A Visit to the Newtonian N-body Problem via Elementary Complex Variables

DONALD G. SAARI, *Northwestern University, Evanston, Il 60208*

DON SAARI received his undergraduate degree from Michigan Technological University and his Ph.D. in mathematics from Purdue University. His Ph.D. dissertation, written under the guidance of Harry Pollard, concerned the qualitative behavior of dynamics in the general Newtonian N-body problem. After a postdoctoral position in the Yale Astronomy Department, Saari moved to Northwestern University where he currently is Professor of Mathematics. His main research interests revolve around applications of dynamical systems to celestial mechanics and physics and to issues coming from the social sciences.

The study of how N celestial bodies move under gravitational forces is an old one. If one is willing to acknowledge the work of the ancient astrologers and shepherds—two groups that carefully plotted the positions of the stars and planets —then this subject area traces its origins to the earliest reaches of mankind. Indeed, had the title not been already preempted, one might suggest that the study of the N-body problem is "the world's oldest profession." If it isn't the oldest, then, most surely, it is "the second oldest."

How do the heavenly bodies move? With the help of elementary complex variables, certain selective orbits will be described where, as it turns out, even "simple" motion can be surprisingly complicated. Also, some of the history of the Newtonian N-body problem will be related with an emphasis on the myth that only the two body problem has been solved. For practical purposes this assertion is correct, but at the turn of the century the Finnish mathematician K. Sundman "solved" the three body problem in an accepted sense of that time. While explaining why Sundman's result isn't widely known, it will be indicated how collisions, both the real type where two or more particles hit each other and the complex type where imaginary collisions occur at complex values of time, affect the behavior of the system.

To start, consider the question that, at some time, probably bothered many of us. Namely, why did the early astronomers have so much trouble predicting the motion of the planets? Armed with a 20th-century education, we know that a reasonable approximation for the motion of a planet is uniform circular motion about the Sun. What is so difficult about describing motion as predictable as this? This question is closely related to the fable where Galileo silently disavowed his forced recanting of a Sun-centered solar system. Probably generations of irreverent school children silently wondered, "Who cares? What difference does it make if the Sun or if the Earth is the center?" It does matter. This simple change of variables introduces a significantly different perspective of the solar system, it explains the difficulties faced by the astronomers of antiquity, and it underscores the critical importance of the Copernican revolution.

To understand why an Earth-centered prejudice creates problems, consider a simplified version of the Sun-Earth-Mars system. Mars is approximately $3/2$ times as far from the Sun as the Earth (1.524 times) and it takes approximately 2 years

(687 Earth days) to complete one revolution of the Sun. To eliminate fractions, replace the traditional astronomical unit with half-astronomical units so that, in the new system, the Earth is distance 2 from the Sun. In this simplified model, the Earth's position is $z_E(t) = 2e^{2\pi it}$ where t is measured in "Earth years," while that of Mars is $z_M(t) = 3e^{\pi it}$. The position of Mars as observed from Earth is

$$z(t) = z_M(t) - z_E(t) = 3e^{\pi it} - 2e^{2\pi it}. \tag{1.1}$$

A convenient way to describe this orbit is with the translation

$$z(t) - 2 = 3e^{\pi it} - 2e^{2\pi it} - 2 = e^{\pi it}(3 - 2e^{\pi it} - 2e^{-\pi it}) = (3 - 4\cos \pi t)e^{\pi it}$$

The graph of this equation—the polar coordinate representation of a limaçon with a loop—depicts the orbit of Mars relative to Earth.[1] (See FIGURE 1.)

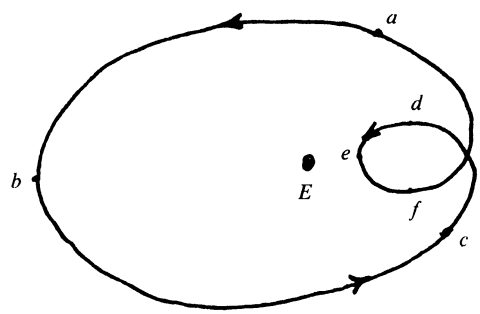

FIG. 1.

It is clear from FIGURE 1 why the pre-Copernican, Earth-centered prejudice made it so difficult to predict the motion of the planets and to develop a "Newtonian Theory." While the orbit from a to b to c is not overly complex, it becomes quite complicated once Mars passes through point c. Mars continues in a counterclockwise direction until point d where it appears to *stop* and then change to a clockwise motion until point f. At f, Mars again reverses direction to return to a counterclockwise motion. Imagine what complications this motion presented to astronomers trying to predict positions. Next, imagine a theorist trying to determine the governing laws of motion. While Mars goes from a to b to c to d, a theorist might persuasively argue for a law of attraction. But, how does one explain the orbit from e to f to a? How does one justify the sudden law of repulsion? What is there about the Earth that drives Mars away?

Actually, the motion of Mars in FIGURE 1 is simpler than that of the other planets. Similar elementary complex variable arguments prove that the orbits of Saturn, of Jupiter, and of the other planets that take much longer to circle the Sun, admit several loops similar to the one in FIGURE 2. This can be proved as above, but

[1] The trigonometric version of this argument involves nothing more difficult than the double angle formulas, so it can be used to motivate several topics in calculus and precalculus classes. For instance, I find that this description of the orbit of Mars serves as a more persuasive illustration of the relevance of limaçons and cardioids than many of the standard examples used in calculus.

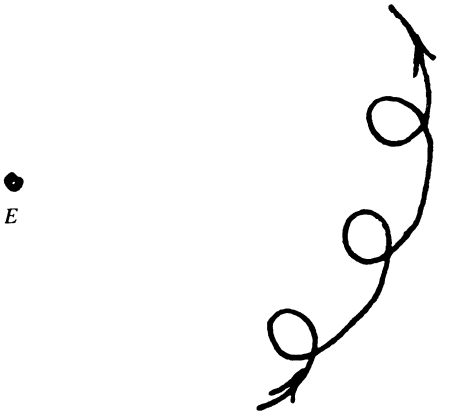

FIG. 2.

now the limaçon is relative to a rotating circle. (Instead of finding the orbit relative to a fixed translation, find it relative to a rotating one chosen so that the exponential terms collapse to a cosine term.) As an alternative way to demonstrate this, express the relative orbit, $z(t) = ae^{\alpha\pi it} - 2e^{2\pi it}$, $a > 2$, in the standard complex form of $z(t) = r(t)e^{i\theta(t)}$. The direction of the motion, whether it is clockwise or counterclockwise, is determined by the sign of θ'. This is the imaginary part of $(\ln z(t))' = z'/z = r'/r + i\theta'$. But $z'/z = \pi i(a\alpha - 4e^{(2-\alpha)\pi it})/(a - 2e^{(2-\alpha)\pi it})$, so after an elementary computation, it follows that the sign of θ' changes if $a\alpha < 4$. This inequality holds because Kepler's third law asserts that $a^3\alpha^2 = k$ where k is a constant; thus $a\alpha = (k/a)^{1/2}$ is a decreasing function of a.

Pictures of these looping orbits depicting the motion of the distant planets relative to the earth can be found in several books on the history of astronomy. Even more impressive are the photos of the planets taken over a several year period that are superimposed on one plate. When one compares this kind of dynamical behavior with the uniform motion obtained through a Sun centered system, it becomes clear why the Copernican change of variables has had such a profound, simplifying impact on science and astronomy.

The Copernican representation significantly simplifies the description of the motion of the planets, but is it a correct theory? This is the essence of Cardinal Barberini's query when he contested Galileo (at least in Brecht's play *Life of Galileo*), "*Are you sure... you astronomers aren't just trying to make astronomy a little easier for yourselves?... You like to think in circles or ellipses and in uniform velocities, in simple motions commensurate with your minds. But what if God had been pleased to make His stars move like this?*" where Barberini moves his finger through the air in a complicated course that presumably resembles FIGURE 2. Responding with his version of Occam's razor, Galileo argued for Copernicus's theory by asserting, "*if God had created the world like this* [He retraces Barberini's course.] *He would have constructed our minds like this too* [He repeats the same course.] *to enable them to recognize these courses as the simplest. I believe in reason.*"

Elementary complex variables disclose these somewhat surprising, relative orbits for the outer planets by capturing the effects of the rotating coordinate system defined by the motion of the Earth. There are other natural rotating systems in

astronomy, so one might correctly suspect that a related approach would exhibit other surprising astronomical behavior. To demonstrate this, consider the planet Mercury. It takes Mercury about 90 Earth days (87.967) to circle the Sun. Based on observations, it was believed until the 1960s that Mercury took about the same length of time to rotate on its axis. If this were true, then, as asserted in most of the older texts on astronomy, the same face of Mercury always would face the Sun—a solar day (or solar night) on Mercury would last forever. However, with radar observations, we now know that Mercury takes about 60 Earth days to rotate on its axis; this shortens a Mercury solar day to about 176 Earth days, or two Mercury years. Moreover, as shown next, when a more accurate representation for the orbit of Mercury is used, an observer on Mercury would find the apparent orbit of the Sun to be quite peculiar.

To obtain a sharper approximation for the orbit of Mercury, or any other planet about the Sun, treat the orbit as an ellipse with eccentricity ε; e.g., for the Earth, $\varepsilon = 0.0167$, while for Mercury, $\varepsilon = 0.2056$. The position of the planet on the ellipse is given by

$$r(\theta) = a/(1 - \varepsilon\cos(\theta)) \approx a(1 + \varepsilon\cos(\theta)), \tag{1.2}$$

where a is a positive constant and $\theta(t)$ is the angle of the planet determined by a reference line. Equation 1.2 has the complex representation $z(\theta) \approx a(1 + \varepsilon\cos(\theta))e^{i\theta} = a(1 + (\varepsilon/2)e^{i\theta} + (\varepsilon/2)e^{-i\theta})e^{i\theta}$, or

$$z(\theta) \approx a\varepsilon/2 + ae^{i\theta} + (a\varepsilon/2)e^{2i\theta}. \tag{1.3}$$

The rotation of Mercury is given by $e^{12\pi i}$. It follows that the apparent position of the Sun on Mercury is

$$Z(t) = -z(\theta(t))e^{-12\pi it}, \tag{1.4}$$

so

$$\text{Arg}(Z(t)) = \text{Arg}(z(\theta(t)) - 12\pi t + \pi = \theta(t) - 12\pi + \pi. \tag{1.5}$$

The apparent motion of the Sun changes directions whenever $\text{Arg}(Z(t))'$ changes sign. This happens, and the reason is based on Kepler's second law, which asserts that $r^2\theta'$ is constant valued. Thus, if ε is sufficiently large, as it is for Mercury, $\theta(t)$ cannot be represented by uniform motion. In particular, when r is at perihelion (i.e., $r(t) = |z(t)|$ is at a minimum), θ' is at its maximum value. Indeed, using Kepler's second law and eq. 1.2, we have that $\theta' = n/(1 - \varepsilon\cos(\theta))^2$ where n is the mean motion; e.g., for Mercury, $n = 8\pi$. Consequently, at perihelion, $\theta' = 8\pi/(0.7944)^2 > 12\pi$. This same argument shows that $\text{Arg}(Z(t))' \geqslant 0$ if $|\theta| \leqslant 26.8°$; otherwise it is negative. From this, a schematic representation of the apparent motion of the Sun, given in FIGURE 3, is easy to determine. Each of the two loops corresponds to Mercury's "yearly" passage through perihelion. (A similar argument shows that the apparent motion of the Sun would have no reversal of direction had the orbit of Mercury been circular enough so that $\varepsilon < 0.1835$.)

The change in the sign of $\text{Arg}(Z(t))$ creates an interesting phenomenon. There are locations on Mercury where after the Sun rises in the east on a Mercury morning, it reaches only a certain point in the sky before it *stops* to retrace its motion and to set in the *east*. Then, the Sun rises a second time this "morning," but this time it progresses normally throughout the long day. That's not all; after it sets

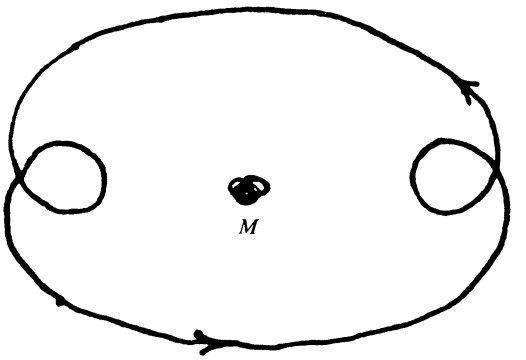

FIG. 3.

in the *west*, this changeable Sun reverses behavior once more to rise again in the west for a brief time before it finally sets for the long Mercury night.

2. Epicycles. An important planetary theory was developed by Ptolemy. One can appreciate the genius of Ptolemy by examining the orbits depicted in FIGURES 1 and 2. Recall the problem facing him: a scientific theory of planetary motion needed to satisfy the governing prejudices of the time. For Ptolemy, this meant a theory needed to accommodate the Earth as the center of the solar system, and we've just seen the complexities associated with this assumption. The next obstacle, left over from Aristotle, was that the circle is the most perfect figure in geometry. Obviously, heavenly bodies are "incorruptible," so their motions should be described by uniform circular motion. But how does one build a predictive, planetary theory incorporating these assumptions?

In his *Almagest*, written around A.D. 130, Ptolemy invented the ingenious approach of epicycles. This is where the position of the particles satisfies Aristotle's constraint of uniform circular motion. The clever idea is that the point indicated by the motion of the first circle, the deferent, does *not* represent the location of the planet. It locates the center of a *second* circle spinning with uniform motion. The location of the planet is given by the moving point on the second circle—the epicycle. (See FIGURE 4.) Today this approach may seem to be hopelessly naive, but remember that variations of this theory dominated astronomy for more than a millennium—an incredibly long time for any scientific theory. Even Newton's

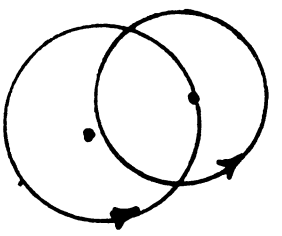

FIG. 4.

equations didn't enjoy such a long reign before being challenged by Einstein's relativity.

The long success of Ptolemy's approach can be understood by using complex variables. Let a_j be the radius of the jth circle where the period of the uniform motion is $1/2b(j)$, $j = 1, 2$. This means that the motion of a planet, as described through epicycles, is given by

$$z(t) = a_1 e^{b(1)\pi it} + a_2 e^{b(2)\pi it}. \tag{2.1}$$

From this equation, the source of the success of the epicycles becomes obvious. Namely, with the appropriate choices of a_j and $b(j)$, $j = 1, 2$, eq. 2.1 is the same as eq. 1.1. In other words, the epicycle approach can be viewed as recapturing the change of variables used to convert the Sun centered motion of a planet to an Earth centered one. With the appropriate choice of constants, this epicycle expression is the correct one for any planet.

As time marched on, the accuracy of Ptolemy's theory didn't always satisfy the increasing demands of astrology and astronomy. In part, this is because the actual orbits are on perturbed ellipses rather than on circles about the Sun. (Some of Ptolemy's orbits already incorporated elliptic behavior.) To achieve more accurate results, modifications of the same idea could be used. The approach is simple: instead of treating the location of the moving point on the second circle as the location of the planet, treat it as the center of a third (or fourth, or fifth, or ...) circle moving in uniform circular motion. The complex variable representation of these approximations is

$$z(t) = \sum_j a_j e^{b(j)\pi it} \tag{2.2}$$

where $j = 1, \ldots, k$, and k is the number of circles being used. As already illustrated by eq. 1.3, such an approach can achieve a higher degree of accuracy. Also, imagine how this never ending problem of determining the next values of a_k and $b(k)$ could serve as the source of an infinite number of Ph.D. dissertations.

Epicycles were abandoned long ago. We now know, after extensive development of sophisticated mathematical theories applied to Newton's equations, that there are situations where the orbits of planets are either quasi-periodic or almost periodic. What is quasi-periodic motion? It is motion represented by eq. 2.2, the epicycles, while almost periodic motion is the limit of this summation as $k \to \infty$.

3. Collisions and spinors. Today, Newton's equations are used to study the N-body problem. One approach to solving these equations might be to find the series solution. As we know from complex variables, the radius of convergence for this series is determined by the distance to the nearest singularity. Such a singularity appears to be a collision. (For $N \geqslant 4$, the situation is more complicated; see [**13**].) Thus, as a first step toward finding series solutions of Newton's equations, we need to understand the properties of collisions. For instance, one might ask whether it is possible to mathematically continue a solution through a collision? If so, this could extend the radius of convergence of a series.

To develop the behavior of collisions, start with the one body, or central force problem. If $\mathbf{r}(t)$ represents the vector position of the particle, then, with the

appropriate units of mass, time, and distance, the equations of motion are

$$\mathbf{r}(t)'' = -\mathbf{r}/r^3, \tag{3.1}$$

where $r = |\mathbf{r}|$. The solutions, given by eq. 1.2, are ellipses, parabolas, or hyperbolas. But, what happens near or at a collision? Can we define the motion through $r = 0$? This question was posed and answered by Sundman [**16**] in 1913. An improved, simplified theory was developed by the Italian geometer Levi-Civita [**4**] in 1920.

To see the ideas, start with a family of orbits that have a collision in the limit, i.e., a family represented by eq. 1.2 where ε tends to unity. Such a family is depicted in FIGURE 5. The key point is that when a particle has a "close approach," it rapidly spins around the central body. Therefore, if the dynamics can be mathematically extended beyond a collision, the colliding particle must make a 2π angular change —we should expect the particle to hit and then rebound from the central force. To remove this collision singularity from the equations of motion, this abrupt change needs to be removed by straightening out the orbit.

FIG. 5.

One way to see how to do this is to restrict the motion to the plane. Reexpress the position vector $\mathbf{r} = (x, y)$ as a complex variable $z = x + iy$, so the equations of motion are $z'' = -z/r^3$, where $r = |z|$. The collision occurs at $r = 0$, so this is where a change of variables is needed to convert the abrupt, 2π polar angle change into a form where the motion is on a straight line. To do this, the change of dependent variables must take half of the angular change at this point. With the complex variable representation of the coordinates, the appropriate change of variables is obvious; use the square root. The coordinate change is $w = u_1 + iu_2 = z^{1/2}$, or

$$w^2 = z. \tag{3.2}$$

The equations of motion for w, when accompanied with the change of independent variables $ds = dt/r(t)$ introduced by Sundman, not only are well defined at $w = 0$, a collision, but they assume the particularly simple form

$$\mathbf{u}'' + a\mathbf{u} = \mathbf{0} \tag{3.3}$$

where a is some positive constant and $\mathbf{u} = (u_1, u_2)$ is the vector representation for w. In other words, *this transformation converts the nonlinear Newtonian equations into the linear equations for a harmonic oscillator.*

To see how eq. 3.3 arises, note that the change of the independent variable defines the operator $d^2/dt^2 = [rd^2/ds^2 - r'd/ds]/r^3$, where $(')$ is differentiation with respect to s. An important contribution to the derivation of eq. 3.3 is the r^3

term in the denominator of the operator; it cancels the r^3 term in the Newtonian force. Thus, the equations of motion and the energy integral ($\mathbf{v}^2 = 2(r^{-1} + h)$) are, respectively,

$$rz'' - r'z' + z = 0, \qquad |z'|^2 = 2(r + hr^2) \tag{3.4}$$

where h is the constant of energy. When the new dependent variable is included, a cancellation of terms occurs if r, r' are replaced with their representation in terms of w and its conjugate. The new equations of motion and energy integral are

$$2w'' - \frac{w}{r}(2|w'|^2 - 1) = 0, \qquad 2|w'|^2 = 1 + hr. \tag{3.5}$$

Making the obvious substitution with the energy integral, the equations of motion become $w'' - (h/2)w = 0$. Thus, both this substitution and the coefficient $a = -h/2$ in eq. 3.3 reflect the fact that the transformed equations are on a fixed energy surface. Because the change of the dependent variables is not 1-1, two $\mathbf{u}$ values correspond to a single $\mathbf{r}$ value. The exception is $\mathbf{u} = \mathbf{0}$ which corresponds to $\mathbf{r} = \mathbf{0}$.

In addition to being able to analyze the behavior of orbits near a binary collision, another advantage of eq. 3.3 is that they are linear equations with purely complex eigenvalues. From this it follows that all solutions are stable; a small perturbation has only a small effect on the solution. As such, numerical solutions of these equations retain the properties of the actual solution. This isn't true for eq. 3.1. Here, a small numerical error can force the numerical solution onto a different energy surface, which, in turn, alters the frequency of the motion. Then, as is true for two pendulums with close, but different frequencies, the true and the computed solutions eventually will be at opposite ends of the orbits.

Several important dynamical systems, such as the Earth satellite problem, are perturbed forms of eq. 3.1. (The perturbations for the equations of the Earth satellite problem reflect the Earth's "middle age bulge" around the equator.) Many of these systems cannot be solved analytically; instead they are numerically integrated. Consequently, it is natural to question whether perturbed forms of eq. 3.1 also can be converted into a form that inherits some of the stability properties of the harmonic oscillator. However, "real" problems, such as the Earth satellite problem, are in a three-dimensional space; they can't be restricted to the "Flatland" setting of a fixed two-dimensional plane. Therefore, to carry out the program of converting the satellite problem into a perturbed form of the three-dimensional harmonic oscillator, one first must be able to convert the three-dimensional eq. 3.1 into a three-dimensional harmonic oscillator. It appears from comments in the literature (e.g., see [**15**, p. 23]) that Levi-Civita unsuccessfully tried to find such a three-dimensional extension of his two-dimensional solution.

As the world moved into the Space Age, this issue of finding a harmonic oscillator form for the three-dimensional, two-body problem added practical importance to its theoretical interest. This question was raised by E. Stiefel at the 1964 Oberwolfach conference he organized. Attending this conference was P. Kustaanheimo, a Finnish astronomer, who had been using spinors to analyze properties of his theory of relativity and other problems from physics. Spinors are a natural generalization of complex variables, and, sure enough, by mimicking Levi-Civita's approach with spinors, Kustaanheimo solved the problem during this conference. More specifically, let $w = u_1 + iu_2 + ju_3 + ku_4$, $z = x_1 + ix_2 + jx_3 +$

kx_4, and $\mathbf{u} = (u_1, u_2, u_3, u_4)$. With the change of dependent variables, $w^2 = z$, along with Sundman's change of the independent variable, Kustaanheimo [2] converted the system into the equations

$$\mathbf{u}'' + a\mathbf{u} = 0. \tag{3.6}$$

Of course, the true problem is in R^3, not in R^4. Fortunately, as indicated later, this change of variables maps each solution of eq. 3.6 to a fixed plane $x_4 = c$. Thus, x_4 is a dummy variable that can be set equal to zero, and solutions of eq. 3.6 are mapped to solutions of eq. 3.1 in the plane $x_4 = 0$. Moreover, because a point z in the plane $x_4 = 0$ determines the magnitude of the corresponding $\mathbf{u}$, it follows from the dimensional difference that each point in the plane $x_4 = 0$ corresponds to a circle in the $\mathbf{u}$ variables rather than the two points in the Levi-Civita transformation. The only exception is when $\mathbf{r} = \mathbf{u} = \mathbf{0}$, where this circle degenerates into a point.

Now that the spinor transformation solved this decades old problem, Kustaanheimo and Steifel [3] joined forces to use this approach to analyze perturbed problems of equation 3.1. Among other conclusions, they showed that this transformation does provide numerical advantage. (Extensions are in [15].) This transformation, which is quite widely used today (more so in Europe than in the U.S.), is known as the KS transformation.

What was the source of the difficulties in extending Levi-Civita's approach? Why can't we find a similar relation for R^3? One way to see this is to express Levi-Civita's change of dependent variables, $x = u_1^2 - u_2^2$, $y = 2u_1u_2$ in the differential form

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = 2\begin{pmatrix} u_1 & -u_2 \\ u_2 & u_1 \end{pmatrix}\begin{pmatrix} du_1 \\ du_2 \end{pmatrix}. \tag{3.7}$$

The first column in this orthogonal matrix can be viewed as locating a point in space, say, on the unit circle $u_1^2 + u_2^2 = 1$, while the second column defines a tangent vector. Thus, the Levi-Civita transformation can be identified with a vector field of unit length on S^1—the unit circle in R^2. Presumably, the appropriate R^3 change of dependent variables leading to the harmonic oscillator would be given by a 3×3 orthogonal matrix where the first column locates a point on S^2, the unit ball in R^3, while the remaining two columns correspond to tangent vectors to the sphere. But the existence of any such matrix is frustrated by the "hairy billiard ball" effect—one can't comb a hairy billiard ball without getting a "cowlick" where at least one hair stands upright. In other terms, as we know from Poincaré and Brouwer, there doesn't exist a smooth, nonzero, tangent vector field on S^2. In our setting, this means there doesn't exist a 3×3 orthogonal matrix of the appropriate form, or, in turn, these topological reasons proscribe the existence of the desired change of variables.

On the other hand, going back to his Ph.D. dissertation and his earlier work in algebraic topology, E. Stiefel [15, see Chp 11] was aware of the fact that while it is impossible to comb a hairy S^2, one can comb a hairy S^3, the unit ball in R^4, in several ways. This unit ball admits a smooth frame where if (u_1, u_2, u_3, u_4) determines a point on S^3, then the remaining three mutually orthogonal, unit tangent vectors are $(-u_2, u_1, u_4, -u_3)$, $(-u_3, -u_4, u_1, u_2)$, and $(u_4, -u_3, u_2, -u_1)$. Once these four vectors are expressed as columns of a 4×4 matrix, much as in eq. 3.7, this defines the KS transformation of the dependent variables. Note that the last

row of this matrix is orthogonal to $\mathbf{u}$; this ensures that $x_4 = 0$. Indeed, the resulting mapping is one of the Hopf maps from S^3 to S^2.

Subsequent to the KS transformation, other representations of the two-body motion have been found to eliminate the difficulties of collisions. A geometric approach is given by J. Moser [7], where he showed that two-body motion in a d-dimensional physical space can be related to geodesic flow on S^d. Thus in the specific case $d = 3$, the problem in R^3 is related, in a different manner, to motion on S^3. In Moser's representation, the positions on the sphere correspond to the velocity of the particle through stereographic projections. Hence, the north pole of S^d represents the collision where the velocity becomes infinite. J. Milnor [6] has a nice exposition of this and related ideas. A second important development is due to R. McGehee [8]. Remember, Sundman and Levi-Civita found transformed equations that included the point $\mathbf{r} = \mathbf{0}$ as a regular point. These approaches don't extend for triple collisions. However, McGehee found a way to "paste on" the behavior of orbits at triple collisions when $r = 0$. He did this by expressing the equations of motion in a spherical coordinate framework, and by introducing an appropriate change of independent variable. The next step was to exploit the singularity in spherical coordinates that occurs when the radius is zero but all of the angles are admitted. In his representation, the "angles" correspond to the configurations formed by the particles while the radius of the sphere measured the distance between colliding particles. In this manner, the motion in the collinear three-body problem can be extended to a "collision manifold" corresponding to fictitious motion where the radius of the system is zero; i.e., this is the limiting motion when the distances between the particles approaches zero. R. Devaney [1] has a nice exposition of these ideas.

4. Sundman theory. At the beginning of this century, K. Sundman made several important contributions to our understanding of the N-body problem—contributions that have stood the test of time. Ironically, one of his major conclusions killed interest in a line of inquiry, so this particular result is not very well known. It should be; it is where Sundman "solved" the three-body problem according to accepted standards of the late 1800s and early 1900s. Indeed, in the late 1800s the King of Sweden and Norway established a prize for anyone who could find the solution of the N-body problem. The prize was awarded to Poincaré in 1889 even though he hadn't solved the original problem. (On the other hand, Poincaré's prizewinning paper contains a wealth of ideas that remain influential.) The originally stated problem finally was solved in 1913 by Sundman [16] when he found a converging series solution for the three-body problem. Unfortunately, his series converges so slowly that, essentially, it is useless for any practical purpose. Consequently, this nice result is not as well known as it should be. Still, Sundman's work remains a beautiful and important combination of several deep ideas that have a continuing influence on celestial mechanics. Sundman's results can be described via complex variables, so these variables will be used to outline what Sundman did and why his series converges so slowly. Part of this description of Sundman's contributions is based on Cauchy's theorem showing that the radius of convergence of a power series is determined by the location of the nearest singularity.

A serious complication hindering achievement of a convergent power series solution for the three-body problem are the singularities caused by binary collisions.

We now know that collisions of any kind are improbable [**11**]. But collisions exist, and when they do, they restrict the accompanying radius of convergence for any power series solution. One way to avoid these complications might be to restrict attention to those conditions with collision-free orbits. The fault with such an approach is that we don't know how to characterize these conditions; in general, we don't know whether an initial condition will, or will not lead to a collision-free trajectory.

As already indicated, Sundman cleverly avoided the problem of binary collisions by converting the equations to a new system where a binary collision is *not* a singularity—it becomes a regular point. Because of the central role played by this result, another complex variable argument will be outlined to suggest why we should expect this conclusion and why Sundman's change of independent variable is a "natural" one.

To start, consider colliding particles in the collinear, central force problem. The motion is on the positive half of the x-axis so the defining equations of motion are $x'' = -x^{-2}$. Because the right-hand side of this equation always is negative, the solution, $x(t)$, is concave down. This forces the system to have a collision either forwards or backwards in time, and if $x'(t) \leqslant 0$, then x' remains negative in the future. To determine how the collision occurs, multiply both sides of the equations of motion by x' and integrate to obtain the energy integral

$$(x')^2 = 2x^{-1} + 2h, \tag{4.1}$$

where h is a constant of integration.

If there is a collision at time t_0, then $x \to 0$ as $t \to t_0$. This changes the energy integral to $x(x')^2 = 2 + 2hx \sim 2$. In turn, this means that $x^{1/2}x' \sim -2^{1/2}$, or that $x^{3/2}(t) \sim A(t_0 - t)$. Consequently,

$$x(t) \sim A(t_0 - t)^{2/3} \qquad \text{as } t \to t_0 \tag{4.2}$$

where A is some positive constant.

It turns out [**10, 12**] for the general N-body problem that if there is a collision of any kind at $t = t_0$, then the colliding particles must approach each other like $(t_0 - t)^{2/3}$; i.e., the rate of approach for collinear collisions extends to all possible kinds of collisions. This is important information about collisions, but it doesn't explain the complex nature of this singularity. Is it an algebraic branch point? Is it a logarithmic singularity? As indicated next, binary collisions always correspond to algebraic branch points. This is true even if several different binary collisions occur at the same time [**12**].

One way to show this for the collinear problem is to "blow-up" the singularity by defining $X(t)$ as $X(t)\,(t_0 - t)^{2/3} = x(t)$. To simplify the equations, assume that the time of collision is at $t_0 = 0$ (this just defines the origin on the time axis) and that we approach the collision through positive values. (The system is time reversible; the equations of motion are invariant with respect to the change of independent variable $-t$.) By substituting $[t^{2/3}X(t)]''$ into the defining equations of motion, the equations for $X(t)$ become

$$t^2X'' + (4/3)tX' - (2/9)X = -X^{-2}. \tag{4.3}$$

To convert this Euler differential equation into a form that admits analytic solutions, use the change of the independent variable $s = t^{1/3}$ to obtain

$$s^2X'' + 2sX' - 2X = -9X^{-2}, \tag{4.4}$$

where the primes now denote differentiation with respect to s. By using series methods from elementary ordinary differential equations and a standard majoriant argument to justify convergence, it follows that this system has an analytic solution in s. Moreover, in a neighborhood about the time of collision the solution is

$$x(t) = \sum_k a_k t^{2k/3}. \tag{4.5}$$

It follows from eq. 4.5 that *a binary collision for the system is an algebraic branch point* where $s = t^{1/3}$ serves as a local uniformizing variable. As t passes through zero, the solution runs through the sheets of the Reimann surface to emerge as though the collision corresponds to an exact rebound or a perfectly elastic collision. Consequently, we should expect that this collision singularity can be removed, and that it doesn't constitute a serious obstacle to the analysis of the system. (The same assertion does not hold for triple or more complicated collisions. Here, some of the exponents in the expansion from eq. 4.3 or 4.4 depend continuously on the value of the masses of the colliding particles. Thus, logarithmic singularities result. See **[14]** for triple collisions and **[12]** for a general discussion.)

Incidentally, eq. 4.5 provides the basis for an "intuitive" argument to explain why Sundman's change of independent variable, $ds = dt/r(t)$, works. As in our derivation of eq. 4.5, if we know that the time of the collision is t_0, then we could substitute $s = (t - t_0)^{1/3}$, or $ds = dt/3(t - t_0)^{2/3}$, to obtain an analytic solution. But, we don't know, a priori, when or even whether a collision will occur. However, perhaps one way to avoid this complication about when a collision occurs is to base the time change on the appropriate power of $r(t)$. After all, a collision occurs if and only if r approaches zero, so the growth properties of r determine when a collision happens. If such an approach is to be successful, then the growth properties of r need to be further exploited by using the appropriate choice of α so that r^α effectively replaces the $(t - t_0)^{2/3}$ term in the change of the independent variable. The asymptotic expression given in eq. 4.2 shows that $\alpha = 1$, so a natural choice for the change of variables is $ds = dt/r(t)$. This is the Sundman change of the independent variable.

After Sundman eliminated binary collisions from the equations of motion, the only remaining, real singularity for the transformed three-body problem is a triple collision. Triple collisions can be avoided by appealing to a result proved by Sundman, and already known to Weierstrass (see **[9**, p. 66]). This result uses the integral of angular momentum $\sum_i m_i \mathbf{r}_i x \mathbf{v}_i = \mathbf{c}$ where $\mathbf{c}$ is a vector constant of integration and where m_i, $\mathbf{r}_i$, and $\mathbf{v}_i$ are, respectively, the mass, the position vector, and the velocity vector of the ith particle. The Weierstrass-Sundman theorem asserts that if $\mathbf{c} \neq \mathbf{0}$, then the system cannot have a complete collapse. (For an elegant proof, see **[9**, p. 66].) Namely, if $N = 3$, then triple collisions cannot occur off of the algebraic variety $\mathbf{c} = \mathbf{0}$.

By restricting attention to $\mathbf{c} \neq \mathbf{0}$, we've removed all real singularities from the transformed equations for the three-body problem. One last constraint on the radius of convergence of a series solution is the possibility of complex-valued singularities;

are there imaginary collisions of the particles at imaginary values of time? There are. To show this and to develop some insight about how they are related to real behavior, consider elliptical solutions of the central force problem eq. 3.1. The solution $r(t) = (\mathbf{r}(t), \mathbf{r}(t))^{1/2}$ does not have a closed form representation. Instead, it is defined implicitly through the *Kepler equations*

$$r(u) = a(1 - \varepsilon \cos u) \qquad t = u - \varepsilon \sin u, \tag{4.6}$$

where ε is the eccentricity of the ellipse, a is the length of the semimajor axis, and u is a variable.

Elliptical solutions are well behaved without any possibility of a collision, so one might guess that the power series converges for all values of time. It doesn't; this is because there are complex singularities. By letting $u = u_1 + iu_2$ and $t = t_1 + it_2$, a computation that just involves the expansion of sine and cosine for complex values proves that there is an "imaginary collision" where $r = 0$ at

$$u_k = 2k\pi \pm i \cosh^{-1}(1/\varepsilon), \qquad k = 1, 2, \ldots, \qquad t_k = u_k - e \sin u_k. \tag{4.7}$$

Thus, not only do complex singularities occur, but $\mathrm{Re}(t_k)$ is determined precisely where, on the real axis, $r(t)$ assumes its minimum value. Moreover, the smaller the value of this minimum (i.e., the larger the value of ε), the closer this complex singularity is to the real axis. This suggests that, if in some manner, the three-body system can be forced to be "bounded" away from a triple collision, then perhaps the complex singularities are bounded away from the real axis.

This is what Sundman proved. His first step, which is a significant improvement over the Weierstrass-Sundman Theorem, was to show that if $\mathbf{c} \neq \mathbf{0}$, then for each value of t, $\max\{|\mathbf{r}_i(t) - \mathbf{r}_j(t)|\} > D(\mathbf{c}) > 0$. In other words, if $\mathbf{c} \neq \mathbf{0}$, then not only can there be no triple collision, but the system is strictly bounded away from a triple collision. (Subsequently, Sundman's result has been extended to the N-body problem for all values of N. See [5].) Using this fact along with the Cauchy existence theorem for differential equations, Sundman proved that the system has no complex singularities in a strip (depending on the value of $\mathbf{c}$) in the complex plane centered around the real axis. With this, the remainder of his proof is immediate. All one needs to do is to conformally map this strip to the unit disk. For instance, if the strip is $|\mathrm{Im}\, s| \leqslant \beta$, then the change of independent variables is $\sigma = (e^{\pi s/2\beta} - 1)/(e^{\pi s/2\beta} + 1)$. In this new system, the equations of motion have no singularities in the unit disk, so the resulting power series converges. Along the real axis in the unit disk, σ corresponds, in a 1–1 fashion, to all real values of time. Thus, in this way, Sundman proved the existence of power series solutions for the three-body problem that converges for all real values of time.

Sundman solved the three-body problem but, unfortunately, the series solution is of little practical value because it converges too slowly and it requires too many terms to achieve any interesting degree of accuracy. It is fairly easy to see why this is so. Remember, Sundman used two changes of the independent variable. The first was to eliminate the singularities due to binary collisions. As we see from eq. 4.2, this involves a change that has the form $s = (t_0 - t)^{1/3}$ near a binary collision. This change of variables "slows down" the dynamics with the accompanying effect that a numerical value of s tends to be larger than the corresponding numerical value of t. (This is similar to the situation in numerical integration where smaller step sizes are

used to achieve accuracy with rapid changes in the dynamics. Accompanying the smaller steps is an increase in the number of steps; this number of steps corresponds to the value of s.) Next, this larger value of s is exponentially mapped to the unit disk in the complex σ-plane. Thus, it follows that fairly small values of t can be identified with values of σ near the boundary of the unit disk. When this happens, we should expect the convergence to be slow. This occurs. In other words, while Sundman successfully solved the three-body problem, the theoretical properties of his analysis demonstrate that, in general, an universal power series solution of the three-body problem is impractical. His success in solving an age-old problem killed interest in this line of inquiry. Nevertheless, many of Sundman's results used to achieve his series solution continue to play a significant role in the development of celestial mechanics.

5. A footnote. The emphasis in this brief description of aspects of the Newtonian N-body problem has been on the role played by complex variables. It is appropriate to end with a historical aside about how certain problems from celestial mechanics played a role in the development of some of the classical results in complex variables.

As mentioned in Section 3, the solution for the simple, central force problem, $r(t)$, given by eq. 4.6 for the elliptic case, does not have a closed form representation. This solution is only defined implicitly through Kepler's equations. Consequently, several mathematicians sought to find a useful direct expression, even a power series solution, for r. A. Wintner [**17**], in his book *Analytical Foundations of Celestial Mechanics*, notes in a footnote on page 217, "*The direct proof of* [*this aspect of Kepler's equation*] *played an important historical role in the theory of analytic functions* . . . [*A*] *principal impetus for Cauchy's discoveries in complex function theory was his desire to find a satisfactory treatment for Lagrange's series* [*for Kepler's equation*]. *Cauchy was led to his fundamental theorem connecting the radius of convergence with the location of the nearest singularity, as well as to his maximum principle, precisely in his papers dealing* [*with this problem*]. *Also the facts usually referred to as the argument principle and Rouche's theorem were first observed in connection with this problem concerning Kepler's equation.*"

Acknowledgement. Some of the material described in this essay is known to experts in celestial mechanics (e.g., see S. Sternberg's *Celestial Mechanics*, Part I, Benjamin, 1969), but it doesn't seem to be well known outside of this community. I've used aspects of this material in my courses on celestial mechanics, on complex variables, and even in calculus. Indeed, this article is based on my introductory lecture for a "short course" on the Newtonian N-body problem I gave at Instituto Matematica Pura e Aplicada, Rio de Janeiro, Brazil during January and February of 1987. I would like to thank the institute and my host, J. Palis, Jr., for their hospitality. This paper was written with partial support from an NSF grant.

REFERENCES

1. R. Devaney, Blowing up singularities in classical mechanical systems, this MONTHLY, 89 (1982), 535–552.
2. P. Kustaanheimo, Spinor regularization of the Kepler motion, *Ann. Univ. Turku, Ser.* AI. 73 (1964).
3. P. Kustaanheimo and E. Stiefel, Perturbation theory of Kepler motion based on spinor regularization, *J. Reine Angew. Math.*, 218, (1965) 204–219.
4. T. Levi-Civita, Sur la regularisation du probleme des troi corps, *Acta Math.*, 42 (1920) 99–144.

5. C. Marchal and D. G. Saari, On the final evolution of the N-body problem, *J. Diff. Eqs.*, 20 (1976) 150–186.
6. J. Milnor, On the geometry of the Kepler problem, this MONTHLY, 90 (1983) 353–364.
7. J. Moser, Regularization of Kepler's problem and the averaging method on a manifold, *Comm. Pure Appl. Math.*, 23 (1970) 609–636.
8. R. McGehee, Triple collisions in the collinear three body problem, *Invent. Math.*, 27 (1974) 191–227.
9. Harry Pollard, Celestial Mechanics, Carus Math Monograph # 18, MAA, 1976.
10. H. Pollard and D. G. Saari, Singularities of the N-body problem, I, *Arch. Rat. Mech. Anal.*, 30 (1968) 263–269.
11. D. G. Saari, Improbability of collisions in Newtonian gravitational systems, *Trans. Amer. Math. Soc.*, 162 (1971), 267–271; Paper II, *Trans. Amer. Math. Soc.*, 181 (1973) 351–368.
12. _____, Manifold structure for collisions and for hyperbolic-parabolic orbits in the N-body problem, *J. Diff. Eqs.*, 55 (1984) 300–329.
13. _____, A global existence theorem for the four body problem of Newtonian mechanics, *J. Diff. Eqs.*, 26 (1977) 80–111.
14. C. L. Siegel, Der Dreierstross, *Ann. Math.*, 42 (1941) 127–168.
15. E. L. Stiefel, and G. Scheifel, Linear and Regular Celestial Mechanics, Springer-Verlag, New York, 1971.
16. K. F. Sundman, Recherches sur le probleme des trois corps, *Acta Soc. Sci. Fennicae*, 34 no. 34 (1907).
17. A. Wintner, The Analytical Foundations of Celestial Mechanics, Princeton University Press, 1941.