter examples in which the problem is not merely to solve a triangle, but to gain some experience in setting up a problem and deciding which triangle has to be solved.

Example 1

The coordinates of the Dominion Astrophysical Observatory, near Victoria, British Columbia, are

Latitude $48^\circ 31'.3\text{N}$ Longitude $123^\circ 25'.0\text{W}$

and the coordinates of the David Dunlap Observatory, near Toronto, Ontario, are

Latitude $43^\circ 51'.8\text{N}$ Longitude $79^\circ 25'.3\text{W}$

How far is Toronto from Victoria, and what is the azimuth of Toronto relative to Victoria?

The triangle to be drawn and solved is the triangle PVT, where P is the Earth's north pole, V is Victoria, and T is Toronto. On figure $\text{III}.16$ are marked the colatitudes of the two cities and the difference between their longitudes.

The great circle distance ω between the two observatories is easily given by the cosine formula:

$$\cos \omega = \cos 41^\circ 28'.7 \cos 46^\circ 08'.2 + \sin 41^\circ 28'.7 \sin 46^\circ 08'.2 \cos 43^\circ 59'.7 \tag{3.5.26}$$

From this, we find $\omega = 30^\circ 22'.7$ or 0.53021 radians. The radius of the Earth is 6371 km, so the distance between the observatories is 3378 km or 2099 miles.

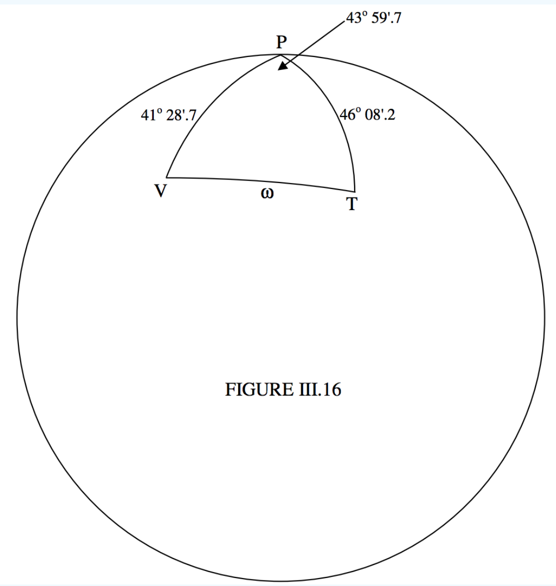


FIGURE III.16

Now that we have found ω, we can find the azimuth, which is the angle V, from the sine formula:

$$\sin V = \frac{\sin 46^\circ 08'.2 \ \sin 43^\circ 59'.7}{\sin 30^\circ 22'.7} = 0.990\,275 \tag{3.5.27}$$

and hence

$$V = 82^\circ 00'.3 \tag{3.5.28}$$

But we should now remember that $\sin^{-1} 0.990\,275$ has two values between 0° and 180°, namely $82^\circ 00'.3$ and $97^\circ 59'.7$.

Usually it is obvious from inspection of a drawing which of the two values of $\sin^{-1}$ is the required one. Unfortunately, in this case, both values are close to 90°, and it may not be immediately obvious which of the two values we require. However, it will be noticed that Toronto has a more southerly latitude than Victoria, and this should easily resolve the ambiguity.

We could, of course, have found the azimuth V by using the cotangent formula, without having to calculate ω first. Thus

$$\cos 41^\circ 28'.7 \cos 43^\circ 59'.7 = \sin 41^\circ 28'.7 \cot 46^\circ 08'.2 - \sin 43^\circ 59'.7 \cot V \tag{3.5.29}$$

There is only one solution for V between 0° and 180°, and it is the correct one, namely $82^\circ 00'.3$. A good drawing will show the reader why the correct solution was the acute rather than the obtuse angle (in our drawing the angle was made to be close to 90° in order not to bias the reader one way or the other), but in any case all readers, especially those who were trapped into choosing the obtuse angle, *should take careful note of the difficulties that can be caused by the ambiguity of the function* $\sin^{-1}$. *Indeed it is the strong advice of the author never to use the sine formula, in spite of the ease of memorizing it. The cotangent formula is more difficult to commit to memory, but it is far more useful and not so prone to quadrant mistakes.*

Example 2

Consider two points, A and B, at latitude 20°N, longitude 25°E, and latitude 72°N, longitude 44°E. Where are the poles of the great circle passing through these two points? We shall present two methods of doing the problem. First, by solving spherical triangles. And second, kindly suggested to me by Achintya Pal, using the methods of algebraic coordinate geometry.

Let us call the colatitude and longitude of the first point (θ_1, ϕ_1) and of the second point (θ_2, ϕ_2) We shall consider the question answered if we can find the coordinates (θ_0, ϕ_0) of the poles Q and Q' of the great circle passing through the two points. In figure $\text{III}.17$, P is the north pole of the Earth, A and B are the two points in question, and Q is one of the two poles of the great circle joining A and B. The figure also shows the triangle PQA. We'll suppose that the origin for longitudes ("Greenwich") is behind the plane of the paper. The east longitudes of Q, A and B are, respectively, ϕ_0, ϕ_1, ϕ_2; and their colatitudes are $\theta_0, \theta_1, \theta_2$.

$$0 = \cos\theta_0 \cos\theta_1 + \sin\theta_0 \sin\theta_1 \cos(\phi_1 - \phi_0), \tag{3.5.16}$$

from which

$$\tan\theta_0 = -\frac{1}{\tan\theta_1 \cos(\phi_1 - \phi_0)}. \tag{3.5.17}$$

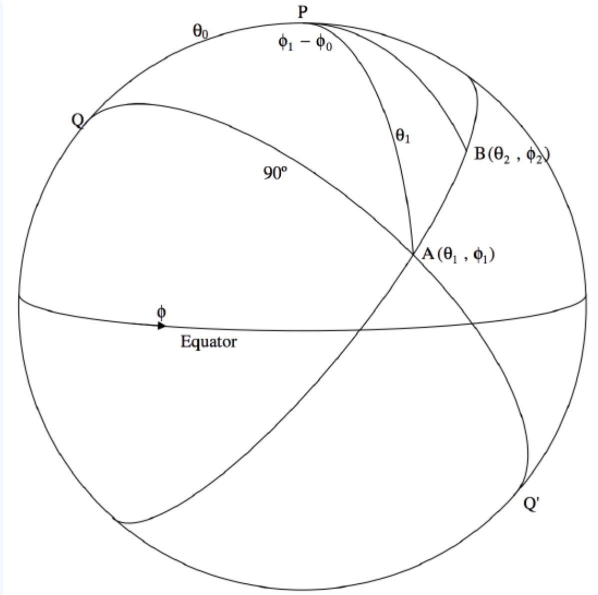


FIGURE III.17

Similarly from triangle PQB we would obtain

$$\tan\theta_0 = -\frac{1}{\tan\theta_2\cos(\phi_2-\phi_0)}. \tag{3.5.18}$$

These are two Equations in θ_0 and ϕ_0, so the problem is in principle solved. Equate the righthand sides of the two Equations, expand the terms $\cos(\phi_1-\phi_0)$ and $\cos(\phi_2-\phi_0)$, gather the terms in $\sin\phi_0$ and $\cos\phi_0$, eventually to obtain

$$\tan\phi_0 = \frac{\tan\theta_1\cos\phi_1-\tan\theta_2\cos\phi_2}{\tan\theta_2\sin\phi_2-\tan\theta_1\sin\phi_1}. \tag{3.5.19}$$

If we substitute the angles given in the original problem, we obtain

$$\tan\phi_0 = \frac{\tan 70^\circ\cos 25^\circ-\tan 18^\circ\cos 44^\circ}{\tan 18^\circ\sin 44^\circ-\tan 70^\circ\sin 25^\circ} = -2.412\,091\,0 \tag{3.5.30}$$

from which

$$\phi_0 = 112^\circ 31'.1 \quad \text{or} \quad 292^\circ 31'.1 \tag{3.5.31}$$

Note that we get two values for ϕ_0 differing by 180°, as expected.

We then use either of the Equations for $\tan\theta_0$ to obtain θ_0 (It is good practice to use both of them as a check on the arithmetic.) The north polar distance, or colatitude, must be between 0° and 180°, so there is no ambiguity of quadrant.

With $\phi_0 = 112^\circ 31'.1$, we obtain $\theta_0 = 96^\circ 47'.1$, i.e. latitude $6^\circ 47'.1$ S.
and with $\phi_0 = 292^\circ 31'.1$, we obtain $\theta_0 = 83^\circ 12'.9$, i.e. latitude $6^\circ 47'.1$ N.

and these are the coordinates of the two poles of the great circle passing through A and B. The reader is strongly urged actually to carry out these computations numerically in order to be quite sure that the quadrants are correct and unambiguous. Indeed, dealing with the quadrant problem may be regarded as the most important part of the exercise.

We arrived at Equation 3.5.17 and 3.5.18 by solving two spherical triangles by the methods of spherical trigonometry. The second method, suggested, as mentioned above, by Achintya Pal, uses the methods of algebraic coordinate geometry in three dimensions to arrive at the same Equations. We refer coordinates to axes Oxyz. O is the centre of the Earth, taken to be of unit radius. OP is the z-axis. The Ox and Oy axes are not drawn in figure III.17, but the x-axis may be taken to be directed somewhere to the rear of the drawing (away from the reader), and the y-axis somewhere in the front of the drawing, both being, of course, in the plane of the equator.

Let us write the Equation to the plane containing A and B in the form

$$lx + my + nz = 0 \tag{3.5.20}$$

Here (l, m, n) are the direction cosines of the normal to the plane AB, and are given by

$$l = \sin\theta_0 \cos\phi_0 \quad m = \sin\theta_0 \sin\phi_0 \quad n = \cos\theta_0 \tag{3.5.21a,b,c}$$

The (x, y, z) coordinates of the point A are

$$x = \sin\theta_1 \cos\phi_1 \quad y = \sin\theta_1 \sin\phi_1 \quad x = \cos\theta_1 \tag{3.5.22a,b,c}$$

On substitution of Equations 3.5.21a,b,cand 3.5.22a,b,cinto Equation 3.5.20we obtain:

$$\sin\theta_0 \cos\phi_0 \sin\theta_1 \cos\phi_1 + \sin\theta_0 \sin\phi_0 \sin\theta_1 \sin\phi_1 + \cos\theta_0 \cos\theta_1 = 0 \tag{3.5.23}$$

After some very modest algebraic manipulation (e.g., start by dividing by $\sin\theta_1 \cos\theta_0$) we very soon arrive again at Equation 3.5.17, and in a similar manner at Equation 3.5.18.

As a bonus we note that any point having spherical coordinates (θ, ϕ) lying on the great circle whole pole is at (θ_0, ϕ_0) satisfies the Equation

$$\cot\theta = -\tan\theta_0 \cos(\phi - \phi) \tag{3.5.24}$$

This Equation may be regarded as the (θ, ϕ) Equation to the great circle AB, and it answers the problem converse to the one originally posed: What is the Equation to the great circle whose pole is at (θ_0, ϕ_0) ?

Example 3

Here is a challenging exercise and an important one in meteor astronomy. Two shower meteors are seen, diverging from a common radiant. One starts at right ascension 6 hours, declination +65 degrees, and finishes at right ascension 1 hour, declination +75 degrees. The second starts at right ascension 5 h, declination +35 degrees, and finishes at right ascension 3 hours, declination +15 degrees. Where is the radiant?

The assiduous student will make a good drawing of the celestial sphere, illustrating the situation as accurately as possible. The calculation will require some imaginative manipulation of spherical triangles. After arriving at what you believe to be the correct answer, look at your drawing to see whether it is reasonable. The next step might be to develop a general trigonometrical expression for the answer in terms of the original data, or to program the calculation for a computer, so that it is henceforth available for any similar calculation. Or one can go yet further, and write a computer program that will give a least-squares solution for the radiant for many more than two meteors in the shower. I find for the answer to the above problem that the radiant is at right ascension 7.26 hours and declination +43.8 degrees.

Uniqueness of Solutions

The reader who has by now worked through a variety of problems in the solution of a triangle will have noticed that, given three elements of a triangle, sometimes there is a unique solution, whereas sometimes there are two possible triangles that satisfy the original data. Yet again, it may sometimes be found that there is no possible solution, meaning that there is no possible triangle that satisfies the given data, which must therefore be presumed incorrect. I am very much indebted to Alan Johnstone for lengthy discussions on this problem, and indeed for pointing out that some of the "solutions" given in an earlier version of these notes were in fact invalid (and have now been corrected). I believe the following criteria determine how many valid solutions there are for a given triplet of data, for plane triangles and for spherical triangles.

We may be given three elements of a triangle,

Thus

i. Three sides: a, b, c,
ii. Two sides and the included angle: b, c, A.
iii. Two sides and a nonincluded angle: a, b, A.
iv. Two angles and a common side: a, B, C.
v. Two angles and another side: A, B, a.
vi. Three angles: A, B, C.

Question:

Which of these give a unique solution, and which admit of two solutions? And which are impossible triangles? I believe the answers are as follows:

Plane Triangles

i. Let $d = a+b-c, \quad e = b+c-a, \quad f = c+a-b$

For a valid triangle, d, e and f must all be positive. If so, there is a unique solution.

ii. There is a unique solution.

iii. If $a > b$ there is a unique solution.

If $a = b$, there is a unique solution if $A < 90^\circ$. Otherwise there is no valid triangle.

If $a < b$ there are zero, one or two solutions, according as to whether

$\sin A > \frac{a}{b}, \sin A = \frac{a}{b}$ or $\sin A < \frac{a}{b}$.

iv. There is a unique solution.

v. There is a unique solution.

vi. There is a unique solution except that only the relative lengths of the sides are determined.

Spherical Triangles

i. Let $d = a+b-c, \quad e = b+c-a, \quad f = c+a-b$

For a valid triangle, d, e and f must all be positive. If so, there is a unique solution.)

ii. There is a unique solution.

iii. If $\sin A > \frac{\sin a}{\sin b}$, there is no real solution.

If $A = a = b = 90^\circ$, then $B = 90^\circ$, and c and C are equal but indeterminate.

Otherwise:

If $a > b$ there is a unique solution.

If $a = b$, there is a unique solution if $a < 90^\circ$. Otherwise there is no real solution.

If $a < b$ there are one or two solutions, according as to whether

$$\sin A = \frac{\sin a}{\sin b} \text{ or } \sin A < \frac{\sin a}{\sin b}. \tag{3.5.32}$$

iv. There is a unique solution.

v. If $\sin A = \frac{\sin A}{\sin B}$, there is no real solution.

If $A = B = a = 90^\circ$, then $b = 90^\circ$, and c and C are equal but indeterminate.

Otherwise:

If $A > B$ there is a unique solution.

If $A = B$, there is a unique solution if $a < 90^\circ$. Otherwise there is no real solution.

If $A < B$ there are one or two solutions, according as to whether

$$\sin a = \frac{\sin A}{\sin B} \text{ or } \sin a < \frac{\sin A}{\sin B}. \tag{3.5.33}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

3.6: Rotation of Axes, Two Dimensions

In this section we consider the following problem. Consider two sets of orthogonal axes, $\mathrm{O}x$, $\mathrm{O}y$, and $\mathrm{O}x'$, $\mathrm{O}y'$, such that one set makes an angle θ with respect to the other. See figure (a) below. A point P can be described either by its coordinates (x, y) with respect to one "basis set" $\mathrm{O}x$, $\mathrm{O}y$, or by its coordinates with respect to the other basis set $\mathrm{O}x'$, $\mathrm{O}y'$. The question is, what is the relation between the coordinates (x, y) and the coordinates (x', y')? See figure III.18.

We see that $\mathrm{OA} = x,\ \mathrm{AP} = y,\ \mathrm{ON} = x',\ \mathrm{PN} = y',\ \mathrm{OM} = x\cos\theta,\ \mathrm{MN} = y\sin\theta$,

$$\therefore \quad x' = x\cos\theta + y\sin\theta. \tag{3.6.1}$$

Also $\mathrm{MA} = \mathrm{NB} = x\sin\theta, \quad \mathrm{PB} = y\cos\theta$,

$$\therefore \quad y' = -x\sin\theta + y\cos\theta. \tag{3.6.2}$$

These two relations can be written in matrix form as

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.6.3}$$

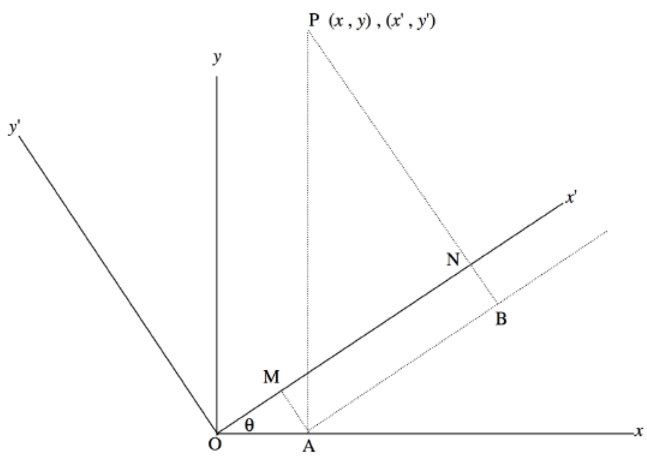


FIGURE III.18

There are several ways of obtaining the converse relations; that is, Equations for x and y in terms of x' and y'. One way would be to design drawings similar to (b) and (c) that show the converse relations clearly, and the reader is encouraged to do this. Another way is merely to solve the above two Equations (which can be regarded as two simultaneous Equations in x and y) for x and y. Less tedious is to interchange the primed and unprimed symbols and change the sign of θ. Perhaps the quickest of all is to recognize that the determinant of the matrix

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{3.6.1}$$

is unity and therefore the matrix is an orthogonal matrix. One important property of an orthogonal matrix $\mathbf{M}$ is that its reciprocal $\mathbf{M}^{-1}$ is equal to its transpose $\tilde{\mathbf{M}}$ (formed by transposing the rows and columns). Therefore the converse relation that we seek is

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}^{-1} \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}. \tag{3.6.4}$$

The reader might like to try all four methods to ensure that they all arrive at the same result.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

3.7: Rotation of Axes, Three Dimensions. Eulerian Angles

We now consider two sets of orthogonal axes $\mathrm{O}x$, $\mathrm{O}y$, $\mathrm{O}z$ and $\mathrm{O}x'$, $\mathrm{O}y'$, $\mathrm{O}z'$ in three-dimensional space and inclined to each other. A point in space can be described by its coordinates (x, y, z) with respect to one basis set or (x', y', z') with respect to the other. What is the relation between the coordinates (x, y, z) and the coordinates (x', y', z')?

We first need to describe exactly how the primed axes are inclined with respect to the unprimed axes. In the figure below are shown the axes $\mathrm{O}x$, $\mathrm{O}y$ and $\mathrm{O}z$. Also shown are the axes $\mathrm{O}x'$ and $\mathrm{O}z'$; the axis $\mathrm{O}y'$ is directed behind the plane of the paper and is not drawn. The orientation of the primed axes with respect to the unprimed axes is described by three angles θ, ϕ and ψ, known as the Eulerian angles, and they are shown in figure $\mathrm{III}.19$.

The precise definitions of the three angles can be understood by three consecutive rotations, illustrated in figures $\mathrm{III}.20,21,22$

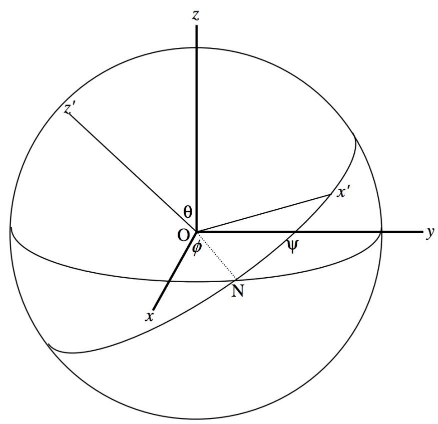


FIGURE III.19

First, a rotation through ϕ counterclockwise around the $\mathrm{O}z$ axis to form a set of intermediate axes $\mathrm{O}x_1$, $\mathrm{O}y_1$, $\mathrm{O}z_1$, as shown in figure $\mathrm{III}.20$. The $\mathrm{O}z$ and $\mathrm{O}z_1$ axes are identical. Part (b) shows the rotation as seen when looking directly down the $\mathrm{O}z$ (or $\mathrm{O}z_1$) axis.

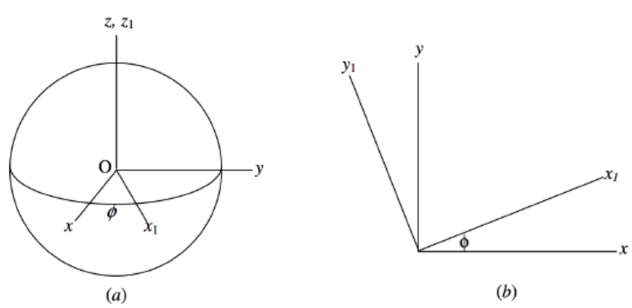


FIGURE III.20

The relation between the (x, y, z) and $(x_1,\ y_1,\ z_1)$ coordinates is

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.7.1}$$

Next, a rotation through θ counterclockwise around the $\mathrm{O}x_1$ axis to form a set of axes $\mathrm{O}x_2$, $\mathrm{O}y_2$, $\mathrm{O}z_2$. The $\mathrm{O}x_1$ and $\mathrm{O}x_2$ axes are identical (Figure $\mathrm{III}.21$). Part (b) of the figure shows the rotation as seen when looking directly towards the origin along the $\mathrm{O}x_1$ (or $\mathrm{O}x_2$) axis.

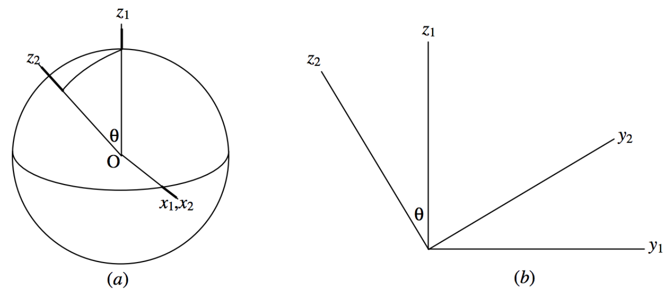


FIGURE III.21

The relation between the (x_1, y_1, z_1) and (x_2, y_2, z_2) coordinates is

$$\begin{pmatrix} y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} y_1 \\ z_1 \end{pmatrix}. \tag{3.7.2}$$

Lastly, a rotation through ψ counterclockwise around the $\mathrm{O}z_2$ axis to form the set of axes $\mathrm{O}x'$, $\mathrm{O}y'$, $\mathrm{O}z'$ (figure III.22). The $\mathrm{O}z_2$ and $\mathrm{O}z'$ axes are identical. Part (b) of the figure shows the rotation as seen when looking directly down the $\mathrm{O}z_2$ (or $\mathrm{O}z'$) axis.

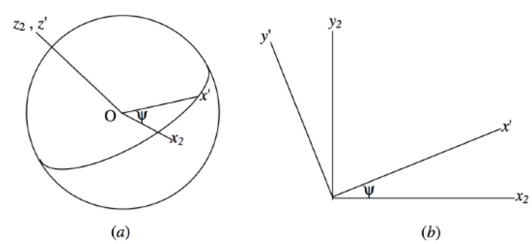


FIGURE III.22

The relation between the (x_2, y_2, z_2) and (x', y', z') coordinates is

$$\begin{pmatrix} x' & y' & z' \end{pmatrix} = \begin{pmatrix} \cos\psi & \sin\psi & 1 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix}. \tag{3.7.3}$$

Thus we have for the relations between (x', y', z') and (x, y, z)

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{3.7.4}$$

On multiplication of these matrices, we obtain

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\psi\cos\phi - \cos\theta\sin\phi\sin\psi & \cos\psi\sin\phi + \cos\theta\cos\phi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\theta\sin\phi\cos\psi & -\sin\psi\sin\phi + \cos\theta\cos\phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{3.7.5}$$

The inverse of this may be found, as in the two-dimensional case, either by solving these three Equations for x, y and z (which would be rather tedious); or by interchanging the primed and unprimed quantities and reversing the order and signs of all operations (replace ψ with $-\phi$, θ with $-\theta$, and ϕ with $-\psi$) which is less tedious; or by recognizing that the determinant of the matrix is unity and therefore its reciprocal is its transpose, which is hardly tedious at all. The reader should verify that the determinant of the matrix is unity by multiplying it out and making use of trigonometric identities. The reason that the determinant must be unity, however, and that the rotation matrix must be orthogonal, is that rotation of axes cannot change the magnitude of a vector.

Each element of the matrix is the cosine of the angle between an axis in one basis set and an axis in the other basis set. For example, the second element in the first row is the cosine of the angles between $\mathrm{O}x'$ and $\mathrm{O}y$. The first element of the third row is the cosine of the angles between $\mathrm{O}z'$ and $\mathrm{O}x$. The matrix can be referred to as the matrix of direction cosines between the axes of one basis set and the axes of the other basis set, and the relations between the coordinates can be written

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{3.7.6}$$

$$\mathbf{R}' = \mathbf{CR}. \tag{3.7.7}$$

You will note the similarity of the forms of the direction cosines to the cosine formula for the solution of a spherical triangle, and indeed the direction cosines can all be derived by drawing and solving the relevant spherical triangles. You might (or might not!)enjoy trying to do this.

The matrix $\mathbf{C}$ of the direction cosines is orthogonal, and the properties of an orthogonal matrix are as follows. The reader should verify this using the formulas for the direction cosines in terms of the Eulerian angles. The properties also apply, of course, although more trivially, to the rotation matrix in two dimensions.

a. det $\mathbf{C} = \pm 1$ and (det $\mathbf{C} = -1$ implies that the two basis sets are of opposite chirality or "handedness"; that is, if one basis set is right-handed, the other is left-handed.)
b. The sum of the squares of the elements in any row or any column is unity. This merely means that the magnitudes of unit orthogonal vectors are indeed unity.
c. The sum of the products of corresponding elements in any two rows or any two columns is zero. This is merely a reflection of the fact that the scalar o dot product of any two unit orthogonal vectors is zero.
d. Every element is equal to its own cofactor. This a reflection of the fact that the vector or cross product of any two unit orthogonal vectors in cyclic order is equal to the third.
e. $\mathbf{C}^{-1} = \tilde{\mathbf{C}}$, or the reciprocal of an orthogonal matrix is equal to its transpose.

The first four properties above can be (and should be) used in a numerical case to verify that the matrix is indeed orthogonal, and they can be used for detecting and for correcting mistakes.

For example, the following matrix is supposed to be orthogonal, but there are, in fact, two mistakes in it. Using properties (b) and (c) above, locate and correct the mistakes. (It will become clear when you do this why verification of property (b) alone is not sufficient.) When you have corrected the matrix, see if you can find the Eulerian angles θ, ϕ and ψ without ambiguity of quadrant. As a hint, start at the bottom right hand side of the matrix and note, from the way in which the Eulerian angles are set up, that θ must be between 0° and 180°, so that there is no ambiguity of quadrant. The other two angles, however, can lie between 0° and 360° and must be determined by examining the signs of their sines and cosines. When you have calculated the Eulerian angles, a further useful exercise would be to prepare a drawing showing the orientation of the primed axes with respect to the unprimed axes.

$$\begin{pmatrix} +0.075\,284\,882\,7 & -0.518\,674\,468\,2 & +0.851\,650\,739\,6 \\ -0.553\,110\,473\,2 & -0.732\,363\,000\,8 & +0.397\,131\,261\,9 \\ -0.829\,699\,337\,5 & +0.442\,158\,963\,2 & -0.342\,020\,143\,3 \end{pmatrix} \tag{3.7.1}$$

Note, as a matter of good computational practice, that the numbers are written in groups of three separated by half spaces after the decimal point, all numbers, positive and negative, are signed, and leading zeroes are not omitted.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

3.8: Trigonometrical Formulas

I gather here merely for reference a set of commonly-used trigonometric formulas. It is a matter of personal preference whether to commit them to memory. It is probably fair to remark that anyone who is regularly engaged in problems in celestial mechanics or related disciplines will be familiar with most of them, at least from frequent use, whether or not any conscious effort was made to memorize them. At the very least, the reader should be aware of their existence, even if he or she has to look to recall the exact formula.

$$\frac{\sin A}{\cos A} = \tan A \tag{3.8.1}$$

$$\sin^2 A + \cos^2 A = 1 \tag{3.8.2}$$

$$1 + \cot^2 A = \csc^2 A \tag{3.8.3}$$

$$1 + \tan^2 A = \sec^2 A \tag{3.8.4}$$

$$\sec A \csc A = \tan A + \cot A \tag{3.8.5}$$

$$\sec^2 A \csc^2 A = \sec^2 A + \csc^2 A \tag{3.8.6}$$

$$\sin(A \pm B) = \sin A \cos B \pm \cos A \sin B \tag{3.8.7}$$

$$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B \tag{3.8.8}$$

$$\tan(A \pm B) = \frac{\tan A \pm \tan B}{1 \mp \tan A \tan B} \tag{3.8.9}$$

$$\sin 2A = 2 \sin A \cos A \tag{3.8.10}$$

$$\cos 2A = \cos^2 A - \sin^2 A = 2\cos^2 A - 1 = 1 - 2\sin^2 A \tag{3.8.11}$$

$$\tan 2A = \frac{2 \tan A}{1 - \tan^2 A} \tag{3.8.12}$$

$$\sin \frac{1}{2} A = \sqrt{\frac{1 - \cos A}{2}} \tag{3.8.13}$$

$$\cos \frac{1}{2} A = \sqrt{\frac{1 + \cos A}{2}} \tag{3.8.14}$$

$$\tan \frac{1}{2} A = \sqrt{\frac{1 - \cos A}{1 + \cos A}} = \frac{1 - \cos A}{\sin A} = \frac{\sin A}{A + \cos A} = \csc A - \cot A \tag{3.8.15}$$

$$\sin A + \sin B = 2 \sin \frac{1}{2} S \cos \frac{1}{2} D, \tag{3.8.16}$$

where

$$S = A + B \quad \text{and} \quad D = A - B \tag{3.8.17}$$

$$\sin A - \sin B = 2 \cos \frac{1}{2} S \sin \frac{1}{2} D \tag{3.8.18}$$

$$\cos A + \cos B = 2 \cos \frac{1}{2} S \cos \frac{1}{2} D \tag{3.8.19}$$

$$\cos A - \cos B = -2 \sin \frac{1}{2} S \sin \frac{1}{2} D \tag{3.8.20}$$

$$\sin A \sin B = \frac{1}{2} (\cos D - \cos S) \tag{3.8.21}$$

$$\cos A \cos B = \frac{1}{2} (\cos D + \cos S) \tag{3.8.22}$$

$$\sin A\cos B = \frac{1}{2}(\sin S + \sin D) \tag{3.8.23}$$

$$\sin A = \frac{T}{\sqrt{1+T^2}} = \frac{2T}{1+t^2}, \tag{3.8.24}$$

where

$$T = \tan A \text{ and } t = \tan\frac{1}{2}A \tag{3.8.25}$$

$$\cos A = \frac{1}{\sqrt{1+T^2}} = \frac{1-t^2}{1+t^2} \tag{3.8.26}$$

$$\tan A = T = \frac{2t}{1-t^2} \tag{3.8.27}$$

$$s = \sin A, \quad c = \cos A \tag{3.8.28}$$

$$\begin{array}{ll}
\cos A = c & \sin A = s \\
\cos 2A = 2c^2 - 1 & \sin 2A = 2cs \\
\cos 3A = 4c^3 - 3c & \sin 3A = 3s - 4s^3 \\
\cos 4A = 8c^4 - 8c^2 + 1 & \sin 4A = 4c(s - 2s^3) \\
\cos 5A = 16c^5 - 20c^3 + 5c & \sin 5A = 5s - 20s^3 + 16s^5 \\
\cos 6A = 32c^6 - 48c^4 + 18c^2 - 1 & \sin 6A = 2c(3s - 16s^3 + 16s^5) \\
\cos 7A = 64c^7 - 112c^5 + 56c^3 - 7c & \sin 7A = 7s - 56s^3 + 112s^5 - 64s^7 \\
\cos 8A = 128c^8 - 256c^6 + 160c^4 - 32c^2 + 1 & \sin 8A = 8c(s - 10s^3 + 24s^5 - 16s^7)
\end{array} \tag{3.8.29}$$

$$\begin{array}{l}
\cos^2 A = \frac{1}{2}(\cos 2A + 1) \\
\cos^3 A = \frac{1}{4}(\cos 3A + 3\cos A) \\
\cos^4 A = \frac{1}{8}(\cos 4A + 4\cos 2A + 3) \\
\cos^5 A = \frac{1}{16}(\cos 5A + 5\cos 3A + 10\cos A) \\
\cos^6 A = \frac{1}{32}(\cos 6A + 6\cos 4A + 15\cos 2A + 10) \\
\cos^7 A = \frac{1}{64}(\cos 7A + 7\cos 5A + 21\cos 3A + 35\cos A) \\
\cos^8 A = \frac{1}{128}(\cos 8A + 8\cos 6A + 28\cos 4A + 56\cos 2A + 35)
\end{array} \tag{3.8.30}$$

$$\begin{array}{l}
\sin^2 A = \frac{1}{2}(1 - \cos 2A) \\
\sin^3 A = \frac{1}{4}(3\sin A - \sin 3A) \\
\sin^4 A = \frac{1}{8}(\cos 4A - 4\cos 2A + 3) \\
\sin^5 A = \frac{1}{16}(\sin 5A - 5\sin 3A + 10\sin A) \\
\sin^6 A = \frac{1}{32}(10 - 15\cos 2A + 6\cos 4A - \cos 6A) \\
\sin^7 A = \frac{1}{64}(35\sin A - 21\sin 3A + 7\sin 5A - \sin 7A) \\
\sin^8 A = \frac{1}{128}(\cos 8A - 8\cos 6A + 28\cos 4A - 56\cos 2A + 35)
\end{array} \tag{3.8.31}$$

$$\sin A = A - \frac{A^3}{3!} + \frac{A^5}{5!} - \ldots \tag{3.8.32}$$

$$\cos A = 1 - \frac{A^2}{2!} + \frac{A^4}{4!} - \ldots \tag{3.8.33}$$

$$\int_0^{\pi/2} \sin^m\theta\cos^n\theta d\theta = \frac{(m-1)!!(n-1)!!X}{(m+n)!!}, \quad \text{where } X = \pi/2 \text{ if } m \text{ and } n \text{ are both even, and } X = 1 \text{ otherwise.} \tag{3.8.34}$$

$e^{ni\theta} = e^{in\theta}$ (de Moivre's theorem - the only one you need know. All others can be deduced from it.)

Plane triangles:

$$\frac{a}{\sin A}=\frac{b}{\sin B}=\frac{c}{\sin C} \tag{3.8.35}$$

$$a^2=b^2+c^2-2bc\cos A \tag{3.8.36}$$

$$a\cos B+b\cos A=c \tag{3.8.37}$$

$$s=\frac{1}{2}(a+b+c) \tag{3.8.38}$$

$$\sin\frac{1}{2}A=\sqrt{\frac{(s-b)(s-c)}{s(s-a)}} \tag{3.8.39}$$

$$\cos\frac{1}{2}A=\sqrt{\frac{s(s-a)}{bc}} \tag{3.8.40}$$

$$\tan\frac{1}{2}A=\sqrt{\frac{(s-b)(s-c)}{s(s-a)}} \tag{3.8.41}$$

Spherical triangles

$$\frac{\sin a}{\sin A}=\frac{\sin b}{\sin B}=\frac{\sin c}{\sin C} \tag{3.8.42}$$

$$\cos a=\cos b\cos c+\sin b\sin c\cos A \tag{3.8.43}$$

$$\cos A=-\cos B\cos C+\sin B\sin C\cos a \tag{3.8.44}$$

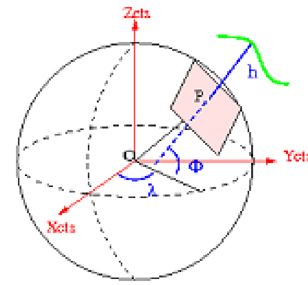
$$\cos(\mathrm{IS})\cos(\mathrm{IA})=\sin(\mathrm{IS})\cot(\mathrm{OS})-\sin(\mathrm{IA})\cot(\mathrm{OA}) \tag{3.8.45}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

4: COORDINATE GEOMETRY IN THREE DIMENSIONS



4.1: Introduction

Various geometrical figures in three-dimensional space can be described relative to a set of mutually orthogonal axes Ox, Oy, Oz, and a point can be represented by a set of rectangular coordinates (x, y, z). The point can also be represented by cylindrical coordinates (ρ, ϕ, z) or spherical coordinates (r, θ, ϕ), which were described in Chapter 3. In this chapter, we are concerned mostly with (x, y, z). The rectangular axes are usually chosen so that when you look down the z-axis towards the xy-plane, the y-axis is 90° counterclockwise from the x-axis. Such a set is called a right-handed set. A left-handed set is possible, and may be useful under some circumstances, but, unless stated otherwise, it is assumed that the axes chosen in this chapter are right-handed.

An Equation connecting x, y and z, such as

$$f(x, y, z) = 0 \tag{4.1.1}$$

or

$$z = z(x, y) \tag{4.1.2}$$

describes a two-dimensional surface in three-dimensional space. A line (which need be neither straight nor two-dimensional) can be described as the intersection of two surfaces, and hence a line or curve in three-dimensional coordinate geometry is described by two Equations, such as

$$f(x, y, z) = 0 \tag{4.1.3}$$

and

$$g(x, y, z) = 0. \tag{4.1.4}$$

In two-dimensional geometry, a single Equation describes some sort of a plane curve. For example,

$$y^2 = 4qx \tag{4.1.5}$$

describes a parabola. But a plane curve can also be described in parametric form by two Equations. Thus, a parabola can also be described by

$$x = qt^2 \tag{4.1.6}$$

and

$$y = 2qt \tag{4.1.7}$$

Similarly, in three-dimensional geometry, a line or curve can be described by three Equations in parametric form. For example, the three Equations

$$x = a\cos t \tag{4.1.8}$$

$$y = a\sin t \tag{4.1.9}$$

$$z = ct \tag{4.1.10}$$

describe a curve in three-space. Think of the parameter t as time, and see if you can imagine what sort of a curve this is.

We shall be concerned in this chapter mainly with six types of surface: the plane, the ellipsoid, the paraboloid, the hyperboloid, the cylinder and the cone.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.2: Planes and Straight Lines

The geometry of the plane and the straight line is, of course, rather simple, so that we can dispose of them in this brief introductory section in a mere 57 Equations.

The Equation

$$Ax + By + Cz + D = 0 \tag{4.2.1}$$

represents a plane. If $D \neq 0$ it is often convenient, and saves algebra and computation with no loss of information, to divide the Equation through by D and re-write it in the form

$$ax + by + cz = 1. \tag{4.2.2}$$

The coefficients need not by any means all be positive. If $D = 0$, the plane passes through the origin of coordinates, and it may be convenient to divide the Equation 4.2.1 by C and hence to rewrite it in the form

$$ax + by + z = 0. \tag{4.2.3}$$

The plane represented by Equation 4.2.2 intersects the yz-, zx- and xy-planes in the straight lines

$$by + cz = 1 \tag{4.2.4}$$

$$cz + ax = 1 \tag{4.2.5}$$

$$ax + by = 1 \tag{4.2.6}$$

and it intersects the x-, y- and z-axes at

$$x = x_0 = 1/a \tag{4.2.7}$$

$$y = y_0 = 1/b \tag{4.2.8}$$

$$z = z_0 = 1/c \tag{4.2.9}$$

The geometry can be seen in figure IV.1

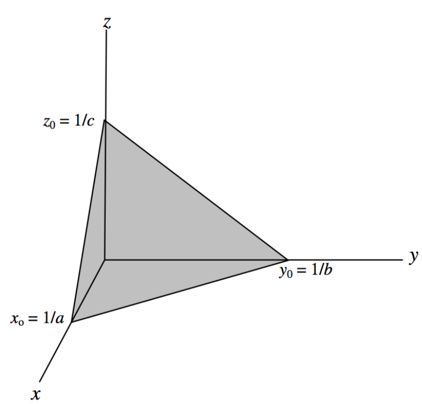


FIGURE IV.1

Another way of writing the Equation to the plane would be

$$\frac{x}{x_0} + \frac{y}{y_0} + \frac{z}{z_0} = 1. \tag{4.2.10}$$

In this form, x_0, y_0 and z_0 are the intercepts on the x-, y- and z-axes.

Distance of a point from the plane

We now consider this problem. Let $\mathrm{P}_1\ (x_1, y_1, z_1)$ be some point in space. What is the perpendicular distance from P_1 to the plane $ax + by + cz = \frac{x}{x_0} + \frac{y}{y_0} + \frac{z}{z_0} = 1$?

[The algebra in the following paragraphs may seem a little heavy. If all you are interested in is the distance of the plane from the origin, simply substitute $x_1 = y_1 = z_1 = 0$, and the algebra will be considerably eased.]

Let $\mathrm{P}(x, y, z)$ be a point on the plane. The distance s between P_1 and P is given by

$$s^2 = (x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2 \tag{4.2.11}$$

But since (x, y, z) is on the plane, we can write s^2 in terms of x and y alone, by substituting for z from Equation 4.2.2:

$$s^2 = (x - x_1)^2 + (y - y_1)^2 + \left(\frac{1 - ax - by}{c} - z_1\right)^2 \tag{4.2.12}$$

This distance (from P to P_1) is least for a point on the plane such that $\frac{\partial s^2}{\partial x}$ and $\frac{\partial s^2}{\partial y}$ are both zero. These two conditions result in

$$(a^2 + c^2)x = a + c^2 x_1 - acz_1 - aby \tag{4.2.13}$$

$$(b^2 + c^2)y = b + c^2 y_1 - bcz_1 - abx \tag{4.2.14}$$

These, combined with Equation 4.2.2, result in

$$x = \frac{(b^2 + c^2)x_1 + a(1 - by_1 - cz_1)}{a^2 + b^2 + c^2} \tag{4.2.15}$$

$$y = \frac{(c^2 + a^2)y_1 + b(1 - cz_1 - ax_1)}{a^2 + b^2 + c^2} \tag{4.2.16}$$

$$z = \frac{(a^2 + b^2)z_1 + c(1 - ax_1 - by_1)}{a^2 + b^2 + c^2} \tag{4.2.17}$$

These are the coordinates of the point P in the plane that is nearest to P_1. The perpendicular distance between P and P_1 is

$$p = \frac{1 - ax_1 - by_1 - cz_1}{\sqrt{a^2 + b^2 + c^2}} \tag{4.2.18}$$

This is positive if P_1 is on the same side of the plane as the origin, and negative if it is on the opposite side. If the perpendicular distances of two points from the plane, as calculated from Equation 4.4.18, are of opposite signs, they are on opposite sides of the plane. If $p = 0$, or indeed if the numerator of Equation 4.4.18 is zero, the point $\mathrm{P}_1(x_1, y_1, z_1)$ is, of course, in the plane.

It is worthwhile to repeat these results for the case where the point P_1 coincides with the origin O. In that case we find that the coordinates of the point P on the plane that is nearest to the origin are

$$x = \frac{a}{a^2 + b^2 + c^2}, \quad y = \frac{b}{a^2 + b^2 + c^2}, \quad z = \frac{c}{a^2 + b^2 + c^2}, \tag{4.2.19a,b,c}$$

and the perpendicular distance from the origin to the plane (i.e. from O to P) is

$$p = \frac{1}{\sqrt{a^2 + b^2 + c^2}} \tag{4.2.20}$$

Further, OP is normal to the plane, and the *direction cosines* (see Chapter 3, especially section 3.3) of OP, i.e. of the normal to the plane, are

$$\frac{a}{\sqrt{a^2 + b^2 + c^2}}, \quad \frac{b}{\sqrt{a^2 + b^2 + c^2}}, \quad \frac{c}{\sqrt{a^2 + b^2 + c^2}} \tag{4.2.21}$$

The coefficients a, b, c are direction ratios of the normal to the plane; that is to say, they are numbers that are proportional to the direction cosines.

Example: Consider the plane

$$0.5x + 0.25y + 0.20z = 1 \tag{4.2.22}$$

The plane intersects the x-, y- and z-axes at $(2,0,0)$, $0,4,0)$ and $(0,0,5)$. The point on the plane that is closest to the origin is $(1.4184, 0.7092, 0.5674)$The perpendicular distance of the origin from the plane is 1.6843. The direction cosines of the normal to the plane are $(0.8422, 0.4211, 0.3369)$

An Equation for the plane containing three specified points can be found as follows. Let (x_1, y_1), (x_2, y_2), (x_3, y_3) be the three specified points, and let (x, y) be any point in the plane that contains these three points. Each of these points must satisfy an Equation of the form 4.2.1. That is,

$$xA + yB + zC + D = 0 \tag{4.2.24}$$

$$x_1A + y_1B + z_1C + D = 0 \tag{4.2.25}$$

$$x_2A + y_2B + z_2C + D = 0 \tag{4.2.26}$$

$$x_3A + y_3B + z_3C + D = 0 \tag{4.2.27}$$

In these Equations, we are treating A, B, C, D as unknowns, and the x, y, z, x_1, y_1... as coefficients. We have four linear Equations in four unknowns, and no constant term. From the theory of Equations, these are consistent only if each is a linear combination of the other three. This is satisfied only if the determinant of the coefficients is zero:

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0 \tag{4.2.28}$$

and this is the Equation to the required plane containing the three points. The reader will notice the similarity of this Equation to Equation 2.2.4 for a line passing between two points in two-dimensional geometry. The reader might like to repeat the argument, but requiring instead the four points to satisfy an Equation of the form 4.2.2. There will then be four linear Equations in three unknowns. Otherwise the argument is the same.

We now move on to the question of finding the area of a triangle whose vertices are given. It is straightforward to do this with a numerical example, and the reader is now encouraged to write a computer program, in whatever language is most familiar, to carry out the following tasks. Read as data the x-y-z coordinates of three points A, B, C. Calculate the lengths of the sides a, b, c, a being opposite to A, etc. Calculate the three angles at the vertices of the triangle, in degrees and minutes, and check for correctness by verifying that their sum is 180°. If an angle is obtuse, make sure that the computer displays its value as a positive angle between 90° and 180°. Finally, calculate the area of the triangle.

The data for several triangles could be written into a data file, which your program reads, and then writes the answers into an output file. Alternatively, you can type the coordinates of the vertices of one triangle and ask the computer to read the data from the monitor screen, and then to write the answers on the screen followed by a message such as "Do you want to try another triangle (1) or quit (2)?". Your program should also be arranged so that it writes an appropriate message if the three points happen to be collinear.

It should be easy to calculate the sides. The angles can then be calculated from Equation 3.2.2 and the area from each of the four Equations 3.2.3 and 3.2.4. They should all yield the correct answer, of course, but the redundant calculations serve as an important check on the correctness of your programming, as also does your check that the three angles add to 180°. Where there are two of more ways of performing a calculation, a careful calculator will do all of them as a check against mistakes, whether the calculation is done by hand or by computer.

Example. If the coordinates of the vertices are

$$\mathrm{A}(7,4,3), \quad \mathrm{B}(11,6,2), \quad \mathrm{C}(9,2,4) \tag{4.2.1}$$

the sides are

$$a = 4.899, \quad b = 3.000, \quad c = 4.583, \tag{4.2.2}$$

and the angles are

$$A = 65^\circ 55', \quad B = 36^\circ 42', C = 77^\circ 23', \tag{4.2.3}$$

which add up to 180°. The area is 6.708.

Example. If the coordinates of the vertices are

$$\mathrm{A}(6,4,9), \quad \mathrm{B}(2,6,17), \quad \mathrm{C}(8,3,5) \tag{4.2.4}$$

the area of the triangle is zero and the points are collinear.

The foregoing showed that it was not difficult to calculate numerically the area of the triangle from the coordinates of its vertices. Is it easy to find a simple explicit algebraic formula for the area in terms of (x_1, y_1, z_1), (x_2, y_2, z_2) and (x_3, y_3, z_3)? On referring to figure $\mathrm{IV}.2$, we can proceed as follows.

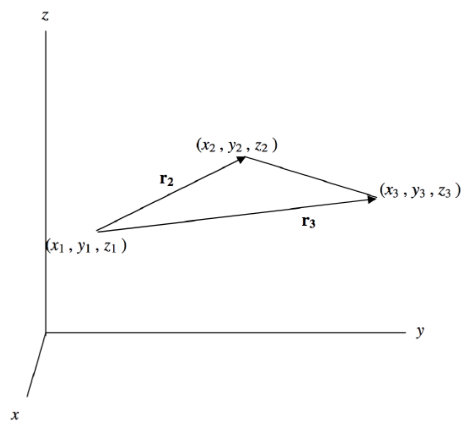


FIGURE IV.2

The vectors $\mathbf{r}_2$ and $\mathbf{r}_3$ can be written

$$\mathbf{r}_2 = (x_2 - x_1)\mathbf{i} + (y_2 - y_1)\mathbf{j} + (z_2 - z_1)\mathbf{k} \tag{4.2.29}$$

$$\mathbf{r}_3 = (x_3 - x_1)\mathbf{i} + (y_3 - y_1)\mathbf{j} + (z_3 - z_1)\mathbf{k} \tag{4.2.30}$$

where $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are unit vectors parallel to the x-, y- and z- axes.

The cross product of $\mathbf{r}_2$ and $\mathbf{r}_3$ gives the (vector) area of the parallelogram of which they form two sides. The area $\mathbf{A}$ of the triangle is half of this, so that

$$2\mathbf{A} = \mathbf{r}_2 \times \mathbf{r}_3 \tag{4.2.5}$$

$$= [(y_2 - y_1)(z_3 - z_1) - (y_3 - y_1)(z_2 - z_1)]\mathbf{i} \tag{4.2.6}$$

$$+[(z_2 - z_1)(x_3 - x_1) - (z_3 - z_1)(x_2 - x_1)]\mathbf{j} \tag{4.2.7}$$

$$+[(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)]\mathbf{k} \tag{4.2.31}$$

The magnitude of this vector can be found in the usual way, to obtain

$$4A^2 = [(y_1(z_2 - z_3) + y_2(z_3 - z_1) + y_3(z_1 - z_2)]^2 \tag{4.2.8}$$

$$+[(z_1(x_2 - x_3) + z_2(x_3 - x_1) + z_3(x_1 - x_2)]^2 \tag{4.2.9}$$

$$+[(x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2)]^2. \tag{4.2.32}$$

The reader can verify that, if $z_1 = z_2 = z_3$, this reduces to Equation 2.2.12 for the area of a triangle the xy-plane. Equation 4.1.32 can also be written

$$4A^2 = \begin{vmatrix} y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \\ 1 & 1 & 1 \end{vmatrix}^2 + \begin{vmatrix} z_1 & z_2 & z_3 \\ x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{vmatrix}^2 + \begin{vmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{vmatrix}^2 \tag{4.2.33}$$

This gives the area explicitly in terms of the coordinates of the vertices. If it is zero, the points are collinear.

The volume of a tetrahedron is $\frac{1}{6} \times \text{base} \times \text{height}$. By combining Equation 4.2.33 for the area of a triangle with Equation 4.2.14 for the perpendicular distance of a point from a plane, we can determine that the volume of the tetrahedron whose vertices are

$$(x_0, y_0, z_0),\ (x_1, y_1, z_1),\ (x_2, y_2, z_2),\ (x_3, y_3, z_3) \tag{4.2.10}$$

is

$$\frac{1}{6}\begin{vmatrix} x_0 & y_0 & z_0 & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} \tag{4.2.34}$$

If this determinant is zero, the four points are coplanar.

In three-dimensional coordinate geometry, a straight line is described by two Equations, being the intersection of two planes:

$$a_1x + b_1y + c_1z = 1 \tag{4.2.35}$$

$$a_2x + b_2y + c_2z = 1 \tag{4.2.36}$$

If $a_1/a_2 = b_1/b_2 = c_1/c_2$,the normals to the two planes have the same direction ratios, so the planes are parallel and do not intersect. Otherwise the normals to the two planes have different direction ratios (a_1, b_1, c_1), (a_2, b_2, c_2), and, since the line of intersection of the planes is at right angles to both normals, the direction ratios of the line are found from the cross product of vectors normal to the planes. The direction ratios of the line of intersection are therefore

$$(b_1c_2 - b_2c_1,\ c_1a_2 - c_2a_1,\ a_1b_2 - a_2b_1) \tag{4.2.37}$$

The line crosses the yz-, zx- and xy- planes at

$$y = \frac{c_2 - c_1}{b_1c_2 - b_2c_1} \quad z = \frac{b_1 - b_2}{b_1c_2 - b_2c_1} \tag{4.2.38}$$

$$z = \frac{a_2 - a_1}{c_1a_2 - c_2a_1} \quad x = \frac{c_1 - c_2}{c_1a_2 - c_2a_1} \tag{4.2.39}$$

$$x = \frac{b_2 - b_1}{a_1b_2 - a_2b_1} \quad y = \frac{a_1 - a_2}{a_1b_2 - a_2b_1} \tag{4.2.40}$$

An example of computing a straight line from the intersection of two planes occurs in meteor astronomy. We can assume a flat Earth, which is tantamount to supposing that the height of a meteor is negligible compared with the radius of Earth, and the height of an observer above sea level is negligible compared with the height of the meteor. Since the heights of meteors are typically a few tens of km, both of these approximations are reasonable, at least for noninstrumental eyewitness accounts.

We suppose that, relative to an arbitrary origin O on the surface of Earth, a witness A is 15 km east and 5 km north of the origin. He sees a fireball start at an angle $\theta = 25^\circ.5$ from his zenith and at an azimuth $\phi = 54^\circ.5$ counterclockwise from his east, and it finishes at $\theta = 36^\circ.7$, $\phi = 16^\circ.7$. (See figure IV.3.)

Show that the plane containing the witness and the meteor is

$$0.0363x + 0.0911y - 0.0454z = 1 \tag{4.2.41}$$

A second witness, 30 km east and 15 km north of O, estimates the zenith distance and azimuth of two points on the meteor track to be $\theta = 29^\circ.6$, $\phi = 202^\circ.9$ and $\theta = 33^\circ.6$, $\phi = 242^\circ.9$

Show that the plane containing this second witness and the meteor is

$$0.0257x + 0.0153y + 0.0168z = 1 \tag{4.2.42}$$

These two Equations describe the path of the fireball through the air. Show that, if the meteoroid carries on moving in a straight line, it will strike the ground as a meteorite 42.4 km east and 6.0 km south of the origin O.

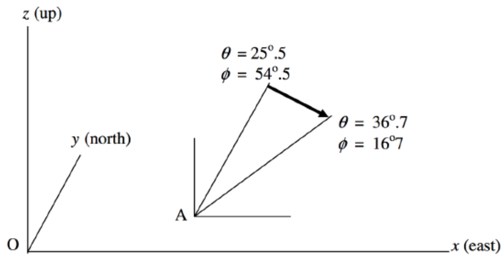


FIGURE IV.3

As we have just discussed, two nonparallel planes intersect in a straight line. Usually, three nonparallel planes intersect at a single unique point; for, if L is a line formed from the intersection of planes P_1 and P_2, L will usually intersect the plane P_3 at a point.

Example: The planes

$$2x + 3y + 4z - 9 = 0 \tag{4.2.43}$$

$$x + y - 8z + 6 = 0 \tag{4.2.44}$$

$$5x + 6y - 12z + 1 = 0 \tag{4.2.45}$$

intersect at $(1, 1, 1)$.

It will be recalled from the theory of linear Equations that three Equations

$$A_1x + B_1y + C_1z + D_1 = 0 \tag{4.2.46}$$

$$A_2x + B_2y + C_2z + D_2 = 0 \tag{4.2.47}$$

$$A_3x + B_3y + C_3z + D_3 = 0 \tag{4.2.48}$$

have a unique solution only if

$$\Delta = \begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_2 & C_3 \end{vmatrix} \neq 0 \tag{4.2.49}$$

and, in the geometrical interpretation, this is the condition that three planes meet in a single point. Consider, however, the three planes

$$2x + 3y + 4z - 9 = 0 \tag{4.2.50}$$

$$x + y - 8z + 6 = 0 \tag{4.2.51}$$

$$5x + 6y - 20z + 12 = 0 \tag{4.2.52}$$

The direction ratios of the three lines found by combining the planes in pairs (see Equation $4.2.37$) are

$$(-28, 20, -1) \quad (-84, 60, -3) \quad (28, -20, 1) \tag{4.2.11}$$

It will be observed that each is a multiple of either of the others, and the direction cosines of each of the three lines are identical apart from sign: $(\mp 0.813, \pm 0.581, \mp 0.029)$

The three lines are, in fact, parallel, and the three planes enclose a prism. A condition for this is that

$$\Delta = 0. \tag{4.2.53}$$

But consider now the planes

$$2x + 3y + 4z - 9 = 0 \tag{4.2.54}$$

$$x + y - 8z + 6 = 0 \tag{4.2.55}$$

$$5x + 6y - 20z + 9 = 0 \tag{4.2.56}$$

Not only does $\Delta = 0$, but also

$$\Delta' = \begin{vmatrix} A_1 & B_1 & D_1 \\ A_2 & B_2 & D_2 \\ A_3 & B_3 & D_3 \end{vmatrix} = 0 \tag{4.2.57}$$

The three lines obtained by combining Equations 4.2.54,55,56 in pairs are in fact identical, and the three planes meet in a single line. Each of Equations 4.2.54,55,56 is a linear combination of the other two.

In summary, three nonparallel planes meet in a single line if $\Delta \neq 0$. They meet in a single point if $\Delta = \Delta' = 0$. They enclose a prism if $\Delta = 0, \Delta' \neq 0$.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.3: The Ellipsoid

Consider the Equation

$$\frac{x^2}{a^2}+\frac{z^2}{c^2}=1, \tag{4.3.1}$$

with $a>c$, in the xz-plane. The length of the semi major axis is a and the length of the semi minor axis is c. If this figure is rotated through 360° about its minor (z-) axis, the three- dimensional figure so obtained is called an *oblate spheroid*. The figure of the Earth is not exactly spherical; it approximates to a very slightly oblate spheroid, the ellipticity $(c-a)/a$ being only 0.00335. (The actual figure of the Earth, mean sea level, is often referred to as the *geoid*.)

The Equation to the oblate spheroid referred to above is

$$\frac{x^2}{a^2}+\frac{y^2}{a^2}+\frac{z^2}{c^2}=1. \tag{4.3.2}$$

If the ellipse 4.3.1 is rotated through 360° about its major (x-) axis, the figure so obtained is called a *prolate* spheroid. A rugby football (or, to a lesser extent, a North American football, which is a bit too pointed) is a good approximation to a prolate spheroid.

The Equation to the prolate spheroid just described is

$$\frac{x^2}{a^2}+\frac{y^2}{c^2}+\frac{z^2}{c^2}=1. \tag{4.3.3}$$

Either type of spheroid can be referred to as an "ellipsoid of revolution".

The figure described by the Equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1 \tag{4.3.4}$$

is a *tri-axial ellipsoid*. Unless stated otherwise, I shall adopt the convention $a>b>c$, and choose the coordinate axes such that the major, intermediate and minor axes are along the x-, y- and z-axes respectively. A tri-axial ellipsoid is not an ellipsoid of revolution; it cannot be obtained by rotating an ellipse about an axis.

The special case $a=b=c$:

$$x^2+y^2+z^2=a^2 \tag{4.3.5}$$

is, of course, a sphere.

Figure $\text{IV}.4$ shows the cross-section of a tri-axial ellipse in the yz- plane (a), the xz-plane (b) and (twice - (c), (d)) the xy-plane. If you imagine your eye wandering in the xz-plane from the x-axis (a) to the z-axis (c), you will be convinced that there is a direction in the xz-plane from which the

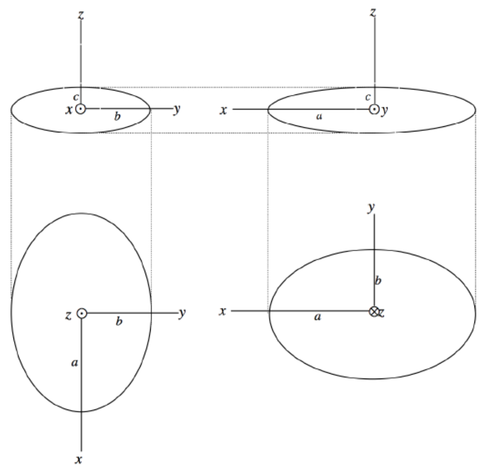


FIGURE IV.4

cross-section of the ellipse is a circle. There are actually two such directions, symmetrically situated on either side of the z-axis, but there are no such directions in either the xy- or the yz-planes from which the cross-section of the ellipsoid appears as a circle. Expressed otherwise, there are two planes that intersect the ellipsoid in a circle. This fact is of some importance in the description of the propagation of light in a bi-axial crystal, in which one of the wavefronts is a tri-axial ellipsoid.

Let us refer the ellipsoid 4.3.4 to a set of axes $\mathrm{O}x'y'z'$ such that the angles $z'\mathrm{O}z$ and $x'\mathrm{O}x$ are each θ, and the y'- and y-axes are identical. The Equation of the ellipsoid referred to the new axes is (by making use of the usual formulas for the rotation of axes)

$$\frac{(z'\sin\theta+x'\cos\theta)^2}{a^2}+\frac{y'^2}{b^2}+\frac{(z'\cos\theta-x'\sin\theta)^2}{c^2}=1. \tag{4.3.6}$$

The cross-section of the ellipsoid in the $x'y'$-plane (i.e. normal to the z'-axis) is found by putting $z'=0$:

$$\frac{(x'\cos\theta)^2}{a^2}+\frac{y'^2}{b^2}+\frac{(x'\sin\theta)^2}{c^2}=1. \tag{4.3.7}$$

This is a circle if the coefficients of x' and y' are equal. Thus it is a circle if

$$\cos^2\theta=\frac{a^2(b^2-c^2)}{b^2(a^2-c^2)}. \tag{4.3.8}$$

Thus, a plane whose normal is in the xz-plane (i.e. between the major and minor axis) and inclined at an angle θ to the minor (z-) axis, cuts the tri-axial ellipsoid in a circle. As viewed from either of these directions, the cross-section of the ellipsoid is a circle of radius b.

As an asteroid tumbles over and over, its brightness varies, for several reasons, such as its changing phase angle, the directional reflective properties of its regolith, and, of course, the cross-sectional area presented to the observer. The number of factors that affect the light-curve of a rotating asteroid is, in fact, so large that it is doubtful if it is possible, from the light-curve alone, to deduce with much credibility or accuracy the true shape of the asteroid. However, it is obviously of some interest for a start in any such investigation to be able to calculate the cross-sectional area of the ellipsoid 4.3.3 as seen from some direction (θ,ϕ).

Let us erect a set of coordinate axes $\mathrm{O}x'y'z'$ such that $\mathrm{O}z'$ is in the direction (θ,ϕ), first by a rotation through ϕ about $\mathrm{O}z$ to form intermediate axes $\mathrm{O}x_1y_1z_1$, followed by a rotation through θ about $\mathrm{O}y_1$. The (x',y',z') coordinates are related to the (x,y,z) coordinates by

$$\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}\cos\phi & -\sin\phi & 0\\ \sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{pmatrix}\begin{pmatrix}x'\\y'\\z'\end{pmatrix} \tag{4.3.9}$$

If we substitute for $x,\ y,\ z$ in Equation 4.3.4 from Equation 4.3.9, we obtain the Equation to the ellipsoid referred to the $\mathrm{O}x'y'z'$ coordinate systems. And if we put $z'=0$, we see the elliptical crosssection of the ellipsoid in the plane normal to $\mathrm{O}z'$. This will be of the form

$$Ax'^2+2Hx'y'+By'^2=1, \tag{4.3.10}$$

where

$$A=\cos^2\theta\left(\frac{\cos^2\phi}{a^2}+\frac{\sin^2\phi}{b^2}\right)+\frac{\sin^2\theta}{c^2} \tag{4.3.11}$$

$$2H=2\cos^2\theta\sin\phi\cos\phi\left(\frac{1}{b^2}-\frac{1}{a^2}\right), \tag{4.3.12}$$

$$B=\frac{\sin^2\phi}{a^2}+\frac{\cos^2\phi}{b^2}. \tag{4.3.13}$$

This is an ellipse whose axes are inclined at an angle ψ from $\mathrm{O}x'$ given by

$$\tan 2\psi=\frac{2H}{A-B}. \tag{4.3.14}$$

By replacing x' and y' by x'' and y'', where

$$\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{pmatrix}\cos\psi & -\sin\psi\\ \sin\psi & \cos\psi\end{pmatrix}\begin{pmatrix}x''\\y''\end{pmatrix} \tag{4.3.15}$$

we shall be able to describe the ellipse in a coordinate system $\mathrm{O}x''y''$ whose axes are along the axes of the ellipse, and the Equation will be of the form

$$\frac{x''^2}{a''^2}+\frac{y''^2}{b''^2}=1 \tag{4.3.16}$$

and the area of the cross-section is $\pi a''b''$.

For example, suppose the semi axes of the ellipsoid are $a=3,\ b=2,\ y=1$, and we look at it from the direction $\theta=60^\circ$, $\phi=45^\circ$. Following Equations 4.4.9,10,11,12, we obtain for the Equation of the elliptical cross-section referred to the system $\mathrm{O}x'y'z'$

$$0.79513\dot{8}x'^2+0.069\dot{4}x'y'+0.180\dot{5}y'^2=1. \tag{4.3.17}$$

From Equation 4.4.13 we find $\psi=3^\circ$ .22338. Equation 4.4.14 then transforms Equation 4.4.16 to

$$0.797094x''^2+0.178600y''^2=1 \tag{4.3.18}$$

or

$$\frac{x''^2}{(1.1201)^2}+\frac{y''^2}{(2.3662)^2}=1. \tag{4.3.19}$$

The area is

$$\pi\times 1.1201\times 2.3662=8.362. \tag{4.3.1}$$

It is suggested here that the reader could write a computer program in the language of his or her choice for calculating the cross-sectional area of an ellipsoid as seen from any direction. As an example, I reproduce below a Fortran program for an ellipse with $(a,b,c)=(3,2,1)$. It is by no means the fastest and most efficient Fortran program that could be written, but is sufficiently straightforward that anyone familiar with Fortran and probably many who are not should be able to follow the steps.

```
A=3.
B=2.
C=1.
```

```
A2=A*A
B2=B*B
C2=C*C
READ(5,*)TH,PH
TH=TH/57.29578
PH=PH/57.29578
STH=SIN(TH)
CTH=COS(TH)
SPH=SIN(PH)
CPH=COS(PH)
STH2=STH*STH
CTH2=CTH*CTH
SPH2=SPH*SPH
CPH2=CPH*CPH
AA=CTH2*(CPH2/A2+SPH2/B2)+STH2/C2
TWOHH=2.*CTH*STH*CPH*(1./B2−1./A2)
BB=SPH2/A2+CPH2/B2
PS=.5*ATAN2(TWOHH,AA−BB)
SPS=SIN(PS)
CPS=COS(PS)
AAA=CPS*(AA*CPS+TWOHH*SPS)+BB*SPS*SPS
BBB=SPS*(AA*SPS−TWOHH*CPS)+BB*CPS*CPS
SEMAX1=1./SQRT(AAA)
SEMAX2=1./SQRT(BBB)
AREA=3.1415927*SEMAX1*SEMAX2
WRITE(6,1)AREA
1 FORMAT(' Area = ',F7.3)
STOP
END
```

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.4: The Paraboloid

The Equation $x^2 = 4qz = 2lz$ is a parabola in the xz-plane. The distance between vertex and focus is q, and the length of the semi latus rectum $l = 2q$. The Equation can also be written

$$\frac{x^2}{a^2} = \frac{z}{h} \tag{4.4.1}$$

Here a and h are distances such that $x = a$ when $z = h$, and the length of the semi latus rectum is $l = a^2/(2h)$.

If this parabola is rotated through 360° about the z-axis, the figure swept out is a *paraboloid of revolution*, or *circular paraboloid*. Many telescope mirrors are of this shape. The Equation to the circular paraboloid is

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} = \frac{z}{h}. \tag{4.4.2}$$

The cross-section at $z = h$ is a circle of radius a.

The Equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \frac{z}{h}, \tag{4.4.3}$$

in which we shall choose the x- and y-axes such that $a > b$, is an elliptic paraboloid and, if $a \neq b$, is not formed by rotation of a parabola. At $z = h$, the cross section is an ellipse of semi major and minor axes equal to a and b respectively. The section in the plane $y = 0$ is a parabola of semi latus rectum $a^2/(2h)$. The section in the plane $x = 0$ is a parabola of semi latus rectum $b^2/(2h)$. The elliptic paraboloid lies entirely above the xy-plane.

The Equation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = \frac{z}{h} \tag{4.4.4}$$

is a hyperbolic paraboloid, and its shape is not quite so easily visualized. Unlike the elliptic paraboloid, it extends above and below the plane. It is a saddle-shaped surface, with the saddle point at the origin. The section in the plane $y = 0$ is the "nose down" parabola $x^2 = a^2 z/h$ extending above the xy-plane. The section in the plane $x = 0$ is the "nose up" parabola $y^2 = -b^2 z/h$ extending below the xy-plane. The section in the plane $z = h$ is the hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{4.4.5}$$

The section with the plane $z = -h$ is the conjugate hyperbola

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = -1. \tag{4.4.6}$$

The section with the plane $z = 0$ is the asymptotes

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0. \tag{4.4.7}$$

The surface for $a = 3$, $b = 2$, $h = 1$ is drawn in figure IV.4.

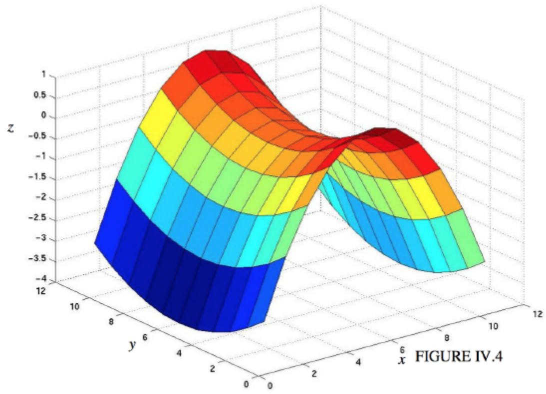


FIGURE IV.4

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.5: The Hyperboloid

The Equation

$$\frac{x^2}{a^2}-\frac{z^2}{c^2}=1 \tag{4.5.1}$$

is a hyperbola, and a is the semi transverse axis. (As described in Chapter 2, c is the semi transverse axis of the conjugate hyperbola.)

If this figure is rotated about the z-axis through 360°, the surface swept out is a *circular hyperboloid* (or *hyperboloid of revolution*) *of one sheet*. Its Equation is

$$\frac{x^2}{a^2}+\frac{y^2}{a^2}-\frac{z^2}{c^2}=1. \tag{4.5.2}$$

Imagine two horizontal rings, one underneath the other. The upper one is fixed. The lower one is suspended from the upper one by a large number of vertical strings attached to points equally spaced around the circumference of each ring. Now twist the lower one through a few degrees about a vertical axis, so that the strings are no longer quite vertical, and the lower ring rises slightly. These strings are generators of a circular hyperboloid of one sheet.

If the figure is rotated about the x-axis through 360°, the surface swept out is a *circular hyperboloid* (or *hyperboloid of revolution*) *of two sheets*. Its Equation is

$$\frac{x^2}{a^2}-\frac{y^2}{c^2}-\frac{z^2}{c^2}=1. \tag{4.5.3}$$

The Equations

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \tag{4.5.4}$$

and

$$\frac{x^2}{a^2}-\frac{y^2}{b^2}-\frac{z^2}{c^2}=1 \tag{4.5.5}$$

represent hyperbolas of one and two sheets respectively, but are not hyperbolas of revolution, since their cross sections in the planes $z=$ constant and $x=$ constant $>a$ respectively are ellipses rather than circles. The reader should imagine what the cross- sections of all four hyperboloids are like in the planes $x=0,\ y=0$ and $z=0$.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.6: The Cylinder

In three-dimensional solid geometry the Equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{4.6.1}$$

represents a cylinder of elliptical cross-section, whose axis coincides with the z-axis. The Equation to a cylinder with an axis in another position and with another orientation can be obtained by the usual processes of translation and rotation of axes (see Section 3.7).

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.7: The Cone

The Equation

$$x^2+y^2=a^2z^2 \tag{4.7.1}$$

represents a circular cone whose vertex in at the origin and whose axis coincides with the z-axis. The semi-vertical angle α of the cone is given by

$$\alpha=\tan^{-1}a. \tag{4.7.2}$$

In this context, the word "vertical" has nothing to do with "upright"; it merely means "of or at the vertex". A little knowledge of Latin is useful even today!

The Equation

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=z^2 \tag{4.7.3}$$

is a cone with elliptical cross-section.

If the vertex of the cone remains at the origin, but the axis is in some arbitrary direction (described, for example, by the direction cosines, or by spherical angles θ and ϕ) the Equation can be derived by a rotation of coordinate axes. This will introduce terms in yz, zx and xy, but it will not produce any terms in x, y or z, nor will it introduce a constant. Therefore the Equation to such a cone will have only second degree terms in it. The Equation will be of the form

$$ax^2+by^2+cz^2+2fyz+2gzx+2hxy=0 \tag{4.7.4}$$

With one proviso, the converse is usually true, namely that Equation 4.7.4 represents a cone with vertex at the origin. The one proviso is that if

$$\begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix}=0 \tag{4.7.5}$$

Equation 4.7.4 will factorize into two linear expressions, and will represent two planes, which intersect in a line that contains the origin - and this could be regarded as a special case of a cone, with zero vertical angle. If, however, the two linear factors are identical, the two planes are coincident.

The Equation to a cone of semi vertical angle α whose vertex is at the origin and whose axis has direction cosines (l,m,n) can be found as follows. The Equation

$$lx+my+nz+h=0 \tag{4.7.6}$$

represents a plane that is perpendicular to the axis of the cone and is at a distance h from the origin (i.e. from the vertex of the cone). Let P (x,y,z) be a point in the plane and also on the surface of the cone, at a distance r from the origin. The semi vertical angle of the cone is then given by

$$\cos\alpha=h/r. \tag{4.7.7}$$

But

$$r^2=x^2+y^2+z^2 \tag{4.7.8}$$

and, from Equation 4.7.6,

$$h^2=(lx+my+nz)^2 \tag{4.7.9}$$

Thus

$$(x^2+y^2+z^2)\cos^2\alpha=(lx+my+nz)^2 \tag{4.7.10}$$

is the required Equation to the cone.

It might be thought that the Equation to an inclined cone is unlikely to find much application in astronomy. Here is an application.

A fireball (i.e. a bright meteor, potentially capable of depositing a meteorite) moves down through the atmosphere with speed V along a straight line trajectory with direction cosines (l, m, n) referred to some coordinate system whose xy-plane is on the surface of Earth (assumed flat). If v is the speed of sound, and $V > v$, the meteoroid will generate a conical shock front of semi vertical angle α given by

$$\sin \alpha = v/V. \tag{4.7.11}$$

At a time t before impact, the coordinates of the vertex will be (lVt, mVt, nVt), and the Equation to the conical shock front will then be

$$\left[(x+lVt)^2 + (y+mVt)^2 + (z+nVt)^2\right] \cos^2 \alpha = \left[l(x+lVt) + m(y+mVt) + n(z+nVt)\right]^2 \tag{4.7.12}$$

Part of this shock front (at time t before impact) has already reached ground level, and it intersects the ground in a conic section given by putting $z = 0$ in Equation 4.7.12:

$$(\cos^2 \alpha - l^2)x^2 + (\cos^2 \alpha - m^2)y^2 - 2lmxy - 2Vt \sin^2 \alpha (lx + my) - V^2 t^2 \sin^2 \alpha = 0 \tag{4.7.13}$$

and witnesses on the ground on any point on this conic section will hear the shock front at the same time. Further details on this can be found in Tatum, J.B., *Meteoritics and Planetary Science*, **34**, 571 (1999) and Tatum, J.B., Parker, L. C. and Stumpf, L. L., *Planetary and Space Science* **48**, 921 (2000).

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.8: The General Second Degree Equation in Three Dimensions

The general second degree Equation in three dimensions is

$$ax^2+by^2+cz+2fyz+2gzx+2hxy+2ux+2vy+2wz+d=0 \tag{4.8.1}$$

This may represent a plane or pair of planes (which, if not parallel, define a straight line), or an ellipsoid, paraboloid, hyperboloid, cylinder or cone. The Equation could, if convenient, be divided through by d (or any of the other constants), and there are in reality only nine independent constants. Therefore nine points in space are sufficient to determine the second degree surface on this they lie.

If d is zero, the surface contains the origin. If u, v and w are all zero, and the surface is an ellipsoid, hyperbolic paraboloid or a hyperboloid, the origin is at the centre of the figure. If the figure is an elliptic paraboloid, the origin is at the vertex. If u, v, w and d are all zero, the surface is a cone with the proviso mentioned in section 4.7. If a, b, c, f, g, h are all zero, the surface is a plane.

Let us consider a particular example:

$$3x^2-4y^2+6z^2+8yz-2zx+4xy+14x-10y-4z+5=0 \tag{4.8.2}$$

What sort of a surface is this?

We need to do two things. First we need to rotate the coordinate axes so that they are parallel to the figure axes. The Equation referred to the figure axes will have no terms in yz, zx or xy. Then we need to translate the axes so that the origin is at the centre of the figure (or at the vertex, if it is an elliptical paraboloid).

Mathematically, we need to find the eigenvectors of the matrix

$$\begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix} = \begin{vmatrix} 3 & 2 & -1 \\ 2 & -4 & 4 \\ -1 & 4 & 6 \end{vmatrix} \tag{4.8.3}$$

Some readers will readily know how to do this. Others may not, and may not even be quite certain what an eigenvector is. Section 4.9 may be of interest to either group of readers. In any case, the eigenvectors are found to be

$$\begin{pmatrix} l_{11} \\ l_{12} \\ l_{13} \end{pmatrix} = \begin{pmatrix} -0.069\,5481 \\ +0.318\,8310 \\ +0.945\,2565 \end{pmatrix} \quad \begin{pmatrix} l_{12} \\ l_{22} \\ l_{32} \end{pmatrix} = \begin{pmatrix} -0.240\,6405 \\ +0.914\,2071 \\ -0.326\,0635 \end{pmatrix} \quad \begin{pmatrix} l_{13} \\ l_{23} \\ l_{33} \end{pmatrix} = \begin{pmatrix} -0.968\,1194 \\ -0.2501441 \\ +0.013\,1423 \end{pmatrix} \tag{4.8.1}$$

with corresponding eigenvalues $7.422\,7590,\ 5.953\,0969,\ 3.530\,3380$

The elements of the eigenvectors are the direction cosines of the present coordinate axis with respect to the figure axes. To express the Equation to the surface relative to coordinate axes that are parallel to the figure axes, we replace

$$x \text{ by } l_{11}x+l_{12}y+l_{13}z \tag{4.8.2}$$

$$y \text{ by } l_{21}x+l_{22}y+l_{23}z \tag{4.8.3}$$

$$z \text{ by } l_{31}x+l_{32}y+l_{33}z \tag{4.8.4}$$

This will make the terms in yz, zx and xy vanish; this should be checked numerically, particularly as it is easy to rotate the axes in the wrong sense. When the substitutions are made, the Equation is found to be

$$7.422\,7590x^2-5.9530969y^2+3.530\,3380z^2-7.9430994x-11.2067840y-11.1047998z+5=0. \tag{4.8.4}$$

Notice that there are now no terms in yz, zx or xy.

Now we need to translate the origin of coordinates to the centre of the figure (or to the vertex if it is an elliptic paraboloid). It will readily be seen that this can be done by substituting

$$x-\alpha \quad \text{for } x \tag{4.8.5}$$

$$y-\beta \quad \text{for } y \tag{4.8.6}$$

$$z-\gamma \quad \text{for } z \tag{4.8.7}$$

where

$$\alpha=(\text{coefficient of } x)/(\text{twice the coefficient of } x^2)=-0.535\ 050\ 336 \tag{4.8.8}$$

$$\beta=(\text{coefficient of } y)/(\text{twice the coefficient of } y^2)=+0.941\ 256\ 643 \tag{4.8.9}$$

$$\gamma=(\text{coefficient of } z)/(\text{twice the coefficient of } z^2)=-1.572\ 767\ 224 \tag{4.8.10}$$

.

The Equation then becomes

$$7.422\ 7590x^2-5.953\ 0969y^2+3.530\ 3380z^2-0.583\ 3816=0 \tag{4.8.11}$$

or

$$\frac{x^2}{(0.280\ 346)^2}-\frac{y^2}{(0.313\ 044)^2)}+\frac{z^2}{(0.406\ 507)^2}=1. \tag{4.8.12}$$

The surface is a hyperboloid of one sheet, elliptical in any $y=$ constant cross-section.

The surfaces described by second-degree Equations in three dimensions - ellipsoids, paraboloids, hyperboloids, cones and cylinders - are generally called quadric surfaces. The surface described by the Equation

$$\left[(x^2+y^2)^{\frac{1}{2}}-a\right]^2+z^2=b^2, \quad b<a \tag{4.8.13}$$

is not one of the quadric surfaces. If the square root is isolated and squared, the resulting Equation will contain terms of degree four. The surface is a fairly familiar one, and the reader should try to imagine what it is. Failing that, if your computer skills are up to it, you might try to draw the surface in three-dimensional space. The only hint I give is to suggest that you put $y=0$ in Equation ??? to see what the section is in the z-x plane.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

4.9: Matrices

This book has assumed a knowledge of matrices, which may or may not be justified. In this section, we do not attempt a thorough treatment of the subject, which must be sought elsewhere, but the remarks may be of use both to novices and to the more experienced.

However you may have the opportunity of learning about the manipulation of matrices, it is suggested that you should aim to understand and know how to carry out at least the following operations on matrices whose elements are real numbers:

Multiply a vector by a matrix
Multiply a matrix by a matrix
Calculate the determinant of a square matrix
Invert a square matrix
Diagonalize a symmetric matrix
Test a matrix for orthogonality

Numerous other aspects of matrix manipulation are possible, and the subject expands greatly if we allow the elements to be complex numbers. These six, however, are particularly useful for many applications. It might be mentioned, however, that solving simultaneous linear Equations by Kramers' Rule or by inverting a matrix are very inefficient ways of solving such Equations, and that is not the main purpose of acquiring the above skills.

Most, or doubtless all, of the above operations are available in many modern mathematical computer packages. This is not what I mean, however, by "understand and know how to carry out". The student should carry out at least once by hand calculator, step by step, each of the above operations, and, at each step, try to understand not only the algebraic and arithmetic steps, but also try to visualize the geometric interpretation, particularly when rotating axes and calculating eigenvectors. After doing a hand calculation, you should then write a series of short computer programs (rather than one vast, all-encompassing matrix package) for each operation, so that when, in future, you need to do any of these things, you can instantly obtain the answer without having to go through tedious calculations. For example, in the previous section, when I needed the eigenvectors of the matrix, I was able to generate them with a single word "EIGEN" on a computer; a considerable amount of arithmetic was actually performed by the computer.

On the question of testing a matrix for orthogonality, the usual application in mechanical and geometrical problems is to test a matrix of direction cosines. The tests can not only detect mistakes, but it can locate them and even suggest what the correct element should be. Tests for orthogonality are as follows. The student should try to think of the geometric interpretation of each.

The sum of the squares of the elements in any row or any column is unity. (This test does not guard against mistakes in signs of the elements.)

The sum of the products of corresponding elements in any two rows or in any two columns is zero.

Every element is equal to plus or minus its cofactor.

The determinant of the matrix is plus or minus one.

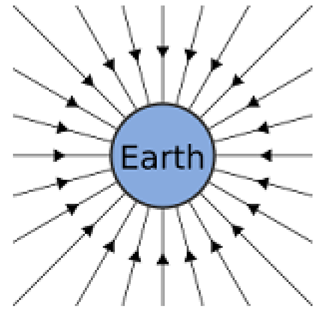
A minus sign in the last two tests indicates that the two sets of axes differ in chirality (handedness). This usually does not matter, and can easily be dealt with by reversing the signs of the elements of one eigenvector and of its corresponding eigenvalue.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

5: GRAVITATIONAL FIELD AND POTENTIAL



This chapter deals with the calculation of gravitational fields and potentials in the vicinity of various shapes and sizes of massive bodies. The reader who has studied electrostatics will recognize that this is all just a repeat of what he or she already knows.

5.10: NABLA, GRADIENT AND DIVERGENCE

We are going to meet, in this section, the symbol ∇ . In North America it is generally pronounced "del", although in the United Kingdom and elsewhere one sometimes hears the alternative pronunciation "nabla", called after an ancient Assyrian harp-like instrument of approximately that shape.

5.11: LEGENDRE POLYNOMIALS

In this section we cover just enough about Legendre polynomials to be useful in the following section.

5.12: GRAVITATIONAL POTENTIAL OF ANY MASSIVE BODY

5.13: PRESSURE AT THE CENTRE OF A UNIFORM SPHERE

What is the pressure at the centre of a sphere of radius a and of uniform density ρ?

5.1: Introduction

This chapter deals with the calculation of gravitational fields and potentials in the vicinity of various shapes and sizes of massive bodies. The reader who has studied electrostatics will recognize that this is all just a repeat of what he or she already knows. After all, the force of repulsion between two electric charges q_1 and q_2 a distance r apart *in vacuo* is

$$\frac{q_1 q_2}{4\pi\varepsilon_0 r^2}, \tag{5.1.1}$$

where ε_0 is the permittivity of free space, and the attractive force between two masses M_1 and M_2 a distance r apart is

$$\frac{GM_1M_2}{r^2}, \tag{5.1.2}$$

where G is the gravitational constant, or, phrased another way, the *repulsive* force is

$$-\frac{GM_1M_2}{r^2}. \tag{5.1.3}$$

Thus all the Equations for the fields and potentials in gravitational problems are the same as the corresponding Equations in electrostatics problems, provided that the charges are replaced by masses and $4\pi\varepsilon_0$ is replaced by $-1/G$.

I can, however, think of two differences. In the electrostatics case, we have the possibility of both positive and negative charges. As far as I know, only positive masses exist. This means, among other things, that we do not have "gravitational dipoles" and all the phenomena associated with polarization that we have in electrostatics.

The second difference is this. If a particle of mass m and charge q is placed in an electric field $\mathbf{E}$, it will experience a force $q\mathbf{E}$, and it will accelerate at a rate and in a direction given by $q\mathbf{E}/m$. If the same particle is placed in a gravitational field $\mathbf{g}$, it will experience a force $m\mathbf{g}$ and an acceleration $m\mathbf{g}/m = \mathbf{g}$, irrespective of its mass or of its charge. All masses and all charges in the same gravitational field accelerate at the same rate. This is not so in the case of an electric field.

I have some sympathy for the idea of introducing a "rationalized" gravitational constant Γ, given by $\Gamma = 1/(4\pi G)$, in which case the gravitational formulas would look even more like the SI (rationalized MKSA) electrostatics formulas, with 4π appearing in problems with spherical symmetry, 2π in problems with cylindrical symmetry, and no π in problems involving uniform fields. This is unlikely to happen, so I do not pursue the idea further here.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.2: Gravitational Field

The region around a gravitating body (by which I merely mean a mass, which will attract other masses in its vicinity) is a *gravitational field*. Although I have used the words "around" and "in its vicinity", the field in fact extents to infinity. All massive bodies (and by "massive" I mean any body having the property of mass, however little) are surrounded by a gravitational field, and all of us are immersed in a gravitational field.

If a test particle of mass m is placed in a gravitational field, it will experience a *force* (and, if released and subjected to no additional forces, it will *accelerate*). This enables us to define quantitatively what we mean by the *strength* of a gravitational field, which is merely the *force experienced by unit mass* placed in the field. I shall use the symbol $\mathbf{g}$ for the gravitational field, so that the force $\mathbf{F}$ on a mass m situated in a gravitational field $\mathbf{g}$ is

$$\mathbf{F} = m\mathbf{g}. \tag{5.2.1}$$

It can be expressed in newtons per kilogram, $\mathrm{N\,kg^{-1}}$. If you work out the *dimensions* of g, you will see that it has dimensions $\mathrm{LT^{-2}}$, so that it can be expressed equivalently in $\mathrm{m\,s^{-2}}$. Indeed, as pointed out in section 5.1, the mass m (or indeed any other mass) will accelerate at a rate g in the field, and the strength of a gravitational field is simply equal to the rate at which bodies placed in it will accelerate.

Very often, instead of using the expression "strength of the gravitational field" I shall use just "the gravitational field" or perhaps the "field strength" or even just the "field". Strictly speaking, the "gravitational field" means the region of space surrounding a gravitating mass rather than the field strength, but I hope that, when I am not speaking strictly, the context will make it clear.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.3: Newton's Law of Gravitation

Newton noted that the ratio of the centripetal acceleration of the Moon in its orbit around the Earth to the acceleration of an apple falling to the surface of the Earth was inversely as the squares of the distances of Moon and apple from the centre of the Earth. Together with other lines of evidence, this led Newton to propose his universal law of gravitation:

Every particle in the Universe attracts every other particle with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them. In symbols:

$$F = \frac{GM_1M_2}{r^2}. \quad \text{N} \tag{5.3.1}$$

Here, G is the Universal Gravitational Constant. The word "universal" implies an assumption that its value is the same anywhere in the Universe, and the word "constant" implies that it does not vary with time. We shall here accept and adopt these assumptions, while noting that it is a legitimate cosmological question to consider what implications there may be if either of them is not so.

Of all the fundamental physical constants, G is among those whose numerical value has been determined with least precision. Its currently accepted value is $6.6726 \times 10^{-11}\text{N m}^2 \text{ kg}^{-2}$. It is worth noting that, while the product GM for the Sun is known with very great precision, the mass of the Sun is not known to any higher degree of precision than that of the gravitational constant.

Exercise. Determine the *dimensions* (in terms of M, L and T) of the gravitational constant. Assume that the period of pulsation of a variable star depends on its mass, its average radius and on the value of the gravitational constant, and show that the period of pulsation must be inversely proportional to the square root of its average density.

The gravitational field is often held to be the weakest of the four forces of nature, but to aver this is to compare incomparables. While it is true that the electrostatic force between two electrons is far, far greater than the gravitational force between them, it is equally true that the gravitational force between Sun and Earth is far, far greater than the electrostatic force between them. This example shows that it makes no sense merely to state that electrical forces are stronger than gravitational forces. Thus any statement about the relative strengths of the four forces of nature has to be phrased with care and precision.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4: The Gravitational Fields of Various Bodies

In this section we calculate the fields near various shapes and sizes of bodies, much as one does in an introductory electricity course. Some of this will not have much direct application to celestial mechanics, but it will serve as good introductory practice in calculating fields and, later, potentials.

Topic hierarchy

5.4.1: Field of a Point Mass

Equation 5.3.1, together with the definition of field strength as the force experienced by unit mass, means that the field at a distance r from a point mass M is

$$g=\frac{GM}{r^2} \quad \mathrm{N\,kg^{-1}\ or\ m\,s^{-2}} \tag{5.4.1}$$

In vector form, this can be written as

$$\mathbf{g}=-\frac{GM}{r^2}\hat{\mathbf{r}} \quad \mathrm{N\,kg^{-1}\ or\ m\,s^{-2}} \tag{5.4.2}$$

Here $\hat{\mathbf{r}}$ is a dimensionless *unit* vector in the radial direction.

It can also be written as

$$\mathbf{g}=-\frac{GM}{r^3}\mathbf{r} \quad \mathrm{N\,kg^{-1}\ or\ m\,s^{-2}} \tag{5.4.3}$$

Here $\mathbf{r}$ is a vector of magnitude r − hence the r^3 in the denominator.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.2: Field on the Axis of a Ring

Before starting, one can obtain a qualitative idea of how the field on the axis of a ring varies with distance from the centre of the ring. Thus, the field at the centre of the ring will be zero, by symmetry. It will also be zero at an infinite distance along the axis. At other places it will not be zero; in other words, the field will first increase, then decrease, as we move along the axis. There will be some distance along the axis at which the field is greatest. We'll want to know where this is, and what is its maximum value.

FIGURE V.1

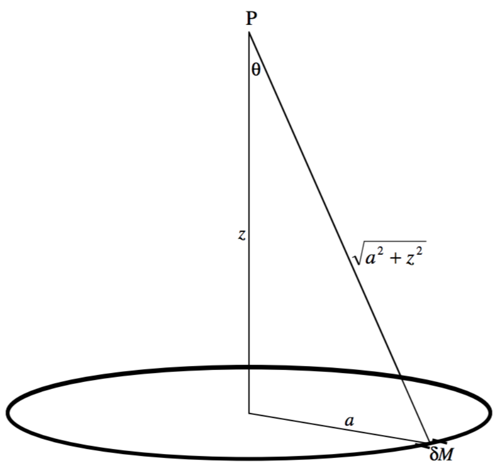


Figure V.1 shows a ring of mass M, radius a. The problem is to calculate the strength of the gravitational field at P. We start by considering a small element of the ring of mass δM. The contribution of this element to the field is

$$\frac{G\delta M}{a^2+z^2}, \tag{5.4.2.1}$$

directed from P towards δM. This can be resolved into a component along the axis (directed to the centre of the ring) and a component at right angles to this. When the contributions to all elements around the circumference of the ring are added, the latter component will, by symmetry, be zero. The component along the axis of the ring is

$$\frac{G\delta M}{a^2+z^2}\cos\theta = \frac{G\delta M}{a^2+z^2}\cdot\frac{z}{\sqrt{a^2+z^2}} = \frac{G\delta M z}{(a^2+z^2)^{3/2}}. \tag{5.4.2.2}$$

On adding up the contributions of all elements around the circumference of the ring, we find, for the gravitational field at P

$$g = \frac{GMz}{(a^2+z^2)^{3/2}} \tag{5.4.4}$$

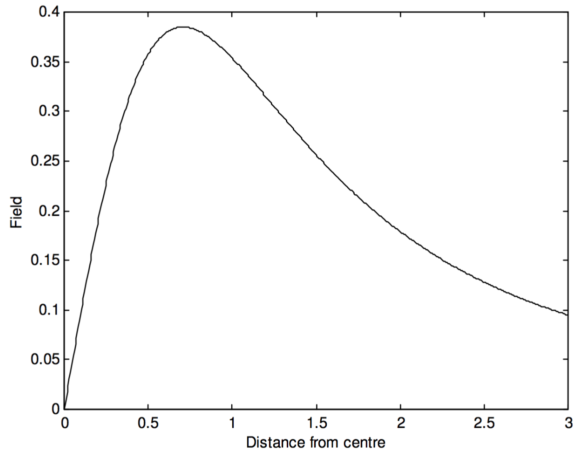
directed towards the centre of the ring. This has the property, as expected, of being zero at the centre of the ring and at an infinite distance along the axis. If we express z in units of a, and g in units of GM/a^2, this becomes

$$g = \frac{z}{(1+z^2)^{3/2}}. \tag{5.4.5}$$

This is illustrated in figure V.2.

Excercise: Show that the field reaches its greatest value of $\frac{\sqrt{12}GM}{9a^2} = \frac{0.385GM}{a^2}$ where $z = a/\sqrt{2} = 0.707a$. Show that the field has half this maximum value where $z = 0.2047a$ and $z = 1.896a$.

FIGURE V.2

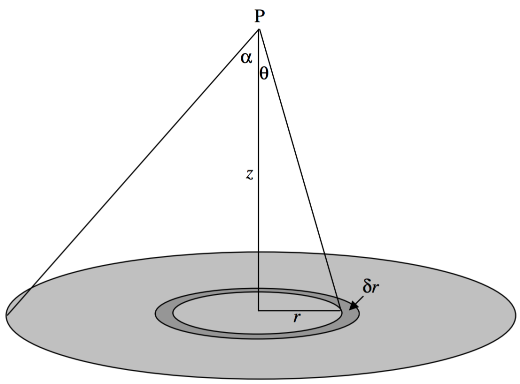


Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.3: Plane discs

FIGURE V.2A



Consider a disc of surface density (mass per unit area) σ, radius a, and a point P on its axis at a distance z from the disc. The contribution to the field from an elemental annulus, radii r, $r+\delta r$, mass $2\pi\sigma\ r\ \delta r$ is (from Equation 5.4.1)

$$\delta g = 2\pi G\sigma \frac{zr\delta r}{\left(z^2+r^2\right)^{3/2}}. \tag{5.4.6}$$

To find the field from the entire disc, just integrate from $r=0$ to a, and, if the disc is of uniform surface density, σ will be outside the integral sign. It will be easier to integrate with respect to θ (from 0 to α), where $r=z\tan\theta$. You should get

$$g = 2\pi G\sigma(1-\cos\alpha), \tag{5.4.7}$$

or, with $M=\pi a^2\sigma$,

$$g = \frac{2GM(1-\cos\alpha)}{a^2}. \tag{5.4.8}$$

Now $2\pi(1-\cos\alpha)$ is the solid angle ω subtended by the disc at P. (Convince yourself of this – don't just take my word for it.) Therefore

$$g = G\sigma\omega. \tag{5.4.9}$$

This expression is also the same for a uniform plane lamina of any shape, for the downward component of the gravitational field. For, consider figure V.3.

The downward component of the field due to the element δA is $\frac{G\sigma\delta A\cos\theta}{r^2} = G\sigma\delta\omega$. Thus, if you integrate over the whole lamina, you arrive at $G\sigma\omega$.

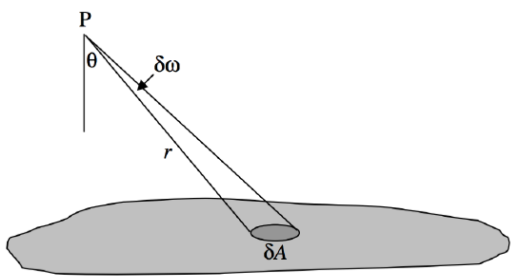


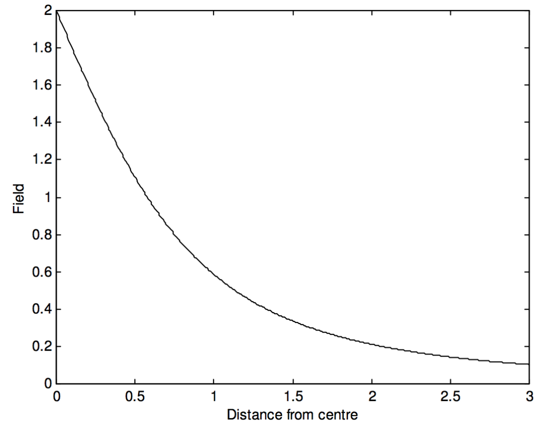
FIGURE 5.3

Returning to Equation 5.4.8, we can write the Equation in terms of z rather than α. If we express g in units of GM/a^2 and z in units of a, the Equation becomes

$$g = 2\left(1 - \frac{z}{\sqrt{1+z^2}}\right). \tag{5.4.10}$$

This is illustrated in figure V.4.

FIGURE V.4



The field is greatest immediately above the disc. On the opposite side of the disc, the field changes direction. In the plane of the disc, at the centre of the disc, the field is zero. For more on this, see Subsection 5.4.7.

If you are calculating the field on the axis of a disc that is not of uniform surface density, but whose surface density varies as $\sigma(r)$, you will have to calculate

$$M = 2\pi\int_0^a \sigma(r)rdr \tag{5.4.11}$$

and

$$g = 2\pi Gz\int_0^a \frac{\sigma(r)rdr}{(z^2+r^2)^{3/2}}. \tag{5.4.12}$$

You could try, for example, some of the following forms for $\sigma(r)$:

$$\sigma_0\left(1-\frac{kr}{a}\right), \quad \sigma_0\left(1-\frac{kr^2}{a}\right), \quad \sigma_0\sqrt{1-\frac{kr}{a}}, \quad \sigma_0\sqrt{1-\frac{kr^2}{a^2}}. \tag{5.4.3.1}$$

If you are interested in galaxies, you might want to try modelling a galaxy as a central spherical bulge of density ρ and radius a_1, plus a disc of surface density $\sigma(r)$ and radius a_2, and from there you can work your way up to more sophisticated models.

In this section we have calculated the field on the *axis* of a disc. As soon as you move off axis, it becomes much more difficult.

Exercise. Starting from Equations 5.4.1 and 5.4.10, show that at vary large distances along the axis, the fields for a ring and for a disc each become GM/z^2. All you have to do is to expand the expressions binomially in a/z. The field at a large distance r from any finite object will approach GM/r^2.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.4: Infinite Plane Laminas

For the gravitational field due to a uniform infinite plane lamina, all one has to do is to put $\alpha = \pi/2$ in Equation 5.4.7 or $\omega = 2\pi$ in Equation 5.4.9 to find that the gravitational field is

$$g = 2\pi G\sigma. \tag{5.4.13}$$

This is, as might be expected, independent of distance from the infinite plane. The lines of gravitational field are uniform and parallel all the way from the surface of the lamina to infinity.

Suppose that the surface density of the infinite plane is not uniform, but varies with distance in the plane from some point in the plane as $\sigma(r)$, we have to calculate

$$g = 2\pi Gz \int_0^\infty \frac{\sigma(r) r dr}{\left(z^2 + r^2\right)^{3/2}}. \tag{5.4.14}$$

Try it, for example, with $\sigma(r)$ being one of the following:

$$\sigma_0 e^{-kr}, \quad \sigma_0 e^{-k^2 r^2}, \quad \frac{\sigma_0}{1 + k^2 r^2}. \tag{5.4.4.1}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.5: Hollow Hemisphere

Exercise. Find the field at the centre of the base of a hollow hemispherical shell of mass M and radius a.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.6: Rods

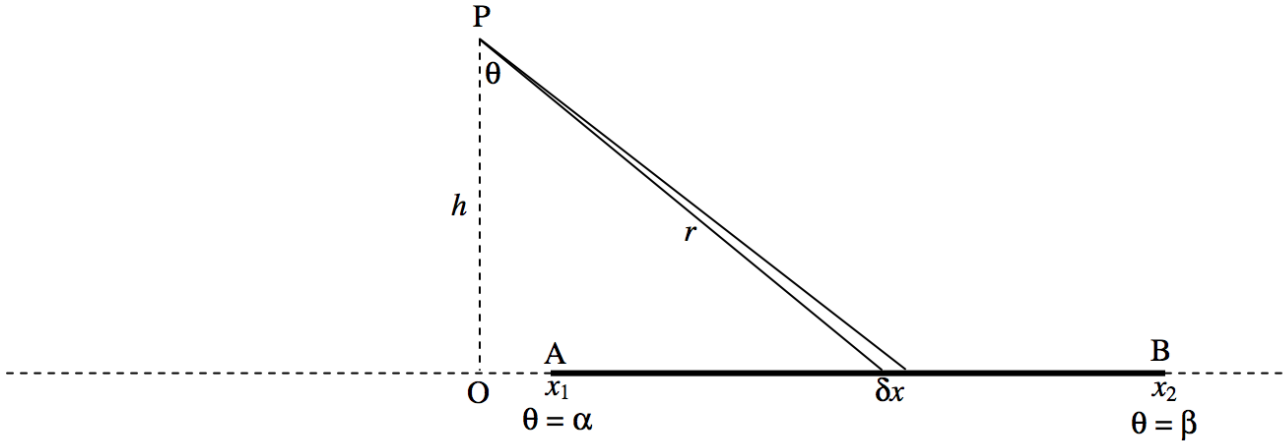


FIGURE V.5

Consider the rod shown in figure V.5, of mass per unit length λ. The field at P due to the element δx is $G\lambda\delta x/r^2$. But $x=\tan\theta$, $\delta x = h\sec^2\theta\delta\theta$, $r = h\sec\theta$ so the field at P is $G\lambda\delta\theta/h$ This is directed from P to the element δx.

The x-component of the field due to the whole rod is

$$\frac{G\lambda}{h}\int_\alpha^\beta \sin\theta\, d\theta = \frac{G\lambda}{h}(\cos\alpha - \cos\beta). \tag{5.4.15}$$

The y-component of the field due to the whole rod is

$$-\frac{G\lambda}{h}\int_\alpha^\beta \cos\theta\, d\theta = -\frac{G\lambda}{h}(\sin\beta - \sin\alpha). \tag{5.4.16}$$

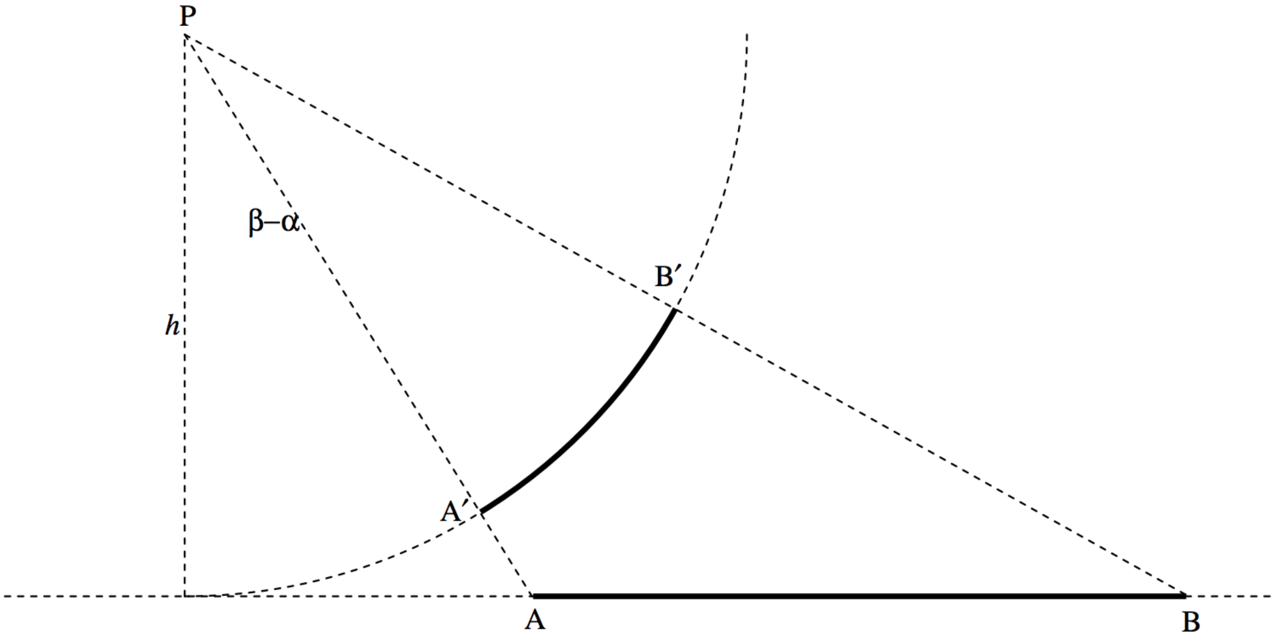
The total field is the orthogonal sum of these, which, after use of some trigonometric identities (do it!), becomes

$$g = \frac{2G\lambda}{h}\sin\frac{1}{2}(\beta-\alpha) \tag{5.4.17}$$

at an angle $\frac{1}{2}(\alpha+\beta)$ - i.e. bisecting the angle APB.

If the rod is of infinite length, we put $\alpha = -\pi/2$ and $\beta = \pi/2$, and we obtain for the field at P

$$g = \frac{2G\lambda}{h}. \tag{5.4.18}$$



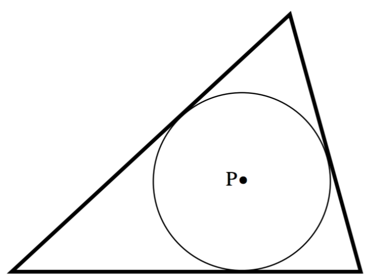
$textFIGUREV.6$

Consider an arc $\mathrm{A'B'}$ of a circle of radius h, mass per unit length λ, subtending an angle $\beta-\alpha$ at the centre P of the circle.

Exercise: Show that the field at P is $g=\frac{2G\lambda}{h}\sin\frac{1}{2}(\beta-\alpha)$. This is the same as the field due to the rod AB subtending the same angle. If $\mathrm{A'B'}$ is a semicircle, the field at P would be $g=\frac{2G\lambda}{h}$, the same as for an infinite rod.

An interesting result following from this is as follows.

FIGURE V.7



Three massive rods form a triangle. P is the incentre of the triangle (i.e. it is equidistant from all three sides.) The field at P is the same as that which would be obtained if the mass were distributed around the incircle. I.e., it is zero. The same result would hold for any quadrilateral that can be inscribed with a circle – such as a cyclic quadrilateral.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.7: Solid Cylinder

We do this not because it has any particular relevance to celestial mechanics, but because it is easy to do. We imagine a solid cylinder, density ρ, radius a, length l. We seek to calculate the field at a point P on the axis, at a distance h from one end of the cylinder (figure V.8).

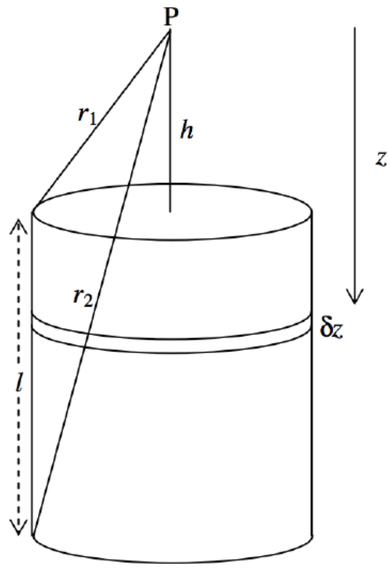


FIGURE V.8

The field at P from an elemental disc of thickness δz a distance z below P is (from Equation 5.4.9)

$$\delta g = G\rho\delta z\omega. \tag{5.4.19}$$

Here ω is the solid angle subtended at P by the disc, which is $2\pi\left[1-\frac{z}{(z^2+a^2)^{1/2}}\right]$. Thus the field at P from the entire cylinder is

$$g = 2\pi G\rho\int_h^{l+h}\left[1-\frac{z}{(z^2+a^2)^{1/2}}\right]dz, \tag{5.4.20}$$

or

$$g = 2\pi G\rho\left(l-\sqrt{(l+h)^2+a^2}+\sqrt{h^2+a^2}\right), \tag{5.4.21}$$

or

$$g = 2\pi G\rho(l-r_2+r_1). \tag{5.4.22}$$

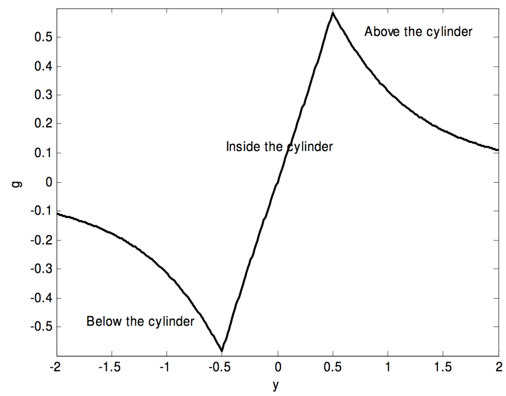
It might also be of interest to express g in terms of the height $y(=\frac{1}{2}l+h)$ of the point P above the mid-point of the cylinder. Instead of Equation 5.4.21, we then have

$$g = 2\pi G\rho\left(l-\sqrt{(y+\frac{1}{2}l)^2+a^2}+\sqrt{(y-\frac{1}{2}l)^2+a^2}\right). \tag{5.4.23}$$

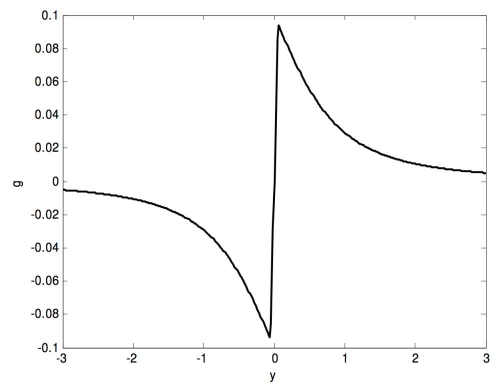
If the point P is *inside* the cylinder,at a distance h below the upper end of the cylinder, the limits of integration in Equation 5.4.20are h and $l-h$, and the distance y is $\frac{1}{2}l-h$. In terms of y the gravitational field at P is then

$$g = 2\pi G\rho\left(2y-\sqrt{(y+\frac{1}{2}l)^2+a^2}+\sqrt{(y-\frac{1}{2}l)^2+a^2}\right). \tag{5.4.24}$$

In the graph below I have assumed, by way of example, that l and a are both 1, and I have plotted g in units of $2\pi G\rho$ (counting g as positive when it is directed downwards) from $y=-1$ to $y=+1$. The portion inside the cylinder $(-\frac{1}{2} \le y \le \frac{1}{2} l)$, represented by Equation 5.4.24, is almost, but not quite, linear. The field at the centre of the cylinder is, of course, zero.



Below, I draw the same graph, but for a thin disc, with $a=1$ and $l=0.1$. We see how it is that the field reaches a maximum immediately above or below the disc, but is zero at the centre of the disc.



Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.8: Hollow Spherical Shell

We imagine a hollow spherical shell of radius a, surface density σ, and a point P at a distance r from the centre of the sphere. Consider an elemental zone of thickness δx. The mass of this element is $2\pi a\sigma\,\delta x$. (In case you doubt this, or you didn't know, "the area of a zone on the surface of a sphere is equal to the corresponding area projected on to the circumscribing cylinder".)

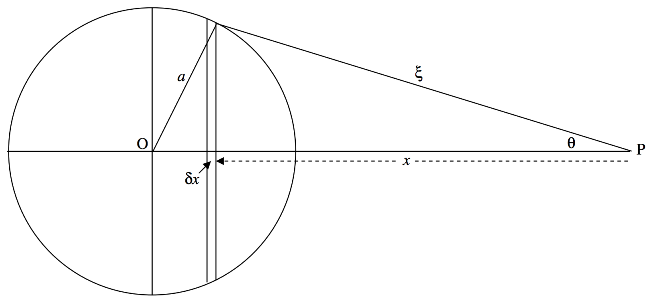


FIGURE V.9

The field due to this zone, in the direction PO is

$$\frac{2\pi a\sigma G\cos\theta\delta x}{\xi^2}. \tag{5.4.8.1}$$

Let's express this all in terms of a single variable, ξ. We are going to have to express x and θ in terms of ξ.

We have

$$a^2 = r^2 + \xi^2 - 2r\xi\cos\theta = r^2 + \xi^2 - 2rx, \tag{5.4.8.2}$$

from which

$$\cos\theta = \frac{r^2 - a^2 + \xi^2}{2r\xi} \qquad \delta x = \frac{\xi\delta\xi}{r}. \tag{5.4.8.3}$$

Therefore the field at P due to the zone is $\frac{\pi a G\sigma}{r^2}\left(1+\frac{r^2-a^2}{\xi^2}\right)\delta\xi$.

If P is an external point, in order to find the field due to the entire spherical shell, we integrate from $\xi = r-a$ to $r+a$. This results in

$$g = \frac{GM}{r^2}. \tag{5.4.8.4}$$

However, if P is an *internal* point, in order to find the field due to the entire spherical shell, we integrate from $\xi = a-r$ to $a+r$, which results in $g = 0$.

> *Thus we have the important result that the field at an external point due to a hollow spherical shell is exactly the same as if all the mass were concentrated at a point at the centre of the sphere, whereas the field inside the sphere is zero.*

Caution. The field inside the sphere is zero only if there are no other masses present. The hollow sphere will not shield you from the gravitational field of any other masses that might be present. Thus in figure V.10, the field at P is the sum of the field due to the hollow sphere (which is indeed zero) and the field of the mass M, which is not zero. Anti-grav is a useful device in science fiction, but does not occur in science fact.

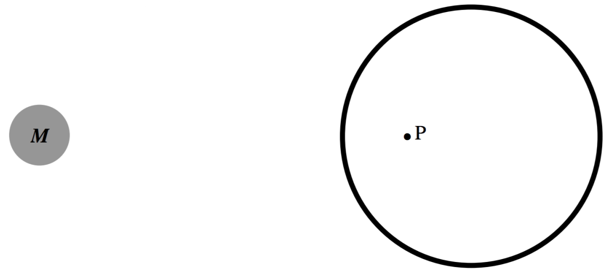


FIGURE V.10

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.9: Solid Sphere

A solid sphere is just lots of hollow spheres nested together. Therefore, the field at an external point is just the same as if all the mass were concentrated at the centre, and the field at an internal point P is the same is if all the mass *interior* to P, namely M_r, were concentrated at the centre, the mass *exterior* to P not contributing at all to the field at P. This is true not only for a sphere of uniform density, but of any sphere in which the density depends only of the distance from the centre – i.e., any spherically symmetric distribution of matter.

If the sphere is uniform, we have $\frac{M_r}{M} = \frac{r^3}{a^3}$, so the field inside is

$$g = \frac{GM_r}{r^2} = \frac{GMr}{a^3}. \tag{5.4.24}$$

Thus, inside a uniform solid sphere, the field increases linearly from zero at the centre to GM/a^2 at the surface, and thereafter it falls off as GM/r^2.

If a uniform hollow sphere has a narrow hole bored through it, and a small particle of mass m is allowed to drop through the hole, the particle will experience a force towards the centre of $GMmr/a^3$, and will consequently oscillate with period P given by

$$P^2 = \frac{4\pi^2}{GM}a^3. \tag{5.4.25}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.4.10: Bubble Inside a Uniform Solid Sphere

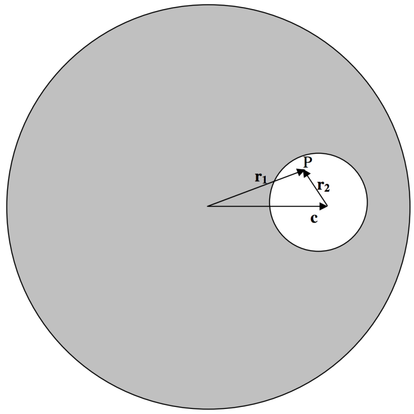


FIGURE V.11

P is a point inside the bubble. The field at P is equal to the field due to the entire sphere minus the field due to the missing mass of the bubble. That is, it is

$$\mathbf{g} = -\frac{4}{3}\pi G\rho\mathbf{r}_1 - (-\frac{4}{3}\pi G\rho\mathbf{r}_2) = -\frac{4}{3}\pi G\rho(\mathbf{r}_1 - \mathbf{r}_2) = -\frac{4}{3}\pi G\rho\mathbf{c}. \tag{5.4.26}$$

That is, the field at P is uniform (i.e. is independent of the position of P) and is parallel to the line joining the centres of the two spheres.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.5: Gauss's Theorem

Much of the above may have been good integration practice, but we shall now see that many of the results are immediately obvious from Gauss's Theorem – itself a trivially obvious law. (Or shall we say that, like many things, it is trivially obvious *in hindsight*, though it needed Carl Friedrich Gauss to point it out!)

First let us define gravitational *flux* Φ as an extensive quantity, being the product of gravitational field and area:

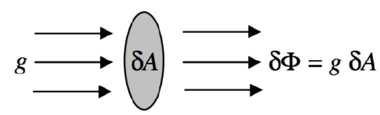


FIGURE V.12

If $\mathbf{g}$ and $\delta\mathbf{A}$ are not parallel, the flux is a scalar quantity, being the scalar or dot product of $\mathbf{g}$ and $\delta\mathbf{A}$:

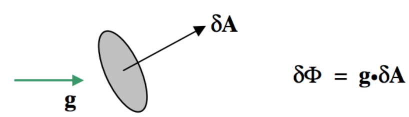


FIGURE V.13

If the gravitational field is threading through a large finite area, we have to calculate $\mathbf{g} \bullet \delta\mathbf{A}$ for each element of area of the surface, the magnitude and direction of $\mathbf{g}$ possibly varying from point to point over the surface, and then we have to integrate this all over the surface. In other words, we have to calculate a *surface integral*. We'll give some examples as we proceed, but first let's move toward Gauss's theorem.

In figure $\text{V}.14$, I have drawn a mass M and several of the gravitational field lines converging on it. I have also drawn a sphere of radius r around the mass. At a distance r from the mass, the field is GM/r^2. The surface area of the sphere is $4\pi r^2$. Therefore the total inward flux, the product of these two terms, is $4\pi GM$, and is independent of the size of the sphere. (It is independent of the size of the sphere because the field falls off inversely as the square of the distance. Thus Gauss's theorem is a theorem that applies to inverse square fields.) Nothing changes if the mass is not at the centre of the sphere. Nor does it change if (figure $\text{V}.15$) the surface is not a sphere. If there were several masses inside the surface, each would contribute $4\pi G$ times its mass to the total normal inwards flux. Thus the total normal inward flux through any closed surface is equal to $4\pi G$ times the total mass enclosed by the surface. Or, expressed another way:

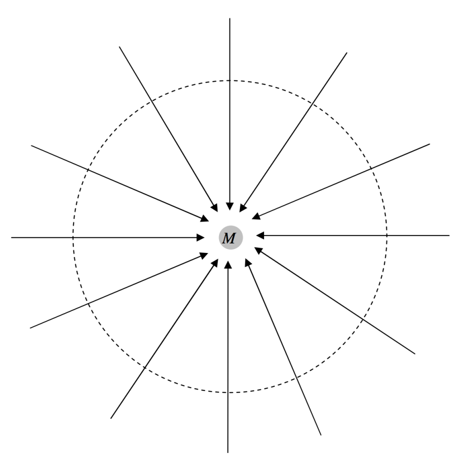


FIGURE V.14

The total normal outward gravitational flux through a closed surface is equal to $-4\pi G$ times the total mass enclosed by the surface.

This is Gauss's theorem.

Mathematically, the flux through the surface is expressed by the surface integral $\int\int \mathbf{g}\cdot d\mathbf{A}$. If there is a continuous distribution of matter inside the surface, of density ρ which varies from point to point and is a function of the coordinates, the total mass inside the surface is expressed by $\int\int\int \rho dV$. Thus Gauss's theorem is expressed mathematically by

$$\int\int \mathbf{g}\cdot d\mathbf{A} = -4\pi G\int\int\int \rho dV. \tag{5.5.1}$$

You should check the dimensions of this Equation.

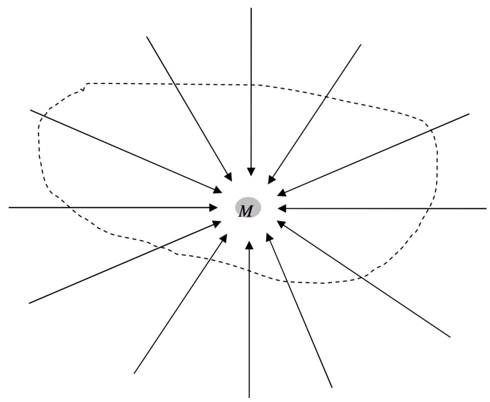


FIGURE V.15

In figure V.16 I have drawn gaussian spherical surfaces of radius r outside and inside hollow and solid spheres. In a and c, the outward flux through the surface is just $-4\pi G$ times the enclosed mass M; the surface area of the gaussian surface is $4\pi r^2$. This the outward field at the gaussian surface (i.e. at a distance r from the centre of the sphere is $-GM/r^2$. In b, no mass is inside the gaussian surface, and therefore the field is zero. In d, the mass inside the gaussian surface is M_r, and so the outward field is $-GM_r/r^2$.

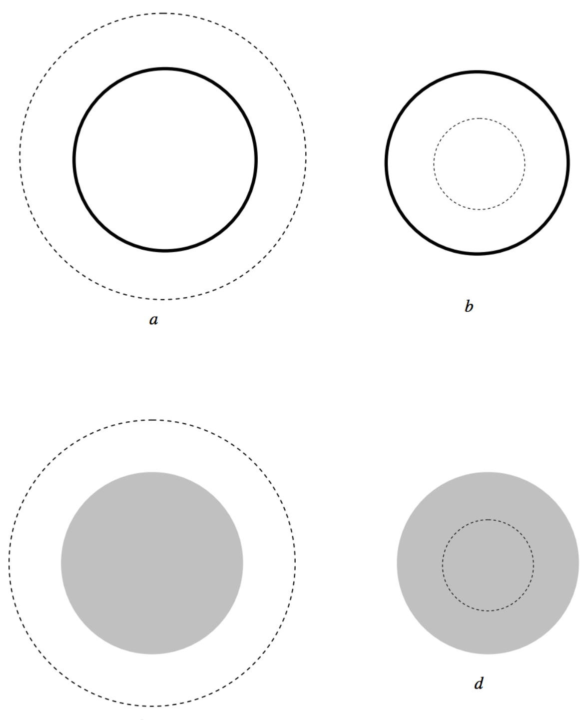


FIGURE V.16

In figure V.17 I draw (part of an) infinite rod of mass λ per unit length, and a cylindircal gaussian surface of radius h and length l around it.

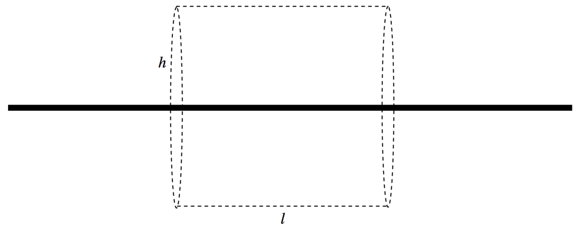


FIGURE V.17

The surface area of the curved surface of the cylinder is $2\pi hl$, and the mass enclosed within it is λl. Thus the outward field at the surface of the gaussian cylinder (i.e. at a distance h from the rod) is $-4\pi G\times\lambda l\div 2\pi hl=-2G\lambda/h$, in agreement with Equation 5.4.18.

In figure $\text{V}.18$ I have drawn (part of) an infinite plane lamina of surface density σ, and a cylindrical gaussian surface or cross-sectional area A and height $2h$.

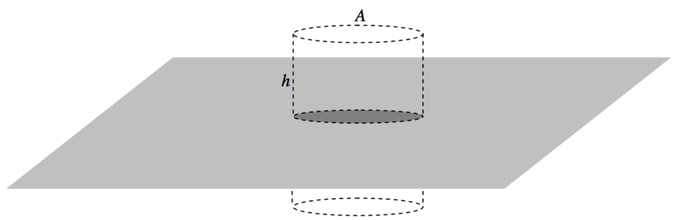


FIGURE V.18

The mass enclosed by the cylinder is σA and the area of the two ends of the cylinder is $2A$. The outward field at the ends of the cylinder (i.e. at a distance h from the plane lamina) is therefore $-4\pi G\times\sigma A\div 2A=-2\pi G\sigma$, in agreement with Equation 5.4.13.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.6: Calculating Surface Integrals

While the concept of a surface integral sounds easy enough, how do we actually calculate one in practice? In this section I do two examples.

Example 5.6.1

In Figure V.19 I show a small mass m, and I have surrounded it with a cylinder or radius a and height $2h$. The problem is to calculate the surface integral $\int \mathbf{g} \cdot d\mathbf{A}$ through the entire surface of the cylinder. Of course we already know, from Gauss's theorem, that the answer is $= -4\pi Gm$, but we would like to see a surface integral actually carried out.

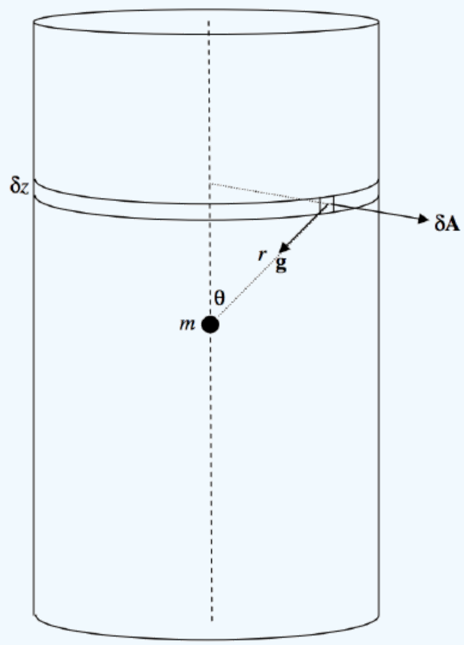


FIGURE V.19

I have drawn a small element of the surface. Its area δA is dz times $a\delta\phi$, where ϕ is the usual azimuthal angle of cylindrical coordinates. That is, $\delta A = a\,\delta z\,\delta\phi$. The magnitude g of the field there is Gm/r^2, and the angle between $\mathbf{g}$ and $d\mathbf{A}$ is $90^\circ + \theta$. The outward flux through the small element is

$$\mathbf{g} \cdot \delta\mathbf{A} = \frac{Gma\cos(\theta + 90^\circ)\delta z\delta\phi}{r^2}. \tag{5.6.1}$$

(This is negative – i.e. it is actually an inward flux – because $\cos(\theta + 90^\circ) = -\sin\theta$.) When integrated around the elemental strip δz, this is $-\frac{2\pi Gma\sin\theta\delta z}{r^2}$. To find the flux over the total curved surface, let's integrate this from $z = 0$ to h and double it, or, easier, from $\theta = \pi/2$ to α and double it, where $\tan\alpha = a/h$. We'll need to express z and r in terms of θ (that's easy:- $z = a\cot\theta$ and $r = a\csc\theta$),and the integral becomes

$$4\pi Gm \int_{\pi/2}^{\alpha} \sin\theta\, d\theta = -4\pi Gm\cos\alpha \tag{5.6.1}$$

Let us now find the flux through one of the flat ends of the cylinder.

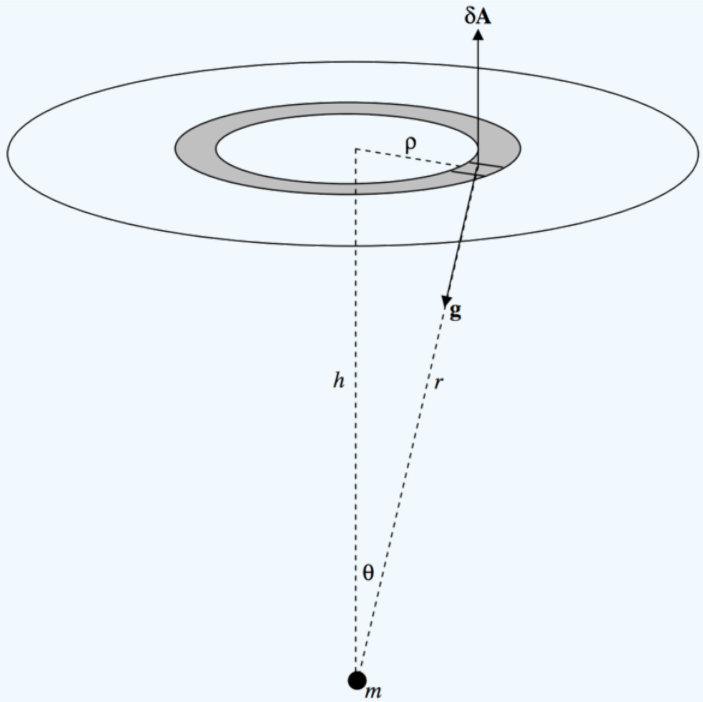


FIGURE V.20

This time, $\delta A=\rho\,\delta\rho\,\delta\phi$, $g=Gm/r^2$ and the angle between $\mathbf{g}$ and $\delta\mathbf{A}$ is $180^\circ-\theta$. The outwards flux through the small element is $\frac{Gm\rho\cos(180^\circ-\theta)\delta\rho\delta\phi}{r^2}$ and when integrated around the annulus this becomes $-\frac{2\pi Gm\cos\theta\,\rho\delta\rho}{r^2}$. We now have to integrate this from $\rho=0$ to a, or, better, from $\theta=0$ to α. We have $r=h\sec\theta$ and $\rho=h\tan\theta$, and the integral becomes

$$-2\pi Gm\int_0^\alpha\sin\theta d\theta=-2\pi Gm(1-\cos\alpha). \tag{5.6.2}$$

There are two ends, so the total flux through the entire cylinder is twice this plus Equation 5.6.1 to give

$$\phi=-4\pi Gm, \tag{5.6.3}$$

as expected from Gauss's theorem.

Example 5.6.2

In figure V.21 I have drawn (part of) an infinite rod whose mass per unit length is λ. I have drawn around it a sphere of radius a. The problem will be to determine the total normal flux through the sphere. From Gauss's theorem, we know that the answer must be $-8\pi G\alpha\lambda$.

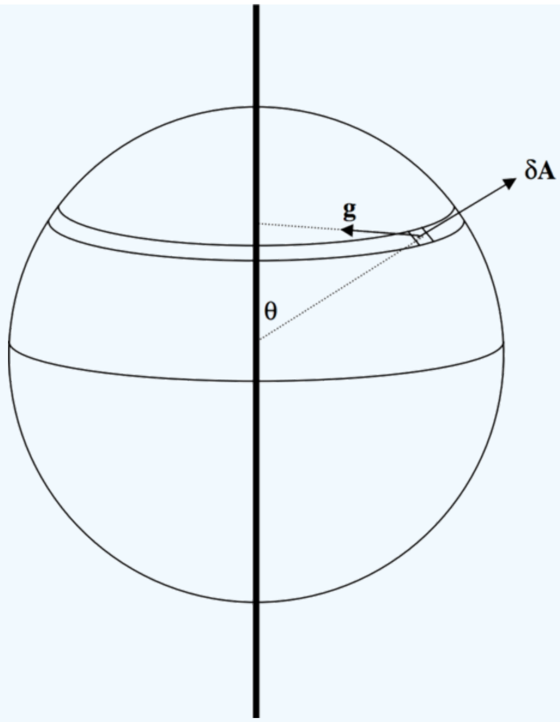


FIGURE V.21

The vector $\delta \mathbf{A}$ representing the element of area is directed away from the centre of the sphere, and the vector $\mathbf{g}$ is directed towards the nearest point of the rod. The angle between them is $\theta+90^{\circ}$. The magnitude of $\delta \mathbf{A}$ in spherical coordinates is $a^{2} \sin \theta \delta \theta \delta \phi$, and the magnitude of $\mathbf{g}$ is (see Equation 5.4.15) $\frac{2 G \lambda}{a \sin \theta}$. The dot product $\mathbf{g} \cdot \delta \mathbf{A}$ is

$$\frac{2 G \lambda}{a \sin \theta} \cdot a^{2} \sin \theta \delta \theta \delta \phi \cdot \cos \left(\theta+90^{\circ}\right)=-2 G \lambda a \sin \theta \delta \theta \delta \phi. \tag{5.6.4}$$

To find the total flux, this must be integrated from $\phi=0$ to 2π and from $\theta=0$ to π. The result, as expected, is $-8 \pi G \alpha \lambda$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.7: Potential

If work is required to move a mass from point A to point B, there is said to be a gravitational *potential difference* between A and B, with B being at the higher potential. The work required to move unit mass from A to B is called the *potential difference* between A and B. In SI units it is expressed in J kg^{-1}.

We have defined only the potential *difference* between *two points*. If we wish to define *the potential at a point*, it is necessary arbitrarily to define the potential at a particular point to be zero. We might, for example define the potential at floor level to be zero, in which case the potential at a height h above the floor is *gh*; equally we may elect to define the potential at the level of the laboratory bench top to be zero, in which case the potential at a height z above the bench top is gz. Because the value of the potential at a point depends on where we define the zero of potential, one often sees that the potential at some point is equal to some mathematical expression *plus an arbitrary constant*. The value of the constant will be determined once we have decided where we wish to define zero potential.

In celestial mechanics it is usual to assign zero potential to all points at an *infinite distance* from any bodies of interest.

Suppose we decide to define the potential at point A to be zero, and that the potential at B is then ψ J kg^{-1}. If we move a point mass m from A to B, we shall have to do an amount of work equal to $m\psi$ J. The *potential energy* of the mass m when it is at B is then $m\psi$. In these notes, I shall usually use the symbol ψ for the potential at a point, and the symbol V for the potential energy of a mass at a point.

In moving a point mass from A to B, it does not matter what *route* is taken. All that matters is the potential difference between A and B. Forces that have the property that the work required to move from one point to another is route-independent are called *conservative forces*; gravitational forces are conservative. The potential at a point is a *scalar* quantity; it has no particular direction associated with it.

If it requires work to move a body from point A to point B (i.e. if there is a potential difference between A and B, and B is at a higher potential than A), this implies that there must be a gravitational field directed from B to A.

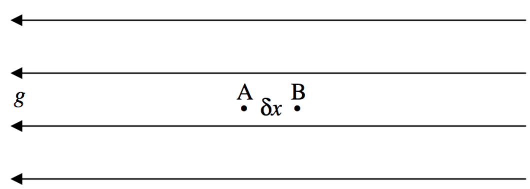


FIGURE V.22

Figure $\text{V}.22$ shows two points, A and B, a distance δx apart, in a region of space where the gravitational field is g directed in the *negative* x direction. We'll suppose that the potential difference between A and B is $\delta\psi$. By definition, the work required to move unit mass from A to B is $\delta\psi$. Also by definition, the force on unit mass is g, so that the work done on unit mass is $g\delta x$. Thus we have

$$g = -\frac{d\psi}{dx}. \tag{5.7.1}$$

The minus sign indicates that, while the potential increases from left to right, the gravitational field is directed to the left. In words, the gravitational field is minus the potential gradient.

This was a one-dimensional example. In a later section, when we discuss the vector operator ∇, we shall write Equation 5.7.1 in its three-dimensional form

$$\mathbf{g} = -\mathbf{grad}\psi = -\nabla\psi. \tag{5.7.2}$$

While ψ itself is a scalar quantity, having no directional properties, its *gradient* is, of course, a vector.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8: The Gravitational Potentials Near Various Bodies

Because potential is a scalar rather than a vector, potentials are usually easier to calculate than field strengths. Indeed, in order to calculate the gravitational field, it is sometimes easier first to calculate the potential and then to calculate the gradient of the potential.

Topic hierarchy

5.8.1: Potential Near a Point Mass

We shall define the potential to be zero at infinity. If we are in the vicinity of a point mass, we shall always have to *do work* in moving a test particle *away from* the mass. We shan't reach zero potential until we are an infinite distance away. It follows that the potential at any finite distance from a point mass is *negative*. The potential at a point is the work required to move unit mass *from infinity to the point*; i.e., it is negative.

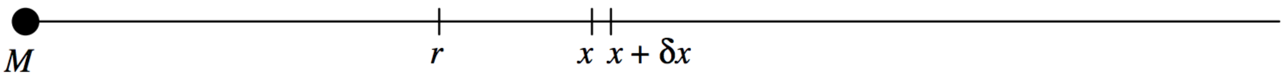


FIGURE V.23

The magnitude of the field at a distance x from a point mass M (figure V.23) is GM/x^2, and the force on a mass m placed there would be GMm/x^2. The work required to move m from x to $x+\delta x$ is $GMm\delta x/x^2$. The work required to move it from r to infinity is

$$GMm\int_r^\infty \frac{dx}{x^2} = \frac{GMm}{r}. \tag{5.8.1.1}$$

The work required to move *unit* mass *from* ∞ to r, which is the potential at r is

$$\psi = -\frac{GM}{r}. \tag{5.8.1}$$

The *mutual potential energy of two point masses a distance r apart*, which is the work required to bring them to a distance r from an infinite initial separation, is

$$V = -\frac{GMm}{r}. \tag{5.8.2}$$

I here summarize a number of similar-looking formulas, although there is, of course, not the slightest possibility of confusing them. Here goes:

Force between two masses:

$$F = \frac{GMm}{r^2}. \quad \text{N} \tag{5.8.3}$$

Field near a point mass:

$$g = \frac{GM}{r^2}, \quad \text{N kg}^{-1} \textit{ or } \text{m s}^{-2} \tag{5.8.4}$$

which can be written in vector form as:

$$\mathbf{g} = -\frac{GM}{r^2}\hat{\mathbf{r}} \quad \text{N kg}^{-1} \textit{ or } \text{m s}^{-2} \tag{5.8.5}$$

or as:

$$\mathbf{g} = -\frac{GM}{r^3}\mathbf{r}. \quad \text{N kg}^{-1} \textit{ or } \text{m s}^{-2} \tag{5.8.6}$$

Mutual potential energy of two masses:

$$V = -\frac{GMm}{r}. \quad \text{J} \tag{5.8.7}$$

Potential near a point mass:

$$\psi = -\frac{GM}{r}. \quad \text{J kg}^{-1} \tag{5.8.8}$$

I hope that's crystal clear.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.2: Potential on the Axis of a Ring

We can refer to figure V.1. The potential at P from the element δM is $-\frac{G\delta M}{(a^2+z^2)^{1/2}}$. This is the same for all such elements around the circumference of the ring, and the total potential is just the scalar sum of the contributions from all the elements. Therefore the total potential on the axis of the ring is:

$$\psi = -\frac{GM}{(a^2+z^2)^{1/2}}. \tag{5.8.9}$$

The z-component of the field (its only component) is $-d/dz$ of this, which results in $g = -\frac{GMz}{(a^2+z^2)^{3/2}}$. This is the same as Equation 5.4.1 except for sign. When we derived Equation 5.4.1 we were concerned only with the magnitude of the field. Here $-d\psi/dz$ gives the z-component of the field, and the minus sign correctly indicates that the field is directed in the negative z-direction. Indeed, since potential, being a scalar quantity, is easier to work out than field, the easiest way to calculate a field is first to calculate the potential and then differentiate it. On the other hand, sometimes it is easy to calculate a field from Gauss's theorem, and then calculate the potential by integration. It is nice to have so many easy ways of doing physics!

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.3: Plane Discs

Refer to figure $\mathrm{V}.2\mathrm{A}$. The potential at P from the elemental disc is

$$d\psi=-\frac{G\delta M}{\left(r^2+z^2\right)^{1/2}}=-\frac{2\pi G\sigma r\delta r}{\left(r^2+z^2\right)^{1/2}}. \tag{5.8.10}$$

The potential from the whole disc is therefore

$$\psi=-2\pi G\sigma\int_0^a\frac{rdr}{\left(r^2+z^2\right)^{1/2}}. \tag{5.8.11}$$

The integral is trivial after a brilliant substitution such as $X=r^2+z^2$ or $r=z\tan\theta$, and we arrive at

$$\psi=-2\pi G\sigma\left(\sqrt{z^2+a^2}-z\right). \tag{5.8.12}$$

This increases to zero as $z\rightarrow\infty$. We can also write this as

$$\psi=-\frac{2\pi Gm}{\pi a^2}\cdot\left[z\left(1+\frac{a^2}{z^2}\right)^{1/2}-z\right], \tag{5.8.13}$$

and, if you expand this binomially, you see that for large z it becomes, as expected, $-Gm/z$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.4: Infinite Plane Lamina

The field above an infinite uniform plane lamina of surface density σ is $-2\pi G\sigma$. Let A be a point at a distance a from the lamina and B be a point at a distance b from the lamina (with $b > a$), the potential difference between B and A is

$$\psi_{\mathrm{B}} - \psi_{\mathrm{A}} = 2\pi G\sigma(b-a). \tag{5.8.14}$$

If we elect to call the potential zero at the surface of the lamina, then, at a distance h from the lamina, the potential will be $+2\pi G\sigma h$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.5: Hollow Hemisphere

Any element of mass, δM on the surface of a hemisphere of radius a is at a distance a from the centre of the hemisphere, and therefore the potential due to this element is merely $-G\delta M/a$. Since potential is a scalar quantity, the potential of the entire hemisphere is just $-GM/a$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.6: Rods

Refer to figure V.5. The potential at P due to the element δx is $-\frac{G\lambda\delta x}{r} = -G\lambda \sec\theta\delta\theta$. The total potential at P is therefore

$$\psi = -G\lambda \int_{\alpha}^{\beta} \sec\theta d\theta = -G\lambda \ln\left[\frac{\sec\beta + \tan\beta}{\sec\alpha + \tan\alpha}\right]. \tag{5.8.15}$$

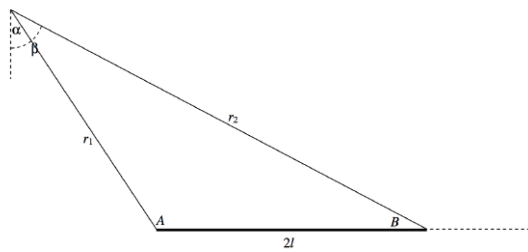


FIGURE V.24

Refer now to figure V.24, in which $A = 90^\circ + \alpha$ and $B = 90^\circ - \beta$.

$$\frac{\sec\beta + \tan\beta}{\sec\alpha + \tan\alpha} = \frac{\cos\alpha(1+\sin\beta)}{\cos\beta(1+\sin\alpha)} = \frac{\sin A(1+\cos B)}{\sin B(1-\cos A)} = \frac{2\sin\frac{1}{2}A\cos\frac{1}{2}A.\, 2\cos^2\frac{1}{2}B}{2\sin\frac{1}{2}B\cos\frac{1}{2}B.\, 2\sin^2\frac{1}{2}A} = \cot\frac{1}{2}A\cot\frac{1}{2}B \tag{5.8.6.1}$$

$$= \sqrt{\frac{s(s-r_2)}{(s-r_1)(s-2l)}} \cdot \sqrt{\frac{s(s-r_1)}{(s-2l)(s-r_2)}},$$

where $s = \frac{1}{2}(r_1 + r_2 + 2l)$. (You may want to refer here to the formulas on pp. 37 and 38 of Chapter 2.)

Hence

$$\psi = -G\lambda \ln\left[\frac{r_1 + r_2 + 2l}{r_1 + r_2 - 2l}\right]. \tag{5.8.16}$$

If r_1 and r_2 are very large compared with l, they are nearly equal, so let's put $r_1 + r_2 = 2r$ and write Equation 5.8.17 as

$$\psi = -\frac{Gm}{2l}\ln\left[\frac{2r\left(1+\frac{2l}{2r}\right)}{2r\left(1-\frac{2l}{2r}\right)}\right] = -\frac{Gm}{2l}\left[\ln\left(1+\frac{l}{r}\right) - \ln\left(1-\frac{l}{r}\right)\right]. \tag{5.8.6.2}$$

Maclaurin expand the logarithms, and you will see that, at large distances from the rod, the potential is, expected, $-Gm/r$.

Let us return to the near vicinity of the rod and to Equation 5.8.16. We see that if we move around the rod in such a manner that we keep $r_1 + r_2$ constant and equal to $2a$, say − that is to say if we move around the rod in an *ellipse* (see our definition of an ellipse in Chapter 2, Section 2.3) − the potential is constant. In other words the equipotentials are confocal ellipses, with the foci at the ends of the rod. Equation 5.8.16 can be written

$$\psi = -G\lambda \ln\left(\frac{a+l}{a-l}\right). \tag{5.8.17}$$

For a given potential ψ, the equipotential is an ellipse of major axis

$$2a = 2l\left(\frac{e^{\psi/(G\lambda)} + 1}{e^{\psi/(G\lambda)} - 1}\right), \tag{5.8.20}$$

where $2l$ is the length of the rod. This knowledge is useful if you are exploring space and you encounter an alien spacecraft or an asteroid in the form of a uniform rod of length $2l$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.7: Solid Cylinder

Refer to figure V.8. The potential from the elemental disc is

$$d\psi = -2\pi G\rho\delta z\left[\left(z^2+a^2\right)^{1/2}-z\right] \tag{5.8.21}$$

and therefore the potential from the entire cylinder is

$$\psi = const. -2\pi G\rho\left[\int_h^{h+l}\left(z^2+a^2\right)^{1/2}dz - \int_h^{h+1} zdz\right]. \tag{5.8.22}$$

I leave it to the reader to carry out this integration and obtain a final expression. One way to deal with the first integral might be to try $z = a\tan\theta$. This may lead to $\int\sec^3\theta d\theta$. From there, you could try something like $\int\sec^3\theta = \int\sec\theta d\tan\theta = \sec\theta\tan\theta - \int\tan\theta d\sec\theta = \sec\theta\tan\theta - \int\sec\theta\tan^2\theta d\theta = \sec\theta\tan\theta - \int\sec^3\theta + \int\sec\theta d\theta$, and so on.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.8: Hollow Spherical Shell

Outside the sphere, the field and the potential are just as if all the mass were concentrated at a point in the centre. The potential, then, outside the sphere, is just $-GM/r$. *Inside* the sphere, the field is zero and therefore the potential is uniform and is equal to the potential at the surface, which is $-GM/a$. The reader should draw a graph of the potential as a function of distance from centre of the sphere. There is a discontinuity in the slope of the potential (and hence in the field) at the surface.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.8.9: Solid Sphere

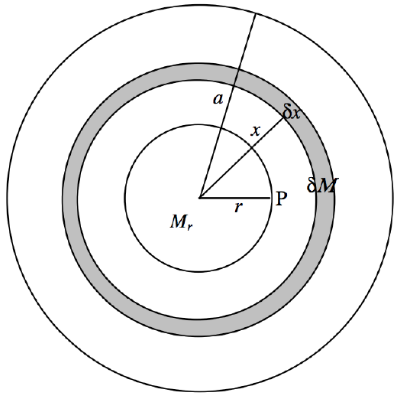


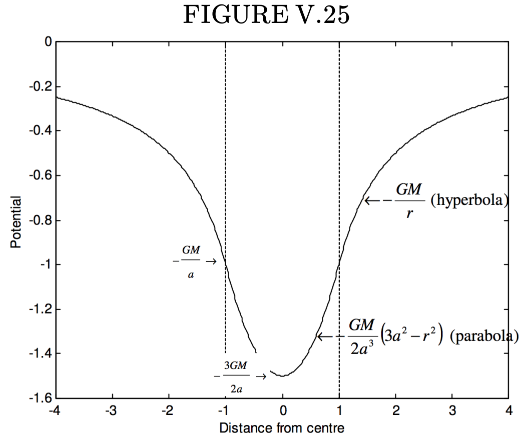
FIGURE V.24A

The potential *outside* a solid sphere is just the same as if all the mass were concentrated at a point in the centre. This is so, even if the density is not uniform, and long as it is spherically distributed. We are going to find the potential at a point P inside a uniform sphere of radius a, mass M, density ρ, at a distance r from the centre ($r < a$). We can do this in two parts. First, there is the potential from that part of the sphere "below" P. This is $-GM_r/r$, where $M_r = \frac{r^3 M}{a^3}$ is the mass within radius r. Now we need to deal with the material "above" P. Consider a spherical shell of radii x, $x + \delta x$. Its mass is $\delta M = \frac{4\pi x^2 \delta x}{\frac{4}{3}\pi a^3} \cdot M = \frac{3Mx^2 \delta x}{a^3}$. The potential from this shell is $-G\delta M/x = -\frac{3GMx\delta x}{a^3}$. This is to be integrated from $x = 0$ to a, and we must then add the contribution from the material "below" P. The final result is

$$\psi = -\frac{GM}{2a^3}\left(3a^2 - r^2\right). \tag{5.8.23}$$

Figure $\text{V}.25$ shows the potential both inside and outside a uniform solid sphere. The potential is in units of $-GM/r$, and distance is in units of a, the radius of the sphere.

FIGURE V.25



Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.9: Work Required to Assemble a Uniform Sphere

Let us imagine a uniform solid sphere of mass M, *density* ρ and radius a. In this section we ask ourselves, how much work was done in order to assemble together all the atoms that make up the sphere if the atoms were initially all separated from each other by an infinite distance? Well, since massive bodies (such as atoms) attract each other by gravitational forces, they will naturally eventually congregate together, so in fact you would have to do work in dis-assembling the sphere and removing all the atoms to an infinite separation. To bring the atoms together from an infinite separation, the amount of work that you do is *negative*.

Let us suppose that we are part way through the process of building our sphere and that, at present, it is of radius r and of mass $M_r = \frac{4}{3}\pi r^3\rho$. The potential at its surface is

$$-\frac{GM_r}{r} = -\frac{G}{r}\cdot\frac{4\pi r^3\rho}{3} = -\frac{4}{3}\pi G\rho r^2. \tag{5.9.1}$$

The amount of work required to add a layer of thickness δr and mass $4\pi\rho r^2\delta r$ to this is

$$-\frac{4}{3}\pi G\rho r^2\times 4\pi r^2\rho\delta r = -\frac{16}{3}\pi^2 G\rho^2 r^4\delta r. \tag{5.9.2}$$

The work done in assembling the entire sphere is the integral of this from $r = 0$ to a, which is

$$-\frac{16\pi^2 G\rho^2 a^5}{15} = -\frac{3GM^2}{5a}. \tag{5.9.1}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.10: Nabla, Gradient and Divergence

We are going to meet, in this section, the symbol ∇. In North America it is generally pronounced “del”, although in the United Kingdom and elsewhere one sometimes hears the alternative pronunciation “nabla”, called after an ancient Assyrian harp-like instrument of approximately that shape.

In section 5.7, particularly Equation 5.7.1, we introduced the idea that the gravitational field g is minus the gradient of the potential, and we wrote $g=-d\psi/dx$. This Equation refers to an essentially one-dimensional situation. In real life, the gravitational potential is a three dimensional scalar function $\psi(x,y,z)$, which varies from point to point, and its *gradient* is

$$\mathbf{grad}\psi=\mathbf{i}\frac{\partial\psi}{\partial x}+\mathbf{j}\frac{\partial\psi}{\partial y}+\mathbf{k}\frac{\partial\psi}{\partial x}, \tag{5.10.1}$$

which is a vector field whose magnitude and direction vary from point to point. The gravitational field, then, is given by

$$\mathbf{g}=-\mathbf{grad}\psi. \tag{5.10.2}$$

Here, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$ are the unit vectors in the x-, y- and z-directions.

The operator ∇ is $\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial x}$, so that Equation 5.10.2 can be written

$$\mathbf{g}=-\nabla\psi. \tag{5.10.3}$$

I suppose one could write a long book about ∇, but I am going to try to restrict myself in this section to some bare essentials.

Let us suppose that we have some vector field, which we might as well suppose to be a gravitational field, so I’ll call it $\mathbf{g}$. (If you don’t want to be restricted to a gravitational field, just call the field $\mathbf{A}$ as some sort of undefined or general vector field.) We can calculate the quantity

$$\nabla\cdot\mathbf{g}=\left(\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial x}+\mathbf{k}\frac{\partial}{\partial x}\right)\cdot(\mathbf{i}g_x+\mathbf{j}g_y+\mathbf{k}g_z). \tag{5.10.4}$$

When this is multiplied out, we obtain a *scalar* field called the *divergence* of $\mathbf{g}$:

$$\nabla\cdot\mathbf{g}=\text{div}\mathbf{g}=\frac{\partial g_x}{\partial x}+\frac{\partial g_y}{\partial y}+\frac{\partial g_z}{\partial z}. \tag{5.10.5}$$

Is this of any use?

Here’s an example of a possible useful application. Let us imagine that we have some field $\mathbf{g}$ which varies in magnitude and direction through some volume of space. Each of the components, g_x, g_y, g_z can be written as functions of the coordinates. Now suppose that we want to calculate the surface integral of g through the closed boundary of the volume of space in question. Can you just imagine what a headache that might be? For example, suppose that $g=x^2\mathbf{i}-xy\mathbf{j}-xz\mathbf{k}$, and I were to ask you to calculate the surface integral over the surface of the ellipsoid $\frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1$. It would be hard to know where to begin.

Well, there is a theorem, which I am not going to derive here, but which can be found in many books on mathematical physics, and is not particularly difficult, which says:

The surface integral of a vector field over a closed surface is equal to the volume integral of its divergence.

In symbols:

$$\int\int\mathbf{g}\cdot d\mathbf{A}=\int\int\int div\mathbf{g}dV. \tag{5.10.6}$$

If we know g_x, g_y and g_z as functions of the coordinates, then it is often very simple and straightforward to calculate the divergence of $\mathbf{g}$, which is a scalar function, and it is then often equally straightforward to calculate the volume integral. The example I gave in the previous paragraph is trivially simple (it is a rather artificial example, designed to be ridiculously simple) and you will readily find that div $\mathbf{g}$ is everywhere zero, and so the surface integral over the ellipsoid is zero.

If we combine this very general theorem with Gauss's theorem (which applies to an inverse square field), which is that the surface integral of the field over a closed volume is equal to $-4\pi G$ times the enclosed mass (Equation 5.5.1) we understand immediately that the divergence of $\mathbf{g}$ at any point is related to the density at that point and indeed that

$$\text{div }\mathbf{g} = \nabla\cdot\mathbf{g} = -4\pi G\rho. \tag{5.10.7}$$

This may help to give a bit more physical meaning to the divergence. At a point in space where the local density is zero, div $\mathbf{g}$, of course, is also zero.

Now Equation 5.10.2tells us that $\mathbf{g} = -\nabla\psi$, so that we also have

$$\nabla\cdot(-\nabla\psi) = -\nabla\cdot(\nabla\psi) = -4\pi G\rho. \tag{5.10.8}$$

If you write out the expressions for ∇ and for $\nabla\psi$ in full and calculate the dot product, you will find that $\nabla\cdot(\nabla\psi)$, which is also written $\nabla^2\psi$, is $\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{dy^2}+\frac{\partial^2\psi}{\partial z^2}$. Thus we obtain

$$\nabla^2\psi = \frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2} = 4\pi G\rho. \tag{5.10.9}$$

This is *Poisson's Equation*. At any point in space where the local density is zero, it becomes

$$\nabla^2\psi = 0 \tag{5.10.10}$$

which is *Laplace's Equation*. Thus, no matter how complicated the distribution of mass, the potential as a function of the coordinates must satisfy these Equations.

We leave this topic here. Further details are to be found in books on mathematical physics; our aim here was just to obtain some feeling for the physical meaning. I add just a few small comments. One is, yes, it is certainly possible to operate on a vector field with the operator $\nabla\times$. Thus, if $\mathbf{A}$ is a vector field, $\nabla\times\mathbf{A}$ is called the $\mathbf{curl}$ of $\mathbf{A}$. The $\mathbf{curl}$ of a gravitational field is zero, and so there is no need for much discussion of it in a chapter on gravitational fields. If, however, you have occasion to study fluid dynamics or electromagnetism, you will need to become very familiar with it. I particularly draw your attention to a theorem that says

*The line integral of a vector field around a closed plane circuit is equal to the surface integral of its **curl**.*

This will enable you easily to calculate two-dimensional line integrals in a similar manner to that in which the divergence theorem enables you to calculate threedimensional surface integrals.

Another comment is that very often calculations are done in spherical rather than rectangular coordinates. The formulas for $\mathbf{grad}$, div, $\mathbf{curl}$ and ∇^2 are then rather more complicated than their simple forms in rectangular coordinates.

Finally, there are dozens and dozens of formulas relating to nabla in the books, such as "$\mathbf{curl}\,\mathbf{curl} = \mathbf{grad}$ div minus nabla-squared". While they should certainly never be memorized, they are certainly worth becoming familiar with, even if we do not need them immediately here.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.11: Legendre Polynomials

In this section we cover just enough about Legendre polynomials to be useful in the following section. Before starting, I want you to expand the following expression, by the binomial theorem, for $|x| < 1$, up to x^4 :

$$\frac{1}{(1-2x\cos\theta+x^2)^{1/2}}. \tag{5.11.1}$$

Please do go ahead and do it. Well, you probably won't, so I'd better do it myself:

I'll start with

$$(1-X)^{-1/2} = 1+\frac{1}{2}X+\frac{3}{8}X^2+\frac{5}{16}X^3+\frac{35}{128}X^4\dots \tag{5.11.2}$$

and therefore

$$[1-x(2\cos\theta-x)]^{-1/2} = 1+\frac{1}{2}x(2\cos\theta-x)+\frac{3}{8}x^2(2\cos\theta-x)^2+\frac{5}{16}x^3(2\cos\theta-x)^3+\frac{35}{128}x^4(2\cos\theta-x)^4\dots \tag{5.11.3}$$

$$= 1+x\cos\theta-\frac{1}{2}x^2+\frac{3}{8}x^2(4\cos^2\theta-4x\cos\theta+x^2)+\frac{5}{16}x^3(8\cos^3\theta-12x\cos^2\theta+6x^2\cos\theta-x^3) + \frac{35}{128}x(16\cos^4\theta-32x\cos^3\theta+24x^2\cos^2\theta-8x^3\cos\theta+x^4)\dots \tag{5.11.4}$$

$$= 1+x\cos\theta+x^2(-\frac{1}{2}+\frac{3}{2}\cos^2\theta)+x^3(-\frac{3}{2}\cos\theta+\frac{5}{2}\cos^3\theta)+x^4(\frac{3}{8}-\frac{15}{4}\cos^2\theta+\frac{35}{8}\cos^4\theta)\dots \tag{5.11.5}$$

The coefficients of the powers of x are the *Legendre polynomials* $P_l(\cos\theta)$, so that

$$\frac{1}{(1-2x\cos\theta+x^2)^{1/2}} = 1+xP_1(\cos\theta)+x^2P_2(\cos\theta)+x^3P_3(\cos\theta)+x^4P_4(\cos\theta)+\dots \tag{5.11.6}$$

The Legendre polynomials with argument $\cos\theta$ can be written as series of terms in powers of $\cos\theta$ by substitution of $\cos\theta$ for x in Equations 1.12.5 in Section 1.12 of Chapter 1. Note that x in Section 1 is not the same as x in the present section. Alternatively they can be written as series of cosines of multiples of θ as follows.

$$\begin{aligned}
P_0 &= 1\\
P_1 &= \cos\theta\\
P_2 &= \frac{1}{4}(3\cos 2\theta+1)\\
P_3 &= \frac{1}{8}(5\cos 3\theta+3\cos\theta)\\
P_4 &= \frac{1}{64}(35\cos 4\theta+20\cos 2\theta+9)\\
P_5 &= \frac{1}{128}(63\cos 5\theta+35\cos 3\theta+30\cos\theta)\\
P_6 &= \frac{1}{512}(231\cos 6\theta+126\cos 4\theta+105\cos 2\theta+50)\\
P_7 &= \frac{1}{1024}(429\cos 7\theta+231\cos 5\theta+189\cos 3\theta+175\cos\theta)\\
P_8 &= (6435\cos 8\theta+3432\cos 6\theta+2772\cos 4\theta+2520\cos 2\theta+1225)/2^{14}
\end{aligned} \tag{5.11.7}$$

For example, $P_6(\cos\theta)$ can be written either as given by Equation 5.11.7, or as given by Equation 1, namely

$$P_6 = \frac{1}{16}(231c^6-315c^4+105c^2-5),\ \text{where } c=\cos\theta. \tag{5.11.8}$$

The former may look neater, and the latter may look "awkward" because of all the powers. However, the latter is far faster to compute, particularly when written as nested parentheses:

$$P_6 = (-5+C(105+C(-315+231C)))/16,\ \text{where } C=\cos^2\theta. \tag{5.11.9}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.12: Gravitational Potential of any Massive Body

You might just want to look at **Chapter 2** of Classical Mechanics (Moments of Inertia) before proceeding further with this chapter.

In figure VIII.26 I draw a massive body whose centre of mass is C, and an external point P at a distance R from C. I draw a set of $Cxyz$ axes, such that P is on the z-axis, the coordinates of P being $(0,0,z)$. I indicate an element δm of mass, distant r from C and l from P. I'll suppose that the density at δm is ρ and the volume of the mass element is $\delta\tau$, so that $\delta m=\rho\delta\tau$.

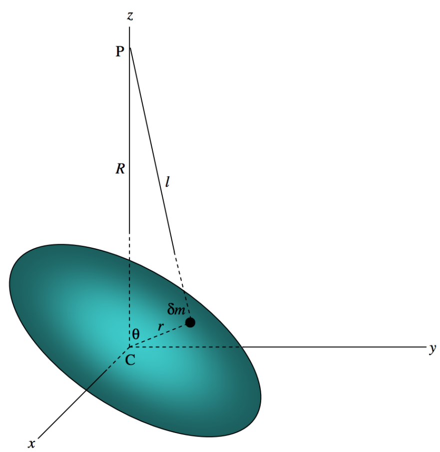


FIGURE V.26

The potential at P is

$$\psi=-G\int\frac{dm}{l}=-G\int\frac{\rho d\tau}{l}. \tag{5.12.1}$$

But $l^2=R^2+r^2-2Rr\cos 2\theta$,

so

$$\psi=-G\left[\frac{1}{R}\int\rho d\tau+\frac{1}{R^2}\int\rho r\cos\theta d\tau+\frac{1}{R^3}\int\rho r^2P_2(\cos\theta)d\tau+\frac{1}{R^4}\int\rho r^3P_3(\cos\theta)d\tau\ldots\right]. \tag{5.12.2}$$

The integral is to be taken over the entire body, so that $\int\rho d\tau=M$, where M is the mass of the body. Also $\int\rho r\cos\theta d\tau=\int zdm$, which is zero, since C is the centre of mass. The third term is

$$\frac{1}{2R^3}\int\rho r^2(3\cos^2\theta-1)d\tau=\frac{1}{2R^3}\int\rho r^2(2-3\sin^2\theta)d\tau. \tag{5.12.3}$$

Now

$$\int 2\rho r^2d\tau=\int 2r^2dm=\int\left[(y^2+z^2)+(z^2+x^2)+(x^2+y^2)\right]dm=A+B+C \tag{5.12.1}$$

where A, B and C are the second moments of inertia with respect to the axes Cx, Cy, Cz respectively. But $A+B+C$ is invariant with respect to rotation of axes, so it is also equal to $A_0+B_0+C_0$, where A_0, B_0, C_0 are the *principal moments of inertia*.

Lastly, $\int\rho r^2\sin^2\theta d\tau$ is equal to C, the moment of inertia with respect to the axis Cz.

Thus, if R is sufficiently larger than r so that we can neglect terms of order $(r/R)^3$ and higher, we obtain

$$\psi=-\frac{GM(2MR^2+A_0+B_0+C_0-3C)}{2R^3}. \tag{5.12.4}$$

In the special case of an *oblate symmetric top*, in which $A_0 = B_0 < C_0$, and the line CP makes an angle γ with the principal axis, we have

$$C = A_0 + (C_0 - A_0)\cos^2\gamma = A_0 + (C_0 - A_0)Z^2/R^2, \tag{5.12.5}$$

so that

$$\psi = -\frac{G}{R}\left[M + \frac{C_0 - A_0}{2R^2}\left(1 - \frac{3Z^2}{R^2}\right)\right]. \tag{5.12.6}$$

Now consider a uniform oblate spheroid of polar and equatorial diameters $2c$ and $2a$ respectively. It is easy to show that

$$C_0 = \frac{2}{5}Ma^2. \tag{5.12.7}$$

Exercise 5.12.1

Confirm Equation 5.12.7.

It is slightly less easy to show (*Exercise*: Show it.) that

$$A_0 = \frac{1}{5}M\left(a^2 + c^2\right). \tag{5.12.8}$$

For a symmetric top, the integrals of the odd polynomials of Equation 5.12.2 are zero, and the potential is generally written in the form

$$\psi = -\frac{GM}{R}\left[1 + \left(\frac{a}{R}\right)^2 J_2P_2(\cos\gamma) + \left(\frac{a}{R}\right) J_4P_4(\cos\gamma)\ldots\right] \tag{5.12.9}$$

Here γ is the angle between CP and the principal axis. For a uniform oblate spheroid, $J_2 = \frac{C_0 - A_0}{Mc^2}$. This result will be useful in a later chapter when we discuss precession.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

5.13: Pressure at the Centre of a Uniform Sphere

What is the pressure at the centre of a sphere of radius a and of uniform density ρ?

(Preliminary thought: Show by dimensional analysis that it must be something times $G\rho^2a^2$.)

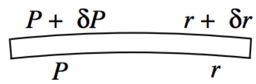


FIGURE V.27

Consider a portion of the sphere between radii r and $r+\delta r$ and cross-sectional area A. Its volume is $A\delta r$ and its mass is $\rho A\delta r$. (Were the density not uniform throughout the sphere, we would here have to write $\rho(r)A\delta r$.) Its weight is $\rho gA\delta r$, where $g = GM_r/r^2 = \frac{4}{3}\pi G\rho r$. We suppose that the pressure at radius r is P and the pressure at radius $r+\delta r$ is $P+\delta P$. (δP is negative.) Equating the downward forces to the upward force, we have

$$A(P+\delta P)+\frac{4}{3}\pi AG\rho^2 r\delta r = AP. \tag{5.13.1}$$

That is:

$$\delta P = -\frac{4}{3}\pi G\rho^2 r\delta r. \tag{5.13.2}$$

Integrate from the centre to the surface:

$$\int_{P_0}^{0} dP = -\frac{4}{3}\pi G\rho^2\int_0^a r dr. \tag{5.13.3}$$

Thus:

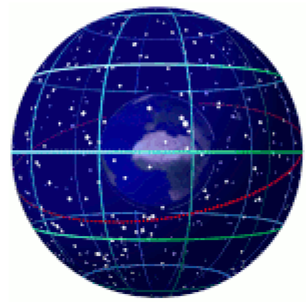
$$P = \frac{2}{3}\pi G\rho^2 a^2. \tag{5.13.4}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

6: THE CELESTIAL SPHERE



If you look up in the sky, it appears as if you are at the centre of a vast crystal sphere with the stars fixed on its surface. This sphere is the celestial sphere. It has no particular radius; we record positions of the stars merely by specifying angles. We see only half of the sphere; the remaining half is hidden below the horizon. In this section we describe the several coordinate systems that are used to describe the positions of stars and other bodies on the celestial sphere.

6.9: THE LENGTH OF THE YEAR

The calendar that we use in everyday life is the Gregorian Calendar, in which there are 365 days in most years, but 366 days in years that are divisible by 4 unless they are also divisible by 100 other than those that are also divisible by 400. The Anomalistic Year is the interval between consecutive passages of the Earth through perihelion and is a little longer than the sidereal year.

6.1: Introduction to the Celestial Sphere

If you look up in the sky, it appears as if you are at the centre of a vast crystal sphere with the stars fixed on its surface. This sphere is the *celestial sphere*. It has no particular radius; we record positions of the stars merely by specifying angles. We see only half of the sphere; the remaining half is hidden below the *horizon*. In this section we describe the several coordinate systems that are used to describe the positions of stars and other bodies on the celestial sphere, and how to convert between one system and another. In particular, we describe *altazimuth*, *equatorial* and *ecliptic coordinates* and the relations between them. The relation between ecliptic and equatorial coordinates varies with time owing to the *precession of the equinoxes* and *nutation*, which are also described in this chapter.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.2: Altazimuth Coordinates

In figure $\mathrm{VI.1}$ we see the celestial sphere with the observer O at its centre. The point immediately overhead, Z, is the *zenith*. The point directly underneath, Z', is the *nadir*. The points marked N, E, S are the *north*, *east* and *south points of the horizon*. The west point of the horizon is behind the plane of the paper (or of your computer screen) and is not drawn. The great circle NESW is, of course, the *horizon*.

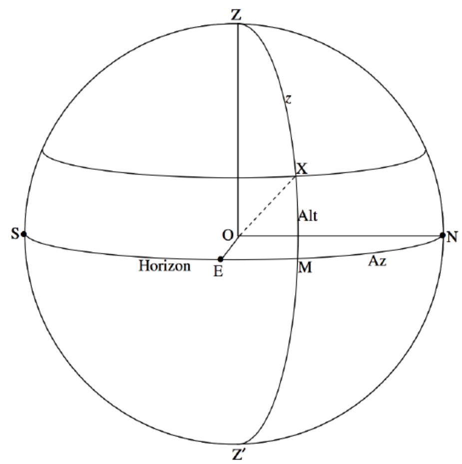


FIGURE VI.1

Any great circle passing through Z and Z' is called a *vertical circle*. The vertical circle passing through S and N, the south and north points of the horizon, is the *meridian*. The vertical circle passing through the east and west points of the horizon (which I have not drawn) is the *prime vertical*. X is the position of a star on the celestial sphere, and I have drawn the vertical circle ZXMZ' passing through the star. The angle MX is the *altitude* of the star (also referred to in some contexts as its "elevation"). The complement of its altitude, the angle z, is the zenith distance (also called, not unreasonably, the "zenith angle").

A small circle of constant altitude – i.e. a small circle parallel to the horizon – has the curious name of an *almucantar*, and I have drawn the almucantar through the star X. An almucantar can also be called a parallel of altitude.

The angle NM that I have denoted by Az on figure $\mathrm{VI.1}$ is called the *azimuth* (or "bearing") of the star. As drawn on the figure, it is measured eastwards from the north point of the horizon. This is perhaps the most common convention for observers in the northern hemisphere. However, for stars that are west of the meridian, it may often be convenient to express azimuth as measured westwards from the north point. I don't know what the custom is of astronomers who live in the southern hemisphere, but it would not surprise me if often they express azimuth as measured from the south point of their horizon. In any case, it is important not to assume that there is some universal convention that will be understood by everybody, and it is *essential* when quoting the azimuth of a star to add a phrase such as "measured from the north point eastwards". If you merely write "an azimuth of 32 degrees", it is almost certain that you will be either misunderstood or not understood at all.

In the altazimuth system of coordinates, the position of a star is uniquely specified by its azimuth and either its altitude or its zenith distance.

Of course the altitude and azimuth of a star are changing continuously all the time, and they are also different for all observers at different geographical locations.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.3: Equatorial Coordinates

If you live in the northern hemisphere and if you face north, you will observe that the entire celestial sphere is rotating slowly counterclockwise about a point in the sky close to the star Polaris (α Ursae Minoris). The point P about which the sky appears to rotate is the *North Celestial Pole*. If you live in the southern hemisphere and if you face south you will see the entire sky rotating clockwise about a point Q, the *South Celestial Pole*. There is no bright star near the south celestial pole; the star σ Octantis is close to the south celestial pole, but it is only just visible to the unaided eye provided you are dark adapted and if you have a clear sky free of light pollution. The great circle that is 90° from either pole is the *celestial equator*, and it is the projection of Earth's equator on to the celestial sphere.

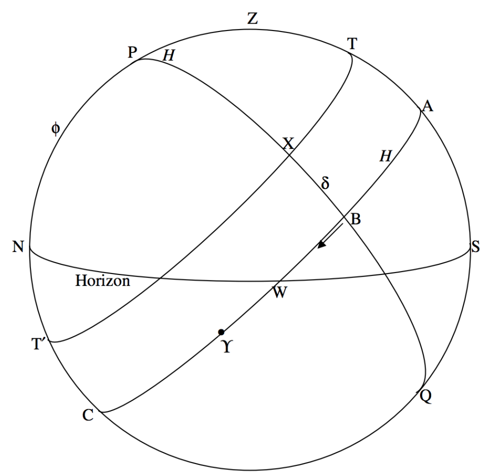


FIGURE VI.2

In figure $\mathrm{VI.2}$ I have drawn the celestial sphere from the opposite side from the drawing of figure $\mathrm{VI.1}$, so that, this time, you can see the west point of the horizon, but not the east point. The celestial equator is the great circle $\mathrm{ABW\Upsilon C}$.

You might possibly have noticed that, in section 2, I had not properly defined the north point of the horizon other than by saying that it was the point marked N in figure VI.1. We see now that the north and south points of the horizon are the points where the vertical circle that passes through the celestial poles (i.e. the meridian) meets the horizon.

The altitude ϕ of the north celestial pole is equal to the geographical north latitude of the observer. Thus for an observer at Earth's north pole, the north celestial pole is at the zenith, and for an observer at Earth's equator, the north celestial pole is on the horizon.

You will see that a star such as X transits across the meridian twice. *Lower meridian transit* occurs at the point $\mathrm{T'}$, when the star is north of the observer and is directly below the north celestial pole. For the star X of figure $\mathrm{VI.2}$, lower meridian transit is also below the horizon, and it cannot be seen. The star reaches its highest point in the sky (i.e. it *culminates*) at upper meridian transit.

The first quantitative astronomical observation I ever did was to see how long the celestial sphere takes to rotate through 360°. This is best done by timing the interval between two consecutive upper meridian transits of a star. It will be found that this interval is $23^\mathrm{h}\ 56^\mathrm{m}\ 04^\mathrm{s}.099$ of mean solar time, although of course it requires more than a casual observation to determine the interval to that precision. The rotation of the celestial sphere is, of course, a reflection of the rotation of Earth on its axis. In other words, this interval is the sidereal (i.e. relative to the stars) rotation period of Earth.

We are now in a position to describe the position of a star on the celestial sphere in *equatorial coordinates*. The angle δ in figure $\mathrm{VI.2}$ is called the *declination* of the star. It is usually expressed in degrees, arcminutes and arcseconds, from 0° to $+90^\circ$ for stars on or north of the equator, and from 0° to -90° for stars on or south of the equator. When quoting the declination of a star, the sign of the declination must always be given.

When the star X in figure $\mathrm{VI.2}$ is at lower meridian transit, it is below the horizon and is not visible. However, if the declination of a star is greater than $90^\circ - \phi$, the star will not reach the horizon and it will never set. Such stars are called *circumpolar* stars.

The second coordinate is the angle H in figure $\mathrm{VI.2}$. It is measured westward from the meridian. It will immediately be noticed that, while the declination of a star does not change through the night, its hour angle continuously increases, and also the hour angle of a star at any given time depends on the geographical longitude of the observer. While hour angle could be expressed in either radians or degrees, it is customary to express the hour angle in hours, minutes and seconds of time. Thus hour angle goes from 0^{h} to 24^{h}. When a star has an hour angle of, for example, 3^{h}, it means that it is three sidereal hours since it transited (upper transit) the meridian. Conversion factors are

$$1^{\mathrm{h}} = 15^\circ \quad 1^{\mathrm{m}} = 15' \quad 1^{\mathrm{s}} = 15'' \quad 1^\circ = 4^{\mathrm{m}} \quad 1' = 4^{\mathrm{s}}. \tag{6.3.1}$$

(The reader may have noticed that I have just used the term “sidereal hours”. For the moment, just read this as “hours” – but a little later on we shall say what we mean by “sidereal” hours, and you may then want to come back and re-read this.)

While it is useful to know the hour angle of a star at a particular time for a particular observer, we still need a coordinate that is fixed on the celestial sphere. To do this, we refer to a point on the celestial equator, which I shall define more precisely later on, denoted on figure $\mathrm{VI.2}$ by the symbol Υ. This is the astrological symbol for the sign Aries, and it was originally in the constellation Aries, although at the present time it is in the constellation Pisces. In spite of its present location, it is still called the *First Point of Aries*. The angle measured eastward from Υ to the point B is called the right ascension of the star X, and is denoted by the symbol α. This does not change (at least not very much – but we shall deal with small refinements later) during the night or from night to night. Thus we can describe the position of a star on the celestial sphere by the two coordinates δ, its declination, and α, its right ascension, and since its right ascension does not change (at least not very much), we can list the right ascensions as well as the declinations of the stars in our catalogues. The right ascension of the First Point of Aries is, of course, 0^h.

I have hinted in the last paragraph that the right ascension of a star, although it doesn’t change “very much” during a night, does change quite perceptibly over a year. We shall have to return to this point later. I have not as yet precisely defined where the point Υ is or how it is defined, but we shall later learn that it is not quite fixed on the equator, but it moves slightly in a manner that I shall have to describe in due course. Thus the entire system of equatorial coordinates, and the right ascensions and declinations of the stars, depends on where this mysterious First Point of Aries is. For that reason, it is always necessary to state the epoch to which right ascensions and declinations are referred. For much of the twentieth century, equatorial coordinates were referred to the epoch 1950.0 (strictly it was $\mathrm{B}1950.0$, but I shall have to postpone explaining the meaning of the prefix B). At present catalogues and atlases refer right ascensions and declinations to the epoch $\mathrm{J}2000.0$, where again I shall have to defer an explanation of the prefix J. While there is evidently some further explanation yet to come, suffice it to say at this point that, when giving the right ascension and declination of any object, it is *essential* that the epoch also be given. The First Point of Aries moves very, very slowly westward relative to the stars, so that the right ascensions of all the stars are increasing at a rate of about $0^s.008$ per day. This does not amount to much for day-to-day purposes, but it does emphasize why it is always necessary to state the epoch to which right ascensions and declinations of stars are quoted. It also means that, if you were able to observe two consecutive upper transits of Υ across the meridian, the interval would be $0^s.008$ shorter than the sidereal rotation period of Earth. It would be, in fact, $23^{\mathrm{h}}\ 56^{\mathrm{m}}\ 04^{\mathrm{s}}.091$. This interval between two consecutive upper meridian transits of the First Point of Aries, is called a *sidereal day*. (It might be thought that, since the word “sidereal” implies “relative to the stars”, this is not a particularly good term. I would have sympathy with this view, and would prefer to call the interval an “equinoctial day”. However, the term *sidereal day* is so firmly entrenched that I shall use that term in these notes.) A sidereal day is divided into 24 *sidereal hours*, which are shorter than mean solar hours by a factor of 0.99726957. We shall discuss the motion of Υ in more detail in a later section. At this stage no great harm is done by considering Υ in the first approximation to be fixed relative to the stars.

Now some more words. Small circles parallel to the celestial equator (such as the small circle $\mathrm{T'XT}$ in figure $\mathrm{VI.2}$) are *parallels of declination*. Great circles that pass through the north and south celestial poles (for example the great circle PXBQ of figure $\mathrm{VI.2}$) and which are fixed on and rotate with the celestial sphere are called by a variety of names. Some call them *declination circles*, because you measure declination up and down these circles. Others call them *hour circles*, because the hour angle or right ascension is constant along them. For those who find it confusing that a given circle can be called either a declination circle or an hour circle, you can get around this difficulty by calling them *colures*. The colure that passes through the First Point of Aries and the diametrically opposite point on the celestial sphere, and which therefore has right ascensions

0^h and 12^h, is the equinoctial colure. The colure that is 90° from this (or, rather, 6 hours from this) and which has right ascensions 6^h and 18^h, is the *solstitial colure*.

The time that has elapsed, in sidereal hours, since the First Point of Aries transited (upper transit) the meridian, that is to say the hour angle of the first point of Aries, or the angle from A to Υ in figure $\mathrm{VI.2}$, is called the *Local Sidereal Time*. It is evident from figure $\mathrm{VI.2}$ that the Local Sidereal Time is also equal to $\Upsilon\mathrm{B}+\mathrm{AB}$. But $\Upsilon\mathrm{B}$ is the right ascension of the star X and AB is its hour angle. Therefore *the local sidereal time* (the hour angle of the First Point of Aries) *is equal to the right ascension of any star plus its hour angle*.

The sidereal time at the longitude of Greenwich (0° longitude) is tabulated daily in the *Astronomical Almanac* and the local sidereal time at your location is equal to the local sidereal time at Greenwich minus your geographical longitude. Most observatories have two clocks running in the dome at all times. One gives Universal Time, while the other, which runs a little faster, gives the local sidereal time. But you always have a sidereal clock available, for a glance at figure $\mathrm{VI.2}$ will tell you that the local sidereal time is equal to the right ascension of stars at upper meridian transit.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.4: Conversion Between Equatorial and Altazimuth Coordinates

Whereabouts in the sky will a given star be at a certain time? This as a typical problem involving conversion between equatorial and altazimuth coordinates. We have to solve a spherical triangle. That is no problem – we already know how to do that. The problem is: which triangle?

The first problem, however, arises from the phrase "at a certain time". In particular, if we want to know where a star is, for example, at 2002 November 24, at 10:00 p.m. Pacific Standard Time as seen from Victoria, whose longitude is $123^\circ\ 25'.0\ \mathrm{W}$, we need to know the *local sidereal time* at that instant.

The calculation might go something like this.

From the *Astronomical Almanac* we find that the local sidereal times at Greenwich at 0^h UT on November 25 and 26, 2002, are

November 25: $04^\mathrm{h}\ 16^\mathrm{m} 59^\mathrm{s}$
November 26: $04\quad 20\quad 56$

We want the local sidereal time at November $24^\mathrm{d}\ 22^\mathrm{h}\ 00^\mathrm{m}$ PST
= November $25^\mathrm{d}\ 06^\mathrm{h}\ 00^\mathrm{m}$ UT

By interpolation we find that the local sidereal time at Greenwich at that instant is $10^\mathrm{h}\ 17^\mathrm{m}\ 58^\mathrm{s}$.

The longitude of Victoria is $08^\mathrm{h}\ 13^\mathrm{m}\ 40^\mathrm{s}$, and therefore the local sidereal time at Victoria is $02^\mathrm{h}\ 04^\mathrm{m}\ 18^\mathrm{s}$.

We have overcome the first obstacle, and we now know the local sidereal time (LST).

We'll ask ourselves now what are the altitude and azimuth of a star whose right ascension and declination are α and δ. We also need the latitude of the observer (= altitude of the north celestial pole), which I'll call ϕ. The hour angle H of the star is LST $-\ \alpha$.

The triangle that we have to solve is the triangle PZX. Here P, Z and X are, respectively, the north celestial pole, the zenith and the star. That is, we solve the triangle formed by *the star and the poles of the two coordinate systems* of interest. I draw the celestial sphere in figure VI.3 as seen from the west. I have marked in the hour angle H, the codeclination $90^\circ - \delta$, the altitude ϕ of the pole, the zenith distance z and the azimuth A measured from the north point westwards.

In triangle PZX, we know ϕ, δ and H, so we immediately find the zenith distance z by application of the cosine formula (Equation 3.5.2) and the azimuth A from the cotangent formula (Equation 3.5.5).

Problem. Show that the hour angle H of a star of declination δ when it sets for an observer at latitude ϕ is given by $\cos H = -\tan\delta\tan\phi$. This will enable you now to find the Local Sidereal Time of starset, since $\mathrm{LST} = \text{hour angle plus right ascension}$, and then you can convert to your zone solar time.

Show also that the azimuth A of starset, westward from the north point, is given by $\tan A = -\sin\phi\tan H$.

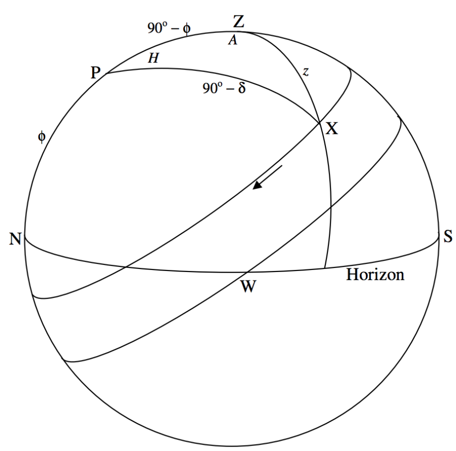


FIGURE VI.3

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.5: Ecliptic Coordinates

In figures $\mathrm{VI}.2$ and 3 we were concerned mainly with the daily rotation of the celestial sphere. In figure $\mathrm{VI}.4$ we shall be concerned mainly with the annual motion of the Sun relative to the stars on the celestial sphere. In contrast to figures $\mathrm{VI}.2$ and 3, I have drawn the celestial equator, not the observer's horizon, horizontally, and the north celestial pole, not the observer's zenith, is at the top of the diagram. It is found that, for an observer on Earth, the Sun moves eastward relative to the stars during the course of the year, its right ascension continuously increasing; this apparent motion of the Sun relative to the stars is, of course, a consequence of the Earth revolving around the Sun.

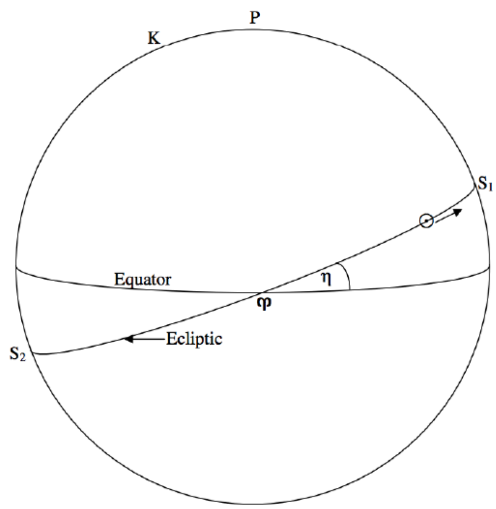


FIGURE VI.4

Relative to the stars, it is found that, during the course of a year, the Sun moves eastward along a great circle that is inclined to the equator at an angle of about $23^\circ.4$. This great circle is called the *ecliptic*, and it is the projection of the plane of Earth's orbit on the celestial sphere. The angle between the ecliptic and the equator is called the *Obliquity of the Ecliptic*. The ecliptic crosses the equator at two points. The Sun reaches one of these points on about March 22 each year on its way north at which time the Sun's declination changes from negative to positive. This point, the ascending node of the Sun's path on the equator, is the *First Point of Aries*, which we introduced in Section 6.3. As mentioned there, and for reasons that will be explained in section 6.7, it is actually in the constellation Pisces rather than Aries. Nevertheless it is still known as the First Point of Aries and is denoted by the astrological symbol Υ for Aries. It is the point from which right ascensions are measured. The instant of time when the Sun crosses the equator from north to south at the First Point of Aries is the *March Equinox*. Days and nights are of equal length all over the world on that date ("equinox" = "equal night"), and that date marks the first day of Spring in the northern hemisphere. For that reason it is also called the "vernal equinox" (Latin verna = "spring") – but that is hardly fair to southern hemisphere astronomers, for it marks the beginning of the southern autumn.

About three months later, on or near June 21, the Sun reaches the point S_1 at the *June Solstice* (called by those who live in the Northern hemisphere, the summer solstice). The declination of the Sun is then at its highest point, $+23.4$ degrees. At that instant the rate of change of the Sun's declination is zero, which explains the origin of the word "solstice", which implies that the Sun is momentarily standing still. The Sun is then in the constellation Gemini. After a further three months, the Sun has descended back to the equator on its way south, at the *September equinox* (the "autumnal equinox" for northerners) on or near September 23, when the Sun is in the constellation Virgo. And after a further three months the Sun reaches its most southerly declination at the *December solstice* ("winter solstice" to northerners) on or near December 21, when the Sun is in the constellation Sagittarius.

Also drawn in figure $\mathrm{VI}.4$ is the *North Pole of the Ecliptic*, K, which is in Draco. The *South Pole of the Ecliptic* is in Dorado.

The ecliptic and its pole K form the basis of the ecliptic coordinate system, illustrated in figure $\mathrm{VI}.5$. The *ecliptic longitude* λ and the *ecliptic latitude* β of a star X are shown in figure $\mathrm{VI}.5$, which should be self explanatory. In order to convert between equatorial and ecliptic coordinates, the triangle to solve is triangle PKX. The arc KX is $90^\circ - \beta$ and the angle PKX is $90^\circ - \lambda$. What are the arc PK, the arc PX and the angle KPX?

[Answers: $\mathrm{PK} = \eta$ $\mathrm{PX} = 90^\circ - \delta$ $\mathrm{KPX} = 90^\circ + \alpha$]

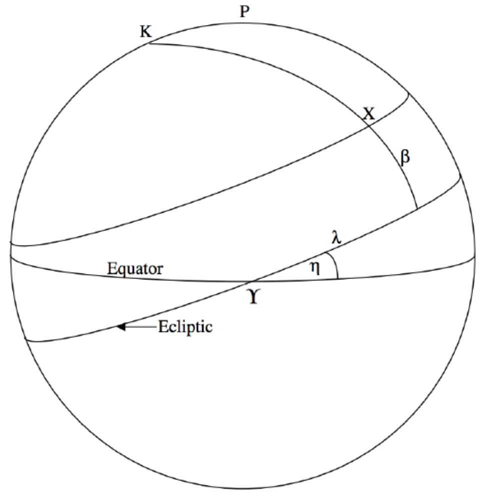


FIGURE VI.5

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.6: The Mean Sun

The bright yellow (or white) ball of fire that *appears* in the sky and which you could see with your eyes if ever you were foolish enough to look directly at it is the *Apparent Sun*. It is moving eastward along the ecliptic, and its right ascension is increasing all the time. Consequently consecutive upper transits across the meridian take about four minutes longer than consecutive transits of a star or of the First Point of Aries. The hour angle of the Apparent Sun might have been called the local apparent solar time, except that we like to start our days at midnight rather than at midday. Therefore the *Local Apparent Solar Time* is the *hour angle of the Apparent Sun plus twelve hours*. It is "local", because the hour angle of the apparent Sun depends continuously on the longitude of the observer. It is the time indicated by a sundial. In order to convert it to a standard zone time, we must know, among other things, our longitude.

The Apparent Sun has some drawbacks as an accurate timekeeper, particularly because *its right ascension does not increase at a uniform rate throughout the year*. The motion of the Apparent Sun, is, of course, just a reflection of Earth's annual orbital motion around the Sun. The Earth moves rather faster at perihelion (on or near January 4) than at aphelion (on or near July 4); consequently the Apparent Sun moves faster along the ecliptic in January than in July. Even if this were not so, however, and the Sun were to move at a uniform rate along the ecliptic, its right ascension would not increase at a uniform rate. This is because right ascension is measured along the celestial equator rather than along the ecliptic. If the Sun were moving uniformly along the ecliptic, its right ascension would be increasing faster at the solstices (where its motion is momentarily parallel to the equator) than at the equinoxes, (where is motion is inclined at $23^\circ.4$ to the ecliptic). So there are these two reasons why the right ascension of the apparent Sun does not increase uniformly throughout the year.

To get over these two difficulties we have to invent two imaginary suns. One of them accompanies the apparent (i.e. the real!) Sun in its journey around the ecliptic. The two start together at perihelion. This Dynamic Sun moves at a constant rate, so that the Apparent Sun (which moves faster in January when Earth is at perihelion) moves ahead of the imaginary sun. By the time Earth reaches aphelion in July, however, the Apparent Sun is slowing down, and the Dynamic Sun manages to catch up with the Apparent Sun. After that, the Dynamic Sun surges ahead, leaving the Apparent Sun behind. But the Apparent Sun starts to gain speed again, and catches up again with the Dynamic Sun at perihelion in January. The Apparent Sun and the Dynamic Sun coincide twice per year, at perihelion and at aphelion.

Now we imagine a second imaginary sun – a rather important one, known as the Mean Sun. The Mean Sun moves at a constant rate *along the equator*, its right ascension moving uniformly all through the year. It coincides with the Dynamic Sun at Υ. At this time, the right ascension of the Dynamic Sun is increasing rather slowly, because it is moving along the ecliptic, at an angle to the equator. Its right ascension increases most rapidly at the solstices, and by the time of the first solstice it has caught up with the Mean Sun. After that, it moves ahead of the Mean Sun for a while, but it soon slows down as its motion begins to make an ever steeper angle to the equator, and Dynamic Sun and the Mean Sun coincide again at the second equinox. Indeed these two suns coincide four times a year – at each of the equinoxes and solstices.

Local Mean Solar Time is the *hour angle of the Mean Sun plus twelve hours*, and the difference Local Apparent Solar Time minus Local Mean Solar Time is called the *Equation of Time*. The Equation of time is the sum of two periodic functions. One is the *Equation of the centre*, which is the difference in right ascensions of the Apparent Sun and the Dynamic Sun, and it has a period of one year. The second is the *reduction to the equator*, which has a period of half a year. The value of the Equation of time varies through the year, and it can amount to a little more than 16 minutes in early November. Local Mean Solar Time, while uniform (or as uniform as the rotation of the Earth) still depends on the longitude of the observer. For that reason, all the inhabitants of a zone on Earth roughly between longitudes $7^\circ.5$ East and West agree to use a standard the Local Mean Solar Time at Greenwich, also called Greenwich Mean Time, GMT, or Universal Time, UT. Similar zones about 15 degrees wide have been established around the world, within each of which the time differs by an integral; number of hours from Greenwich Mean Time.

We shall discuss in Chapter 7 small distinctions between various versions of Universal Time as well as Ephemeris Time and Terrestrial Dynamical Time.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.7: Precession

The First Point of Aries is the point where the ecliptic crosses the equator at the point occupied by the Sun at the March equinox. It is the point from which right ascensions are measured. We have hitherto treated it as if it were fixed relative to the stars, although we have hinted from time to time that this is not exactly so. Indeed we have said that it is essential, when stating the right ascension and declination of a star, to state the date of the equinox to which it refers.

In figure $\mathrm{VI}.6$, I have drawn the ecliptic horizontally, and the celestial equator inclined at an angle of $23^\circ.4$. You can see the north pole of the ecliptic, K, and the north celestial pole P. The great circle $\mathrm{P}\Upsilon$ (not drawn) is the equinoctial colure, and the right ascension of Υ is 0^h. The right ascension and declination of K are 18^h, $+66^\circ.6$, which is a point between the stars δ and ζ Draconis.

Neither the north celestial pole P nor the "First Point of Aries" Υ are fixed, however. The north celestial pole P describes a small circle of radius $23^\circ.4$ around K, and the equinox regresses westwards along the ecliptic in a period of 25,800 years. This motion, called the precession of the equinoxes (or just "precession" for short) is not quite uniform, but is nearly so and will be treated as such in this section. The complete cycle of 25,800 years corresponds to a westward regression of Υ along the ecliptic of $50''.2$ per year or $0''.137$ per day. The component of that motion along the celestial equator is $0''.137 \cos 23^\circ.4 = 0''.126 = 0^{\mathrm{s}}.008$ per day. That is why the length of the mean sidereal day (which is defined as the interval between two consecutive upper meridian transits of the first point of Aries) is $0^{\mathrm{s}}.008$ shorter than the sidereal rotation period of Earth.

The precession of P around K means that the entire system of equatorial coordinates (right ascension and declination) moves continuously, and the right ascensions and declinations of all the stars are continuously changing. No matter where P is in its journey around K, however, the equatorial coordinates of Υ and of K are always 0^{h}, 0^{o} and 18^{h}, $+66^{\mathrm{o}}.5$. However, equatorial coordinates of the stars must always be referred to the equinox and equator of a stated epoch.

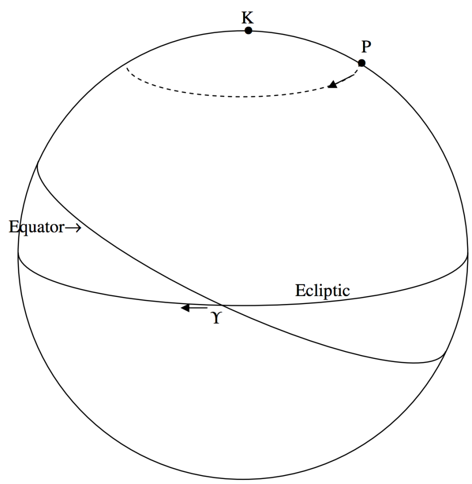


FIGURE VI.6

During much of the twentieth century, the epoch referred to by many catalogues and atlases was $\mathrm{B}1950.0$. That is the beginning of the *Besselian Year* of 1950, at the instant (shortly before midnight on the night of 1949 Dec 31 / 1950 Jan 1) when the right ascension of the Mean Sun was $18^{\mathrm{h}}40^{\mathrm{m}}$. Most catalogues since 1984 have referred right ascensions and declinations to the mean equinox and equator of $\mathrm{J}2000.0$. That is the beginning of the *Julian Year* 2000, at the instant when Greenwich Mean Time (UT) indicated midnight. For example, in the older catalogues, the right ascension and declination of Arcturus would be given as

$$\alpha_{1950.0} = 14^{\mathrm{h}} \quad 13^{\mathrm{m}}.4 \quad \delta_{1950.0} = +19^\circ 26', \tag{6.7.1}$$

whereas in more recent catalogues they are given as

$$\alpha_{2000.0} = 14^{\mathrm{h}} \quad 15^{\mathrm{m}}.8 \quad \delta_{2000.0} = +19^\circ 11'. \tag{6.7.2}$$

Thus it can be seen that for precise work the difference is not at all negligible, and to state the equatorial coordinates of an object without also stating the epoch of the equinox and equator to which the coordinates are referred is not generally useful. Of course, when setting the circles of a telescope for the night's observations, what one needs are the right ascension and declination referred to the equinox and equator *of date* – i.e. for the date in question. It is therefore essential for a practical observer to know how to correct for precession.

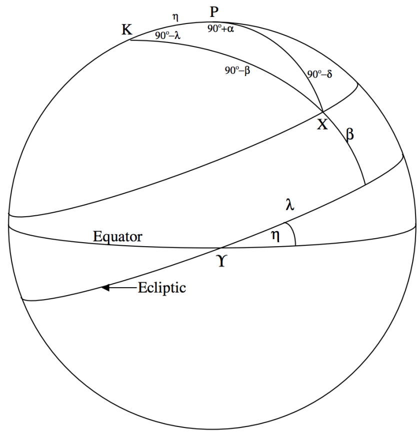


FIGURE VI.7

Apply the cosine formula (Equation 3.5.2) to triangle PKX to obtain

$$\sin\delta = \cos\eta\sin\beta + \sin\eta\cos\beta\sin\lambda. \tag{6.7.1}$$

Since Υ is regressing down the ecliptic, the ecliptic longitude λ of the star X is increasing. If it is increasing at a rate $\dot{\lambda}$ $(= 50''.2$ per year), the rate of change of its declination can be obtained by differentiation of Equation 6.7.1 with respect to time, bearing in mind that β and η are constant:

$$\cos\delta\dot{\delta} = \sin\eta\cos\beta\cos\lambda\dot{\lambda}. \tag{6.7.2}$$

But $(\cos\beta\cos\lambda)/\cos\delta$ is obtained from the sine formula (Equation 3.5.1):

$$\frac{\cos\beta}{\cos\alpha} = \frac{\cos\delta}{\cos\lambda}. \tag{6.7.3}$$

Hence we obtain for the rate of change of declination of a star due to precession:

$$\dot{\delta} = \dot{\lambda}\sin\eta\cos\alpha. \tag{6.7.4}$$

To obtain the rate of change of right ascension, we can write Equation 6.7.3 as

$$\cos\alpha = \cos\beta\sec\delta\cos\lambda \tag{6.7.5}$$

and then differentiate with respect to time:

$$-\sin\alpha\dot{\alpha} = \cos\beta\sec\delta(\tan\delta\dot{\delta}\cos\lambda - \sin\lambda\dot{\lambda}), \tag{6.7.6}$$

which I am going to write as

$$-\sin\alpha\dot{\alpha} = \cos\beta\sec\delta\cos\lambda(\tan\delta\dot{\delta} - \tan\lambda\dot{\lambda}). \tag{6.7.7}$$

We can get $\cos\beta\sec\delta\cos\lambda$ from Equation 6.7.5, and of course we have $\dot{\delta}$ from Equation 6.7.4, but we still need to find an expression for $\tan\lambda$ in terms of equatorial coordinates. We can do this from the cotangent formula (Equation 3.5.4), in which the inner angle is $90^\circ + \alpha$ and the inner side is η:

$$-\cos\eta\sin\alpha = \sin\eta\tan\delta - \cos\alpha\tan\lambda. \tag{6.7.8}$$

On substitution of Equations 6.7.4, 6.7.5 and 6.7.8 into Equation 6.7.7 we obtain, after a very small amount of algebra, for the rate of change of right ascension of a star due to precession:

$$\dot{\alpha}=\dot{\lambda}(\cos\eta+\sin\alpha\tan\delta\sin\eta). \tag{6.7.9}$$

With $\dot{\lambda}=50''.2$ per year and $\eta=23^\circ.4$, Equations 6.7.4 and 6.7.9 become

$$\dot{\delta}=19''.9\cos\alpha \quad \text{per year} \tag{6.7.10}$$

and

$$\dot{\alpha}=46''.1+19''.9\sin\alpha\tan\delta \quad \text{per year} \tag{6.7.11}$$

or

$$\dot{\alpha}=3^s.07+1^s.33\sin\alpha\tan\delta \quad \text{per year.} \tag{6.7.12}$$

These formulae should be adequate for all but very precise calculations.

Problem: Use Equations 6.7.10 and 6.7.12 to verify the data about Arcturus – and please let me know if it isn't right!

At the time of Hipparcos (who discovered the phenomenon of precession as long ago as the second century B.C.), the spring equinox was in the constellation Aries – indeed at its eastern boundary. Hence it was called the First Point of Aries. Over the centuries, precession has carried the equinox westward right across the constellation Aries, and because of this, together with the way in which the constellation boundaries were formally fixed in 1928, the equinox is now near the western boundary of Pisces and is only a few degrees from Aquarius. It is still called, however, by its traditional name of the First Point of Aries. Incidentally, the ecliptic actually passes through the constellation Ophiuchus, which is not one of the traditional twelve "Signs of the Zodiac", and it is sometimes said that this is a result of precession over the centuries. This is not the case. Precession does not alter the plane of the ecliptic, and the ecliptic continues to pass through the same constellations regardless of where the equinox is along it. The inclusion of Ophiuchus is merely a result of the way in which the constellation boundaries were formally fixed in 1928.

The Physical Cause of the Precession

The daily motion of the stars around the north celestial pole is, of course, a reflection of Earth's rotation on its axis; and the annual motion of the Sun along the ecliptic, which is inclined at $23^\circ.4$ to the celestial equator, is a reflection of the annual orbital motion of Earth around the Sun, the plane of Earth's rotational equator being inclined at $23^\circ.4$ to the plane of its orbit – i.e. to the ecliptic. Although this obliquity of $23^\circ.4$ is approximately constant, the direction of Earth's rotational axis is not fixed, but it precesses around the normal to the ecliptic plane with a period of 25,800 years.

From the point of view of classical mechanics, Earth is an *oblate symmetric top*. That is to say, it has an axis of symmetry and two of its principal moments of inertia are equal and are less than the moment of inertia about the axis of symmetry. The phenomena of precession of such a body are well understood and are studied in courses of classical mechanics. It is necessary, however, to be clear in one's mind about the distinction between *torque-free precession* and *torque-induced precession*.

The phenomenon of *torque-free precession* is the precession that occurs when a symmetric top is spinning about an axis that does not coincide with its symmetry axis and it is spinning freely with no external torques acting upon it. In such circumstances, the angular momentum vector is fixed in magnitude and direction. The symmetry axis precesses about the fixed angular momentum vector while the instantaneous axis of rotation precesses about the symmetry axis. The rotation of Earth does indeed exhibit this type of behaviour, but this is not the precession that we are talking about in connection with the precession of the equinoxes. The instantaneous axis of rotation of Earth is only a very few metres away from its symmetry axis and the period of the torque-free precession is about 432 days. This gives rise to a phenomenon known as *variation of latitude*, and it results in the latitudes of locations of Earth's surface varying quasi-periodically with an amplitude of less than a fifth of an arcsecond. The precession of the equinoxes that we have been discussing in this section is something entirely different.

The figure of Earth is approximately an oblate spheroid. If we call the equatorial radius a and the polar radius c, the *geometric ellipticity* $(a-c)/a$ is about $1/297.0$. If we call the corresponding principal moments of inertia A and C, the *dynamical ellipticity* $(C-A)/A$ is about $1/305.1$. Earth's equator is inclined to the ecliptic, and, because of the equatorial bulge, the spinning Earth is subject to torques from both the Sun and the Moon (whose orbit is inclined to the ecliptic by about 5 degrees). The magnitude of the torque is proportional to the diameter of Earth times the *gravitational field gradient* $2GM/r^3$, and the direction of the torque vector is perpendicular to the angular momentum vector.

Exercise 6.7.1

Look up the masses of Sun and Moon, and their mean distances from Earth. Show that M/r^3 for the Moon is about twice that for the Sun. Thus the torque on Earth exerted by the Moon is about twice the torque exerted by the Sun.

Now if a symmetric top is spinning about its axis of symmetry with angular momentum $\mathbf{L}$ and if it is subject to an external torque $\boldsymbol{\tau}$, its angular momentum will change (not in magnitude, but in direction), and $\mathbf{L}$ will precess with an angular velocity $\boldsymbol{\Omega}$ given by

$$\boldsymbol{\tau} = \boldsymbol{\Omega} \times \mathbf{L}. \tag{6.7.13}$$

Equation 6.7.13 does not give the direction of $\boldsymbol{\Omega}$ uniquely – that depends on the initial conditions. Figure VI.8 illustrates the situation.

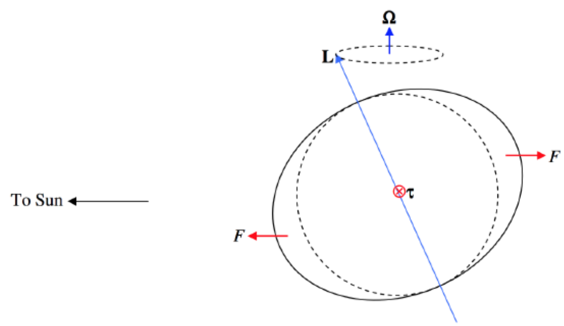


FIGURE VI.8

The equatorial bulge is much exaggerated. The figure is drawn in a reference frame that is revolving around the Sun with the Earth, so there is no net gravitational force on Earth (the gravitational attraction of the Sun is counteracted by the centrifugal force). In this frame, there is a little force F acting towards the Sun on the sunward-facing bulge, and an equal force acting away from the Sun on the opposite side. This amounts to a torque of magnitude $\tau = Fd \sin \eta$, where η is the obliquity of the ecliptic and d is the diameter of Earth. Thus if we equate the magnitudes of both sides of Equation 6.7.13, we obtain for the angular speed of the precession

$$\Omega = Fd/L, \tag{6.7.14}$$

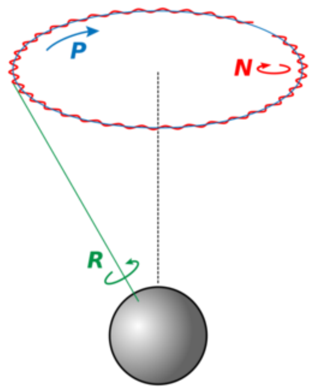
which is independent of η. This, then is the cause of the precession of the equinoxes, except that, for the purpose of figure VI.8, I referred only to the Sun. You have yourself calculated that the influence of the Moon is about twice that of the Sun, and the combined effect of the Moon and the Sun is called the *luni-solar precession*. There is a small additional precession resulting from the influence of the other planets in the solar system.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.8: Nutation

Those who have studied the gyrations of a spinning top will recall that, in addition to precessing, a top may *nutate*, or nod up and down (Latin *nutare*, to nod), the amplitude and type of nutation depending on the initial conditions. Earth's axis does indeed nutate, but not from the same cause. Those who have studied tops will understand that damping more or less rapidly damps out the amplitude of the nutation, and, since Earth is a nonrigid, flexible body, this type of nutation has long ago damped out.



Rotation (green) precession (blue), and nutation (red) in obliquity of a planet. (CC BY-SA 3.0; User Herbye).

Earth's axis of rotation nutates because it is subject to varying torques from Sun and Moon – the former varying because of the eccentricity of Earth's orbit, and the latter because of both the eccentricity and inclination of the Moon's orbit. This means that the equinox Υ does not move at uniform speed along the ecliptic, and the obliquity of the ecliptic varies quasi-periodically. These two effects are known as the *nutation* in *longitude* and the *nutation in the obliquity*. While several effects involving both the Sun and the Moon are involved, the most important term in the general expressions for both nutation in longitude and nutation in obliquity involve the longitude of the nodes of the Moon's orbit, which are known to regress with a period of 18.6 years. Thus both nutations, in the first approximation, have a period of 18.6 years. The nutation in longitude has an amplitude of $17''.2$, and the nutation in the obliquity has an amplitude of $9''.2$. In addition, planetary perturbations cause a secular (i.e. not periodic) decrease in the obliquity of about $0''.47$ per year.

A further point that should be mentioned is that the plane of the ecliptic is not quite invariable. What is invariable in the absence of external torques on the solar system is the direction of the angular momentum vector of the solar system; the plane perpendicular to this is called the *invariable plane* of the solar system.

This section and the previous section have described briefly in a rather qualitative way the motion of the equinox along the ecliptic with a period of 25,800 years (i.e. precession) – a motion that is not quite uniform on account of the nutations in longitude and the obliquity. This brief account may suffice for most purposes of the observational astronomer and for the aim of this chapter, which is a general overview of the celestial sphere. A more thorough and detailed treatment of precession and nutation will have to wait for a special chapter devoted to the subject.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.9: The Length of the Year

The time taken for Earth to revolve around the Sun with respect to the stars, which is the same thing as the time taken for the Apparent Sun to move around the ecliptic with respect to the stars, is a *Sidereal Year*, which is $365^{\mathrm{d}}.25636$, where the "d" denotes a mean solar day. The length of the seasons, however, is determined by the motion of the Apparent Sun relative to Υ. Because Υ is moving westward along the ecliptic, the time that the Apparent Sun takes to move around the ecliptic relative to Υ, which is called the *Tropical Year*, is a little less than the sidereal year. We have seen, however that the motion of Υ along the ecliptic is not quite uniform, and we have to average out the effects of nutation. Thus the *Mean Tropical Year* is the average time for the ecliptic longitude of the Apparent Sun to increase by 360°, which is $365^{\mathrm{d}}.24219$.

The calendar that we use in everyday life is the *Gregorian Calendar*, in which there are 365 days in most years, but 366 days in years that are divisible by 4 unless they are also divisible by 100 other than those that are also divisible by 400. Thus *leap years* (those that have 366 days) include 1996, 2000, 2004, but not 2005 or 1900. (2000 was a leap year because, although it is divisible by 100, it is also divisible by 400.) The average length of the *Gregorian Year* is 365.2425, which is close enough to the Mean Tropical Year for present-day purposes, but which is of concern to calendar reformers and will be of some concern to our remote descendants.

The *Anomalistic Year* is the interval between consecutive passages of the Earth through perihelion. The perihelion of Earth's orbit is slowly advancing in the same direction as the Earth's motion, so the anomalistic year is a little longer than the sidereal year, and is equal to $365^{\mathrm{d}}.25964$.

Figure VI.9 illustrates a way of thinking about the relation between the sidereal and tropical years. We are looking down on the ecliptic from the direction of the north ecliptic pole. We see the Sun moving counterclockwise at angular speed ω_{sid} and moving clockwise at angular speed ω_{Υ}. The angular speed of the Sun relative to Υ

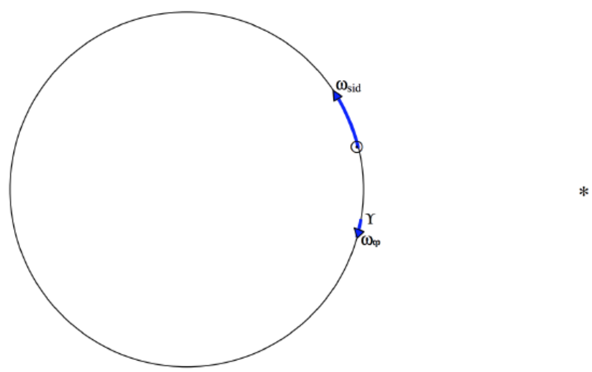


FIGURE VI.9

is $\omega_{\mathrm{trop}} = \omega_{\mathrm{sid}} + \omega_{\Upsilon}$. But period P and angular speed ω are related by $\omega = 2\pi/P$.

Therefore:

$$\frac{1}{P_{\mathrm{trop}}} = \frac{1}{P_{\mathrm{sid}}} + \frac{1}{P_{\Upsilon}}. \tag{6.9.1}$$

Thus $P_{\mathrm{sid}} = 365^{\mathrm{d}}.25636$ and $P_{\Upsilon} = 25800 \text{ years} = 9.424 \times 10^{6} \text{ days}$. Hence $P_{\mathrm{trop}} = 365^{\mathrm{d}}.2422$. Using the same argument, see if you can calculate how long it takes for the perihelion of Earth's orbit to advance by 360° – bearing in mind that the perihelion is advancing, not regressing.

One more point worth noting is that, during a sidereal year, the Sun has upper transited across the meridian 365.25636 times, whereas a fixed star has transited 366.25636 times. Expressed another way, while Earth turns on its axis 365.25636 times relative to the Sun, relative to the stars it has made one extra turn during its revolution around the Sun. Thus

$$\frac{\text{Length of sidereal day}}{\text{Length of solar day}} = \frac{365.25636}{366.25636}. \tag{6.9.1}$$

Thus the length of the sidereal day is $23^{\mathrm{h}}\ 56^{\mathrm{m}}\ 04^{\mathrm{s}}$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.10: Problems

In Section 3.5 of Chapter 5, I suggested that it might be a good ides to write a computer program, which would last you for life, that would solve any problem involving plane or spherical triangles. If you did that, the following problems will be easy. If you didn't, you are now about to suffer.

6.10.1

The equatorial coordinates ($\mathrm{J}2000.0$) of Antares and Deneb are, respectively

Antares $\alpha=16^{\mathrm{h}} 29^{\mathrm{m}} .5$ $\delta=-26^{\circ}\ 26^{\prime}$
Deneb $20\ 37.6+45\ 17$

Calculate the positions of the poles of the great circle joining these two stars.

I put one star in the northern hemisphere, and the other in the south, and I put the stars in the third and fourth quadrants of right ascension, just to be awkward.

6.10.2

The parallax of Antares is $0^{\prime\prime}.00540$, and the parallax of Deneb is $0^{\prime\prime}.00101$. How far apart are the stars (a) in parsecs? (b) in km? (c) in light-years? The speed of light is $2.997\ 92 \times 10^{8}\ \mathrm{m\ s}^{-1}$, the radius of Earth's orbit is $1.495\ 98 \times 10^{8}\ \mathrm{km}$, and a tropical year is 365.24219 mean solar days.

6.10.3

A meteor starts at $\alpha=23^{\mathrm{h}} 24^{\mathrm{m}} .0 \quad \delta=+04^{\circ}\ 00^{\prime}$
and finishes at $\alpha=01^{\mathrm{h}} 36^{\mathrm{m}} .0 \quad \delta=+10^{\circ}\ 00^{\prime}$

A second meteor, from the same shower (i.e. from the same meteoroid stream) starts at

$$\alpha=00^{\mathrm{h}} 06^{\mathrm{m}} .0 \quad \delta=+03^{\circ}\ 00^{\prime}$$

and finishes at $\alpha=02^{\mathrm{h}} 12^{\mathrm{m}} .0 \quad \delta=+05^{\circ}\ 30^{\prime}$.

Calculate the position of the radiant (i.e. the position on the sky where the two paths, projected backwards, intersect).

Again you'll notice that I chose the coordinates to be as awkward as I could.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

6.11: Solutions

6.11.1

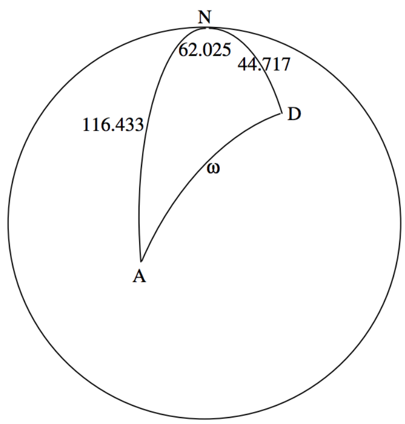
I think the first thing that I would do, would be to convert the coordinates to degrees and decimals (or maybe even radians and decimals, though I do it below in degrees and decimals):

Antares: $\alpha = 247.375 \quad \delta = -26.433$
Deneb $\alpha = 309.400 \quad \delta = +45.283$

We already did a similar problem in Chapter 3, Section 3.5, Example 2, so I shan't do it again. I make the answer:

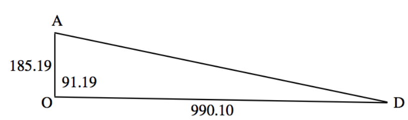
One pole: $\alpha = 11^{\mathrm{h}}47^{\mathrm{m}}.3 \quad \delta = +56^\circ 11'$
The other pole: $\alpha = 23^{\mathrm{h}}47^{\mathrm{m}}.3 \quad \delta = +123^\circ 49'$

6.11.2



I have drawn the North Celestial Pole N, and the colures from N to Antares (A) and to Deneb (D), together with their north polar distances in degrees. I have also marked the difference between their right ascensions, in degrees. We can immediately calculate, from the cosine rule for spherical triangles, Equation 3.5.2, the angular distance ω between the two stars in the sky. I make it $\omega = 91^\circ.190\ 79$.

Now that we know the angle between the stars, we can use a plane triangle to calculate the distance between them:



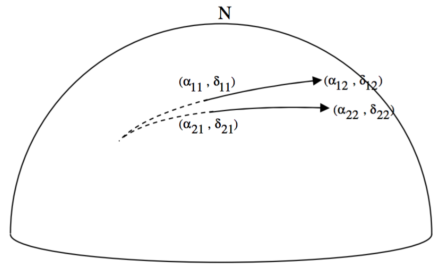
I have marked Antares (A), Deneb (D) and us (O), and the distances from us to the two stars, in parsecs. (That's the reciprocal of their parallaxes in arcsec.) I have also marked the angles, in degrees, between Antares and Deneb. We can now use the cosine rule for planes triangles, Equation 3.2.2, to find the distance AD. I make it 1011 parsecs.

A parsec is the distance at which an astronomical unit (approximately the radius of Earth's orbit) would subtend an angle of one arcsecond. This also means, if you come to think of it, that the number of astronomical units in a parsec is equal to the number of arcseconds in a radian, which is $360 \times 3600 \div (2\pi) = 2.062648 \times 10^5$. The distance between the stars is therefore $1011 \times 2.062648 \times 10^5$ astronomical units. Multiply this by $1.495\ 98 \times 10^8$, to get the distance in km. I make the distance 3.120×10^{16} km.

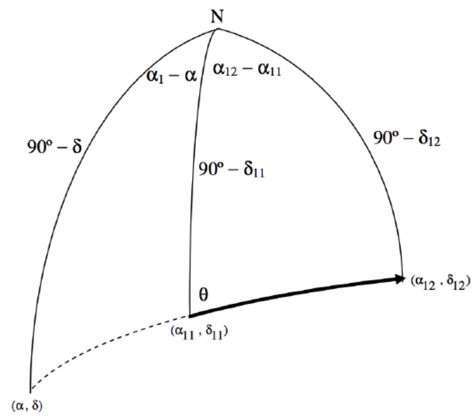
This would take light 1.040596×10^8 seconds to travel, or 3298 years, so the distance between the stars is 3298 light-years.

6.11.3

Let's see if we can develop a formula for a general case. We'll have the first meteor start at $(\alpha_{11},\ \delta_{11})$ and finish at $(\alpha_{12},\ \delta_{12})$. The second meteor starts at $(\alpha_{21},\ \delta_{21})$ and finishes at $(\alpha_{22},\ \delta_{22}$. We have to find the coordinates $(\alpha,\ \delta$ of the point from which the two meteors diverge.



This is not a particularly easy problem – but is one that is obviously useful for meteor observers. I'll just outline some suggestions here, and leave the reader to work out the details. I'll draw below one of the meteors, and the radiant, and the North Celestial Pole:



Use the cotangent rule (Equation 3.5.5) on the righthand triangle to get an expression for $\cot\theta$:

$$\sin\delta_{11}\cos(\alpha_{12}-\alpha_{11}) = \cos\delta_{11}\tan\delta_{12} + \sin(\alpha_{12}-\alpha_{11})\cot\theta. \tag{6.11.1}$$

Equate these two expression for $\cot\theta$ (i.e. eliminate θ between the two Equations). This will give you a single Equation containing the two unknowns, α and δ, everything else in the Equation being a known quantity. (This will be obvious if you are actually doing a numerical example.)

Now do the same thing for the second meteor, and you will get a second Equation in α and δ. In principle you are now home free, though there may be a bit of heavy algebra and trigonometry to go through before you finally get there.

I make the answer as follows:

$$\tan\alpha = \frac{\cos\alpha_{22}\tan\delta_{22} - \cos\alpha_{12}\tan\delta_{12} + a_1\sin\alpha_{12} - a_2\sin\alpha_{22}}{\sin\alpha_{12}\tan\delta_{12} - \sin\alpha_{22}\tan\delta_{22} + a_1\cos\alpha_{12} - a_2\cos\alpha_{22}}, \tag{6.11.2}$$

where

$$a_1 = \frac{\tan\delta_{11}}{\sin(\alpha_{11}-\alpha_{12})} - \frac{\tan\delta_{12}}{\tan(\alpha_{11}-\alpha_{12})} \tag{6.11.3}$$

and

$$a_2 = \frac{\tan\delta_{21}}{\sin(\alpha_{21}-\alpha_{22})} - \frac{\tan\delta_{22}}{\tan(\alpha_{21}-\alpha_{22})} \tag{6.11.4}$$

Then

$$\tan\delta = \cos(\alpha-\alpha_{12})\tan\delta_{12} + \sin(\alpha-\alpha_{12})[\csc(\alpha_{11}-\alpha_{12})\tan\delta_{11} - \cot(\alpha_{11}-\alpha_{12})\tan\delta_{12}] \tag{6.11.5}$$

or

$$\tan\delta = \cos(\alpha-\alpha_{22})\tan\delta_{22} + \sin(\alpha-\alpha_{22})[\csc(\alpha_{21}-\alpha_{22})\tan\delta_{21} - \cot(\alpha_{21}-\alpha_{22})\tan\delta_{22}]. \tag{6.11.6}$$

Either of these two Equations for $\tan\delta$ should give the same result. In the computer program I use for this calculation, I get it to calculate $\tan\delta$ from *both* Equations, just as a check for mistakes.

This may look complicated, but all terms are just calculable numbers for any particular case. If the equinoctial colure gets in the way (as it did – deliberately – in the numerical example I gave), I suggest just add 24 hours to all right ascensions.

For the numerical example I gave, I make the coordinates of the radiant to be:

$$\alpha = 22^{\mathrm{h}}01^{\mathrm{m}}.3 \quad \delta = -00^\circ 37'. \tag{6.11.7}$$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

7: Time

In this chapter we briefly discuss the several time scales that are in use in astronomy, such as Universal Time, Mean Solar Time, Ephemeris Time, Terrestrial Dynamical Time, and the several types of second, hour, day and year that are or have been in use. For some topics it will be assumed that the reader has read the relevant portions of Chapter 6 in order to have a fuller understanding. Some of the items in this chapter will be given only in short note form or single sentence definitions, particularly where they have already been discussed in Chapter 6. Others will require a bit more discussion.

The *Local Apparent Solar Time* at a particular geographical longitude is the hour angle of the Apparent Sun plus 12 hours. It is the time indicated by a sundial. Because the right ascension of the Apparent Sun does not increase uniformly during the year, local apparent solar time does not proceed at a uniform rate. (What is meant by "time proceeding at a uniform rate" is something that can be pondered about. One might indeed ponder for a long time.)

The *Local Mean Solar Time* at a particular longitude is the hour angle of the Mean Sun plus 12 hours. Although, like local apparent solar time, it is local to a particular longitude, it does, at least in some sense, flow uniformly, inasmuch as the right ascension of the Mean Sun increases uniformly. If the reader is wondering whether the sentence "mean solar time flows uniformly because the right ascension of the Mean Sun increases uniformly" is circular logic, and that either part of the sentence follows from the definition of the other, he or she is not alone. Indeed, defining exactly what is meant by "uniformly flowing time" is not easy; I am not sure if anyone has ever fully successfully managed it.

The *Equation of Time* is the difference between Local Apparent Solar Time and Local Mean Solar Time. Whether it is $\text{LAMS} - \text{LMST}$ or $\text{LMST} - \text{LAMS}$ varies from author to author. Thus, whenever you use the phrase in your own writing, be careful to define which sense you intend.

Greenwich Mean Time is the Local Mean Time at the longitude of Greenwich. In a general sense it is the same thing as Universal Time. However, there are some slight refinements of Universal Time of which we should be aware, and which will be discussed later.

Zone Time. Since Local Mean Solar Time is essentially local – i.e. it varies from longitude to longitude − it has been decided, for civil timekeeping purposes, to divide the world into a number of longitude zones approximately 15 degrees (one hour) wide, in which everyone agrees to keep the same time, namely the local mean solar time for a particular longitude within the zone. Here, where I write in Victoria, British Columbia, Canada, within our zone we use Pacific Standard Time during the winter months. This is eight hours behind Greenwich Mean Time. Many jurisdictions advance their zone time by one hour during the summer months; thus in the summer here in Victoria, we use Pacific Daylight-saving Time, which is just seven hours behind Greenwich Mean Time. It needs to be remembered that, to change from the Standard time for a given zone to Daylight-saving time, clocks are advanced by one hour in spring, and that "spring" occurs six months apart in the northern and southern hemispheres! The standard zone time for Sydney, Australia, is 18 hours ahead of the standard time for Victoria, Canada. But in December, it is summer in Australia and winter in Victoria; Sydney is then on Daylight-saving Time while Victoria is on Pacific Standard Time – a difference of 19 hours. In June, Victoria is on Daylight-saving Time while Australia is on Standard Time, a difference of 17 hours. These complications have to be understood by those who are planning international telephone calls!

Local Sidereal Time is the hour angle of the First Point of Aries, and is equal to the hour angle plus right ascension of any star.

A *Mean Solar Day* is the interval between two consecutive upper meridian transits of the Mean Sun.

A *Mean Sidereal Day* is the interval between two consecutive upper meridian transits of the mean equinox. It is equal to $23^{\text{h}}\ 56^{\text{m}}\ 04^{\text{s}}.091$ of mean solar time. The rotation period of Earth relative to the fixed stars is $23^{\text{h}}\ 56^{\text{m}}\ 04^{\text{s}}.099$ of mean solar time. Transits of Υ are slightly closer together because of the westward precessional motion of Υ along the ecliptic.

A *Sidereal Year* is the period of revolution of Earth around the Sun relative to the fixed stars, and it is $365^{\text{d}}.25636$, where "d" denotes "mean solar days".

A *Mean Tropical Year* is the mean time required for the Apparent Sun to increase its ecliptic longitude by 360°. It is the interval that determines the seasons and is equal to $365^{\text{d}}.24219$. It is less that a sidereal year because of the westward motion of Υ along the ecliptic.

An *Anomalistic Year* is the interval between two consecutive passages of Earth through perihelion. It is equal to $365^{\mathrm{d}}.25964$ It is longer that the sidereal year because of the forward motion of perihelion.

In days gone by, when life was simpler, a *second* was merely the fraction $1/86400$ of a mean solar day. As time-keeping became more and more precise, it became evident not only that time could be measured more precisely in the laboratory with atomic clocks than the rotation of Earth could be measured, but that Earth itself was not a perfect timekeeper, because it does not rotate uniformly when measured with an atomic clock. This is presumably because of unpredictable changes within the body of Earth which change its rotational inertia. This again raises the question of what is meant by "uniformly flowing time". Whatever is meant by it, atomic time is presumed to be a better representation of it than an irregularly rotating Earth.

At present, the SI (Système International) definition of the second is the interval of $9192\,631\,770$ periods of the radiation corresponding to the transition between the hyperfine levels $F=0$ and $F=1$ of the ground level $^2\mathrm{S}_{1/2}$ of the caesium isotope $^{137}\mathrm{Cs}$. While it can easily be argued that this definition of the second is superior in numerous respects to the definition based on the rotation of Earth, it must be noticed that this definition is useful (albeit *very* useful) only for determining *intervals of time* – i.e. how many seconds have elapsed between event A and event B. By itself, the definition does nothing to determine the *instant of time* of a single event. It tells us nothing about how far Earth has rotated on its axis (time of day) or how far it has moved in its orbit around the Sun (time of year). There is still a need for a time scale for determining the instant of time of astronomical events and to use as a "uniformly-flowing" time as argument in celestial mechanical calculations and the provision of ephemerides.

International Atomic Time (TAI – from the initial letters of the French name, *Temps Atomique International*) does enable us to define an instant of the time of occurrence of an event, since it is defined by the "ticking" of a caesium atomic clock beating out SI seconds of atomic time, which started at the beginning of the day 1958 January 01. That is, it has a unit of time and a starting point. Many seconds have elapsed since that epoch, however, so you can compare it with the "time of day" or with Greenwich Mean Time, by subtracting 86400 seconds whenever the number of seconds exceeds this number. That is, TAI is the number of seconds that have elapsed since the initial epoch, "modulo 86400". This will not agree exactly with Greenwich Mean Time (i.e. the hour angle of the Mean Sun at Greenwich plus 12 hours) unless Earth rotates uniformly when compared with an atomic clock. It does not, quite, so one may ask which clock is "at fault", or which clock is "running uniformly". Most of us will probably agree that it is the atomic clock that is running uniformly and that the difference between TAI and GMT is caused by irregularities in the rate of rotation of Earth relative to TAI. Thus we may be tempted, for many purposes, to prefer to measure time interval with an atomic clock than to use the rotation of Earth as our time keeper. This is valid indeed, if all we want to do is to measure the interval of time between two events – but it still does not tell us where the Sun (whether Mean or Apparent) is in the sky, and we still need a time scale, whether it is uniform or not, that tells us the hour angle of the Sun.

The required time scale is *Universal Time*, which is the hour angle of the Mean Sun at Greenwich plus 12 hours, and is, for most purposes, the same as Greenwich Mean Time. However, for very precise work, there are several subtly-different varieties of Universal Time. In principle, we could determine UT by measuring the hour angle of the Mean Sun – if only we could see the Mean Sun and record exactly when it crosses the meridian. In practice, UT is determined by recording the transit times of stars, and calculating the Universal Time from the observed Local Sidereal Time. If you do this, you get what is known as $\mathrm{UT0}$. However, small corrections are necessary to account for *variation of latitude*(see section 6.7) and *polar motion* (slippage of Earth's crust with respect to the body of the planet), and when these corrections are made, we arrive at $\mathrm{UT1}$. These corrections are not sufficient, however, to keep Universal Time always in exact agreement with TAI, and whenever $\mathrm{UT1}$ differs from TAI by as much as 0.9 seconds, a *leap second* is added to (or in principle, if not in practice, subtracted from) $\mathrm{UT1}$ to arrive at *Coordinated Universal Time* UTC, which thus never differs from TAI by as much as a second. Leap seconds are typically added at the end of a year, or sometimes in mid-year. The time signals broadcast by radio and over the Internet are UTC, and announcements are made whenever a leap second is inserted. Whenever a leap second is inserted, the minute during which it is inserted has 61 seconds, and an announcement is made. Normally the instant of time at which astronomical events are observed (such as lunar occultations) should be recorded and reported in UTC. Of course, if the observation is not made with a precision of better than a second (e.g. the commencement of a lunar eclipse), one should not pretend that one can distinguish between the various versions of UT, and the time recorded should be "UT". To say "UTC" under such circumstances is to pretend to a greater precision than was actually achieved – rather like quoting a measurement to too many significant figures.

While Universal Time tells us the "time of day" – i.e. how far Earth has rotated on its axis – it is not the argument of time needed in the theoretical calculation of orbital ephemerides. For much of the twentieth century, the time scale used for theoretical ephemeris calculations was *Ephemeris Time*, ET. (I believe a movie was made about Ephemeris Time. At least the title of the movie was *ET*, so I presume that's what it was about, though I haven't actually seen it.) Ephemeris time was based not on the (irregularly rotating) Earth, but in principle on the motion of Earth in its orbit around the Sun, which was presumed to be "uniform". (In practice, ET was calculated from observations of occultations of stars by the Moon, the motion of the Moon in its orbit being supposed to be calculated using a uniformly-flowing Ephemeris Time.) Just as TAI has a unit of time (the SI second) and an initial epoch (1958 January 01), so ET had a unit of time (the mean tropical year) and an initial epoch (1900 January 0^{d} 12^{h} ET).

While ET was much more satisfactory as the "uniformly-flowing" argument of time necessary for ephemeris or other celestial mechanical calculations, it eventually had to be admitted that intervals of *atomic* time could be determined much more precisely than on any other time scale, and consequently ephemeris time (ET) was replaced in 1984 by Terrestrial Dynamical Time (TDT) as the independent argument of supposedly uniformly-flowing time for ephemeris calculations. Unlike ET, the unit of time is not the mean tropical year but it is the SI second of time based on the hyperfine transition of caesium as defined earlier in the chapter. And the starting point for TDT is defined such that at the instant 1977 January 01^{d} 00^{h} 00^{m} 00^{s} TAI was the same instant as 1977 January $01^{\mathrm{d}}.0003725$ Since $0^{\mathrm{d}}.0003725$ is $32^{\mathrm{s}}.184$, TDT is equal to TAI plus $32^{\mathrm{s}}.184$. This was re-named simply Terrestrial Time (TT) in 1991. Like TDT, TT was ahead of TAI by $0^{\mathrm{d}}.003725$ at 1997 January $01^{\mathrm{d}}.0$ TAI, the difference being that the unit of time interval in TT was defined in 1991 as the SI second at mean sea level. (At mean sea level? What has that got to do with it?! Not very much, to be sure, but, for precise timekeeping, it is important because, according to general relativity, the rate of passage of time depends on the gravitational potential.)

In summary, the time signals that are broadcast on short-wave radio or on the Internet are UTC. When you make an observation and record the instant of occurrence of an astronomical phenomenon, you must report the observation in UTC, without converting it to some other scale. The only proviso is that, if your observation is less precise than a second, you should not pretend to greater precision that is warranted by your observation (i.e. you should not pretend that your timing was sufficiently precise that you could distinguish between the various varieties of UT), and it then becomes appropriate to report your observation merely as "UT".

If you are calculating and publishing an ephemeris, the argument of time that you should use in your calculations and which should be published in the ephemeris is TDT. This also applies if you are calculating orbital elements, except that the computer (by which I mean the human being who is doing the calculation or programming a machine to do it) must be aware and cognizant of the fact that the observations that are presented to him or her are given in UTC, and corrections must be made accordingly.

How great is the difference between UTC and TDT? This is given by two quantities, known as ΔT and $\Delta\mathrm{UT1}$, given by

$$\Delta T = \mathrm{TDT} - \mathrm{UT1} \tag{7.1}$$

and

$$\Delta\mathrm{UT1} = \mathrm{UT1} - \mathrm{UTC}, \tag{7.2}$$

from which it follows that

$$\mathrm{TDT} - \mathrm{UTC} = \Delta T + \Delta\mathrm{UT1}, \text{ which is also given the symbol } \Delta\,\mathrm{TT}. \tag{7.3}$$

The values of these corrections are published in *The Astronomical Ephemeris*. They cannot be predicted exactly for a given future year, and their exact values are known only a few years after publication. *The Astronomical Ephemeris* gives a table of $\Delta\mathrm{T}$ since 1620, and a prediction of its value for the current year. In 2000, it was about 63 seconds and increasing at about three-quarters of a second per year.

For most purposes the difference between the UTC used by observers and the TDT used by computers is of little import. After all, from the practical point of view the right ascension and declination of a planet do not change by very much in 63 seconds. An exception may be in the case of a fast-moving near-Earth asteroid. For example, if an asteroid is moving at the very fast rate of 10000 arcseconds per hour, in 63 seconds it will have moved three arcminutes. In principle, an asteroid observer who is "lying in wait" to "ambush" a new asteroid as it "swims into his view" would have to take into account the difference between the TDT of the published ephemeris and the UTC of his clock. In practice, the uncertainties in the elements of a newly-discovered fast-moving asteroid present the observer with more challenges than the challenge of ΔT, so that, even in the case of fast-moving asteroids, it is seldom that the ΔT is the most important difficulty.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

8: PLANETARY MOTIONS

In this chapter, I do not attempt to calculate planetary ephemerides, which will come in a later chapter. Rather, I discuss in an idealistic and qualitative manner how it is that a planet sometimes moves in one direction and sometimes in another. That the treatment in this chapter is both idealistic and qualitative by no means implies that it will be devoid of Equations or of quantitative results, or that the matter discussed in this chapter will have no real practical or observational value.

8.1: Introduction to Planetary Motions

The word "planet" means "wanderer" (πλάνητες αστέρες − wandering stars); in contrast to the "fixed stars", the planets wander around on the celestial sphere, sometimes moving from east to west and sometimes from west to east – and of course there are "stationary points" at the instants when their motions change from one direction to the other.

In this chapter, I do not attempt to calculate planetary ephemerides, which will come in a later chapter. Rather, I discuss in an idealistic and qualitative manner how it is that a planet sometimes moves in one direction and sometimes in another. That the treatment in this chapter is both idealistic and qualitative by no means implies that it will be devoid of Equations or of quantitative results, or that the matter discussed in this chapter will have no real practical or observational value.

I shall assume in this chapter that planets move around the Sun in coplanar circular orbits. Pluto apart, the inclinations of the orbits of the planets are small (Mercury is 7 degrees , Venus 3 degrees and the remainder are smaller), and if you were to draw the most eccentric orbit (Mercury's) to scale, without marking in the position of the Sun, your eye could probably not distinguish the orbit from a circle. Thus these ideal orbits, while not suitable for computing precise ephemerides, are not unrealistic for a general description of the apparent motions of the planets.

I shall assume that the angular speed of Earth in its motion around the Sun, relative to the stars, is $0.017\,202\,098\,95$radians per mean solar day, or $147.841\,150$arcseconds per mean solar hour. In this chapter I shall use the symbol ω_0 for this angular speed, though in many contexts it is also given the symbol k, and is called the gaussian constant.

It may be noted that the *definition of the astronomical unit* (AU) *of distance* is the radius of the orbit of a particle of negligible mass that moves around the Sun in a circular orbit at angular speed $0.017\,202\,098\,95$radians per mean solar day. In other words, the formal definition of the astronomical unit makes no mention of planet Earth. However, to a good approximation, Earth does move around the Sun in a near-circular orbit of about that radius and about that speed, and that is the assumption that will be made in this chapter. [In 2012, the International Astronomical Union redefined the astronomical unit as 149 597 870 700 m exactly, and they recommended the symbol au rather than AU. This makes no substantial difference to the content of this chapter.]

I shall also make the assumption that other planets move around the Sun in coplanar circular orbits at angular speeds that are proportional to $a^{-3/2}$ and hence at linear speeds that are proportional to $a^{-1/2}$, where a is the radius of their orbits. This is, as we shall describe in Chapter 9, Kepler's third law of planetary motion.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

8.2: Opposition, Conjunction and Quadrature

Planets that are closer to the Sun than Earth (i.e. whose orbital radii are less than 1 AU), that is to say the planets Mercury and Venus, are *inferior planets*. (Any asteroids that may be found in such orbits are therefore inferior asteroids, and, technically, any spacecraft that are in solar orbits within that of the orbit of Earth could also be called inferior spacecraft, although it is doubtful whether this nomenclature would ever win general acceptance.) Other planets (i.e. Mars and beyond) are *superior planets*.

In figure VIII.1 I draw the orbits of Earth and of an inferior planet.

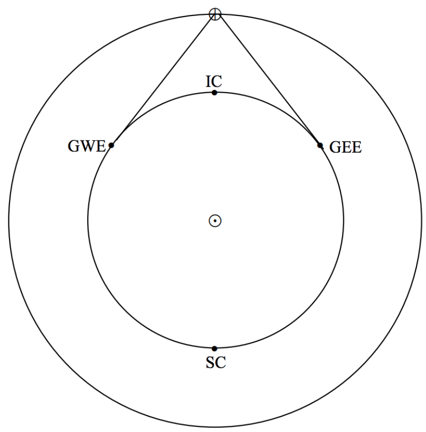


FIGURE VIII.1

The symbol ⊙ denotes the Sun and ⊕ denotes Earth. At \(\text{IC}|), the planet is at *inferior conjunction* with the Sun. At SC, it is at *superior conjunction* with the Sun. At GWE it is at *greatest western elongation* from the Sun. At GEE it is at *greatest eastern elongation*. It should be evident that the sine of the greatest elongation is equal to the radius of the planet's orbit in AU. Thus the radius of Venus's (almost circular) orbit is 0.7233 AU, and therefore its greatest elongation from the Sun is about 46°. Mercury's orbit is relatively eccentric ($e = 0.2056$), so that its distance from the Sun varies from 0.3075 AU at perihelion to 0.4667 at aphelion. Consequently greatest elongations can be from 18° to 28°, depending on where in its orbit they occur.

In figure VIII.2 I draw the orbits of Earth and of a superior planet.

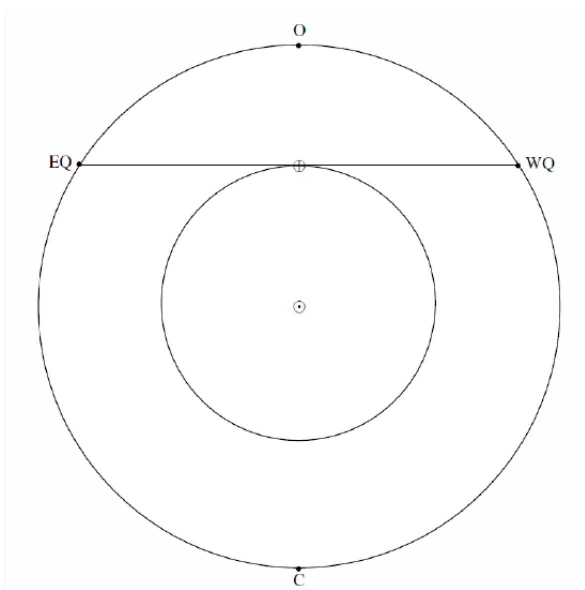


FIGURE VIII.2

At C, the planet is in *conjunction* with the Sun. At O it is in *opposition* to the Sun. The opposition point is very familiar to observers of asteroids. Its right ascension differs from that of the Sun by 12 hours, and it transits across the meridian at midnight local solar time. The points EQ and WQ are *eastern quadrature* and *western quadrature*.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

8.3: Sidereal and Synodic Periods

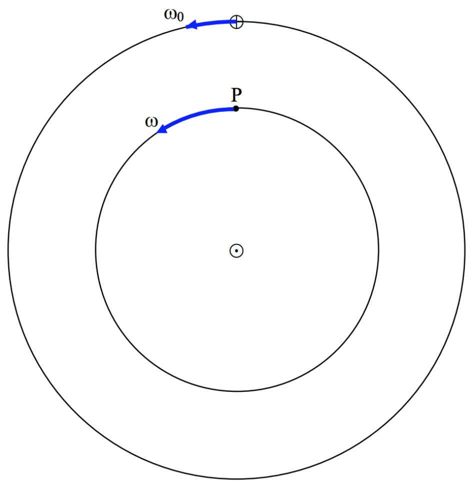


FIGURE VIII.3

Figure VIII.3 shows the orbits of Earth (⊕) and an inferior planet (P). Earth is moving around the Sun at angular speed ω_0 and period $P_0 = 2\pi/\omega_0 = 1$ sidereal year. The planet is moving around the Sun at a faster angular speed ω and shorter period $P_{\text{sid}} = 2\pi/\omega$, which is called the sidereal period of the planet (i.e. the period relative to the fixed stars). The angular speed of the planet with respect to Earth is $\omega_{\text{PE}} = \omega - \omega_0$. The interval between two consecutive inferior conjunctions of the planet is called its *synodic period*, P_{syn}, and is equal to $2\pi/\omega_{\text{PE}}$. Thus, since the relation between angular speed and period is $\omega = 2\pi/P$, we see that

$$\frac{1}{P_{\text{syn}}} = \frac{1}{P_{\text{sid}}} - \frac{1}{P_0}. \quad \text{(inferior planet)} \tag{8.3.1}$$

The reader can draw the situation for a superior planet, and will see that in that case $\omega_{\text{PE}} = \omega_0 - \omega$. The synodic period of the planet is the interval between two consecutive oppositions, and we arrive at

$$\frac{1}{P_{\text{syn}}} = \frac{1}{P_0} - \frac{1}{P_{\text{sid}}}. \quad \text{(superior planet)} \tag{8.3.2}$$

Of all the major planets, Mars has the longest synodic period, namely 780 days, so that it comes to opposition and is easy to observe at intervals of a little more than two years. Mercury has the shortest synodic period, namely 116 days. The synodic periods of all superior planets are greater than one sidereal year. The synodic periods of inferior planets may be less than (Mercury) or greater that (Venus) one sidereal year.

Exercise 8.3.1

An inferior planet in a circular orbit has a synodic period of one sidereal year. What is the radius of its orbit?

Contributor

- Jeremy Tatum (University of Victoria, Canada)

8.4: Direct and Retrograde Motion, and Stationary Points

As seen from the north ecliptic pole, the major planets move counterclockwise around the Sun. Such motion is called *direct* or *prograde* motion. A body moving clockwise (such as some comets) is said to be moving *retrograde*.

In figure VIII.4 I have drawn Earth moving around the Sun at angular speed ω_0 and a superior planet (which I have indicated at opposition and at conjunction) moving with slower angular speed ω.

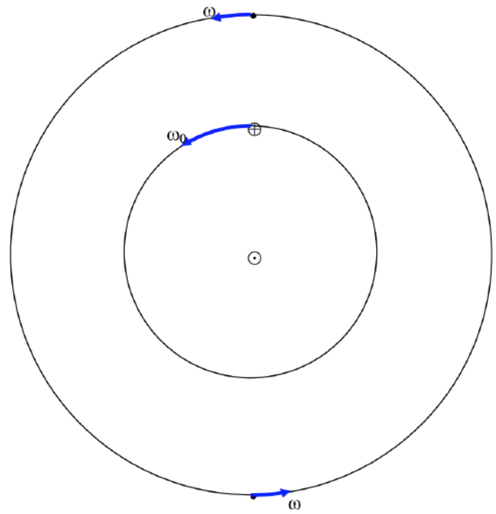


FIGURE VIII.4

In figure VIII.5. I have drawn the same situation but referred to what I call a *synodic* reference frame. That is, a reference frame that is co-rotating with Earth, such that the Earth-Sun line is stationary. In the synodic frame, the planet is moving clockwise at angular speed $\omega - \omega_0$.

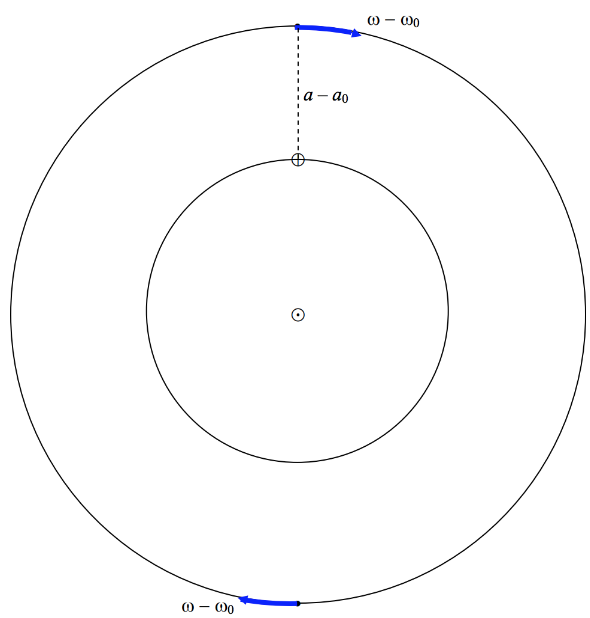


FIGURE VIII.5

Let a_0 and a be the radii or Earth's and the planet's orbit respectively. In that case, the angular speed of the planet in the *sidereal* frame is, by Kepler's third law, $\omega = \omega_0 \left(\frac{a_0}{a}\right)^{3/2}$ counterclockwise, and therefore, in the *synodic* frame, it is $\omega_0 \left[1 - \left(\frac{a_0}{a}\right)^{3/2}\right]$ clockwise. From this point, I am going to express angular speeds in units of ω_0 and distances in astronomical units ($a_0 = 1$). In these units, then, the angular speed of the planet around the Sun in the synodic frame is $1 - a^{-3/2}$ clockwise, and its linear speed in its orbit (of radius a) is $a(1 - a^{-3/2})$.

Now suppose the planet is at opposition, so that its distance from Earth is $a - 1$. The angular speed of the planet *as seen from Earth* is therefore $\frac{a(-a^{-3/2})}{a-1}$ clockwise. For superior planets and asteroids ($a > 1$), this goes from 1.5 to 1.0 as a goes from 1 to ∞. Now in the synodic frame, the celestial sphere with the fixed stars upon it is revolving around Earth at angular speed 1. Therefore, at opposition, the *angular speed of the planet against the background of stars* (also known as the apparent *proper motion*, for which I shall use the symbol p) of the planet is the above expression minus 1, which, after simplification, becomes

$$p = \frac{1 - 1/\sqrt{a}}{a - 1} \tag{8.4.1}$$

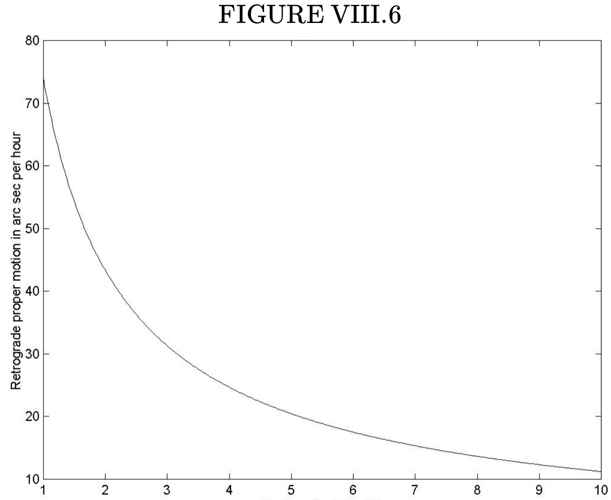
in the direction of decreasing ecliptic longitude or decreasing right ascension – i.e. towards the west. That is to say, at opposition, the planet appears from Earth to be moving in the *retrograde direction*. The reader is reminded that, in Equation 8.4.1, p is the proper motion to the west, in units of $\omega_0 = 147.8$ arcseconds per mean solar hour, and a is the radius of the planet's orbit in AU. A graph of p versus a is shown in figure $\text{VIII}.6$.

Equation 8.4.1 enables us to calculate p given a. The more interesting problem is to calculate a given p. Thus, you are searching for asteroids near the opposition point one night, and a new planet swims into your ken. (That's from a poem by Keats, by the way.) You see that it is moving retrograde with respect to the stars by so many arcseconds per hour. Assuming that it is moving in a circular orbit, what is the radius of its orbit? The quick answer, of course, is to look at figure $\text{VIII}.6$, but you can also keep your hand at high-school algebra in by inverting Equation 8.4.1 to obtain

$$a = \frac{p + 2 - \sqrt{p(p+4)}}{2p}. \tag{8.4.2}$$

Similar considerations for an inferior planet will show that, at inferior conjunction, the angular speed of the planet towards the west is $\frac{a(a^{-3/2} - 1)}{1-a}$, which is the same as the formula for a superior planet at opposition. As a goes from 1 to 0, this goes from 1.5 to ∞. In the synodic frame, the stars are moving westward at angular speed 1, so, relative to the background stars, an inferior planet at inferior conjunction has a retrograde (westward) proper motion given by the same formula as for a superior planet at superior conjunction, namely Equation 8.4.1. A graph of p versus a for an inferior planet drops from ∞ at $a = 0$ to 73.9 arcsec per hour at $a = 1$. Just to keep your algebra skills polished, you can show from Equation 8.4.1 that when $a = 1$, $p = ½$.

FIGURE VIII.6



Thus a *superior* planet at *opposition* moves westward (it "retrogrades") relative to the stars, and an *inferior* planet at *inferior conjunction* also moves westward (it "retrogrades") relative to the stars.

It will, however, be obvious that a superior planet at conjunction, or an inferior planet at superior conjunction, will move eastward ("direct" or "prograde") relative to the stars. Therefore at some point in its orbit a planet will be stationary relative to the stars at the moment when its proper motion changes from direct to retrograde. As seen from Earth, a planet moves generally eastward relative to the stars, except for a short time near opposition (for a superior planet) or inferior conjunction (for an inferior planet) when it briefly retrogrades towards the west. It is small wonder that the ancient astronomers, believing that the Earth was at the centre of the solar system, believed in their system of deferents and epicycles. We would believe the same today if we hadn't read differently in books and on this web site.

Two small words of caution. It is sometimes believed by the unwary that the stationary points in the orbit of an inferior planet occur when the planet is at greatest elongation from the Sun. This is not the case, and indeed there is a small exercise on this point in the penultimate paragraph of this chapter. The second small point to notice is that, for precise work, it is necessary to distinguish between when a planet is stationary (i.e. it is at the moment of changing direction) in right ascension, and when it is stationary in ecliptic longitude. In our simple model of coplanar orbits, we need not make this fine distinction.

In what follows, we are going to calculate (for our concentric circular coplanar model) the angular distance of a superior planet from the opposition point when it is stationary, and the angular distance ("elongation") from the Sun when an inferior planet is stationary. We'll start with a superior planet.

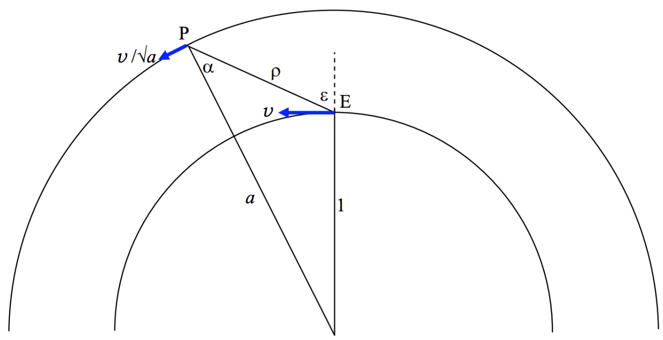


FIGURE VIII.7

Figure VIII.7 shows the Earth E moving in its orbit of radius 1 AU with speed v, and a superior planet or asteroid P moving in its orbit of radius $a \text{ AU}$ with speed $v/\sqrt{a}$. The angle ε is the angular distance of the planet from the opposition point. The angle α is known as the *phase angle*. There is no apparent motion of the planet against the starry background (i.e. the planet is at its stationary point) when the components of the two velocity vectors perpendicular to the line EP are equal. That is, the planet is at a stationary point when $\frac{v}{\sqrt{a}}\cos\alpha = v\cos\varepsilon$, or

$$\cos\alpha = \sqrt{a}\cos\varepsilon. \tag{8.4.3}$$

But from triangle SEP we have

$$a\sin\alpha = \sin\varepsilon. \tag{8.4.4}$$

On elimination of α from Equations 8.4.3 and 8.4.4, we find that the planet is at a stationary point when its angular distance from the opposition point is given by

$$\tan\varepsilon = \frac{a}{\sqrt{1+a}}. \tag{8.4.5}$$

On inversion of this Equation (do it!), we find that the heliocentric distance of a planet which reaches it stationary point at an angular distance ε from the opposition point is

$$a = \frac{1}{2}\left(t + \sqrt{t(t+4)}\right), \tag{8.4.6}$$

where $t = \tan^2\varepsilon$.

The relation between a and ε is shown in figure VIII.8. The least possible angular distance of the stationary point from opposition for a superior planet moving in a circular orbit is $\tan^{-1} 1/(\sqrt{2}) = 35^\circ\ 16' = 02^{\text{h}}\ 21^{\text{m}}$.

FIGURE VIII.8

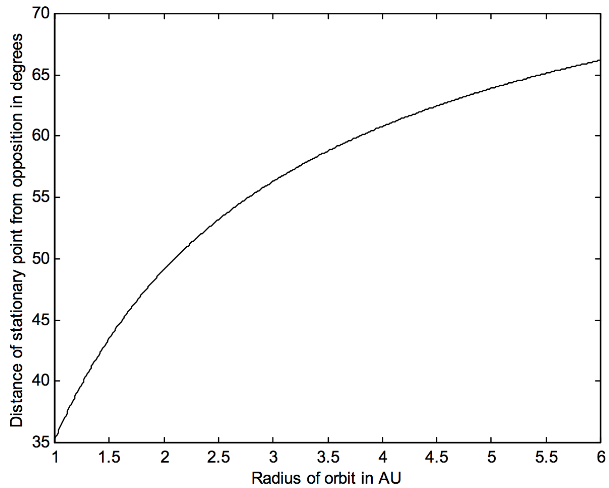
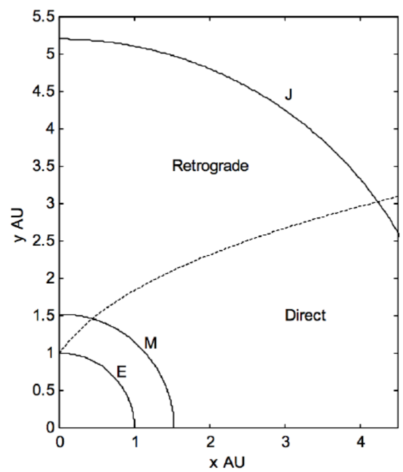


Figure VIII.9, in which the Sun is at the origin, shows the orbits of Earth, Mars and Jupiter, and it divides the area in which asteroids moving in circular orbits will have direct or retrograde proper motions.

FIGURE VIII.9



If the reader carries out the same analysis for inferior planets, he or she will find that Equations 8.4.4 to 8.4.6 apply equally well, except that, in the case of inferior planets (and inferior asteroids, such as the Aten group, of which more are likely to be discovered in the coming years) the angle ε is the angular distance or elongation of the planet from the Sun rather than from the opposition point, and $35^\circ 16'$ is the *greatest* value this may have for the stationary point of an inferior planet in a circular orbit. The Equation corresponding to 8.4.3 becomes, for in inferior planet, $\cos\alpha = -\sqrt{a}\cos\varepsilon$. The elongation of the stationary point is, unsurprisingly, less than the greatest elongation. Also, for an inferior planet, it is to be noted that, for a given elongation (other than greatest elongation) *two* phase angles are possible and *two* geocentric distances are possible. At the stationary point, the obtuse phase angle and the lesser of the two geocentric distances are the correct ones.

Of course in general, we are not likely to be observing an asteroid exactly at the opposition point or exactly at a stationary point. We now tackle the slightly more difficult problem: What is the proper motion of an asteroid whose circular orbital radius is a when it is observed at an angular distance ε from the opposition point (or from the Sun)? Or, conversely, if we observe an asteroid at an angular distance ε from the opposition point, and we see that it has a proper motion p, what is the radius of its (assumed circular) orbit?

In figure 9b we see, in a sidereal reference frame, the orbits of Earth, E, and a superior planet (or asteroid), P, the radii of their orbits being 1 and a AU. The heliocentric and geocentric distances of the planet are a and ρ. The angular distance of the planet from the opposition point is ε and the phase angle is α. Earth is moving with angular speed 1 (in units of ω_0) and the planet is moving (according to Kepler's third law) with angular speed $a^{-3/2}$.

In figure 9c we see the same situation in a synodic reference frame, in which Earth is stationary and the planet is moving clockwise at an angular speed $1-a^{-3/2}$ (in units of ω_0).

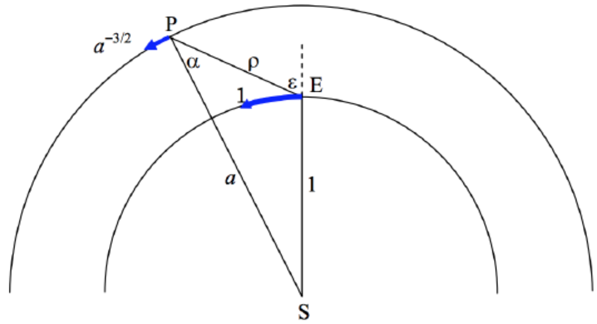


FIGURE VIII.9B

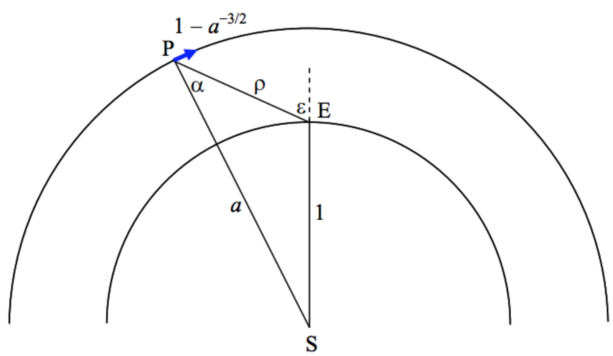


FIGURE VIII.9C

In the synodic frame, the linear speed of the planet (whose angular speed is $1-a^{-3/2}$ and whose heliocentric distance is a) is

$$a(1-a^{-3/2}) = a - \frac{1}{\sqrt{a}}. \tag{8.4.7}$$

The transverse component of this velocity as seen from Earth is

$$\left(a - \frac{1}{\sqrt{a}}\right)\cos\alpha, \tag{8.4.8}$$

so that its angular velocity as seen from Earth is

$$\frac{1}{\rho}\left(a - \frac{1}{\sqrt{a}}\right)\cos\alpha \tag{8.4.9}$$

retrograde.

In the synodic frame, the stars are moving retrograde at angular speed 1. Therefore the planet is moving direct relative to the background stars at angular speed

$$p = 1 - \frac{1}{\rho}\left(a - \frac{1}{\sqrt{a}}\right)\cos\alpha \tag{8.4.10}$$

and this is the required proper motion. In this Equation, geometry shows that

$$a\sin\alpha = \sin\varepsilon \tag{8.4.11}$$

and

$$\rho^2 = 1 + a[a - 2\cos(\varepsilon - \alpha)]. \tag{8.4.12}$$

Thus p can be calculated, given ε and a.

The more interesting and practical problem, however, is that you have observed an asteroid at an angular distance ε from the opposition point, and it is moving at an angular speed p relative to the starry background. (We'll count p as positive if the proper motion is direct – i.e. if the asteroid is moving eastward relative to the stars. The sign of ε does not matter.) You are going to have to invert Equation 8.4.10. I am not sure if this can easily be done algebraically, so your challenge is to write a computer program that will return a numerically given p and ε as input data. It can be done, but I shall not pretend that it is easy.

When you have done this, here are three examples for you:

1. Proper motion = 40 arcsec per hour westward; i.e. $p=-40$ arcsec per hour. $\varepsilon=20^\circ$. Find the heliocentric distance a in AU.

2. $p=+40$ " $/\text{hr}$. $\varepsilon=70^\circ$. Find a.

3. $p=-15$ " $/\text{hr}$. $\varepsilon=40^\circ$. Find a.

I have written my own Fortran program to invert Equation 8.4.10, using Newton-Raphson iteration, and here are the answers it gives me.

1. $a=1.578$ AU
2. $a=1.718$ AU
3. Error message!

My computer failed to do example number 3! In other words, given a proper motion of -15 " $/\text{hr}$ and an opposition distance of 40°, it could not tell me the heliocentric distance!

In figure VIII.10 I have plotted proper motion versus ε for several heliocentric distances, and in figure VIII.11 I have drawn proper motion versus heliocentric distance for several ε. You will find that you can easily find approximate solutions to the first two of these problems from either figure, but you cannot solve the third problem from either figure. In other words, given certain combinations of p and ε, it simply is not possible to determine a. There is a large range of value of a and ε that result in the same proper motion.

If you carry out the same analysis for inferior planets, you will find that the Equations that correspond to Equations 8,4,10-12 are as follows:

$$p=1+\frac{1}{\rho}\left(\frac{1}{\sqrt{a}}-a\right)\cos\alpha. \tag{8.4.13}$$

This is the same as Equation 8.4.10.

$$a\sin\alpha=\sin\varepsilon. \tag{8.4.14}$$

This is the same as Equation 8.4.11, except that, for an inferior planet, ε is the elongation from the Sun and there are two solutions for α, one acute and the other obtuse. As Galileo announced: *Cynthiae figuras aemulatur mater amorum*.

The Equation that corresponds to Equation 8.4.12 is

$$\rho^2=1+a[a+2\cos(\varepsilon+\alpha)], \tag{8.4.15}$$

which differs slightly from Equation 8.4.12

FIGURE VIII.10

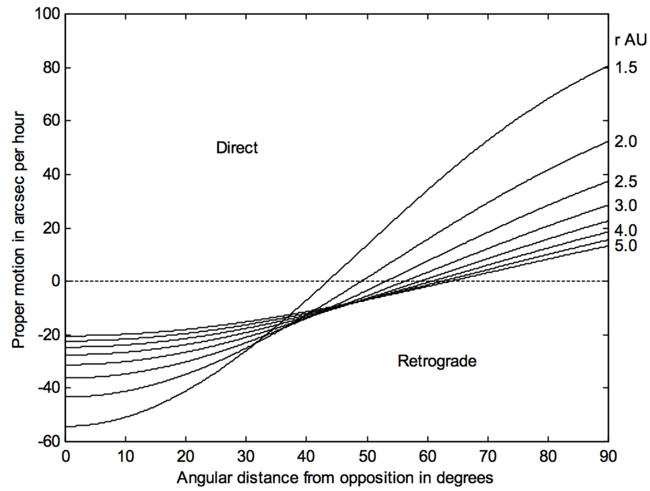
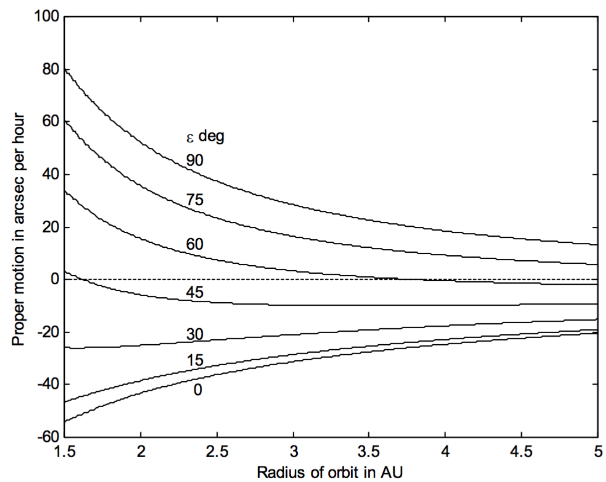
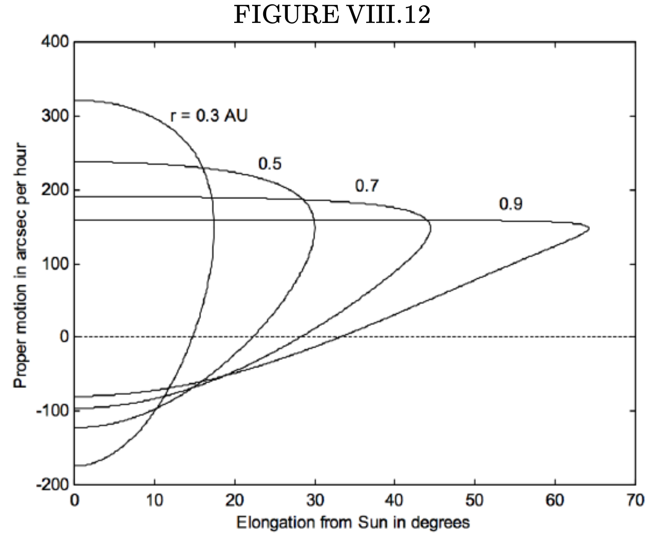


FIGURE VIII.11



In figure VIII.12 I have plotted the proper motion versus elongation from the Sun for several inferior heliocentric distances. You will observe that, for a given elongation and proper motion, there are two possible solutions for a, and there is nothing you can do about it from a single observation of ε and p. For $\varepsilon = 0$ (conjunction with the Sun), the proper motion is positive at superior conjunction and negative at inferior conjunction.

FIGURE VIII.12



As an exercise, you might like to convince yourself – either from the Equations or just from the geometry of the situation − that the proper motion relative to the stars of any inferior planet in a circular orbit at greatest elongation is independent of the radius of the orbit. What is this proper motion in arcsec per hour?

Summary. The graphs and Equations in this section will enable an estimate to be made of the radius of the orbit of an asteroid to be estimated from a single night's observation of its proper motion and angular distance from the opposition point (superior asteroid) or from the Sun (inferior asteroid). The assumptions made are that Earth and asteroid are in coplanar circular orbits. While this is not the case for many asteroids, it is a reasonable approximation for most of the asteroids at least in the main belt. However, there are some combinations of p and ε for which a solution cannot be obtained, and, for inferior asteroids, there are always two possible solutions.

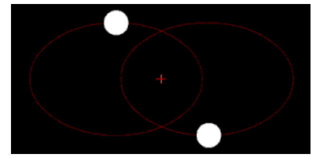
Contributor

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

9: THE TWO BODY PROBLEM IN TWO DIMENSIONS

In this chapter we show how Kepler's laws can be derived from Newton's laws of motion and gravitation, and conservation of angular momentum, and we derive formulas for the energy and angular momentum in an orbit. We show also how to calculate the position of a planet in its orbit as a function of time. The discussion here is limited to two dimensions. The corresponding problem in three dimensions, and how to calculate an ephemeris of a planet or comet in the sky, is discussed elsewhere.



9.1: KEPLER'S LAWS

Kepler's law of planetary motion are as follows: 1. Every planet moves around the Sun in an orbit that is an ellipse with the Sun at a focus. 2. The radius vector from Sun to planet sweeps out equal areas in equal time. 3. The squares of the periods of the planets are proportional to the cubes of their semi major axes.

9.2: KEPLER'S SECOND LAW FROM CONSERVATION OF ANGULAR MOMENTUM

Kepler's second law. that argued a line joining a planet and the Sun sweeps out equal areas during equal intervals of time, can be derived from conservation of angular momentum.

9.3: SOME FUNCTIONS OF THE MASSES

9.4: KEPLER'S FIRST AND THIRD LAWS FROM NEWTON'S LAW OF GRAVITATION

9.5: POSITION IN AN ELLIPTIC ORBIT

9.6: POSITION IN A PARABOLIC ORBIT

When a "long-period" comet comes in from the Oort belt, it typically comes in on a highly eccentric orbit, of which we can observe only a very short arc. Consequently, it is often impossible to determine the period or semi major axis with any degree of reliability or to distinguish the orbit from a parabola. There is therefore frequent occasion to have to understand the dynamics of a parabolic orbit.

9.7: POSITION IN A HYPERBOLIC ORBIT

If an interstellar comet were to encounter the solar system from interstellar space, it would pursue a hyperbolic orbit around the Sun. To date, no such comet with an original hyperbolic orbit has been found, although there is no particular reason why we might not find one some night. However, a comet with a near-parabolic orbit from the Oort belt may approach Jupiter on its way in to the inner solar system, and its orbit may be perturbed into a hyperbolic orbit.

9.8: ORBITAL ELEMENTS AND VELOCITY VECTOR

In two dimensions, an orbit can be completely specified by four orbital elements. Three of them give the size, shape and orientation of the orbit. They are, respectively, a , e and ω . The fourth element is needed to give information about where the planet is in its orbit at a particular time. Usually this is T , the time of perihelion passage.

9.9: OSCULATING ELEMENTS

In practice, an orbit is subject to perturbations, and the planet does not move indefinitely in the orbit that is calculated from the position and velocity vectors at a particular time. The orbit that is calculated from the position and velocity vectors at a particular instant of time is called the osculating orbit, and the corresponding orbital elements are the osculating elements.

9.10: MEAN DISTANCE IN AN ELLIPTIC ORBIT

It is sometimes said that " a " in an elliptic orbit is the "mean distance" of a planet from the Sun. In fact a is the semi major axis of the orbit. Whether and it what sense it might also be the "mean distance" is worth a moment of thought.

9.1: Kepler's Laws

Kepler's law of planetary motion (the first two announced in 1609, the third in 1619) are as follows:

1. Every planet moves around the Sun in an orbit that is an ellipse with the Sun at a focus.
2. The radius vector from Sun to planet sweeps out equal areas in equal times.
3. The squares of the periods of the planets are proportional to the cubes of their semi major axes.

The first law is a consequence of the inverse square law of gravitation. An inverse square law of attraction will actually result in a path that is a *conic section* – that is, an ellipse, a parabola or a hyperbola, although only an ellipse, of course, is a closed orbit. An inverse square law of repulsion (for example, α-particles being deflected by gold nuclei in the famous Geiger-Marsden experiment) will result in a hyperbolic path. An attractive force that is directly proportional to the first power of the distance also results in an elliptical path (a Lissajous ellipse) - for example a mass whirled at the end of a Hooke's law elastic spring - but in that case the centre of attraction is at the centre of the ellipse, rather than at a focus.

We shall derive, in Section 9.5, Kepler's first and third laws from an assumed inverse square law of attraction. The problem facing Newton was the opposite: Starting from Kepler's laws, what is the law of attraction governing the motions of the planets? To start with, he had to invent the differential and integral calculus. This is a far cry from the popular notion that he "discovered" gravity by seeing an apple fall from a tree.

The second law is a consequence of conservation of angular momentum, and would be valid for any law of attraction (or repulsion) as long as the force was entirely radial with no transverse component. We derive it in Section 9.3.

Although a full treatment of the first and third laws awaits Section 9.5, the third law is trivially easy to derive in the case of a *circular* orbit. For example, if we suppose that a planet of mass m is in a circular orbit of radius a around a Sun of mass M, M being supposed to be so much larger than m that the Sun can be regarded as stationary, we can just equate the product of mass and centripetal acceleration of the planet, $ma\omega^2$, to the gravitational force between planet and Sun, GMm/a^2 ; and, with the period being given by $P = 2\pi/\omega$, we immediately obtain the third law:

$$P^2 = \frac{4\pi^2}{GM}a^3. \tag{9.2.1}$$

The reader might like to show that, if the mass of the Sun is not so high that the Sun's motion can be neglected, and that planet and Sun move in circular orbits around their mutual centre of mass, the period is

$$P^2 = \frac{4\pi^2}{G(M+m)}a^3. \tag{9.2.2}$$

Here a is the distance between Sun and planet.

Exercise 9.1.1

Express the period in terms of a_1, the radius of the planet's circular orbit around the centre of mass.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.2: Kepler's Second Law from Conservation of Angular Momentum

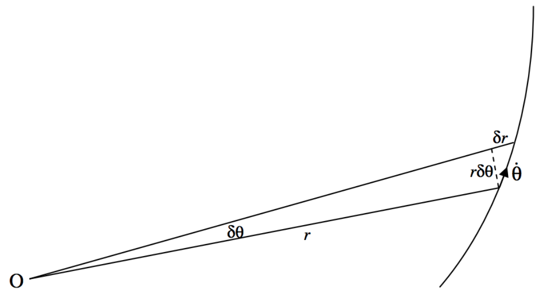


FIGURE IX.1

In figure IX.1, a particle of mass m is moving in some sort of trajectory in which the only force on it is directed towards or away from the point O. At some time, its polar coordinates are (r, θ). At a time δt later these coordinates have increased by δr and $\delta\theta$.

Using the formula one half base times height for the area of a triangle, we see that the area swept out by the radius vector is approximately

$$\delta A = \frac{1}{2}r^2\delta\theta + \frac{1}{2}r\delta\theta\delta r. \tag{9.3.1}$$

On dividing both sides by δt and taking the limit as $\delta t \to 0$, we see that the rate at which the radius vector sweeps out area is

$$\dot{A} = \frac{1}{2}r^2\dot{\theta}. \tag{9.3.2}$$

But the angular momentum is , $mr^2\dot{\theta}$ and since this is constant, the areal speed is also constant. The areal speed, in fact, is half the angular momentum per unit mass.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.3: Some Functions of the Masses

In section 9.5 I am going to consider the motion of two masses, M and m around their mutual centre of mass under the influence of their gravitational attraction. I shall probably want to make use of several functions of the masses, which I shall define here, as follows:

Total mass of the system:

$$\mathbf{M} = M + m. \tag{9.4.1}$$

"Reduced mass"

$$\mathrm{m} = \frac{Mm}{M+m}. \tag{9.4.2}$$

"Mass function":

$$\mathfrak{M} = \frac{M^3}{(M+m)^2}. \tag{9.4.3}$$

No particular name:

$$m_+ = m\left(1 + \frac{m}{M}\right). \tag{9.4.4}$$

Mass ratio:

$$q = m/M. \tag{9.4.5}$$

Mass fraction:

$$\mu = m/(M+m). \tag{9.4.6}$$

The first four are of dimension M; the last two are dimensionless. When $m << M$, $\mathrm{m} \rightarrow m$, $\mathfrak{M} \rightarrow M$ and $m_+ \rightarrow m$.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.4: Kepler's First and Third Laws from Newton's Law of Gravitation

In figure IX.2 I illustrate two masses (they needn't be point masses – as long as they are spherically symmetric, they act gravitationally as if they were point masses) revolving about their common centre of mass C.

FIGURE IX.2

At some time they are a distance r apart, where

$$r=r_1+r_2, \quad r_1=\frac{mr}{M+m}, \quad r_2=\frac{Mr}{M+m} \tag{9.5.1}$$

The Equations of motion of m in polar coordinates (with C as pole) are

- Radial:

$$\ddot{r}_2-r_2\dot{\theta}^2=-GM/r^2. \tag{9.5.2}$$

- Transverse:

$$r_2\ddot{\theta}+2\dot{r}_2\dot{\theta}=0. \tag{9.5.3}$$

Elimination of t between these Equations will in principle give us the Equation, in polar coordinates, of the path.

A slightly easier approach is to write down expressions for the angular momentum and the energy. The angular momentum per unit mass of m with respect to C is

$$h_2=r_2^2\dot{\theta}. \tag{9.5.4}$$

The speed of m is $\sqrt{\dot{r}_2^2+r_2^2\dot{\theta}^2}$ and the speed of M is m/M times this. Some effort will be required of the reader to determine that the total energy E of the system is

$$E=\frac{1}{2}m_+\left(\dot{r}_2^2+r_2^2\dot{\theta}^2\right)-\frac{GM^2\mu}{r^2}. \tag{9.5.5}$$

[It is possible that you may have found this line quite difficult. The reason for the difficulty is that we are not making the approximation of a planet of negligible mass moving around a stationary Sun, but we are allowing both bodies to have comparable masses and the move around their common centre of mass. You might first like to try the simpler problem of a planet of negligible mass moving around a stationary Sun. In that case $r_1=0$ and $r=r_2$ and $\mathrm{m}\rightarrow m$, $\mathfrak{M}\rightarrow M$ and $m_+\rightarrow m$.]

It is easy to eliminate the time between Equations 9.5.4 and 9.5.5. Thus you can write

$$\dot{r}_2=\frac{dr_2}{dt}=\frac{dr_2}{d\theta}\cdot\frac{d\theta}{dt}=\dot{\theta}\frac{dr_2}{d\theta} \tag{9.4.1}$$

and then use Equation 9.5.4 to eliminate $\dot{\theta}$. You should eventually obtain

$$\frac{m_+h_2^2}{r_2^4}\left[\left(\frac{dr_2}{d\theta}\right)^2+r_2^2\right]=2E+\frac{2GM^2\mu}{r^2}. \tag{9.5.6}$$

This is the differential Equation, in polar coordinates, for the path of m. All that is now required is to integrate it to obtain r_2 as a function of θ.

At first, integration looks hopelessly difficult, but it proceeds by making one tentative substitution after another to see if we can't make it look a little easier. For example, we have (if we multiply out the square bracket) r_2 in the denominator three times in the Equation. Let's at least try the substitution $w=1/r_2$. That will surely make it look a little easier. You will have to use

$$\frac{dr_2}{d\theta}=\frac{dr_2}{dw}\frac{dw}{d\theta}=-\frac{1}{w^2}\frac{dw}{d\theta},\tag{9.4.2}$$

and after a little algebra you should obtain

$$\left(\frac{dw}{d\theta}\right)^2+\left(w-\frac{GM^2\mu}{m_+h_2^2}\right)^2=\frac{2E}{m_+h_2^2}+\frac{G^2M^4\mu^2}{m_+^2h_2^4}.\tag{9.5.7}$$

This may at first sight not look like much of an improvement, but the right hand side is just a lot of constants, and, since it is positive, let's call the right hand side H^2. (In case you doubt that the right hand side is positive, the left hand side certainly is!) Also, make the obvious substitution

$$u=w-\frac{GM^2\mu}{m_+h_2^2},\tag{9.4.3}$$

and the Equation becomes almost trivial:

$$\left(\frac{du}{d\theta}\right)^2+u^2=H^2,\tag{9.5.9}$$

from which we proceed to

$$\int d\theta=\pm\int\frac{du}{\sqrt{H^2-u^2}}.\tag{9.5.10}$$

At this stage you can choose either the $+$ or the $-$ and you can choose to make the next substitution $u=H\sin\phi$ or $u=H\cos\phi$; you'll get the same result in the end. I'll choose the plus sign and I'll let $u=H\cos\phi$, and I get $\int d\theta=-\int d\phi$ and hence

$$\theta=-\phi+\omega,\tag{9.5.11}$$

where ω is the arbitrary constant of integration. Now you have to go back and remember what ϕ was, what u was and what w was and what H was. Thus $\theta-\omega=-\phi,\ \therefore\cos(\theta-\omega)=\cos(-\phi)=cos\phi=u/H=\dots$ and so on. Your aim is to get it in the form $r_2=$ function of θ, and, if you persist, you should eventually get

$$r_2=\frac{m_+h_2^2/(GM^2\mu)}{1+\left(1+\frac{2Eh_2^2m_+}{G^2M^4\mu^2}\right)^{1/2}\cos(\theta-\omega)}.\tag{9.5.12}$$

You'll immediately recognize this from Equation 2.3.37 or 2.4.16 or 2.5.18:

$$r=\frac{l}{1+e\cos(\theta-\omega)}\tag{2.3.37}$$

as being the polar Equation to a conic section (ellipse, parabola or hyperbola). Equation 9.5.12 is the Equation of the path of the mass m *about the centre of mass* of the two bodies. The eccentricity is

$$e=\left(1+\frac{2Eh_2^2m_+}{G^2M^4\mu^2}\right)^{1/2},\tag{9.5.13}$$

or, if you now recall what are meant by μ and m_+,

$$e=\left(1+\frac{2Eh_2^2}{G^2}\cdot\frac{(M+m)^3}{M^5m}\right)^{1/2}.\tag{9.5.14}$$

(Check the dimensions of this!)

The eccentricity is less than 1, *equal to* 1, *or greater than* 1 *(i.e. the path is an ellipse, a parabola or a hyperbola) according to whether the total energy E is negative, zero or positive.*

The semi latus rectum of the path of m relative to the centre of mass is of length

$$l_2 = \frac{m_+ h_2^2}{GM^2\mu}, \tag{9.5.15}$$

or

$$l_2 = \frac{h_2^2}{G} \cdot \frac{(M+m)^2}{M^3}. \tag{9.5.16}$$

(Check the dimensions of this!)

We can also write Equations 9.5.15 or 9.5.16 as

$$h_2^2 = G\mathfrak{M} l_2. \tag{9.5.17}$$

At this point it is useful to recall what we mean by $\mathfrak{M}$ and by h_2. $\mathfrak{M}$ is the mass function, given by Equation 9.4.3:

$$\mathfrak{M} = \frac{M^3}{(M+m)^2}. \tag{9.4.3}$$

Let us suppose that the total energy is negative, so that the orbits are elliptical. The two masses are revolving in similar elliptic orbits around the centre of masses; the semi latus rectum of the orbit of m is l_2, and the semi latus rectum of the orbit of M is l_1, where

$$\frac{l_2}{l_1} = \frac{M}{m}. \tag{9.5.18}$$

Relative to M the mass m is revolving in a larger but still similar ellipse with semi latus rectum l given by

$$l = \frac{M+m}{M} l_2. \tag{9.5.19}$$

I am now going to define h as the *angular momentum per unit mass of* m *relative to* M. In other words, we are working in a frame in which M is stationary and m is moving around M in an elliptic orbit of semi latus rectum l. Now angular momentum per unit mass is proportional to the areal speed, and therefore it is proportional to the square of the semi latus rectum. Thus we have

$$\frac{h}{h_2} = \left(\frac{l}{l_2}\right)^2 = \left(\frac{M+m}{M}\right)^2 \tag{9.5.20}$$

Combining Equations 9.5.18, 9.4.3, 9.5.19, 9.5.20 and 9.4.1 we obtain

$$h^2 = G\mathbf{M} l, \tag{9.5.21}$$

where $\mathbf{M}$ is the total mass of the system.

Once again:

The angular momentum per unit mass of m relative to the centre of mass is , $\sqrt{G\mathfrak{M} l_2}$ where l_2 is the semi latus rectum of the orbit of m relative to the centre of mass, and it is $\sqrt{G\mathbf{M} l}$ relative to M, where l is the semi latus rectum of the orbit of m relative to M.

If you were to start this analysis with the assumption that $m << M$, and that M remains stationary, and that the centre of mass coincides with M, you would find that either Equation 9.5.17 or 9.5.21 reduces to

$$h^2 = GMl. \tag{9.5.22}$$

The period of the elliptic orbit is area ÷ areal speed. The area of an ellipse is $\pi ab = \pi a^2\sqrt{1-e^2}$, and the areal speed is half the angular momentum per unit mass (see section 9.3) $= \frac{1}{2}h = \frac{1}{2}\sqrt{G\mathbf{M} l} = \frac{1}{2}\sqrt{G\mathbf{M} a(1-e^2)}$. Therefore the period is $P = \frac{2\pi}{\sqrt{G\mathbf{M}}} a^{3/2}$, or

$$P^2 = \frac{4\pi^2}{G\mathbf{M}} a^3, \tag{9.5.23}$$

which is Kepler's third law.

We might also, while we are at it, express the eccentricity (Equation) in terms of h rather than h_2, using Equation 9.5.20. We obtain:

$$e = \left(1 + \frac{2Eh^2}{G^2 Mm(M+m)}\right)^{1/2}. \tag{9.5.24}$$

If we now substitute for h^2 from Equation 9.5.21, and invert Equation 9.5.24, we obtain, for the energy of the system

$$E = \frac{Gm(M+m)(e^2-1)}{2l}, \tag{9.5.25}$$

or for the energy or the system per unit mass of m:

$$\mathrm{E} = \frac{G\mathbf{M}\left(e^2-1\right)}{2l}. \tag{9.5.26}$$

Here $\mathbf{M}$ is the mass of the system – i.e. $M+m$. E in Equation 9.5.25 is the total energy of the system, which includes the kinetic energy of both masses as well as the mutual potential energy of the two, while E in Equation 9.5.26 is merely E/m. That is, it is, as stated, the energy of the system per unit mass of m.

Equations 9.5.21 and 9.5.26 apply to any conic section. For the different types of conic section they can be written:

For an ellipse:

$$h = \sqrt{G\mathbf{M}a(1-e^2)}, \quad \mathrm{E} = -\frac{G\mathbf{M}}{2a} \tag{9.5.27a,b}$$

For a parabola:

$$h = \sqrt{2G\mathbf{M}q}, \quad \mathrm{E} = 0 \tag{9.5.28a,b}$$

For a hyperbola:

$$h = \sqrt{G\mathbf{M}a(e^2-1)}, \quad \mathrm{E} = +\frac{G\mathbf{M}}{2a} \tag{9.5.29a,b}$$

We see that the energy of an elliptic orbit is determined by the semi major axis, whereas the angular momentum is determined by the semi major axis and by the eccentricity. For a given semi major axis, the angular momentum is greatest when the orbit is circular.

Still referring the orbit of m with respect to M, we can find the speed V of m by noting that

$$\mathrm{E} = \frac{1}{2}V^2 - \frac{G\mathbf{M}}{r} \tag{9.5.30}$$

and, by making use of the b-parts of Equations 9.5.27-29, we find the following relations between speed of m in an orbit versus distance from M:

Ellipse:

$$V^2 = G\mathbf{M}\left(\frac{2}{r} - \frac{1}{a}\right). \tag{9.5.31}$$

Parabola:

$$V^2 = \frac{2G\mathbf{M}}{r}. \tag{9.5.32}$$

Hyperbola:

$$V^2 = G\mathbf{M}\left(\frac{2}{r} + \frac{1}{a}\right). \tag{9.5.33}$$

Circle:

$$V^2 = \frac{G\mathbf{M}}{a}. \tag{9.5.34}$$

Exercise: Show that in an elliptic orbit, the speeds at perihelion and aphelion are, respectively, $\sqrt{\frac{GM}{a}\left(\frac{1+e}{1-e}\right)}$ and $\sqrt{\frac{GM}{a}\left(\frac{1-e}{1+e}\right)}$ and that the ratio of perihelion to aphelion speed is, therefore, $\frac{1+e}{1-e}$.

It might be noted at this point, from the definition of the astronomical unit (Chapter 8, section 8.1), that if distances are expressed in astronomical units, periods and time intervals in sidereal years, GM (where M is the mass of the Sun) has the value $4\pi^2$. The mass of a comet or asteroid is much smaller than the mass of the Sun, so that $\mathbf{M} = M + m \simeq M$. Thus, using these units, and to this approximation, Equation 9.5.23becomes merely $P^2 = a^3$.

A Delightful Construction

I am much indebted to Dr Bob Rimmer, for the following delightful construction. Dr Rimmer found it in the recent book *Feynman's Lost Lecture, The Motion of the Planets Around the Sun*, by D.L. and J.R. Goodstein, and Feynman in his turn ascribed it to a passage (Section IV, Lemma XV) in the *Principia* of Sir Isaac Newton. It has no doubt changed slightly with each telling, and I present it here as follows.

C is a circle of *radius* $2a$ (Figure IX.3). F is the centre of the circle, and F' is a point inside the circle such that the distance $\mathrm{FF}' = 2ae$, where $e < 1$. Join F and F' to a point Q on the circle. MP' is the perpendicular bisector of $\mathrm{F}'\mathrm{Q}$, meeting FQ at P.

The reader is invited to show that, as the point Q moves round the circle, the point P describes an ellipse of eccentricity e, with F and F' as foci, and that MP' is tangent to the ellipse.

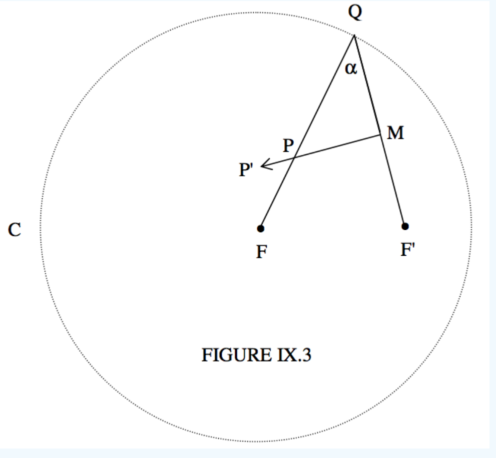


FIGURE IX.3

Hint: It is *very* easy – no math required! Draw the line $\mathrm{F}'\mathrm{P}$, and let the lengths of FP and $\mathrm{F}'\mathrm{P}$ be r and r' respectively. It will then become very obvious that $r + r'$ is always equal to $2a$, and hence P describes an ellipse. By looking at an isosceles triangle, it will also be clear that the angles $\mathrm{F}'\mathrm{PM}$ and FPP' are equal, thus satisfying the focus-to-focus reflection property of an ellipse, so that MP' is tangent to the ellipse.

But there is better to come. You are asked to find the length QF' in terms of a, e and r', or a, e and r.

An easy way to do it is as follows. Let $\mathrm{QF}' = 2p$, so that $\mathrm{QM} = p$. From the right-angled triangle QMP we see that $\cos\alpha = p/r'$. Apply the cosine rule to triangle QFF' to find another expression for $\cos\alpha$, and eliminate $\cos\alpha$ from your two Equations. You should quickly arrive at

$$p^2 = a^2(1-e^2) \times \frac{r'}{2a - r'}. \tag{9.5.35}$$

And, since $r' = 2a - r$, this becomes

$$p = a\sqrt{(1-e)^2} \times \sqrt{\frac{2a-r}{r}} = a^{3/2}\sqrt{(1-e)^2} \times \sqrt{\frac{2}{r} - \frac{1}{a}}. \tag{9.5.36}$$

Now the speed at a point P on an elliptic orbit, in which a planet of negligible mass is in orbit around a star of mass M is given by

$$V = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)}. \tag{9.5.37}$$

Thus we arrive at the result that the length of $\text{F}'\text{Q}$ (or of $\text{F}'\text{M}$) is proportional to the speed of a planet P moving around the Sun F in an elliptic orbit, and of course the direction MP', being tangent to the ellipse, is the direction of motion of the planet. Figure $\text{IX}.4$ shows the ellipse.

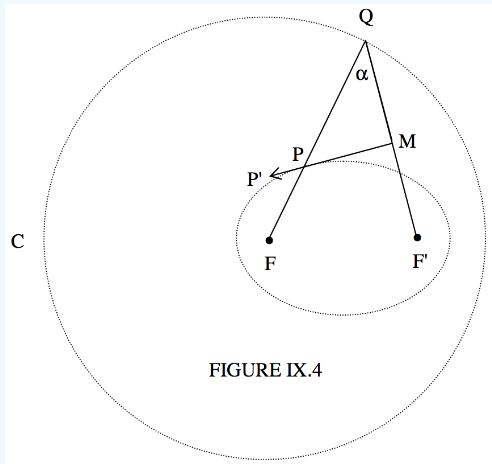


FIGURE IX.4

It is left to the reader to investigate what happens it F' is outside, or on, the circle

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.5: Position in an Elliptic Orbit

The reader might like to refer back to Section 2.3, especially the part that deals with the polar Equation to an ellipse, to be reminded of the meanings of the angles θ, ω and v, which, in an astronomical context, are called, respectively, the *argument of latitude*, the *argument of perihelion* and the *true anomaly*. In this section I shall choose the initial line of polar coordinates to coincide with the major axis of the ellipse, so that ω is zero and $\theta = v$. The Equation to the ellipse is then

$$r = \frac{l}{1+e\cos v}. \tag{9.6.1}$$

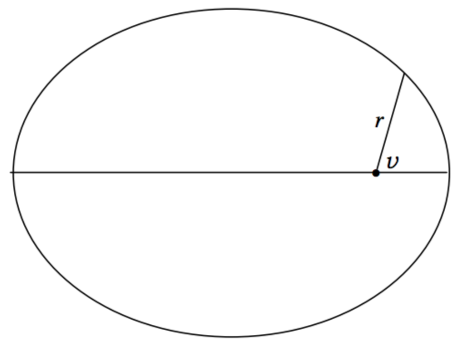


FIGURE IX.5

I'll suppose that a planet is at perihelion at time $t = T$, and the aim of this section will be to find v as a function of t. The semi major axis of the ellipse is a, related to the semi latus rectum by

$$l = a(1-e^2) \tag{9.6.2}$$

and the period is given by

$$P^2 = \frac{4\pi^2}{G\mathbf{M}}a^3. \tag{9.6.3}$$

Here the planet, of mass m is supposed to be in orbit around the Sun of mass M, and the origin, or pole, of the polar coordinates described by Equation 9.6.1 is the Sun, rather than the centre of mass of the system. As usual, $\mathbf{M} = M + m$.

The radius vector from Sun to planet does not move at constant speed (indeed Kepler's second law states how it moves), but we can say that, over a complete orbit, it moves at an *average* angular speed of $2\pi/P$. The angle $\frac{2\pi}{P}(t-T)$ is called the *mean anomaly* of the planet at a time $t-T$ after perihelion passage. It is generally denoted by the letter M, which is already overworked in this chapter for various masses and functions of the masses. For mean anomaly, I'll try this font: $\mathcal{M}$. Thus

$$\mathcal{M} = \frac{2\pi}{P}(t-T). \tag{9.6.4}$$

The first step in our effort to find v as a function of t is to calculate the *eccentric anomaly* E from the mean anomaly. This was defined in figure II.11 of Chapter 2, and it is reproduced below as figure IX.6.

In time $t-T$, the area swept out by the radius vector is the area FBP, and, because the radius vector sweeps out equal areas in equal times, this area is equal to the fraction $(t-T)/P$ of the area of the ellipse. In other words, this area is $\frac{(t-T)\pi ab}{P}$. Now look at the area FBP′. Every ordinate of that area is equal to a/b times the corresponding ordinate of FBP, and therefore the area of FBP′ is $\frac{(t-T)\pi a^2}{P}$. The area FBP′ is also equal to the sector OP′B minus the triangle OP′F. The area of the sector OP′B is $\frac{E}{2\pi}\times\pi a^2 = \frac{1}{2}Ea^2$, and the area of the triangle OP′F is $\frac{1}{2}ae\times a\sin E = \frac{1}{2}a^2 e\sin E$.

$$\therefore \frac{(t-T)\pi a^2}{P} = \frac{1}{2}Ea^2 - \frac{1}{2}a^2 e\sin E. \tag{9.5.1}$$

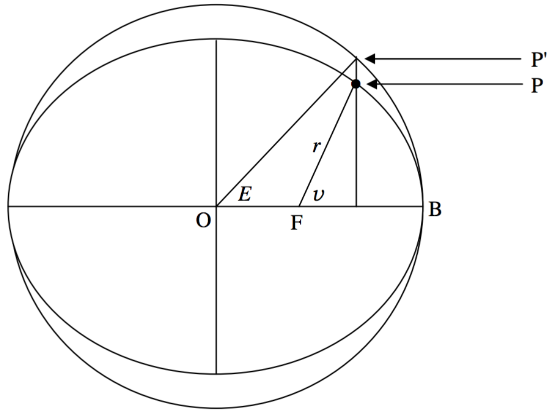


FIGURE IX.6

Multiply both sides by $2/a^2$, and recall Equation 9.6.4, and we arrive at the required relation between the mean anomaly $\mathcal{M}$ and the eccentric anomaly E:

$$\mathcal{M} = E - e\sin E. \tag{9.6.5}$$

This is *Kepler's Equation*.

The first step, then, is to calculate the mean anomaly $\mathcal{M}$ from Equation 9.6.4, and then calculate the eccentric anomaly E from Equation 9.6.5. This is a transcendental Equation, so I'll say a word or two about solving it in a moment, but let's press on for the time being. We now have to calculate the true anomaly v from the eccentric anomaly. This is done from the geometry of the ellipse, with no dynamics, and the relation is given in Chapter 2, Equations 2.3.16 and 2.3.17c, which are reproduced here:

$$\cos v = \frac{\cos E - e}{1 - e\cos E}. \tag{2.3.16}$$

From trigonometric identities, this can also be written

$$\sin v = \frac{\sqrt{1-e^2}\,\sin E}{1 - e\cos E}, \tag{2.3.17a}$$

or

$$\tan v = \frac{\sqrt{1-e^2}\,\sin E}{\cos E - e} \tag{2.3.17b}$$

or

$$\tan\frac{1}{2}v = \sqrt{\frac{1+e}{1-e}}\tan\frac{1}{2}E. \tag{2.3.17c}$$

If we can just solve Equation 9.6.5 (Kepler's Equation), we shall have done what we want – namely, find the true anomaly as a function of the time.

The solution of Kepler's Equation is in fact very easy. We write it as

$$f(E) = E - e\sin E - \mathcal{M} \tag{9.6.6}$$

from which

$$f'(E) = 1 - e\cos E, \tag{9.6.7}$$

and then, by the usual Newton-Raphson process:

$$E = \frac{\mathcal{M} - e(E\cos E - \sin E)}{1 - e\cos E}. \tag{9.6.8}$$

The computation is then extraordinarily rapid (especially if you store cos E and don't calculate it twice!).

Example 9.5.1

Suppose $e = 0.95$ and that $M = 245^\circ$. Calculate E. Since the eccentricity is very large, one might expect the convergence to be slow, and also E is likely to be very different from $\mathcal{M}$, so it is not easy to make a first guess for E. You might as well try 245° for a first guess for E. You should find that it converges to ten significant figures in a mere four iterations. Even if you make a mindlessly stupid first guess of $E = 0^\circ$, it converges to ten significant figures in only nine iterations.

There are a few exceptional occasions, hardly ever encountered in practice, and only for eccentricities greater than about 0.99, when the Newton-Raphson method will not converge when you make your first guess for E equal to $\mathcal{M}$. Charles and Tatum (*Celestial Mechanics and Dynamical Astronomy* **69**, 357 (1998)) have shown that the Newton-Raphson method will *always* converge if you make your first guess $E = \pi$. Nevertheless, the situations where Newton-Raphson will not converge with a first guess of $E = \mathcal{M}$ are unlikely to be encountered except in almost parabolic orbits, and usually a first guess of $E = \mathcal{M}$ is faster than a first guess of $E = \pi$. The chaotic behaviour of Kepler's Equation on these exceptional occasions is discussed in the above paper and also by Stumpf (Cel. Mechs. and Dyn. Astron. **74**, 95 (1999)) and references therein.

Exercise 9.5.1

Show that a good first guess for E is

$$E = \mathcal{M} + x(1 - \frac{1}{2}x^2), \tag{9.6.9}$$

where

$$x = \frac{e\sin\mathcal{M}}{1 - e\cos\mathcal{M}}. \tag{9.6.10}$$

Exercise 9.5.2

Write a computer program in the language of your choice for solving Kepler's Equation. The program should accept e and $\mathcal{M}$ as input, and return E as output. The Newton-Raphson iteration should be terminated when $|(E_{\text{new}} - E_{\text{old}})/E_{\text{old}}$ is less than some small fraction to be determined by you.

Exercise 9.5.1

An asteroid is moving in an elliptic orbit of semi major axis 3AU and eccentricity 0.6. It is at perihelion at time = 0. Calculate its distance from the Sun and its true anomaly one sidereal year later. You may take the mass of the asteroid and the mass of Earth to be negligible compared with the mass of the Sun. In that case, Equation 9.6.3 is merely

$$P^2 = \frac{4\pi^2}{GM}a^3, \tag{9.5.2}$$

where M is the mass of the Sun, and, if P is expressed in sidereal years and a in AU, this becomes just $P^2 = a^3$. Thus you can immediately calculate the period in years and hence, from Equation 9.6.4 you can find the mean anomaly. From there, you have to solve Kepler's Equation to get the eccentric anomaly, and the true anomaly from Equation 2.3.16 or 17. Just make sure that you get the quadrant right.

Exercise 9.5.1

Write a computer program that will give you the true anomaly and heliocentric distance as a function of time since perihelion passage for an asteroid whose elliptic orbit is characterized by a, e. Run the program for the asteroid of the previous exercise for every day for a complete period.

You are now making some real progress towards ephemeris computation!

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.6: Position in a Parabolic Orbit

When a "long-period" comet comes in from the Oort belt, it typically comes in on a highly eccentric orbit, of which we can observe only a very short arc. Consequently, it is often impossible to determine the period or semi major axis with any degree of reliability or to distinguish the orbit from a parabola. There is therefore frequent occasion to have to understand the dynamics of a parabolic orbit.

We have no mean or eccentric anomalies. We must try to get v directly as a function of t without going through these intermediaries.

The angular momentum per unit mass is given by Equation 9.5.28a:

$$h = r^2\dot{v} = \sqrt{2G\mathbf{M}q}, \tag{9.7.1}$$

where v is the true anomaly and q is the perihelion distance.

But the Equation to the parabola (see Equation 2.4.16) is

$$r = \frac{2q}{1+\cos v}, \tag{9.7.2}$$

or (see section 3.8 of Chapter 3), by making use of the identity

$$\cos v = \frac{1-u^2}{1+u^2}, \quad \text{where} \quad u = \tan\frac{1}{2}v, \tag{9.7.3a,b}$$

the Equation to the parabola can be written

$$r = q\sec^2\frac{1}{2}v. \tag{9.7.4}$$

Thus, by substitution of Equation 9.7.4 into 9.7.1 and integrating, we obtain

$$q^2\int_0^v \sec^4(\frac{1}{2}v)dv = \sqrt{2G\mathbf{M}q}\int_T^t dt. \tag{9.7.5}$$

Upon integration (drop me an email if you get stuck!) this becomes

$$u + \frac{1}{3}u^3 = \frac{\sqrt{\frac{1}{2}G\mathbf{M}}}{q^{3/2}}(t-T). \tag{9.7.6}$$

This Equation, when solved for u (which, remember, is $\tan\frac{1}{2}v$), gives us v as a function of t. As explained at the end of section 9.5, if q is in astronomical units and $t-T$ is in sidereal years, and if the mass of the comet is negligible compared with the mass of the Sun, this becomes

$$u + \frac{1}{3}u^3 = \frac{\pi\sqrt{2}(t-T)}{q^{3/2}} \tag{9.7.7}$$

or

$$3u + u^3 - C = 0, \quad \text{where} \quad C = \frac{\pi\sqrt{18}(t-T)}{q^{3/2}}. \tag{9.7.8a,b}$$

There is a choice of methods available for solving Equation 9.7.8a,b, so it might be that the only difficulty is to decide which of the several methods you want to use! The constant $\frac{1}{3}C$ is sometimes called the "parabolic mean anomaly".

Method 1: Just solve it by Newton-Raphson iteration. Thus $f = 3u + u^3 - C = 0$ and $f' = 3(1+u^2)$, so that the Newton-Raphson $u = u - f/f'$ becomes

$$u = \frac{2u^3 + C}{3(1+u^2)}, \tag{9.7.9}$$

which should converge quickly. For economy, calculate u^2 only once per iteration.

Method 2:

Let

$$u = x - 1/x \quad \text{and} \quad C = c - 1/c. \tag{9.7.10a,b}$$

Then Equation 9.7.8a becomes

$$x = c^{1/3}. \tag{9.7.11}$$

Thus, as soon as c is found, x, u and v can be calculated from Equations 9.7.11, 10a, and 3a or b, and the problem is finished – as soon as c is found!

So, how do we find c? We have to solve Equation 9.7.10b.

Method 2a:

Equation 9.7.10b can be written as a quadratic Equation:

$$c^2 - Cc - 1 = 0. \tag{9.7.12}$$

Just be careful that you choose the correct root; you should end with v having the same sign as $t - T$.

Method 2b:

Let

$$C = 2\cot 2\phi \tag{9.7.13}$$

and calculate ϕ. But by a trigonometric identity,

$$2\cot 2\phi = \cot\phi - 1/\cot\phi \tag{9.7.14}$$

so that, by comparison with Equation 9.7.10b, we see that

$$c = \cot\phi. \tag{9.7.15}$$

Again, just make sure that you choose the right quadrant in calculating ϕ from Equation 9.7.13, so as to be sure that you end with v having the same sign as $t - T$.

Method 3.

I am told that Equation 9.7.8 has the exact analytic solution

$$u = \frac{1}{2}w^{\frac{1}{3}} - 2w^{-\frac{1}{3}}, \tag{9.7.16}$$

where

$$w = 4C + \sqrt{64 + 16C^2}. \tag{9.7.17}$$

I haven't verified this for myself, so you might like to have a go.

Example: Solve the Equation $3u + u^3 = 1.6$ by all four methods. (Methods 1, 2a, 2b and 3.)

Example: A comet is moving in an elliptic orbit with perihelion distance $0.9\ \text{AU}$. Calculate the true anomaly and heliocentric distance 20 days after perihelion passage. (A sidereal year is 365.25636 days.)

Exercise: Write a computer program that will return the true anomaly as a function of time, given the perihelion distance of a parabolic orbit. Test it with your answer for the previous example.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.7: Position in a Hyperbolic Orbit

If an interstellar object were to encounter the solar system from interstellar space, it would pursue a hyperbolic orbit around the Sun. The first known such object with an original hyperbolic orbit was detected in 2017, and was given the name Oumuamua. However, a comet with a near-parabolic orbit from the Oort belt may approach Jupiter on its way in to the inner solar system, and its orbit may be perturbed into a hyperbolic orbit. This will result in its ultimate loss from the solar system. Several examples of such cometary orbits are known. There is evidence, from radar studies of meteors, of meteoroidal dust encountering Earth at speeds that are hyperbolic with respect to the Sun, although whether these are on orbits that are originally hyperbolic (and are therefore from interstellar space) or whether they are of solar system origin and have been perturbed by Jupiter into hyperbolic orbits is not known.

I must admit to not having actually carried out a calculation for a hyperbolic orbit, but I think we can just proceed in a manner similar to an ellipse or a parabola. Thus we can start with the angular momentum per unit mass:

$$h = r^2\dot{v} = \sqrt{G\mathbf{M}l}, \tag{9.8.1}$$

where

$$r = \frac{l}{1+e\cos v} \tag{9.8.2}$$

and

$$l = a(e^2-1). \tag{9.8.3}$$

If we use astronomical units for distance and mass, we obtain

$$\int_0^v \frac{dv}{(1+e\cos v)^2} = \frac{2\pi}{a^{3/2}(e^2-1)^{3/2}}\int_T^t dt. \tag{9.8.4}$$

Here I am using astronomical units of distance and mass and have therefore substituted $4\pi^2$ for $G\mathbf{M}$.

I'm going to write this as

$$\int_0^v \frac{dv}{(1+e\cos v)^2} = \frac{2\pi(t-T)}{a^{3/2}(e^2-1)^{3/2}} = \frac{Q}{(e^2-1)^{3/2}} \tag{9.8.5}$$

where $Q = \frac{2\pi(t-T)}{a^{3/2}}$. Now we have to integrate this.

Method 1

Guided by the elliptical case, but bearing in mind that we are now dealing with a hyperbola, I'm going to try the substitution

$$\cos v = \frac{e-\cosh E}{e\cosh E - 1} \tag{9.8.6}$$

If you try this, I think you'll end up with

$$e\sinh E - E = Q. \tag{9.8.7}$$

This is just the analogy of Kepler's Equation.

The procedure, then, would be to calculate Q from Equation 9.8.5. Then calculate E from Equation 9.8.7. This could be done, for example, by a Newton-Raphson iteration in quite the same way as was done for Kepler's Equation in the elliptic case, the iteration now taking the form

$$E = \frac{Q + e(E\cosh E - \sinh E)}{e\cosh E - 1}. \tag{9.8.8}$$

Then v is found from Equation 9.8.6, and the heliocentric distance is found from the polar Equation to a hyperbola:

$$r = \frac{a(e^2-1)}{1+e\cos v}. \tag{9.8.9}$$

Method 2

Method 1 should work all right, but it has the disadvantage that you may not be as familiar with sinh and cosh as you are with sin and cos, or there may not be a sinh or cosh button your calculator. I believe there are SINH and COSH functions in FORTRAN, and there may well be in other computing languages. Try it and see. But maybe we'd like to try to avoid hyperbolic functions, so let's try the brilliant substitution

$$\cos v = -\frac{u(u-2e)+1}{u(eu-2)+e}. \tag{9.8.10}$$

You may have noticed, when you were learning calculus, that often the professor would make a brilliant substitution, and you could see that it worked, but you could never understand what made the professor think of the substitution. I don't want to tell you what made me think of this substitution, because, when I do, you'll see that it isn't really very brilliant at all. I remembered that

$$\cosh E = \frac{1}{2}(\mathrm{e}^E + \mathrm{e}^E) \tag{9.8.11}$$

and then I let $\mathrm{e}^E = u$, so

$$\cosh E = \frac{1}{2}(u+1/u), \tag{9.8.12}$$

and I just substituted this into Equation 9.8.6 and I got Equation 9.8.10. Now if you put the expression 9.8.10 for $\cos v$ into Equation 9.8.5, you eventually, after a few lines, get something that you can integrate. Please do work through it. In the end, on integration of Equation 9.8.5, you should get

$$\frac{1}{2}e(u-\frac{1}{u}) - \ln u = Q. \tag{9.8.13}$$

You already know from Chapter 1 how to solve the Equation $f(x) = 0$, so there is no difficulty in solving Equation 9.8.13 for u. Newton-Raphson iteration results in

$$u = \frac{2u[e-u(1-Q\ln u)]}{u(eu-2)+2}, \tag{9.8.14}$$

and this should converge in the usual rapid fashion.

So the procedure in method 2 is to calculate Q from Equation 9.8.5, then calculate u from Equation 9.8.14, and finally v from Equation 9.8.10 – all very straightforward.

Exercise 9.7.1

Set yourself a problem to make sure that you can carry through the calculation. Then write a computer program that will generate v and r as a function of t.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.8: Orbital Elements and Velocity Vector

In two dimensions, an orbit can be completely specified by *four orbital elements*. Three of them give the size, shape and orientation of the orbit. They are, respectively, a, e and ω. We are familiar with the semi major axis a and the eccentricity e. The angle ω, the argument of perihelion, was illustrated in figure $\text{II}.19$, which is reproduced here as figure $\text{IX}.7$. It is the angle that the major axis makes with the initial line of the polar coordinates. Figure $\text{II}.19$ reminds us of the relation between the argument of perihelion ω, the argument of latitude θ and the true anomaly v. We remind ourselves here of the Equation to a conic section

$$r = \frac{l}{1+e\cos v} = \frac{l}{1+e\cos(\theta-\omega)}, \tag{9.9.1}$$

where the semi latus rectum l is $a(1-e^2)$ for an ellipse, and $a(e^2-1)$ for a hyperbola. For a hyperbola, the parameter a is usually called the semi transverse axis. For a parabola, the size is generally described by the perihelion distance q, and $l = 2q$.

The fourth element is needed to give information about where the planet is in its orbit at a particular time. Usually this is T, the time of perihelion passage. In the case of a circular orbit this cannot be used. One could instead give the time when $\theta = 0$, or the value of θ at some specified time.

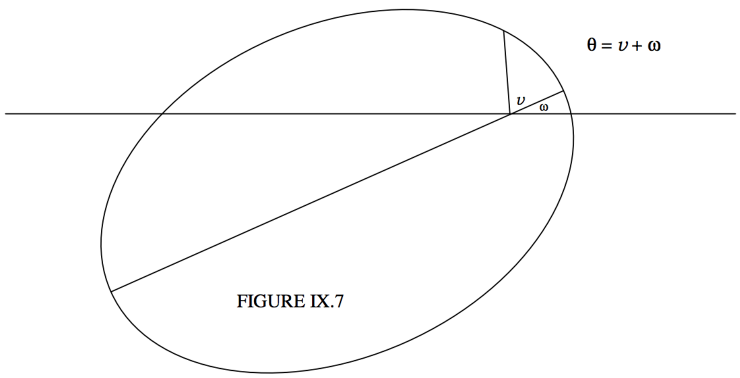


FIGURE IX.7

Refer now to figure IX.8.

We'll suppose that at some time t we know the coordinates $(x,\ y)$ or $(r,\ \theta)$ of the planet, and also the velocity – that is to say the speed and direction, or the x- and y- or the radial and transverse components of the velocity. That is, we know four quantities. The subsequent path of the planet is then determined. In other words, given the four quantities (two components of the position vector and two components of the velocity vector), we should be able to determine the four elements a, e, ω and T. Let us try.

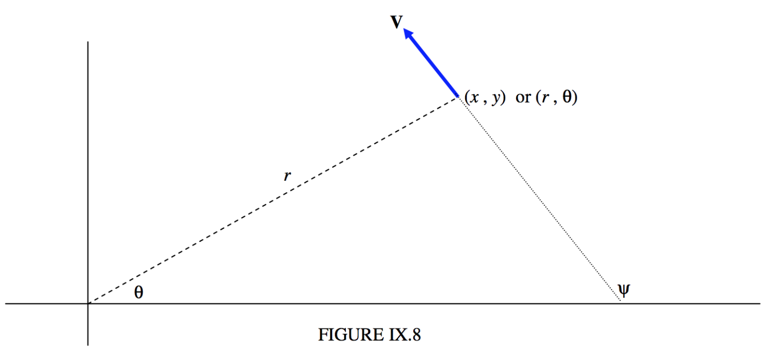


FIGURE IX.8

The semi major axis is easy. It's determined from Equation 9.5.31:

$$V^2 = G\mathbf{M}\left(\frac{2}{r} - \frac{1}{a}\right). \tag{9.5.31}$$

If distances are expressed in AU and if the speed is expressed in units of $29.7846917 \text{ km s}^{-1}$, $G\mathbf{M} = 1$, so that the semi major axis in AU is given by

$$a = \frac{r}{2 - rV^2}. \tag{9.9.2}$$

In other words, if we know the speed and the heliocentric distance, the semi major axis is known. If a turns out to be infinite - in other words, if $V^2 = 2/r$ - the orbit is a parabola; and if a is negative, it is a hyperbola. For an ellipse, of course, the period in sidereal years is given by $P^2 = a^3$.

From the geometry of figure IX.8, the transverse component of $\mathbf{V}$ is $\mathrm{V}\sin(\psi - \theta)$, which is known, the magnitude and direction of $\mathbf{V}$ being presumed known. Therefore the angular momentum per unit mass is r times this, and, for an elliptic orbit, this is related to a and e by Equation 9.5.27a:

$$h = \sqrt{G\mathbf{M}a(1-e^2)}. \tag{9.5.27a}$$

$$rV\sin(\psi - \theta) = \sqrt{G\mathbf{M}a(1-e^2)}. \tag{9.9.3}$$

Again, if distances are expressed in AU and V in units of $29.7846917 \text{ km s}^{-1}$, $G\mathbf{M} = 1$, and so

$$rV\sin(\psi - \theta) = \sqrt{a(1-e^2)}. \tag{9.9.4}$$

Thus e is determined.

The Equation to an ellipse is

$$r = \frac{a(1-e^2)}{1 + e\cos(\theta - \omega)}, \tag{9.9.5}$$

so, provided the usual care is take in choosing the quadrant, ω is now known.

From there we proceed:

$$v = \theta - \omega, \quad \cos E = \frac{e + \cos v}{1 + e\cos v}, \quad \mathcal{M} = E - e\sin E = \frac{2\pi}{P}(t - T), \tag{9.8.1}$$

and T is found. The procedure for a parabola or a hyperbola is similar.

Example 9.8.1

At time $t = 0$, a comet is at $x = +3.0$, $y = +6.0$ AU and it has a velocity with components $\dot{x} = -0.2$, $\dot{y} = +0.4$ times $29.7846917 \text{ km s}^{-1}$. Find the orbital elements a, e, ω and T. (In case you are wondering, a particle of negligible mass moving around the Sun in an unperturbed circular orbit of radius one astronomical unit, moves with a speed of $29.7846917 \text{ km s}^{-1}$. This follows from the definition of the astronomical unit of length.)

Solution

Note in what follows that, although I am quoting numbers to only a few significant figures, the calculation at all times carries all ten figures that my hand calculator allows. You will not get exactly the same results unless you do likewise. Do not prematurely round off. I am using astronomical units of distance, sidereal years for time and speed in units of 29.8 km s^{-1}.

$$r = 6.708, \ \theta = 63^\circ 26', \ V = 0.4472, \ \psi = 116^\circ 34' \tag{9.8.2}$$

Be sure to get the quadrants right!

$$\underline{a = 10.19 \text{ AU}} \quad P = 32.5 \quad \underline{e = 0.6593} \quad \cos(\theta - \omega) = -0.21439 \tag{9.8.3}$$

And now we are faced with a dilemma. $\theta - \omega = 102^\circ\ 23'$ or $257^\circ\ 37'$. Which is it? This is a typical "quadrant problem", and it cannot be ignored. The two possible solutions give $\omega = 321^\circ\ 03'$ or $165^\circ\ 49'$, and we have to decide which is correct.

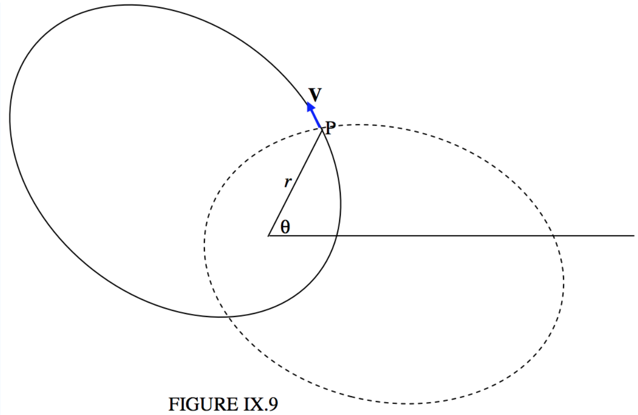


FIGURE IX.9

The two solutions are drawn in Figure IX.9. The continuous curve is the ellipse for $\omega = 321^\circ\ 03'$ and the dashed curve is the curve for $\omega = 165^\circ\ 49'$. I have also drawn in the velocity vector at $(r,\ \theta)$, and it is clear from the drawing that the continuous curve with $\omega = 321^\circ\ 03'$ is the correct ellipse. We now have

$$\underline{\omega = 321^\circ\ 03'} \tag{9.8.4}$$

Is there a way of deducing this from the Equations rather than going to the trouble of drawing the ellipses? I offer the following. I am going to find the slope (gradient) of each ellipse at the point P. The correct ellipse is the one for which $\psi = 116^\circ\ 34'$, i.e. $dy/dx = -2$. The Equation to the ellipse is

$$r = \frac{1}{1 + e\cos(\theta - \omega)} = \frac{a(1-e^2)}{1 + e\cos(\theta - \omega)}, \tag{9.9.6}$$

from which

$$\frac{dr}{d\theta} = \frac{le\sin(\theta - \omega)}{[(1 + e\cos(\theta - \omega)]^2}. \tag{9.9.7}$$

The expression for $\frac{dx}{dy}(= tan\psi)$ in polar coordinates is

$$\frac{dx}{dy} = \frac{\tan\theta\frac{dr}{d\theta} + r}{\frac{dr}{d\theta} - r\tan\theta}, \tag{9.9.8}$$

and of course

$$r = \sqrt{x^2 + y^2}. \tag{9.9.9}$$

From these, I obtain, in our numerical example,

for $\omega = 165^\circ\ 49',\ \psi = 190^\circ$, and for $\omega = 321^\circ\ 03',\ \psi = 165^\circ 49'$,

so clearly the latter is correct.

From this point we go:

$$v = 102^\circ\ 23', \quad \cos E = 0.51817, \tag{9.8.5}$$

and again we are presented with a dilemma, for this gives $E = 58^\circ\ 47'$ or $301^\circ\ 13'$, and we have to decide which one is correct. From the geometrical meaning of v and E, we can understand that they are equal when each of them is either 0° or 180°. Since $v < 180^\circ$, E must also be less than 180°, so the correct choice is $E = 58^\circ\ 47' = 1.0261$ rad. From there, we have

$$\mathcal{M} = 26^\circ\ 29' = 0.46218 \text{ rad}, \quad \underline{T = -2.392\text{sidreal years,}} \tag{9.8.6}$$

and the elements are now completely determined.

Exercise 9.8.1

Write a computer program, in the language of your choice, in which the input data are x, y, $\dot{x}$, $\dot{y}$, and the output is a, e, ω and T. You will probably want to keep it simple at first, and deal only with ellipses. Therefore, if the program calculates that a is not positive, exit the program then. I'm not sure how you will solve the quadrant problems. That will be up to your ingenuity. Don't forget that many languages have an $\mathrm{ARCTAN2}$ function. Later, you will want to expand the program and deal with *any* set of x, y, $\dot{x}$, $\dot{y}$, with a resulting orbit that may be any of the conic sections. Particularly annoying cases may be those in which the planet is heading straight for the Sun, with no transverse component of velocity, so that it is moving in a straight line, or a circular orbit, in which case T is undefined.

Notice that the problem we have dealt with in this section is the opposite of the problem we dealt with in Sections 9.6, 9.7 and 9.8. In the latter, we were given the elements, and we calculated the position of the planet as a function of time. That is, we calculated an ephemeris. In the present section, we are given the position and velocity at some time and are asked to calculate the elements. Both problems are of comparable difficulty. Perhaps the latter is slightly easier than the former, since we don't have to solve Kepler's Equation. This might give the impression that calculating the orbital elements of a planet is of comparable difficulty to, or even slightly easier than, calculating an ephemeris from the elements. This is, in practice, very far from the case, and in fact calculating the elements from the observations is very much more difficult than generating an ephemeris. In this section, we have calculated the elements, given the position and velocity vectors. In real life, when a new planet swims into our ken, we have no idea of the distance or of the speed or the direction of motion. All we have is a set of positions against the starry background, and the most difficult part of the problem of determining the elements is to determine the distance.

The next chapter will deal with generating an ephemeris (right ascension and declination as a function of time) from the orbital elements in the real three-dimensional situation. Calculating the elements from the observations will come much later.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.9: Osculating Elements

We have seen that, if we know the position and velocity vectors at a particular instant of time, we can calculate the orbital elements of a planet with respect to the Sun. If the Sun and one planet (or asteroid or comet) are the only two bodies involved, and if the Sun is spherically symmetric and if we can ignore the refinements of general relativity, the planet will pursue that orbit indefinitely. In practice, however, the orbit is subject to *perturbations*. In the case of most planets moving around the Sun, the perturbations are caused mostly by the gravitational attractions of the other planets. For Mercury, the refinements of general relativity are important. The asphericity of the Sun is unimportant, although for satellites in orbit around aspherical planets, the asphericity of the planet becomes important. In any case, for one reason or another, in practice, an orbit is subject to perturbations, and the planet does not move indefinitely in the orbit that is calculated from the position and velocity vectors at a particular time. The orbit that is calculated from the position and velocity vectors at a particular instant of time is called the *osculating orbit*, and the corresponding orbital elements are the *osculating elements*. The instant of time at which the position and velocity vectors are specified is the *epoch of osculation*. The osculating orbit touches ("kisses") the real, perturbed orbit at the epoch of osculation. The verb "to osculate", from the Latin *osculare*, means "to kiss".

For the time being, then, we shall be satisfied with calculating an *osculating orbit*, and with generating an ephemeris from the *osculating* elements. In computing practice, for asteroid work, people compute elements for an epoch of osculation that is announced by and changed by the Minor Planet Center of the International Astronomical Union every 200 days.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

9.10: Mean Distance in an Elliptic Orbit

It is sometimes said that "a" in an elliptic orbit is the "mean distance" of a planet from the Sun. In fact a is the semi major axis of the orbit. Whether and it what sense it might also be the "mean distance" is worth a moment of thought.

It was the late Professor C. E. M Joad whose familiar answer to the weighty questions of the day was "It all depends what you mean by..." And the "mean distance" depends on whether you mean the distance averaged over the true anomaly v or over the time. The mean distance averaged over the true anomaly is $\frac{1}{\pi}\int_0^\pi rdv$, where $r = l/(1 + ecosv)$. If you are looking for some nice substitution to help you to integrate this, Equation 2.13.6 does very nicely, and you soon find the unexpected result that the mean distance, averaged over the mean anomaly, is b, the semi *minor* axis.

On the other hand, the mean distance averaged over the time is $\frac{1}{2}P\int_0^{\frac{1}{2}P} rdt$. This one is slightly more tricky, but, following the hint for evaluating $\frac{1}{\pi}\int_0^\pi rdv$, you could try expressing r and v in terms of the eccentric anomaly. It will take you a moment or so, but you should eventually find that the mean distance averaged over the time is $a(1 + \frac{1}{2}e^2)$.

It is often pointed out that, because of Kepler's second law, a planet spends more time far from the Sun that it does near to the Sun, which is why we have longer summers than winters in the northern hemisphere. An easy exercise would be to ask you what fraction of its orbital period does a planet spend on the sunny side of a latus rectum. A slightly more difficult exercise would be to ask: What fraction of its orbital period does a planet spend closer to the Sun than its mean (time-averaged) distance? You'd first have to ask, what is the true anomaly when $r = a(1 + \frac{1}{2}e^2)$? Then you need to calculate the fraction of the area of the orbit. Area in polar coordinates is $\frac{1}{2}\int r^2 dv$. I haven't tried this, but, if it proves difficult, I'd try and write r and v in terms of the eccentric anomaly E and see if that helped.

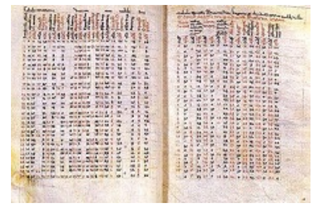
Contributor

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

10: COMPUTATION OF AN EPHEMERIS

10.1: INTRODUCTION TO AN EPHEMERIS



The entire enterprise of determining the orbits of planets, asteroids and comets is quite a large one, involving several stages. From the available observations, the orbit of the body has to be determined; in particular we have to determine the orbital elements, a set of parameters that describe the orbit. For a new body, one determines preliminary elements from the initial few observations that have been obtained. As more observations are accumulated, so will the calculated preliminary elements

10.2: ELEMENTS OF AN ELLIPTIC ORBIT

10.3: SOME ADDITIONAL ANGLES

10.4: ELEMENTS OF A CIRCULAR OR NEAR-CIRCULAR ORBIT

10.5: ELEMENTS OF A PARABOLIC ORBIT

The eccentricity, of course, is unity, so only five elements are necessary. In place of the semi major axis, one usually specifies the parabola by the perihelion distance q . Presumably no orbit is ever exactly parabolic, which implies an eccentricity of exactly one. However, many long-distance comets move in large and eccentric orbits, and we see them over such a short arc near to perihelion that it is not possible to calculate accurate elliptic orbits, and we usually then fit a parabolic orbi

10.6: ELEMENTS OF A HYPERBOLIC ORBIT

10.7: CALCULATING THE POSITION OF A COMET OR ASTEROID

We suppose that we are given the orbital elements of an asteroid or comet. Our task is to be able to predict, from these, the right ascension and declination of the object in the sky at some specified future (or past) date. If we can do it for one date, we can do it for many dates - e.g. every day for a year if need be. In other words, we will have constructed an ephemeris.

10.8: QUADRANT PROBLEMS

10.9: COMPUTING AN EPHEMERIS

10.10: ORBITAL ELEMENTS AND VELOCITY VECTOR

10.11: HAMILTONIAN FORMULATION OF THE EQUATIONS OF MOTION

10.1: Introduction to an Ephemeris

The entire enterprise of determining the orbits of planets, asteroids and comets is quite a large one, involving several stages. New asteroids and comets have to be searched for and discovered. Known bodies have to be found, which may be relatively easy if they have been frequently observed, or rather more difficult if they have not been observed for several years. Once located, images have to be obtained, and these have to be measured and the measurements converted to usable data, namely right ascension and declination. From the available observations, the orbit of the body has to be determined; in particular we have to determine the *orbital elements*, a set of parameters that describe the orbit. For a new body, one determines *preliminary elements* from the initial few observations that have been obtained. As more observations are accumulated, so will the calculated preliminary elements. After all observations (at least for a single opposition) have been obtained and no further observations are expected at that opposition, a *definitive orbit* can be computed. Whether one uses the preliminary orbit or the definitive orbit, one then has to compute an *ephemeris* (plural: *ephemerides*); that is to say a day-to-day prediction of its position (right ascension and declination) in the sky. Calculating an ephemeris from the orbital elements is the subject of this chapter. Determining the orbital elements from the observations is a rather more difficult calculation, and will be the subject of a later chapter.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.2: Elements of an Elliptic Orbit

Six numbers are necessary and sufficient to describe an elliptic orbit in three dimensions. These include the four (a, e, ω and T) that we described in section 9.9 for the two dimensional case. Two additional angles, which will be given the symbols i and Ω, will be needed to complete the description of the orbit in 3-space.

The six elements of an elliptic orbit, then, are as follows.

a the semi major axis, usually expressed in *astronomical units* (AU).
e the eccentricity
i the inclination
Ω the longitude of the ascending node
ω the argument of perihelion
T the time of perihelion passage

The three angles, i, Ω and ω must always be referred to the equinox and equator of a stated epoch. For example, at present they are usually referred to the mean equinox and equator of $\mathrm{J2000.0}$ The meanings of the three angles are explained in figure $\mathrm{X.1}$ and the following paragraphs.

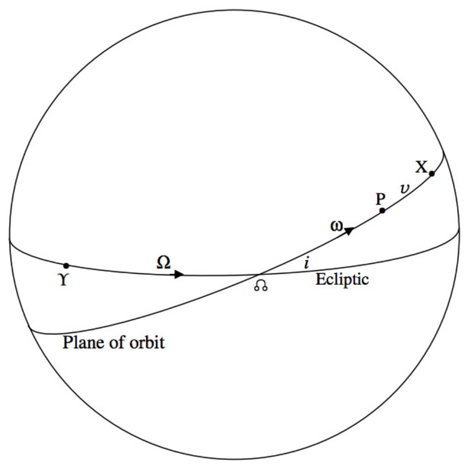


FIGURE X.1

In figure $\mathrm{X.1}$ I have drawn a celestial sphere centred on the Sun. The two great circles are intended to represent the plane of Earth's orbit (i.e. the ecliptic) and the plane of a planet's orbit – (i.e. not the orbit itself, but its projection on to the celestial sphere.) The point P is the projection of the perihelion point of the orbit on to the celestial sphere, and the point X is the projection of the planet on to the celestial sphere at some time. The two points where the plane of the ecliptic and the plane of the planet's orbit intersect are called the *nodes*, and the point marked is the ascending node. The descending node, , not shown in the figure, is on the far side of the sphere. The symbol is the First point of Aries (now in the constellation Pisces), where the ecliptic crosses the equator. As seen from the Sun, Earth is at on Υ or near September 22. (For the benefit of personal computer users, I found the symbols , and Υ in Bookshelf Symbol 3.) —⊸

The inclination i is the angle between the plane of the object's orbit and the plane of the ecliptic (i.e. of Earth's orbit). It lies in the range o o 0 ≤ i < 180 . An inclination greater than 90o implies that the orbit is retrograde – i.e. that it is moving around the Sun in a direction opposite to that of Earth's motion.

The angle Ω, measured eastward from Υ to $mulitmapiny$ is the *ecliptic longitude of the ascending node*. (The word "ecliptic" is usually omitted as understood.) It goes from 0° to 360°.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.3: Some Additional Angles

The sum of the two angles Ω and ω is often given the symbol ϖ (a form of the Greek letter pi), and is called (not entirely accurately) the *longitude of perihelion*. It is the sum of two angles measured in different planes.

The angle v, measured from perihelion to the planet, is the *true anomaly* of the planet at some time. We imagine, in addition to the true planet, a "mean" planet, which moves at constant angular speed $2\pi/P$, so that the angle from perihelion to the mean planet at time t is $M=\frac{2\pi(t-T)}{P}$, which is called the *mean anomaly* at time t. The words "true" and "mean" preceding the word "anomaly" refer to the "true" planet and the "mean" planet.

The angle $\theta=\omega+v$, measured from FIND SYMBOL, is the *argument of latitude* of the planet at time t.

The angle $l=\Omega+\theta=\Omega+\omega+v=\varpi+v$ measured in two planes, is the *true longitude* of the planet. This is a rather curious term, since, being measured in two planes, it is not really the true longitude at all. The word "true" refers to the "true" planet rather than to the longitude.

Likewise the angle $L=\Omega+\omega+M=\varpi+M$ is the *mean longitude* (i.e. the "longitude" of the "mean" planet.).

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.4: Elements of a Circular or Near-circular Orbit

For a near-circular orbit (such as the orbits of most of the major planets), the position of perihelion and the time of perihelion passage are ill-defined, and for a perfectly circular orbit they cannot be defined at all. For a near-circular orbit, the argument of perihelion ω (or sometimes the "longitude of perihelion", ϖ) is retained as an element, because there is really no other way of expressing the position of perihelion, though of course the more circular the orbit the less the precision to which ω can be determined. However, rather than specify the time of perihelion passage T, we usually specify some instant of time called the *epoch*, which I denote by t_0, and then we specify either the mean anomaly at the epoch, M_0, or the mean longitude at the epoch, L_0, or the true longitude at the epoch, l_0. For the meanings of mean anomaly, mean longitude and true longitude, refer to section 3, especially for the meanings of "mean" and "true" in this context. Of the three, only l_0 makes no reference whatever to perihelion.

Note that you should not confuse the epoch for which you specify the mean anomaly or mean longitude or true longitude with the equinox and equator to which the angular elements i, Ω and ω are referred. These may be the same, but they need not be (and usually are not). Thus it is often convenient to refer i, Ω and ω to the standard epoch $\text{J}2000.0$, but to give the mean longitude for an epoch during the current year.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.5: Elements of a Parabolic Orbit

The eccentricity, of course, is unity, so only five elements are necessary. In place of the semi major axis, one usually specifies the parabola by the perihelion distance q. Presumably no orbit is ever exactly parabolic, which implies an eccentricity of exactly one. However, many long-distance comets move in large and eccentric orbits, and we see them over such a short arc near to perihelion that it is not possible to calculate accurate elliptic orbits, and we usually then fit a parabolic orbit to the observations.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.6: Elements of a Hyperbolic Orbit

In place of the semi major axis, we have the semi transverse axis, symbol a. This amounts to just a name change, although some authors treat a for a hyperbola as a negative number, because some of the formulas, for example for the speed in an orbit, $V^2 = GM\left(\frac{2}{r} - \frac{1}{a}\right)$, are then identical for an ellipse and for a hyperbola.

Although there is no fundamental reason why the solar system should not sometime receive a cometary visitor from interstellar space, as yet we know of no comet with an original hyperbolic orbit around the Sun. Some comets, initially in elliptic orbits, are perturbed into hyperbolic orbits by a close passage past Jupiter, and are then lost from the solar system. Such orbits are necessarily highly perturbed and one cannot in general compute a reliable ephemeris by treating it as a simple two-body problem; the instantaneous osculating elements will not predict a reliable ephemeris far in advance.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.7: Calculating the Position of a Comet or Asteroid

We suppose that we are given the orbital elements of an asteroid or comet. Our task is to be able to predict, from these, the right ascension and declination of the object in the sky at some specified future (or past) date. If we can do it for one date, we can do it for many dates - e.g. every day for a year if need be. In other words, we will have constructed an ephemeris. Nowadays, of course, we can obtain ephemeris-generating programs and ephemerides with a few deft clicks on the Web, without knowing so much as the difference between a sine and a cosine; but that way of doing it is not the purpose of this section.

For example, according to the Minor Planet Center, the osculating elements for the minor planet (1) Ceres for the epoch of osculation $t_0 = 2002$ May 6.0 TT are as follows:

$a = 2.766\ 412\ 2$ AU $\quad \Omega = 80^\circ.486\ 32$
$e = 0.079\ 115\ 8 \quad \omega = 73^\circ.984\ 40$
$i = 10^\circ.583\ 47 \quad M_0 = 189^\circ.275\ 00$

i, Ω and ω are referred to the equinox and equator of $\mathrm{J}2000.0$

Calculate the right ascension and declination (referred to $\mathrm{J}2000.0$) at 2002 July 15.0 TT.

We have already learned how to achieve much of our aim from Chapter 9. Thus, from the elements a, e, ω and T for an elliptic orbit (or the corresponding elements for a parabolic or hyperbolic orbit) we can already compute the *true anomaly* v and the heliocentric distance r as a function of time. These are the heliocentric polar coordinates of the body (henceforth "asteroid"). In order to find the right ascension and declination (i.e. geocentric coordinates with the celestial equator as xy-plane) all we have to do is to find the coordinates relative to the ecliptic, rotate the coordinate system from ecliptic to equatorial, and shift the origin of coordinates from Sun to Earth,. We just have to do some straightforward geometry, and no further dynamics.

Let's start by doing what we already know how to do from Chapter 9, namely, we'll calculate the true anomaly and the heliocentric distance.

- Mean anomaly at the epoch ($t_0 = \text{May } 6.0$) is $M_0 = 189^\circ.275\ 00$.
- Mean anomaly at time t ($= \text{July } 15$. which is 70 days later) is given by

$$M - M_0 = \frac{2\pi}{P}(t - t_0). \tag{10.7.1}$$

The quantity $2\pi/P$ is called the mean motion (actually the average orbital angular speed of the planet), usually given the symbol n. We can calculate P in sidereal years from $P^2 = a^3$, and, given that a sidereal year is $365^{\mathrm{d}}.25636$ and that 2π radians is 360 degrees, we can calculate the mean motion in its usual units of degrees per day. We find that $n = 0.214\ 205$ degrees per day. In fact the Minor Planet Center, as well as giving the orbital elements, also lists, for our convenience, the mean motion, and they give $n = 0.214\ 204\ 57$ degrees per day. The small discrepancy between the n given by the Minor Planet Center and the value that we have calculated from the published value of a presumably arises because the published values of the elements have been rounded off for publication, and the Minor Planet Center presumably carries all digits in its calculations. I would recommend using the value of n published by the Minor Planet Center, and I do so here. By July 15, then, Equation 10.7.1 tells us that the mean anomaly is $M = 204^\circ.269342$. (I'm carrying six decimal places, even though M_0 is given only to five, just to be sure that I'm not accumulating rounding-off errors in the intermediate calculations. I'll round off properly when I reach the final result.)

We now have to find the eccentric anomaly from Kepler's Equation $M = E - e\sin E$. Easy. (See chapter 9 if you've forgotten how.) We find $E = 202^\circ.5322784$ and, from Equations 2.3.16 and 17, we obtain the true anomaly $v = 200^\circ.8540289$. The polar Equation to an ellipse is , $r = \frac{a(1-e^2)}{1+e\cos v}$. so we find that the heliocentric distance is $r = 2.968\ 5716$ au (The Minor Planet Centre gives r, to four significant figures, as 2.969 au) So much we could already do from Chapter 9. Note also that $\omega + v$, known as the argument of latitude and often given the symbol θ, is $274^\circ.838\ 429$.

We are going to have to make use of three heliocentric coordinate systems and one geocentric coordinate system.

1. *Heliocentric plane-of-orbit*. $\odot xyz$ with the ?x axis directed towards perihelion. The polar coordinates in the plane of the orbit are the heliocentric distance r and the true anomaly v. The z-component of the asteroid is necessarily zero, and $x = r\cos v$ and $y = r\sin v$.

2. Heliocentric ecliptic. $\odot XYZ$ with the $\odot X$ axis directed towards the *First Point of Aries*, where Earth, as seen from the Sun, will be situated on or near September 22. The spherical coordinates in this system are the heliocentric distance r, the ecliptic longitude λ, and the ecliptic latitude β, such that $X = r\cos\beta\cos\lambda$, $Y = r\cos\beta\sin\lambda$ and $Z = r\sin\beta$.

INSERT FIGURE HERE
FIGURE X.2

3. *Heliocentric equatorial coordinates*. $\odot\xi\eta\zeta$ with the $\odot\xi$ axis directed towards the First Point of Aries and therefore coincident with the X axis . The angle between the Z axis and the ζ axis is ε, the obliquity of the ecliptic. This is also the angle between the XY-plane (plane of the ecliptic, or of Earth's orbit) and the ξη-plane (plane of Earth's equator). See figure X.4.

4. *Geocentric equatorial coordinates*. ⊕xyz with the ⊕x axis directed towards the First Point of Aries. The spherical coordinates in this system are the geocentric distance Δ, the right ascension α and the declination δ, such that $x = \Delta\cos\delta\cos\alpha$, $y = \Delta\cos\delta\sin\alpha$ and $z = \Delta\sin\delta$.

In figure X.2, the arc ϒN is the heliocentric ecliptic longitude λ of the asteroid, and so NN is $\lambda - \Omega$. The arc NX is the heliocentric ecliptic latitude β. By two applications of Equation 3.5.5 we find

$$\cos(\lambda-\Omega)\cos i = \sin(\lambda-\Omega)\cot(\omega+v) - \sin i\cot 90^\circ \tag{10.7.2}$$

and

$$\cos(\lambda-\Omega)\cos 90^\circ = \sin(\lambda-\Omega)\cot\beta - \sin 90^\circ\cot i. \tag{10.7.3}$$

These reduce to

$$\tan(\lambda-\Omega) = \cos i\tan(\omega+v) \tag{10.7.4}$$

and

$$\tan\beta = \sin(\lambda-\Omega)\tan i. \tag{10.7.5}$$

In our particular example, we obtain (if we are careful to watch the quadrants),

$$\lambda-\Omega = 274^\circ.921\,7550, \quad \lambda = 355^\circ.408\,0750, \quad \beta = -10^\circ.545\,3234$$

Now, we'll take the X-axis for the heliocentric ecliptic coordinates through ϒ and the Y-axis 90° east of this. Then, by the usual formulas for converting between spherical and rectangular coordinates, that is, $X = r\cos\beta\cos\lambda$, $Y = r\cos\beta\sin\lambda$ and $Z = r\sin\beta$, we obtain

$$X = +2.909\,0661, \quad Y = -0.233\,6453, \quad Z = -0.543\,2880 \quad \text{au.}$$

(Check: $X^2+Y^2+Z^2 = r^2$.)

Exercise 10.7.1

Show, by elimination of λ and β, or otherwise, that:

$$X = r(\cos\Omega\cos\theta - \sin\Omega\sin\theta\cos i) \tag{10.7.6}$$

$$Y = r(\sin\Omega\cos\theta + \cos\Omega\sin\theta\cos i) \tag{10.7.8}$$

$$Z = r\sin\theta\sin i. \tag{10.7.9}$$

This will provide a more convenient way of calculating the coordinates. Verify that these give the same numerical result as before. Here are some suggestions for doing it "otherwise"

Refer to Figure $X.3$, in which K is the pole of the ecliptic, and X is the asteroid. The radius of the celestial sphere can be taken as equal to r, the heliocentric distance of the asteroid. The rectangular heliocentric ecliptic coordinates are

$X = r\cos \Upsilon \odot \mathrm{X} \quad Y = r\cos \mathrm{R} \odot \mathrm{X} \quad Z = r\cos \mathrm{K} \odot \mathrm{X}$

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.8: Quadrant Problems

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.9: Computing an Ephemeris

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.10: Orbital Elements and Velocity Vector

Contributor

- Jeremy Tatum (University of Victoria, Canada)

10.11: Hamiltonian Formulation of the Equations of Motion

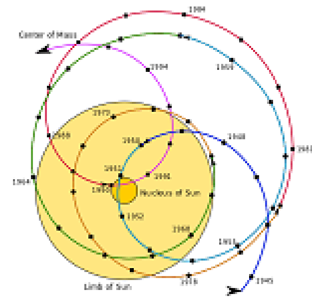
Contributor

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

11: PHOTOGRAPHIC ASTROMETRY

Astrometry is the art and science of measuring positions of celestial objects, and indeed the first step in determining the orbit of a new asteroid or comet is to obtain a set of good astrometric positions. For much of the twentieth century, most astrometric positions were determined photographically.



11.1: Introduction to Photographic Astrometry

Astrometry is the art and science of measuring positions of celestial objects, and indeed the first step in determining the orbit of a new asteroid or comet is to obtain a set of good astrometric positions. For much of the twentieth century, most astrometric positions were determined photographically, although transit circle measurements were (and still are in some applications) important. A photographic plate or film would be baked for several hours in an oven in an atmosphere of dry hydrogen and nitrogen. This “hypersensitization” was known to increase the sensitivity of the emulsion in long exposures. The film would then be exposed through a telescope to an area of the sky containing the asteroid. An hour or so later, a second photograph would be exposed, the asteroid presumably having moved slightly between the exposures. Exposure times would be from several minutes to an hour or even more, and the telescope had to be carefully guided throughout the long exposure. After exposure, the film had to be developed in a chemical solution in a dark-room, then “fixed” in another solution, washed under running water, and hung up to dry. After these procedures, which took some hours, preparation for measurement could start. The first thing to do would be to identify the asteroid. (In Mrs Beecham’s words, “First catch your hare”.) To do this, the two photographs would be viewed rapidly one after the other with a blink comparator (in which case the asteroid would move to and fro) or viewed simultaneously with a stereocomparator (in which case the asteroid would appear to be suspended in air above the film). Next, a number of comparison stars would have to be identified. This would be done by consulting a star catalogue and laboriously plotting the positions of the stars on a sheet of paper and comparing the pattern with what was seen on the photographs.

Each photograph would then be placed in a “measuring engine”, or two-coordinate measuring microscope, and the x- and y-coordinates of the stars and the asteroid would be measured. Tedious calculations would be performed to convert the measurements to right ascension and declination. The results of this process, which would typically take several hours, would then be sent by mail to the Minor Planet Center of the International Astronomical Union in Cambridge, Massachusetts.

Starting in the early 1990s, photographic astrometry started to be superseded by CCD (charge coupled device) astrometry, and today almost no astrometry is done photographically, the CCD having taken over more or less completely. Everyone knows that the quantum efficiency of a CCD is far superior to that of a photographic emulsion, so that one can now image much fainter asteroids and with much shorter exposures. But that is only the beginning of the story – the CCD and other modern technologies have completely changed the way in which astrometry is carried out. For example, vast catalogues containing the positions of hundreds of millions of faint stars are stored in computer files, and the computer can automatically compare the positions of the stars in its catalogue with the star images on the CCD; thus the hitherto laborious process of identifying the comparison stars is carried out automatically and almost instantaneously. Further, there is no measurement to be done – each stellar image is already sitting on a particular pixel (or group of pixels), and all that has to be done is to read which pixels contain the stellar images. The positional measurements are all inherently completed as soon as the CCD is exposed. The positional measurements (of dozens of stars rather than a mere half-dozen) can then be automatically transferred into a computer program that carries out the necessary trigonometrical calculations to convert them to right ascension and declination, and the results can then be automatically sent by electronic mail to the Minor Planet Center. The entire process, which formerly took many hours, can now be done in less than a minute, to much higher precision than formerly, and for much fainter objects.

Why, then, would you ever want to read a chapter on photographic astrometry? Well, perhaps you won’t. After all, to convert your observations to right ascension and declination today, a single key on your computer keyboard will do it all. But this is because someone, somewhere, and usually a very anonymous person, has written for you a highly efficient computer program that carries out all the necessary calculations, so that you can do useful astrometry even if you don’t know the difference between a sine and a cosine. Thus you can probably safely bypass this chapter.

However, for those who wish to plod through it, this chapter describes how to convert the positional measurements on a photographic film (or on a CCD) to right ascension and declination – a process that is carried out by modern computer software, even if you are unaware of it. Much of this chapter is based on an article by the author published in the *Journal of the Royal Astronomical Society of Canada* $\mathbf{76}$, 97 (1982), and you may want to consult that in the hope that I might have made it clear in either one place or the other.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.2: Standard Coordinates and Plate Constants

We shall suppose that the optic axis of the telescope, whose effective focal length is F, is pointing to a point C on the celestial sphere, whose right ascension and declination are $(A,\ D)$. The stars, as every astrophysicist knows, are scattered around on the surface of the celestial sphere, which is of arbitrary radius, and I shall take the radius to be equal to F, the focal length of the telescope. In figure $\mathrm{XI.1}$, I have drawn the tangent plane to the sky at C, which is what will be recorded on the photograph. In the tangent plane (which is similar to the plane of the photographic plate or film) I have drawn two orthogonal axes: $\mathrm{C}\xi$ to the east and $\mathrm{C}\eta$ to the north. I have drawn a star, Q, whose coordinates are (α, δ), on the surface of the celestial sphere, and its projection, Q', on the tangent plane, where its coordinates are (ξ, η). Every star is similarly mapped on to the tangent plane by a similar projection. The coordinates (ξ, η) are called the *standard coordinates* of the star, and our first task is to find a relation between the equatorial coordinates (α, δ) on the surface of the celestial sphere and the standard coordinates (ξ, η) on the tangent plane or the photograph.

In figure $\mathrm{XI.2}$, I have re-drawn figure $\mathrm{XI.1}$, and, in addition to the star Q and its projection Q', I have also drawn the north Celestial Pole P and its projection P'. The point P' is on the η axis. The spherical triangle PQC maps onto the plane triangle $\mathrm{P'Q'C}$. On the spherical triangle PQC, the side $\mathrm{PQ} = 90^\circ - \delta$ and the side $\mathrm{PC} = 90^\circ - D$.

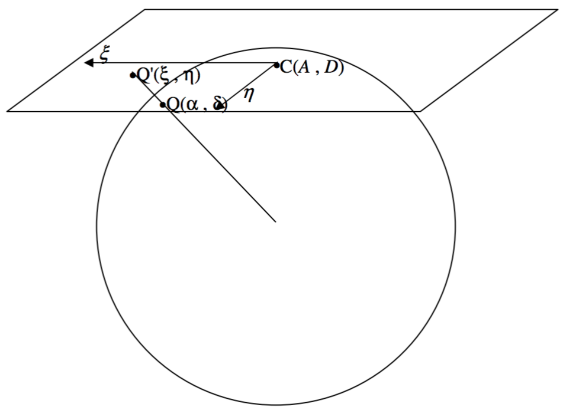


FIGURE XI.1

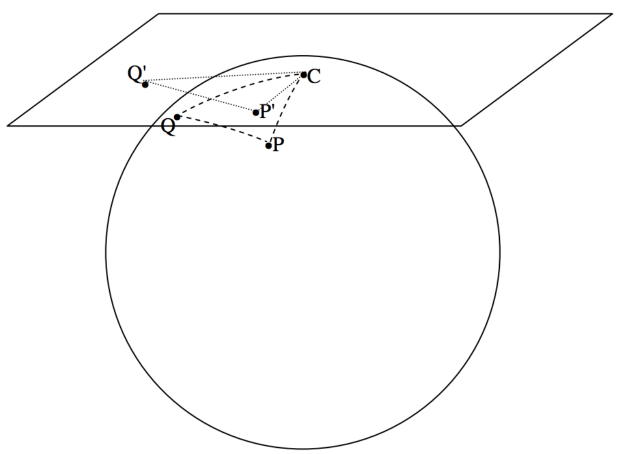


FIGURE XI.1

The angle PCQ in the spherical triangle PCQ is equal to the angle $\mathrm{P'CQ'}$ in the plane triangle $\mathrm{P'CQ'}$, and I shall call that angle γ. I shall call the arc CQ in the spherical triangle ε. In figure $\mathrm{XI.3}$ I draw the tangent plane, showing the $\xi-$ and η-axes and the projections, P' and Q' of the pole P and the star Q, as well as the plane triangle $\mathrm{P'CQ'}$.

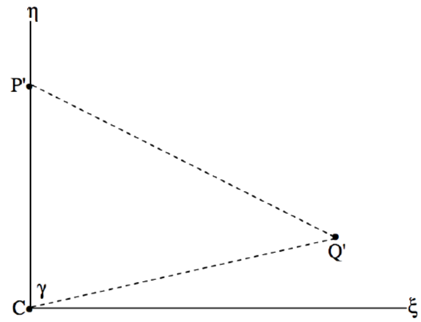


FIGURE XI.3

The ξ and η coordinates of Q' are $(\mathrm{CQ}'\sin\gamma,\ \mathrm{CQ}'\cos\gamma)$. And by staring at figures XI.1 and XI.2 for a while, you can see that $\mathrm{CQ}' = F\tan\varepsilon$. Thus the standard coordinate of the image Q' of the star on the photograph, in units of the focal length of the telescope, are $(\tan\varepsilon\sin\gamma,\ \tan\varepsilon\cos\gamma)$. It remains now to find expressions for $\tan\varepsilon\sin\gamma$ and $\tan\varepsilon\cos\gamma$ in terms of the right ascensions and declinations of Q and of C. I draw now, in figure XI.4, the spherical triangle PCQ.

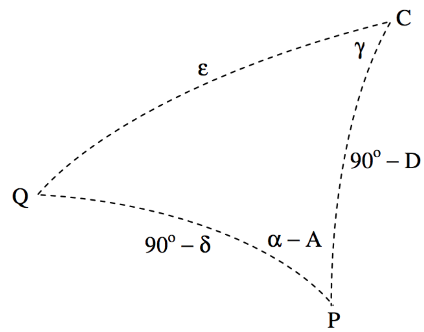


FIGURE XI.4

It is easy, from the usual formulas for spherical triangles, to obtain expressions for $\cos\varepsilon$ and for $\tan\gamma$:

$$\cos\varepsilon = \sin\delta\sin D + \cos\delta\cos D\cos(\alpha - A) \tag{11.2.1}$$

and

$$\tan\gamma = \frac{\sin(\alpha - A)}{\cos D\tan\delta - \sin D\cos(\alpha - A)}, \tag{11.2.2}$$

from which one can (eventually) calculate the standard coordinates (ξ, η) of the star. It is also possible to calculate explicit expressions for $\tan\varepsilon\sin\gamma$ and for $\tan\varepsilon\cos\gamma$. Thus, by further applications of the spherical triangle formulas, we have

$$\tan\varepsilon = \frac{\cos D}{\sin D\cos\gamma + \cot(\alpha - A)\sin\gamma}. \tag{11.2.3}$$

Multiplication of Equation 11.2.3 by $\sin\gamma$ gives $\tan\varepsilon\sin\gamma$ except that $\tan\gamma$ appears on the right hand side. This, however, can be eliminated by use of Equation 11.2.2, and one obtains, after some algebra:

$$\xi = \tan\varepsilon\sin\gamma = \frac{\sin(\alpha - A)}{\sin D\tan\delta + \cos D\cos(\alpha - A)}. \tag{11.2.4}$$

In a similar way, you can multiply Equation 11.2.3 by $\cos\gamma$, and again eliminate $\tan\gamma$ and eventually arrive at

$$\eta = \tan\varepsilon\cos\gamma = \frac{\tan\delta - \tan D\cos(\alpha - A)}{\tan D\tan\delta + \cos(\alpha - A)}. \tag{11.2.5}$$

These give the standard coordinates of a star or asteroid at (α, δ) in units of the focal length F.

Now it would seem that all we have to do is to measure the standard coordinates (ξ, η) of an object, and we can immediately determine its right ascension and declination by inverting Equations 11.2.4 and 11.2.5:

$$\tan(\alpha - A) = \frac{\xi}{\cos D - \eta \sin D} \tag{11.2.6}$$

and

$$\tan\delta = \frac{(\eta\cos D + \sin D)\sin(\alpha - A)}{\xi}. \tag{11.2.7}$$

Indeed in principle that is what we have to do – but in practice we are still some way from achieving our aim.

One small difficulty is that we do not know the effective focal length F (which depends on the temperature) precisely. A more serious problem is that we do not know the exact position of the plate centre, nor do we know that the directions of travel of our two-coordinate measuring engine are parallel to the directions of right ascension and declination.

The best we can do is to start our measurements from some point near the plate centre and measure (in mm rather than in units of F) the horizontal and vertical distances (x, y) of the comparison stars and the asteroid from our arbitrary origin. These (x, y) coordinates are called, naturally, the *measured coordinates*.

The measured coordinates will usually be expressed in millimetres (or perhaps in pixels if a CCD is being used), and the linear distance s between any two comparison star images is found by the theorem of Pythagoras. The angular distance ω between any two stars is given by solution of a spherical triangle as

$$\cos\omega = \sin\delta_1 \sin\delta_2 + \cos\delta_1 \cos\delta_2 \cos(\alpha_1 - \alpha_2) \tag{11.2.8}$$

The focal length F is then s/ω, and this can be calculated for several pairs of stars and averaged. From that point the standard coordinates can then be expressed in units of F.

The *measured coordinates* (x, y) are displaced from the *standard coordinates* (ξ, η) by an unknown translation and an unknown rotation (figure XI.4) , but the relation between them, if unknown, is at least linear (but see subsection $11.3.5$) and thus of the form:

$$\xi - x = ax + by + c, \tag{11.2.9}$$

$$\eta - y = dx + ey + f. \tag{11.2.10}$$

The constants a– f are the *plate constants*. They are determined by measuring the standard coordinates for a minimum of three comparison stars whose right ascensions and declinations are known and for which the standard coordinates can therefore be calculated. Three sets of Equations 11.2.9 and 10 can then be set up and solved for the plate constants. In practice more than three comparison stars should be chosen, and a least squares solution determined. For how to do this, see either section 8 of chapter 1, or the article cited in section 1 of this chapter. In the photographic days, just a few (perhaps half a dozen) comparison stars were used. Today, when there are catalogues containing hundreds of millions of stars, and CCD measurement and automatic computation are so much faster, several dozen comparison stars may be used, and any poor measurements (or poor catalogue positions) can quickly be identified and rejected.

Having determined the plate constants, Equations 11.2.9 and 10 can be used to calculate the standard coordinates of the asteroid, and hence its right ascension and declination can be calculated from Equations 11.2.6 and 7.

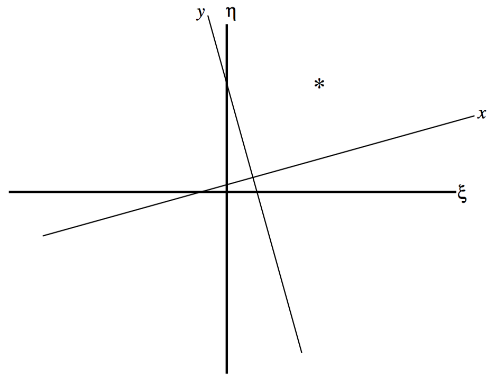


FIGURE 11.5

It should be noted that the position of the asteroid that you have measured – and should report to the Minor Planet Center, is the *topocentric* position (i.e. as measured from your position on the surface of Earth) rather than the *geocentric* position (as seen from the centre of Earth). The Minor Planet Center expects to receive from the observer the topocentric position; the MPC will know how to make the correction to the centre of Earth.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3: Refinements and Corrections

Topic hierarchy

For precise work there are a number of refinements that should be considered, some of which should be implemented, and some which probably need not be. Things that come to mind include parallax and proper motion of the comparison stars, refraction, aberration, optical distortion, mistakes – which include such things as poor measurements, blends, poor or erroneous catalogue positions or any of a number of mistakes caused by human or instrumental frailty. If you write your own reduction programs, you will know which of these refinements you have included and which you have left out. If you use a "pre-packaged" program, you may not always know whether a given correction has been included.

Let us now look at some of these refinements.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.1: Parallaxes of the Comparison Stars

Unless you are unlucky enough to choose as one of your comparison stars Proxima Centauri (whose parallax is much less than an arcsecond), the parallaxes of the comparison stars are not normally something that the asteroid astrometrist has to worry about.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.2: Proper Motions of the Comparison Stars

Corrections for the proper motions of the comparison stars should certainly be made if possible.

Until a quarter of a century or so ago, a typical stellar catalogue used by asteroid observers was the *Smithsonian Astrophysical Stellar Catalog* containing the positions *and proper motions* of about a quarter of a million stars down to about magnitude 9. This catalogue gives the position (right ascension and declination referred to the equinox and equator of $\mathrm{B1950.0}$) of each star at the time of the original epoch when the photograph on which the catalogue was based, and also the position of each star corrected for proper motion to the epoch 1950.0, as well as the proper motion of each star. Thus for the star $\mathrm{SAO013800}$ the position (referred to the equinox and equator of $\mathrm{B1950.0}$) *at the original epoch* is given as

$$\alpha_{1950.0} = 08^{\mathrm{h}}\ 14^{\mathrm{m}}\ 40^{\mathrm{s}}.390 \quad \delta_{1950.0} = +65^{\circ}\ 09'\ 18''.87 \tag{11.3.2.1}$$

and the proper motion is given as

$$\mu_{\alpha} = -0^{\mathrm{s}}.0058 \quad \mu_{\delta} = -0''.085 \text{ per year} \tag{11.3.2.2}$$

The epoch of the original source is not immediately readable from the catalogue, but can be deduced from information therein. In any case the catalogue gives the position (referred to the equinox and equator of $\mathrm{B1950.0}$) corrected for proper motion to the epoch 1950.0:

$$\alpha_{1950.0} = 08^{\mathrm{h}}\ 14^{\mathrm{m}}\ 40^{\mathrm{s}}.274 \quad \delta_{1950.0} = +65^{\circ}\ 09'\ 17''.16 \tag{11.3.2.3}$$

Now, suppose that you had taken a photograph in 1980. At that time we were still referring positions to the equinox and equator of $\mathrm{B1950.0}$ (today we use $\mathrm{J2000.0}$), but you would have to correct the position for proper motion to 1980; that is, you need to apply the proper motion for the 30 years since 1950. The position, then, in 1980, referred to the equinox and equator of $\mathrm{B1950.0}$) was

$$\alpha_{1950.0} = 08^{\mathrm{h}}\ 14^{\mathrm{m}}\ 40^{\mathrm{s}}.100 \quad \delta_{1950.0} = +65^{\circ}\ 09'\ 09\ 14''.61 \tag{11.3.2.4}$$

and this is the position of the star that should be used in determining the plate constants.

One problem with this was that the proper motions were not equally reliable for all the stars (although the catalogue does list the formal standard errors in the proper motions), and there are a few stars in which the proper motion is even given with the wrong sign! In such cases, correcting for proper motion obviously does more harm than good. However, the stars with the “worst” proper motions are generally also those with the smallest proper motions; it can probably be assumed that the stars with significant proper motions also have proper motions that are well determined.

The situation changed in the 1990s with the widespread introduction of CCDs and the publication of the *Guide Star Catalog* containing positions of about half a billion stars down to about magnitude 21. With modern instrumentation one would never normally consider using comparison stars anything like as bright as magnitude 9 (the faint limit of the SAO Catalog). You now have the opportunity of choosing many more comparison stars, and faint ones, whose positions can be much more precisely measured than bright stars. Also, the *Guide Star Catalog* gives positions referred to the equinox and equator of $\mathrm{J2000.0}$, which is the present-day norm for reporting astrometric positions. A difficulty is, however, that the GSC positions were obtained at only one epoch, so that proper motions are not available for the GSC stars, and hence proper motions cannot be applied. The standard response to this drawback is that, since faint stars (magnitude 16 and fainter) can be used, proper motions are negligible. Further, the epoch at which the GSC positions were obtained is recent, so again the proper motion correction is negligible. One always had certain qualms about accepting this assurance, since the apparent magnitude of a star depends not only on its distance but also on its absolute luminosity. Stars are known to have an enormous range in luminosity, and it is probable that stars of low luminosity stars are the commonest stars in the Galaxy, and consequently many of the apparently faint stars in the GSC may also be intrinsically faint stars that are nearby and may have appreciable proper motions. Furthermore, as time marches inexorably on, the epoch of the GSC becomes less and less “recent” and one cannot go on indefinitely declaring that proper motion corrections are negligible.

Today, however, the catalogue favoured for astrometric observations of asteroids is the $\mathrm{USNU\text{-}B}$ Catalog. (USNO = United States Naval Observatory.) This has positions and proper motions for more than a billion objects, so there is no longer any excuse for not applying proper motion corrections to the comparison stars.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.3: Refraction

Refraction of starlight as it passes through Earth's atmosphere displaces the position of the star towards the zenith. The amount of the refraction is not large close to the zenith, but it amounts to about half a degree near the horizon. Earth's atmosphere is but a thin skin compared with the radius of Earth, and, provided that the star is not close to the horizon, we may treat the atmosphere as a plane-parallel atmosphere. The situation is illustrated in figure XI.5.

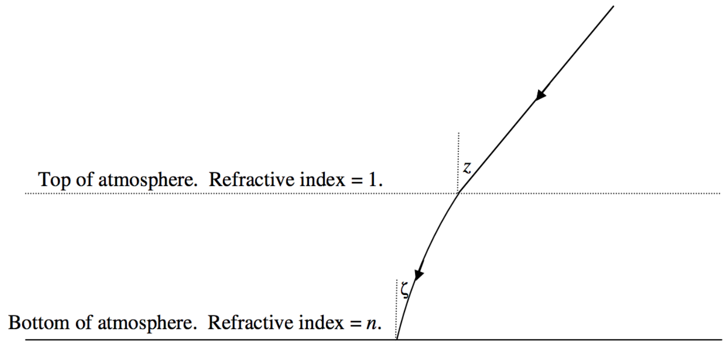


FIGURE XI.5

The angle z is the *true zenith distance* – i.e. the zenith distance it would have in the absence of an atmosphere. The angle ζ is the *apparent zenith distance*. By application of Snell's law, we have $\sin z = n \sin \zeta$, and if we let $\varepsilon = z - \zeta$, this becomes

$$\sin \zeta \cos \varepsilon + \cos \zeta \sin \varepsilon = n \sin \zeta. \tag{11.3.1}$$

Divide both sides by $\sin \zeta$ and make the approximations (correct to first order in ε) $\sin \varepsilon \approx \varepsilon$, $\cos \varepsilon \approx 1$, and we obtain

$$\varepsilon = z - \zeta = (n-1) \tan \zeta. \tag{11.3.2}$$

The refractive index at ground level varies a little with temperature and pressure, but it averages about $n - 1 = 58''.2$. (You didn't know that refractive index was expressed in arcseconds, did you?)

We have made some approximations in deriving Equation 11.3.2, but it must be borne in mind that, as far as astrometry is concerned, what is important is the *differential refraction* between the bottom and top of the detector (photographic film or CCD), and Equation 11.3.2 should be more than adequate – unless one is observing very close to the horizon. The only time when one is likely to be observing close to the horizon would be for a bright comet, for which it is very difficult to make precise measurements anyway. The differential refraction between top and bottom obviously amounts to

$$\delta \varepsilon = (n-1) \sec^2 \zeta \delta \zeta, \tag{11.3.3}$$

where $\delta \zeta$ is the range of zenith distance covered by the detector. In the table below I show the differential refraction between top and bottom of a detector (such as a photographic film) with a 5-degree field, and for a detector (such as a CCD) with a 20-arcminute field, for four zenith distances. Obviously, the correction for differential refraction should be made for the 5-degree photographic field. It might be argued that, for the relatively small field of a 20-arcminute CCD, the correction for differential refraction is unimportant. However, the precision expected for modern CCD astrometry is rather higher than the precision that was expected during the photographic era, and certainly, for large zenith distances, if one hopes for sub-arcsecond astrometry, a correction for differential refraction is desirable. Bear in mind, too, that CCDs are becoming larger as technology advances, and that the larger the CCD, the more important will be the refraction correction.

Zenith distance in degrees	$\delta\varepsilon$ in arcseconds for 5° field	$\delta\varepsilon$ in arcseconds for $20'$ field
15	5.5	0.4
30	6.8	0.5
45	10.2	0.7
60	20.4	1.4

(11.3.3.1)

The most straightforward way of correcting for differential refraction is to calculate the true zenith distance z and azimuth A of each comparison star by the usual methods of spherical astronomy:

$$\cos z=\sin \phi \sin \delta+\cos \phi \cos \delta \cos H \tag{11.3.4}$$

and

$$\tan A=\frac{\sin H}{\cos \phi \tan \delta-\sin \phi \cos H} . \tag{11.3.5}$$

Here ϕ is the observer's latitude, and H is the hour angle of the star, to be found from its right ascension and the local sidereal time. Having found z, then calculate the apparent zenith distance ζ from Equation 11.3.2 (refraction does not, of course, change the azimuth), and then invert Equations 11.3.4 and 11.3.5 to obtain the apparent hour angle H' (and hence apparent right ascension α') and apparent declination δ' of the star. Do this for all the comparison stars. (By hand, this might sound long and tedious, but of course when a computer is programmed to do it, it is all automatic and instantaneous.)

$$\sin \delta'=\sin \phi \cos \zeta+\cos \phi \sin \zeta \cos A \tag{11.3.6}$$

and

$$\tan H'=\frac{\sin A \tan \zeta}{\cos \phi-\sin \phi \cos A \tan \zeta} . \tag{11.3.7}$$

You can then carry out the measurements and from them calculate the apparent right ascension and declination of the asteroid. From these, calculate the apparent zenith distance. Correct this to obtain the true zenith distance, and finally calculate the true right ascension and declination of the asteroid – again all of this is done instantaneously once you have correctly programmed the computer.

Another aspect of refraction that might be considered is that blue (early-type) stars are refracted more than red (late-type) stars. In principle, therefore, one should use only comparison stars that are of the same colour as the asteroid. In practice, I imagine that few astrometrists always do this. If, by ill-fortune, one of the comparison stars is very red or very blue, this may result in a large residual for that star, and the star can be detected and rejected, as described in subsection 11.3.6. Yet another aspect is that, because of *dispersion*, the light from the star – especially if it is low down near the horizon − will be drawn out into a short spectrum, with the red end closer to the horizon than the blue end, and there is then a problem of how to measure the position of the star. The answer is probably to leave asteroids that are close to the horizon to observers who are at a more favourable latitude. As mentioned above, the only time you are likely to observe very low down would be for a long-period comet, on which you cannot set extremely precisely in any case.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.4: Aberration of Light

By "aberration" I am not referring to optical aberrations produced by lenses and mirrors, such as coma and astigmatism and similar optical aberrations, but rather to the *aberration of light* resulting from the vector difference between the velocity of light and the velocity of Earth. (In these notes, the word "velocity" is used to mean "velocity" and the word "speed" is used to mean "speed". The word "velocity" is not to be used merely as a longer and more impressive word for "speed".)

The effect of aberration is to displace a star towards the *Apex of the Earth's Way*, which is the point on the celestial sphere towards which Earth is moving. The apex is where the ecliptic intersects the observer's meridian at 6 hours local apparent solar time. The amount of the aberrational displacement varies with position on the sky, being greatest for stars 90° from the apex. It is then of magnitude v/c, where v and c are the speeds of Earth and light respectively. This amounts to 20.5 arcseconds. (You didn't know that the speed of Earth could be expressed in arcseconds, did you?) But what matters in astrometry is the *differential aberration* between one edge of the detector (photographic film or CCD) and the other. Evidently this is going to be a much smaller effect than differential refraction.

Let us examine the effect of aberration in figures $\mathrm{XI}.6\mathrm{a}$ and b.

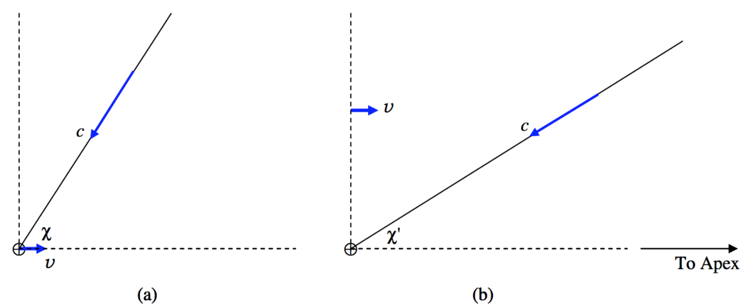


FIGURE XI.6

Part (a) of the figure shows a stationary reference frame. By "stationary" I mean a frame in which Earth, $\oplus$, is moving towards the apex at speed $v = 29.8\ \mathrm{km\ s^{-1}}$. Light from a star is approaching Earth at speed c from a direction that makes an angle χ, which I shall call the true apical distance, with the direction to the apex.

Part (b) shows the same situation referred to a frame in which Earth is stationary; that is the frame (b) is moving towards the apex with speed v relative to the frame (a). Referred to this frame, the speed of light is c, and it is coming from a direction χ', which I shall call the *apparent apical distance*.

I refer to the difference $\varepsilon = \chi - \chi'$ as the *aberrational displacement*.

For brevity I shall refer to the direction to the apex as the "x-direction" and the upwards direction in the figures as the "y-direction".

Referred to frame (a), the x-component of the velocity of light is $-c\cos\chi$, and referred to frame (b), the x-component of the velocity of light is $-c\cos\chi'$. These are related by the Lorentz transformation between velocity components:

$$c\cos\chi' = \frac{c\cos\chi + v}{1 + (v/c)\cos\chi}. \tag{11.3.8}$$

Referred to frame (b), the y-component of the velocity of light is $-c\sin\chi$, and referred to frame (b), the y-component of the velocity of light is $-c\sin\chi'$. These are related by the Lorentz transformation between velocity components:

$$c\sin\chi' = \frac{c\sin\chi}{\gamma(1 + (v/c)\cos\chi)}, \tag{11.3.9}$$

in which, if need be, a c can be cancelled from each side of the Equation. In Equation 11.3.9, γ is the Lorentz factor $1/\sqrt{1-(v/c)^2}$.

Equations 11.3.8 and 9 are not independent; indeed one may be regarded as just another way of writing the other. One easy way to show this, for example, is to show that $\sin^2\chi' + \cos^2\chi' = 1$. In any case, either of them gives χ' as a function of χ

and v/c.

Figure XI.7 shows χ' as a function of χ for $v/c = 0.125,\ 0.250,\ 0.375,\ 0.500,\ 0.625,\ 0.750$ and 0.875.

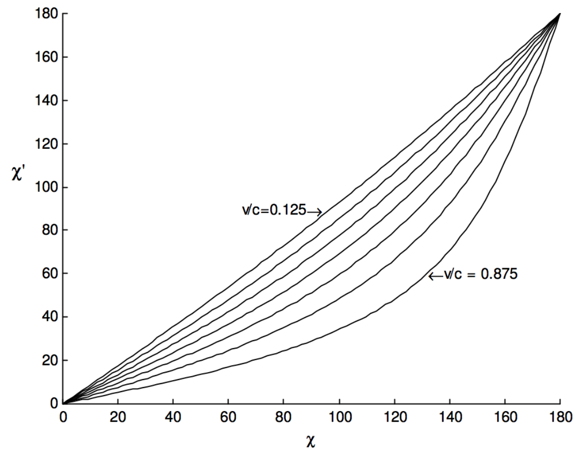


FIGURE XI.7

For Earth in orbit around the Sun, $v = 29.8\ \text{km s}^{-1}$ and $v/c = 9.9 \times 10^{-5}$, which corresponds to an angle of $20''.5$. Thus the aberrational displacement is very small. If we write $\varepsilon = \chi - \chi'$, Equation 11.3.8 takes the form to first order in ε:

$$\cos(\chi - \varepsilon) = \cos\chi + \varepsilon\sin\chi = \frac{\cos\chi + (v/c)}{1 + (v/c)\cos\chi}, \tag{11.3.10}$$

from which, after a very little algebra, we find

$$\varepsilon = \frac{(v/c)\sin\chi}{1 + (v/c)\chi}, \tag{11.3.11}$$

or, since $v/c << 1$,

$$\varepsilon \approx \frac{v\sin\chi}{c}, \tag{11.3.12}$$

Thus we see that the aberrational displacement is zero at the apex and at the antapex, and it reaches is greatest value, $20''.5$, ninety degrees from the apex.

As with refraction, however, it is the *differential* aberration that counts, and if the diameter of the detector field is $\delta\chi$, the difference $\delta\varepsilon$ in the aberrational displacement across the field is

$$\delta\varepsilon = \frac{v\cos\chi\delta\chi}{c}. \tag{11.3.13}$$

Notice that the *differential* aberration is *greatest* at the apex and antapex, and is zero ninety degrees from the apex. It might be noted that the opposition point, where perhaps the majority of asteroid observations are made, is ninety degrees from the apex.

The following table, similar to the one shown for differential refraction, shows the differential aberration across five-degree and 20-arcminute fields for various apical distances.

Apical distance in degrees	$\delta\varepsilon$ in arcseconds for 5° field	$\delta\varepsilon$ in arcseconds for $20'$ field
0	1.8	0.12
15	1.7	0.12
30	1.5	0.10
45	1.3	0.09
60	0.9	0.06
90	0.0	0.00

(11.3.4.1)

It might be concluded that the effect of differential aberration is so small as to be scarcely worth worrying about in most circumstances. However, the expectations for the precision of asteroid astrometry are now rather stringent and are likely to become more exacting as time progresses, and for precise work the correction should be made. One of the problems with pre-packaged astrometry programs is that the user does not always know what corrections are included in the package. The surest way is to do it oneself.

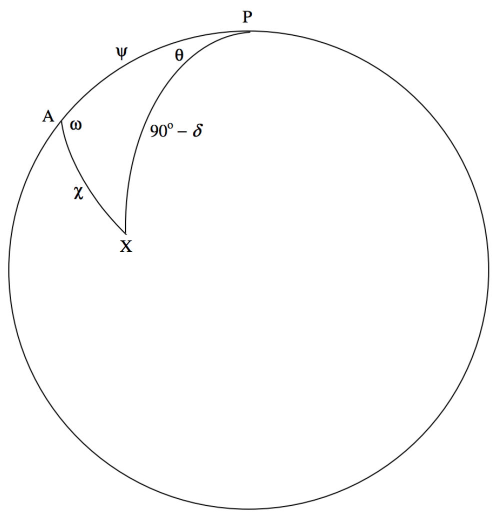


FIGURE XI.8

In figure XI.8, P is the north celestial pole, A is the apex of the Earth's way, and X is a star of true equatorial coordinates (α, δ). The apical distance AX is χ. The angle θ is α(X) − α(A), the angle PAX is ω, and ψ is the distance from pole to apex. It is assumed that the observer knows how to calculate ω and ψ by the usual formulas of spherical astronomy, and hence that all angles in figure XI.8 are known.

From the cotangent formula, we have

$$\cos\psi\cos\omega = \sin\psi\cot\chi - \sin\omega\cos\theta. \tag{11.3.14}$$

If χ is increased by $\delta\chi$, the corresponding increase in θ is given by

$$\sin\omega\sin\theta\delta\theta = \sin\psi\csc^2\chi\delta\chi. \tag{11.3.15}$$

Here $\delta\theta = \alpha'(\mathrm{X}) - \alpha(\mathrm{X})$, where α and α' are, respectively, the true and apparent right ascensions of the star, and $\delta\chi$ is $\chi' - \chi$, which is $-\varepsilon$. It is easy to err in sign at this point, so I re-write Equation 11.3.15more explicitly:

$$(\alpha'(X) - \alpha(X)) \,.\, \sin\omega .\, \sin(\alpha(X) - \alpha(P)) = -\varepsilon\sin\psi\csc^2\chi. \tag{11.3.16}$$

Here ε is the aberrational displacement of X towards A given by Equation 11.3.12. On substitution of Equation 11.3.12into Equation 11.3.16, this becomes, then,

$$(\alpha'(X) - \alpha(X)) \,.\, \sin\omega .\, \sin(\alpha(X) - \alpha(P)) = -\frac{v}{c}\sin\psi\csc\chi. \tag{11.3.17}$$

This enables us to calculate the apparent right ascension of the star.

The declination is obtained from an application of the cosine formula:

$$\sin\delta = \cos\chi\cos\psi + \sin\chi\sin\psi\cos\omega, \tag{11.3.18}$$

from which

$$\cos\delta\delta\delta = (-\cos\psi\sin\chi + \sin\psi\cos\omega\cos\chi)\,\delta\chi. \tag{11.3.19}$$

Here again, as in the usual convention of calculus, $\delta\chi$ represents an increase in χ and $\delta\delta$ is the corresponding increase in δ. But aberration results in a decrease of apical distance, so that $\delta\chi = -\varepsilon$.

Equation 11.3.19 enables us to calculate the apparent declination of the star.

From the measurements of the positions of the comparison stars and the asteroid, we can now calculate the apparent right ascension and declination of the asteroid, and, by inversion of Equations 11.3.17and 11.3.19, we can determine the true right ascension and declination of the asteroid.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.5: Optical Distortion

I refer here to pincushion or barrel distortion introduced by the optical system. This results in a displacement of the stellar images towards or away from the plate centre. Unlike differential refraction or aberration of light, the stellar displacements are symmetric with respect to inversion through the plate centre. This is also true of the optical aberration known as coma. A comatic stellar image results in a displacement of the centre of the stellar image away from the plate centre. Thus we can deal with distortion and coma in a similar manner.

This can be best dealt with by assuming a quadratic relation for the difference between true and measured coordinates:

$$\xi - x = ax^2 + 2hxy + by^2 + 2gx + 2fy + c, \tag{11.3.20}$$

$$\eta - y = a'x^2 + 2h'xy + b'y^2 + 2g'x + f'y + c'. \tag{11.3.21}$$

There are six plate constants in each coordinate, and therefore a minimum of six comparison stars are necessary to solve for them. If more that six are used (which is highly desirable) a least squares solution can be obtained for the plate constants. One then follows the same procedure as in the linear case.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

11.3.6: Errors, Mistakes and Blunders

I was once told that the distinction between errors, mistakes and blunders was roughly as follows. *Errors* are the inevitable small variations caused by imprecision of measurement, or, in the case of computation, the small random errors produced by rounding off (which, incidentally, should not be done before the final “answer” is arrived at). *Mistakes* are things such as writing a 3 instead of 4, or 56 instead of 65, or writing 944 instead of 994 (this is a common one), or reading a poorly-handwritten 6 as a 0 or a 4, or writing a plus sign instead of a minus (this sort of mistake can be quite large!), or thinking that six times eight is 42. A *blunder* is a complete misconception of the entire problem!

Even with the greatest care, errors and mistakes can occur during measurement and reduction of an astrometric plate. The important thing is to find them and either correct or reject them. A stellar image can be contaminated by blending with another star or with a blemish on the plate. A star can be misidentified. There may be a mistake in the catalogued position, or the proper motion may be poor. A measurement can be poor simply because of fatigue or carelessness.

If only the minimum number of comparison stars are used (i.e. three for a linear plate solution, six for a quadratic plate solution), there is no way of detecting errors and mistakes other than carefully repeating the entire measurement and calculation. Error and mistake detection requires an overdetermination of the solution, by using more than the minimum number of comparison stars.

What has to be done is as follows. Once the plate constants have been determined, the right ascension and declination of each of the comparison stars must be calculated, and compared with the right ascension and declination given in the catalogue. The difference $(\mathrm{O}-\mathrm{C})$ is determined for each star, and the standard deviation of the residuals is calculated. Any star with a residual of more that two or three standard deviations should be rejected. The exact criterion for rejection will depend on how many stars we used. Statistical tests will determine the probability that a given residual is a random or gaussian deviation from zero. A full and proper statistical test is slightly laborious (although a computer can make short work of it), and many measurers may decide to reject any star whose residual is more than 2.5 standard deviations from zero, even if this is not strictly the correct statistical way of doing it.

Contributor

- Jeremy Tatum (University of Victoria, Canada)

12: CCD Astrometry

It is now many years since CCD (charge-coupled device) astrometry replaced the photographic plate for astrometry. In practice all astrometry these days is performed with CCD and associated technology, the one possible exception being the measurement of photographs of meteors, which are still commonly recorded on photographic film – although it is probable that the CCD or similar technology will soon replace photographic films even for the measurement of meteors. This chapter, therefore, ought to have been a high priority chapter in the series. It has, unfortunately been delayed while my attention has been occupied with other matters, and in the meantime I have allowed modern methods in astrometry to slip by me and I am not at all qualified to write an authoritative or detailed account of this important subject. However, from time to time correspondents have urged me to fill the gap in this series of topics in celestial mechanics. I shall respond not with an authoritative account of the detailed observing and reduction techniques, but rather with a few general remarks. These remarks will include a comparison of the new methods of CCD astrometry with the older photographic methods. While such a comparison may be of interest to some, the younger generation may be bewildered by it, for, to the modern CCD astrometrist, the CCD is not "new" technology at all; it is not only well established but it is the only technology they have ever known. Many have never handled photographic materials, and indeed photographic emulsions to them are part of the early history of astronomy. Nevertheless some comparison with the old and the new may be of interest.

When CCDs first came into use in astrometry, it was early evident that useful images could be obtained on a CCD far faster than on a photographic emulsion, that far fainter stars could be reached, and higher precision was obtainable. Initial misgivings were that the devices were small and covered only a small area of sky, so that only a few comparison stars were available. Available star catalogues contained positions of only a few hundred thousand stars. As time passed, catalogues were produced that contained many more stars, but there were still misgivings because the newer catalogues, while containing many more stars than the earlier traditional ones, were single-epoch catalogues lacking proper motion data. Against this objection it would be argued that the many faint stars in the newer catalogues were so distant that their proper motions were negligible. This was something of an act of faith, because it is by no means improbable that our Galaxy contains a large number of intrinsically faint stars that are relatively close to us and which may therefore have appreciable proper motions. A further misgiving was that CCDs were relatively insensitive to the blue end of the spectrum – the opposite situation from photographic emulsions, which were typically more sensitive to blue light than to red.

These early perceived drawbacks are now a thing of the past. Modern catalogues suitable for astrometry are available "on line", and contain billions of star positions, and even the initial lack of proper motions is being rapidly remedied.

Let us recall what was involved in obtaining usable astrometric positions of, for example, asteroids, in the photographic era, and compare the situation with the methods in common use today.

In what follows I describe the several steps involved in obtaining and measuring an astrometric position of an asteroid. Under each step I outline what was done (a) in the photographic days and (b) with modern CCD methods.

1. (a) First you had to obtain a photograph of the asteroid. (As Mrs Beeton would have written: "First catch your hare".) To do this would require an exposure of many minutes, or even an hour or even more. During this long exposure time, it was difficult – and tiring – to ensure that the telescope was tracking the stars accurately over such a long time. You could not just allow the telescope to be driven, unattended, by its sidereal drive, but the observer had to stay at the eyepiece for the whole duration of the exposure, constantly vigilant against any small departures from perfect tracking. Of course you would need a second photograph – because the asteroid could only be identified by its motion against the background of the fixed stars. Typically one would wait about an hour before taking the second photograph.

   During a long time exposure, an asteroid would often appear as a short streak, while the stars were (almost) point-like. For faint asteroids, for which an orbit and ephemeris were at least approximately known, a useful (though not particularly easy) technique would be to move the telescope not at the sidereal rate but to follow the predicted motion of the asteroid. That way, the asteroid image would build up, and would appear on the photograph as a point. Thus images of faint asteroid could be obtained. The stars images, of course, then appeared as streaks, and this then made it difficult to measure the streaked stellar images during subsequent analysis of the photograph.

   (b) Today, a CCD still has to be exposed, but exposures are typically just a very few minutes, and the interval between the first and second exposures are again typically measured in minutes. Indeed, because of the speed at which exposures are obtained and the small interval needed between exposures, it is almost universal practice to make at least three exposures in rapid succession, rather than just two with an hour between each.

The corresponding technique for faint asteroids is to take a series (perhaps a dozen or more) of short exposures of the required field, keeping the telescope at sidereal rate. The several images can then be stacked electronically, either (according to choice) so that the stellar images are all stacked upon one another and the asteroid appears as a (barely visible) row of dots, or the several images can be offset before they are stacked, in such a manner that the several asteroid images are stacked upon each other to form an easily-visible pointlike image, and the stars appear as a row of dots. The asteroid position can then be easily measured relative to one of the pointlike stellar images, which remain perfectly usable for astrometric measurement (unlike the streaked stellar images in the photographic method).

2. (a) The photograph had to be developed. This not only meant "messing around" in the darkroom, but one had to wait for hours (after a long night of observing) while the film was first washed and then dried before one could start measurement.

   (b) It is true that a CCD image doesn't have to be "developed" in the same sense that a photographic film had to be – but the CCD observer doesn't quite get off scot-free here. There is a certain amount of "image-processing" that has to be done, and this requires a not inconsiderable amount of experience and know-how. A beginner doing this for the first time may well find it difficult, bewildering and time-consuming. But, once the process has been learned, it becomes very quick and automatic – whereas the process of developing, fixing, washing and drying a photographic plate never gets any easier or faster.
3. (a) Any asteroids on the photograph have to be found. This was done using either a blink comparator or a stereocomparator. In the former the two photographs could be viewed – either through a microscope or projected on to a screen – one after the other in rapid succession. An asteroid would have moved its position relative to the stars between the two exposures, and its presence on the two photographs could be detected because the image of the asteroid would hop to and fro as first one photograph and then the other was viewed. In a stereocomparator, the two photographs would be viewed simultaneously through a stereo binocular microscope. An asteroid that had moved relative to the stars between the two exposures would appear to the eyes, because of a stereoscopic effect to stand up above the plane of the stellar images. These methods were exceedingly effective, but nevertheless a thorough search of a pair of photographs with either of these instruments was time-consuming and tiring.

   (b) The blink technique is also used in CCD astrometry. As mentioned above, it is usual to obtain three images rather than two. The three images can be displayed, one after another in rapid succession, on a computer screen, and any asteroid image will be seen hopping across the screen and back over and over again. In a variation of this technique the three images are obtained through three coloured filters, perhaps red, green and blue. The three images are then stacked on top of each other on the screen, so that the star images appear white. A moving asteroid appears on the screen as three coloured dots (or short dashes) and can be seen very quickly. In yet another technique possible with CCD images, two exposures of a star field can be superimposed on the screen, one positive and the other negative. Thus one image is subtracted from the other, and the computer screen appears blank – except for an asteroid that has moved between exposures. The asteroid appears as two adjacent spots on the screen – one white and one black. Although any of these three techniques is far quicker and less tiring for the measurer than "blinking" or "stereoing" a pair of photographic films, they are by no means the last word in locating asteroid images on CCD exposures, for computer software is available that can detect any object that has moved between two exposures, and can indicate any such objects to the operator.

   One problem with CCD images is that the occasionally faulty pixel on a CCD array can look like as asteroid image on the screen, and also it is common for several pixels to be hit by a cosmic ray particle during the exposure, and this also produces a blemish on the image which looks a bit like an asteroid. However any operator who has measured a few asteroid positions very soon gets to recognize the characteristic appearance of either a bad pixel or a cosmic ray hit, and to distinguish either of these on sight from a real asteroid image. Computer software that is also used to scan pairs of images to detect moving objects can also be programmed to recognize these blemishes, so that in practice they are no real problem to an experienced operator.
4. (a) When we have located an asteroid image on a photographic plate or film, we are not yet ready to start the actual measurement. We have to identify enough comparison stars on the photograph, and look up and write down their right

ascensions, declinations and proper motions by comparison of the photographs with star charts and catalogues. This was always a laborious, tiring and time-consuming part of the procedure, and could occupy a couple of hours or so after a long night of observing and as the evening of the next night rapidly approached.

(b) In the CCD age, this formerly tiresome procedure is over in seconds. All that need be done is to click on the image as many stars as one would like to use as comparison stars. Not just half-a-dozen as in the photographic era, but two or three dozen if you like. The astrometric software in use has access to an enormous catalogue of billions of stars, and instantaneously reads their positions from the catalogue and marks each "clicked" star with a circle for the operator to see. The operator has no need to write down or even to see the positional data of his comparison stars.

5. (a) When we have, after a couple of hours or so, managed to identify the asteroid and the comparison stars on a film or plate, we are at last ready to start the measurement. The film is carefully positioned on the stage of a measuring microscope or "measuring engine" as it was called in the old days. Several settings of a microscope crosshair , in both the x- and the ydirections, were made on the asteroid and the comparison stars. After each setting, a reading of the position was made on a vernier scale that was part of the measuring engine and was duly recorded with pencil and paper. After the asteroid and all the stars had been measured, the film had to be reversed in the measuring engine, and all measurements repeated, in order to allow for systematic measuring errors. The process was very laborious and took several hours for every photograph. In the latter days shortly before CCD astrometry took over, we introduced some quite effective labour-saving devices. We directed a laser beam at a corner reflector attached to the movable microscope stage. The reflected laser beam was interfered with the incident beam to form a system of standing light waves. As the microscope stage moved, a phototransistor counted the number of half-waves, and hence it recorded the position of the microscope stage to a precision, in principle, of half a wavelength. As each setting was made, the position of the microscope stage was sent automatically to the computer that was to be used subsequently to perform the necessary calculations. Apart from greatly increasing the precision of the measurements, the measurer did not have to read a vernier scale, nor even did he have to write down the position. While this device greatly increased the efficiency of the operation, nevertheless several hours were still needed to measure each photograph.

   (b) So how does one measure the positions of the asteroid and the very numerous comparison stars on a CCD? How tedious is the measurement? The astonishing answer is that there is no measurement to be made! The measuring process is bypassed entirely! The reason is that the image of every star sits already on a certain pixel, and all that has to be done is for the computer to read which row and which column that pixel is on. As soon as the exposure is made, the position is already determined! In fact, the situation is even better than that. As described, the positional precision of the measurement is determined by the pixel size. If the pixel measures one arc second by one arcsecond at the focal plane of the telescope, then the precision of the measurement, as we have described it, will be no better than one arcsecond. But this is not the case at all. In practice, a stellar image is spread out over several pixels in two dimensions, each of several pixels holding a certain number of photons. (Not literally photons, of course, but electron-hole pairs, each of which has been generated by a single photon.) The software reads the number of photons in each of the pixels over which the stellar image is distributed, it fits a statistical distribution function (such as a two-dimensional gaussian function) to the image, and calculates the "centre of gravity" of the image to a position of typically about a tenth of a pixel. And so, as soon as the exposure is made, we have the position of the asteroid and of dozens of comparison stars already determined for us to a tenth of an arcsecond or better. Furthermore, the right ascensions and declinations of the comparison stars used are automatically read from an on-line star catalogue, and the calculations to determine the right ascension of the asteroid (or, more probably, of several asteroids recorded on the CCD) are instantaneously computed.

Since all of these calculations can be done instantly by any of several available computer packages, they can be done by anyone with little mathematical training. This has obvious advantages, though the availability of "do-it-yourself" computer packages to the untrained or the unwary may also have some drawbacks. For example, does a given astrometric computer package include such corrections as differential refraction and aberration, proper motion, and so on? Perhaps some do, and some don't. How can one tell – or how can a nonmathematically-trained user determine what corrections are included in the package? For the experienced professional scientist, this may not be a problem, but there are pitfalls to be wary of when a prepackaged program is in the hands of an untrained user, who just wants the "answer" as quickly as possible, without necessarily wanting to know how that answer is obtained.

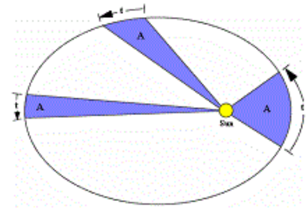
Contributor

- Jeremy Tatum (University of Victoria, Canada)

Thumbnail: Motion of barycenter of solar system relative to the Sun. (CC BY-SA 3.0; Carl Smith derivative work: Rubik-wuerfel).

CHAPTER OVERVIEW

13: CALCULATION OF ORBITAL ELEMENTS



13.1: Introduction to Calculating Orbital Elements

We have seen in Chapter 10 how to calculate an ephemeris from the orbital elements. This chapter deals with the rather more difficult problem of determining the orbital elements from the observations.

We saw in Chapter 2 how to fit an ellipse (or other conic section) to five points in a plane. In the case of a planetary orbit, we need also to know the orientation of the plane, which will require two further bits of information. Thus we should be able to determine the shape, size and orientation of the ellipse from seven pieces of information.

This, however, is not quite the same problem facing us in the determination of a planetary orbit. Most importantly, we do not know *all* of the coordinates of the planet at the time of *any* of the observations. We know two of the coordinates – namely the right ascension and declination – but we have no idea at all of the distance. All that an observation gives us is the direction to the planet in the sky at a given instant of time. Finding the geocentric distance at the time of a given observation is indeed one of the more difficult tasks; once we have managed to do that, we have broken the back of the problem.

However, although we do not know the geocentric (or heliocentric) distances, we do have some additional information to help us. For one thing, we know where one of the foci of the conic section is. The Sun occupies one of them – though we don't immediately know which one. Also, we know the instant of time of each observation, and we know that the radius vector sweeps out equal areas in equal times. This important keplerian law is of great value in computing an orbit.

To determine an orbit, we have to determine a set of six orbital elements. These are, as previously described, $a,\ e,\ i,\ \Omega, \omega$ and T for a sensibly elliptic orbit; for an orbit of low eccentricity one generally substitutes an angle such as M_0, the mean anomaly at the epoch, for T. Thus we can calculate the orbit from six pieces of information. We saw in Chapter 10 how to do this if we know the three heliocentric spatial coordinates and the three heliocentric velocity components – but this again is not quite the problem facing us, because we certainly do not know any of these data for a newly-discovered planet.

If, however, we have three suitably-spaced observations, in which we have measured three directions $(\alpha,\ \delta)$ at three instants of time, then we have six data, from which it may be possible to calculate the six orbital elements. It should be mentioned, however, that three observations are *necessary* to obtain a credible solution, but they may not always be *sufficient*. Should all three observations, for example, be on the ecliptic, or near to a stationary point, or if the planet is moving almost directly towards us for a while and consequently hardly appears to move in the sky, it may not be possible to obtain a credible solution. Or again, observations always have some error associated with them, and small observational errors may under some circumstances translate into a wide range of possible solutions, or it may not even be possible to fit a single set of elements to the slightly erroneous observations.

In recent years, the computation of the orbits of near-Earth asteroids has been a matter of interest for the public press, who are likely to pounce on any suggestion that the observations might have been "erroneous" and the orbit "wrong" – as if they were unaware that all scientific measurement always have error associated with them. There is a failure to distinguish *errors* from *mistakes*.

When a new minor planet or asteroid is discovered, as soon as the requisite minimum number of observations have been made that enable an approximate orbit to be computed, the elements and an ephemeris are distributed to observers. The purpose of this *preliminary orbit* is not to tell us whether planet Earth is about to be destroyed by a cataclysmic collision with a near-Earth asteroid, but is simply to supply observers with a good enough ephemeris that will enable them to find the asteroid and hence to supply additional observations. Everyone who is actively involved in the process of observing asteroids or computing their orbits either knows or ought to know this, just as he also knows or ought to know that, as additional observations come in, the orbit will be *revised* and *differential corrections* will be made to the elements. Further, the computed orbit is generally an *osculating orbit*, and the elements are *osculating elements* for a particular *epoch of osculation*. In order to allow for planetary perturbations, the epoch of osculation is changed every 200 days, and new osculating elements are calculated. All of this is routine and is to be expected. And yet there has been an unfortunate tendency in recent years for not only the press but also for a number of persons who would speak for the scientific community, but who may not themselves be experienced in orbital computations, to attribute the various necessary revisions to an orbit to "mistakes" or "incompetence" by experienced orbit computers.

When all the observations for a particular apparition have been amassed, and no more are expected for that apparition, a definitive orbit for that apparition is calculated from all available observations. Even then, there will be small variations in the elements obtained by different computers. This is because, among other things, each observation has to be critically assessed and weighted. Some observations may be photographic; the majority these days will be higher-precision CCD observations, which will receive a higher weight. Observations will have been made with a variety of telescopes with very different focal lengths, and there will be variations in the experience of the observers involved. Some observations will have been made in a great hurry in the night immediately following a new discovery. Such observations are valuable for computing the preliminary orbit, but may merit less weight in the definitive orbit. There is no unique way for dealing with such problems, and if two computers come up with slightly different answers as a result of weighting the observations differently it does not mean than one of them is “right” and that the other has made a “mistake”. All of this should be very obvious, though some words that have been spoken or written in recent years suggest that it bears repeating.

There are a number of small problems involving the original raw observations. One is that the instant of time of an observation is recorded and reported by an observer in Universal Time. This is the correct thing for an observer to do, and is what is expected of him or her. The computer, however, uses as the argument for the orbital calculation the best representation of a uniformly-flowing dynamical time, which at present is TT, or Terrestrial Time (see chapter 7). The difference for the current year is never known exactly, but has to be estimated. Another difficulty is that observations are not made from the centre of Earth, but from some point on the surface of Earth – a point that is moving as Earth rotates. Thus a small parallactic correction has to be made to the observations – but we do not know how large this correction is until we know the distance of the planet. Or again, the computer needs to know the position of the planet when the sunlight reflected from it left the planet, not when the light eventually arrived at Earth twenty or so minutes later – but we do not know how large the light travel-time correction is until we know the distance of the planet.

There is evidently a good deal involved in computing orbits, and this could be a very long chapter indeed, and never written to perfection to cover all contingencies. In order to get started, however, I shall initially restrict the scope of this chapter to the basic problem of computing elliptical elements from three observations. If and when the spirit moves me I may at a later date expand the chapter to include parabolic and hyperbolic orbits, although the latter pose special problems. Computing hyperbolic elements is in principle no more difficult than computing elliptic orbits; in practice, however, any solar system orbits that are sensibly hyperbolic have been subject to relatively large planetary perturbations, and so the problem in practice is not at all a simple one. Carrying out differential corrections to a preliminary orbit is also something that will have to be left to a later date.

In the sections that follow, I am much indebted to Carlos Montenegro of Argentina who went line-by-line with me through the numerical calculations, resulting in a number of corrections to the original text. Any remaining mistakes (I hope there are few, if any) are my own responsibility.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.2: Triangles

I shall start with a geometric theorem involving triangles, which will be useful as we progress towards our aim of computing orbital elements.

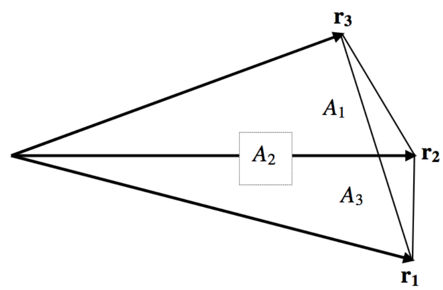


FIGURE XIII.1

Figure XIII.1 shows three coplanar vectors. It is clearly possible to express $\mathbf{r}_2$ as a linear combination of the other two. That is to say, it should be possible to find coefficients such that

$$\mathbf{r}_2 = a_1\mathbf{r}_1 + a_3\mathbf{r}_3. \tag{13.2.1}$$

The notation I am going to use is as follows:

- The area of the triangle formed by joining the tips of $\mathbf{r}_2$ and $\mathbf{r}_3$ is A_1.
- The area of the triangle formed by joining the tips of $\mathbf{r}_3$ and $\mathbf{r}_1$ is A_2.
- The area of the triangle formed by joining the tips of $\mathbf{r}_1$ and $\mathbf{r}_2$ is A_3.

To find the coefficients in Equation 13.2.1, multiply both sides by $\mathbf{r}_1\times$:

$$\mathbf{r}_1 \times \mathbf{r}_2 = a_3\mathbf{r}_1 \times \mathbf{r}_3. \tag{13.2.2}$$

The two vector products are parallel vectors (they are each perpendicular to the plane of the paper), of magnitudes $2A_3$ and $2A_2$ respectively. ($2A_3$ is the area of the *parallelogram* of which the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ form two sides.)

$$\therefore a_3 = A_3/A_2. \tag{13.2.3}$$

Similarly by multiplying both sides of Equation 13.2.1 by $\mathbf{r}_3\times$ it will be found that

$$a_1 = A_1/A_2. \tag{13.2.4}$$

Hence we find that

$$A_2\mathbf{r}_2 = A_1\mathbf{r}_1 + A_3\mathbf{r}_3. \tag{13.2.5}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.3: Sectors

Figure XIII.2 shows a portion of an elliptic (or other conic section) orbit, and it shows the radii vectores of the planet's position at instants of time t_1, t_2 and t_3.

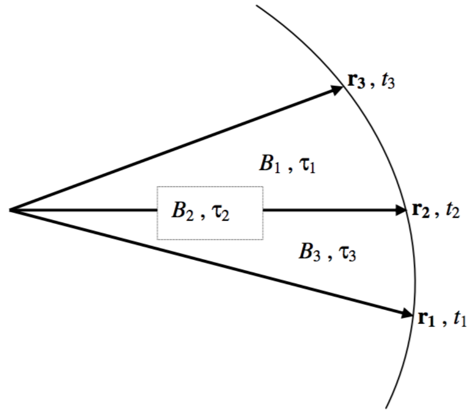


FIGURE XIII.2

The notation I am going to use is as follows:

- The area of the sector formed by joining the tips of $\mathbf{r}_2$ and $\mathbf{r}_3$ around the orbit is B_1.
- The area of the sector formed by joining the tips of $\mathbf{r}_3$ and $\mathbf{r}_1$ around the orbit is B_2.
- The area of the sector formed by joining the tips of $\mathbf{r}_1$ and $\mathbf{r}_2$ around the orbit is B_3.

- The time interval $t_3 - t_2$ is τ_1.
- The time interval $t_3 - t_1$ is τ_2.
- The time interval $t_2 - t_1$ is τ_3.

Provided the arc is fairly small, then to a good approximation (in other words we can approximate the sectors by triangles), we have

$$B_2\mathbf{r}_2 \approx B_1\mathbf{r}_1 + B_3\mathbf{r}_3. \tag{13.3.1}$$

That is,

$$\mathbf{r}_2 \approx b_1\mathbf{r}_1 + b_3\mathbf{r}_3, \tag{13.3.2}$$

where

$$b_1 = B_1/B_2 \tag{13.3.3}$$

and

$$b_3 = B_3/B_2 \tag{13.3.4}$$

The coefficients b_1 and b_3 are the *sector ratios*, and the coefficients a_1 and a_3 are the *triangle ratios*.

By Kepler's second law, the sector areas are proportional to the time intervals.

That is

$$b_1 = \tau_1/\tau_2 \tag{13.3.5}$$

and

$$b_3 = \tau_3/\tau_2. \tag{13.3.6}$$

Thus the coefficients in Equation 13.3.2 are *known*. Our aim is to use this approximate Equation to find approximate values for the heliocentric distances at the instants of the three observations, and then to refine them in order to satisfy the exact

Equation 13.2.5. We shall embark upon our attempt to do this in Section 13.6, but we should first look at the following three sections.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.4: Kepler's Second Law

In section 13.3 we made use of Kepler's second law, namely that the radius vector sweeps out equal areas in equal times. Explicitly,

$$\dot{B} = \frac{1}{2}h = \frac{1}{2}\sqrt{GMl}. \tag{13.4.1}$$

We are treating this as a two-body problem and therefore ignoring planetary perturbations. It is nevertheless worth reminding ourselves – from section 9.5 of chapter 9, especially Equations 9.5.17, 9.4.3, 9.5.19, 9.5.20 and 9.5.21, of the precise meanings of the symbols in Equation 13.4.1. The symbol h is the angular momentum per unit mass of the orbiting body, and l is the semi latus rectum of the orbit. If we are referring to the centre of mass of the two-body system as origin, then h and l are the angular momentum per unit mass of the orbiting body and the semi latus rectum relative to the centre of mass of the system, and M is the *mass function* $M^3/(M+m)^2$ of the system, M and m being the masses of Sun and planet respectively. In chapter 9 we used the symbol $\mathfrak{M}$ for the mass function. If we are referring to the centre of the Sun as origin, then h and l are the angular momentum per unit mass of the planet and the semi latus rectum of the planet's orbit relative to that origin, and M is the sum of the masses of Sun and planet, for which we used the symbol $\mathbf{M}$ in chapter 9. In any case, for all but perhaps the most massive asteroids, we are probably safe in regarding the mass of the orbiting body as being negligible compared with the mass of the Sun. In that case there is no distinction between the centre of the Sun and the centre of mass of the two-body system, and the M in Equation 13.4.1 is then merely the mass of the Sun. (Note that I have n*ot* said that the barycentre of the entire solar system coincides with the centre of the Sun. The mass of Jupiter, for example, is nearly one thousandth of the mass of the Sun, and that is by no means negligible.)

The symbol G, of course, stands for the universal gravitational constant. Its numerical value is not known to any very high precision, and consequently the mass of the Sun is not known to any higher precision than G is. Approximate values for them are $G = 6.672 \times 10^{-11}\ \mathrm{N\,m^2\,kg^{-2}}$ and $M = 1.989 \times 10^{30}\ \mathrm{kg}$. The *product* GM, is known to considerable precision; it is $1.327\ 124\ 38 \times 10^{20}\ \mathrm{m^3\ s^{-2}}$.

Definition: Until June 2012 the *astronomical unit of distance* (au) was defined as the radius of a circular orbit in which a body of negligible mass will, in the absence of planetary perturbations, move around the Sun at an angular speed of exactly $0.017\ 202\ 098\ 95$radians per mean solar day, or $1.990\ 983\ 675 \times 10^{-7}\ \mathrm{rad\,s^{-1}}$, or $0.985\ 607\ 668\ 6$degrees per mean solar day. This angular speed is sometimes called the *gaussian constant* and is given the symbol k. With this definition, the value of the astronomical unit is approximately $1.495\ 978\ 70 \times 10^{11}$ m.

However, in June 2012 the International Astronomical Union re-defined the astronomical unit as $149\ 597\ 870\ 700$ metres exactly. This means that a body of negligible mass moving around the Sun in a circular orbit will, in the absence of planetary perturbations, move at an angular speed of approximately $0.017\ 202\ 098\ 95$radians per mean solar day, This angular speed is the *gaussian constant* k - but, with the new definition of the au, it is no longer regarded as one of the fundamental astronomical constants. The IAU also recommended that the official abbreviation for the astronomical unit should be au.

If we equate the centripetal acceleration of the hypothetical body moving in a circular orbit of radius $1\ \mathrm{au}$ at angular speed k to the gravitational force on it per unit mass, we see that $ak^2 = GM/a^2$, so that

$$GM = k^2a^3, \tag{13.4.2}$$

where a is the length of the astronomical unit and k is the gaussian constant.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.5: Coordinates

We need to make use of several coordinate systems, and I reproduce here the descriptions of them from section 10.7 of chapter 10. You may wish to refer back to that chapter as a further reminder.

1. *Heliocentric plane-of-orbit*. $\odot xyz$ with the $\odot x$ axis directed towards perihelion. The polar coordinates in the plane of the orbit are the heliocentric distance r and the true anomaly v. The z-component of the asteroid is necessarily zero, and $x = r\cos v$ and $y = r\sin v$.

2. *Heliocentric ecliptic*. $\odot XYZ$ with the $\odot X$ axis directed towards the First Point of Aries Υ, where Earth, as seen from the Sun, will be situated on or near September 22. The spherical coordinates in this system are the heliocentric distance r, the ecliptic longitude λ, and the ecliptic latitude β, such that $X = r\cos\beta\cos\lambda$, $Y = r\cos\beta\sin\lambda$ and $Z = r\sin\beta$.

3. *Heliocentric equatorial coordinates*. $\odot\xi\eta\zeta$ with the $\odot\xi$ axis directed towards the First Point of Aries and therefore coincident with the $\odot X$ axis . The angle between the $\odot Z$ axis and the $\odot\zeta$ axis is ε, the obliquity of the ecliptic. This is also the angle between the XY-plane (plane of the ecliptic, or of Earth's orbit) and the $\xi\eta$-plane (plane of Earth's equator). See figure X.4.

4. *Geocentric equatorial coordinates*. $\oplus\mathfrak{x}\mathfrak{y}\mathfrak{z}$ with the $\oplus\mathfrak{x}$ axis directed towards the First Point of Aries. The spherical coordinates in this system are the geocentric distance Δ, the right ascension α and the declination δ, such that $\mathfrak{x} = \Delta\cos\delta\cos\alpha$, $\mathfrak{y} = \Delta\cos\delta\sin\alpha$ and $\mathfrak{z} = \Delta\sin\delta$.

A summary of the relations between them is as follows

$$\mathfrak{x} = \Delta\cos\alpha\cos\delta = l\Delta = \mathfrak{x}_o + \xi, \tag{13.5.1}$$

$$\mathfrak{y} = \Delta\sin\alpha\cos\delta = m\Delta = \mathfrak{y}_o + \eta, \tag{13.5.2}$$

$$\mathfrak{z} = \Delta\sin\delta = n\Delta = \mathfrak{z}_o + \zeta. \tag{13.5.3}$$

Here, $(l,\ m,\ n)$ are the direction cosines of the planet's geocentric radius vector. They offer an alternative way to $(\alpha,\ \delta)$ for expressing the direction to the planet as seen from Earth. They are not independent but are related by

$$l^2 + m^2 + n^2 = 1. \tag{13.5.4}$$

The symbols $\mathfrak{x}_o$, $\mathfrak{y}_o$ and $\mathfrak{z}_o$ are the geocentric equatorial coordinates of the Sun.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.6: Example

As we proceed with the theory, we shall try an actual numerical example as we go. We shall suppose that the following three observations are available:

0^h TT	R.A. (J2000.0)	Dec. (J2000.0)
2002 Jul 10	$21^h\ 15^m.40$ $= 318^\circ.8500$ $= 5.564\ 982$ rad	$+16^\circ\ 13'.8$ $= +16^\circ.2300$ $= +0.283\ 267$ rad
2002 Jul 15	$21^h\ 12^m.44$ $= 318^\circ.1100$ $= 5.552\ 067$ rad	$+16^\circ\ 03'.5$ $= +16^\circ.0583$ $= +\ 0.280\ 271$ rad
2002 Jul 25	$21^h\ 05^m.60$ $= 316^\circ.4000$ $= 5.522\ 222$ rad	$+15^\circ\ 24.8$ $= +15^\circ.4133$ $= +\ 0.269\ 013$ rad

We shall suppose that the times given are 0^h TT, and that the observations were made by an observer at the centre of Earth. In practice, an observer will report his or her observations in Universal Time, and from the surface of Earth. We shall deal with these two refinements at a later time.

The "observations" given above are actually from an ephemeris for the minor planet 2 Pallas published by the Minor Planet Center of the International Astronomical Union. They will not be expected to reproduce exactly the elements also published by the MPC, because the ephemeris positions are rounded off to $0^m.01$ and $0'.1$, and of course the MPC elements are computed from all available observations, not just three. But we should be able to compute elements close to the correct ones. Observations are usually given to a precision of about 0.1 arcsec. For the purposes of the illustrative calculation let us start the calculation with the right ascensions and declinations given above to six decimal places as exact.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.7: Geocentric and Heliocentric Distances - First Attempt

Let us write down the three heliocentric equatorial components of Equation 13.2.1:

$$\xi_2 = a_1\xi_1 + a_3\xi_3, \tag{13.7.1}$$

$$\eta_2 = a_1\eta_1 + a_3\eta_3, \tag{13.7.2}$$

$$\zeta_2 = a_1\zeta_1 + a_3\zeta_3. \tag{13.7.3}$$

Now write $l\,\Delta - \mathfrak{x}_o$ for ξ, etc., from Equations 13.5.1,2,3 and rearrange to take the solar coordinates to the right hand side:

$$l_1 a_1\,\Delta_1 - l_2\,\Delta_2 + l_3 a_3\,\Delta_3 = a_1\mathfrak{x}_{o1} - \mathfrak{x}_{o2} + a_3\mathfrak{x}_{o3}, \tag{13.7.4}$$

$$m_1 a_1\,\Delta_1 - m_2\,\Delta_2 + m_3 a_3\,\Delta_3 = a_1\mathfrak{y}_{o1} - \mathfrak{y}_{o2} + a_3\mathfrak{y}_{o3}, \tag{13.7.5}$$

$$n_1 a_1\,\Delta_1 - n_2\,\Delta_2 + n_3 a_3\,\Delta_3 = a_1\mathfrak{z}_{o1} - \mathfrak{z}_{o2} + a_3\mathfrak{z}_{o3}. \tag{13.7.6}$$

As a very first, crude, approximation, we can let $a_1 = b_1$ and $a_3 = b_3$, for we know b_1 and b_3 (in our numerical example, $b_1 = \frac{2}{3}$, $b_3 = \frac{1}{3}$), so we can solve Equations 13.7.4,5,6 for the three geocentric distances. However, we shall eventually need to find the correct values of a_1 and a_3.

When we have solved these Equations for the geocentric distances, we can then find the heliocentric distances from Equations 13.5.1,2 and 3. For example,

$$\xi_1 = l_1\,\Delta_1 - \mathfrak{x}_{o1} \tag{13.7.7}$$

and of course

$$r_1^2 = \xi_1^2 + \eta_1^2 + \zeta_1^2. \tag{13.7.8}$$

In our numerical example, we have

$$l_1 = +0.722\,980\,907$$
$$l_2 = +0.715\,380\,933$$
$$l_3 = +0.698\,125\,992$$

$$m_1 = -0.631\,808\,343$$
$$m_2 = -0.641\,649\,261$$
$$m_3 = -0.664\,816\,398$$

$$n_1 = +0.279\,493\,876$$
$$n_2 = +0.276\,615\,882$$
$$n_3 = +0.265\,780\,465$$

As a check on the arithmetic, the reader can - and should - verify that

$$l_1^2 + m_1^2 + n_1^2 = l_2^2 + m_2^2 + n_2^2 = l_3^2 + m_3^2 + n_3^2 = 1$$

This does not verify the signs of the direction cosines, for which care should be taken.

From *The Astronomical Almanac* for 2002, we find that

$\mathfrak{x}_{o1} = -306\,728\,3$ $\quad\mathfrak{y}_{o1} = +0.889\,290\,0$ $\quad\mathfrak{z}_{o1} = +0.385\,549\,5$ AU
$\mathfrak{x}_{o2} = -386\,194\,4$ $\quad\mathfrak{y}_{o2} = +0.862\,645\,7$ $\quad\mathfrak{z}_{o2} = +0.373\,999\,6$
$\mathfrak{x}_{o3} = -536\,330\,8$ $\quad\mathfrak{y}_{o3} = +0.791\,387\,2$ $\quad\mathfrak{z}_{o2} = +0.343\,100\,4$

(For a fraction of a day, which will usually be the case, these coordinates can be obtained by nonlinear interpolation – see chapter 1, section 1.10.)

Equations 13.7.4,5,6 become

$$+0.481\,987\,271\,\Delta_1 - 0.\,715\,380\,933\,\Delta_2 + 0.\,232\,708\,664\,\Delta_3 = 0.002\,931\,933$$

$$-0.421\,205\,562\,\Delta_1 + 0.641\,649\,261\,\Delta_2 - 0.221\,605\,466\,\Delta_3 = -0.005\,989\,967$$

$+0.186\ 329\ 251\ \Delta_1\ -0.276\ 615\ 882\ \Delta_2\ -0.088\ 593\ 488\ \Delta_3 = -0.002\ 599\ 800$

I give below the solutions to these Equations, which are our first crude approximations to the geocentric distances in AU, together with the corresponding heliocentric distances. I also give, for comparison, the correct values, from the published MPC ephemeris

$$\begin{array}{llll} \multicolumn{2}{c}{\text{First crude estimates}} & \multicolumn{2}{c}{\text{MPC}} \\ \Delta_1 = 2.725\ 71 & r_1 = 3.485\ 32 & \Delta_1 = 2.644 & r_1 = 3.406 \\ \Delta_2 = 2.681\ 60 & r_2 = 3.481\ 33 & \Delta_2 = 2.603 & r_2 = 3.404 \\ \Delta_3 = 2.610\ 73 & r_3 = 3.474\ 71 & \Delta_3 = 2.536 & r_3 = 3.401 \end{array} \tag{13.7.1}$$

This must justifiably give cause for some satisfaction, because we now have some idea of the geocentric distances of the planet at the instants of the three observations, though it is a little early to open the champagne bottles. We still have a little way to go, for we must refine our values of a_1 and a_3. Our first guesses, $a_1 = b_1$ and $a_3 = b_3$, are not quite good enough.

The key to finding the geocentric and heliocentric distances is to be able to determine the triangle ratios $a_1 = A_1/A_2$, $a_3 = A_3/A_2$ and the triangle/sector ratios a/b. The sector ratios are found easily from Kepler's second law. We have made our first very crude attempt to find the geocentric and heliocentric distances by assuming that the triangle ratios are equal to the sector ratios. It is now time to improve on that assumption, and to obtain better triangle ratios. After what may seem like a considerable amount of work, we shall obtain approximate formulas, Equations 13.8.35a,b, for improved triangle ratios. The reader who does not wish to burden himself with the details of the derivation of these Equations may proceed directly to them, near the end of Section 13.7

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.8: Improved Triangle Ratios

The Equation of motion of the orbiting body is

$$\ddot{\mathbf{r}} = -\frac{GM}{r^3}\mathbf{r}. \tag{13.8.1}$$

If we recall Equation 13.4.2, this can be written

$$\ddot{\mathbf{r}} = -k^2\left(\frac{a^3}{r^3}\right)\mathbf{r}. \tag{13.8.2}$$

If we now agree to express r in units of a (i.e. in Astronomical Units of length (au)) and time in units of $1/k$ ($1/k = 58.132\,440\,87$ mean solar days), this becomes merely

$$\ddot{\mathbf{r}} = -\frac{1}{r^3}\mathbf{r}. \tag{13.8.3}$$

In these units, GM has the value 1.

Now write the x- and y- components of this Equation, where $(x,\ y)$ are heliocentric coordinates in the plane of the orbit (see sections 13.5 or 10.7).

$$\ddot{x} = -\frac{x}{r^3} \tag{13.8.4}$$

and

$$\ddot{y} = -\frac{y}{r^3}, \tag{13.8.5}$$

where

$$x^2 + y^2 = r^2. \tag{13.8.6}$$

The areal speed is $\frac{1}{2}h = \frac{1}{2}\sqrt{GMl}$ or, in these units, $\frac{1}{2}\sqrt{l}$ where l is the semi latus rectum of the orbit in au

Let the planet be at $(x,\ y)$ at time t. Then at time $t+\delta t$ it will be at $(x+\delta x, y+\delta y)$, where

$$\delta x = \dot{x}\delta t + \frac{1}{2!}\ddot{x}(\delta t)^2 + \frac{1}{3!}\dddot{x}(\delta t)^3 + \frac{1}{4!}\ddddot{x}(\delta t)^4 + \dots \tag{13.8.7}$$

and similarly for y.

Now, starting from Equation 13.8.4 we obtain

$$\dddot{x} = \frac{3x\dot{r}}{r^4} - \frac{\dot{x}}{r^3} \tag{13.8.8}$$

and

$$\ddddot{x} = 3\left(\frac{\dot{x}\dot{r}}{r^4} + \frac{x\ddot{r}}{r^4} - \frac{4x\dot{r}^2}{r^5}\right) - \frac{r^3\ddot{x} - 3r^2\dot{x}\dot{r}}{r^6}. \tag{13.8.9}$$

(The comment in the paragraph preceding Equation 3.4.16 may be of help here, in case this is heavy-going.)

Now $\ddot{x}$ and x are related by Equation 13.8.4, so that we can write Equation 13.8.9 with no time derivatives of x higher than the first, and indeed it is not difficult, because Equation 13.8.4 is just $r^3\ddot{x} = -x$. We obtain

and

$$\ddddot{x} = x\left(\frac{1}{r^6} - \frac{12\dot{r}^2}{r^5} + \frac{3\ddot{r}}{r^4}\right) + \frac{6\dot{x}\dot{r}}{r^4}. \tag{13.8.10}$$

In a similar fashion, because of the relation 13.8.4, all higher time derivatives of x can be written with no derivatives of x higher than the first, and a similar argument holds for y.

Thus we can write Equation 13.8.7 as

$$x+\delta x=Fx+G\dot{x} \tag{13.8.11}$$

and similarly for y:

$$y+\delta y=Fy+G\dot{y}, \tag{13.8.12}$$

where

$$F=1-\frac{1}{2r^3}(\delta t)^2+\frac{\dot{r}}{2r^4}(\delta t)^3+\frac{1}{24}\left(\frac{1}{r^6}-\frac{12\dot{r}^2}{r^5}+\frac{3\ddot{r}}{r^4}\right)(\delta t)^4+\ldots \tag{13.8.13}$$

and

$$G=\delta t-\frac{1}{6r^3}(\delta t)^3+\frac{\dot{r}}{4r^4}(\delta t)^4+\ldots \tag{13.8.14}$$

Now we are going to look at the triangle and sector areas. From figure XIII.1 we can see that

$$\mathbf{A}_1=\frac{1}{2}\mathbf{r}_2\times\mathbf{r}_3,\ \mathbf{A}_2=\frac{1}{2}\mathbf{r}_1\times\mathbf{r}_3,\ \mathbf{A}_3=\frac{1}{2}\mathbf{r}_1\times\mathbf{r}_2. \tag{13.8.15a,b,c}$$

Also, angular momentum per unit mass is $\mathbf{r}\times\mathbf{v}$ and Kepler's second law tells us that areal speed is half the angular momentum per unit mass and that it is constant and equal to $\frac{1}{2}\sqrt{l}$ (in the units that we are using), so that

$$\dot{\mathbf{B}}_1=\frac{1}{2}\mathbf{r}_1\times\dot{\mathbf{r}}_1=\frac{1}{2}\mathbf{r}_2\times\dot{\mathbf{r}}_2=\frac{1}{2}\mathbf{r}_3\times\dot{\mathbf{r}}_3. \tag{13.8.16a,b,c}$$

All four of these vectors are parallel and perpendicular to the plane of the orbit, to that their magnitudes are just equal to their z-components. From the usual formulas for the components of a vector product we have, then,

$$A_1=\frac{1}{2}(x_2y_3-y_2x_3),\quad A_2=\frac{1}{2}(x_1y_3-y_1x_3),\quad A_3=\frac{1}{2}(x_1y_2-y_1x_2) \tag{13.8.17a,b,c}$$

and

$$\frac{1}{2}\sqrt{l}=\frac{1}{2}(x_1\dot{y}_1-y_1\dot{x}_1)=\frac{1}{2}(x_2\dot{y}_2-y_2\dot{x}_2)=\frac{1}{2}(x_3\dot{y}_3-y_3\dot{x}_3). \tag{13.8.18a,b,c}$$

Now, start from the second observation (x_2,y_2) at instant t_2. We shall try to predict the third observation, using Equations 13.8.11-14, in which $x+\delta x$ is x_3 and δt is t_3-t_2, which we are calling (see section 13.3) τ_1. I shall make the subscripts for F and G the same as the subscripts for τ. Thus the F and G that connect observations 2 and 3 will have subscript 1, just as we are using the notation τ_1 for t_3-t_2.

Thus we have

$$x_2=F_1x_2+G_1\dot{x}_2 \tag{13.8.19}$$

and

$$y_3=F_1y_2+G_1\dot{y}_2, \tag{13.8.20}$$

where

$$F_1=1-\frac{1}{2r_2^3}\tau_1^2+\frac{\dot{r}_2}{2r_2^4}\tau_1^3+\frac{1}{24}\left(\frac{1}{r_2^6}-\frac{12\dot{r}_2^2}{r_2^5}+\frac{3\ddot{r}_2}{r_2^4}\right)\tau_1^4+\ldots \tag{13.8.21}$$

and

$$G_1=\tau_1-\frac{1}{6r_2^3}\tau_1^3+\frac{\dot{r}_2}{4r_2^3}\tau_1^4+\ldots\quad. \tag{13.8.22}$$

Similarly, the first observation is given by

$$x_1 = F_3 x_2 + G_3 \dot{x}_2 \tag{13.8.23}$$

and

$$y_1 = F_3 y_2 + G_3 \dot{y}_2, \tag{13.8.24}$$

where, by substitution of $-\tau_3$ for δt,

$$F_3 = 1 - \frac{1}{2r_2^3}\tau_3^2 - \frac{\dot{r}_2}{2r_2^4}\tau_3^3 + \frac{1}{24}\left(\frac{1}{r_2^6} - \frac{12\dot{r}_2^2}{r_2^5} + \frac{3\ddot{r}_2}{r_2^4}\right)\tau_3^4 + \ldots \tag{13.8.25}$$

and

$$G_3 = -\tau_3 + \frac{1}{6r_2^3}\tau_3^3 + \frac{\dot{r}_2}{4r_2^4}\tau_3^4 + \ldots \quad . \tag{13.8.26}$$

From Equations 13.8.17,18,19,20,23,24, we soon find that

$$A_1 = \frac{1}{2}G_1\sqrt{l}, \quad A_2 = \frac{1}{2}(F_3G_1 - F_1G_3)\sqrt{l}, \quad A_3 = -\frac{1}{2}G_3\sqrt{l}. \tag{13.8.27a,b,c}$$

Now we do not yet know $\dot{r}$ or $\ddot{r}$, but we can take the expansions of F and G as far as τ^2. We then obtain, correct to τ^3 :

$$A_1 = \frac{1}{2}\sqrt{l}\tau_1\left(1 - \frac{\tau_1^2}{6_2^3}\right), \tag{13.8.28}$$

$$A_2 = \frac{1}{2}\sqrt{l}\tau_2\left(1 - \frac{\tau_1^2}{6r_2^3}\right), \tag{13,8.29}$$

and

$$A_3 = \frac{1}{2}\sqrt{l}\tau_3\left(1 - \frac{\tau_3^2}{6r_2^3}\right). \tag{13.8.30}$$

Thus the triangle ratio $a_1 = A_1/A_2$ is

$$a_1 = \frac{\tau_1}{\tau_2}\left(1 - \frac{\tau_1^2}{6r_2^3}\right)\left(1 - \frac{\tau_2^2}{6r_2^3}\right)^{-1}, \tag{13.8.31}$$

which, to order τ^3, is

$$a_1 = \frac{\tau_1}{\tau_2}\left(1 + \frac{(\tau_2^2 - \tau_1^2)}{6r_2^3}\right), \tag{13.8.32}$$

or, with $\tau_2 - \tau_1 = \tau_3$,

$$a_1 = \frac{\tau_1}{\tau_2}\left(1 + \frac{\tau_3(\tau_2 + \tau_1)}{6r_2^3}\right). \tag{13.8.33}$$

Similarly,

$$a_3 = \frac{\tau_3}{\tau_2}\left(1 + \frac{\tau_1(\tau_2 + \tau_3)}{6r_2^3}\right). \tag{13.8.34}$$

Further, with $\tau_1/\tau_2 = b_1$ and $\tau_3/\tau_2 = b_3$,

$$a_1 = b_1 + \frac{\tau_1\tau_3}{6r_2^3}(1 + b_1) \quad \text{and} \quad a_3 = b_3 + \frac{\tau_1\tau_3}{6r_2^3}(1 + b_3). \tag{13.8.35a,b}$$

These will serve as better approximations for the triangle ratios. Be aware, however, that Equations 13.8.35a,b are only approximations, and do not give the *exact* values for a_1 and a_3.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.9: Iterating

We can now use Equations 13.8.35a,b and get a better estimate of the triangle ratios. The numerical data are

$$b_1 = 2/3, \quad b_3 = 1/3, \quad r_2 = 3.481\,33,$$

$\tau_1 = t_3 - t_2 = 10$ mean solar days and $\tau_3 = t_2 - t_1 = 5$ mean solar days, but recall that we are expressing time intervals in units of $1/k$, which is $58.132\,440\,87$mean solar days, and therefore

$$\tau_1 = 0.172\,021 \quad \text{and} \quad \tau_3 = 0.086\,010.$$

Equations 13.8.35 then result in

$$a_1 = 0.666\,764, a_3 = 0.333\,411$$

Now we can go back to Equation 13.7.4 and start again with our new values for the triangle ratios – und so weiter − until we obtain new values for Δ_1, Δ_2, Δ_3 and r_2. I show below in the first two columns the first crude estimates (already given above), in the 16 second two columns the results of the first iteration, and, in the last two columns, the values given in the published IAU ephemeris.

	First crude estimates		First iteration		MPC	
	Δ	r	Δ	r	Δ	r
1	2.72571	3.48532	2.65825	3.41952	2.644	3.406
2	2.68160	3.48133	2.61558	3.41673	2.603	3.404
3	2.61073	3.47471	2.54579	3.41082	2.536	3.401

(13.9.1)

We see that we have made a substantial improvement, but we are not there yet. We can now calculate new values of a_1 and a_3 from Equations 13.8.35a,b to get

$$a_1 = 0.666\,770 \quad a_3 = 0.333\,416$$

We *could* (if we so wished) now go back to Equations 13.7.4,5,6, and iterate again. However, this will result in only small changes to a_1, a_3, Δ and r, and we have to bear in mind that Equations 13.8.35a,b are only approximations (to order τ^3). Therefore, even if successive iterations converge, they will still not give precise correct answers for Δ and r.

To anticipate, eventually we shall arrive at some exact Equations (Equations 13.12.25 and 13.12.26) that will allow us to solve the problem. But these Equations will not be easy to solve. They have to be solved by iteration using a reasonably good first guess. It is our present aim to obtain a reasonably good first guess for a_1, a_3, Δ and r, in order to prepare for the solution of the exact Equations 13.12.25 and 13.12.26. Our current values of a_1 and a_3, while not exact, will enable us to solve Equations 13.12.25 and 13.12.26 exactly, so we should now, rather than going back again to Equations 13.7.4,5,6, proceed straight to Sections 13.11, 13.12 and 13.13.

Nevertheless, in the following section, we provide (in Equations 13.10.9 and 13.10.10), after considerable effort, higher-order expansions for a_1 and a_3. These may be useful, but for reasons explained in the previous paragraph, it may be easier to skip Section 13.10 entirely.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.10: Higher-order Approximation

The reason that we made the approximation to order τ^3 was that, in evaluating the expressions for F_1, G_1, F_3 and G_3, we did not know the radial velocity $\dot{r}_2$. Perhaps we can now evaluate it.

Exercise. Show that the radial velocity of a particle in orbit around the Sun, when it is at a distance r from the Sun, is

$$\text{Ellipse}: \quad \dot{r} = \mp\sqrt{\frac{GM}{a_0}}\left(\frac{a^2e^2-(a-r)^2}{ar^2}\right)^{1/2}, \tag{13.10.1}$$

$$\text{Parabola}: \quad \dot{r} = \mp\sqrt{\frac{GM(r-q)}{a_0}}, \tag{13.10.2}$$

$$\text{Hyperbola}: \quad \dot{r} = \mp\sqrt{\frac{GM}{a_0}}\left(\frac{(a+r)^2-a^2e^2}{ar^2}\right)^{1/2}. \tag{13.10.3}$$

Show that the radial velocity is greatest at the ends of a latus rectum.

Here a_0 is the astronomical unit, a is the semi major axis of the elliptic orbit or the semi transverse axis of the hyperbolic orbit, q is the perihelion distance of the parabolic orbit, and e is the orbital eccentricity. The $-$ sign is for pre-perihelion, and the $+$ sign is for post-perihelion.

Unfortunately, while this is a nice exercise in orbit theory, we do not know the eccentricity, so these formulas at present are of no use to us.

However, we can calculate the heliocentric distances at the times of the first and third observations by exactly the same method as we used for the second observation. Here are the results for our numerical example, after one iteration. The units, of course, are au. Also indicated are the instants of the observations, taking $t_2=0$ and expressing the other instants in units of $1/k$ (see section 13.8).

$$t_1 = -\tau_3 = -0.086\,010\,494\,75 \quad r_1 = 3.419\,52$$

$$t_2 = 0 \quad r_2 = 3.416\,73$$

$$t_3 = +\tau_1 = +0.172\,020\,989\,5 \quad r_3 = 3.410\,82$$

We can fit a quadratic expression to this, of the form:

$$r = c_0 + c_1 t + c_2 t^2 \tag{13.10.4}$$

With our choice of time origin $t_2 = 0$, c_0 is obviously just equal to r_2, so we have just two constants, c_1 and c_2 to solve for. We can then calculate the radial velocity at the time of the second observation from

$$\dot{r}_2 = c_1 + 2c_2t_2. \tag{13.10.5}$$

We can calculate A_1, A_2 and A_3 in the same manner as before, up to τ^4 rather than just τ^3. The algebra is slightly long and tedious, but straightforward. Likewise, the results look long and unwieldy, but there is no difficulty in programming them for a computer, and the actual calculation is, with a modern computer, virtually instantaneous. The results of the algebra that I give below are taken from the book *Determination of Orbits* by A.D. Dubyago (which has been the basis of much of this chapter). I haven't checked the algebra myself, but the conscientious reader will probably want to do so himself or herself.

$$A_1 = \frac{1}{2}\sqrt{l}\tau_1\left(1 - \frac{\tau_1^2}{6r_2^3} + \frac{\tau_1^3}{4r_2^4}\dot{r}_2\right), \tag{13.10.6}$$

$$A_2 = \frac{1}{2}\sqrt{l}\tau_2\left(1 - \frac{\tau_2^2}{6r_2^3} + \frac{\tau_2^2(\tau_1-\tau_3)}{4r_2^4}\dot{r}_2\right), \tag{13.10.7}$$

$$A_3 = \frac{1}{2}\sqrt{l}\tau_3\left(1 - \frac{\tau_3^2}{6r_2^3} - \frac{\tau_3^3}{4r_2^4}\dot{r}_2\right). \tag{13.10.8}$$

And from these,

$$a_1 = \frac{\tau_1}{\tau_2}\left(1 + \frac{\tau_3(\tau_2+\tau_1)}{6r_2^3} + \frac{\tau_3(\tau_3(\tau_1+\tau_3)-\tau_1^2)}{4r_2^4}\dot{r}_2\right). \tag{13.10.9}$$

and

$$a_3 = \frac{\tau_3}{\tau_2}\left(1 + \frac{\tau_1(\tau_2+\tau_3)}{6r_2^3} - \frac{\tau_1(\tau_1(\tau_1+\tau_3)-\tau_3^2}{4r_2^4}\dot{r}_2\right). \tag{13.10.10}$$

This might result in slightly better values for a_1 and a_3. I have not calculated this for our numerical example here, for reasons given in Section 13.9. We can move on to the next section, using our current vales of a_1 and a_3, namely

$$a_1 = 0.666\,770 \quad \text{and} \quad a_3 = 0.333\,416.$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.11: Light-time Correction

Before going further, however, our current estimates of the geocentric distances are now sufficiently good that we should make the light-time corrections. The observed positions of the planet were not the positions that they occupied at the instants when they were observed. It actually occupied these observed positions at times $t_1 - \Delta_1/c$, $t_2 - \Delta_2/c$ and $t_3 - \Delta_3/c$. Here, c is the speed of light, which, as everyone knows, is 10065.320 astronomical units per $1/k$. The calculation up to this point can now be repeated with these new times. This may seem tedious, but of course with a computer, all one needs is a single statement telling the computer to go to the beginning of the program and to do it again. I am not going to do it with our particular numerical example, since the "observations" that we are using are in fact predicted positions from a Minor Planet Center ephemeris.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.12: Sector-Triangle Ratio

We recall that it is easy to determine the ratio of adjacent sectors swept out by the radius vector. By Kepler's second law, it is just the ratio of the two time intervals. What we really need, however, are the triangle ratios, which are related to the heliocentric distance by Equation 13.2.1. Oh, wouldn't it just be so nice if someone were to tell us the ratio of a sector area to the corresponding triangle area! We shall try in this section to do just that.

$$Notation: \text{ Triangle ratios}: \quad a_1 = A_1/A_2, \quad a_3 = A_3/A_2. \tag{13.12.1a,b}$$

$$\text{Sector ratios}: \quad b_1 = B_1/B_2, \quad b_3 = B_3/B_2. \tag{13.12.2a,b}$$

$$\text{Sector-triangle ratios}: \quad R_1 = \frac{B_1}{A_1}, \quad R_2 \frac{B_2}{A_2}, \quad R_3 = \frac{B_3}{A_3}, \tag{13.12.3a,b,c}$$

from which it follows that

$$a_1 = \frac{R_2}{R_1} b_1, \quad a_3 = \frac{R_2}{R_3} b_3. \tag{13.12.4a,b}$$

We also recall that subscript 1 for areas refers to observations 2 and 3; subscript 2 to observations 3 and 1; and subscript 3 to observations 1 and 2. Let us see, then, whether we can determine R_3 from the first and second observations.

Readers who wish to avoid the heavy algebra may proceed direct to Equations 13.12.25 and 13.12.26, which will enable the calculation of the sector-triangle ratios.

Let (r_1, v_1) and (r_2, v_2) be the polar coordinates (i.e. heliocentric distance and true anomaly) in the plane of the orbit of the planet at the instant of the first two observations. In concert with our convention for subscripts involving two observations, let

$$2f_3 = v_2 - v_1. \tag{13.12.5}$$

We have $R_3 = B_3/A_3$. From Equation 13.4.1, which is Kepler's second law, we have, in the units that we are using, in which $GM = 1$, $\dot{B} = \frac{1}{2}\sqrt{l}$ and therefore $B_3 = \frac{1}{2}\sqrt{l}\tau_3$. Also, from the z-component of Equation 13.8.15c, we have $A_3 = \frac{1}{2} r_1 r_2 \sin(v_2 - v_1)$.

Therefore

$$R_3 = \frac{\sqrt{l}\tau_3}{r_1 r_2 \sin(v_2 - v_1)} = \frac{\sqrt{l}\tau_3}{r_1 r_2 \sin 2f_3}. \tag{13.12.6a}$$

In a similar manner, we have

$$R_1 = \frac{\sqrt{l}\tau_1}{r_2 r_3 \sin(v_3 - v_2)} = \frac{\sqrt{l}\tau_1}{r_2 r_3 \sin 2f_1} \tag{13.12.6b}$$

$$R_2 = \frac{\sqrt{l}\tau_2}{r_3 r_1 \sin(v_3 - v_1)} = \frac{\sqrt{l}\tau_2}{r_3 r_1 \sin 2f_2}. \tag{13.12.6c}$$

I would like to eliminate l from here.

I now want to recall some geometrical properties of an ellipse and a property of an elliptic orbit. By glancing at figure II.11, or by multiplying Equations 2.3.15 and 2.3.16, we immediately see that $r\cos v = a(\cos E - e)$, and hence by making use of a trigonometric identity we find

$$r\cos^2 \frac{1}{2} v = a(1-e)\cos^2 \frac{1}{2} E, \tag{13.12.7}$$

and in a similar manner it is easy to show that

$$r\sin^2 \frac{1}{2} v = a(1+e)\sin^2 \frac{1}{2} E. \tag{13.12.8}$$

Here E is the eccentric anomaly.

Also, the mean anomaly at time t is defined as $\frac{2\pi}{P}(t-T)$ and is also equal (via Kepler's Equation) to $E-e\sin E$. The period of the orbit is related to the semi major axis of its orbit by Kepler's third law: $P^2=\frac{4\pi^2}{GM}a^3$. (This material is covered on Chapter 10.) Hence we have (in the units that we are using, in which $GM=1$):

$$E-e\sin E=\frac{t-T}{a^{3/2}}, \tag{13.12.9}$$

where T is the instant of perihelion passage.

Now introduce

$$2f_3=v_2-v_1, \tag{13.12.10}$$

$$2F_3=v_2+v_1, \tag{13.12.11}$$

$$2g_3=E_2-E_1, \tag{13.12.12}$$

$$2G_3=E_2+E_1. \tag{13.12.13}$$

From Equation 13.12.7 I can write

$$\sqrt{r_1r_2}\cos\frac{1}{2}v_1\cos\frac{1}{2}v_2=a(1-e)\cos\frac{1}{2}E_1\cos\frac{1}{2}E_2 \tag{13.12.14}$$

and from Equation 13.12.8 I can write

$$\sqrt{r_1r_2}\sin\frac{1}{2}v_1\sin\frac{1}{2}v_2=a(1+e)\sin\frac{1}{2}E_1\sin\frac{1}{2}E_2. \tag{13.12.15}$$

I now make use of the sum of the sum-and-difference formulas from page 38 of chapter 3, namely $\cos A\cos B=\frac{1}{2}(\cos S+\cos D)$ and $\sin A\sin B=\frac{1}{2}(\cos D-\cos S)$, to obtain

$$\sqrt{r_1r_2}(\cos F_3+\cos f_3)=a(1-e)(\cos G_3+\cos g_3) \tag{13.12.16}$$

and

$$\sqrt{r_1r_2}(\cos f_3-\cos F_3)=a(1+e)(\cos g_3-\cos G_3). \tag{13.12.17}$$

On adding these, we obtain

$$\sqrt{r_1r_2}\cos f_3=a(\cos g_3-e\cos G_3). \tag{13.12.18}$$

I leave it to the reader to derive in a similar manner (also making use of the formula for the semi latus rectum $l=a(1-e^2)$)

$$\sqrt{r_1r_2}\sin f_3=\sqrt{a}\sqrt{l}\sin g \tag{13.12.19}$$

and

$$r_1+r_2=2a(1-e\cos g_3\cos G_3). \tag{13.12.20}$$

We can eliminate $e\cos G$ from Equations 13.12.18 and 13.12.20:

$$r_1+r_2-2\sqrt{r_1r_2}\cos f_3\cos g_3=2a\sin^2 g_3 \tag{13.12.21}$$

Also, if we write Equation 13.12.9 for the first and second observations and take the difference, and then use the formula on page 35 of chapter 3 for the difference between two sines, we obtain

$$2(g_3-e\sin g_3\cos G_3)=\frac{\tau_3}{a^{3/2}}. \tag{13.12.22}$$

Eliminate $e\cos G_3$ from Equations 13.12.18 and 13.12.22:

$$2g_3-\sin 2g_3+\frac{2\sqrt{r_1r_2}}{a}\sin g_3\cos f_3=\frac{\tau_3}{a^{3/2}}. \tag{13.12.23}$$

Also, eliminate l from Equations 13.12.6 and 13.12.19:

$$R_3 = \frac{\tau_3}{2\sqrt{a}\sqrt{r_1 r_2}\cos f_3 \sin g_3}. \tag{13.12.24}$$

We have now eliminated F_3, G_3 and e, and we are left with Equations 13.12.21, 23 and 24, the first two of which I now repeat for easy reference:

$$r_1 + r_2 - 2\sqrt{r_1 r_2}\cos f_3 \cos g_3 = 2a \sin^2 g_3 \tag{13.12.21}$$

$$2g_3 - \sin 2g_3 + \frac{2\sqrt{r_1 r_2}}{a}\sin g_3 \cos f_3 = \frac{\tau_3}{a^{3/2}}. \tag{13.12.23}$$

In these Equations we already know an approximate value for f_3 (we'll see how when we resume our numerical example); the unknowns in these Equations are R_3, a and g_3, and it is R_3 that we are trying to find. Therefore we need to eliminate a and g_3. We can easily obtain a from Equation 13.12.24, and, on substitution in Equations 13.12.21 and 23 we obtain, after some algebra:

$$R_3^2 = \frac{M_3^2}{N_3 - \cos g_3} \tag{13.12.25}$$

and

$$R_3^3 - R_3^2 = \frac{M_3^2 (g_3 - \sin g_3 \cos g_3)}{\sin^3 g_3}, \tag{13.12.26}$$

where

$$M_3 = \frac{\tau_3}{2\left(\sqrt{r_1 r_2}\cos f_3\right)^{3/2}} \tag{13.12.27}$$

and

$$N_3 = \frac{r_1 + r_2}{2\sqrt{r_1 r_2}\cos f_3}. \tag{13.12.28}$$

Similar Equations for R_1 and R_2 can be obtained by cyclic permutation of the subscripts. Equations 13.12.25 and 26 are two simultaneous Equations in R_3 and g_3. Their solution is given as an example in section 1.9 of chapter 1, so we can now assume that we can calculate the sector-triangle ratios.

We can then calculate better triangle ratios from Equations 13.12.4 and return to Equations 13.7.4, 5 and 6 to get better geocentric distances. From Equations 13.7.8 and 9 calculate the heliocentric distances. Make the light-time corrections. (I am not doing this in our numerical example because our original positions were not actual observations, but rather were ephemeris positions.) Then go straight to this section (13.12) again, until you get to here again. Repeat until the geocentric distances do not change.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.13: Resuming the Numerical Example

Let us start with our previous iteration

$$\Delta_1 = 2.65825 \quad r_1 = 3.41952$$
$$\Delta_2 = 2.61558 \quad r_2 = 3.41673$$
$$\Delta_3 = 2.54579 \quad r_3 = 3.41082$$

- or rather with the more precise values that will at this stage presumably be stored in our computer.

These are the values that we had reached when we last left the numerical example.

I promised to say how we know f_3. We defined $2f_3$ as $v_2 - v_1$, and this is the angle between the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Thus

$$\cos 2f_3 = \frac{\xi_1\xi_2 + \eta_1\eta_2 + \zeta_1\zeta_2}{r_1 r_2}. \tag{13.13.1}$$

The heliocentric coordinates can be obtained from Equations 13.5.1, 2 and 3. For example,

$$\xi_1 = l_1 \Delta_1 - \mathfrak{x}_{01}, \tag{13.13.2}$$

and of course

$$r_1 = \sqrt{\xi_1^2 + \eta_1^2 + \zeta_1^2}. \tag{13.13.3}$$

We know how to find the components $(\xi,\ \eta,\ \zeta)$ of the heliocentric radius vector (see Equations 13.7.8 and 9), and so we can now find f_3. I obtain

$$\cos 2f_3 = 0.999\ 929\ 1, \quad \cos f_3 = 0.999\ 982\ 3\ .$$

This means that the true anomaly is advancing at about $0^\circ.68$ in five days. It is interesting to see whether we are on the right track. According to the MPC, Pallas has a period of 4.62 years, which means that, on average, it will move through $1^\circ.067$ in five days. But Pallas has a rather eccentric orbit (according to the MPC, $e = 0.23$). The semi major axis of the orbit must be $P^{2/3} = 2.77$ AU (which agrees with the MPC), and therefore its aphelion distance $a(1+e)$ is about 3.41 AU. Thus Pallas must be close to aphelion in July 2002. By conservation of angular momentum, its angular motion at aphelion must be less than its mean motion by a factor of $(1+e)^2$ so the increase in the true anomaly in five days should be about $1^\circ.067/1.23^2$ or $0^\circ.71$. Thus we do seem to be on the right track.

We can now calculate M_3 and N_3 from Equations 13.12.27 and 28:

$$M_3^2 = 0.000\ 046\ 313\ 0$$
$$N_3 = 1.000\ 018$$

and so we have the following Equations 13.12.25 and 26 for the sector-triangle ratios:

$$R_3^2 = \frac{0.000\ 046\ 313\ 0}{1.000\ 018 - \cos g_3}$$

$$\text{and } R_3^3 - R_3^2 = \frac{0.000\ 046\ 313\ 0(g_3 - \sin g_3 \cos g_3)}{\sin^3 g_3}.$$

Since we discussed how to solve these Equations in section 1.9 of chapter 1, I merely give the solutions here. The one useful hint worth giving is that you can make the first guess for the iteration for g_3 equal to f_3, which we know ($\cos f_3 = .0999\ 9823$), and $R_3 = 1$.

$$\cos g_3 = 0.999\ 972, \quad R_3 = 1.000\ 031$$

We can proceed similarly with R_1 and R_2.

Here is a summary:

subscript	$\cos f$	M^2	N	$\cos g$	R
1	0.999 928 7	$1.859\,91 \times 10^{-4}$	1.000 072	0.999 886	1.000 124
2	0.999 839 9	$4.180\,80 \times 10^{-4}$	1.000 161	0.999 743	1.000 279
3	0.999 982 3	$4.631\,30 \times 10^{-5}$	1.000 018	0.999 972	1.000 031

(13.13.1)

Our new triangle ratios will be

$$a_1 = \frac{R_2}{R_1} b_1 = \frac{1.000\,279}{1.000\,124} \times \frac{2}{3} = 0.666\,770$$

$$\text{and } a_3 = \frac{R_2}{R_3} b_3 = \frac{1.000\,279}{1.000\,031} \times \frac{1}{2} = 0.333\,416.$$

We can now go back to Equations 13.7.4,5 and 6, and calculate the geocentric and heliocentric distances anew. Skip sections 13.8, 13.9 and 13.10, and calculate new sector- triangle ratios and hence new triangle ratios, and repeat until convergence is obtained. After three iterations, I obtained convergence to six significant figures and after seven iterations I obtained convergence to 11 significant figures. The results to six significant figures are as follows:

$$\Delta_1 = 2.65403 \quad r_1 = 3.41539$$
$$\Delta_2 = 2.61144 \quad r_2 = 3.41268$$
$$\Delta_3 = 2.54172 \quad r_3 = 3.40681$$

This is not to be expected to agree exactly with the published MPC values, which are based on all available Pallas observations, whereas we arbitrarily chose three approximate ephemeris positions, but, based on these three positions, we have now broken the back of the problem and have found the geocentric and heliocentric distances.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.14: Summary So Far

1. Gather together the three observations $(t,\ \alpha,\ \delta)$.
2. Convert t from UT to TT. (See Chapter 7.)
3. Calculate or look up and interpolate the solar coordinates.
4. Calculate the geocentric direction cosines of the planet. (Equations 13.5.1-3)
5. Calculate the first approximation to the geocentric distances, using $a_1 = b_1$, $a_3 = b_3$. (Equations 13.7.4-6)
6. Calculate the heliocentric distances. (Equations 13.7.7-8)
7. Improve a_1 and a_3. (Equations 13.8.32-34) Do steps 6 and 7 again.
8. Optional. Calculate $\dot{r}^2$ (Equation 13.10.4) and improve a_1 and a_3 again (Equations 13.10.9-10) and again repeat steps 6 and 7.
9. Make the light travel time corrections for the planet, and go back to step 3! Repeat 6 and 7 but of course with your best current a_1 and a_3.
10. Calculate $f_1,\ f_2,\ f_3$ and the three values of M^2 and N. (Equations 13.13.1, 13.12.27-28) and solve Equations 13.12.25-26 for the sector-triangle ratios. The method of solution of these Equations is given in chapter 1, section 1.9.
11. Calculate new triangle ratios (Equations 13.12.4a,b) – and start all over again!

By this stage we know the geocentric and heliocentric distances, and it is fairly straightforward from this point, at least in the sense that there are no further iterations, and we can just proceed from step to step without having to repeat it all over again. The main problem in computing the angular elements is likely to be in making sure that the angles you obtain (when you calculate inverse trigonometric functions such as arcsin, arccos, arctan) are in the correct quadrant. If your calculator or computer has an $\mathrm{ATAN2}$ facility, make good use of it!

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.15: Calculating the Elementss

We can now immediately calculate the semi latus rectum from Equation 13.12.6a (recalling that $2f_3 = v_2 - v_1$, so that everything except l in the Equation is already known.) In fact we have three opportunities for calculating the semi latus rectum by using each of Equations 13.12.6a,b,c, and this serves as a check on the arithmetic. For our numerical example, I obtain

$$l = 2.61779$$

identically (at least to eleven significant figures) for each of the three permutations.

Now, on referring to Equation 2.3.37, we recall that the polar Equation to an ellipse is

$$r = \frac{l}{1 + e\cos v}. \tag{13.15.1}$$

We therefore have, for the first and third observations,

$$e\cos v_1 = l/r_1 - 1 \tag{13.15.2}$$

and, admitting that $v_3 = v_1 + 2f_2$,

$$e\cos(v_1 + 2f_2) = l/r_3 - 1. \tag{13.15.3}$$

We observe that, in Equations 13.15.2and 13.15.3, the only quantities we do not already know are v_1 and e – so we are just about to find our first orbital element, the eccentricity!

A hint for solving Equations 13.15.2and 3: Expand $cos(v_1 + 2f_2)$. Take $e\sin v$ to the left hand side, and Equation 13.15.3 will become

$$e\sin v_1 = \frac{(l/r_1 - 1).\cos 2f_2 - (l/r_3 - 1)}{\sin 2f_2}. \tag{13.15.4}$$

After this, it is easy to solve Equations 13.15.2 and 13.15.4 for e and for v_1. The other true anomalies are given by $v_2 = v_1 + 2f_3$ and $v_3 = v_1 + 2f_2$. A check on the arithmetic may (and should) be performed by carrying out the same calculation for the first and second observations and for the second and third observations. For all three, I obtained

$$\underline{e = 0.23875}$$

We have our first orbital element!

(The MPC value for the eccentricity for this epoch is 0.22994– but this is based on all available observations, and we cannot expect to get the MPC value from just three hypothetical “observations”.)

The true anomalies at the times of the three observations are

$$v_1 = 191^\circ.99814 \quad v_2 = 192^\circ.68221 \quad v_3 = 194^\circ.05377$$

After that, the semi major axis is easy from Equation 2.3.10, $l = a(1 - e^2)$, for the semi latus rectum of an ellipse. We find

$$\underline{a = 2.77602 \text{ au}}$$

The period in sidereal years is given by $P^2 = a^3$, and is therefore 4.62524 sidereal years. This is not one of the six independent elements, since it is always related to the semi major axis by Kepler’s third law, so it doesn’t merit the extra dignity of being underlined. However, it is certainly worth converting it to mean solar days by multiplying by 365.25636 We find that $P = 1689.39944$days.

The next element to yield will be the time of perihelion passage. We find the eccentric anomalies for each of the three observations from any of Equations 2.3.16, 17a, 17b or 17c. For example:

$$\cos E = \frac{e + \cos v}{1 + e\cos v}. \tag{13.15.5}$$

Then the time of perihelion passage will come from Equations 9.6.4 and 9.6.5:

$$T=t-\frac{P}{2\pi}(E-e\sin E)+nP. \tag{13.15.6}$$

With $n=1$ I make this $\underline{T=t_1+756^{\mathrm{d}}.1319}$

The next step is to calculate the Ps and Qs. These are defined in Equation 10.9.40. They are the direction cosines relating the heliocentric plane-of-orbit basis set to the heliocentric equatorial basis set.

Exercise. Apply Equation 10.9.50 to the first and third observations to show that

$$P_x=\frac{\xi_1 r_3\sin v_3-\xi_3 r_1\sin v_1}{r_1 r_3\sin 2f_2} \tag{13.15.7}$$

and

$$Q_x=\frac{\xi_3 r_1\cos v-\xi_1 r_3\cos v_3}{r_1 r_2\sin 2f_2} \tag{13.15.8}$$

From Equations 10.9.51 and 52, find similar Equations for P_y, Q_y, P_z, Q_z.

The numerical work can and should be checked by calculating these direction cosines also from the first and second, and from the second and third, observations. Check also that $P_x^2+P_y^2+P_z^2=Q_x^2+Q_y^2+Q_z^2=1$. I get

$P_x=-0.48044 \quad P_y=+0.86568 \quad P_z=-0.14059$

$Q_x=-0.87392 \quad Q_y=-0.45907 \quad Q_z=+0.15978$

(Remember that my computer is carrying all significant figures to double precision, though I print out here only a limited number of significant figures. You will not get exactly my numbers unless you, too, carry all significant figures and do not prematurely round off.)

The direction cosines are related to the Eulerian angles, of course, by Equations 10.9.41- 46 (how could you possibly forget?!). All (!) you have to do, then, is to solve these six Equations for the Eulerian angles. (You need six Equations to remove quadrant ambiguity from the angles. Remember the $\mathrm{ATAN2}$ function on your computer – it's an enormous help with quadrants.)

Exercise. Show that (or verify at any rate) that:

$$\sin\omega\sin i=P_z\cos\varepsilon-P_y\sin\varepsilon \tag{13.15.9}$$

and

$$\cos\omega\sin i=Q_z\cos\varepsilon-Q_y\sin\varepsilon. \tag{13.15.10}$$

You can now solve this for the argument of perihelion ω. Don't yet try to solve it for the inclination. (Why not?!) Using $\varepsilon=23^\circ.438\,960$ for the obliquity of the ecliptic of date (calculated from page B18 of the 2002 *Astronomical Almanac*), I get

$$\underline{\omega=304^\circ.81849}$$

Exercise. Show that (or verify at any rate) that:

$$\sin\Omega=(P_y\cos\omega-Q_y\sin\omega)\sec\varepsilon \tag{13.15.11}$$

and

$$\cos\Omega=P_x\cos\omega-Q_x\sin\omega. \tag{13.15.12}$$

From these, I find:

$$\underline{\Omega=172^\circ.64776}$$

One more to go!

Exercise. Show that (or verify at any rate) that:

$$\cos i=-(P_x\sin\omega+Q_x\cos\omega)\csc\Omega. \tag{13.15.13}$$

You can now solve this with Equation 13.15.9or 13.15.10(or both, as a check on the arithmetic) for the inclination. I get

$$\underline{i = 35^\circ.20872}$$

Here they are, all together:

$$\begin{array}{ll} a = 2.77602 \text{ AU} & i = 35^\circ.20872 \\ e = 0.23875 & \Omega = 172^\circ.64776 \\ T = t_1 + 756^{\text{d}}.1319 & \omega = 304^\circ.81849 \end{array} \tag{13.15.1}$$

Have we made any mistakes? Well, presumably after you read chapter 10 you wrote a program to generate an ephemeris. So now, use these elements to see whether they will reproduce the original observations! Incidentally, to construct an ephemeris, there is no need actually to use the elements – you can use the Ps and Qs instead.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.16: Topocentric-Geocentric Correction

In section 13.1 I indicated two small (but not negligible) corrections that needed to be made, namely the ΔT correction (which can be made at the very start of the calculation) and the light-time correction, which can be made as soon as the geocentric distances have been determined – after which it is necessary to recalculate the geocentric distances from the beginning! I did not actually make these corrections in our numerical example, but I indicated how to do them.

There is another small correction that needs to be made. The diameter of Earth subtends an angle of $17''.6$ at 1 au, so the observed position of an asteroid depends appreciably on where it is observed from on Earth's surface. Observations are, of course, reported as *topocentric* – i.e. from the place ($\tau o \pi o \varsigma$) where the observer was situated. They must be corrected by the computer to *geocentric* positions – but of course that can't be done until the distances are known. As soon as the distances are known, the light-time and the topocentric-geocentric corrections can be made. Then, of course, one has to return to the beginning and recompute the distances – possibly more than once until convergence is reached. This section shows how to make the topocentric-geocentric correction.

We have used the notation $\mathfrak{x}, \mathfrak{y}, \mathfrak{z}$ for geocentric coordinates, and I shall use $\mathfrak{x}'$, $\mathfrak{y}'$, $\mathfrak{z}'$ for topocentric coordinates. In figure $\mathrm{XIII}.3$ I show Earth from a point in the equatorial plane, and from above the north pole. The radius of Earth is R, and the radius of a small circle of latitude ϕ (where the observer is situated) is $R\cos\phi$. The $\mathfrak{x}$- and $\mathfrak{x}'-$ axes are directed towards the first point of Aries, Υ.

It should be evident from the figure that the corrections are given by

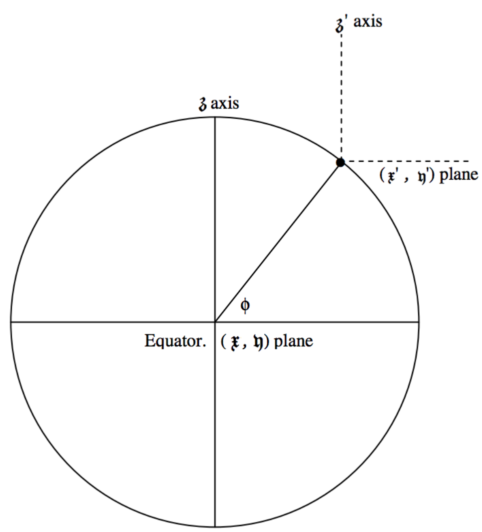
$$\mathfrak{x}' = \mathfrak{x} - R\cos\phi\cos\mathrm{LST}, \tag{13.16.1}$$

$$\mathfrak{y}' = \mathfrak{y} - R\cos\phi\sin\mathrm{LST} \tag{13.16.2}$$

and

$$\mathfrak{z}' = \mathfrak{z} - R\sin\phi. \tag{13.16.3}$$

Any observer who submits observations to the Minor Planet Center is assigned an *Observatory Code*, a three-digit number. This code not only identifies the observer, but, associated with the Observatory Code, the Minor Planet Center keeps a record of the quantities $R\cos\phi$ and $R\sin\phi$ in AU. These quantities, in the notation employed by the MPC, are referred to as $-\Delta_{xy}$ and $-\Delta_z$ respectively. They are unique to each observing site.



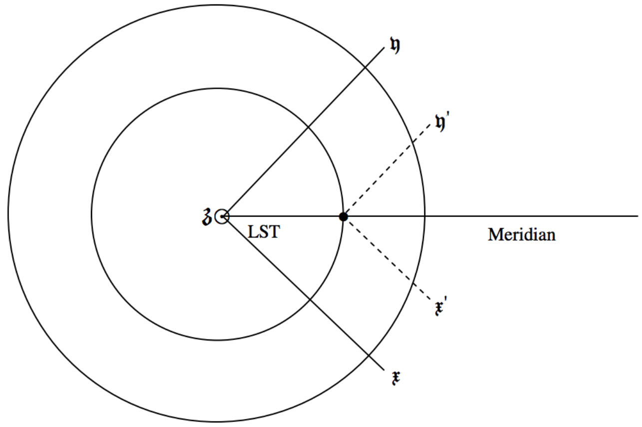


FIGURE XIII.3

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

13.17: Concluding Remarks

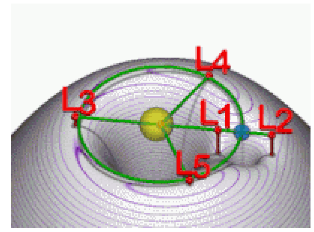
Anyone who has done the considerable work of following this chapter in detail is now capable of determining the elements of an elliptic orbit from three observations, if the orbit is an ellipse and if indeed elliptical elements can be obtained from the observations (which is not always the case). No one arriving at this stage would possibly think of himself or herself as an expert in orbit calculations. There is much, much more to be learned, and much of it will come with experience, and be self-taught or picked up from others. There are questions about how to handle more than the requisite three observations, how to correct the elements differentially as new observations become available, how to apply planetary perturbations, how to handle parabolic or hyperbolic orbits. Some of this material may (or may not!) be discussed in future chapters. However, often the most difficult thing is getting started, and calculating one's very first orbit from the minimum data. It is hoped that this chapter has helped the reader to attain this.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

14: GENERAL PERTURBATION THEORY



14.1: Introduction to General Perturbation Theory

A particle in orbit around a point mass – or a spherically symmetric mass distribution – is moving in a gravitational potential of the form $-GM/r$. In this potential it moves in a keplerian ellipse (or hyperbola if its kinetic energy is large enough) that can be described by the six orbital elements $a, e, i, \Omega, \omega, T$, or any equivalent set of six parameters.

If the potential is a little different from $-GM/r$, say $-(GM/r+R)$, the orbit will be *perturbed*, and R is described as a *perturbation*. As a result it will no longer move in a perfect keplerian ellipse. Perturbations may be *periodic* or *secular*. For example, the elements such as a, e or i may vary in a periodic fashion, while there may be secular changes (i.e. changes that are not periodic but constantly increase or decrease in the same direction) in elements such as Ω and ω. (That is, the line of nodes and the line of apsides may monotonically precess; they may *advance* or *regress*.)

In some situations it may be possible to express the perturbation in terms of a simple algebraic formula. An example would be a particle in orbit around a slightly oblate planet, where it is possible to express the potential algebraically. The aim of this chapter will be to try to find general expressions for the rates of change of the orbital elements in terms of the perturbing function, and we shall use the orbit around an oblate planet as an example.

In other situations it is not easily possible to express the perturbation in terms of a simple algebraic function. For example, a planet in orbit around the Sun is subject not only to the gravitational field of the Sun, but to the perturbations caused by all the other planets in the solar system. These special perturbations have to be treated numerically, and the techniques for doing so will be described in chapter 15.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

14.2: Contact Transformations and General Perturbation Theory

(Before reading this section, it may be well to re-read section 10.11 of Chapter 10.)

Suppose that we have a simple problem in which we know the hamiltonian H_0 and that the Hamilton-Jacobi Equation has been solved:

$$H_0\left(q_1, \frac{\partial S}{\partial q_1}, t\right) + \frac{\partial S}{\partial t} = 0. \tag{14.2.1}$$

Now suppose we have a similar problem, but that the hamiltonian, instead of being just H_0 is $H = H_0 - R$, and $K = H + \frac{\partial S}{\partial t}$.

Let us make a contact transformation from $(p_i,\ q_i)$ to $(P_i,\ Q_i)$, where $\dot{Q}_i = \frac{\partial K}{\partial P_i}$ and $\dot{P}_i = -\frac{\partial K}{\partial Q_i}$. In the orbital context, following Section 10.11, we identify Q_i with α_i and P_i with $-\beta_i$, which are functions (given in Section 10.11) of the orbital elements and which can serve in place of the orbital elements. The parameters are constants with respect to the unperturbed problem, but are variables with respect to the perturbing function. They are given, as functions of time, by the solution of Hamilton's Equations of motion, which retain their form under a contact transformation.

$$\dot{\alpha}_i = \frac{\partial R}{\partial \beta_i} \text{ and } \dot{\beta}_i = -\frac{\partial R}{\partial \alpha_i}. \tag{14.2.2a,b}$$

Perturbation theory will show , then, how the α_i and β_i will vary with a given perturbation. The conventional elements $a,\ e,\ i,\ \Omega,\ \omega,\ T$ are functions of $\alpha_i,\ \beta_i$, and our aim is to find how the conventional elements vary with time under the perturbation R.

We can do that as follows. Let A_i be an orbital element, given by

$$A_i = A_i(\alpha_i, \beta_i). \tag{14.2.3}$$

Then

$$\dot{A}_i = \sum_j \frac{\partial A_i}{\partial \alpha_j}\dot{\alpha}_j + \sum_j \frac{\partial A_i}{\partial \beta_j}\dot{\beta}_j. \tag{14.2.4}$$

By Equations 14.2.2a,b, this becomes

$$\dot{A}_i = \sum_j \frac{\partial A_i}{\partial \alpha_j}\frac{\partial R}{\partial \beta_j} - \sum_j \frac{\partial A_i}{\partial \beta_j}\frac{\partial R}{\partial \beta_j}. \tag{14.2.5}$$

But

$$\frac{\partial R}{\partial \alpha_j} = \sum_k \frac{\partial R}{\partial A_k}\frac{\partial A_k}{\partial \alpha_j} \quad \text{and} \quad \frac{\partial R}{\partial \beta_j} = \sum_k \frac{\partial R}{\partial A_k}\frac{\partial A_k}{\partial \beta_j}. \tag{14.2.6a,b}$$

$$\therefore \quad \dot{A}_i = \sum_j\sum_k \frac{\partial R}{\partial A_k}\left(\frac{\partial A_i}{\partial a_j}\frac{\partial A_k}{\partial \beta_j} - \frac{\partial A_i}{\partial \beta_j}\frac{\partial A_k}{\partial \alpha_j}\right) \tag{14.2.6}$$

That is

$$\dot{A}_i = \sum_k \frac{\partial R}{\partial A_k}\sum_j\left(\frac{\partial A_i}{\partial a_j}\frac{\partial A_k}{\partial \beta_j} - \frac{\partial A_i}{\partial \beta_j}\frac{\partial A_k}{\partial \alpha_j}\right) \tag{14.2.7}$$

This can be written, in shorthand:

$$\dot{A}_i = \sum_k \frac{\partial R}{\partial A_k}\{A_i, A_k\}_{\alpha_j,\beta_j}. \tag{14.2.8}$$

Here the symbol $\{A_i, A_k\}_{\alpha_i,\beta_i}$ is called the *Poisson bracket* of A_i, A_k with respect to α_j, β_j. (In the language of the typographer, the symbols (), [] and {} are, respectively, parentheses, brackets and braces; you may refer to Poisson braces if you wish, but the usual term, in spite of the symbols, is Poisson bracket.)

Note the property $\{A_i, A_k\}_{\alpha_j,\beta_j} = -\{A_k, A_i\}_{\alpha_j,\beta_j}$.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

14.3: The Poisson Brackets for the Orbital Elements

A worked example is in order. From Equations 14.2.7 and 14.2.8, we see that the Poisson brackets are defined by

$$\{A_i, A_k\}_{\alpha_j,\beta_j} = \sum_j \left(\frac{\partial A_i}{\partial \alpha_j}\frac{\partial A_k}{\partial \beta_j} - \frac{\partial A_i}{\partial \beta_j}\frac{\partial A_k}{\partial \alpha_j} \right). \tag{14.3.1}$$

The A_i are the orbital elements.

For our example, we shall calculate $\{\Omega, i\}$ and we write out the sum in full:

$$\{\Omega, i\} = \sum_j \left(\frac{\partial \Omega}{\partial \alpha_j}\frac{\partial i}{\partial \beta_j} - \frac{\partial \Omega}{\partial \beta_j}\frac{\partial i}{\partial \alpha_j} \right) \tag{14.3.2}$$

$$= \frac{\partial \Omega}{\partial \alpha_1}\frac{\partial i}{\partial \beta_1} + \frac{\partial \Omega}{\partial \alpha_2}\frac{\partial i}{\partial \beta_2} + \frac{\partial \Omega}{\partial \alpha_3}\frac{\partial i}{\partial \beta_3} - \frac{\partial \Omega}{\partial \beta_1}\frac{\partial i}{\partial \alpha_1} - \frac{\partial \Omega}{\partial \beta_2}\frac{\partial i}{\partial \alpha_2} - \frac{\partial \Omega}{\partial \beta_3}\frac{\partial i}{\partial \alpha_3}. \tag{14.3.3}$$

Refer now to Equations 10.11.27 and 29, and we find

$$\{\Omega, i\} = 0 + 0 + 0 - 0 + \frac{1}{\alpha_3\sqrt{1-\alpha_2^2/\alpha_3^2}} - 0. \tag{14.3.4}$$

Finally, referring to Equations 10.11.20 and 21, we obtain

$$\{\Omega, i\} = \frac{1}{\sqrt{GMm^2 a(1-e^2)}.\sin i}. \tag{14.3.5}$$

Proceeding in a similar manner for the others, we obtain

$$\{a, T\} = -\frac{2a^2}{GMm}, \tag{14.3.6}$$

$$\{e, T\} = -\frac{a(1-e^2)}{GMme}, \tag{14.3.7}$$

$$\{i, \omega\} = \frac{1}{\sqrt{GMm^2 a(1-e^2)}.\tan i}. \tag{14.3.8}$$

$$\{e, \omega\} = -\frac{\sqrt{1-e^2}}{em\sqrt{GMa}}, \tag{14.3.9}$$

In addition, we have, of course,

$$\{i, \Omega\} = -\{\Omega, i\}, \{T, a\} = -\{a, T\}, \{T, e\} = -\{e, T\} \text{ and } \{\omega, i\} = -\{i, \omega\}. \tag{14.3.10}$$

All other pairs are zero.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

14.4: Lagrange's Planetary Equations

We now go to Equation 14.2.8 to obtain *Lagrange's Planetary Equations*, which will enable us to calculate the rates of change of the orbital elements if we know the form of the perturbing function:

$$\dot{a} = -\frac{2a^2}{GMm}\frac{\partial R}{\partial T}, \tag{14.4.1}$$

$$\dot{e} = -\frac{a(1-e^2)}{GMme}\frac{\partial R}{\partial T}, \tag{14.4.2}$$

$$i = -\frac{1}{\sqrt{GMm^2a(1-e^2)}\sin i}\frac{\partial R}{\partial \Omega} - \frac{1}{me}\sqrt{\frac{1-e^2}{GMa}}\frac{\partial R}{\partial \omega}, \tag{14.4.3}$$

$$\dot{\omega} = \frac{1}{me}\sqrt{\frac{1-e^2}{GMa}}\frac{\partial R}{\partial e} - \frac{1}{\sqrt{GMm^2a(1-e^2)}\tan i}\frac{\partial R}{\partial i}, \tag{14.4.4}$$

$$\dot{\Omega} = \frac{1}{\sqrt{GMm^2(1-e^2)}\sin i}\frac{\partial R}{\partial i}, \tag{14.4.5}$$

$$\dot{T} = \frac{2a^2}{GMm}\frac{\partial R}{\partial a} + \frac{a(1-e^2)}{GMme}\frac{\partial R}{\partial e}. \tag{14.4.6}$$

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

14.5: Motion Around an Oblate Symmetric Top

In Section 5.12, we developed an expression (Equation 5.12.6) for the gravitational potential of an oblate symmetric top (e.g. an oblate spheroid). With a slight change of notation to conform to the present context, we obtain for the perturbing function

$$R=\frac{Gm(C-A)}{2r^3}\left(1-\frac{3z^2}{r^2}\right). \tag{14.5.1}$$

This is the negative of the additional potential *energy* of a mass m at a point whose cylindrical coordinates are $(r,\ z)$ in the vicinity of a symmetric top (which I'll henceforth call an oblate spheroid) whose principal second moments of inertia are C (polar) and A (equatorial). This is correct to order r/a, where a is the equatorial radius of the spheroid.

Let us imagine a particle of mass m in orbit around an oblate spheroid – e.g. an artificial satellite in orbit around Earth. Suppose the orbit is inclined at an angle i to the equator, and the argument of perigee is ω. At some instant, when the cylindrical coordinates of the satellite are $(r,\ z)$, its true anomaly is v.

Exercise 14.5.1: Geometry

Show that $z/r=\sin\,i\ \sin(\omega+v)$.

Having done that, we see that the perturbing function can be written

$$R=\frac{Gm(C-A)}{2r^3}\left(1-3\sin^2 i\sin^2(\omega+v)\right). \tag{14.5.2}$$

Here, r and v vary with time, or what amounts to the same thing, with the mean anomaly $\mathbf{M}$. With a (nontrivial) effort, this can be expanded as a series, including a constant (time independent) term plus periodic terms of the form $\cos\mathbf{M}$, $\cos 2\mathbf{M}$, $\cos 3\mathbf{M}$, etc. If the spirit moves me, I may post the details at a later date, but for the present I give the result that, if the expansion is taken as far as e^2 (i.e. we are assuming that the orbit of the satellite is not strongly eccentric), the constant (time-independent) part of the perturbing function is

$$R=\frac{Gm(C-A)}{2a^3}(1+\frac{3}{2}e^2)(1-\frac{3}{2}\sin^2 i). \tag{14.5.3}$$

Now look at Lagrange's Equations, and you see that the secular parts of $\dot{a}$, $\dot{e}$ and $\dot{i}$ are all zero. That is, although there may be periodic variations (which we have not examined) in these elements, to this order of approximation (e^2) there is no secular change in these elements.

On the other hand, application of Equation 14.4.5 gives for the secular rate of change of the longitude of the nodes

$$\dot{\Omega}=-\frac{3\sqrt{GM}}{2}\frac{(C-A)}{M}\frac{1}{a^{7/2}}(1+2e^2)\cos i. \tag{14.5.4}$$

The reader will no doubt be relieved to note that this expression does not contain m, the mass of the orbiting satellite; M is the mass of the Earth. The reader may also note the minus sign, indicating that the nodes regress. To obtain the factor (1 + 2e^2), readers will have to do a little bit of work, and to expand, by the binomial theorem, whatever expression in e they get, as far as e^2.

Let a be the equatorial radius of Earth. Multiply top and bottom of Equation 14.5.4 by $\mathrm{a}^{7/2}$, and the Equation becomes

$$\dot{\Omega}=-\frac{3}{2}\sqrt{\frac{GM}{\mathrm{a}^3}}\frac{(C-A)}{M\mathrm{a}^2}\left(\frac{\mathrm{a}}{a}\right)^{7/2}(1+2e^2)\cos i. \tag{14.5.5}$$

Here M is the mass of Earth (not of the orbiting satellite), a is the semi major axis of the satellite's orbit, and a is the equatorial radius of Earth.

[If we assume Earth is an oblate spheroid of uniform density, then, according to example 1.iii of Section 2.20 of Chapter 2 of our notes on Classical Mechanics, . $C=\frac{2}{5}M\mathrm{a}^2$. In that case, Equation 14.5.5 becomes

$\dot{\Omega} = -\frac{3}{5}\sqrt{\frac{GM}{\mathrm{a}^3}}\frac{(C-A)}{C}\left(\frac{\mathrm{a}}{a}\right)^{7/2}(1+2e^2)\cos i$. But the density of Earth is not uniform, so we'll leave Equation 14.5.5 as it is.]

For a nearly circular orbit, Equation 14.5.5 becomes just

$$\dot{\Omega} = -\frac{3}{2}\sqrt{\frac{GM}{\mathrm{a}^3}}\frac{(C-A)}{Ma^2}\left(\frac{\mathrm{a}}{a}\right)^{7/2}\cos i. \tag{14.5.6}$$

This tells us that the line of nodes of a satellite in orbit around an oblate planet (i.e $C > A$) *regresses*. From the rate of regression of the line of nodes, we can deduce the difference, $C - A$ between the principal moments of inertia, though we cannot deduce either moment separately. (If we *could* determine the moment of inertia from the rate of regression of the nodes – which we cannot – how well can we determine the density distribution inside Earth? See Problem 14 in Chapter A of our Classical Mechanics notes to determine the answer to this. It will be found that knowledge of the moment of inertia places only weak constraints on the core size and density.)

Numerically it is known for Earth that the quantity $\frac{3}{2}\sqrt{\frac{GM}{\mathrm{a}^3}}\frac{(C-A)}{M\mathrm{a}^2}$ is about 2.04 rad s^{-1}, or about 10.1 degrees per day. Thus the rate of regression of the nodes of a satellite in orbit around Earth in a near-circular orbit is about

$$\dot{\Omega} = -10.1\left(\frac{\mathrm{a}}{a}\right)^{7/2}\cos i \quad \text{degrees per day.}$$

We can refer to Equations 14.4.4 and 14.5.3 to determine the rate of motion of the line of apsides, $\dot{\omega}$. After some algebra, and neglect of terms of order e^2 and higher, we find

$$\dot{\omega} = \frac{3(C-A)}{4a^{7/2}}\sqrt{\frac{G}{M}}(5\cos^2 i - 1). \tag{14.5.7}$$

or, if we multiply top and bottom by $\mathrm{a}^{7/2}$,

$$\dot{\omega} = \frac{3(C-A)}{4M\mathrm{a}^2}\sqrt{\frac{GM}{\mathrm{a}^3}}\left(\frac{\mathrm{a}}{a}\right)^{7/2}(5\cos^2 i - 1). \tag{14.5.8}$$

Thus we find that the line of apsides advances if the inclination of the orbit to the equator is less than 63° and it regresses if the inclination is greater than this.

In this section, I have demanded a fair amount of work from the reader – in particular for the expansion of Equation 14.5.2. While the work requires some patience and persistence, it is straightforward, and the resolute reader will be able to work out the expansion in terms of the mean anomaly and the time, and hence, by making use of Lagrange's planetary Equations, will be able to predict the periodic variations in a, e and i. For the time being, I am not going to do this, since no new principles are involved, the aim of the chapter being to give the reader a start on how to start to calculate the changes in the orbital elements if one can express the perturbing function analytically.

For the effect of the perturbation of a planetary orbit by the presence of other planets, we have to solve the problem numerically by the techniques of *special perturbations*, which, I hope, some time in the future, may be the subject of an additional chapter.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

15: SPECIAL PERTURBATIONS

[This chapter is under development and it may be a rather long time before it is complete. It is the intention that it may deal with special perturbations, differential corrections, and the computation of a definitive orbit. However, it will probably proceed rather slowly and whenever the spirit moves me.]

15.1: Introduction

Chapter 14 dealt with the subject of *general perturbations*. That is, if the perturbation R can be expressed as an explicit algebraic function, the rates of change of the orbital elements with time can be calculated by explicit algebraic expressions known as *Lagrange's Planetary Equations*. By way of example we derived Lagrange's Equations for the case of a satellite in orbit around an oblate planet, in which the departure of the gravitational potential from that of a spherically symmetric planet could be expressed in simple algebraic form.

Lagrange's Equations are important and interesting from a theoretical point of view. However, in the practical matter of calculating the perturbations of the orbit of an asteroid or a comet resulting from the gravitational field of the other planets in the solar system, that is not how it is done. The perturbing forces are functions of time which must be computed numerically rather than from a simple formula. Such perturbations are generally referred to as *special perturbations*. While long-established computer programs, such as RADAU15, may be available to carry out the necessary rather long computations without the user having to understand the details, it is the intention in this chapter to indicate in principle how such a program may be developed from scratch.

Jupiter is by far the greatest perturber, but for high-precision work it may be necessary to include perturbations from the other major planets, Mercury to Neptune. Pluto may also be considered. However, it is now known that Pluto is a good deal less massive than it was once estimated to be, so it is a nice question as to whether or not to include Pluto. Besides, Pluto is probably not the most massive of the transneptunian objects - Eris is believed to be a little larger and hence possibly more massive. The main belt object Ceres may be more important than either of these. The total mass of the remaining asteroids is usually considered negligible in this context.

It will be evident that any computer program intended to compute special perturbations will have to include, as subroutines, programs for calculating, day-by-day, the positions and distances of each of the perturbing planets to be included in the computation. Computer programs are available to provide these. In what follows, it will be assumed that the reader has access to such a program (I do!) or is otherwise able to compute the planetary positions, and we move on from there to see how we calculate the planetary perturbations.

15.2: Orbital elements and the position and velocity vector

The six elements used to describe the orbit of an asteroid are the familiar

$$a, e, i, \Omega, \omega, T$$

Because of the precession and nutation of Earth, the angular elements must, of course, be referred to a particular equinox and equator, usually chosen to be that of the standard epoch J2000.0, which means 12h 00m TT on 2000 January 01. (The "J" stands for "Julian Year".)

The element T is the instant of perihelion passage. If the orbit is nearly circular, the instant of perihelion passage is ill-defined, and if the orbit is exactly circular, it is not defined at all. In such cases, instead of T, we may give either the *mean anomaly* M_0 or the *mean longitude* L_0 at a specified epoch (see Chapter 10). This epoch need not be (and usually is not) the same as the standard epoch referred to in the previous paragraph.

Suppose that, at some instant of time (to be known, for reasons to be explained later, as the epoch of osculation), the heliocentric ecliptic coordinates of an asteroid or comet in an elliptic orbit are (X, Y, Z) and the components of the velocity vector are ($\dot{X}, \dot{Y}, \dot{Z}$). We have shown in Chapter 10, Section 10.10) how to calculate, from these, the six elements $a, e, i, \Omega, \omega, T$ of the orbit at that instant. Conversely, given the orbital elements, we could reverse the calculation and calculate the components of the position and velocity vectors. Thus an orbit may equally well be described by the six numbers

$$X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}$$

That is to say the components, at some specified instant of time, of the position and velocity vector in heliocentric ecliptic coordinates.

We could equally well give the components, at some instant of time, of the position and velocity vectors in heliocentric equatorial coordinates:

$$\xi, \eta, \zeta, \dot{\xi}, \dot{\eta}, \dot{\zeta}$$

We saw in Section 10.9 that yet another set of six numbers,

$$P_x, Q_x, P_y, Q_y, P_z, Q_z$$

will also suffice to describe an orbit.

It is assumed here that the reader is familiar with all four of these alternative sets of elements, and can convert between them. Indeed, before reading on, it may be a useful exercise to prepare a computer program that will convert instantly between them. This may not be a trivial task, but I strongly recommend doing so before reading further. The facility to convert instantly between one set and another is an enormous help. To convert between ecliptic and equatorial coordinates, you will need, of course, the obliquity of the ecliptic at that instant - it varies, of course, with time.) The reader will have noticed the frequent occurrence of the phrase "at that instant" in the previous paragraphs. If the asteroid were not subject to perturbations from the other planets, it would retain its orbital elements forever. However, because of the planetary perturbations, the elements a, e i,, Ω,ω,T computed from $X, Y, Z, \dot{X}, \dot{Y}, \dot{Z}$ or from $\xi, \eta, \zeta, \dot{\xi}, \dot{\eta}, \dot{\zeta}$ at a particular instant of time are valid only for that instant. The elements will change with time. Therefore in quoting the elements of an asteroidal orbit, it is entirely necessary to state clearly and without ambiguity the instant of time to which these elements are referred. The unperturbed orbit, and the real perturbed orbit, will coincide in position and velocity at that instant. The real and unperturbed orbits will "kiss" or osculate at that instant, which is therefore known as the epoch of osculation.

The elements $a, e, i, \Omega, \omega, T$ calculated for a particular epoch of osculation may suffice for the computation of an ephemeris for weeks to come. But after months the observed position of the object will start to deviate from its calculated ephemeris position. It is then necessary to calculate a new set of elements for a later epoch of osculation. Depending on circumstances, orbital elements may be recalculated every year, or every 200 days or every 40 days or every 10 days, or at some other convenient interval. It will be the purpose in what follows to do the following. Given that at some instant (i.e. at some epoch of osculation) the elements are $a, e, i, \Omega, \omega, T$ (or the position and velocity vectors are $\xi, \eta, \zeta, \dot{\xi}, \dot{\eta}, \dot{\zeta}$, how do we calculate the elements at some subsequent epoch, taking into account planetary perturbations?

As pointed out at the end of Section 15.1, we shall need to know the positions and distances of the major planets as a function of time. We suppose that we have subroutines in our program that we can call upon to calculate these data at any date. As mentioned above, the Equations of motion can be written in equatorial or ecliptic coordinates, though it is more likely that, for the positions of the major planets, we shall have available their positions in *equatorial coordinates*.

15.3: The equations of motion

First let us consider the motion of an asteroid under the gravitational influence of the Sun alone, ignoring perturbations from the other planets. We take the mass of the Sun to be M and the mass of the asteroid to be m. The force on the asteroid − and, of course, by Newton's third law, the force on the Sun − is $\frac{GMm}{r^2}$, where r is the distance between the two bodies. The two bodies are, of course, in motion around their common centre of mass, which, in the case of an asteroid, is very close to the centre of the Sun.

The acceleration of the asteroid towards the centre of mass is $\frac{GM}{r^2}$, and the acceleration of the Sun towards the centre of mass is $\frac{Gm}{r^2}$. If we refer the motion to the Sun as origin, we see that the acceleration of the asteroid towards the Sun is $\frac{G(M+m)}{r^2}$. In vector form we may write this as

$$\ddot{\mathbf{r}} = -\frac{G(M+m)}{r^3}\mathbf{r}, \tag{15.3.1}$$

where $\mathbf{r}$ is a vector directed from the Sun towards the asteroid, with heliocentric rectangular components (x, y, z). These heliocentric coordinates could be either ecliptic coordinates, for which we have hitherto used the symbols (X, Y, Z); or they could be equatorial coordinates, for which we have hitherto used the symbols (ξ, η, ζ). The symbols (x, y, z) will be understood here to refer to either, at our convenience. It is more likely that we shall have available the *equatorial* rather than the ecliptic coordinates. The direction cosines of r are $\left(\frac{x}{r}, \frac{y}{r}, \frac{z}{r}\right)$, and consequently the rectangular components of Equation 15.3.1 are

$$\ddot{x} = -\frac{G(M+m)}{r^3}x \tag{15.3.2}$$

$$\ddot{y} = -\frac{G(M+m)}{r^3}y \tag{15.3.3}$$

$$\ddot{z} = -\frac{G(M+m)}{r^3}z \tag{15.3.4}$$

These are the Equations of motion of the asteroid with respect to the Sun as origin. The quantities $x, y, z, r\left(=\sqrt{x^2+y^2+z^2}\right)$ are, of course, functions of time. The solution of these Equations describe the elliptical (or other conic section) orbits of the asteroid and all the other properties that we have discussed in previous chapters.

If we are using ecliptic coordinates (X, Y, Z), the X-axis is directed towards the First Point of Aries, the Y-axis is directed along the direction of increasing ecliptic longitude, and the Z-axis is directed towards the north pole of the ecliptic.

If we are using equatorial coordinates (ξ, η, ζ), the ξ-axis is directed towards the First Point of Aries, the η-axis is directed along the direction of 6 hours right ascension, and the ζ-axis is directed towards the north celestial pole. The Earth will be on the X- or ξaxis in September (not March).

Now let us introduce a third body, a perturbing planet, such as, perhaps, Jupiter. We'll suppose that its mass is m_1, that its distance from the Sun is r_1 and its distance from the asteroid is ρ_1 (see figure XV.I, in which S is the Sun, A is the asteroid, and P is the perturbing planet). This is now a three-body problem and a general solution in terms of algebraic functions is not possible, and it has to be solved by numerical computation.

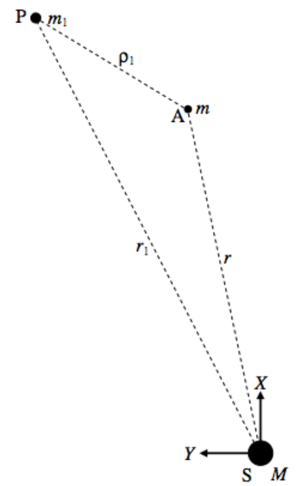


FIGURE XV.I

In addition to the accelerations of the asteroid towards the Sun and the Sun towards the asteroid described on page 3, used in developing Equations 15.3.1-4, we now have also to consider the accelerations of the asteroid and the Sun towards the perturbing planet, as indicated in figure XV.II.

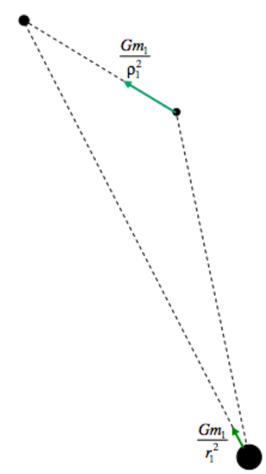


FIGURE XV.II

The x-components of these are $\frac{Gm_1}{\rho_1^2} \times \frac{x_1 - x}{\rho_1}$ and $\frac{Gm_1}{r_1^2} \times \frac{x_1}{r_1}$, and so the additional acceleration of A, relative to the Sun, in the X-direction is $Gm_1\left(\frac{x_1-x}{\rho_1^3} - \frac{x_1}{r_1^3}\right)$, and this has now to be added to the right hand side of Equation 15.3.2:

$$\ddot{x} = -\frac{G(M+m)}{r^3}x + Gm_1\left(\frac{x_1 - x}{\rho_1^3} - \frac{x_1}{r_1^3}\right) \tag{15.3.5}$$

Neither G nor M are known to great precision, but the product GM is known to very great precision. Indeed in computational practice we make use of the *Gaussian constant* $k = \sqrt{\frac{GM}{a_0}}$, where a_0 is the astronomical unit of length. This constant has dimension T^{-1} and is equal to the angular velocity of a particle of negligible mass in circular orbit of radius 1 au around the Sun, which is 0.017 202 098 95 radians per mean solar day. Therefore in computational practice, Equation 15.3.5 is generally written as

$$\ddot{x} = -\frac{k^2(1+m)}{r^3}x + k^2 m_1\left(\frac{x_1 - x}{\rho_1^3} - \frac{x_1}{r_1^3}\right), \tag{15.3.6}$$

in which the units of mass, length and time are, respectively, solar mass, astronomical unit, and mean solar day. Recall that *m* is the mass of the asteroid whose orbit we are computing, and m_1 is the mass of the perturbing planet, and that the origin of coordinates is the centre of the Sun. Similar Equations apply to the *y*- and *z*-components:

$$\ddot{y}=-\frac{k^2(1+m)}{r^3}y+k^2m_1\left(\frac{y_1-y}{\rho_1^3}-\frac{y_1}{r_1^3}\right) \tag{15.3.7}$$

$$\ddot{z}=-\frac{k^2(1+m)}{r^3}z+k^2m_1\left(\frac{z_1-z}{\rho_1^3}-\frac{z_1}{r_1^3}\right) \tag{15.3.8}$$

If we add the perturbations from all the major planets from Mercury (M) to Neptune (N), these Equations become, of course,

$$\ddot{x}=-\frac{k^2(1+m)}{r^3}x+k^2\sum_{i=M}^{N}m_i\left(\frac{x_i-x}{\rho_i^3}-\frac{x_i}{r_i^3}\right) \tag{15.3.9}$$

and similar Equations in *y* and *z*.

In the case of an asteroid or a comet, it may be permissible to neglect *m* in this Equation (i.e. set *m* = 0), but not, of course, m_1. We shall do that here, so the Equation of motion in *x* becomes

$$\ddot{x}=-k^2\frac{x}{r^3}+k^2\sum_{i=\mathrm{M}}^{\mathrm{N}}m_i\left(\frac{x_i-x}{\rho_i^3}-\frac{x_i}{r_i^3}\right), \tag{15.3.10}$$

with similar Equations in y and z.

The *x*, x_i, ρ_i, r_i, etc., are numerical data, which have to be supplied by independent computations (subroutines) for all the planets. As stated at the end of the previous Section, we suppose that we have subroutines in our program that we can call upon to calculate these data at any date. We also pointed out that the Equations of motion are valid for either ecliptic or equatorial coordinates, although the coordinates of the planets are more likely to be available is *equatorial* rather than ecliptic coordinates. They are all functions of time, so that, in effect, we have to develop numerical methods for integrating Equations of the form, where *f(t)* is not an algebraic expression, but rather a table of numerical values.

$$\ddot{x}=f(t) \tag{15.3.11}$$

That is to say

$$\frac{d\dot{x}}{dt}=f(t). \tag{15.3.12}$$

We suppose that we know x& at the epoch of osculation. Then we can find $\dot{x}$ at any subsequent date by any standard technique of numerical integration, such as Simpson's or Weddle's Rules, or Gaussian quadrature, or by a Runge-Kutta process. Thus we now have a table of $\dot{x}$ as a function of time:

$$\dot{x}=g(t) \tag{15.3.13}$$

That is to say

$$\frac{dx}{dt}=g(t) \tag{15.3.14}$$

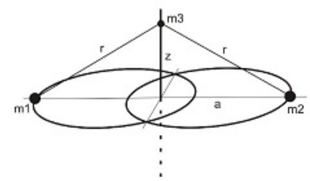
We integrate a second time, until we arrive at both x and $\dot{x}$ at some subsequent epoch of osculation (perhaps 200, or 40, days into the future). Repeat with the y and z components, so we eventually have a new set of $(x, y, z, \dot{x}, \dot{y}, \dot{z})$ for a later epoch, and hence also of *a*, *e*, *i*, Ω, ω, *T*.

CHAPTER OVERVIEW

16: EQUIVALENT POTENTIAL AND THE RESTRICTED THREE-BODY PROBLEM

16.1: INTRODUCTION

The collinear lagrangian points any points on the line passing through the two masses where a third body of negligible mass could orbit around C with the same period as the other two masses; i.e. it would remain on the line joining the two main masses? The collinear points were discussed by Euler before Lagrange, but Lagrange took the problem further and discovered an additional two points not collinear with the masses, and the five points today are generally all known as the lagrangian points.



16.2: MOTION UNDER A CENTRAL FORCE

There is no general analytical solution in terms of simple algebraic functions for the problem of three gravitating bodies of comparable masses. Except in a few very specific cases the problem has to be solved numerically. However in the restricted three-body problem, we imagine that there are two bodies of comparable masses revolving around their common center of mass, and a third body of negligible mass moves in the field of the other two.

16.1: Introduction

We are going to consider the following problem. Two masses, M_1 and M_2 are revolving around their mutual centre of mass C in circular orbits, at a constant distance a apart.

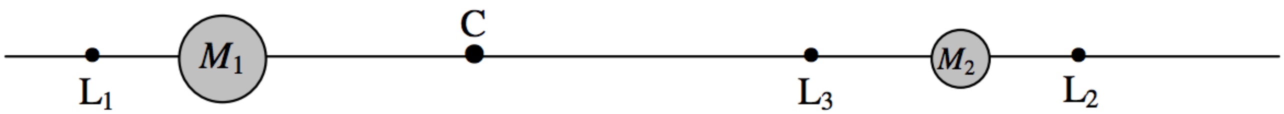


FIGURE IV.4

The orbital period is given by

$$P^2 = \frac{4\pi^2 a^3}{G(M_1 + M_2)} \tag{16.1.1}$$

and the angular orbital speed is given by

$$\omega^2 = \frac{G(M_1 M_2)}{a^3}. \tag{16.1.2}$$

I establish the following notation.

Mass ratio:

$$\frac{M_1}{M_2} = q. \tag{16.1.3}$$

Mass fraction:

$$\frac{M_1}{M_1 + M_2} = \mu. \tag{16.1.4}$$

They are related by

$$q = \frac{\mu}{1 - \mu} \tag{16.1.5}$$

and

$$\mu = \frac{q}{1 + q}. \tag{16.1.6}$$

We note the following distances:

$$\mathrm{M_1C} = (1 - \mu)a, \quad \mathrm{M_2C} = \mu a. \tag{16.1.7}$$

We ask ourselves the following question: Are there any points on the line passing through the two masses where a third body of negligible mass could orbit around C with the same period as the other two masses; i.e. it would remain on the line joining the two main masses?

In fact there are three such points, and they are known as the *collinear lagrangian points*. (The collinear points were discussed by Euler before Lagrange, but Lagrange took the problem further and discovered an additional two points not collinear with the masses, and the five points today are generally all known as the lagrangian points. We shall discuss the additional points in Section 16.2.) I have marked the three points in figure XVI.4 with the letters $\mathrm{L_1}$, $\mathrm{L_2}$ and $\mathrm{L_3}$.

Nomenclature

There are evidently $3! = 6$ ways in which I could choose the subscripts. Often today, the inner lagrangian point is labelled $\mathrm{L_1}$ and the outer points are labelled $\mathrm{L_2}$ and $\mathrm{L_3}$. This seems to me to lack logic, and I choose to label the inner point $\mathrm{L_3}$, and the outer points associated with M_1 and M_2 are then L_1 and L_2 respectively. Incidentally, I am not making any assumption about which of the two main bodies is the more massive.

Let us deal first with L_1. Let us suppose that the distance from C to L_1 is xa.

A particle of mass m at L_1 is subject (in a co-rotating reference frame) to three forces, namely the gravitational attractions from the two main bodies, and the centrifugal force acting away from C. If this body is to be in equilibrium, we must have

$$\frac{GM_1m}{[(x-1+\mu)a]^2}+\frac{GM_2m}{[(x+\mu)a]^2}=mxa\omega^2. \tag{16.1.8}$$

On making use of Equations 16.1.2 and 16.1.4, we find that this Equation becomes

$$\frac{\mu}{(x-1+\mu)^2}+\frac{1-\mu}{(x+\mu)^2}=x. \tag{16.1.9}$$

After manipulation, this becomes

$$a_0+a_1x+a_2x^2+a_3x^3+a_4x^4+x^5=0, \tag{16.1.10}$$

where

$$a_0=-1+3\mu-3\mu^2, \tag{16.1.11}$$

$$a_1=2-4\mu+\mu^2-2\mu^3+\mu^4, \tag{16.1.12}$$

$$a_2=-1+2\mu-6\mu^2+4\mu^3, \tag{16.1.13}$$

$$a_3=1-6\mu+6\mu^2 \tag{16.1.14}$$

and

$$a_4=-2+4\mu. \tag{16.1.15}$$

Although Equation 16.1.10is a quintic Equation, it has just one real root for positive μ.

The positions of L_2 and L_3 can be found by exactly similar arguments – you just have to take care with the directions and distances of the two gravitational forces.

For L_2, the coefficients are the same as for L_1, except

$$a_1=-2+4\mu+\mu^2-2\mu^3+\mu^4, \tag{16.1.6}$$

$$a_2=-1-2\mu+6\mu^2-4\mu^3 \tag{16.1.17}$$

and

$$a_4=2-4\mu. \tag{16.1.18}$$

For L_3 the coefficients are

$$a_0=1-3\mu+3\mu^2-2\mu^3, \tag{16.1.1}$$

$$a_1=2-4\mu+5\mu^2-2\mu^3+\mu^4, \tag{16.1.2}$$

$$a_2=1-4\mu+6\mu^2-4\mu^3, \tag{16.1.3}$$

$$a_3=1-6\mu+6\mu^2 \tag{16.1.4}$$

$$a_4=2-4\mu. \tag{16.1.5}$$

(Reminder: When computing any of these polynomials, write them in terms of nested parentheses. See Chapter 1, Section 1.5.)

It is also of interest to see the equivalent potential (gravitational plus centrifugal). The expression for gravitational potential energy is, as usual, $-GMm/r$, where r is the distance from the mass M. The expression for the centrifugal potential energy is

$-\frac{1}{2}m\omega^2 r^2$, where r is the distance from the centre of mass. The negative of the derivative of this expression is $m\omega^2 r$ which is the usual expression for the centrifugal force. When we apply these principles to the system of two masses under consideration, we obtain the following expression for the equivalent potential (which, in this section, I'll just call V rather than V').

$$V = -\frac{GM_1}{|x+1-\mu|a} - \frac{GM_2}{|x-\mu|a} - \frac{1}{2}x^2a^2\omega^2. \tag{16.1.24}$$

On making use of Equations 16.1.2 and 16.1.4, we find that this Equation becomes

$$W = -\frac{\mu}{|x+1-\mu|} - \frac{1-\mu}{|x-\mu|} - \frac{x^2}{2}, \tag{16.1.25}$$

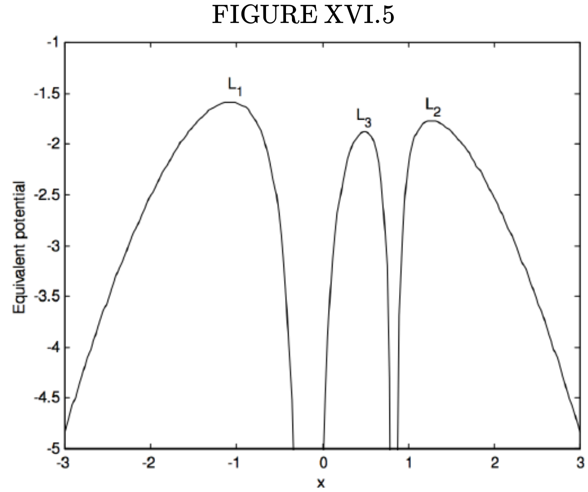
where

$$W = V \div \left(\frac{G(M_1+M_2)}{a}\right). \tag{16.1.26}$$

Setting the derivatives of this expression to zero gives, of course, the positions of the lagrangian points, for these are equilibrium points where the derivative of the potential is zero. Figure XVI.5 shows the potential for a mass ratio $q = 5$. Note that, in the line joining the two masses, the equivalent potential at the lagrangian points is a maximum, and therefore these points, while equilibrium points, are unstable. We shall see in Section 16.6 that the points are actually saddle points. While several spacecraft are in orbit or are planned to be in orbit around the collinear lagrangian points (e.g. SOHO at the interior lagrangian point, and MAP at L_2), continued small expenditure of fuel is presumably needed to keep them there.

It will be of interest to see how the positions of the lagrangian points vary with mass fraction. Indeed mass can be transferred from one member of a binary star system to the other during the evolution of a binary star system. We shall discuss a little later how this can happen. For the time being, without worrying about the exact mechanism, we'll just vary the mass fraction and see how the positions of the lagrangian points vary as we do so. However, if mass is transferred from one member of a binary star system to the other,

FIGURE XVI.5



and if there are no external torques on the system, the angular momentum L of the system will be conserved, and, to ensure this, the separation a of the two stars changes with mass fraction.

Example 16.1.1

Show that, for a given orbital angular momentum L of the system, the separation a of the components varies with mass fraction according to

$$a = \frac{L^2}{GM^3\mu^2(1-\mu)^2}. \tag{16.1.27}$$

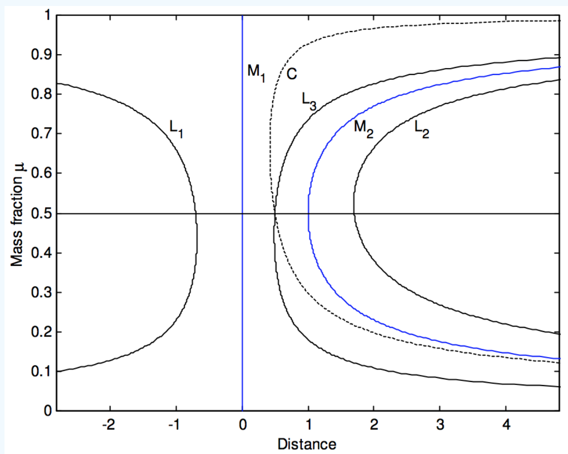
Solution

Here $M = M_1 + M_2$ is the total mass of the system. In figure XVI.6 I have used this Equation, plus Equations 16.1.10 and 16.1.7, to compute the distances of M_2, C, and the three lagrangian points from M_1 as a function of mass fraction. The unit of distance in figure XVI.6 is $16L^2/(GM^3)$, which is the separation of the two masses when the two masses are equal. Each of these distances has a minimum value for a particular mass fraction. These minimum distances, and the mass fractions for which they occur, are as follows:

$$\begin{array}{lcc} & \text{Least value} & \text{Mass fraction} \\ \mathrm{M_1C} & 0.421875^* & 0.66666\dot{6} \\ \mathrm{M_1L_2} & 1.690392 & 0.524579 \\ \mathrm{M_1M_2} & 1.000000 & 0.500000 \\ \mathrm{M_1L_3} & 0.489038 & 0.446273 \\ \mathrm{M_1L_1} & 0.677756 & 0.436062 \end{array} \tag{16.1.6}$$

$^*0.421875 = 27/64$ exactly

FIGURE XVI.6



How can mass transfer actually occur in a binary star system? Well, stars are not points – they are large spherical bodies. When the hydrogen is exhausted in the core by thermonuclear reactions, a star expands hugely ("leaves the main sequence") and when it expands so much that the outer layers of its atmosphere reach the inner lagrangian point, matter from the large star spills over into the other star. The more massive of the two stars in a binary system generally evolves faster; it is the first to leave the main sequence and to expand so that its atmosphere reaches the inner lagrangian points. One can imagine the more massive star gradually filling up its potential well of figure XVI.5, until it overflows and drips over the potential hill of the inner point, and then falls into the potential well of its companion.

One way of interpreting figure XVI.6 is to imagine that M_1 starts with a large mass fraction close to 1, and therefore near the top of figure XVI.6. Now imagine that this star loses mass to its companion, so that the mass fraction decreases. We start moving down the M_1 line of figure XVI.6. We see the inner point L_3 coming closer and closer. If the surface of the star meets L_3 while L_3 is still approaching (i.e. if the mass fraction is still greater than 0.446273), then further mass transfer will make L_3 approach ever faster, and mass transfer will therefore be rapid. When the mass fraction is less than 0.5, the star that was originally the more massive star is by now less massive than its companion. When the mass fraction has been reduced below 0.446273, further mass transfer will push L_3 *away*, and therefore further mass transfer will be slow.

In the calculations of Example 16.1.1, I assumed that the stars can be treated gravitationally as if they are point sources – and so they can be, however large they are, as long as they are spherically symmetric. By the onset of mass transfer, the mass-losing star is quite distorted and is far from spherical. However, this distortion affects mostly the outer atmosphere of the star, and, provided that the greater bulk of the star is contained within a roughly spherically-symmetric volume, the point source approximation should continue to be good. The other assumption I made was that orbital angular momentum is conserved.

There are two reasons why this might not be so – but for both of them there is likely to be very little loss of orbital angular momentum. One possibility is that mass might be lost from the system – through one or other or both of the external collinear lagrangian points. However, figure XVI.5 shows that the potentials of these points are appreciably higher than the internal point; therefore mass transfer takes place well before mass loss. Another reason why orbital angular momentum might be conserved is as follows. When matter from the mass-losing star is transferred through the inner point to the mass-gaining star, or flows over the inner potential hill, it does not move in a straight line directly towards the second star. This entire analysis has been referred to a corotating reference frame, and when matter moves from M_1 towards M_2, it is subject to a *Coriolis force* (see section 4.9 of Classical Mechanics), which sends it around M_2 in an *accretion disc*. During this process the total angular momentum of the system is conserved (provided no mass is lost from the system) but this must now be shared between the orbital angular momentum of the two stars and the angular momentum of the accretion disc. However, as long as the latter is a relatively small contribution to the total angular momentum, conservation of orbital angular momentum remains a realistic approximation.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

16.2: Motion Under a Central Force

There is no general analytical solution in terms of simple algebraic functions for the problem of three gravitating bodies of comparable masses. Except in a few very specific cases the problem has to be solved numerically. However in the *restricted* three-body problem, we imagine that there are two bodies of comparable masses revolving around their common centre of mass C, and a third body of negligible mass moves in the field of the other two. We considered this problem partially in Section 16.1, except that we restricted our interest yet further in confining our attention to the line joining to two principal masses. In this section we shall widen our attention. One question that we asked in section 16.1 was: Are there any points where a third body of negligible mass could orbit around C with the same period as the other two masses? We found three such points, the collinear lagrangian points, on the line joining the two principal masses. In this section we shall discover two additional points, the fourth and fifth lagrangian points. They are not collinear with M_1 and M_2, but are such that the three masses are at the corners of an equilateral triangle.

We shall work in a co-rotating reference frame in which there are two deep hyperbolic potential wells of the form $-GM_1/r_1$ and $-GM_2/r_2$ from the gravitational field of the two principal masses sunk into the nose-up paraboloidal potential of the form $-\frac{1}{2}\rho^2\omega^2$, whose negative derivative is the centrifugal force per unit mass. Here ρ is the usual cylindrical coordinate, and $\omega^2 = G(M_1+M_2)/a^3$.

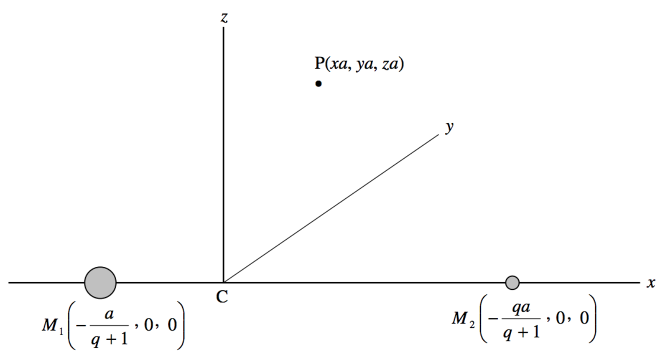


FIGURE XVI.7

In figure XVI.7 we see a coordinate system which is rotating about the z-axis, in such a manner that the two principal masses remain on the x-axis, and the origin of coordinates is the centre of mass C. The mass ratio $M_1/M_2 = q$, so the coordinates of the two masses are as shown in the figure. The constant distance between the two masses is a. P is a point whose coordinates are $(xa,\ ya,\ za)$, x, y and z being dimensionless. The gravitational-plus-centrifugal effective potential V at P is

$$V = -\frac{GM_1}{a\left[\left(x+\frac{1}{q+1}\right)^2+y^2+z^2\right]^{1/2}} - \frac{GM_2}{a\left[\left(x-\frac{q}{q+1}\right)^2+y^2+z^2\right]^{1/2}} - \frac{G(M_1+M_2)(x^2+y^2)}{2a}. \tag{16.2.1}$$

Let $W = \frac{Va}{G(M_1+M_2)}$ (dimensionless). Then

$$W = -\frac{q}{\left[(1+x(q+1))^2+(y^2+z^2)(q+1)^2\right]^{1/2}} - \frac{1}{\left[(q-x(q+1))^2+(y^2+z^2)(q+1)^2\right]^{1/2}} - \frac{x^2+y^2}{2}. \tag{16.2.2}$$

I shall write this for short:

$$W = -Aq - B - \frac{1}{2}(x^2+y^2). \tag{16.2.3}$$

Here A and B are functions with obvious meaning from comparison with Equation 16.2.2.

We are going to need the first and second derivatives, so I list them here, in which, for example, W_{xy} is short for $\partial^2 W/\partial x\partial y$.

$$W_x = -(q+1)[-q(1+x)(q+1))A^3 + (q-x)(q+1))B^3] - x, \tag{16.2.4}$$

$$W_y = (q+1)^2 y[qA^3+B^3] - y, \tag{16.2.5}$$

$$W_z = (q+1)^2 z[qA^3+B^3], \tag{16.2.6}$$

$$W_{xx}=-(q+1)^2[3q\,(1+x(q+1))\,^2A^5-qA^3+3\,(q-x(q+1))\,^2B^5-B^3]-1, \tag{16.2.7}$$

$$W_{yy}=-(q+1)^2[3q(q+1)^2y^2A^5-qA^3+3(q+1)^2y^2B^5-B^3]-1, \tag{16.2.8}$$

$$W_{zz}=-(q+1)^2[3q(q+1)^2z^2A^5-qA^3+3(q+1)^2z^2B^5-B^3], \tag{16.2.9}$$

$$W_{yz}=W_{zy}=-3(q+1)^4yz(qA^5+B^5), \tag{16.2.10}$$

$$W_{zx}=W_{xz}=-3(q+1)^3z[q\,(1+x(q+1))\,A^5-(q-x(q+1))\,B^5], \tag{16.12.11}$$

$$W_{xy}=W_{yz}=-3(q+1)^3y[q\,(1+x(q+1))\,A^5-(q-x(q+1))\,B^5]. \tag{16.2.12}$$

It is a little difficult to draw $W(x,y,z)$, but we can look at the plane $z=0$ and there look at $W(x,y)$. Figure $\text{XVI}.8$ is a contour plot of the surface, for $q=5$, plotted by *Mathematica* by Mr Max Fairbairn of Sydney, Australia. We have already seen, in figure $\text{XVI}.5$, a section along the x-axis.

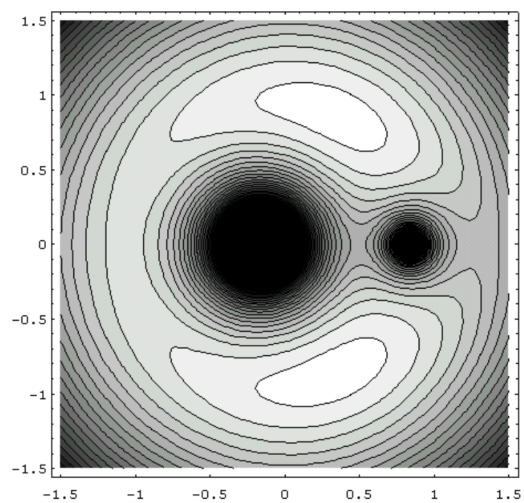


FIGURE XVI.8

Figure $\text{XVI}.9\text{a}$ shows a three-dimensional drawing of the equivalent potential surface in the plane, also plotted by *Mathematica* by Mr Fairbairn. Figure $\text{XVI}.9\text{b}$ is a model of the surface, seen from more or less above. This was constructed of wood by Mr David Smith of the University of Victoria, Canada, and photographed by Mr David Balam, also of the University of Victoria. The mass ratio is $q=5$.

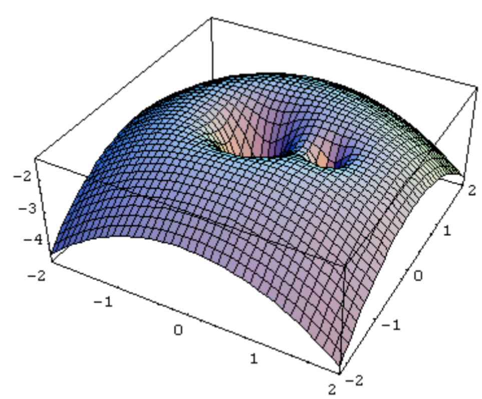


FIGURE XVI.9A

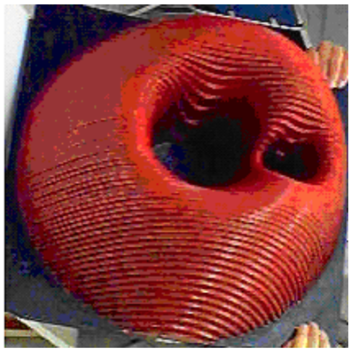


FIGURE XVI.9B

We can imagine the path taken by a small particle in the field of the two principal masses by imagining a small ball rolling or sliding on the equivalent potential surface. It might roll into one of the two deep hyperbolic potential wells representing the gravitational attraction of the two masses. Or it might roll down the sides of the big paraboloid – i.e. it might be flung outwards by the effect of centrifugal force. We must remember, however, that the surface represents the equivalent potential referred to a co-rotating frame, and that, whenever the particle moves relative to this frame, it experiences a Coriolis force at right angles to its velocity.

The three collinear lagrangian points are actually saddle points. Along the x-axis (figure XVI.5, they are maxima, but when the potential is plotted parallel to the y-axis, they are minima. However, in this section, we shall be particularly interested in the equilateral points, whose coordinates (verify this) are . $x_{\mathrm{L}}=\frac{1}{2}\left(\frac{q-1}{q+1}\right)$, $y_{\mathrm{L}}=\pm\frac{\sqrt{3}}{2}$. You may verify from Equations 16.2.4 and 5, (though you may need some patience to do so) that the first derivatives are zero there. Even more patience and determination would be needed to determine from the second derivatives that the equivalent potential is a maximum there – though you may prefer to look at figures XVI.8 and 9 rather than wade through that algebra. I have done the algebra and I can tell you that the first derivatives at the equilateral points are indeed zero and the second derivatives are as follows.

$$W_{xx}=-\frac{3}{4}, \quad W_{yy}=-\frac{9}{4}, \quad W_{zz}=+1, \quad W_{yz}=W_{zx}=0, \quad W_{xy}=-\frac{3\sqrt{3}}{4}\left(\frac{q-1}{q+1}\right). \tag{16.2.1}$$

Because $W_{zz}=+1$, the potential at the equilateral points goes through a minimum as we cross the plane; in the plane, however, W is a maximum, and it has the value there of

$$-\frac{3q^2+5q+3}{2(q+1)^2}. \tag{16.2.2}$$

In the matter of notation, the equilateral points are often called the fourth and fifth lagrangian points, denoted by L_4 and L_5. The question arises, then, which is L_4 and which is L_5? Most authors label the equilateral point that leads the less massive of the two principal masses by 60° L_4 and the one that trails by 60° L_5. This would be unambiguous if we were to restrict our interest, for example, to Trojan asteroids of planets in orbit around the Sun, or Calypso which leads Tethys in orbit around Saturn and Telesto which follows Tethys. There would be ambiguity, however, if the two principal bodies had equal masses, or if the two principal bodies were the members of a close binary pair of stars in which mass transfer led to the more massive star becoming the less massive one. In such special cases, we would have to be careful to make our meaning clear. For the present, however, I shall assume that the two principal bodies have unequal masses, and the equilateral point that precedes the less massive body is L_4.

In figure XVI.10 we are looking in the xy-plane. I have marked a point P, with coordinates (x, y, z); these are expressed in units of a, the constant separation of the two principal masses. The origin of coordinates is the centre of mass C, and the coordinates (in units of a) of the two masses are shown. The angular momentum vector **ω** is directed along the direction of increasing z.

Now imagine a particle of mass m at P. It will be subject to a force given by the negative of the gradient of the potential energy, which is m times the potential. Thus in the x-direction, $ma\ddot{x}=-m\frac{\partial V}{a\partial x}$. In addition to this force, however, whenever it is in motion relative to the co-rotating frame it is subject to a Coriolis force $2m\mathbf{v}\times\boldsymbol{\omega}$. Thus the x-component of the Equation of motion is $ma\ddot{x}=-m\frac{\partial V}{a\partial x}+2m\omega a\dot{y}$. Dividing through by ma we find for the Equation of motion in the x-direction

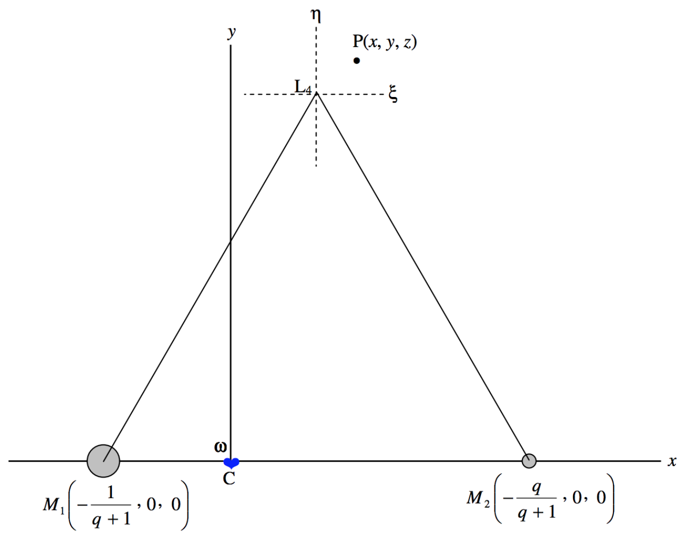


FIGURE XVI.10

$$\ddot{x} = -\frac{1}{a^2}\frac{\partial V}{\partial x} + 2\omega\dot{y}. \tag{16.2.13}$$

Similarly in the other two directions, we have

$$\ddot{y} = -\frac{1}{a^2}\frac{\partial V}{\partial y} - 2\omega\dot{x} \tag{16.2.14}$$

and

$$\ddot{z} = -\frac{1}{a^2}\frac{\partial V}{\partial z}. \tag{16.2.15}$$

These, then, are the differential Equations that will track the motion of a particle moving in the vicinity of the two principal orbiting masses. For large excursions, they are best solved numerically. However, solutions close to the equilateral points lend themselves to a simple analytical solution, which we shall attempt here. Let us start, then, by referring positions to coordinates with origin at an equatorial lagrangian point. The coordinates of the point P with respect to the lagrangian point are (ξ, η, ζ), where $\xi = x - x_L,\ \eta = y - y_L, \zeta = z$. Note also that $\dot{\xi} = \dot{x}\ \ddot{\xi} = \ddot{x}$, etc. We are going to need the derivatives of the potential near to the lagrangian points, and, by Taylor's theorem (or just common sense!) these are given by

$$V_x = (V_x)_\mathrm{L} + \xi(V_{xx})_\mathrm{L} + \eta(V_{yx})_\mathrm{L} + \zeta(V_{zx})_\mathrm{L}, \tag{16.2.16}$$

$$V_y = (V_y)_\mathrm{L} + \xi(V_{xy})_\mathrm{L} + \eta(V_{yy})_\mathrm{L} + \zeta(V_{zy})_\mathrm{L}, \tag{16.2.17}$$

$$V_z = (V_z)_\mathrm{L} + \xi(V_{xz})_\mathrm{L} + \eta(V_{yz})_\mathrm{L} + \zeta(V_{zz})_\mathrm{L}, \tag{16.2.18}$$

We have already worked out the derivatives at the lagrangian points (the first derivatives are zero), so now we can put these expressions into Equations 16.2.13,14 and 15, to obtain

$$\ddot{\xi} - 2\omega\dot{\eta} = \omega^2\left(\frac{3}{4}\xi + \frac{3\sqrt{3}(q-1)}{4(q+1)}\eta\right), \tag{16.2.19}$$

$$\ddot{\eta} + 2\omega\dot{\xi} = \omega^2\left(\frac{3\sqrt{3}(q-1)}{4(q+1)}\xi + \frac{9}{4}\eta\right) \tag{16.2.20}$$

and

$$\ddot{\zeta} = -\omega^2\zeta. \tag{16.2.21}$$

The last of these Equations tells us that displacements in the z-direction are periodic with period equal to the period of the two principal orbiting bodies. The motion is bounded and stable perpendicular to the plane. An orbit inclined to the plane of the orbits containing M_1 and M_2 is stable.

For ξ and η, let us seek periodic solutions of the form

$$\ddot{\xi}=n^2\xi \text{ and } \ddot{\eta}=n^2\eta \tag{16.2.22a,b}$$

so that

$$\dot{\xi}=in\xi \text{ and } \dot{\eta}=in\eta \tag{16.2.23a,b}$$

where n is real and therefore n^2 is positive.

Substitution of these in Equations 16.2.19-21 gives

$$(n^2+\frac{3}{4}\omega^2)\xi+\left(w\omega ni+\frac{3\sqrt{3}}{4}\left(\frac{q-1}{q+1}\right)\omega^2\right)\eta=0 \tag{16.2.24}$$

and

$$\left(2\omega ni-\frac{3\sqrt{3}}{4}\left(\frac{q-1}{q+1}\right)\omega^2\right)\xi-(n^2+\frac{9}{4}\omega^2)\eta=0. \tag{16.2.25}$$

A trivial solution is $\xi=\eta=0$; that is, the particle is stationary at the lagrangian point. While this is indeed a possible solution, it is unstable, since the potential is a maximum there. Nontrivial solutions are found by setting the determinant of the coefficients equal to zero. Thus

$$n^4-\omega^2n^2+\frac{27q\omega^4}{4(q+1)^2}=0. \tag{16.2.26}$$

This is a quadratic Equation in n^2, and for real n^2 we must have $b^2>4ac$, or $1>\frac{27q}{(q+1)^2}$, or $q^2-25q+1>0$. That is, $q>24.959\ 935\ 8$ or $q<1/24.959\ 935\ 8=0.040\ 064\ 206$ We also require n^2 to be not only real but positive. The solutions of Equation $16.2.26$are

$$2n^2=\omega^2\left(1\pm\sqrt{1-27q/(1+q)^2}\right). \tag{16.2.27}$$

For any mass ratio q that is less than $0.040\ 064\ 206$or greater than $24.959\ 935\ 8$both of these solutions are positive. Thus stable elliptical orbits (in the co-rotating frame) around the equilateral lagrangian points are possible if the mass ratio of the two principal masses is greater than about 25, but not otherwise.

If we consider the Sun-Jupiter system, for which $q=1047.35$, we have that

$$n=0.996757\omega \text{ or } n=0.0804645\omega.$$

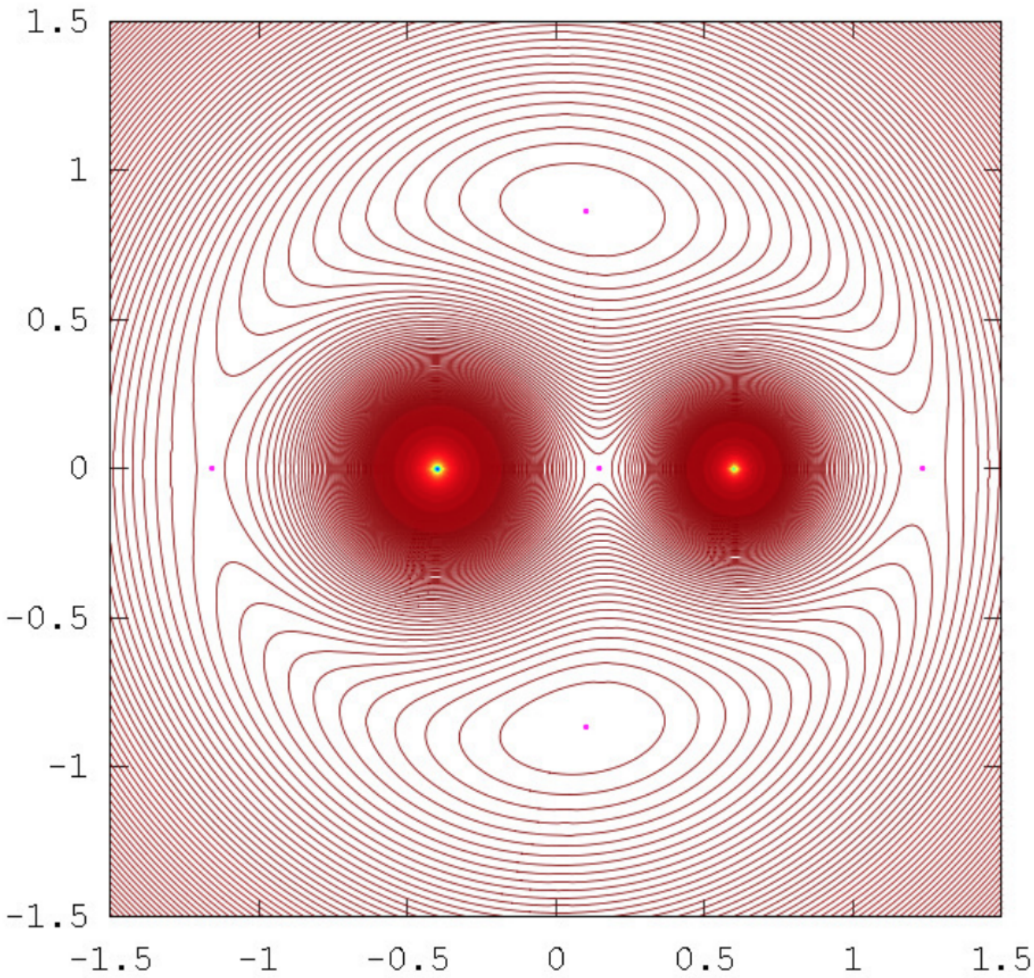
The period of the motion around the lagrangian point is then

$$P=1.0033P_{\text{J}} \text{ or } P=12.428P_{\text{J}}$$

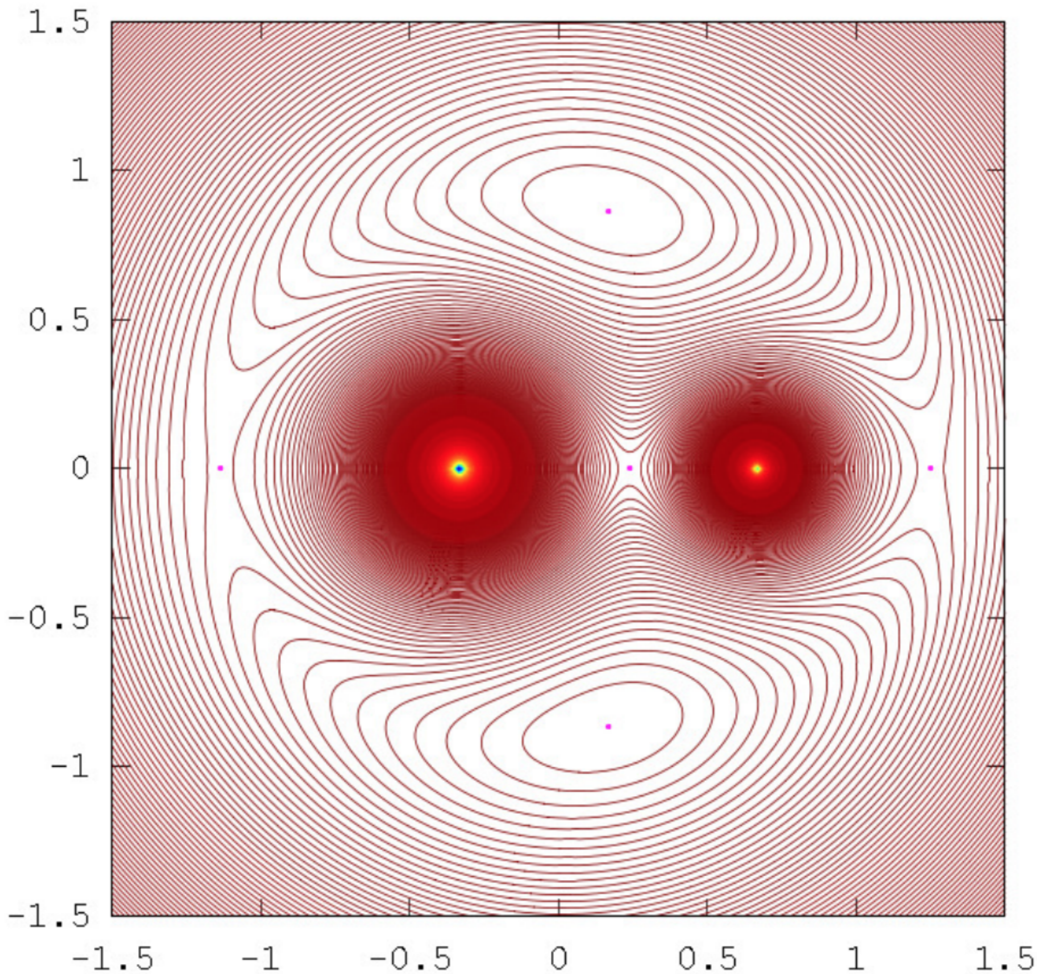
This description of the motion applies to asteroids moving closely around the equilateral lagrangian points, and the approximation made in the analysis appeared in the Taylor expansion for the potential given by Equations 16.2.16-18. For more distant excursions one might try analytical solutions by expanding the Taylor series to higher-order terms (and of course working out the higher-order derivatives) or it might be easier to integrate Equations 16.2.19 and 20 numerically. Many people have had an enormous amount of fun with this. The orbits do not follow the equipotential contours exactly, of course, but in general shape they are not very different in appearance from the contours. Thus, for larger excursions from the lagrangian points the orbits become stretched out with a narrow tail curving towards L_1; such orbits bear a fanciful resemblance to a tadpole shape and are often referred to as tadpole orbits. For yet further excursions, an asteroid may start near L_4 and roll downhill, veering around the back of the more massive body, through the L_1 point and then upwards towards L_5; then it slips back again, goes again through L_1 and then up to L_4 again – and so on. This is a so-called horseshoe orbit.

The drawings below show the equipotential contours for a number of mass ratios. These drawings were prepared using *Octave* by Dr Mandayam Anandaram of Bangalore University, and are dedicated by him to the late Max Fairbairn of Sydney, Australia, who prepared figures $\text{XVI}.8$ and $\text{XVI}.9\text{a}$ for me shortly before his untimely death. Anand and Max were my first graduate students at the University of Victoria, Canada, many years ago. These drawings show the gradual evolution from tadpole-shaped contours to horseshoe-shaped contours. The mass-ratio $q=24.959\ 935\ 8$ is the critical ratio below which stable orbits around the equilateral points L_4 and L_5 are not possible. The massratios $q=81.3$ and 1047 are the ratios for the Earth-Moon and Sun-Jupiter systems respectively. The reader will notice that, in places where the contours are closely-spaced, in particular close to the deep potential well of the larger mass, Moiré fringes appear. These fringes appear where the contour separation is comparable to the pixel size, and the reader will recognize them as Moiré fringes and, we think, will not be misled by them.

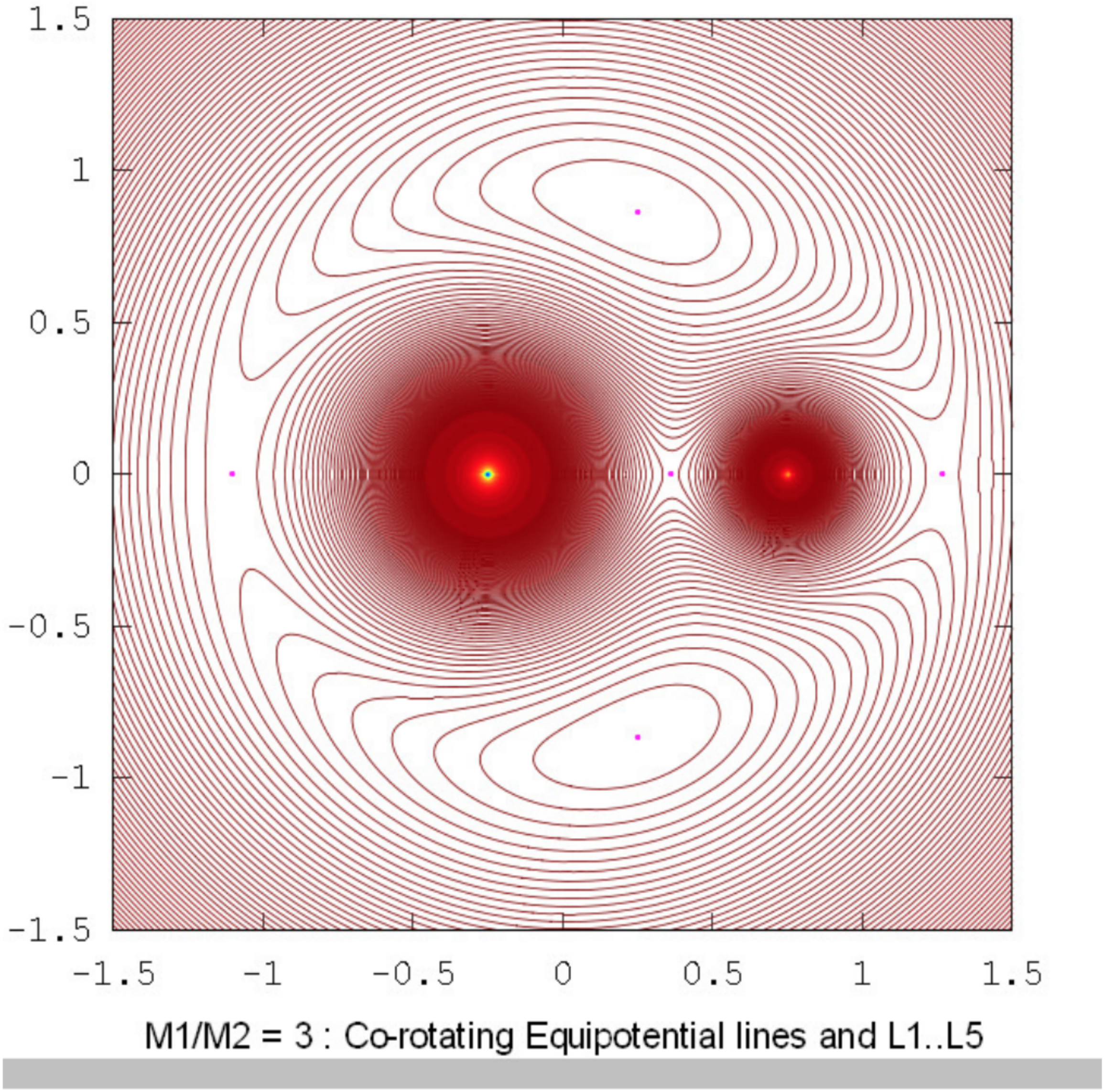
Dr Anandaram has also prepared a number of fascinating drawings in which sample orbits are superimposed, in a second colour, on the equipotential contours. These include tadpole orbits in the vicinity of the equilateral points; “triangular” orbits of the Hilda asteroids, which are in 2 : 3 resonance with Jupiter; the almost “square” orbit of Thule, which is in 3 : 4 resonance with Jupiter; and half of a complete 9940 year libration period of Pluto, which is in 3 : 2 resonance with Neptune. It is proposed to publish these in a separate paper dedicated to Max, the reference to which will in due course be given in these notes.

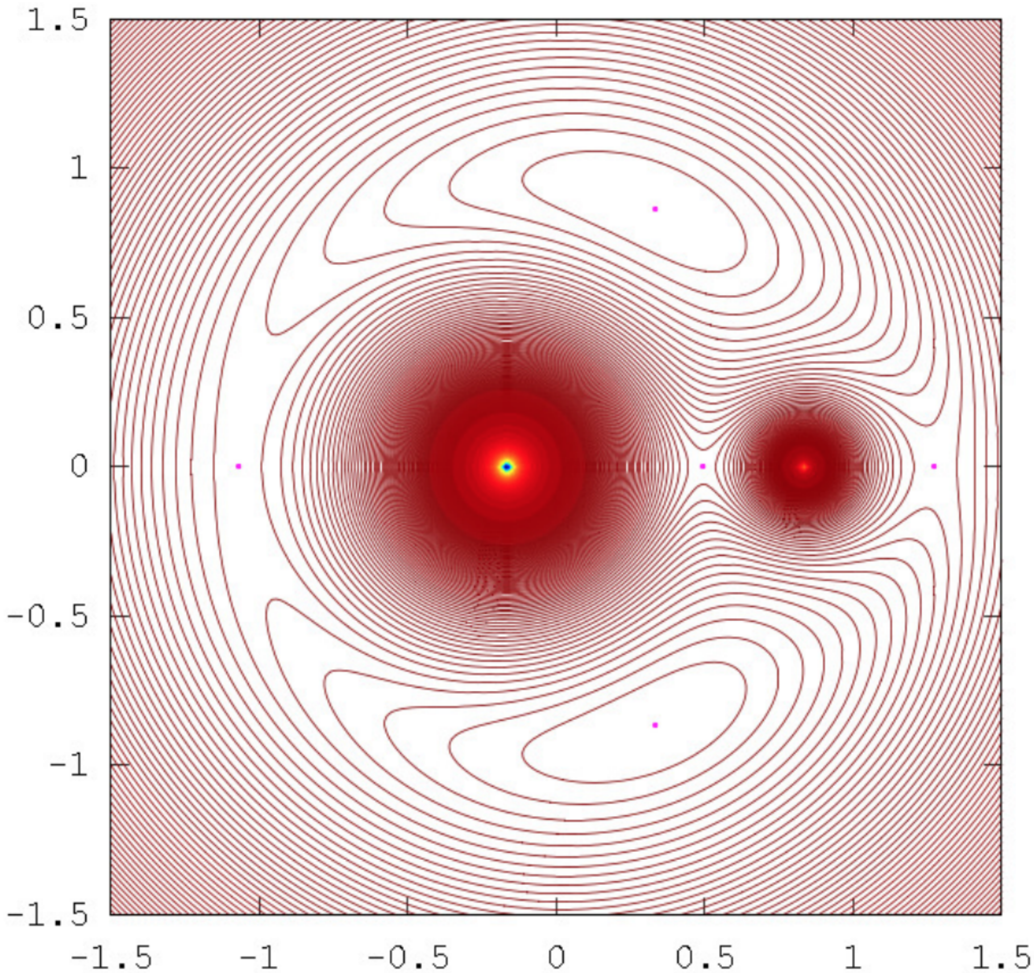


q = M1/M2 = 1.5 : Equipotential contours and Lagrangian points

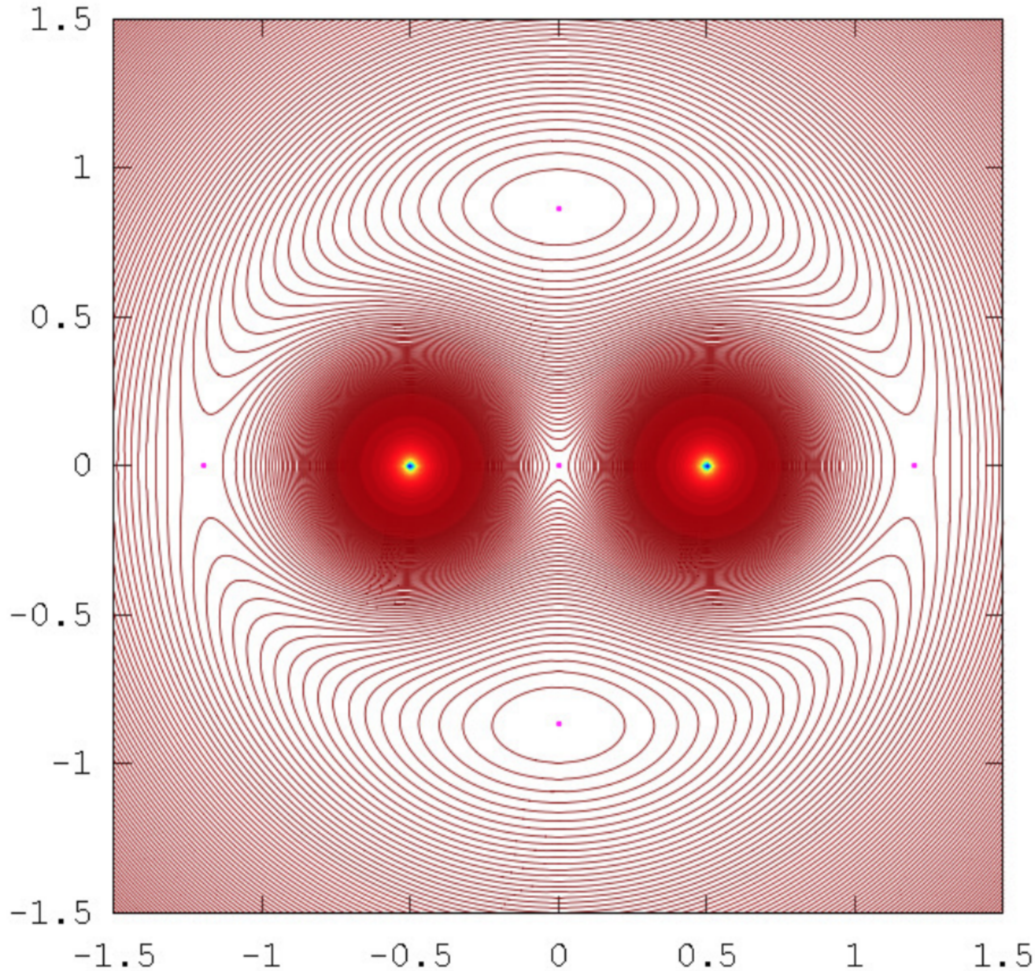


q = M1/M2 = 2 : Equipotential contours and Lagrangian points

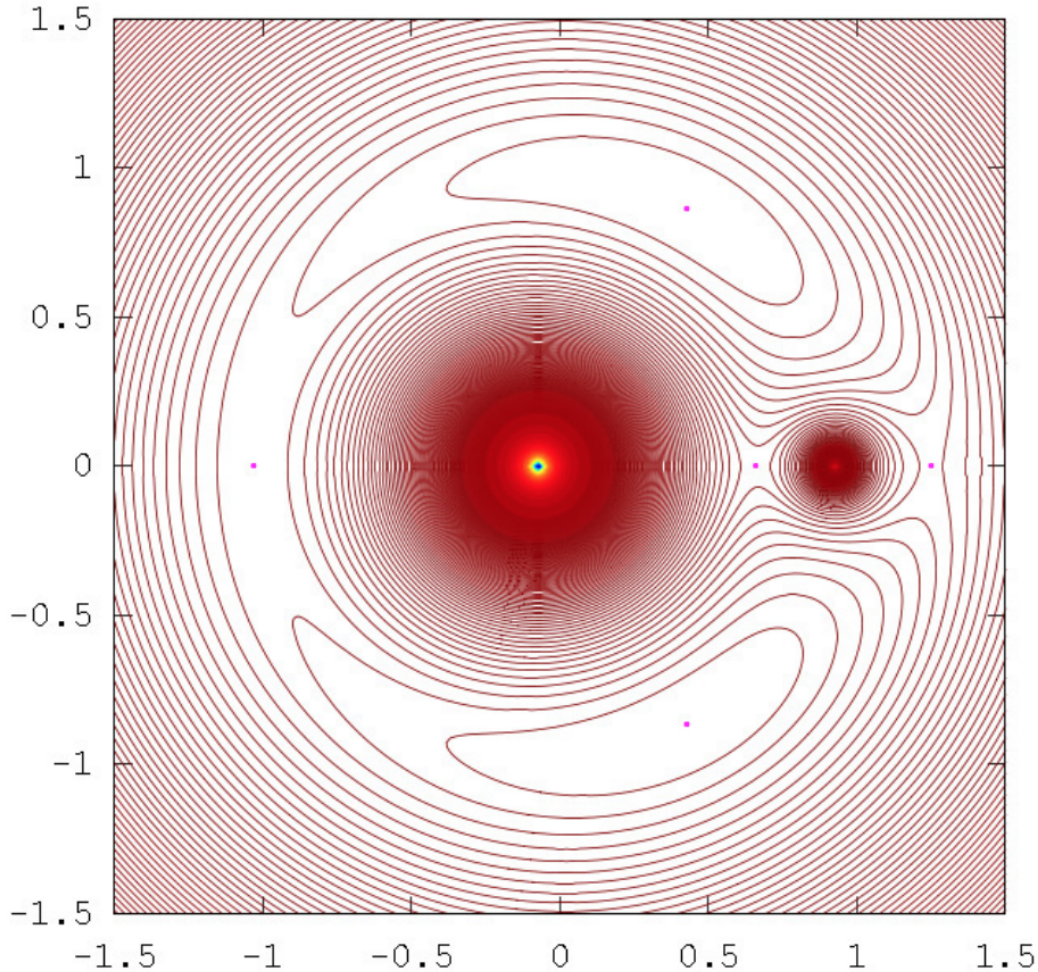

1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5
M1/M2 = 3 : Co-rotating Equipotential lines and L1..L5



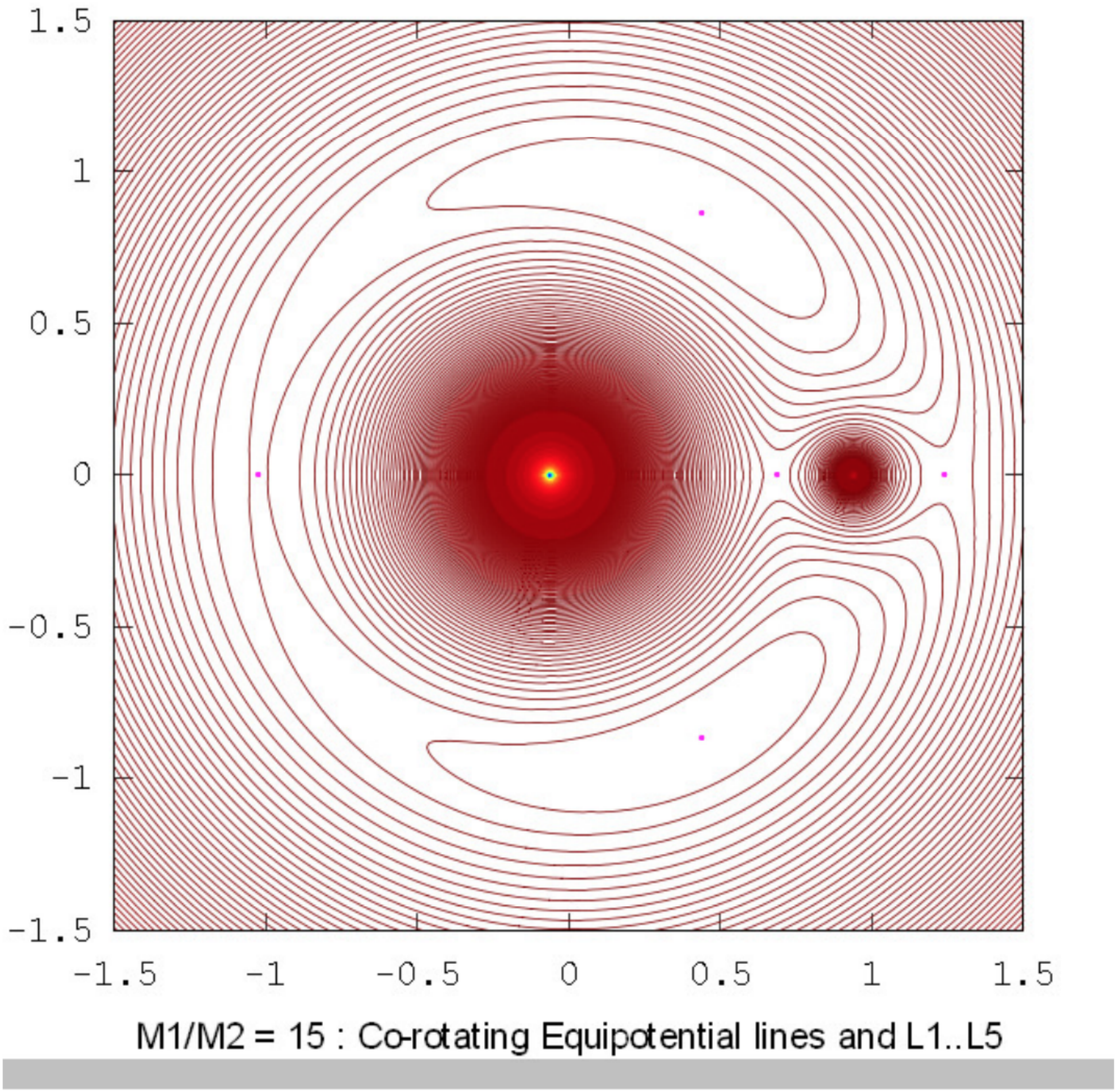
q = M1/M2 = 5 : Equipotential contours and Lagrangian points

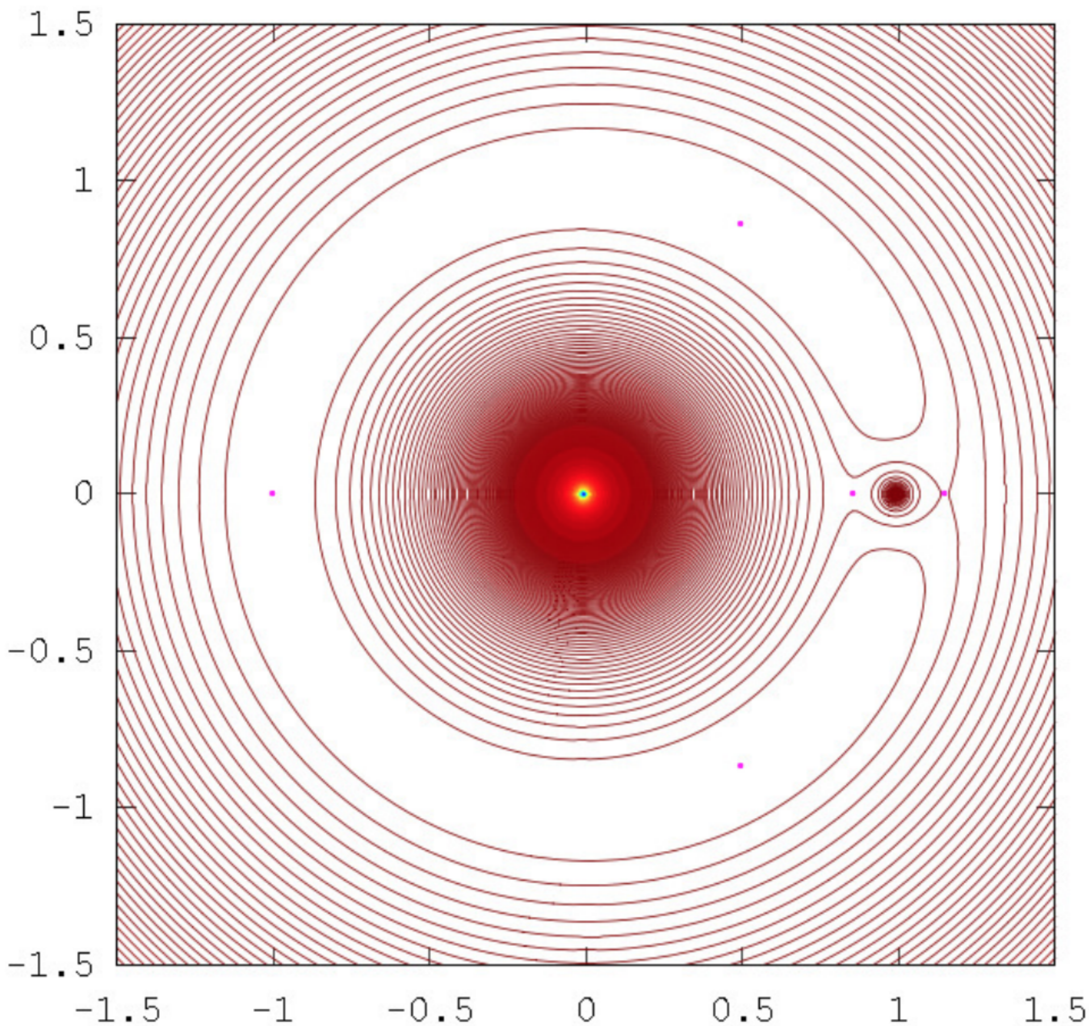


q = M1/M2 = 1 : Equipotential contours and Lagrangian points

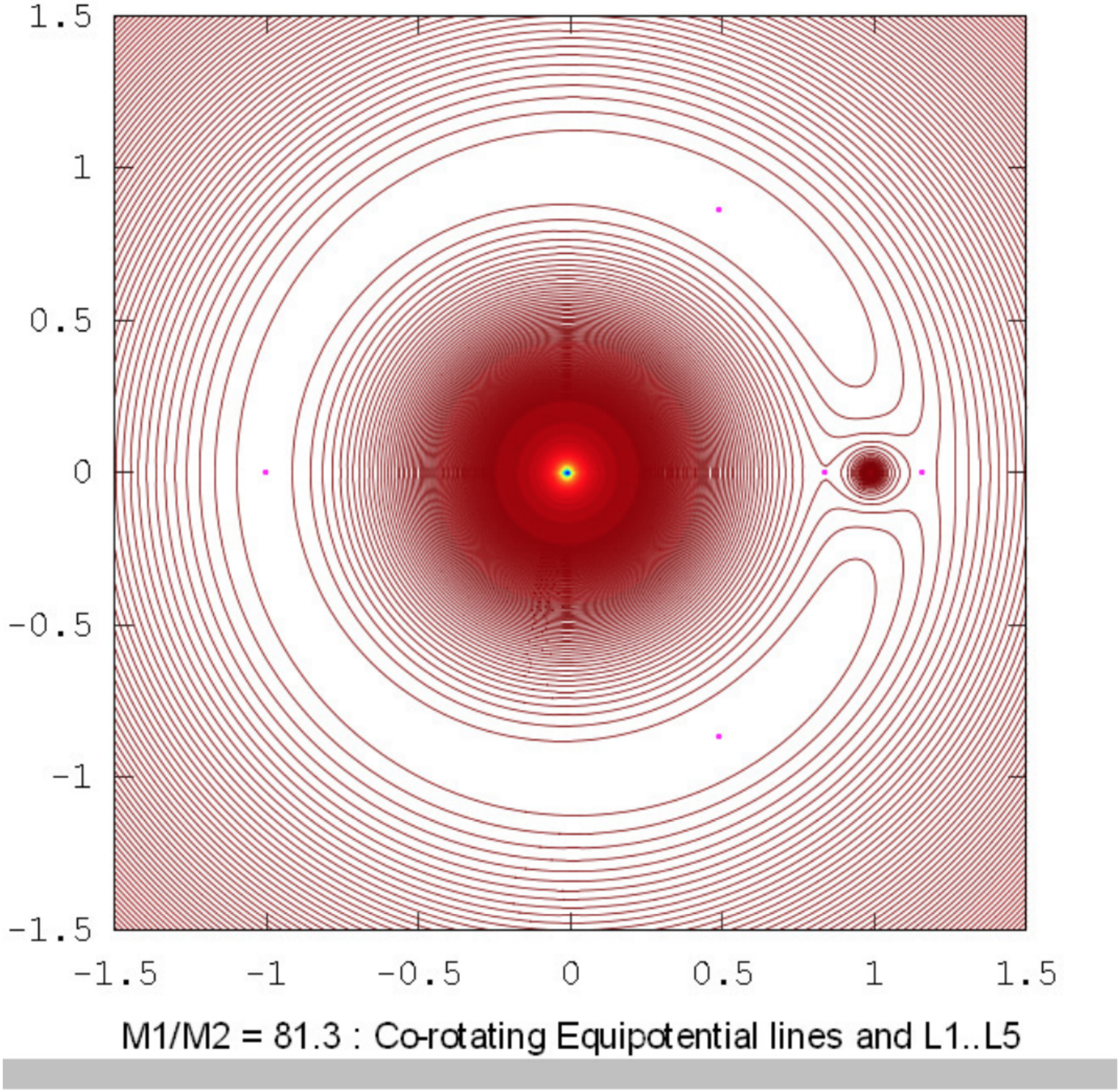


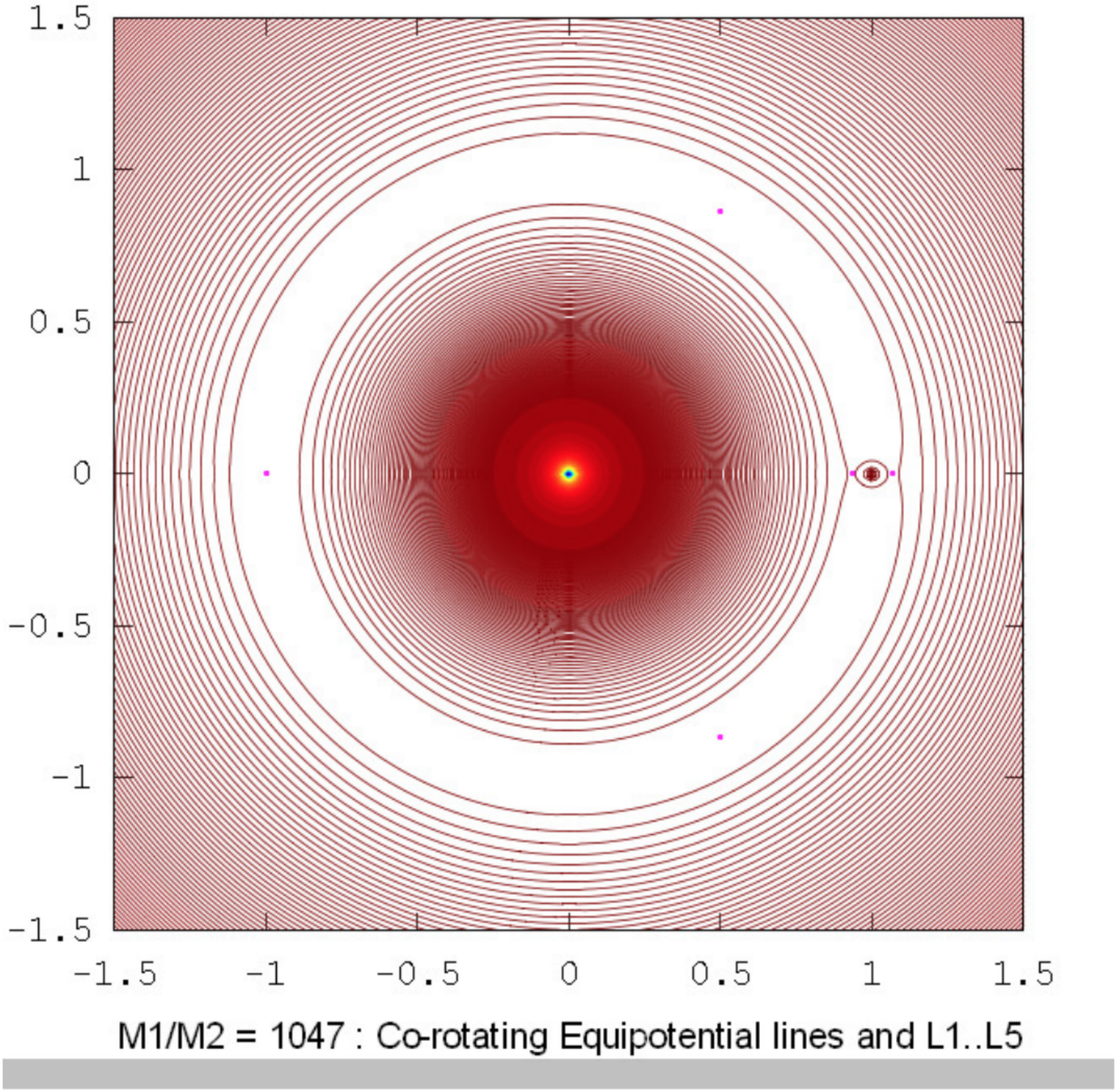
q = M1/M2 = 12.5 : Equipotential contours and Lagrangian points

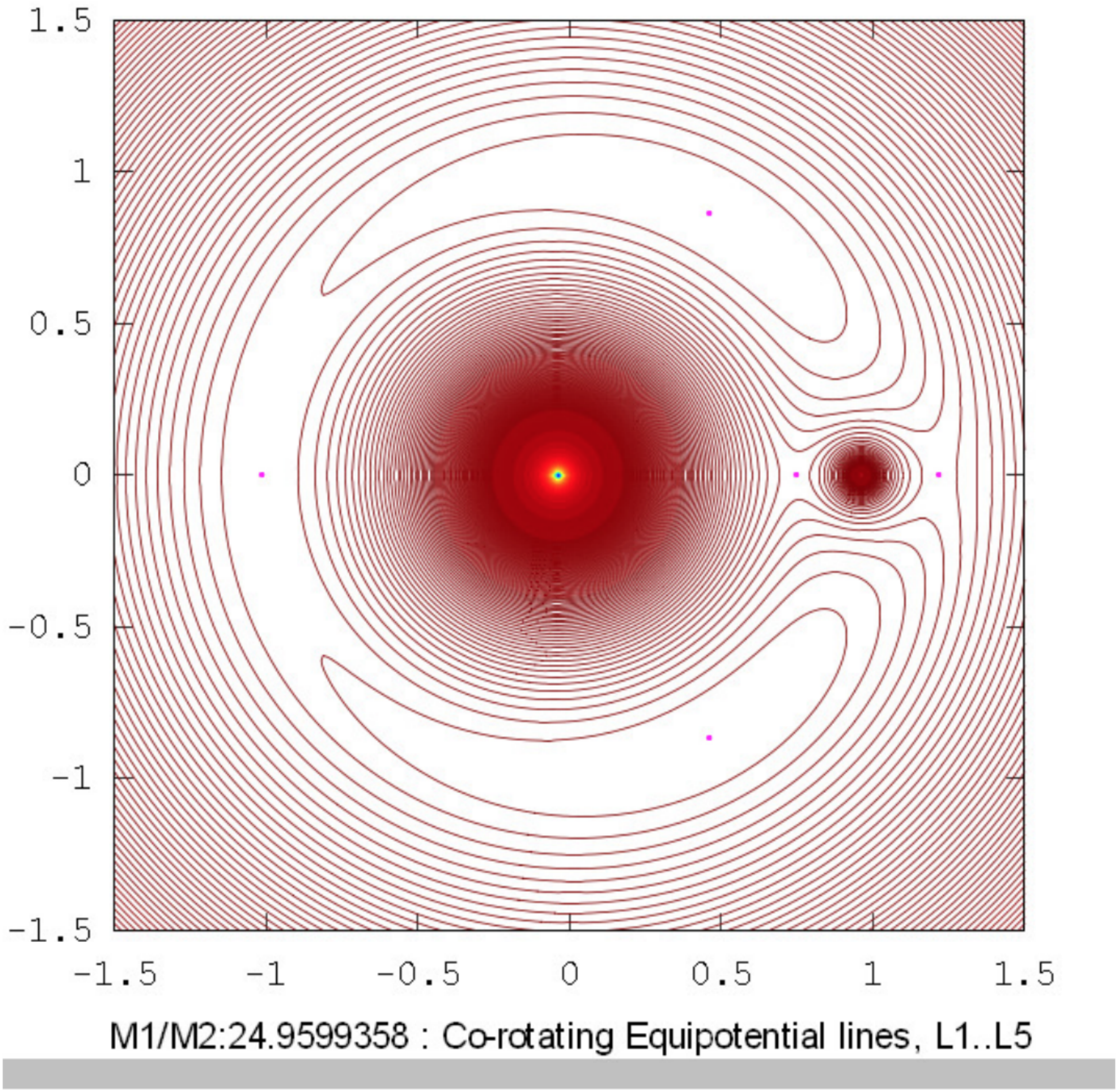

1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5
M1/M2 = 15 : Co-rotating Equipotential lines and L1..L5



q = M1/M2 = 100 : Equipotential contours and Lagrangian points


1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5
M1/M2 = 81.3 : Co-rotating Equipotential lines and L1..L5


1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5
M1/M2 = 1047 : Co-rotating Equipotential lines and L1..L5



Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

16.3: Inverse Square Attractive Force

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

16.4: Hooke's Law

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

16.5: Inverse Fourth Power Force

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

Welcome to the Physics Library. This Living Library is a principal hub of the LibreTexts project, which is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning. The LibreTexts approach is highly collaborative where an Open Access textbook environment is under constant revision by students, faculty, and outside experts to supplant conventional paper-based books.

Courses

Bookshelves

16.7: The Equilateral Lagrangian Points

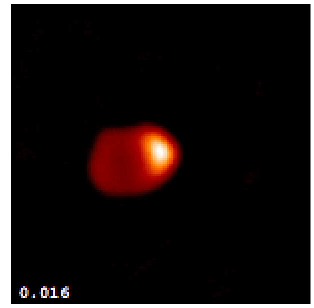
Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

17: VISUAL BINARY STARS

A visual binary is a gravitationally bound system that can be resolved into two stars. These stars are estimated, via Kepler's 3rd law, to have periods ranging from a number of years to thousands of years. A visual binary consists of two stars, usually of a different brightness.



17.1: Introduction to Visual Binary Stars

Many stars in the sky are seen through a telescope to be two stars apparently close together. By the use of a filar micrometer it is possible to measure the position of one star (the fainter of the two, for example) with respect to the other. The position is usually expressed as the angular distance ρ (in arcseconds) between the stars and the position angle θ of the fainter star with respect to the brighter. (The separation can be determined in kilometres rather than merely in arcseconds if the distance from Earth to the pair is known.) The position angle is measured counterclockwise from the direction to north. See figure XVII.1.

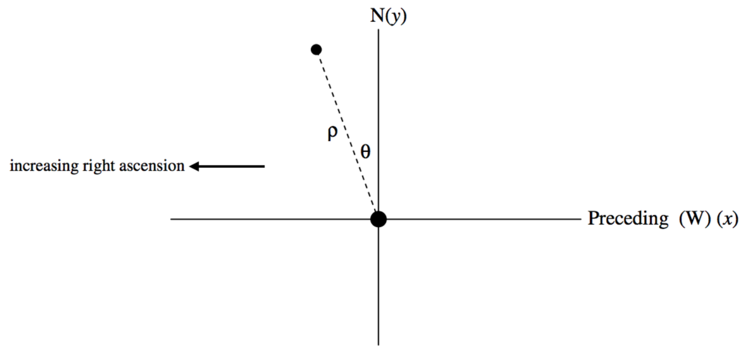


FIGURE XVII.1

These coordinates (ρ, θ) of one star with respect to the other can, of course, easily be converted to (x, y) coordinates. In any case, after the passage of many years (sometimes longer that the lifetime of an astronomer) one ends up with a table of coordinates as a function of time. Because the orbital period is typically of the order of many years, and the available observations are correspondingly spread out over a long period of time, it needs to be pointed out that all position angles, which are measured with respect to the equator of date, need to be adjusted so as to refer to a standard equator, such as that of J2000.0 I don't wish to interrupt the flow of thought here by discussing this point (important though it is) in detail; suffice it to say that

$$\theta_{2000.0} = \theta_t + 20'' \times (2000 - t) \sin\alpha \sec\delta, \tag{17.1.1}$$

where t is the epoch of the observation in years, and the position angles are expressed in arcseconds.

If one star appears to move in a straight line with respect to the other, it is probable that the two stars are not physically connected but they just happen to lie almost in the same line of sight. Such a pair is called an *optical pair* or an *optical double*.

However, if one star appears to describe an ellipse relative to the other, then the two stars are physically connected and are moving around their common centre of mass.

The angular separation between the two stars is usually very small, of the order of arcseconds or less, and is not easy to measure. Much more difficult to measure would be the distances of the two stars individually from their mutual centre of mass. Close pairs are usually measured visually with a filar micrometer, and it is then almost invariably the case that what is measured is the position of the secondary with respect to the primary. Wider pairs, however, can be measured from photographs, or, today, from CCD images. In that case, not only are the measurements more precise, but it is possible to measure the position of each component with respect to background calibration stars, and hence to measure the position of each component with respect to the centre of mass of the system. This enables us to determine the mass ratio of the two components. Pairs that are sufficiently wide apart for photographic measurements, however, come with their own set of problems. If their angular separation is large, this could mean either that the real, linear separation in kilometres is large, or else that the stars are not very far from the Sun. In the former case, we may have to wait rather a long time (perhaps more than an average human lifetime) for the two stars to describe a complete orbit. In the latter case, we may have to take account of complications such as proper motion or annual parallax.

The brighter of the two stars is the *primary*, and the fainter is the *secondary*. This will nearly always mean (though not necessarily so) that the primary star is also the more massive of the pair, but this cannot be assumed without further evidence. If the two stars are of equal brightness, it is arbitrary which one is designated the primary. If the two stars are of equal brightness, it can sometimes happen that, when they become very close to each other, they merge and cannot be distinguished until their separation is sufficiently great for them to be resolved again. It may then not be obvious which of the two had been designated the "primary".

The orbit of the secondary around the primary is, of course, a keplerian ellipse. But what one sees is the *projection* of this orbit on the "plane of the sky". (The "plane of the sky" is the phrase almost universally used by observational astronomers, and there is no substantial objection to it; formally it means the tangent plane to the celestial sphere at the position of the primary component.) The projection of the *true orbit* on the plane of the sky is the *apparent orbit*, and both are ellipses. The centre of the true ellipse maps on to the centre of the apparent ellipse, but the foci of the true ellipse do *not* map on to the foci of the apparent ellipse. The primary star is at a focus of the true ellipse, but it is not at a focus of the apparent ellipse. The radius vector in the true orbit sweeps out equal areas in equal times, according to Kepler's second law. In projection to the plane of the sky, all areas are reduced by the same factor $(\cos i)$. Consequently the radius vector in the apparent orbit also sweeps out equal areas in equal times, even though the primary star is not at a focus of the apparent ellipse.

Having secured the necessary observations over a long period of time, the astronomer faces two tasks. First the apparent orbit has to be determined; then the true orbit has to be determined.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

17.2: Determination of the Apparent Orbit

The apparent orbit may be said to be determined if we can determine the size of the apparent ellipse (i.e. its semi major axis), its shape (i.e. its eccentricity), its orientation (i.e. the position angle of its major axis) and the two coordinates of the centre of the ellipse with respect to the primary star. Thus there are five parameters to determine.

The general Equation to a conic section (see Section 2.7 of Chapter 2) is of the form

$$ax^2 + 2hxy + by^2 + 2gx + 2fy + 1 = 0, \tag{17.2.1}$$

so that we can equally say that the apparent orbit has been determined if we have determined the five coefficients $a,\ h,\ b,\ g,\ f$. Sections 2.8 and 2.9 described how to determine these coefficients if the positions of five or more points were given, and section 2.7 dealt with how to determine the semi major axis, the eccentricity, the orientation and the centre given $a,\ h,\ b,\ g$ and f.

We may conclude, therefore, that in order to determine the apparent ellipse all that need be done is to obtain five or more observations of (ρ, θ) or of (x, y), and then just apply the methods of section 2.8 and 2.9 to fit the apparent ellipse. Of course, although five is the minimum number of observations that are essential, in practice we need many, many more (see section 2.9), and in order to get a good ellipse we really need to wait until observations have been obtained to cover a whole period. But merely to fit the best ellipse to a set of (x, y) points is not by any means making the best use of the data. The reason is that an observation consists not only of (ρ, θ) or of (x, y), but also the time, t. In fact the separation and position angle are quite difficult to measure and will have quite considerable errors, while the *time* of each observation is known with great precision. We have so far completely ignored the one measurement that we know for certain!

We need to make sure that the apparent ellipse that we obtain *obeys* Kepler's *second law*. Indeed it is more important to ensure this than blindly to fit a least-squares ellipse to n points.

If I were doing this, I would probably plot two separate graphs – one of ρ (or perhaps ρ^2) against time, and one of θ against time. One thing that this would immediately achieve would be to identify any obviously bad measurements, which we could then reject. I would draw a smooth curve for each graph. Then, for equal time intervals I would determine from the graphs the values of ρ and dθ/dt and I would then calculate $\rho^2\ d\theta/dt$. According to Kepler's second law, this should be constant and independent of time. I would then adjust my preliminary attempt at the apparent orbit until Kepler's second law was obeyed and $\rho^2\ d\theta/dt$ was constant. A good question now, is, which should be adjusted, ρ or θ ? There may be no hard and fast invariable answer to this, but, generally speaking, the measurement of the separation is more uncertain than the measurement of the position angle, so that it would usually be best to adjust ρ.

If we are eventually satisfied that we have the best apparent ellipse that satisfies as best as possible not only the positions of the points, but also their times, and that the apparent ellipse satisfies Kepler's law of areas, our next task will be to determine the elements of the true ellipse.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

17.3: The Elements of the True Orbit

Unless we are dealing with photographic measurements in which we have been able to measure the positions of both components with respect to their mutual centre of mass, I shall assume that we are determining the orbit of the secondary component with respect to the primary as origin and focus.

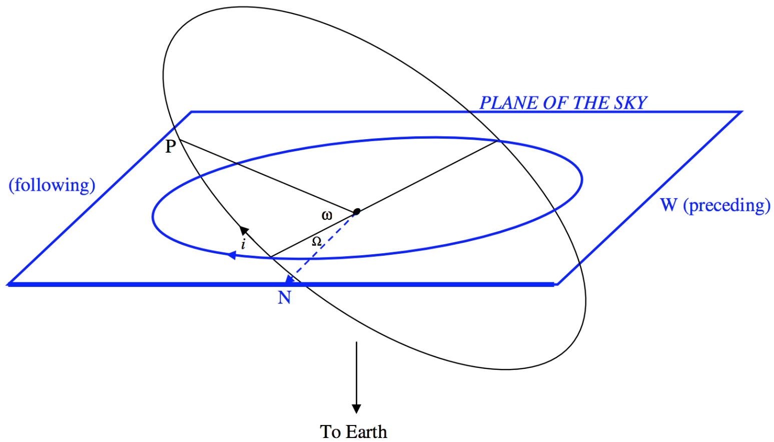


FIGURE XVII.2

In figure XVII.2, which has tested my artistic talents and computer skills to the full, the blue plane is intended to represent the plane of the sky, as seen from "above" – i.e. from outside the celestial sphere. Embedded in the plane of the sky is the apparent orbit of the secondary with respect to the primary as origin and focus. The dashed arrow shows the colure (definition of "colure" – Section 6.4 of Chapter 6) through the primary, and points to the north celestial pole. The primary star is not necessarily at a focus of the apparent ellipse, as discussed in the previous section. As drawn, the position angle of the star is increasing with time – though of course in a real case it is equally likely to be increasing or decreasing with time.

The black ellipse is the true orbit, and of course the primary is at a focus of it. If it does not appear so in figure XVII.2, this is because the true orbit is being seen in projection.

The elements of the true orbit to be determined (if possible) are

- a the semi major axis;
- e the eccentricity;
- i the inclination of the plane of the orbit to the plane of the sky;
- Ω the position angle of the ascending node;
- ω the argument of periastron;
- T the epoch of periastron passage.

All of these will be familiar to those who have read Chapter 10, section 10.2. Some comments are necessary in the context of the orbit of a visual binary star.

Ideally, the semi major axis would be expressed in kilometres or in astronomical units of distance – but this is not possible unless the distance from Earth to the binary star is known. If the distance is not known (as will often be the case), the semi major axis is customarily expressed in arcseconds.

It is sometimes said that, from measurements of separation and position angle alone, and with no further information, and in particular with no spectroscopic measurements of radial velocity, it is not possible to determine the *sign* of the inclination of the true orbit of a visual binary star. This may be a valid view, but, as the late Professor Joad might have said, it all depends on what you mean by "inclination". As with the orbits of planets around the Sun, as described in Chapter 10, Section 10.2, we take the point of view here that the inclination of the orbital plane to the plane of the sky is an angle that lies between 0° and 180° inclusive; that is to say, the inclination is positive, and the question of its sign does not arise. After all an inclination of, say, "-30°" is no different from an inclination of $+150^\circ$. Thus we cannot be ignorant of the "sign" of the inclination. What we do *not* know, however, is which node is the ascending node and which is the descending node.

The Ω that is usually recorded in the analysis of the orbit of a visual binary unsupported by spectroscopic radial velocities is the node for which the position angle is less than 180° – and it is not known whether this is the ascending or descending node.

If the inclination of the orbital plane is less than 90°, the position angle of the secondary will increase with time, and the orbit is described as *direct* or *prograde*. If the position angle decreases with time, the orbit is *retrograde*.

The orbital inclination of a *spectroscopic* binary cannot be determined from spectroscopic observations alone. The inclination of a *visual* binary *can* be determined, although, as discussed above, it is not known which node is ascending and which is descending. If the binary is both a visual binary and a spectroscopic binary, not only can the inclination be determined, but the ambiguity in the nodes is removed. In addition, it may be possible to determine the masses of the stars; this aspect will be dealt with in the chapter on spectroscopic binary stars.

Binary stars that are simultaneously visual and spectroscopic binaries are rare, and they are a copious source of valuable information when they are found. Visual binary stars, unless they are relatively close to Earth, have a large true separation, and consequently their orbital speeds are usually too small to be measured spectroscopically. Spectroscopic binary stars, on the other hand, move fast in their orbits, and this is because they are close together – usually too close to be detected as visual binaries. Binaries that are both visual and spectroscopic are usually necessarily relatively close to Earth.

The element ω, the argument of periastron, is measured from the ascending node (or the first node, if, as is usually the case, the type of node is unknown) from 0° to 360° in the direction of motion of the secondary component.

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

17.4: Determination of the Elements of the True Orbit

I am assuming at this stage that we have used all the observations plus Kepler's second law and have determined the apparent orbit well, and can write it in the form

$$ax^2+2hxy+by^2+2gx+2fy+c=0. \tag{17.4.1}$$

[The coefficients a and b here, and e in Equation 17.4.3, do not, of course, mean the semi major axis a, the semi minor axis b and eeccentricity e of the true ellipse. It is thought that the reader will be unlikely confused by this, but I have nevertheless used slightly different fonts for them.]

The origin of coordinates here is the primary star, which, although it is at the focus of the true ellipse, is not at the focus of the apparent ellipse. The x-axis points west (to the right) and the y-axis points north (upwards), and position angle θ (measured counterclockwise from north) is given by $\tan\theta=-x/y$. Our task is now to find the elements of the true orbit.

During the analysis we are going to be obliged, on more than one occasion, to determine the coordinates of the points where a straight line $y=mx+d$ intersects the ellipse, so it will be worth while to prepare for that now and write a quick program for doing it instantly. The x-coordinates of these points are given by solution of $\mathrm{a}+2hm+\mathrm{b}m^2)x^2+(2hd+2\mathrm{b}md+2g+2fm)x+\mathrm{b}d^2+2fd+c=0,\ 17.4.2$ and the y-coordinates are given by solution of the Equation $(\mathrm{b}+2hn+\mathrm{a}n^2)y^2+(2h\mathrm{e}+2\mathrm{a}n\mathrm{e}+2f+2gn)y+\mathrm{ae}^2+2g\mathrm{e}+c=0,\ 17.4.3$ where $n=/1m$ and $\mathrm{e}=-d/m$. If m is positive the larger solution for y corresponds to the larger solution for x; If m is negative the larger solution for y corresponds to the smaller solution for x.

If the line passes through F, so that $d=0$, these Equations reduce to

$$(\mathrm{a}+2hm+bm^2)x^2+(2g+2fm)x+c=0, \tag{17.4.4}$$

and

$$(\mathrm{b}+2hn+\mathrm{a}n^2)y^2+(2f+2gn)y+c=0. \tag{17.4.5}$$

In figure XVII.3 I draw the *true ellipse* in the plane of the orbit. F is the primary star at a focus of the true ellipse. C is the centre of the ellipse. I have drawn also the auxiliary circle, the major axis (with periastron P at one end and apastron A at the other end), the latus rectum MN through F and the semi minor axis CK. The ratio FC/PC is the eccentricity e of the true ellipse, and the ratio of minor axis to major axis is $\sqrt{1-e^2}$. This is also the ratio of any ordinate on the auxiliary circle to the corresponding ordinate on the ellipse. Thus I have extended the latus rectum and the semi minor axis by the reciprocal of this factor to meet the auxiliary circle in M′, N′ and K′.

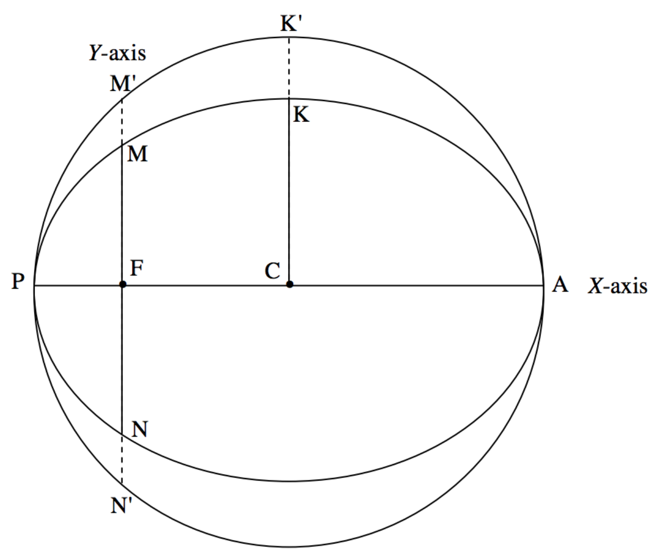


FIGURE 17.3

Now, in figure XVII.4, we are going to look at the same thing as seen projected on the plane of the sky.

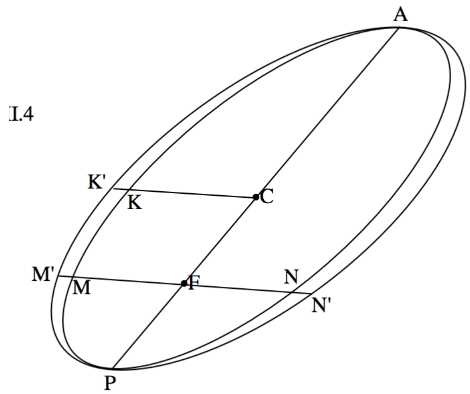


FIGURE XVII.4

The *true ellipse* has become the *apparent ellipse*, and the *auxiliary circle* has become the *auxiliary ellipse*. At the start of the analysis, we know only the apparent ellipse, which is given by Equation 17.4.1, and the position of the focus F, which is at the origin of coordinates, (0 , 0). F is not at a focus of the apparent ellipse, but C *is* at the centre of the apparent ellipse.

From section 2.7, we can find the coordinates $(\bar{x}, \bar{y})$ of the centre C. These are $(\bar{g}/\bar{c},\ \bar{f}/\bar{c})$, where the bar denotes the cofactor in the determinant of coefficients. Thus the slope of the line FC, which is a portion of the true major axis, is $\bar{f}/\bar{g}$. We can now write the Equation of the true major axis in the form $y = mx$ hence, by use of Equations 17.4.4 and 5, we can determine the coordinates of periastron P and apastron A. We can now find the distances FC and PC; and the ratio FC/PC, which has not changed in projection, is the eccentricity e of the true ellipse.

<u>Thus e has been determined.</u>

Our next step is going to be to find the slope of the projected latus rectum MN and the projected semi minor axis CK, which is, of course, parallel to the latus rectum. If the Equation to the projected latus rectum is $y = mx$, we can find the x-coordinates of M and N by use of Equation 17.4.4. But if MN is a latus rectum, it is of course bisected by the major axis and therefore the length FM and FN are equal. That is to say that the two solutions of Equation 17.4.4 are equal in magnitude and opposite in sign, which in turn implies that the coefficient of x is zero. Thus the slope of the latus rectum (and of the minor axis) is $-g/f$.

(It is remarked in passing that the projected major and minor axes are *conjugate diameters* of the apparent ellipse, with slopes $\bar{f}/\bar{g}$ and $-g/f$ respectively.)

Now that we have determined the slope of the projected latus rectum, we can easily calculate the coordinates of M and N by solution of Equations 17.4.4 and 17.4.5. Further, CK has the same slope and passes through C, whose coordinates we know, so it is easy to write the Equation to the projected minor axis in the form $y = mx + d$ (d is $\bar{y} - m\bar{x}$), and then solve Equations 17.4.2 and 17.4.3 to find the coordinates of K.

Now we want to extend FM, FN, CK to M', N' and K' . For M' and N' this is done and simply by replacing x and y by kx and ky, where k is the factor $1/\sqrt{1-e^2}$. For K', it is done by replacing x and y by $\bar{x} + k(x - \bar{x})$ and $\bar{y} + k(y - \bar{y})$ respectively.

We now have five points, P, A, M', N' and K', whose coordinates are known and which are on the auxiliary ellipse. This is enough for us to determine the Equation to the auxiliary ellipse in the form of Equation 17.4.1. A quick method of doing this is described in section 2.8 of Chapter 2.

The slopes of the major and minor axis *of the auxiliary ellipse* (written in the form of Equation 17.4.1) are given by

$$\tan 2\theta = \frac{2h}{a-b}. \tag{17.4.6}$$

This Equation has two solutions for θ, differing by 90°, the tangents of these being the slopes of the major and minor axes of the auxiliary ellipse. Now that we know these slopes, we can write the Equation to these axes in the form $y = mx + d$ (d is

$\bar{y} - m\bar{x}$) and so we can determine where the axes cut the auxiliary ellipse and hence we can determine the lengths of the both axes of the auxiliary ellipse.

This has been hard work so far, but we are just about to make real progress. The major axis of the auxiliary ellipse is the only diameter of the auxiliary circle that has not been foreshortened by projection, and therefore it is equal to the diameter of the auxiliary circle, and hence the major axis of the auxiliary ellipse is also equal to the major axis of the true ellipse.

Thus a has been determined.

The ratio of the lengths of the minor to major axes of the auxiliary ellipse is equal to the amount by which the auxiliary circle has been flattened by projection. That is, the ratio of the lengths of the axes is equal to $|\cos i|$. Since the lengths of the axes are essentially positive, we obtain only $|\cos i|$, not $\cos i$ itself. However, by our definition of i, it lies between 0° and 180° and is less than or greater than 90° according to whether the position angle of the secondary component is increasing or decreasing with time. For example, if $|\cos i| = \frac{1}{2}$, i is 60° or 120°, to be distinguished by the sense of motion of the secondary component.

The *line of nodes* passes through F and *is parallel to the major axis of the auxiliary ellipse*. This indeed is the reason why the major axis of the auxiliary ellipse was unchanged from its original diameter of the auxiliary circle. We therefore already know the slope of the line of nodes and hence we know the position angle of the first node.

Thus Ω has been determined.

In figure XVII.5 I have added the line of nodes, parallel to the (not drawn) major axis of the auxiliary ellipse. I have used the symbols N and N' for the first and second nodes, but we do not know (and cannot know without further information) which of these is ascending and which is descending.

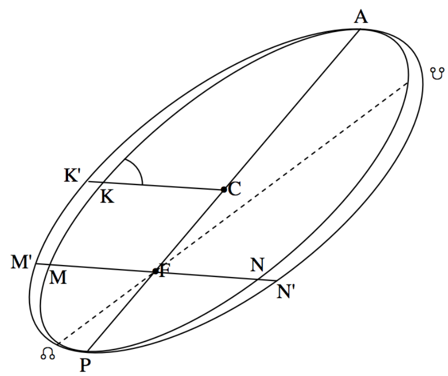


FIGURE XVII.5

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

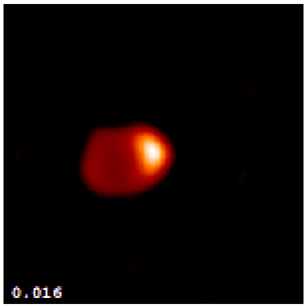
17.5: Construction of an Ephemeris

Contributors and Attributions

- Jeremy Tatum (University of Victoria, Canada)

CHAPTER OVERVIEW

18: SPECTROSCOPIC BINARY STARS



There are many binary stars whose angular separation is so small that we cannot distinguish the two components even with a large telescope – but we can detect the fact that there are two stars from their spectra. In favorable circumstances, two distinct spectra can be seen. It might be that the spectral types of the two components are very different – perhaps a hot A-type star and a cool K-type star, and it is easy to recognize that there must be two stars there.

18.1: Introduction to Spectroscopic Binary Stars

The orbital elements of a binary star system are described in Chapter 17, and are *a, e* , *i*, Ω, ω and *T*. However, on thinking about the meaning of the element W, the position angle of the ascending node, the reader will probably agree that we cannot tell the position angle of either node from radial velocity measurements of an unresolved binary star. We have no difficulty, however, in determining which component is receding from the observer and which is approaching, and therefore we can determine which node is ascending and which is descending, and the sign of the inclination. Thus we can determine some things for a spectroscopic binary that we cannot determine for a visual binary, and *vice versâ*. If a binary star is both spectroscopic and visual (by which I mean that we can see the two components separately, and we can detect the periodic changes in radial velocity from the spectra of each), then we can determine almost anything we wish about the orbits without ambiguity. But such systems are rare – and valuable. Usually (unless the system is very close to us) the linear separation between the pairs of a visual binary is very large (that's why we can see them separately) and so the speeds of the stars in their orbits are too slow for us to measure the changes in radial velocity. Typically, orbital periods of visual binary stars are of the order of years – perhaps many years. Stars whose binarity is detected spectroscopically are necessarily moving fast (typically their orbital periods are of the order of days), which means they are close together – too close to be detected as visual binaries.

Of course, in addition to the periodic variations in radial velocity, which give rise to periodic Doppler shifts in the spectra, the system as a whole may have a radial velocity towards or away from the Sun. The radial velocity of the system – or its centre of mass – relative to the Sun is called, naturally, the *systemic velocity*, and is one of the things we should be able to determine from spectroscopic observations. I shall be using the symbol V_0 for the systemic velocity, though I have seen some authors use the symbol g and even refer to it as the "gamma velocity". [By the way have you noticed the annoying tendency of the semi-educated these days to use technical words that they don't know the meaning of? An annoying example is that people often talk of "systemic discrimination", presumably because they think that the word "systemic" sounds scientific, when they really mean "systematic discrimination".] We must also bear in mind that the actual observations of the star are made not from the Sun, but from Earth, and therefore corrections must be made to the observed radial velocity for the motion of Earth around the Sun as well as for the rotation of Earth around its axis.

18.2: The Velocity Curve from the Elements

In this section, we calculate the velocity curve (i.e. how the radial velocity varies with time) to be expected from a star with given orbital elements. Of course, the practical situation is quite the opposite: we observe the velocity curve, and from it, we wish to determine the elements. We'll deal with that later.

I'm going to use the convenient phrase "plane of the sky" to mean a plane tangent to the celestial sphere, or normal to the line of sight from observer to the centre of mass of the system. The centre of mass C of the system, then, is stationary in the plane of the sky. The plane of the orbits of the two stars around their centre of mass is inclined at an angle *i* to the plane of the sky. I am going to follow the adventures of star 1 about the centre of mass C. And I am going to assume that Chapter 9 is all fresh in your mind!

The semi major axis of the orbit of star 1 about C is a_1, and the semi latus rectum $l_1 = a_1\left(1-e^2\right)$. The angular momentum per unit mass of star 1 about C is $r_1^2\dot{v} = \sqrt{GMl_1}$, where v is the true anomaly and $\mathrm{M} = m_2^3/(m_1+m_2)^2$. The orbital period P is given by $P^2 = \frac{4\pi^2}{GM}a_1^3$. The mean motion n is 2p/P, and hence $n^2 a_1^3 = GM$. Therefore the angular momentum per unit mass is

$$r_1^2\dot{v} = na_1^2\sqrt{1-e^2}. \tag{18.2.1}$$

In figure XVIII.1 we see the star 1 (labelled S) in orbit around C, and at some time the *argument of latitude* of S is θ, and its distance from C is r_1. Its distance above the plane of the sky is *z*, and r_1 sin β. The inclination of the plane of the orbit to the plane of the sky is *i*, and, in order to find an expression for β in terms of the argument of latitude and the inclination, I'm just going to draw, in figure VIII.2, these angles on the surface of a sphere. The sphere is centred at C, and is of arbitrary radius.

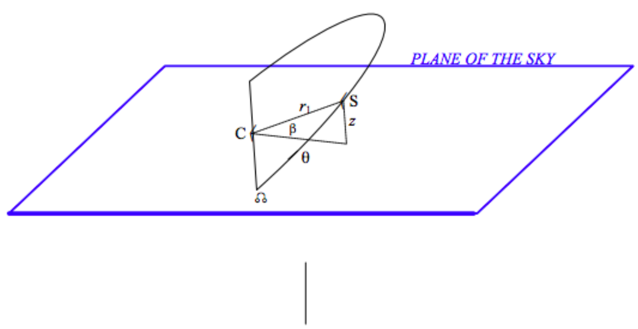


FIGURE XVIII.1

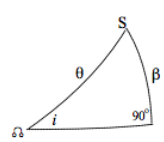


FIGURE XVIII.2

We see from this triangle that sin β = sin *i* sin θ. Also, the argument of latitude θ = ω + *v*, (ω = argument of periastron, *v* = true anomaly), and so

$$z = r_1 \sin i \sin(\omega+v). \tag{18.2.2}$$

At this moment, the radial velocity *V* of star 1 relative to the Sun is given by

$$V = V_0 + \dot{z}, \tag{18.2.3}$$

where V_0 is the radial velocity of the centre of mass C, or the systemic velocity. Differentiation of Equation 18.2.1 with respect to time gives

$$\dot{z}=\sin[\dot{r}_1\sin(\omega+v)+r_1\dot{v}\cos(\omega+v)]. \tag{18.2.4}$$

I would like to express this entirely in terms of the true anomaly v instead of v and r_1. The Equation to the ellipse is

$$r_1=\frac{l_1}{1+e\cos v}=\frac{a_1\left(1-e^2\right)}{1+e\cos v}, \tag{18.2.5}$$

where l_1 is the semi latus rectum, and so

$$\dot{r}_1=\frac{l_1 e\dot{v}\sin v}{(1+e\cos v)^2}=\frac{r_1 e\dot{v}\sin v}{1+e\cos v}. \tag{18.2.6}$$

which helps a bit. Thus we have

$$\frac{\dot{z}}{\sin i}=r_1\dot{v}\left(\frac{e\sin v\sin(\omega+v)}{1+e\cos v}+\cos(\omega+v)\right). \tag{18.2.7}$$

We can also make use of Equation 18.2.1, and, with some help from Equation 18.2.5, we obtain

$$\frac{\dot{z}}{\sin i}=\frac{na_1(1+e\cos v)}{\sqrt{1-e^2}}\left(\frac{e\sin v\sin(\omega+v)}{1+e\cos v}+\cos(\omega+v)\right) \tag{18.2.8}$$

or

$$\frac{\dot{z}}{\sin i}=\frac{na_1}{\sqrt{1-e^2}}(e\sin v\sin(\omega+v)+(1+e\cos v)\cos(\omega+v)). \tag{18.2.9}$$

Now $e\sin v\sin(\omega+v)+e\cos v\cos(\omega+v)=e\cos\omega$,so we are left with

$$\dot{z}=\frac{na_1\sin i}{\sqrt{1-e^2}}(\cos(\omega+v)+e\cos\omega). \tag{18.2.10}$$

The quantity $\frac{na_1\sin i}{\sqrt{1-e^2}}$, which has the dimensions of speed, is generally given the symbol K_1, so that

$$\dot{z}=K_1(\cos(\omega+v)+e\cos\omega), \tag{18.2.11}$$

and so the radial velocity (including the systemic velocity) as a function of the true anomaly and the elements is given by

$$V=V_0+K_1(\cos(\omega+v)+e\cos\omega). \tag{18.2.12}$$

You can see that $\dot{z}$ varies between $K_1(1\ e\cos\omega)$ and $-K_1(1-e\cos\omega)$, and that K_1 is the *semi-amplitude of the radial velocity curve*.

Equation 18.2.12 gives the radial velocity as a function of the true anomaly. But we really want the radial velocity as a function of the time. This is easy, or at least straightforward, because we already know how to calculate the true anomaly as a function of time. I give here the relevant Equations. I have retained their original numbering, so that you can locate them in the earlier chapters.

$$M=\frac{2\pi}{P}(t-T) \tag{18.2.13}$$

$$M=E-e\sin E \tag{18.2.14}$$

$$\cos v=\frac{\cos E-e}{1-e\cos E} \tag{18.2.15}$$

From trigonometric identities, this can also be written

$$\sin v=\frac{\sqrt{1-e^2}\sin E}{1-e\cos E} \tag{18.2.16}$$

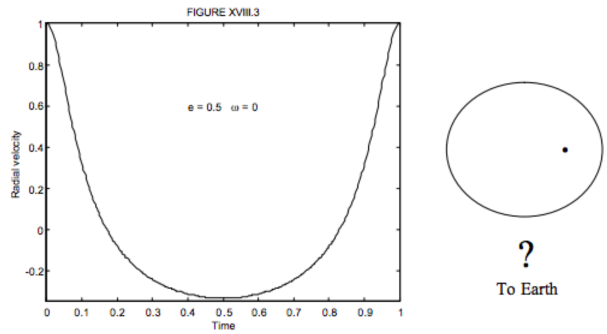
or

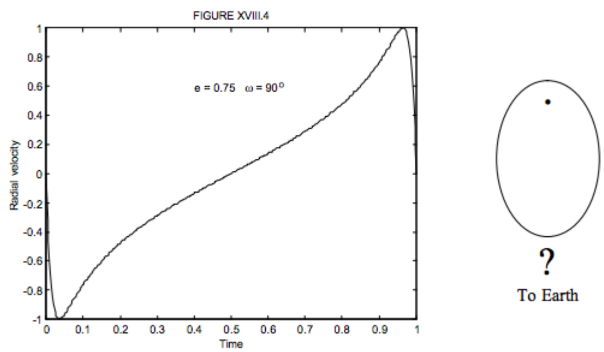
$$\tan v=\frac{\sqrt{1-e^2}\sin E}{\cos E-e} \tag{18.2.17}$$

or

$$\tan \frac{1}{2} v = \sqrt{\frac{1+e}{1-e}} \tan \frac{1}{2} E \tag{18.2.18}$$

I show in figures XIII.3 and XIII.4 two examples of velocity curves. Figure XIII.3 is computed for e = 0.5, ω = 0$^{\text{o}}$. Figure XIII.4 is computed for e = 0.75, ω = 90$^{\text{o}}$.





In order to draw these two figures, it will correctly be guessed that I have written a computer program that will calculate Equations 9.6.4, 9.6.5, 2.3.16 and 18.2.12 in order, for the chosen values of e and ω. This is perfectly straightforward except that Equation 9.6.5, Kepler's Equation, requires some iteration. The solution of Kepler's Equation was discussed in Section 9.6. If I were seriously going to be interested in computing the orbits of spectroscopic binary stars I would at this stage use this program to generate and print out 360 radial velocity curves for 36 values of ω going from 0º to 350º and ten value of e going from 0.0 to 0.9. Then, when I had a real radial velocity curve of a real spectroscopic binary star to analyse, I would be able to compare it with my set of theoretical curves and hence be able to get a least a rough first approximation to the eccentricity and argument of periastron.

I have drawn figures XVIII.3 and 4 for a *systemic velocity* V_0 of zero. A real star will not have a zero systemic velocity and indeed one of the aims must be to determine the systemic velocity.

Thus in figure XVIII.5 I have drawn a radial velocity curve (I'm not saying what the values of ω and e are), but this time I have not assumed a zero systemic radial velocity. It will be noticed that the observed star spends much longer moving towards us than away from us. If we draw a horizontal line *Radial Velocity* = V_0 across the figure, this line must be drawn such that the area between it and the radial velocity curve above it is equal to the area between it and the radial velocity curve below it. How to position this line? That is a good question. If nothing else, you can count squares on graph paper. That at least will give you a first rough idea of what the systemic velocity is.

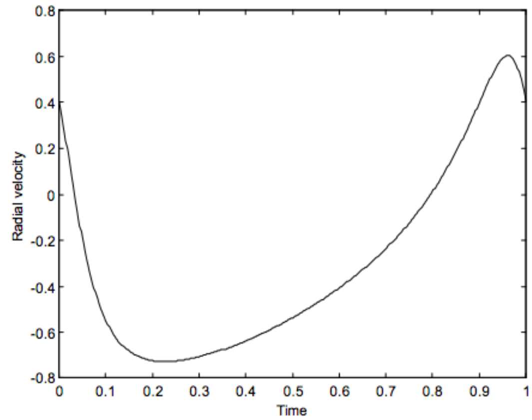


FIGURE XVIII.5

If you have a *double-lined* binary, you will have two radial velocity curves. They are not quite mirror images of each other; the semiamplitude of each component is inversely proportional to its mass. But the systemic velocity is then easy, because the two curves cross when the radial velocity of each is equal to the radial velocity of the system.

18.3: Preliminary Elements from the Velocity Curve

We have seen in the previous section how to calculate the velocity curve given the elements. The more practical problem is the inverse: In this section, we assume that we have obtained a velocity curve observationally, and we want to determine the elements. The assumption that we have obtained a precise radial velocity curve is, of course, rather a large one; but, for the present, let us assume that this has been done and we are trying to determine what we can about the orbit. We limit ourselves in this section to determining from the curve only very rough first estimates of the elements. This will also serve the purpose of establishing what information is *obtainable in principle* from the velocity curve. A later section will deal with refining our estimates and obtaining precise values.

The assumption that we have already obtained the radial velocity curve implies that we already know the period P of the orbit.

The radial velocity curve is given by Equation 18.2.12:

$$V = V_0 + K_1(\cos(\omega + v) + e\cos\omega). \tag{18.3.1}$$

Here $v = v(t, T, e)$. Thus, from the radial velocity curve, we should be able to determine V_0, K_1, e, ω and T. We shall remind ourselves a little later of the meaning of K_1, but in the meantime we can note that the radial velocity varies between a maximum of $V_{max} = V_0 + K_1 (e \cos \omega + 1)$ and a minimum of $V_{min} = V_0 + K_1 (e \cos \omega - 1)$. The difference between these two is $2K_1$. Thus K_1 is the *semiamplitude of the radial velocity curve*, regardless of the shape of the curve and the values of ω and e, and so (again assuming that we have a well-determined radial velocity curve) K_1 can be readily determined.

The systemic velocity V_0 is such that the area under the radial velocity curve above it is equal to the area above the radial velocity curve below it. Thus at least a rough preliminary estimate can be made of V_0, regardless of the shape of the curve and of the values of ω and e.

The shape of the radial velocity curve (as distinct from its amplitude and phase) is determined by ω and e. As suggested in the previous section, we can prepare a set of, say, 360 theoretical curves covering 36 values of ω from 0 to 350º and 10 values of e from 0.0 to 0.9. (By making use of symmetries, one need cover ω only from 0 to 90º, but computers are so fast today that one might as well go from 0 to 350º) By comparing the observed curve with these theoretical curves, we get a first estimate of ω and e. We could then I suppose, take advantage of today's fast computers and prepare a set of velocity curves with much finer intervals around one's first estimate. This would not, of course, allow us to calculate definitive precise values of ω and e, but it would give us a pretty good first guess.

I have already pointed out that

$$V_{\max} = V_0 + K_1(e\cos\omega + 1) \tag{18.3.2}$$

and

$$V_{\min} = V_0 + K_1(e\cos\omega - 1) \tag{18.3.3}$$

From these we see that

$$e\cos\omega = \frac{V_{\max} + V_{\min}}{2K_1}. \tag{18.3.4}$$

This allows us to determine $e\cos\omega$ without reference to the slightly uncertain V_0, and we will want to see that our estimates of e and ω from the shape of the curve are consistent with Equation 18.3.3.

The velocity curve also allows us to determine T, the time of periastron passage. For example, the sample theoretical velocity curves I have drawn in figures XIII.3, 4 and 5 all start at periastron at the left hand limit of each curve.

Note that we have been able to determine K_1, which is $\frac{na_1 \sin i}{\sqrt{1-e^2}}$, and we can determine e and n, which is $2\pi/P$. This means that we can determine $a_1 \sin i$, but that is as far as we can go without additional information; we cannot separate a_1 from i.

18.4: Masses

In Section 18.3 we saw that we could obtain approximate values of P, V_0, K_1, e, ω and T. But, apart from its being the semi-amplitude of the velocity curve, we have forgotten the meaning of K_1. We remind ourselves. It was defined just after Equation 18.2.10 as

$$K_1 = \frac{na_1 \sin i}{\sqrt{1-e^2}}. \tag{18.4.1}$$

Here n is the mean motion $2\pi/P$. Thus, since we know P (hence n), e and K_1, we can determine $a_1 \sin i$ – but we cannot determine a_1 or i separately.

Now the mean motion n is given just before Equation 18.2.1 as

$$n^2 a_1^3 = GM, \tag{18.4.2}$$

where

$$M = m_2^3/(m_1+m_2)^2. \tag{18.4.3}$$

(A reminder: The subscript 1 refers, for a single-lined binary, to the star whose spectrum we can observe, and the subscript 2 refers to the star that we cannot observe.) All of this put together amounts to

$$K_1 = \frac{G}{(1-e^2)\,a_1 \sin i} \times \frac{m_2^3 \sin^3 i}{(m_1+m_2)^2}. \tag{18.4.4}$$

Thus we can determine the *mass function* $\frac{m_2^3 \sin^3 i}{(m_1+m_2)^2}$. We cannot determine the separate masses, or their ratio or sum, or the inclination.

In recent years, it has become possible to measure very small radial velocities of the order of a few metres per second, and a number of single-lined binary stars have been detected with very small values of K_1; that is to say, very small radial velocity amplitudes. These could, of course, refer to stars with small orbital inclinations, so that the plane of the orbit is almost perpendicular to the line of sight. It has been held, however, (on grounds that are not entirely clear to me) that many of these single-lined binary stars with small radial velocity variations are actually single stars with a planet (or planets) in orbit around them. The mass of the star that we can observe (m_1) is very much larger than the mass of the planet, which we cannot observe (m_2). To emphasize this, I shall use the symbol M instead of m_1 for the star, and m instead of m_2 for the planet. The mass function that can be determined is, then

$$\frac{m^3 \sin^3 i}{(M+m)^2}.$$

If m (the mass of the unseen body – the supposed planet) is very much smaller than the star (of mass M) whose radial velocity curve has been determined, then the mass function (which we can determine) is just

$$\frac{m^3 \sin^3 i}{M^2}.$$

And if, further, we have a reasonable idea of the mass M of the star (we know its spectral type and luminosity class from its spectrum, and we can suppose that it obeys the well-established relation between mass and luminosity of main-sequence stars), then we can determine $\mathrm{m}^3 \sin^3 i$ and hence, of course $m \sin i$. It is generally recognized that we cannot determine i for a spectroscopic binary star, and so it is conceded that the mass of the unseen body (the supposed planet) is uncertain by the unknown factor sin i.

However, the entire argument, it seems to me, is fundamentally and rather blatantly unsound, since, in order to arrive at $m \sin i$ and to hence to claim that m is of typically planetary rather than stellar mass, *the assumption that m is small and i isn't has already been made in approximating the mass function by* $\frac{m^3 \sin^3 i}{M^2}$. Unless there is *additional evidence* of a *different kind*, the observation of a velocity curve of small amplitude is not sufficient to indicate the presence of an unseen companion of planetary mass. Equally well (without additional evidence) the unseen companion could be of stellar mass and the orbital inclination could be small.

If the system is a *double-lined* spectroscopic binary system, we can determine the mass function for each component. That is, we can determine $\frac{m_1^3 \sin^3 i}{(m_1+m_2)^2}$ and $\frac{m_2^3 \sin^3 i}{(m_1+m_2)^2}$. The reader should now convince him- or herself that, since we now know these two mass functions, we can determine the *mass ratio* and we can also determine $m_1 \sin^3 i$ and $m_2 \sin^3 i$ separately. But we cannot determine m_1, m_2 or i.

18.5: Refinement of the Orbital Elements

By finding the best fit of the observational values of radial velocity to a set of theoretical radial velocity curves, we have by now determined, if only graphically, a preliminary estimate of the orbital elements. We now have to refine these estimates in order to obtain the best set of elements that we can from the data.

Let us remind ourselves of the theoretical Equation (Equation 18.2.12) that we developed for the radial velocity:

$$V=V_0+K_1(\cos(\omega+V)+e\cos\omega). \tag{18.5.1}$$

Here

$$K_1=\frac{na_1\sin i}{\sqrt{1-e^2}} \tag{18.5.2}$$

and

$$n=2\pi/P. \tag{18.5.3}$$

Also v is a function of the time and the elements T and e, through Equations 9.6.4, 9.6.5 and 2.3.16 cited in Section 18.2. Thus Equation 18.5.1 expresses the radial velocity as a function of the time (hence true anomaly) and of the orbital elements V_0, K_1, ω, e, n and T:

$$V=V\left(t;V_0,K_1,\omega,e,n,T\right). \tag{18.5.4}$$

For each observation (i.e for each time t), we can use our preliminary elements to calculate what the radial velocity should be at that time, and compare it with the observed radial velocity at that time. Our aim is going to be to adjust the orbital elements so that the sum of the squares of the differences V_{obs} - V_{calc} is least.

If we were to change each of the elements of Equation 18.4.4 by a little, the corresponding change in V would be, to first order,

$$\delta V=\frac{\partial V}{\partial V_0}\delta V_0+\frac{\partial V}{\partial K_1}\delta K_1+\frac{\partial V}{\partial\omega}\delta\omega+\frac{\partial V}{\partial e}\delta e+\frac{\partial V}{\partial n}\delta n+\frac{\partial V}{\partial T}\delta T. \tag{18.5.5}$$

When the differentiations have been performed, this becomes

$$\begin{aligned}&\delta V=\delta V_0+(\cos(V+\omega)+e\cos\omega)\delta K_1-K_1(\sin(V+\omega)+e\sin\omega)\delta\omega\\&+K_1\left(\cos\omega-\tfrac{(2+e\cos v)\sin(V+\omega)\sin v}{1-e^2}\right)\delta e\\&-\tfrac{\sin(v+\omega)(1+e\cos v)^2K_1(t-T)}{(1-e^2)^{3/2}}\delta n+\tfrac{K_1n\sin(v+\omega)(1+e\cos v)^2}{(1-e^2)^{3/2}}\delta T\end{aligned}\quad. \tag{18.5.6}$$

In this Equation, δV is V_{obs} - V_{calc}. There will be one such Equation for each observation, and hence, if there are N (> 6) observations there will be N *Equations of condition*. From these, six *normal Equations* will be formed in the manner described in Section 1.8 and solved for the increments in the orbital elements. These are then subtracted from the preliminary elements to form an improved set of elements, and the process can be repeated until there is no significant change.

This process can be highly automated by computer, but in practice the calculation is best overseen by an experienced human orbit computer. While a computer may produce a formal solution, there are a number of situations that may result in a solution that is unrealistic or even quite wrong. Much depends on the distribution of the observations, and on whether the observational errors are normally distributed. Also, if the system has been observed for a long time over many orbital periods, the period may be known to great precision, and the investigator may prefer to keep P (hence n) as a fixed, known constant during the calculation. Or again, if the period is short, the investigator may wish (perhaps on the basis of additional knowledge) to suppose that the two stars are close together and that the orbits of the components are circular, and hence fix e = 0 throughout the calculation. I am always a little uneasy about making an assumption that some element has some desired value; it seems to me that, once one starts this, one might as well assume values for *all* of the elements. This would have the advantage that one need not make any observations or do any calculations and can just assume all the results according to personal taste. Whether an assumption that P or e can be held as fixed and known, or whether one should let the computer do the entire calculation without any intervention, is something that requires the experience of someone who has been calculating orbits for years.

18.6: Finding the Period

The first five sections of this chapter have dealt with calculating the relations between the orbital elements and the radial velocity curve, and that really completes what is necessary in a book whose primary focus is on celestial mechanics. In practice, the celestial mechanics part is the least of the difficulties. The Equations may look forbidding at first sight, but at least the Equations are unambiguous and clear cut. There are lots of problems of one sort of another that in practice occupy much more of the investigator’s time than merely the computation of the orbit, which nowadays is done in the blink of an eye. I mention a few of these only briefly in the remaining sections, partly because they are not particularly concerned with celestial mechanics, and partly because my personal practical experience with them is limited.

If you were able to measure the radial velocity every five minutes throughout a complete period, there would be no difficulty in obtaining a nice velocity curve. In practice, however, you measure a radial velocity “every so often” – with perhaps many orbital periods between consecutive observations. Finding the period, then, is obviously a bit of a problem. (That there is an initial difficulty in finding the period is ultimately compensated for in that, once a preliminary value for the period is found, it can often be calculated to great precision, if the star has been observed over many decades.)

If you have a large number of observations spread out over a long time, it may be possible to identify several observations in which the radial velocity is a maximum, and you might then assume that the least time between consecutive maxima is an integral number of orbital periods. Of course you don’t know what this integral number is, but you might be able to do a little better. For example, you might find that there are 100 days between two consecutive maxima, so that there are an integral number of periods in 100 days. You might also find that two other maxima are separated by 110 days. You now know that there are an integral number of periods in 10 days – which is a great improvement.

A difficulty arises if you observe the star at regular and equal intervals. While there is an obvious answer to this – *i.e.* don’t do it – it may not in practice be so easy to avoid. For example: if you always observe the star when it is highest in the sky, on the meridian, then you are always observing it at an integral number of sidereal days. You then get a stroboscopic effect. Thus, if you have a piece of machinery that is cycling many times per second, you can illuminate it stroboscopically with a light that flashes periodically, and you can then see the machinery moving apparently much more slowly than it really is. The same thing happens if you observe a spectroscopic binary star at precisely regular intervals – it will appear to have a much longer period that is really the case.

It is easier to understand the effect if we work in terms of frequency (reciprocal of the period) rather than period. Thus let n (= $1/P$) be the orbital frequency of the star and let n' (= $1/T$) be the frequency of observation (the frequency of the stroboscope flash, to recall the analogy). Then the apparent orbital frequency ν' of the star is given by

$$|\nu' - \nu| = mn \tag{18.6.1}$$

where m is an integer. Returning to periods, this means that you can be deceived into deducing a spurious period P' given by

$$\frac{1}{P'} = \frac{1}{P} \pm \frac{1}{mn}. \tag{18.6.2}$$

You don’t have to make an observation every single sidereal day to experience this stroboscopic effect. If your stroboscope is defective and it misses a few flashes, the machinery will still appear to slow down. Likewise, if you miss a few observations, you may still get a spurious period.

Once you have overcome these difficulties and have determined the period, in order to construct a radial velocity curve you will have to subtract an integral number of periods from the time of each observation in order to bring all observations on to a single velocity curve covering just one period.

18.7: Measuring the Radial Velocity

In a text primarily concerned with celestial mechanics, I shan't attempt to do justice to the practical details of measuring a spectrum, but one or two points are worth mentioning, if only to draw the reader's attention to them.

To measure the radial velocity, you obtain a spectrum of the star and you measure the wavelength of a number of spectrum lines (i.e. your measure their positions along the length of the spectrum) and you compare the wavelengths with the wavelengths of a comparison laboratory spectrum, such as an arc or a discharge tube, adjacent to the stellar spectrum. If the spectra are obtained on a photographic plate, the measurement is done with a measuring microscope. If they are obtained on a CCD, there is really no "measurement" in the traditional sense to be done – a computer will read the pixels on which the lines fall. If the stellar lines are displaced by $\Delta\lambda$ from their laboratory values λ, then the radial velocity v is given simply by

$$\frac{v}{c} = \frac{\Delta\lambda}{\lambda}. \tag{18.7.1}$$

Note that this formula, in which c is the speed of light, is valid only if $v << c$. This is certainly the case in the present context, though it is not correct for measuring the radial velocities of distant galaxies. (The z in the galaxy context is the measured $\Delta\lambda/\lambda$, and knowledge of both relativity and cosmology is necessary to translate that correctly into radial velocity.)

The accurate measurement of wavelengths in stellar spectra has its own set of difficulties. For example, the spectrum lines of early type stars are broad and diffuse as a result of the high temperatures and quadratic Stark broadening of the lines, as well as the rapid rotation of early type stars. The lines of late-type stars are numerous, closely crowded together and blended. Thus there are difficulties at both ends of the spectral sequence.

One very nice technique for measuring radial velocities involves making use of the entire spectrum rather than the laborious process of measuring the wavelengths of individual lines. Suppose that you are, for example observing a G-type star. You will prepare an opaque mask on which are inscribed, in their correct positions, transparent lines corresponding to the lines expected of a G-type star. During observation, the spectrum of the star is allowed to fall on this mask. Some light gets through the transparent inscribed lines on the mask, and this light is detected by a photoelectric cell behind the mask. The mask is moved parallel to the spectrum until the dark absorption lines in the stellar spectrum fall on the transparent inscribed lines on the mask, and at this moment the amount of light passing through the mask and reaching the photoelectric cell reaches a sharp minimum. Not only does this technique make use of the whole spectrum, but the radial velocity is obtained immediately, *in situ*, at the telescope.

I end by briefly mentioning two little problems that are well known to observers, known as the *rotation effect* and the *blending effect*.

If the orbital inclination is close to 90°, the system, as well as being a spectroscopic binary, might also be an eclipsing binary. In this case, we can in principle get a great deal of information about the system – but there is a danger that the information might not be correct. For example, suppose that the system is a single-lined binary, and that the bright star (the one whose spectrum can be seen) is a rapid rotator and is being partially eclipsed by the secondary. In that case we can see only part of the surface of the primary star – perhaps that part of the star that is (by rotation) moving towards us. This will give us a wrong measurement of the radial velocity.

Or again, suppose that we have a double-lined binary. For much of the orbital period, the lines from one star may be well separated from those of the other. However, there comes a time when the two sets of lines approach each other and become partially blended. I show in figure XVIII.6 two partially blended gaussian profiles. You will see that the minima of the blended profile, shown as a dashed curve, occur closer together than the true minima of the individual lines. If you measure the minima of the blended profile, this will obviously give the wrong radial velocity and will result in a distortion of the velocity curve and corresponding errors in the orbital elements. Many years ago I made some calculations on the amount of the blending effect for gaussian and lorentzian profiles for various separations and relative intensities. These calculations were published in *Monthly Notices of the Royal Astronomical Society*, **141**, 43 (1968).

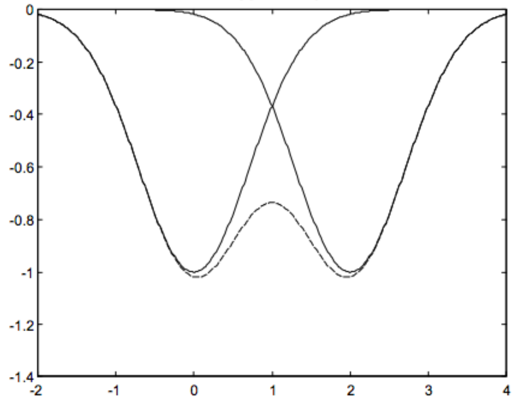


FIGURE XVIII.6
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-2
-1
0
1
2
3
4

Index