Kepler's Area Law in the Principia: Filling in some details in Newton's proof of Proposition 1

Michael Nauenberg
Department of Physics
University of California, Santa Cruz, CA 95064

Abstract

During the past 30 years there has been a controversy regarding the adequacy of Newton's proof of Prop. 1 in Book 1 of the *Principia*. This proposition is of central importance because its proof of Kepler's area law allowed Newton to introduce a geometric measure for time to solve problems in orbital dynamics in the *Principia*. It is shown here that the critics of Prop. 1 have misunderstood Newton's continuum limit argument by neglecting to consider the justification for this limit which he gave in Lemma 3. We clarify the proof of Prop. 1 by filling in some details left out by Newton which show that his proof of this proposition was adequate and well grounded.

Rigor merely sanctions the conquests of sound intuition, Jacques Hadamard

Abstract

Au cours des 30 dernières années, il y a eu une controverse au sujet de la preuve de la Proposition 1 telle qu'elle est formulée par Newton dans le premier livre de ses Principia. Cette proposition est d'une importance majeure puisque la preuve qu'elle donne de la loi des aires de Kepler permit Newton d'introduire une expression géometrique du temps, lui permettant ainsi de résoudre des problèmes dans la domaine de la dynamique orbitale. Nous démontrons ici que les critiques de la Proposition 1 ont mal compris l'argument de Newton relatif aux limites continues et qu'ils ont négligé de considérer la justification pour ces limites donnée par Newton dans son Lemme 3. Nous clarifions la preuve de la Proposition 1 en ajoutant des détails omis par Newton, détails qui montrent que sa preuve de cette proposition et compétent et bien fondée.

Key Words: Kepler's area law, Newton's Principia

1. Introduction

In Prop. 1 of the *Principia* Newton gave a proof that Kepler's empirical area law for planetary orbits and the confinement of these orbits to a plane are consequences of his laws of motion for the special case of central forces. In his words,

> "The areas which bodies made to move in orbits described by radii drawn to an unmoving center of force lie in unmoving planes and are proportional to the times"

This proposition is justifiably regarded as a cornerstone of the Principia, because the proportionality between the area swept out by the radius vector of the orbit and the elapsed time enabled Newton to solve dynamical problems by purely geometrical methods supplemented by continuum limit arguments which he had developed. Although the validity of Newton's proof was not questioned by his contemporaries, an alternative analytic proof of the area law was given later by Jacob Hermann based on the analytic form of the calculus which had been developed by Newton and by Leibniz (Guicciardini 1999). During the past years, however, two influential historians of science, D.T. Whiteside and A.J. Aiton, have criticized Newton's proof, claiming that it was inadequate and that it applied only to an *infinitesimal* arc of the orbit (Whiteside 1974), (Aiton 1989). Whiteside remarked that there were underlying subtleties in the proof that Newton did not fully appreciated, and that Newton continued "to believe in its superficial simplicities" although, Whiteside admitted, not even " Johann Bernoulli, his arch critic ... saw fit to impugn the adequacy of Newton's demonstration" (Whiteside 1991). The basis for the Aiton-Whiteside criticism is that Newton had treated incorrectly the continuum limit of a discrete polygonal orbit due to a sequence of central force impulses. Subsequently, this criticism has been accepted by many Newtonian scholars although some arguments have been presented that it is not valid (Erlichson 1992), (Nauenberg 1998a). For example, in his new translation and guide to Newton's *Principia*, I.B. Cohen warmly endorsed Whiteside's analysis (Cohen 1999), while N. Guicciardini in his new book *Reading the Principia* questioned whether Newton's limit arguments in Prop. 1 are well grounded, although acknowledging dissenting views (Guicciardini 1999). Other authors discussing the *Principia* either neglected to examine the validity of Newton's limit arguments in Prop. 1 (Brackenridge 1995) (Chandrasekhar 1995), or failed to understand them (Densmore 1995). More recently, Pourciau has argued that if one takes a "traditional view what Newton means by orbital motion", Prop. 1 contains in addition to mathematical inadequacies also logical flaws (Pourciau 2002). This is by far the most serious criticism, but it is also not valid. As will be shown here, in 1679 Newton discovered Prop. 1 precisely because at this time he turned to a *non-traditional* view of orbital motion which had been suggested to him by Robert Hooke (Nauenberg 1994b, 1998).

In this paper I review the historical circumstances which led Newton to his momentous discovery that central forces account for Kepler's area law, and I present arguments to refute the criticisms of his proof of this law in Prop. 1. In order to understand this fundamental proposition, and to avoid misconceptions voiced by recent commentators, it is helpful to keep in mind the historical context and the manner in which Newton discovered Prop. 1 outlined in section 3. In section 4, I discuss in detail Newton's mathematical procedure to obtain a continuous orbit as the limit of a discrete polygonal orbit which is the subject of recent controversies discussed in section 2. In particular, I show also how in this limit a sequence of discrete impulses leads to a continuous force with a measure which Newton described subsequently in Prop. 6. In this proposition, Newton started with a continuous orbit which satisfies the area law, and then obtained a measure for the central force by a somewhat different limit argument from the one which he had presented previously in Prop. 1. In Prop. 6 this measure is the acceleration in units of time which according to Prop. 1 are proportional to the area swept by the radius vector. In the following discussions the reader should consult Props. 1 and 6 which are not reproduce here except for some quotations and the diagrams shown in Figs. 1 and 4. These quotations come from a new English translation of the original Latin version of the *Principia* by I.B. Cohen and Ann Whitman (Cohen 1999).

Before discussing the controversy regarding the validity of Newton's proof Prop. 1, it may be helpful to outline below some of the main points emphasized in this paper. After a lengthy correspondence with Robert Hooke in 1679 (Nauenberg, 1994b, 1998), Newton considered the mathematical consequence of thinking of a continuous force as the limit of a series of impulses which give rise to changes of velocity over a vanishingly small interval of time. For impulsive forces, it the orbit consists of the sides of a polygon, which is the discrete version of an orbit such as Hooke had indicated in a letter to him, of

> " ...of a direct motion by the tangent and an attractive motion towards the central body.."

Newton's polygonal construction in Prop. 1 (see Fig. 1), which is discussed in sections 3 and 4, implements mathematically Hooke's idea for orbital motion, provided the force impulses are directed toward a common center S. In this case the resulting polygon is in a plane with its orientation determined by the direction of the initial velocity and the initial position with respect to the center S (when these two initial directions are the same,the orbit degenerates into a line). Newton must have regarded the proof of planarity for impulsive forces as fairly obvious, because all he said about this subject in Prop. 1 is that

> "...all these lines [the sides of his polygon] lie in the same plane"

The proof of the area law is also simple, but requires the application of Euclidean geometry which Newton provided in Prop. 1

The main difficulty with the proof of Prop. 1 arises in connection with the *continuum limit*, when the number of impulses increases indefinitely while the time between them becomes vanishingly small. According to Newton, this limit gives rise to a continuous force, but Newton does not spell out this continuum limit in any detail, referring instead the reader to Cor.4, Lemma 3. When one looks up this lemma one finds a polygonal construction very similar to that in Prop. 1, except that instead of triangles it consists of parallelograms. In this lemma, as in the previous one, Newton describes his method for obtaining the area and perimeter of a curve by considering the continuum limit

> "...as the maximum width [of the parallelograms] is diminished indefinitely."

This is the same language that Newton used in describing the corresponding limit which appears in Prop. 1. In Lemma 3, a curve is *given* , which according to the accompanying figure and arguments decreases monotonically. In this lemma it is clear from the geometry how to construct the associated parallelograms with specified widths. But in the figure associated with Prop. 1 Newton does not show an orbital curve associated with his polygonal construction. One of the main points, which will be elaborated in section 3, is that by referring to Lemma 3, Newton had in mind the existence of such a curve which fixes the location of the vertices of the polygons in Prop. 1, much in the same way that the curve in Lemma 3 determines the location of the upper corners of the parallelograms in this lemma. In this way, Newton's reference to Lemma 3 in support of the continuum limit can be justified, apart from some problems concerning convergence. These problems are outside the scope of the mathematics of the *Principia*, and will be discussed in Appendix A. In appendix B the theorem and proof of Prop. 1 are presented in modern notation. Our summary and conclusions are presented in section 5, which hopefully will help resolve the current controversy over the validity of Newton's proof of Prop. 1.

2. The controversy regarding Prop. 1

The current controversy with Prop. 1 arises mainly because Newton's only statement describing his continuum limit argument in the *Principia* is very succinct:

> " Now let the number of triangles be increased and their widths decreased indefinitely, and their ultimate perimeter ADF will (by lemma. 3, corol. 4) be a curved line ..."

Apparently left unexplained is how this *indefinite* increase in the number of triangles and the corresponding reduction of their widths would have to be tailored to lead to a well defined and unique continuum limit. But to understand Newton's procedure one has to consult Lemma 3 which was given by him as justification for his limit arguments. Remarkably, neither

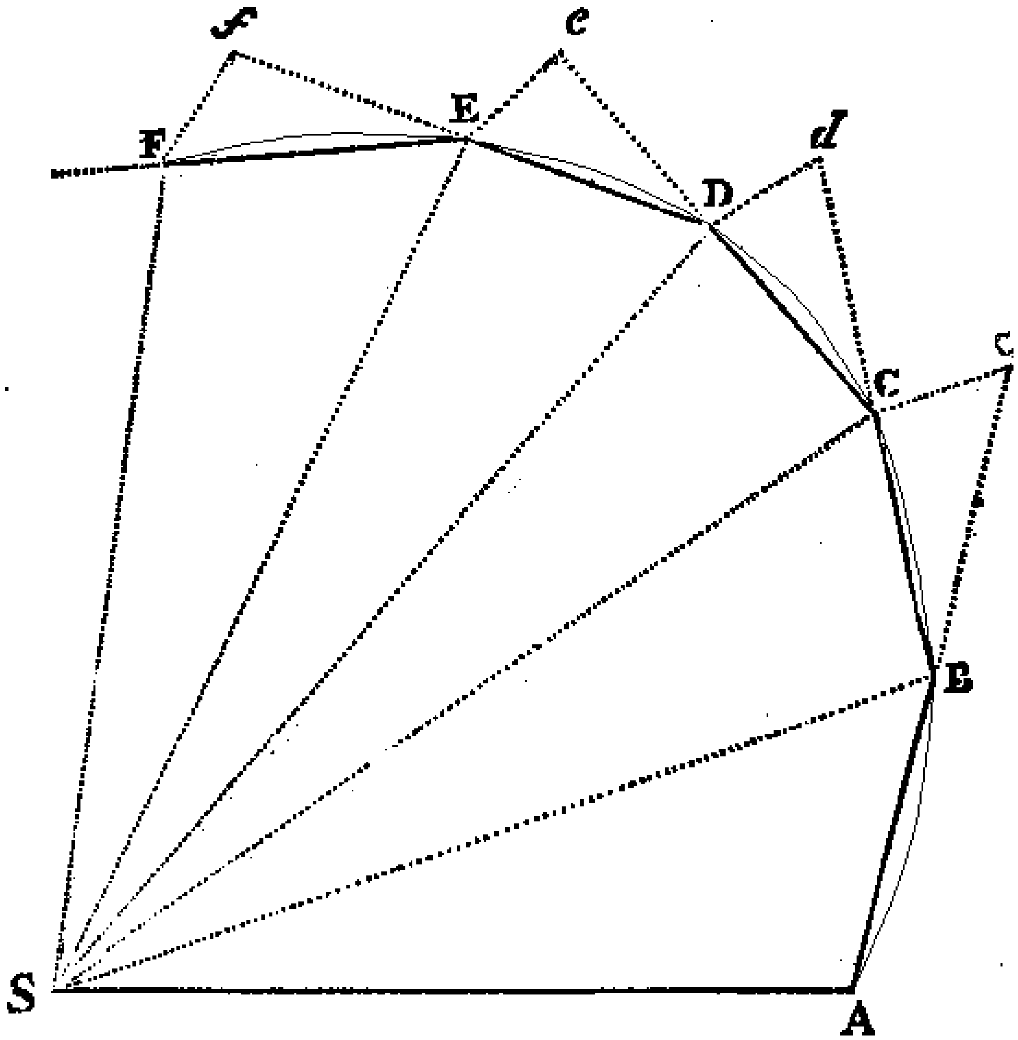

Figure 1: This diagram corresponds to Newton's diagram in Prop. 1, but with the lines Sc, Sd, Se and Sf deleted, and with an additional curve through points $ABCDEF$.

Whiteside (MP 6: footnote 19) (Whiteside 1991) nor Aiton (Aiton 1989) commented on this important Lemma, which was also neglected by one of their critics (Erlichson 1992), and by other recent commentators of the *Principia* (Brackenridge 1995) (Chandrasekhar 1995) (Cohen 1999) (Densmore 1995). In Lemmas 2 and 3 and its corrolaries, Newton described how the area bounded by a given curve and a line and the length of the curve can be approximated by a sequence of rectangles (parallelograms). In Lemma 2, he proved rigorously the existence of a limit for this area by obtaining lower and upper bounds given by the area of the inscribed and circumscribed rectangles of equal width, showing that for a monotonic curve the difference between these two bounds is the area of the first rectangle. Consequently, as the number of these rectangles increases indefinitely while their widths approach zero this difference vanishes, and the sum of the area of these rectangles approach the same limiting value. By definition, this continuum limit is the area under the curve. Indeed, modern calculus books reproduce Newton's proof for the area under a curve, but attribute it to later mathematicians. In Lemma 3 Newton *extended* his proof in Lemma 2 to the case of rectangles of unequal width,

> "The same ultimate ratios are also ratios of equality when the widths AB,BC,CD,... of the parallelograms are *unequal* [my italics] and are all diminished indefinitely".

It is interesting to speculate that this extension was included in the *Principia* primarily to make this Lemma applicable to Prop. 1, because the parallelograms associated with the vertices in the corresponding diagram, taking the initial radial position AS as the horizontal axis, see Fig. 1, would have to have unequal widths. In Cor. 2 of this Lemma Newton asserted that

> " ...the rectilinear figure that is comprehended by the chords of the vanishing arcs...coincides ultimately with the curvilinear figure,"

and in Cor. 4 Lemma 3 he concluded,

> "And therefore these ultimate figures (with respect to their perimeter acE) are not rectilinear, but curvilinear limits of rectilinear figures."

Hence, Newton's reference to Lemma 3 suggests that in Prop. 1 Newton envisioned that the vertices of the polygon in his diagram, Fig. 1, are located on a geometrical curve which remains fixed as the number of vertices in this polygon increased indefinitely. Newton's construction of this polygon requires that these vertices all lie on a plane, and consequently this curve must lie in the *same* plane. This plane is fixed by *initial conditions*, e.g. the initial position radius SA and the direction of the initial velocity AB shown in Fig. 1, which do not change as the continuum limit is approached, although this was not explicitly mentioned by Newton. The limit curve, however, is not shown in the diagram associated with Prop. 1, which has mislead most commentators of this proposition who did not consult Lemma 3. Referring to Fig. 1, it can be seen that with this interpretation Newton's entire polygonal construction is determined by fixing the length of the first chord AB. This construction proceeds as follows: the extension Bc of this chord is set equal in length to AB, and the first deflection Cc is determined by the condition that it is a line parallel to BS starting at c which intersects the given curve at the point C. This procedure is iterated with the next chord BC which is now determined, by setting its extension Cd equal in magnitude to BC and the deflection Dd parallel to CS, starting at d, and ending at the intersection D with the given curve. This iterative process continues until the last point F on the curve is reached. The extension Bc of the chord AB and the deflection Cc must lie on the plane of the initial triangle SAB and therefore the vertex C is also on the same plane. Similarly, this property holds also for all subsequent vertices of this polygon, or as Newton stated in his proposition,

> "...making the body ...describe the individual lines $CD, DE, EF, ...$, all these lines will lie in the same plane"

For these lines to intersect a given curve, this curve must also lie in this plane. The orientation of this plane in space is determined by the initial radial position AS and by the initial chord AB which is in the direction of the initial velocity for the polygonal orbit. Furthermore, in the continuum limit the length of the chord AB vanishes, while the plane's orientation remains unchanged, because the direction of AB must approach that of the initial velocity which is directed along the tangent of the curve at the initial position at A.

In Prop. 1 Newton did not specify directly how the magnitude of the deflections $Cc, Dd, ...$ are obtained, nor how the magnitude must vary with the number n of vertices. We have seen, however, that these deflections can be determined by the assumption that the polygon vertices are attached to a fixed planar curve. An alternative possibility, which in fact was considered by Hooke, (Nauenberg 1994b, 1998b) is that these deflections depend on the radial distance of the vertices of the polygon. Then to obtain the continuum limit a rule has to be given for how these deflections scale with the number of vertices. In this case Newton could have invoked his curvature lemma, Lemma 11, which implies that the deflections scale as the square of the length of the adjacent sides of the polygon, or Lemma 10 that the deflections scale as the square of the time between pulses. But instead, in Prop. 1 Newton cited Lemma 3 which is relevant to the continuum limit when a fixed curve is given.

Newton's discrete construction in Prop. 1 refers to a sequence of triangles rather than rectangles as in Lemma 3, but it is easy to see that this lemma remains also valid in this case. If the rectangles in Lemma 3 are replaced by trapezoids obtained by connecting the intersections with straight lines then the area of these trapezoids is greater than the area of the inscribed rectangles, but smaller than the area of the circumscribed rectangles. Therefore, in the limit that the width of the rectangles is diminished indefinitely, the area of these trapezoids also coincides with the area under the curve. But the total area of the trapezoids is the same as the area of the triangles associated with Prop. 1. There is an important detail, however, regarding this application which needs some clarification. As will be shown below, given a curve of finite length Newton's polygonal construction does not in general cover the entire curve for a finite number of triangles. This problem, however, disappears in the limit that the number of triangles increases indefinitely, and a proof is given in Appendix A.

Newton's description of the continuum limit quoted above continues as follows:

> "...and thus the centripetal force by which the body is continually drawn back from the tangent of this curve will act uninterruptedly..."

In Corols. 3 and 4 to Prop. 1 Newton stated that the ratio of forces at two distinct point on the curve where given by the limit of the ratio of the displacements caused by central force impulses at these points, but he did not show that in the continuum limit the measure of this force is proportional to the measure for force which he gave in Prop. 6. This is another detail that will be discussed after the next section.

3. The historical context of Prop. 1

To understand Newton's proof of Prop. 1 is is necessary to to know what Newton's meant by the term *orbit* in the statement of this proposition (see Introduction). In the *Principia* the term *orbit* is not defined explicitly, but it has been generally understood to mean a geometrical curve which describes the position of a moving body in space. Mathematically an orbit is a continuous curve which is parameterized by the time variable. In Prop. 1, however, Newton had to have a more restrictive definition, because he was dealing there with the special case of the motion of a body under the action of a central force, or in his words,

> " ... bodies made to move in orbits.. by...an unmoving center of force".

To elucidate this point we turn now to the historical circumstances which led Newton to discover this crucial proposition.

One of Newton's early ideas about orbital motion was to consider the action of a continuous force as the limiting case of a sequence of force impulses. As can be seen from his earliest surviving drafts on orbital motion in the *Waste Book* (Herivel 1965) (Whiteside 1991), Newton approximated circular motion by a regular polygon with its vertices located on a circle (see Fig. 2). He also obtained an expression for the continuous force as the limit of force impulses (Brackenridge 1995). But apparently he did not generalize this idea to non-circular motion until shortly after his correspondence in 1679 with Robert Hooke (Nauenberg 1994b), who suggested to him a somewhat similar conceptual scheme to understand the orbital motion of planets moving around the sun. On Nov. 24, 1679 Hooke had written to Newton

> " And particularly if you will let me know your thoughts of that compounding the celestiall motions of the planetts of a direct motion by the tangent and an attractive motion towards the central body..."

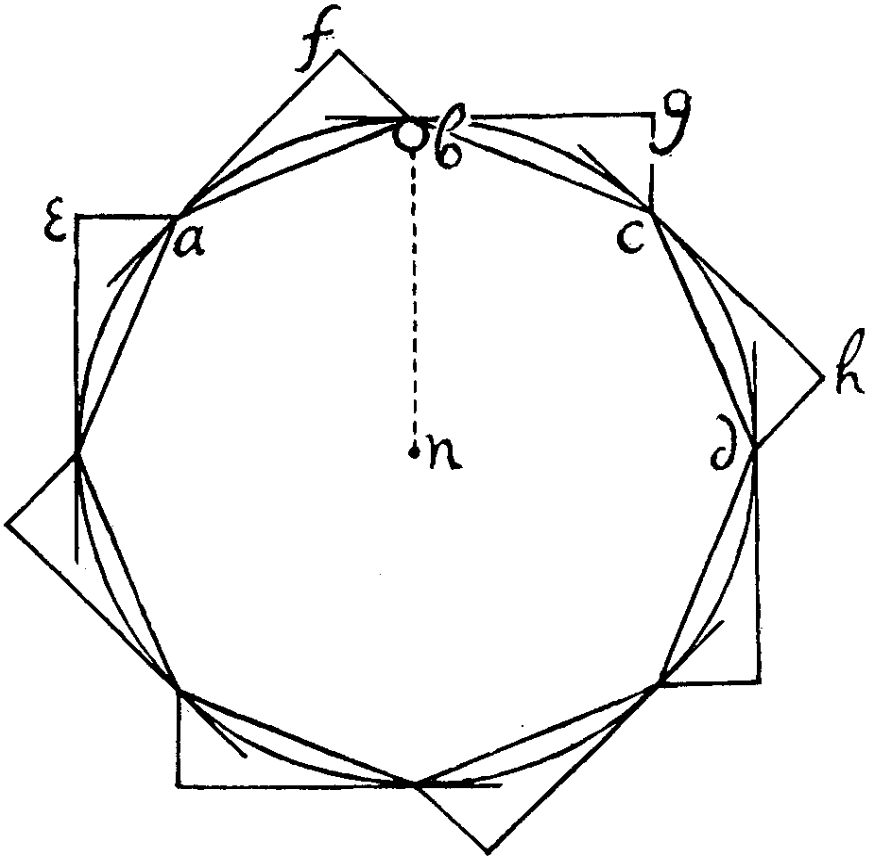

Figure 2: This diagram in Newton's *Waste Book* shows an octagon inscribed in a circle. Here the deflections $gc, hd...$ are shown from the respective tangents bg, ch...rather than from the extensions of the corresponding chords $ab, bc...$ as shown in Fig. 1

Indeed, years later Newton recalled that

> " In the year 1679 in answer to a letter from Dr. Hook ... I found *now* [my italics] that whatsoever was the law of the forces which kept the Planets in their Orbs, the area described by the Radius drawn from them to the Sun would be proportional to the times in which they were described..."

Prop. 1 appeared for the first time as Theorem 1 in a short manuscript, *De Motu*, that Newton had sent in 1684 to Halley containing the beginning draft of what became later his *Principia*. In this manuscript Newton describe the continuum limit in words similar to those he used later,

> " Now let these triangles be infinite in number and infinitely small, so that each individual triangle corresponds to the individual moment of time, the centripetal force acting without diminishing and the proposition will be established"

Apart from the mathematical language, which is far less precise than the language in Prop. 1 quoted in section 1, it is noteworthy that Newton gave no reference here to any lemmas which would justify his limit argument. In *De Motu* he did not give even a hint on how to establish a continuum limit. Nevertheless, Hooke, who was one of the first members of the Royal Society to see this manuscript (Nauenberg 1994b,1998b) recognized that for a finite number of impulses Newton's polygonal construction gave an approximate solution for orbital motion along the lines which he had suggested to Newton in his Nov. 24, 1679 letter quoted above. The best evidence for this supposition is that shortly after the appearance of *De Motu*, Hooke implemented Newton's construction as an *algorithm* to construct the orbit when the magnitude of the force impulses is proportional to the distance from the center. In a manuscript dated Sept. 1685, almost two years before the *Principia* was published, Hooke presented a remarkably accurate graphical drawing of an elliptical orbit (Nauenberg 1994b,1998b) with its center located at the center of force by setting the deflections proportional to the distance from the center. An enlarged version of the upper part of his diagram, excluding some auxiliary lines, is shown in Fig. 3. This enlargement reveals quite clearly the relation of Hooke's diagram to Newton's diagram in Prop. 1, which is shown in Fig. 1. The main difference is that in Hooke's figure the initial position is

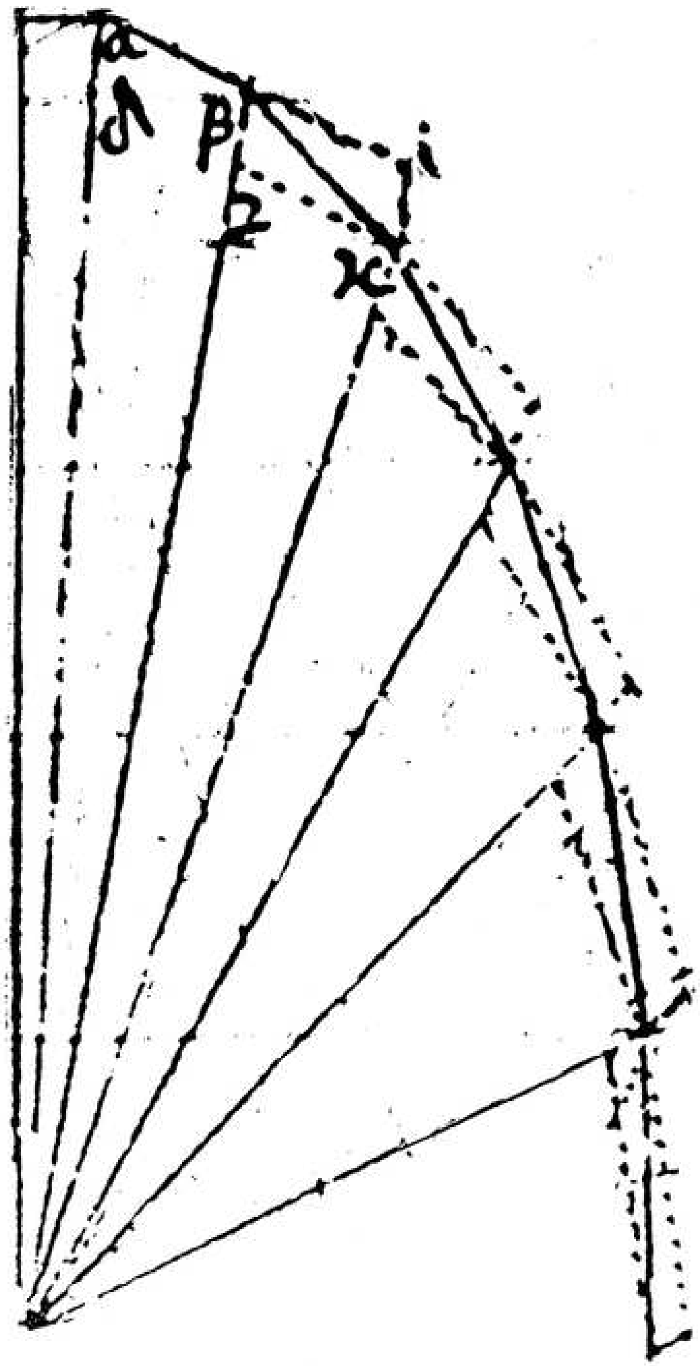

Figure 3: This diagram is a blowup of the upper part of Hooke's Sept. 1685 diagram (see reference [12]) with some auxiliary lines deleted to show more clearly its correspondence with Newton's diagram in Prop. 1, see Fig. 1.

located above the center of force and the initial velocity points to the right, leading to clockwise motion, while in Newton's figure the initial position is to the right of the center of force and the initial velocity is directed upwards leading to counter-clockwise motion. Indeed, Hooke had also conjectured that the force of gravity consisted of discrete pulses. In one of his Cutlerian lectures entitled a *Discourse on the Nature of Comets*, read at a meeting of the Royal Society soon after Michaelmas 1682, but published only after his death, Hooke speculated that bodies emitted periodic gravitational pulse in analogy with his theory of sound and light, and deduced that the intensity decreased with the inverse square distance from the source:

> "This propagated Pulse I take to be the Cause of the Descent of Bodies towards the Earth... Suppose for Instance there should be a 1000 of these pulses in a Second of Time, then must the Grave body receive all those thousand impressions within the space of that Second, and a thousand more the next...(Hooke 1705)"

But the important question of how the magnitude of the deflection caused by the force impulses scales with the size of the triangles, which is an essential ingredient to establish the existence of a continuum limit in this application of Newton's polygonal construction, was not - and could not - have been raised by Hooke. Later, in Lemma 11 on curvature which appears in section 1 of the *Principia*, Newton showed that these deflections scale with the size of the adjacent chord or arc length as the square of these quantities.

From the foregoing it is therefore reasonable to conclude that in Prop. 1 Newton had in mind that any orbit under the action of central forces is the continuum limit of a polygonal orbit *caused* by the action of a sequence of force impulses. In this case the body moves along straight lines between impulses, or by "direct motion" as envisioned by Hooke, where this motion is along the sides of a polygon, while the force impulse give rise to a linear deflection or an "attractive motion" towards the center. Since all the impulses are directed to this common center the resulting polygonal orbit is in a plane as Newton demonstrated in Prop. 1. The orientation of this plane is determined by the direction of the velocity and the position of the body relative to the center of force at some initial time. That Newton was well aware of the important role of initial conditions to fix the orbit is demonstrated by Prop. 17 where he discussed these conditions for the case of elliptical motion under the action of inverse square forces,

> " Suppose that the centripetal force is inversely proportional to the square of the distance of places from the center ... it is required to find the line which a body describes when going forth from a given place with a given velocity along a given straight line."

Setting the sequence of central force impulses at *equal* time intervals, Newton gave a proof in Prop. 1 that the areas of the triangles associated with the resulting polygonal orbit are equal. Since planarity as well as this area law are properties of any polygonal orbit due to central force impulses, it is reasonable to expect that these properties remain also valid in a *properly* defined continuum limit, but Newton did not give any details about how this limit is obtained apart from the brief sentence quoted in our Introduction, and his reference to Cor. 4 Lemma 3. In the next section we will attempt to fill in some of the details left out in Newton's discussion.

In the corrolaries to Prop. 1 Newton referred to lines AB, BC, etc. as chords of arcs. Therefore, it is tempting to argue that this gives further evidence that in order to describe a continuum limit, Newton considered that the vertices of his polygonal construction were attached to a fixed curve as the sides of the polygon decreased indefinetely. But there are problems with this interpretation, because for any finite polygon the discrete times associated with the vertices of this polygon is different from the corresponding times defined by the continuous orbit. The reason is that the area of a finite triangle formed by the chord of an arc of the curve and the center S, which is proportional to the time interval along the chord, differs by a small amount from the area of the "pie" bounded by this arc, which is proportional to the time interval along the arc. But in the corrolaries to Prop. 1 Newton was not clear in making this distinction. For example, he began Cor. 2 of Prop. 1 with the definition

> "... chords AB and BC of two arcs successively described by the same body in equal times..."

and concluded, without proof, that the diagonal BV of the parallelogram $ABCV$ formed from this chords is directed towards the center S in the limit that "those arcs are decreased indefinitely". It appears that "equal times" refers here to the time interval along the continuous orbit, and in this case, for finite arcs, the chord BC does not correspond to the side BC obtained from the polygon construction in Prop. 1, and the diagonal BV is not directed toward S. But this does not correspond to what is illustrated in the figure associated with Prop. 1 which shows BV as a segment of BS. Moreover, starting with the same definition as in Cor. 2, in the next corollary Newton argued that BV is equal to Cc, the deflection "generated by the impulse of the centripetal force at B ", which *is* parallel to BS by construction. But this is true only if the equal time intervals are those associated with the polygonal construction in Prop. 1, and the chords AB and BC are also sides of the polygon.

4. Filling in some details of Newton's proof of Prop. 1

Referring to the diagram in Prop. 1, see Fig.1, we assume that the vertices A, B, C, D, E and F of Newton's polygon are located on a given curve. Since this polygon is planar we expect that this curve should also lie on the same plane We will show that in the continuum limit, Newton's polygonal construction determines a parameterization of this curve as a function of time, describing orbital motion under the action of a central force centered at the point S, and that in this limit the magnitude of the central force obtained from Prop. 1 is equivalent to that defined in Prop. 6. There are some restrictions on the possible planar curves which can support Newton's polygonal construction. For example, the radius vector $\vec{r}$ with origin at S cannot become tangential to the curve, because in the neighborhood of any such point Newton's polygonal approximation cannot be constructed. This construction also fails when the curve crosses this origin, which corresponds to orbital motion when the central force diverges as $1/r^3$ or faster, and when the curvature approaches infinity. Therefore, our discussion will be confined to regions of space where the central force and the curvature of the orbit remain finite.

As shown in the Introduction, given the length of the initial chord AB it is evident that Newton's polygonal construction is uniquely specified by the condition that the vertices of the polygon lie on a *fixed* planar curve. That Newton had such a given curve in mind, although it did not appear in the diagram in Prop. 1, is clear from his reference to Corol. 4, Lemma 3 for the continuum limit, as we argued in detail previously. While in this lemma the approximation to a continuous curve is discussed for a subdivision in rectangles, the extension to triangles is quite straightforward, but there is a detail which needs to be worked out: if A is the initial point of the curve then unless the length of the initial chord AB is suitably chosen the last point F of the polygon will in general not lie at the endpoint of the given curve. But in the continuum limit this is not a problem. Suppose that the last vertex F occurs before the endpoint of the curve. Apply Newton's construction by extending chord EF to g, and draw a line from g parallel to FS. Then either a) this line intersects the curve at a new vertex G or b) it does not intersect the curve at all. In case a) repeat Newton's construction until case b) is reached. When F is the last vertex, and deflections due to the central force impulse are small, it is expected that the distance of F from the endpoint of the curve decreases as the length of the initial chord AB is decreased. Then in the continuum limit all the chord lengths becomes vanishingly small, and the last point F converges to the end point of the curve. A rigorous proof for this assertion is given in Appendix A which is based on a suggestion by Pourciau (private communication). These considerations are valid provided that the curvature of the orbit is finite, in which case the difference between the chord and the arc length is second order in the chord length.

In Prop. 1, the deflection at each vertex due to the central force impulse is equivalent to the deflection that Newton described in Prop. 6. The main difference is that in Prop. 1 the extension of a chord replaces the tangent line to the curve in Prop. 6. For example, referring to Fig. 1, at vertex B the extension Bc of the chord AB in Prop. 1 corresponds to the tangent line PR in Prop. 6 (see Fig. 4), with P equivalent to B, and the deflection Cc parallel to BS in Prop. 1 corresponding to the deflection RQ parallel to PS in Prop. 6. In the limit that the chord length approaches zero, the difference between the tangent line and the chord becomes vanishingly small, and consequently these two constructions become similar, except that the magnitude of the deflection at a vertex in Prop. 1 is twice as large as that in Prop. 6. Hence, in the continuum limit one obtains a measure for central force in Prop. 1 equivalent to that in Prop. 6, by dividing the deflection at a vertex, which depends quadratically on the adjacent chord length, by the square of the area of the triangles. For example, the measure of the force at B is the continuum limit of $Cc/(\Delta SBC)^2$, where $(1/2)\Delta SBC$ is the area of the triangle SBC, is twice the measure of the force at P given in Prop. 6, which is the limit of $QR/(\Delta SPR)^2$, where $(1/2)\Delta SPR = SP \times QT$ is the area of the triangle SPR. (the deviation QR in Prop. 6 is equal to 1/2 the deviation Cc in Prop. 1).

In Cor. 3 of Prop. 1 Newton indicated that

> ... the forces at B and E are to each other in the ultimately ratio of the diagonals BV and EZ...

where BV is equal to the deflection Cc while EZ is equal to the deflection Ff in Fig. 1. In Cor. 4 of Prop. 1, Newton clarified further the relation between impulses and continuum forces, by announcing that

> " The forces ...are to one another as those saggitas of arcs described in equal times ...when the arc are decreased indefinitely. For these sagittas are halves of the diagonals with which we dealt in corol. 3. , "

Hence, in Prop. 1 Newton evaluated the ratio of forces at two different points on the orbit in the continuum limit, but he did not determined the absolute magnitude of the force. It is interesting that these corollaries did not appear in the first edition of the *Principia*, suggesting that afterwards Newton felt the need to explain how continuous forces are obtained as the limit of discrete impulses.

5. Summary and Conclusions

We have shown that apart from some mathematical details which have been discussed here, Newton's polygonal construction for an orbit in Prop. 1 (see Fig. 1), due to the action of discrete impulses, has a well defined continuum limit which is justified by Lemma 3 in section 1 of the *Principia*. This limit is a continuous planar curve as a function of time describing orbital motion under the action of central forces with origin at S. It satisfies Kepler's area law which states that the time interval between any two points on the orbit is proportional to the area swept out by the radius vector between these points. This property of an orbit for central forces was applied by Newton in Prop. 6 to obtain an expression for the magnitude of the force when the orbital curve is known. The planarity property of the orbit is a straightforward consequence of the requirement that this orbit is the continuum limit of a polygonal trajectory due to force impulses, because the vertices of this polygon all lie in the same plane when the impulses are directed to a common center. The orientation of this plane is determined by initial conditions, i.e.,

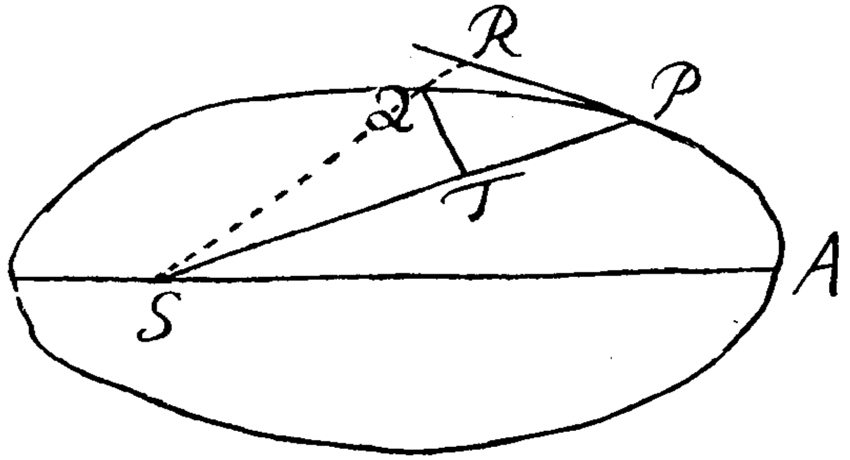

Figure 4: Newton's diagram for Prop.6

the position and velocity vectors of the moving body at some given time. We also have shown that it is straightforward to prove that in the continuum limit these impulses lead to a central force which is proportional to the force measure defined in Prop. 6, but in Prop. 1 Newton established only the ratio of these forces at two different points on the orbit.

Historically, Hooke played an important role in prompting Newton in 1679 to take a new approach to orbital dynamics. This led him to prove the area law for central forces (Nauenberg 1994b) which Kepler had found empirically by fitting the orbit of the planet Mars to the observations of Tycho Brahe. There is evidence that until 1679 Newton had been pursuing a different approach to orbital dynamics based on his development of curvature (Nauenberg 1994a, Brackenridge and Nauenberg 2002). This approach corresponds to a *local* description of central forces in which the area law is not apparent, and evidently Newton was unaware that this law was a consequence of such forces until his correspondence with Hooke. Starting with any such local definition of orbital motion leads also to difficulties with the arguments presented in Prop. 1, particularly with the proof of planarity, as was argued recently by Pourciau (Pourciau, 2002). But in order to interpret correctly the logic of Prop. 1, it is important to understand the historical context in which Newton discovered this crucial proposition. While in Prop. 6 Newton described the action of central forces by a *local* condition which he then applied to Props. 9-17, there is also evidence that he considered orbital motion by a *global* condition, the continuum limit of discrete impulsive forces, which he used to develop a sophisticated three-body perturbation theory. This theory is outlined in Prop. 17, Corollaries 3 and 4, and is described in full detail in the Portsmouth manuscripts (Nauenberg 2000), which remained unpublished until recently (Whiteside 1974). We conclude that apart from some mathematical details left out by Newton, which have been discussed here, Prop. 1 is well grounded, and provides a valid proof that for central forces the orbits are planar and satisfy Kepler's area law.

Appendix A

Following a suggestion by Pourciau (private communication), I give here a rigorous proof in modern notation that for a segment of an orbit with finite curvature all the chord lengths in Newton's polygonal construction in Prop. 1 vanish in a mathematically well defined continuum limit. Moreover, for orbits satisfying certain conditions, this construction covers a finite segment of the orbit. which justifies the application of Lemma 3 cor. 4 which Newton invoked for the existence of this limit in Prop. 1. This proof also demonstrates that the Whiteside-Aiton criticism, that Prop. 1 applies only to an *infinitesimal* arc, is incorrect.

Referring to Fig. 5 for the segment of an orbit, let $s(j)$ be the chord length associated with the jth and (j+1)th vertices of Newton's polygon (see Fig. 1), and $e(j) = s(j+1) - s(j)$ the difference in length between adjacent chords. Then $e(j) \approx d(j)cos[\theta(j)]$ to first order in the ratio $d(j)/s(j)$, where $d(j)$ is the magnitude of the deflection at the j-th vertex which is parallel to the radius vector at this vertex, and $\theta(j)$ is the angle between this deflection and the chord $s(j)$. For an orbit with finite curvature, it follows that

$$d(j) \approx c(j)s(j)^2/sin(\theta(j)), \tag{1}$$

where $c(j)$ is equal to half the curvature of the orbit at the intersection with the j-th vertex. The length of the first chord $s(1)$ determines the length of all the other chords in Newton's polygonal construction, because

$$s(j) = s(1) + e(1) + e(2) + ... + e(j-1) \tag{2}$$

for $j = 2, 3, ...n$. Let $s1 = L/n$ where L is some fixed length. Then

$$e(1) = c'(1) * (L/n)^2, \tag{3}$$

and for $j = 2, 3, ...n$

$$e(j) = c'(j) * (L/n)^2 + o(L/n)^3 \tag{4}$$

where $c'(j) = c(j) * cot(\theta(j))$ and $o(L/n)^3$ refer to all terms proportional to $(L/n)^3$ and higher powers of (L/n). Hence, the sum

$$e(1) + e(2) + ... + e(j) = (L/n)^2 * (c'(1) + c'(2) + ... + c'(j)) + o(L/n)^3 \tag{5}$$

Now suppose that c is the maximum curvature of the given segment of the orbit. Then

$$|e(1) + e(2) + ... + e(j)| < (L/n)^2 * (j-1) * |c'| + o(L/n)^3, \tag{6}$$

and

$$|e(1) + e(2) + ... + e(n)| < (L/n)^2 * (n-1) * |c'| + o(L/n)^2 \tag{7}$$

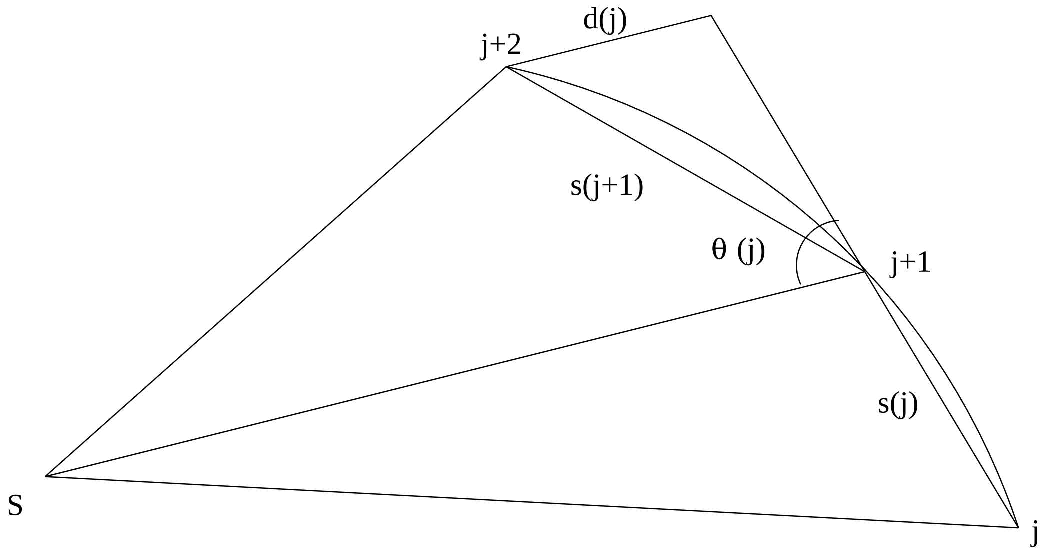

Figure 5:

where $|c'| = cmax(|cot[\theta(j)]|$. The existence of an upper bound for the magnitude of $cot[\theta(j)]$ follows from the constraint discussed previously that an orbit cannot become tangent to the radius vector, which implies that $\theta(j)$ cannot be either 0^0 or 180^0. Hence, as n goes to infinity both of these two sums go to zero, and therefore, according to Eq. 2 all the chords $s(j)$ also vanish in this limit. Therefore in this limit Newton's polygonal construction in Prop. 1 covers a segment of the orbit with a length given by the limit of the sum $s(1) + s(2) + s(3) + ... + s(n)$ as n approaches infinity.

The length of this segment is a function of the parameter L. For the special case that the curvature is a constant, i.e. for a circular orbit with the center of force at the center of the circle, the parameter L is the length of the segment. In this case the angle $\theta(j)$ is equal to 90^o, and the quantities $e(j)$, which are of order $(L/n)^3$, do not contribute to the length of the curve. If the curvature decreases along a segment of the orbit and the center of force is located at the initial center of curvature, the angle $\theta(j)$ becomes smaller than 90^o and the corresponding length of the segment is larger than L. This is the case, for example, for a segment of an elliptic orbit if the initial position is at an apsis of the ellipse where the curvature is a maximum, and the center of force is at one of the two foci. Another example is a spiral curve along the direction where the curvature is decreasing, with the center of force at the center of the spiral. Hence, we have shown that segments of the fundamental orbits which are discussed in sections 2 and 3 of the *Principia* can be regarded as the continuum limit of Newton's polygonal construction.

Appendix B

To help clarify the main points emphasized in this paper, I describe in this appendix the content of Prop. 1 in modern vector notation. I start with a formulation of the main assumptions, which were not spelled out by Newton, and the theorem associated with this proposition.

> a) For central forces, orbital motion is the continuum limit of motion along the sides of a polygon due to the action of impulses directed towards a common center.
>
> b) The continuum limit is constructed (see Cor.2 Lemma 3,) by attaching the vertices of the polygon to a given curve which remains fixed as the number of these vertices increases indefinitely.
>
> Theorem:
>
> For a discrete sequence of central force impulses at *equal time* intervals, the orbital motion is along the sides of a planar polygon. The triangles defined by these sides and the center of force have *equal areas.*
>
> For a central force which is the continuum limit of force impulses, the orbital motion is along a planar curve with radius vector, having origin at the center of force, sweeping equal areas in equal times.

Assuming that there are n impulses and $n+1$ vertices in the polygon shown in the figure associated with Prop. 1, (see Figs. 1 and 5), let $\vec{r}_j$ be the position vector and $\vec{v}_j$ the velocity vector at the j-th vertex where $j = 1, 2, ...n$. Then Newton's construction takes the form

$$\vec{r}_{j+1} = \vec{r}_j + \vec{v}_j \Delta t, \tag{8}$$

and

$$\vec{v}_{j+1} = \vec{v}_j + \vec{\Delta} v_{j+1}, \tag{9}$$

where Δt is the *equal* time interval between impulses, and $\Delta\vec{v}_j$ is the instantaneous velocity change $\Delta\vec{v}_j = \vec{d}_j/\Delta t$ where $\vec{d}_j$ is the deflection at the j-th vertex. The crossed product of Eqs. 8 and 9

$$\vec{r}_{j+1} \times \vec{v}_{j+1} = \vec{r}_j \times \vec{v}_j + \vec{r}_{j+1} \times \Delta\vec{v}_{j+1} \tag{10}$$

where $(1/2)|\vec{r}_j \times \vec{v}_j|$ is the area of the triangle associated with the j-the vertex. Since the deflection $\vec{d}_j$ and corresponding velocity change $\Delta\vec{v}_j$ are parallel to $\vec{r}_j$, which is Newton's *definition* of central force impulses in Prop. 1, the last term in Eq. 10 vanishes, and consequently a) the area of the triangles are equal, and b) the vertices of the polygon lie on a plane. This constitutes Newton's proof of Prop. 1 in the language of vector calculus for discrete impulses.

The problem of the continuum limit is to describe how the time interval Δt and the deflections $\vec{d}_j$ should vary as n approaches infinity. Newton's statement

> "Let the time be divided in equal times...",

corresponds to setting $\Delta t = T/n$, where T is some finite time interval, and his reference to Cor. 4 Lemma 3 implies that the deflections $\vec{d}_j$ are determined by the condition that the vertices of the polygon are located on a *given* curve. Since the polygon is on a plane with orientation determined by initial conditions which are independent of n, this curve must also be on the same plane. It can be described by a vector $\vec{R}(u)$ where u is a scalar parameter which can be chosen arbitrarily. For example, u can be the arc length, or the angular variable in polar coordinates. Then the condition that the j-th vertex is located on this curve is given by

$$\vec{r}_j = \vec{R}(u_j), \tag{11}$$

where u_j is the value of u at the position of this vertex. Associated with the j-th vertex is the time $t_j = jT/n$, and therefore in the continuum limit u becomes a function of t, and $\vec{R}(u)$ describes orbital motion. Since t_j is proportional to the sum of the (equal) areas of the first j triangles of the polygon, in the continuum limit t is proportional to the area swept by the radius vector $\vec{R}(u)$. The deflection due to the force impulses is given by

$$\vec{d}_j = \vec{R}(u_{j+1}) + \vec{R}(u_{j-1}) - 2\vec{R}_j, \tag{12}$$

and as Δt vanishes the ratio

$$\frac{d_j}{(v_j \Delta t)^2} \tag{13}$$

is proportional to the curvature of the orbit at u, where $d_j = |\vec{d}_j|$ and $v_j = |\vec{v}_j|$. For finite curvature, this ratio has a well defined limit as was demonstrated by Newton in Lemma 11, and discussed further in the subsequent Scholium of the *Principia*. Consequently,

$$\frac{\vec{\Delta} v_j}{\Delta t} = \frac{\vec{d}_j}{(\Delta t)^2}. \tag{14}$$

has a continuum limit corresponding to the standard definition in calculus of the acceleration or force/unit mass,

$$\vec{a} = \lim_{\Delta t \to 0} \frac{\vec{\Delta} v_j}{\Delta t} = \frac{d^2 \vec{R}}{dt^2} \tag{15}$$

The existence of this limit follows also from Lemma 10 which implies that when Δt becomes vanishingly small the magnitude $|d_j|$ of the deflections are proportional to $(\Delta t)^2$. But it is evident that in Prop. 1 Newton did not have this condition in mind, because in the proof of this proposition he cited Lemma 3 instead of Lemma 10.

Acknowledgements

I would like to thank Niccolo Guicciardini and Bruce Pourciau for many stimulating exchanges, suggestions and critical comments on the subject of this paper.

References

[1] Aiton, E. 1989. Polygons and Parabolas: Some Problems Concerning the Dynamics of Planetary Orbits, *Centaurus* 31, 207-221

[2] Brackenridge, J.B. 1995. *The Key to Newton's Dynamics: The Kepler Problem and the Principia* University of California Press, pp. 79-85

[3] Brackenridge, J.B and Nauenberg, M. 2002. Curvature in Newton's dynamics. *The Cambridge Companion to Newton* edited by I.B. Cohen and G. Smith Cambridge University Press, pp. 85-137

[4] Chandrasekhar, S. 1995. *Newton's Principia for the Common Reader* Clarendon Press, Oxford, pp. 67-69

[5] Cohen, I.B. and Whitman, A. 1999. *Isaac Newton, The Principia* A new Translation preceded by A Guide to Newton's Principia by I.B. Cohen. University of California Press, p. 115

[6] Densmore, D. 1995. *Newton's Principia: The Central Argument* Green Lion Press, Santa Fe, pp. 98-110

On p. 99 Densmore claims that " the bases of triangles in the proposition (Prop. 1) ... do not circumscribe and are not inscribed in , this ultimate curve, nor do they connect to it or follow it in any other way as a kind of 'ghost curve' ", and she claims that " Newton has not offered an argument that the limiting procedure in the proposition is a unique curve...". But Densmore ignores the fact that Newton invoked cor. 4, Lemma 3 to justify his limiting procedure.

[7] Erlichson, H. 1992. Newton's Polygon Model and the Second Order Fallacy, *Centaurus* 23, 243-258.

Some of the criticism in this paper of Whiteside's analysis of Prop. 1 is not valid, because Whiteside did identify correctly the deflections due to force impulses as " second order infinitesimal magnitudes... ".

[8] Guicciardini, N. 1999. *Reading the Principia*, Cambridge University Press, pp. 211-216

[9] Herivel, J. 1965. *The Background to Newton's Principia, A Study of Newton's Dynamical Researches in the Years 1664-84* Oxford Press

[10] Nauenberg, M. 1998a. The Mathematical Principles Underlying the Principia revisited. *Journal for the History of Astronomy* 29, 286-300, note 4

[11] Nauenberg, M. 1994a. Newton's Early Computational Method for Dynamics, *Archive for History of Exact Sciences* 46, 221-252

[12] Nauenberg, M. 1994b. Hooke, orbital motion, and Newton's Principia *American Journal of Physics* 62, 331-350

[13] Nauenberg, M. 1998b On Hooke's 1685 Manuscript on Orbital Mechanics *Historia Mathematica* 25, 89-93

[14] Nauenberg, M. 2000. Newton's Portsmouth Perturbation Method and its application to Lunar Motion, in *The Foundations of Newtonian Scholarship*, edited by Richard H. Dalitz and Michael Nauenberg, (World Scientific) pp. 167-194

[15] *The Posthumous Works of Robert Hooke, containing his Cutlerian Lectures and other Discourses, Read at the Meeting of the Illustrious Royal Society* (Cass, London, 1971) p. 149

[16] Pourciau, B. *Newton's Argument for the First Proposition of the Principia* To be published in Archives for Exact History of Sciences

[17] Whiteside, D.T. 1974. *The Mathematical Papers of Isaac Newton* v.6 edited by D.T. Whiteside, Cambridge University Press pp. 35-36

[18] Whiteside, D.T. 1991 The Prehistory of the Principia from 1664 to 1686, *Notes and Records of the Royal Society of London* 45, 30