Chapter 1

Introduction

Simply stated, the "Newtonian N-body problem" is the mathematical study of how heavily bodies move in settings where the dynamics are dictated by Newton's law of motion. In practical terms, this area now includes just about any dynamical system that even remotely resembles Newton's law.

Beyond the insight the subject provides for understanding astronomical issues, the Newtonian N-body problem has historically served as a source of mathematical discovery and new problems. The purpose of this book is to introduce the reader to a selective portion of issues about the Newtonian N-body problem while outlining and describing some open problems.[1]

1.1 Mars

How do the heavenly bodies move? A quick introduction can be provided by using elementary complex variables to describe some simple orbits. The ultimate purpose of this exercise is to show how surprising levels of complexity can arise even in particularly "nice" and "well behaved" settings. Later in this chapter, these orbits are used to describe and motivate an open research problem.

Start with a mystery that most surely bothered generations of school kids: it most certainly troubled me when I was in the fourth grade. It involves the story of Galileo being forced to recant his views that the Sun, rather than Earth, is the center of the solar system. Even a child can appreciate the fact that if the church felt it was necessary to force Galileo to recant, then the stakes in the issue must have been high. But, what

[1] A companion book [90] is being prepared that addresses issues other than those described here.

difference does it make if the Sun revolves about the Earth, or the Earth about the Sun? After all, whichever occurs, one forms the center of a circular motion for the other. Why should we care which is which?

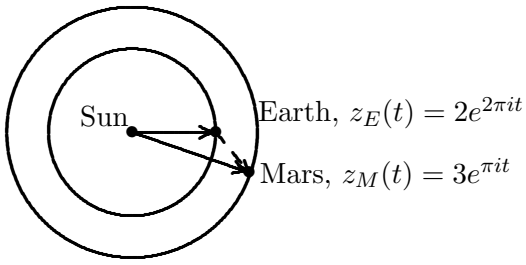


Fig. 1.1. Sun-Earth-Mars coordinates in half-astronomical units

1.1.1 Motion of Mars

To explain the kinds of difficulties that are introduced by an Earth-centered prejudice, start with the Sun as the center of our solar system. A simplified story has Mars approximately 3/2 times (actually, about 1.524 times) as far from the Sun as the Earth, and Mars takes approximately two years (about 687 Earth days) to complete its journey about the Sun.

To keep everything simple, eliminate fractions by replacing the standard astronomical unit (the distance between the Earth and the Sun) with what I call "half-astronomical" units. In the new system, which is depicted in Fig. 1.1, the Earth is two units from the Sun, and Mars is three. Using complex variables, a reasonable description of the motion of the Earth is given by $z_E(t) = 2e^{2\pi it}$ while that of Mars is $z_M = 3e^{\pi it}$.

Finding the orbit of Mars relative to the Earth now is simple; it is

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

To describe this orbit, add and subtract the distance to the Sun to obtain

$$\begin{aligned} z(t) &= 3e^{\pi it} - 2e^{2\pi it} - 2 + 2 = 2 + e^{\pi it}[3 - 2e^{\pi it} - 2^{-\pi it}] \\ &= 2 + [3 - 4\cos(\pi t)]e^{\pi it}. \end{aligned} \tag{1.2}$$

According to Eq. 1.2, the graph of this equation, as given in Fig. 1.2, depicts the surprisingly complicated orbit of Mars when viewed relative to that of the Earth: it is *a limacon* with a nicely defined loop.[2]

[2]In my introductory calculus courses, I often use the trigonometric version of this

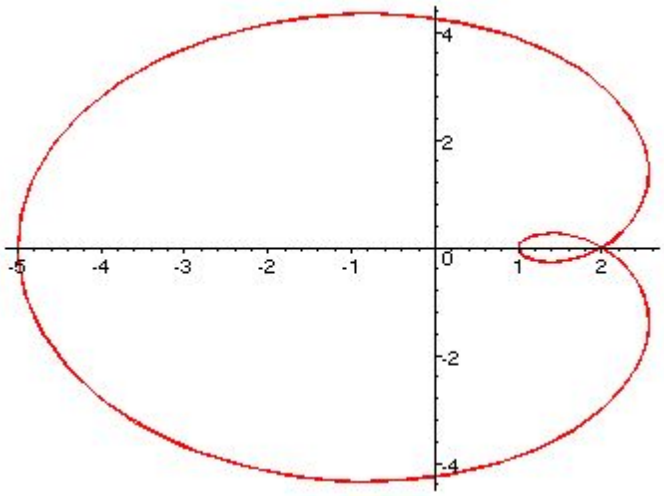


Fig. 1.2. Apparent motion of Mars relative to the Earth

Figure 1.2 makes it clear why the pre-Copernican, Earth-centered prejudice made it so difficult to predict the motion of the planets and to develop a "Newtonian Theory." For a segment of time on this orbit, everything is regular. Indeed, starting at the point where the loop intersects itself, Mars starts on its long journey moving away from the Earth until eventually it is five half-astronomical units away. (This position corresponds to where Earth and Mars are on opposite sides of the Sun.) The interesting, counterintuitive action starts when Mars returns to begin its close approach to the Earth. First, it quickly swoops in a radical plunge toward the Earth. But rather than colliding, Mars suddenly *reverses direction* to swoop out—a motion suggesting that the physics—for some strange reason—suddenly changes to a *law of repulsion* rather than attraction. Finally Mars changes direction once more so that it can repeat its long two-year journey.

Imagine the difficulty in determining the appropriate force law—a law that resembles some form of attraction for most of the journey only to suddenly become a law of repulsion when Mars approaches Earth, and then reverts back into a law of attraction. Other than resorting to bad jokes about the annoyance of Earthling's politics or their behavior, how does one explain the sudden repulsion of Mars when it starts approaching Earth? In other words, the change of variables from a Earth-centered to a Sun-centered system makes a considerable difference: without it, it is difficult to

example to put life into those mandatory reviews of trigonometry. The trigonometric version just uses double angle formulae; e.g., $(3\cos(\pi t), 3\sin(\pi t)) - 2(\cos(2\pi t), 2\sin(2\pi t)) = (2,0) + \rho(cos(\pi t), \sin(\pi t))$ where $\rho = 3 - 4\cos(\pi t)$.

even imagine how Newton's laws of attraction could have been developed.

Incidentally, it is easy to observe this retrograde behavior of Mars. Of course, the change in distance between Earth and Mars cannot be detected by the untrained naked eye, but the change in direction—where Mars appears to be moving in one direction, stops and moves backwards, and then stops again to return to its original direction—is quite apparent over the span of several nights. During those periods when Mars approaches Earth to start its dipping behavior, even a casual observer can notice how at a fixed time each night the position of Mars swings to define, over a period of days, a compressed "Z."

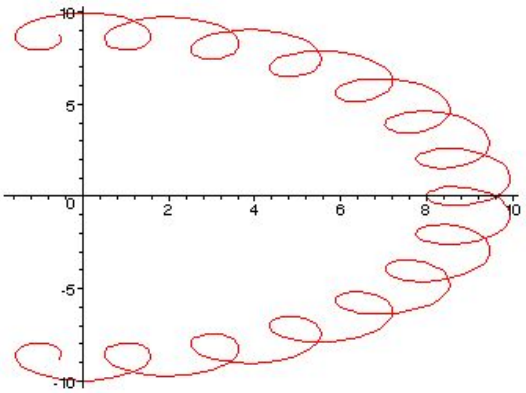


Fig. 1.3. Apparent orbit of a planet 9 times farther from the Sun

While the apparent motion of Mars offers surprising behavior, the orbits of the planets farther from the Sun adopt a much more complicated appearance with the several loops as indicated in Fig. 1.3. This figure depicts the apparent behavior of a planet nine AU away from the Sun: a distance that is a bit short of Saturn's actual orbit. Rather than developing a complicated version of the above description, a different elementary approach is described next.

1.1.2 The "far out" planets

Consider the circular orbit of a far-out planet—Mars, Saturn, or beyond—given by $z_P(t) = ae^{\alpha\pi it}$ where the value of $a \geq 3$ defines the distance from the Sun in our half-astronomical units: the α values are discussed below. After expressing this

$$z(t) = z_P(t) - z_E(t) = ae^{\alpha\pi it} - 2e^{2\pi it}, \tag{1.3}$$

orbit of the planet relative to the Earth in the usual complex variable form of $z(t) = r(t)e^{i\theta(t)}$, a way to determine whether the orbit is moving in a clockwise or counter-clockwise manner (relative to the Earth) is to examine the sign of $\theta'(t)$.

The sign of $\theta'(t)$ is the imaginary part of $(\ln\, z_P(t))' = \frac{z_P'}{z_P} = \frac{r'}{r} + i\theta'$. But since

$$(\ln\, z_P(t))' = \frac{z_P'}{z_P} = \frac{\pi i(a\alpha - 4e^{(2-\alpha)\pi it})}{a - 2e^{(2-\alpha)\pi it}}, \tag{1.4}$$

it follows from the form of the numerator that the sign of θ' must change periodically whenever $a\alpha < 4$.

The reason this $a\alpha < 4$ inequality must hold for all of the planets that are farther from the Sun than the Earth is *Kepler's third law.* This law asserts that

$$a^3\alpha^2 = k \tag{1.5}$$

where k is a constant. Consequently, $a\alpha = (\frac{k}{a})^{1/2}$ is a decreasing function of a: remember, a is the distance of the planet to the Sun. Thus, for a planet sufficiently far from the Sun, we must expect its orbit to experience loops when expressed relative to the Earth. According to Eq. 1.4, the loop occurs whenever the distance between the Earth and the planet decreases toward a (local) minimal value. But because those far-out planets take from decades to a couple of Earth centuries to circle the Sun,[3] it follows that their apparent orbits must exhibit many loops.

A natural related question, which is needed for later purposes, is to determine how far a planet must be beyond the Earth so that its apparent orbit has a loop. Using the units of the Earth, $a = 2, \alpha = 2$, we have that $k = 32$ for Eq. 1.5. Thus, $a^3\alpha^2 = 32$, or the crucial parameter has the value $a\alpha = [32/a]^{\frac{1}{2}}$. Because apparent loops occur when $a\alpha < 4$, it follows that these loops occur when $[32/a]^{\frac{1}{2}} < 4$, or when $a > 2$. Restated in words,

> *the apparent motion of any planet that is farther from the Sun than the Earth has a loop.*

Of course, this assertion holds for all bodies governed by Newton's equation: this fact plays a key role in the discussion about the rings of Saturn given in the last section of this chapter.

Notice how this simple argument just describes a circular uniform motion relative to another circular uniform orbit. The importance of this comment

[3]While Venus takes only about 224 Earth days to circle the Sun, Jupiter takes 4332 (about 11.9 Earth years), Saturn 10,760 (about 29.5 years), Uranus 30,685 (about 84 years), Neptune 60,190 (about 165 years), and Pluto 90,800 days (about 249 years).

derives from the reality that all sorts of circular motions arise in the N-body problem. Consequently, it is reasonable to anticipate that loops and complicated orbits of this type are commonplace. To illustrate with another example, I now turn to the orbit of Mercury.

1.2 Mercury

Mercury, our smallest planet, is only slightly larger than our Moon. Even though this planet was known to the ancient Egyptians, its proximity to the Sun has hampered all attempts to explore Mercury with either telescopic observations or space missions.[4] Yet, enough is known about Mercury to allow it to play an important role in the development of celestial mechanics. We know, for instance, that the perihelion of Mercury (i.e., the closest approach of Mercury to the Sun) deviates 43" of arc per century from that predicted by Newton's laws. What a stunning assertion! When you consider only 43" of arc *per century* you have to join me in being impressed by the precision attained by our nineteenth century colleagues!

This deviant behavior has encouraged all sorts of searches including speculation about the possible existence of another planet called "Vulcan."[5] What a delightful notion: could it be that someone would discover (again!) a new planet strictly through *mathematical computations*? Einstein spoiled the fun by showing that this effect could be explained strictly in terms of his theory of relativity.

Moving on to the orbit of Mercury, we now know that this planet takes

[4]Mercury was visited by NASA's Mariner 10 in March of 1974 where the pictures sent back from its three approaches reveal a planet with plains of frozen lava and a surface pockmarked with craters: the planet resembles our Moon. The Mariner photos also discovered the Caloris Basin; a basin that suggests one of the most cataclysmic events in our planetary system. Mercury is not an inviting destination.

[5]It is highly doubtful that Vulcan exists, yet its colorful history involved important individuals. For instance, the French mathematician Urbain Le Verrier ensured a place in history by using computations about how the path of Uranus deviated from Newtonian predictions to predict the existence and position of Neptune; he did this before Neptune was even observed. Using similar reasoning, in 1860 he wondered whether the deviation of Mercury's perihelion indicated the existence of another planet or an asteroid belt. Accelerating the chase for the discovery of the new planet, which Le Verrier christened "Vulcan," were the claims of the amateur astronomer Lescarbault that he sighted a spot—a planet?—near the Sun. But Le Verrier showed that even if this sighting were the speculated Vulcan, its orbit would not explain Mercury's perihelion problems. Again in 1878 two reputable astronomers, Watson and Swift, suspected they saw "stars" that might be the elusive Vulcan. With the exception of faint objects observed near the Sun during a 1970 solar eclipse, if Vulcan really exists, it has successfully remained hidden.

about 90 Earth days (actually, 87.97) to circle the Sun. If you check the books in astronomy written prior to 1965, they reflect the earlier belief that Mercury took about the same length of time—about 90 Earth days—to rotate on its axis. If this were true, then, as these books asserted, the same face of Mercury would always face the Sun. This would constitute, of course, a phenomenon similar to where the same face of the Moon always faces the Earth.

In 1965 radio astronomers visited this long standing belief. Using the Arecibo radio telescope based in Puerto Rico, they discovered that the sideral rotation (the length of time it takes a planet to revolve once as measured against a fixed background) was about 60 days (actually 59, or two thirds of the length of a Mercury year). In turn, this change in the rotational period significantly shortens a Mercury solar day from eternity to about 176 Earth days, or exactly two Mercury years.

Our interest in these figures is that they identify another astronomical setting where it is reasonable to consider the orbit of a rotating object relative to a point on another rotating object. Stated in simple terms, if someone lived on Mercury, what would the orbit of the Sun look like?

To answer to this question, we need to use a more accurate description for the orbit of Mercury. A sharper approximation for the orbit of any planet is obtained by treating it as an ellipse with eccentricity ϵ rather than a circle where $\epsilon = 0$. For instance, the Earth's orbit is fairly circular as reflected by its eccentricity of $\epsilon = 0.0167$, while the more elliptical orbit of Mercury is manifested by the twelve-fold larger $\epsilon = 0.2056$ value. Indeed, Mercury's more extreme elliptical nature is captured by the difference between its perihelion distance of 0.308 AU and its aphelion (its largest distance from the Sun) of 0.466 AU.

The position of any planet on its ellipse is given by

$$r_P(\theta) = \frac{a}{1 - \epsilon\cos(\theta)} \approx a(1 + \epsilon\cos(\theta)) \tag{1.6}$$

where a is a positive constant (the length of the semi-major axis of the ellipse) and $\theta(t)$ is the angular position of the planet relative to a reference line.[6] This equation has the complex variable representation

$$z_P(\theta) \approx a(1 + \epsilon\cos(\theta))e^{i\theta}. \tag{1.7}$$

By measuring time in Earth years and using the fact Mercury takes about 60 days, or one-sixth of an Earth year, to rotate on its axis, the

[6]For a quick introduction into the basics of the two-body problem, I recommend the first chapter of Pollard's book *Celestial Mechanics* [60].

rotating motion of Mercury about its axis can be approximated by $e^{12\pi it}$. Thus the apparent position of the Sun (relative to a position on the surface of Mercury) can be represented by the product

$$Z_S(t) = -z_P(\theta(t))e^{-12\pi it}. \tag{1.8}$$

In turn, the angular position of the Sun, given by the argument of $Z_S(t)$, is

$$\mathrm{Arg}(Z_S(t)) = \mathrm{Arg}(-z_P(\theta(t))e^{-12\pi it}) = \theta(t) - 12\pi t. \tag{1.9}$$

According to this expression, the Sun's apparent motion changes direction whenever $(\mathrm{Arg}(Z_S(t)))'$ changes sign; that is, whenever θ' passes through the value of 12π. The reason the Sun's orbit relative to a position on Mercury must experience this directional change comes from *Kepler's second law* that captures the angular momentum: this law asserts that

$$r_P^2\theta' = C \tag{1.10}$$

where C is a constant. What we learn from Eq. 1.6 is that if ϵ is sufficiently large, as it is for Mercury, then it is unreasonable to approximate $r_P(t)$ by a constant. In turn, according to Kepler's second law (Eq. 1.10), $\theta'(t)$ cannot be approximated by a constant, so $\theta(t)$ cannot be approximated by uniform motion. In particular, when $r_P(t)$ is at perihelion, θ' achieves its maximum value; when $r_P(t)$ is at apihelion, θ' attains its minimum. This assertion matches intuition gained from those high school physics exercises of swinging a weight on a string where pulling in the string (i.e., r_P is made shorter) makes the object move faster.

Using Kepler's second law and Eq. 1.6, we have that

$$\theta'(t) = \frac{n}{(1-\epsilon\cos(\theta))^2}$$

where the new constant n—a combination of a and C—is called the "mean motion." For Mercury, $n = 8\pi$, so, by setting

$$\theta' = \frac{8\pi}{(1-\epsilon\cos(\theta))^2} = 12\pi, \tag{1.11}$$

it follows immediately that the apparent motion of the Sun moves in different directions depending on whether $(1-\epsilon\cos(\theta))^2$ is greater than, or less than, $\frac{2}{3}$. The first condition requires $r_P(\theta)$ to have a sufficiently large value, the second requires $r_P(\theta)$ to have a sufficiently small value. In turn, this means that we should expect a reversal in direction to occur whenever Mercury moves closer to the Sun.

What remains is to seen whether the orbit of Mercury ever experiences a reversal; namely, can $\theta' - 12\pi$ change sign? The answer follows by examining what happens for $\theta = 0$ —at perihelion, and $\theta = \pi$—at aphelion. A direct computation shows for Mercury that the θ' values at perihelion and aphelion are, respectively,

$$\theta' = \frac{8\pi}{(0.7944)^2} = 12.677\pi > 12\pi, \quad \frac{8\pi}{(1.2056)^2} < 12\pi.$$

Consequently, the apparent orbit of the Sun relative to Mercury must experience reversals.

More specifically, whenever Mercury moves toward its closest approach to the Sun, the Sun's apparent motion *reverses direction.* But there are two Mercury years for each Mercury day,[7] so this unexpected phenomena happens "twice a Mercury day." To add some drama to the description, as indicated in Fig. 1.4, at some location the Sun will rise in the east, only to almost immediately *set again in the east* for a period of time. Then, the Sun does rise a second time in the east for the long Mercury day until it finally appears to set in the west. The correct word is "appears" because shortly after the Sun sets, it *rises again in the west* for a short time, only to finally settle in the west for the long Mercury night.[8]

[7]Mercury's day-year ratio has intrigued mathematicians. Tom Kyner [32] probably was the first to explore this resonance effect by using dynamical systems to show how the orbit is "trapped." Kyner introduced his results at a 1969 conference on mathematical astronomy held in Sao Paulo, Brazil. Amusingly, Kyner presented his paper the day after the first and (probably) only time he observed Mercury: this was at a social gathering held at the Sao Paulo Observatory. (Most of us conference participants were mathematicians, so visiting an observatory was a novelty.) Later C. Robinson and J. Murdock [70] extended the mathematics. But, if this Mercury rotation problem is interesting, a greater challenge comes from the planet Venus. Years ago friends at JPL told me that prior to radio astronomy, astronomers knew nothing about the rotation of Venus. What they discovered through modern technology was surprising: Venus moves in a *retrograde* motion. Why this "backwards" rotation? I expect that an explanation will involve some fascinating mathematics. In any case, it provides an exercise that I leave to the reader: what is the apparent behavior of the Sun for someone on Venus?

[8]I was delighted when in 1991, Chris Fang-Yen, a high-school student supervised by Stan Wagon in a summer research program at the (now defunct) Geometry Center at the University of Minnesota, sent me his project entitled "Sunrise and Sunset on Mercury." After Wagon read my paper (Saari [88]) where I described this phenomenon, he asked Fang-Yen to simulate the motion. In doing so, Fang-Yen discovered that there does not exist a *fixed position* on Mercury where this double-dip behavior can be observed. The reason is that the Sun is too small and, as distinctly suggested by Fig. 1.4, the dips in the apparent orbit are too tight relative to the radius of Mercury for a Mercurian to see over the horizon. Consequently, to observe this phenomenon, our Mercurian would have

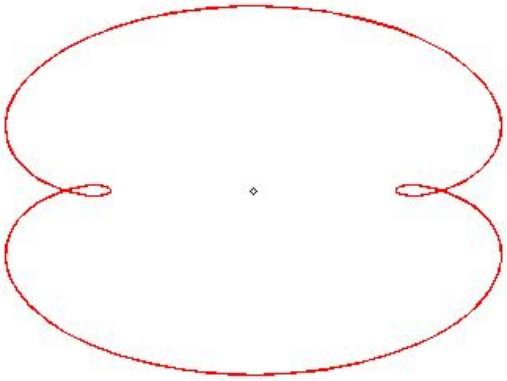


Fig. 1.4. Apparent behavior of the Sun from Mercury

It is clear from Eq. 1.11 that this strange phenomenon is due to the large eccentricity of Mercury. A direct computation, which just involves finding the ϵ value where

$$\theta' = \frac{8\pi}{(1-\epsilon)^2} \leq 12\pi,$$

proves that the apparent motion of the Sun would have no direction reversal had the orbit of Mercury been circular enough so that $\epsilon \leq 0.1835$.

What we have observed in this and in the last section is how different but seemingly unremarkable orbits of two bodies about a central body, or even the orbit of one body about another but with a sufficiently large eccentricity, converts the actual motion into an apparent one with surprisingly complex behavior. A more important observation is that this effect is a direct consequence of describing one circular motion relative to another circular motion. Consequently, we must anticipate this behavior to be reasonably common in celestial mechanics. The interesting message, which helps identify new research issues and opportunities, is that the closer the bodies approach one-another in a Newtonian system, not only does the force between them increase, but it is possible to have an apparent reversal of direction. Thus, as indicated in Figs. 1.2, 1.3, and 1.4, expect surprising and recurring changes in the gravitational forces to arise even in seemingly well-behaved settings. This observation plays a central role in Sect. 1.5.

---

to move to another location. But this should be no problem because he would have "all Mercury day" to do so.

1.3 Epicycles

With an Earth centered system, how does one recognize and represent the motion of the planets? This problem was crucial for astronomers of ancient time. After all, researchers of that epoch needed accurate representations in order to construct astrology tables. Don't scoff at these efforts because, quite frankly, the money paid for these commissioned tables could be viewed as the NSF research funding of that time.

The accuracy needed for this representation problem was achieved through the predictive planetary theory developed by Ptolemy. To appreciate the genius of his work, recall that any theory must adjust to the prejudices of the day; the restrictions facing Ptolemy were the monumental ones established by Aristotle. As a quick, maybe overly simplistic review, Aristotle believed

> (1) that the Earth was the center of the universe, and
>
> (2) that the circle and uniform circular motion were the most virtuous figure and motion.

Of course, since virtue is located in the "heavens,"

> (3) any description of the motion of the planets must be described in terms of uniform circular motion about the Earth.

Fine, but how?

In his *Almagest,* written around 130 A.D., Ptolemy resolved the problem of describing the position of the planets with circular, uniformly moving motion by putting forth his ingenious epicycle approach. The clever idea is that the point indicated by a designated point on the motion of the first circle, the *deferent,* does not represent the location of the planet. Instead, the point merely locates the center of a *second* circle that also is spinning with uniform motion. The location of the planet, then, is given by the moving point on the second circle—the epicycle. (See Fig. 1.5.)

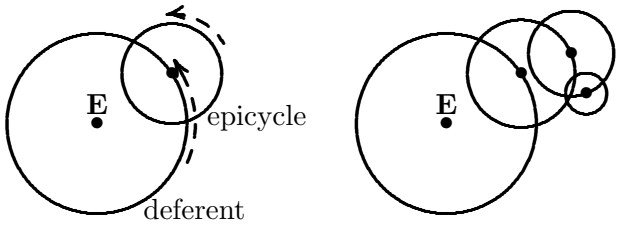
Today this approach may seem to be hopelessly naive. But remember that variations of Ptolemy's theory dominated astronomy for more than a millennium—this is an incredibly long period of time for any scientific theory. Even Newton's theory did not enjoy such a long reign before being challenged by Einstein's relativity, and who knows how long it will take until a serious challenge will force Einstein's theory to be replaced.

The long success of Ptolemy's approach can be understood, again, in terms of elementary complex variables. Let a_j be the radius of the jth circle

where the uniform motion takes c_j Earth years, $j = 1, 2$, to complete one revolution. This means that the motion of a planet as described by epicycles is given by

$$z_P(t) = a_1 e^{b_1 \pi i t} + a_2 e^{b_2 \pi i t} \text{ where } b_j = \frac{2}{c_j},\ j = 1, 2. \tag{1.12}$$

This expression should be familiar: by comparing Eq. 1.12 with the Earth centered expressions Eqs. 1.1, 1.3, we discover that the epicycles can be treated as roughly capturing the conversion of a Sun centered representation of the location of the planets into an Earth centered one. Namely, let the deferent be the location of the Sun relative to the Earth, and let the epicycle describe the location of Mars relative to the Sun. (The actual descriptions were more complicated.) No wonder Ptolemy's approach proved to be so successful!



a. Epicycles **b.** More is better

Fig. 1.5. Epicycle structure

Of course, the demanding needs of astronomers and astrologers to obtain even sharper representations and to correct for accumulated error required a sharper, extended theory. One approach is obvious. Instead of treating a point on the epicycle as the location of a planet, interpret it as the center of still another rotating circle. As indicated in Fig. 1.5b, even that point could be treated as the center for still another rotating circle. Imagine: these attempts to find still better theoretical representations could be the source of an infinite number of Ph.D. theses. By adding more and more terms, we obtain the expression

$$z_P(t) = \sum_j a_j e^{b_j \pi i t}. \tag{1.13}$$

Again, we all accept that the epicycle approach is hopelessly naive. After all, N-body motion is much more complicated where modern theories are accompanied with terms such as "quasi-periodic" or "almost periodic" motion. What is quasi-periodic motion? It is represented by a finite sum in

Eq. 1.13 while almost periodic motion is given by a converging infinite sum. In other words, once we recognize that epicycles geometrically represent motion developed in modern theory, we gain added respect for Ptolemy's insight.

As a brief aside, Eq. 1.13 describes a Fourier series when the b_j's are integer multiples of a specified number. But quasi-periodic and almost periodic motion imposes no constraints on the b_j values. This amazing theory of almost periodic motion was essentially developed by one person, Harald Bohr [6, 7], the younger brother of Niels Bohr.

Before describing more structure of these motions, first recall the story of a mathematician asked to give a general audience talk about mathematics to the parents and teachers of his daughter's elementary school starting off with, "Let X be a non-separable Hilbert Space." The delight of this self-deprecating joke—the kind many mathematicians enjoy—is that many of us cannot quickly provide an example of such a space. To do so while illustrating some of the mathematical structure Bohr developed, define the inner product for complex valued functions on the real line by

$$(f,g) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t)\overline{g(t)}\, dt,$$

and define a Hilbert space in the normal manner. It is easy to show that an orthonormal basis for this space is $\{e^{i\lambda t}\}$ for $\lambda \in R$. As this space—the natural home for epicycles, quasi- and almost periodic motion—has an uncountable basis, it is a natural example of a non-separable Hilbert Space.

1.4 Chaotic behavior

To relate the above story about the behavior of Mars, Mercury, and the other planets to an open research question, allow me a slight digression to describe, in what is intended to be a reader friendly introduction to certain basic concepts from chaotic dynamics. Since this book describes Newtonian mechanics, it is appropriate to describe this behavior in terms of Newton's method for finding zeros of a function.[9] After introducing certain basic points, I briefly touch on that well-known "Period Three implies Chaos" paper by T-Y Li and James Yorke [33].

[9]The material and exposition for this section comes from Saari and Urenko [93] and from Saari [89].

Let me stress that I have no intensions to fully analyze Newton's method nor to describe the subtle features of chaotic dynamics. The intent is strictly to help develop intuition by suggesting what features are indicators of the potential complexity of motion for the Newton's N-body problem. Hopefully this brief, intuitive description will entice readers not familiar with this topic to learn more about these standard tools for the study of the N-body problem and celestial mechanics. More complete descriptions are readily available in books such as Alligood, Sauer, and Yorke [2], Devaney [14] and Robinson [68, 69].

1.4.1 Newton's method

To review, Newton's method for finding a zero of a function $y = f(x)$ starts with an initial guess, x_1, and finds an improved estimate. As indicated in Fig. 1.6, this next choice is found by replacing the specified function with its straight line approximation passing through $(x_1, f(x_1))$. The zero of this linear equation $y = f(x_1) + f'(x_1)(x - x_1)$, denoted by x_2, is the new estimate. If x_2 is not a zero, the process continues. The iteration process has the expression

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{1.14}$$

But rather than using the Eq. 1.14 analytic expression, the geometry of this process as indicated in Fig. 1.6 suffices for our purposes.

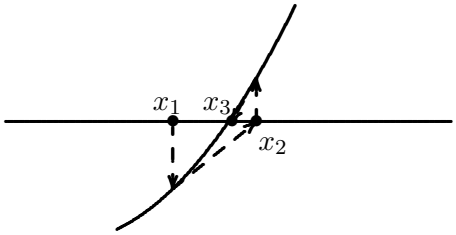


Fig. 1.6. Newton's Method

As it is easy to establish, and as indicated in Fig. 1.6 where the third iterate x_3 already has nearly located the zero, once an iterate is sufficiently near a zero, the process converges to this zero. But as this convergence assertion describes only highly local behavior, it is natural to question what happens globally. Figure 1.7, for instance, represents a polynomial of degree five. The goal is to determine the general dynamical behavior of Newton's method when applied to this function.

Stated in words, other than converging to a zero, what can go wrong?

What else can happen with Newton's method? A little experimentation suggests that there exist period two points: points where Newton's method bounces between them forever. Other points have more serious consequences because they cause Newton's method to cease to exist. In Fig. 1.7, four of these points, depicted by bullets, identify the function's critical points. Since $f'(x_j) = 0$ at the critical points, the horizontal linear approximation never meets the x-axis, so the next iterate is not defined.

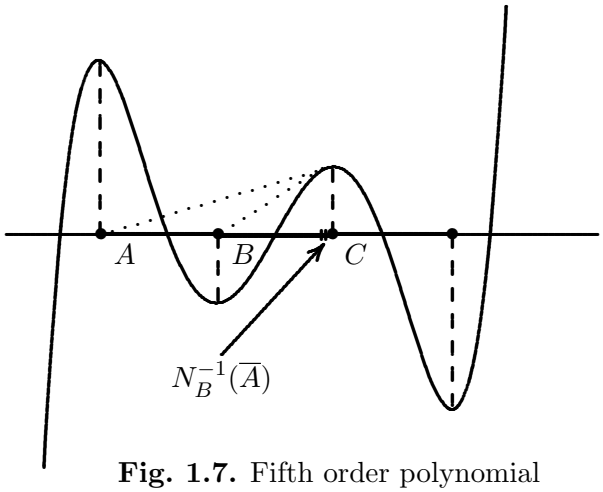


Fig. 1.7. Fifth order polynomial

These four critical points divide the line into five intervals; the two extreme regions are unbounded and the three bounded intervals are labelled A, B, and C. It follows from the properties of polynomials (i.e., the convexity of the curve) that if an iterate ever lands in one of the two unbounded regions, the subsequent iterates converge to the zero of that region. Consequently, all unusual and nonconvergent properties of Newton's method must be limited to the $A \cup B \cup C$ region.

Finding nonconvergent properties

The natural and traditional way to discover "nonconvergent" properties of Newton's method is to examine the behavior of the orbit $x_1, x_2, \ldots, x_n, \ldots$ and experiment with the choice of x_1. If, for instance, iterate x_{101} ends up near x_1, then it is reasonable to expect from continuity considerations that a slight and suitable change in the choice of x_1 will force $x_1 = x_{101}$. Stated in other terms, this means that we should expect the existence of a 100-period orbit. This is an overly simplified, but not inaccurate, description of how various properties about Newton's method were obtained. Have no doubts: while this approach necessarily uncovers only limited conclusions, it can be

technically very difficult.

To motivate an alternative way to examine Newton's method, first remember that anything new or interesting requires keeping all iterates within the region $A \cup B \cup C$. To introduce the approach, after determining an orbit $\{x_1, x_2, x_3, \ldots, x_n, \ldots\}$, replace each iterate with a label identifying the interval in which it belongs. For instance, if $x_1 \in C, x_2 \in A, x_3 \in B, \ldots$ then the initial condition x_1 and its iterates defines the sequence

$$g_f(x_1) = \{C, A, B, \ldots\}.$$

This listing of intervals is not random—it is specified by the dynamic process—so give it status by calling it a *word* or an *itinerary.*

Let $U^3 = \{A, B, C\}^N$ be the *universal* set where N is the set of natural numbers. In words, U^3 consists of the uncountable number of all possible sequences that can be constructed with the entries A, B, and C: call it the *universal set.* With this notation, a word generated by the initial condition x_1 and function f is given by a mapping[10]

$$g_f : A \cup B \cup C \to U^3. \tag{1.15}$$

A way to measure the complexity of Newton's method applied to function f is to determine all entries—all words—in the *dictionary*

$$\mathcal{D}_f = \{g_f(x) \in U^3 \,|\, x \in A \cup B \cup C\}.$$

After all, common sense dictates that if $\mathcal{D}_f$ is a large subset of U^3, then Newton's method admits rich, complex, chaotic dynamics. But if $\mathcal{D}_f$ has only a limited number of words, then it corresponds to a relatively benign dynamic. These comments suggest that a way to measure the complexity of the admissible dynamics is to determine the dictionary $\mathcal{D}_f$. Notice the changed emphasis. Rather than finding particular properties of Newton's method, the ambitious new goal is to completely characterize and catalogue all long term dynamical properties—it is to identify all itineraries in $\mathcal{D}_f$.

The surprising complexity of Newton's method

As shown next (from Saari and Urenko [93]), Newton's method can be highly complex in terms of this complexity measure.

[10] Some choices of x will have their iterates in an unbounded region. Either ignore them as I am doing, or handle them in an obvious manner.

Theorem 1.1 *(Saari and Urenko [93]) For a fifth-order polynomial $y = f(x)$ with distinct real roots*

$$\mathcal{D}_f = U^3.$$

This theorem means that we can choose any sequence consisting of the letters A, B, and C—even a sequence generated by rolling a die—and we are assured that there exists an initial iterate in the specified first interval where the jth Newton method iterate lies in the region specified by the jth entry of the sequence; $j = 1, 2, \ldots$. As indicated below, this means that periodic orbits of any length must exist, as well as far more complex behavior.

The theorem asserts that g_f, defined in Eq. 1.15, is surjective. The method to establish the surjectivity of g_f uses an "iterated inverse image" approach that I illustrate with the sequence

$$w = \{B, A, C, C, A, \ldots\}. \tag{1.16}$$

The approach is to keep refining the set of initial iterates that accomplish each portion of the proposed itinerary.

To see how to do this, let N_k be the portion of Newton's map restricted to interval k where $k = A, B, C$. By using the inverse mappings N_k^{-1}, we have, for instance, that the set of initial iterates starting in the specified first interval B of the Eq. 1.16 sequence and ending in the closed second specified interval $\overline{A}$ is given by $N_B^{-1}(\overline{A})$. The key fact is that N_k maps interval k onto $(-\infty, \infty)$. To see why this is true, notice that as x moves closer to the left endpoint of the bounded interval k, $N_k(x) \to \infty$, while as x moves toward the right-hand endpoint, $N_k(x) \to -\infty$. The conclusion now follows from the continuity of N_k on interval $k = A, B, C$.

Because $N_k : k \to (-\infty, \infty)$ is surjective for $k = A, B, C$, it follows that $N_B^{-1}(\overline{A})$ is a closed subset of B. Actually, as indicated in Fig. 1.7 with the dotted lines, this set is easy to roughly determine: just find the inverse Newton image of the two endpoints of interval A. That is, in interval B find tangent lines to the graph of $y = f(x)$ that terminate on the endpoints of A. The corresponding x values define the endpoints of $N_B^{-1}(\overline{A})$.

Set $N_B^{-1}(\overline{A})$, which identifies all points starting in B that are mapped to A, is much more than we want. After all, our interest is to land only on those points in A that are then mapped to C: we are only interested in the points that are mapped to $N_A^{-1}(\overline{C})$. This set $N_A^{-1}(\overline{C})$ is determined in precisely the same fashion. Thus, refining our set of initial conditions to the closed set $N_B^{-1}(N_A^{-1}(\overline{C}))$ identifies all initial iterates in B that are mapped to A and then to C.

The approach now is obvious. To find all initial points satisfying the future specified in the Eq. 1.16 sequence, continue this iterated inverse image approach to obtain the nested sequence of bounded, closed subsets

$$\overline{B} \supset N_B^{-1}(\overline{A}) \supset N_B^{-1}(N_A^{-1}(\overline{C})) \supset \cdots \supset N_B^{-1}(N_A^{-1}(\ldots N_k^{-1}(\ldots)\ldots)) \supset \ldots. \tag{1.17}$$

By construction, a point in the intersection of all subsets in Eq. 1.17 must satisfy the specified future. But, by appealing to standard results from a first course in real analysis about a nested sequence of bounded, closed sets, we know that such a point must exist. Thus, whatever the envisioned future, it can occur.

Sensitivity and Cantor sets

This construction provides an intuitive description of the source of several of the phrases—"sensitivity to initial conditions," "Cantor sets," etc.—common to this area. To start with the sensitivity phrase, notice how the *expanding* nature of N_k ensures that when N_k is examined in the inverse direction, the inverse image $N_B^{-1}(\overline{A})$ must be a small subset of B. Indeed, by checking Fig. 1.7, it is clear that $N_B^{-1}(\overline{A})$ is a very tiny subinterval.

While set $N_B^{-1}(\overline{A})$ is quite small, its subset $N_B^{-1}(N_A^{-1}(\overline{C}))$ is much smaller. But by construction, this $N_B^{-1}(N_A^{-1}(\overline{C}))$ subset contains *all points* starting in B that Newton's method moves to A and then to C. After the points arrive in C, what happens next? Anything you want: this comment is a direct consequence of the image of N_C being $(-\infty, \infty)$. Stated in words, this surjectivity of N_C along with the small size of $N_B^{-1}(N_A^{-1}(\overline{C}))$ means that even the slightest difference between points in $N_B^{-1}(N_A^{-1}(\overline{C}))$ could result in radically different futures: the dynamic behavior is "sensitive with respect to initial conditions."

Similarly, for each extension of $\{B, A, C, \ldots\}$, each step of the iterated inverse image approach identifies all points that eventually are mapped onto the next specified interval. Included among these points are open intervals that converge to the zero in this interval or in one of the two unbounded regions. Thus, to construct the set of points of nonconvergence of Newton's method, open sets need to be excised at each step—just as in the construction of the "middle thirds" Cantor set. In other words, expect Cantor sets.

With the exception of the behavior of Newton's method on the unbounded regions, nothing in this description restricts the story to polynomials with five real, distinct roots, or even to polynomials. This means, for instance, that the same phenomenon will arise in any polynomial with at

least four real and distinct roots. Just imagine what happens with Newton's method applied to $y = \cos(x)$ with the infinite number of symbols!

All that is needed for this story is that the map incurs an *expansion*—given by the fact that the image of each continuous N_k includes $A \cup B \cup C$—and a *recurrence* effect—captured by those $A \cup B \cup C$ points that are mapped back to $A \cup B \cup C$. It is this *expansion* and *recurrence* combination—a common combination for celestial mechanics— that provides the interest for the N-body problem. Namely, anticipate complex, chaotic behavior in the N-body problem.

1.4.2 Period three and circle maps

To reinforce the basic notions, they are described again using the "Period three implies chaos" title of the influential 1975 paper written by Tien-Yien Li and James Yorke [33]. Beyond the nice mathematics, this paper has historical interest because it is where the term "chaos" originated. As the story goes, after the paper was accepted, the editor of the journal asked the authors to change the title to something mathematically more acceptable and descriptive—maybe something such as "Period three implies topological transitivity." How dull. Fortunately Yorke remained firm in his intent to retain the original title, and the term "chaos" was coined.

Sarkovskii sequence

Unknown to Li and Yorke, eleven years earlier A. N. Sarkovskii [99] published a remarkable and stronger result that if a continuous mapping from the line to the line had a period three orbit, then it also has periodic points of any period. More precisely, Sarkovskii proved for the following sequence, now called the *Sarkovskii sequence*,

$$\begin{array}{ccccccc}
3, & 5, & 7, & 9, & 11, & 13, & \dots \\
2 \cdot 3, & 2 \cdot 5, & 2 \cdot 7, & 2 \cdot 9, & 2 \cdot 11, & 2 \cdot 13, & \dots \\
2^2 \cdot 3, & 2^2 \cdot 5, & 2^2 \cdot 7, & 2^2 \cdot 9, & 2^2 \cdot 11, & 2^2 \cdot 13, & \dots \\
\dots & & & & & & \\
\dots & & & & & & \\
2^n \cdot 3, & 2^n \cdot 5, & 2^n \cdot 7, & 2^n \cdot 9, & 2^n \cdot 11, & 2^n \cdot 13, & \dots \\
\dots & & & & & & \\
\dots & & & & & & \\
\dots & 2^n & \dots & 2^3 & 2^2 & 2^1 & 1
\end{array} \tag{1.18}$$

that a continuous mapping from the line to itself with a periodic point of period k also has periodic points for each period that follow k in the above

listing. Thus, for instance, if such a mapping has a period $k = 3$ point, then, because all positive integers follow 3 in this Sarkovskii sequence, initial points can be found where the same map has a period 1,000,345 point, or a fixed point, or points of any other possible period.

Period three maps

Suppose f is a continuous map from the line back to the line that has the period three point

$$f(x_1) = x_2, \quad f(x_2) = x_3, \quad f(x_3) = x_1.$$

Choose some ordering of these three points, it does not matter what it is, and then plot the three $(x_j, f(x_j))$ points. The assertion is that any way these points can be connected to form a graph of a continuous function, the resulting function admits periodic points of any period along with far more complex behavior. The choice of the mapping selected for Fig. 1.8 provides an unimaginative but minimal straight-line way to connect these points where the selected ordering is $x_1 < x_2 < x_3$.

By definition, this Fig. 1.8 mapping takes x_1 and x_2—the endpoints of the interval A—respectively to points x_2 and x_3—the endpoints of interval B. It now follows from continuity that this mapping—however it may be drawn—must, at the minimum, map interval A onto interval B.

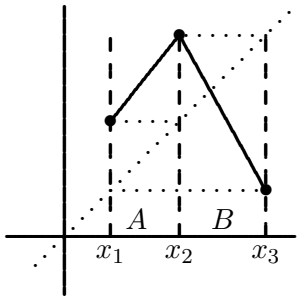


Fig. 1.8. A period three map

Similarly, since the endpoints of B, given by x_2 and x_3, are mapped respectively to x_3 and x_1—these are the endpoints of $A \cup B$—the period-three mapping must experience an *expansion* effect where interval B is reversed and then stretched to be mapped to $A \cup B$. Cataloguing this information as

$$f(\overline{A}) \supset \overline{B}, \quad f(\overline{B}) \supset \overline{A \cup B},$$

it now follows by using the above "iterated inverse image" argument that this f admits a rich variety of different dynamics. In particular, at a minimum ("minimum" because by drawing more expansive maps rather than just connecting the three points, wilder possibilities become possible), we know that any sequence is a word if it satisfies the conditions where

> an A always is followed by a B; a B can be followed by an A or a B.

For any such sequence, we know there exists an initial iterate satisfying this future. Again, it is the combination of expansion and recurrence that creates the complicated dynamics.

Presumably, a period-five point would be given by a sequence that continually repeats the block $ABBBB$, or maybe the block $ABABB$, while a period seven point is obtained from repeated blocks of, say, $ABBABBB$. More complicated orbits that avoid having any periodicity, but skirt arbitrarily close to various period points, are represented by sequences where no block ever repeats itself. One example of this is

$$ABABBABBBABBBBABBBBBA\ldots$$

where each A is followed by even longer sequence of B's.

Circle maps

The preceding paragraph cautiously states that "Presumably, a period-five point would be given by a [repeating] sequence." It does, and a simple way to prove this assertion can be illustrated by using a mapping from the circle to the circle.

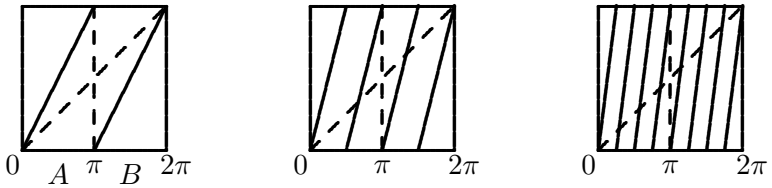
Actually, we already have analyzed a mapping $f : S^1 \to S^1$: a mapping from the circle to the circle. This is Newton's method because, by the usual trick of adding a point at infinity, the infinite line becames a circle where the infinity point is the North Pole. With this representation, Newton's method is continuous as the critical points of a function are mapped to the North Pole.

Suppose we have a simple continuous mapping $f : S^1 \to S^1$ that wraps around the circle twice: for simplicity, consider $f = e^{2\theta}$, $0 \le \theta \le 2\pi$. The goal is to demonstrate that this mapping has periodic points of all periods. By slicing the circle open at the North Pole and flattening it into a line interval $[0, 2\pi)$, the graph becomes as displayed in Fig. 1.9a.

Subintervals A and B are as designated in Fig. 1.9a, and the figure shows that

$$f(\overline{A}) = \overline{A \cup B}, \quad f(\overline{B}) = \overline{A \cup B}.$$

According to the above argument, it now follows that any sequence consisting of A's and B's can be realized by some initial iterate. Presumably, this means that a sequence repeating the block AB is a period-two point while the one repeating, say, the block ABB, is a period three point. A simple argument using the graph shows that, indeed, these are periodic points.



a. $y = f(\theta)$ **b.** $y = f^2(\theta)$ **c.** $y = f^3(\theta)$

Fig. 1.9. Periodic points of circle map

First, if a period two point does exist, then $f(\theta_1) = \theta_2$ and $f(\theta_2) = \theta_1$. Combining the terms leads to the arrangement

$$\theta_1 = f(\theta_2) = f(f(\theta_1)),$$

which means that a period two point is the fixed point for $f^2(\theta)$ where f^2 represents the composition $f \circ f$. So, a way to find the period two points is to create a rough graph for $y = f^2(\theta)$ and find whether it has fixed points. That is, we wish to identify all points where the graph of $y = f^2(\theta)$ crosses the $y = x$ diagonal.

Actually it is easy to create a rough graph of $y = f^2(\theta)$. After all, since $f^2(A) = f(f(A)) = f(A \cup B)$, the graph of f^2 over A must resemble the graph of f over $f(A) = A \cup B$; that is, the graph of f over $[0, 2\pi)$. The approach, then, is to squeeze the Fig. 1.9a graph of $y = f(\theta)$ over the interval A: this is done in the first half of Fig. 1.9b. Similarly, since $f^2(B) = f(f(B)) = f(A \cup B)$ where the orientation is preserved, the graph of $y = f^2$ over B squeezes the full Fig. 1.9a graph of f over interval B. Even if the details are not correct, the number of times this rough graph must cross the $y = x$ diagonal proves that there are two new fixed points for $f^2(\theta)$; let them be θ_1 and $\theta_2 = f(\theta_1)$.

Notice how this construction divides the full interval into four subintervals. These subintervals correspond to terms AA, AB, BA, and BB. So, the first f^2 fixed point corresponds to repeating the sequence AB while the

second one represents a repeating BA. (Notice the tacit use of the fact that f is continuous and monotonic on each interval; in particular, the f image does not reverse direction as it does with the f over B in Fig. 1.8.)

To find the period 3 points, or the fixed points of $y = f^3(\theta) = f(f(f(\theta)))$, notice that $f^3(A) = f^2(f(A)) = f^2(A \cup B)$. In other words, over the $A = [0, \pi)$ region, squeeze in the full graph of Fig. 1.9b. Similarly, since the mappings are orientation preserving, the graph of $y = f^3(B)$ is a squeezing of the graph of $y = f^2$ from Fig. 1.9b over region B. All of this identifies the Fig. 1.9c graph of $y = f^3$ with its six fixed points. Using the above argument, the eight regions from left to right are

$$AAA,\ AAB,\ ABA,\ ABB,\ BAA,\ BAB,\ BBA,\ BBB$$

so the six period three points are identified, respectively, with the repeated blocks of AAB, ABA, ABB, BAA, BAB, BBA.

The same kind of argument applies to Newton's method and the period-three graph of Fig. 1.8. The main difference is that with Fig. 1.8, the image of the B region is reversed, so the "squeezed graph" must be reversed. Of course, while Newton's method reproduces portions of regions, the existence of periodic points follows from the geometry. Simple arguments; nice conclusions!

1.4.3 The forced Van der Pol equations

All of this material is being introduced in order to describe a problem about the rings of Saturn. But, before doing so, I need to outline a nice argument developed by Mark Levi [36, 37] to analyze the periodically forced Van der Pol equation. While a plausibility argument outlining Levi's arguments suffices for my purposes, the reader is strongly encouraged to read Levi's papers to fully enjoy the details.

The periodically forced Van der Pol equations are given by

$$\epsilon x'' + (x^2 - 1)x' + \epsilon x = b \sin(t), \tag{1.19}$$

where ϵ has a small but fixed value and b is the forcing amplitude: these equations arise from the study of electrical circuits containing that ancient device of vacuum tubes. As shown next, solutions for these equations have a nice and regular "beat." Of course, each of us, if we wish to stay alive, consistently experience another kind of "regular, periodic beat" in our chest. Thus it is interesting but not surprising that Van de Pol used versions of these equations to model and understand the heart's behavior. In fact,

friends of mine from the medical field who study the mathematics of body organs know the Van der Pol equations strictly from this medical context.

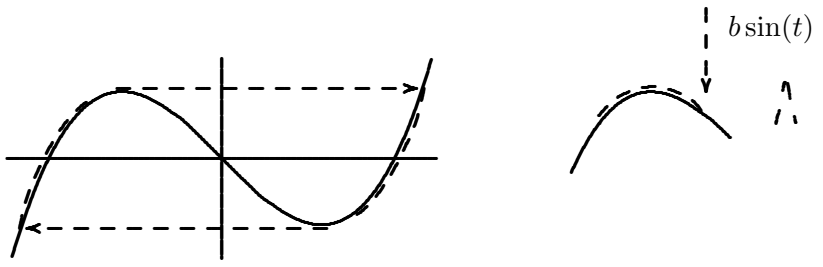
Intuition about the behavior of the Eq. 1.19 system can be obtained by rewriting these equations in the following form that uses the equality $\frac{x^3}{3} - x = \int (x^2 - 1)\, dx$. It follows that

$$x' = \frac{1}{\epsilon}(y - [\frac{x^3}{3} - x]), \quad y' = -\epsilon x + b\sin(t).$$

If $b = 0$ and ϵ has a sufficiently small value, then the intuitive sense derived from the second equation is that the miniscule y' value allows only slow and minor changes in the y value. On the other hand, the large ϵ^{-1} multiplier for x' significantly accentuates differences in the first equation. This argument suggests an active x' change causes the solution to rapidly approach and remain close to the curve given by $x' = 0$; that is, the solution should remain near the curve

$$y = \frac{x^3}{3} - x.$$

The graph of this curve is given in Fig. 1.10a.



a. $y = \frac{x^3}{3} - x$ **b.** Effects of the $b\sin(t)$ force

Fig. 1.10. The forced Van der Pol equation

This story argues that the solution should follow reasonably close to the dashed lines where, when sliding over the two precipices given by the local maximum and minimum of $y = \frac{x^3}{3} - x$, the solution rapidly moves to the other leg in an essentially horizontal manner. (It is essentially horizontal because y' has a small value compared to x'.) This argument has been made precise in different ways.

Now consider the periodic forcing effect of $b\sin(t)$. To develop intuition for what should happen, Levi considered the effect of the dynamics on line segments, or strips, of initial conditions. Without the forcing term, this line

segment moves like a well-behaved water snake sliding along the current given by the dashed Fig. 1.10a curve. Something different, however, can occur with the forcing function and a sufficiently large value of b.

At this stage we need a "thought experiment." Suppose, just suppose, that when the strip—that snake—starts to slide over the edge and prepares for a rapid movement to the right as depicted in Fig. 1.10b, the $\sin(t)$ timing allows a large b value to force portions of the strip downwards. The front part of our snake, then, is forced to twist into a "V." However, only a portion of this strip encounters this force; the $\sin(t)$ term reverses sign so that it no longer is forcing the solution downwards. This means that the rest of the strip is not affected when it crosses the threshold. The combined effect is that our "snake" is chasing its tail to create a compressed and rotated "V" that now slides along the solid Fig. 1.10 curve.

The consequences should be fairly clear. Instead of a long strip monotonically making its rounds of the curve, this strip has a kink as represented by the dashed lines to the right of Fig. 1.10. Thus this kinky affect, somewhat resembling the graph in Fig. 1.8, goes around and around to get further kinks representing higher iterates. Consequently, with appropriate care, an analysis similar to that given about Fig. 1.8 applies showing the chaotic effects allowed by this system.

With this approach of following a strip of initial conditions, Levi reduces portions of periodically forced systems to maps from the interval to the interval. In other words, this kinky mathematics captures subtle expansion and recurrence effects of the motion: it identifies potentially complex dynamical behavior.

What is next?

A message to be taken from the above geometric arguments is that it is possible for the combination of *expansion* and *recurrence* of motion to cause surprisingly complex dynamical behavior. This claim is of particular interest to anyone interested in the N-body problem and celestial mechanics. After all, our solar system enjoys plenty of recurrence with the planets making their regular trips about the Sun. Expansion? Well, the first two sections of this chapter showed how expansion and strong forces accompany close approaches of even well-behaved bodies. Maybe this combination will lead to interesting problems. As indicated in the next section, this is the case.

1.5 The rings of Saturn

This chapter started with Galileo's forced recantation of his views: it ends with a problem linked to his wondrous 1610 discovery of the rings of Saturn by use of his telescope.[11] A problem, of course, is to understand why the rings are there. This is a general issue because, as space exploration has proved, rings have been sighted with other planets. While the rings of other planets are not as dramatic, the important point is that they are there. What are the dynamics? One aspect of these dynamical concerns will be addressed in this section; another will be described at the end of the third chapter.

We now know that while the ring system extends quite a distance from Saturn, the thickness is, in fact, quite thin on the astronomical scale: it is only that of a two story building. The particles forming the rings range from dust to objects about 15 feet in diameter. But rather than forming a nice, circular ring, there are portions that resemble the braiding of hair. The research problem proposed here—a problem that is not necessarily easy but it does seem to be doable— is to develop a dynamical explanation.

1.5.1 Kinky behavior

To provide background, in April, 1973, Pioneer 11 was launched on a long journey to visit Jupiter; later it dropped by Saturn to make the first direct observations of this planet (in 1979). One of the Pioneer 11 discoveries was a new ring for Saturn, imaginatively called the "F ring." What made the discovery unusual, as clarified by pictures from a November 1980 visit by Voyager 1, is that the F-ring appeared to be involved in some kind of "kinky" behavior: two of its three strands provide a braided appearance. Whatever the source, this kinkiness appears to be have been short-lived, rather than a permanent phenomenon, as judged by the pictures of Voyager 2 taken less than a year later. At the later time, the F-ring now was more regular with non-intersecting braided strands.

Adding to the mix are the two shepherds of the F-ring, Prometheus and Pandora.[12] Prometheus, the inner moon, is named after brother of Atlas and Epimetheus—he is the one who stole fire from Zeus and gave

[11] More precisely, Galileo knew he found something, but, because of the limited power of his equipment, he was not completely sure what were the "ear like" appendages that he observed. Were these moons, or, as Galileo initially thought, two stars circling Saturn? It took a stronger telescope to recognize that they are rings.

[12] Both of these moons of Saturn were discovered in 1980 by S. A. Collins and D. Carlson by carefully examining Voyager 1 photos.

it to mankind. Pandora, the outer moon, is named after the first woman who was created by Zeus to punish man for Prometheus' stealing of fire—the punishment consisted of the evils released once Pandora opened her "Pandora's box." The astronomical situation, not the mythological one, is beautifully displayed in one of Voyager's photos given by Fig. 1.11.

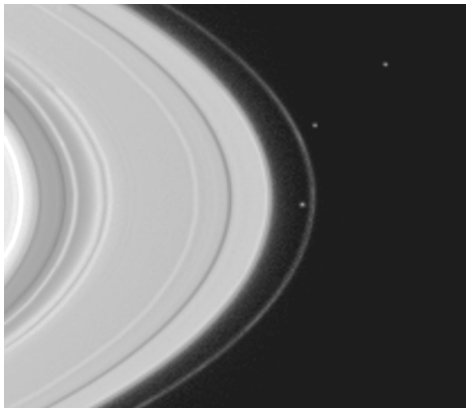


Fig. 1.11. Prometheus, F-ring, and Pandora;
this picture is used thanks to the courtesy of NASA/JPL-Caltech.

The question is whether the motions of these two moons are responsible for the braiding of the F-ring. This is conjecture is so natural that I must assume it has been advanced by many people, yet I do not know of anyone who has provided a mathematical verification of this suggestion. On the other hand, by pulling together all that has been discussed in this chapter, it is possible to propose a natural model—and a mathematical approach—to explain these braids. Let me outline the notions.

1.5.2 A model

The first step is to compute a portion of the obit of Pandora relative to the position of Prometheus (i.e., put the system in a rotating coordinate system based on the motion of Prometheus). According to what we determined earlier (page 5) when discussing the orbit of Mars relative to the Earth, we know that in this system Pandora's orbit must exhibit a loop. Even more: the size of the loop can be slightly enhanced because, although the eccentricity of Pandora's orbit, 0.0042, is small, it is about twice that of Prometheus's eccentricity of 0.0024. By using the earlier argument about the apparent location of the Sun as viewed from Mercury, it can be shown that the loop size of Pandora's apparent orbit can be slightly expanded.

Now consider a dust particle circling Saturn between Prometheus and Pandora. Just because the particle goes about Saturn, its motion has a sense of "recurrence." But, should that particle be somewhere near Pandora when her looping occurs, this proximity creates a stronger pull on the dust—it generates an expansion effect. In other words, the F-ring dynamics, as modified by the presence of Prometheus and Pandora, provides a setting of "expansion and recurrence." This setting is precisely what is needed to suggest the complicated dynamics discussed earlier.

The situation is depicted in Fig. 1.12 where the dot on the left locates the position of Prometheus, the curved line corresponds to a strip of particles in the F-ring, and the loop represents the close incursion of Pandora. To understand the effects of the loop, borrow Levi's approach of using a moving strip—his snake traveling through the region—and consider what happens to various parts of its body. To do so, start with Pandora on the upper part of the loop not yet near its closest approach and where a small F-ring strip is not quite symmetrically centered on the figure because its head is on the circle directly between Pandora and Prometheus. For purposes of argument, assume that both particles in the F-ring and Pandora are moving in a general counter-clockwise direction. (This assumption just corresponds to whether we are looking at the ring from the North or South pole of Saturn.) According to Kepler's second law (Eq. 1.10), particles in the F-ring are moving faster than Pandora.

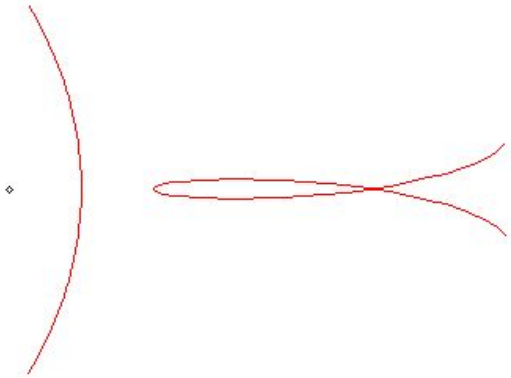


Fig. 1.12. Prometheus, F-ring, and Pandora in a rotating system

The "thought-experiment" shows that with this orientation, Pandora pulls harder on the snake's head than tail. As the snake slithers past, Pandora reaches the bottom of the loop and now exerts an even stronger pull on the snake's midsection. With its faster motion, the snake moves on,

while Pandora is receding in the opposite direction causing the tail to be influenced more by Prometheus than Pandora.

What we have, then, is that various portions of the snake are pulled toward Pandora, and even pulled backwards when the relative motion of Pandora is going backwards. While I know nothing about knitting, this descriptions of how the strip is being pulled and pushed around sounds much like the "Knit one, purl two" phrases uttered by my grandmother.

Now, if two braids that normally would be well behaved run into this looping effect, the differences in speed (again, according to Kepler's law) would require the strips on the braids to encounter this difference in gravitational pull at changing times; we must expect that they would cross and look like a braid. Can this (weak) plausibility argument be made into a more rigorous argument? I don't know, but all the elements suggesting something like this must occur are in place. To continue, notice how this argument suggests that when Prometheus and Pandora are sufficiently separated—a setting where we will not expect this kinky looping behavior to occur— then we must expect the appearance of the F-ring to be more standard: this is consistent with what Voyager 2 reported.

The challenge is given: the technical difficulty derives from the added degrees of freedom that are above that of the Levi example. Yet, I expect that this is the kind of problem that can be solved.