ON THE GEOMETRY OF THE KEPLER PROBLEM

JOHN MILNOR
Institute for Advanced Study, Princeton NJ 08540

Abstract. It will be convenient to use the term *Kepler orbit* for any curve $\mathbf{x} = \mathbf{x}(t)$ in 3-space which arises as a solution to the Newtonian two body problem. Hamilton showed that the velocity vector $\mathbf{v} = d\mathbf{x}/dt$, associated with any nondegenerate Kepler orbit, moves along a circle. Following Györgyi, Moser, Osipov and Belbruno, this *velocity circle* can be interpreted as follows. If we fix the total energy E, then the manifold M_E consisting of all vectors $\mathbf{v}$ with $\mathbf{v} \cdot \mathbf{v} > 2E$ possesses a natural Riemannian metric of constant curvature $-2E$, whose geodesics are precisely the circles associated in this way with Kepler orbits. In other words, M_E can be identified with one of the three classical geometries, that is with spherical, Euclidean or Lobachevsky space, so that each "straight line" in this geometry corresponds to a unique Kepler orbit.

1. The Velocity Circle. In Kepler's first attempts to understand the orbits of the planets, he tried the hypothesis that each orbit is a circle which lies in some plane containing the sun, but is not centered at the sun. This is of course wrong. Yet is does describe the correct answer to a slightly transformed problem.

Consider a particle which is attracted to the origin by a force inversely proportional to the square of the distance, so that its position vector $\mathbf{x} = \mathbf{x}(t)$ satisfies the *Newton differential equation*

$$d^2\mathbf{x}/dt^2 = -k\mathbf{x}/|\mathbf{x}|^3. \tag{1}$$

Here k is some fixed positive constant, and $|\mathbf{x}|$ denotes the Euclidean length $\sqrt{\mathbf{x} \cdot \mathbf{x}}$ of the vector $\mathbf{x}$. (This equation is related to the Newtonian two body problem as follows. If two spherical objects with position coordinates $\mathbf{x}_1$ and $\mathbf{x}_2$ move about their common center of gravity in accordance with Newton's laws, then the difference vector $\mathbf{x} = \mathbf{x}_2 - \mathbf{x}_1$ will satisfy (1).) Solutions of this equation, whether elliptical or otherwise, will be called *Kepler orbits.*

Throughout §1, *we will assume that the position* $\mathbf{x}$ *and velocity* $\mathbf{v} = d\mathbf{x}/dt$ *are linearly independent vectors at time* $t = t_0$.

THEOREM 1 (Hamilton, 1846). *As t varies, the velocity vector $\mathbf{v} = d\mathbf{x}/dt$ moves along a circle C, which lies in some plane P containing the origin, but is not in general centered at the origin. Any such circle can occur, and this "velocity circle" C, together with its orientation, determines the orbit $\mathbf{x} = \mathbf{x}(t)$ uniquely.*

The proof will show that the corresponding orbit is either an ellipse, hyperbola, or parabola according as the origin lies inside, outside, or exactly on the velocity circle C. In the elliptic case, the velocity vector moves around the entire circle, but in the other two cases only that portion of the circle which is convex towards the origin is actually traversed. (Compare Fig. 1.)

To begin the proof, recall that the cross product vector $\mathbf{h} = \mathbf{x} \times \mathbf{v}$ is an invariant of the motion. That is the derivative $d\mathbf{h}/dt$ is identically zero; as one verifies by an easy calculation. Thinking of $\mathbf{x}(t)$ as the orbit of a particle of unit mass, we will call the length $h = |\mathbf{h}|$ the *angular momentum* of this orbit. Note that $h > 0$ by our linear independence hypothesis.

It will be convenient to introduce a new system of cartesian coordinates $\mathbf{x} = (x, y, z)$ so that the vector $\mathbf{h}$ *points along the positive z-axis.* The equation $\mathbf{x}(t) \times \mathbf{v}(t) = \mathbf{h} = (0, 0, h)$ then implies that the vector $\mathbf{x}(t)$ always lies in the (x, y)-plane. Setting $\mathbf{x}(t) = (r\cos\theta, r\sin\theta, 0)$, a straightforward

John Milnor received his Ph.D. at Princeton University in 1954, under the direction of Ralph Fox. As a student, he was also strongly influenced by Norman Steenrod, and by a year in Zürich under Heinz Hopf. After teaching for some years at Princeton University, interrupted by stays in Oxford, Berkeley, U.C.L.A.,and M.I.T., he moved across town to the Institute for Advanced Study. His mathematical work was centered around the topology of manifolds and related areas of algebra.

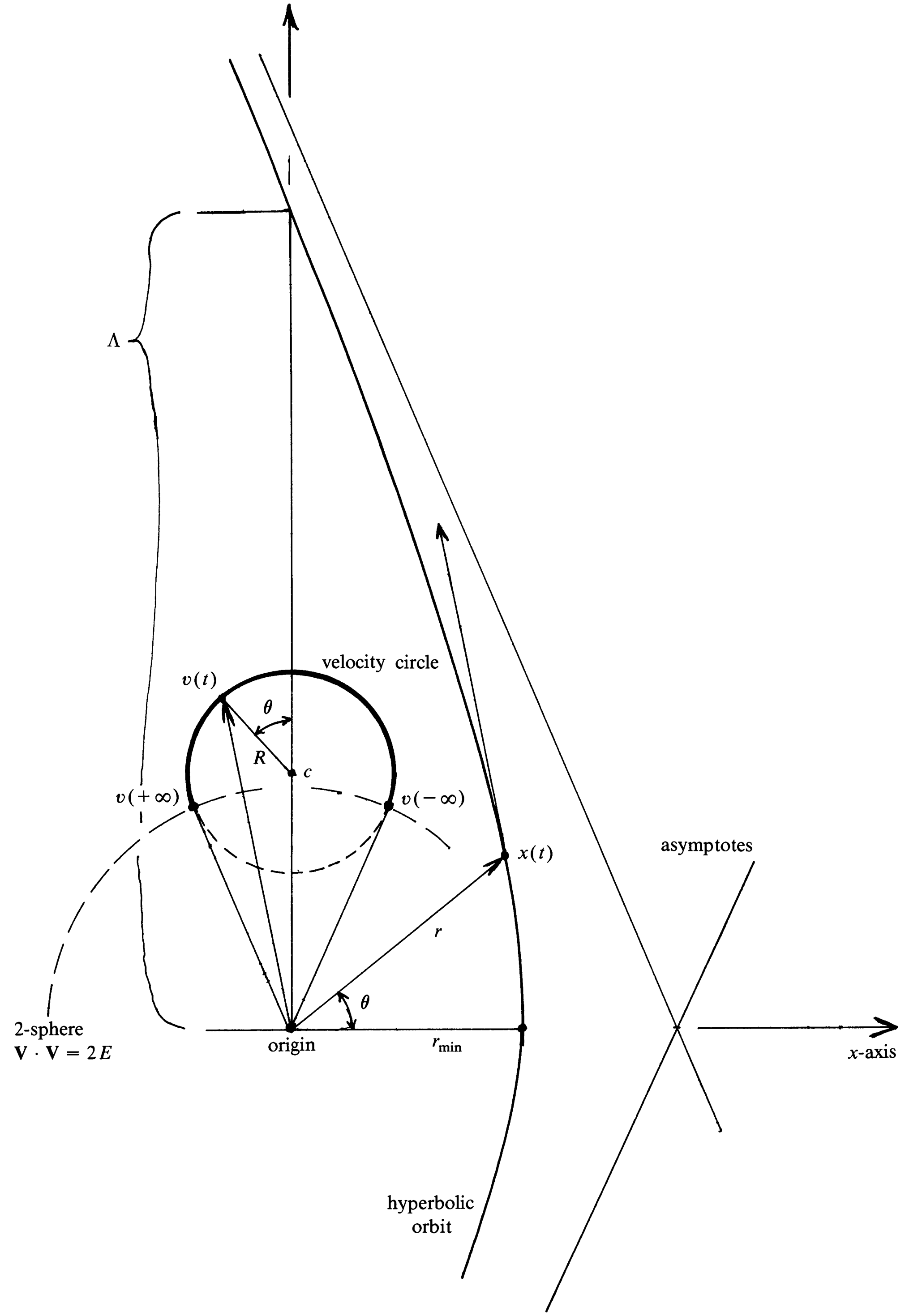
Λ
velocity circle
$v(t)$
θ
R
c
$v(+\infty)$
$v(-\infty)$
$x(t)$
asymptotes
r
θ
2-sphere
$\mathbf{V} \cdot \mathbf{V} = 2E$
origin
$r_{\min}$
x-axis
hyperbolic
orbit

FIG. 1

computation shows that

$$h = r^2\, d\theta/dt, \tag{2}$$

The constancy of this expression $r^2\, d\theta/dt$ is just *Kepler's Second Law*, which asserts that a line segment from the origin to $\mathbf{x}(t)$ traverses equal areas in equal times.

[It is interesting to note that Kepler's interpretation of this law was very different from our modern interpretation, since he had no concept of inertia. He believed that the sun exerts a sideways force, inversely proportional to distance, which pushes the planets around their orbits.]

Let us write the Newton equation as $d\mathbf{v}/dt = -k(\cos\theta, \sin\theta, 0)/r^2$. Dividing by $d\theta/dt = h/r^2$, and setting $R = k/h$, we obtain

$$d\mathbf{v}/d\theta = -R(\cos\theta, \sin\theta, 0),$$

Integrating, we obtain

$$\mathbf{v} = -R(\cos\theta, \sin\theta, 0)\, d\theta = R(-\sin\theta, \cos\theta, 0) + \mathbf{c},$$

where $\mathbf{c} = (c_1, c_2, 0)$ is a constant of integration. *This proves that the velocity vector* $\mathbf{v}$ *moves along a circle centered at* $\mathbf{c}$, *with radius* $R = k/h$ *inversely proportional to angular momentum, and lying in the plane through the origin which is orthogonal to the vector* $\mathbf{h}$.

It is interesting to note that the difference vector $\mathbf{v} - \mathbf{c} = R(-\sin\theta, \cos\theta, 0)$ is always orthogonal to the position vector $\mathbf{x} = r(\cos\theta, \sin\theta, 0)$. *Hence the angle between two velocity vectors, as measured around this circle C, is equal to the angle between corresponding position vectors as seen from the origin.*

The ratio $\varepsilon = |\mathbf{c}|/R$, that is the distance of the center from the origin divided by the radius, is called the *eccentricity* of the circle C with respect to the origin. *It will be convenient to choose our coordinates x, y for the plane so that the center* $\mathbf{c}$ *of the circle C lies on the positive y-axis.* With this convention, we can write

$$\mathbf{v} = R(-\sin\theta, \quad \varepsilon + \cos\theta, 0).$$

Substituting this expression in the equation $\mathbf{h} = \mathbf{x} \times \mathbf{v}$ we obtain the formula $h = rR(1 + \varepsilon\cos\theta)$ for angular momentum; and solving for r we obtain

$$r = \Lambda/(1 + \varepsilon\cos\theta), \tag{3}$$

where $\Lambda = h^2/k$. *This is precisely the equation of a conic section of eccentricity ε, with focus at the origin, in polar coordinates.* (See Appendix 1.) This conic section is either a circle, ellipse, parabola, or hyperbola according as $\varepsilon = 0$, $0 < \varepsilon < 1$, $\varepsilon = 1$, or $\varepsilon > 1$. In the latter two cases, note that the possible values of the angular coordinate θ are constrained by the inequality $1 + \varepsilon\cos\theta > 0$.

Our coordinate system has been chosen so that the point of closest approach, where $r = r_{\min}$, lies on the positive x-axis, with $\theta = 0$. This corresponds to the classical convention that the angular coordinate θ, known as the *anomaly*, should be measured from this point of closest approach. The geometrical constant Λ, proportional to the square of angular momentum, is known classically as the *semi-latus-rectum*. It can be described as the distance from origin to orbit in a direction at right angles to the direction of closest approach.

To complete the proof of Theorem 1, we must show that every such conic section (3) really yields a solution to the Newton equation. But Equation (3) shows that r can be expressed as a function of θ, and Equation (2) implies that $t = \int r(\theta)^2\, d\theta/h$ can also be expressed as a function of θ. It follows from the inverse function theorem that θ and hence $\mathbf{x}$ can be expressed as functions of t; and it is not difficult to check that the functions constructed in this way do indeed satisfy Equation (1). Details will be left to the reader. ∎

Another classical invariant associated with a solution $\mathbf{x} = \mathbf{x}(t)$ of Equation (1) is the *energy* $E = \mathbf{v} \cdot \mathbf{v}/2 - k/r$. A straightforward calculation shows that the derivative dE/dt is zero, so that E is indeed constant along any orbit. Note that the inequality $\mathbf{v} \cdot \mathbf{v} > 2E$ must always be satisfied.

In terms of the velocity circle, with radius R and center $\mathbf{c}$, energy is given by

$$2E = \mathbf{c}\cdot\mathbf{c} - R^2 = -(1-\varepsilon^2)k^2/h^2, \tag{4}$$

as one can check by evaluating E at any point of the orbit. This computation can be expressed more geometrically by the following statement, which is essentially due to Euclid. (Compare Coxeter [4] pp. 8, 81.)

LEMMA 1. *Let* $\mathbf{v}_1$ *and* $\mathbf{v}_2$ *be two points of the circle* C *which lie on a common line through the origin. Then* $\mathbf{v}_1\cdot\mathbf{v}_2 = \mathbf{c}\cdot\mathbf{c} - R^2$; *hence* $\mathbf{v}_1\cdot\mathbf{v}_2$ *is equal to* $2E$.

Proof. Let $\mathbf{w}$ be the point of the circle which is diametrically opposite to $\mathbf{v}_1$. Setting $\mathbf{v}_1 = \mathbf{c}+\mathbf{e}$ and $\mathbf{w} = \mathbf{c}-\mathbf{e}$, we see that

$$\mathbf{v}_1\cdot\mathbf{w} = \mathbf{c}\cdot\mathbf{c} - \mathbf{e}\cdot\mathbf{e} = \mathbf{c}\cdot\mathbf{c} - R^2 = 2E.$$

Since $\mathbf{v}_1, \mathbf{v}_2, \mathbf{w}$ is a right triangle, it follows that $\mathbf{v}_1$ is orthogonal to the difference vector $\mathbf{v}_2 - \mathbf{w}$. Hence $\mathbf{v}_1\cdot\mathbf{v}_2$ is equal to $\mathbf{v}_1\cdot\mathbf{w} = 2E$, as required. ■

In the case of an elliptical orbit, the precise shape and period of the orbit are related to the energy E as follows. First note that the *major axis* $2a$ of the ellipse is given by the computation

$$2a = r_{\min} + r_{\max} = \Lambda/(1+\varepsilon) + \Lambda/(1-\varepsilon) = 2\Lambda/(1-\varepsilon^2).$$

Since $\Lambda = h^2/k$, it follows from (4) that

$$2E = -k/a. \tag{5}$$

Thus the energy E *of an elliptic orbit is negative, and* $|E|$ *is inversely proportional to the major axis of the ellipse.*

The *minor axis* $2b$ of the ellipse, computed by setting $dy/dt = R(\varepsilon + \cos\theta) = 0$, is given by

$$b = y_{\max} = \Lambda/\sqrt{1-\varepsilon^2}.$$

The *area* of the ellipse is then given by

$$A = \pi ab = \pi\Lambda^2/(1-\varepsilon^2)^{3/2}.$$

Applying Kepler's Second Law in the form $dA/dt = h/2$, we see that the *period* T of such an elliptical orbit satisfies $A/T = h/2$. Together with (4), (5), and the definition of Λ, this implies that

$$T = 2\pi k/(-2E)^{3/2} = 2\pi(a^3/k)^{1/2}, \tag{6}$$

which is a modern statement of *Kepler's Third Law. Thus the period is proportional to* $a^{3/2}$ *and is inversely proportional to* $|E|^{3/2}$. For further details of these calculations see, for example, Arnold [1] or Synge and Griffith [21].

Remark. A familiar but incorrect version of this Third Law asserts that the period is proportional to the $3/2$ power of the "mean distance" of $\mathbf{x}(t)$ from the origin. In fact the mean distance, that is the time average $\oint r\,dt/\oint dt$, is not equal to the major semi-axis a but is rather equal to $(1+\varepsilon^2/2)a$. However it is interesting to note that the time average of r^{-1} is precisely equal to a^{-1}. Proofs based on Formula (9) of Appendix 1 are easily supplied.

It is noteworthy that *the period of such a periodic orbit depends only on the energy*. This is true for period solutions of Lagrangian or Hamiltonian differential equations under quite general conditions. Compare Herglotz [9], Wintner [22, §100], and Gordon [6].

2. The Levi-Civita Metric. Moser [16] has given a very pretty picture of the space of all elliptic orbits of fixed energy, or fixed period, in terms of *stereographic projection*. (See also Györgyi [7].) To carry out Moser's construction, we must introduce a 3-dimensional sphere, consisting of all unit vectors $\mathbf{u} = (u_1, u_2, u_3, u_4)$ in Euclidean 4-space, and project it stereographically from its "north pole" $(0,0,0,1)$ onto the 3-dimensional Euclidean space consisting of vectors of the form $\mathbf{v} = (v_1, v_2, v_3, 0)$. Thus each unit vector $\mathbf{u}$ maps to the unique vector $\mathbf{v} = (u_1, u_2, u_3, 0)/(1 - u_4)$ which has last coordinate zero, and lies on the same straight line through the north pole. *Then every circle on the 3-sphere projects onto a circle or line in 3-space.* (Compare Appendix 2.)

A brief computation shows that the *antipodal map* $\mathbf{u} \mapsto -\mathbf{u}$ from the 3-sphere to itself corresponds to the *negative inversion map* $\mathbf{v} \mapsto -\mathbf{v}/|\mathbf{v}|^2$ from Euclidean 3-space to itself. In other words, diametrically opposite points of the 3-sphere correspond to points $\mathbf{v}_1$ and $\mathbf{v}_2$ in Euclidean 3-space which lie on a common line through the origin, and satisfy $\mathbf{v}_1 \cdot \mathbf{v}_2 = -1$. Evidently the images of *great circles* on the 3-sphere are just those circles, or lines through the origin, which map into themselves under this negative inversion operation. Using Lemma 1, we see that a circle in 3-space has this invariance property if and only if it is one of our "velocity circles" with energy $E = (\mathbf{c} \cdot \mathbf{c} - R^2)/2$ equal to $-1/2$, or equivalently with period T equal to $2\pi k$. *Thus stereographic projection carries great circles on the 3-sphere precisely onto the velocity circles associated with elliptic orbits of energy* $-1/2$.

There is one apparent defect in this picture. Namely those great circles which pass through the poles correspond to straight lines through the origin in velocity space. These do not seem to fit into our picture. However this apparent defect is actually a virtue in disguise. For it enables us to give a precise description of the limiting behavior of elliptic orbits of fixed period as the angular momentum h tends to zero. If $h \to 0$, keeping E fixed and negative, it is easy to see that the minor axis of the orbit ellipse tends to zero, so that the ellipse flattens out and tends towards a straight line segment of length $-k/E$ with one end at the origin. In order for the orbit to vary continuously as $h \to 0$, we must adopt the following.

REFLECTION CONVENTION. *If an orbit has angular momentum $h = 0$, and if $dr/dt < 0$ for some values of t so that $\mathbf{x}(t)$ falls freely along a straight line towards the origin, then $\mathbf{x}(t)$ is reflected back along this same line when it hits the origin.*

(Compare Devaney [5], McGehee [14].) In fact it is natural to adopt this same Reflection Convention whether the energy is positive, negative or zero.

By rotating our coordinates, we can put such a singular orbit in the form $t \mapsto (r(t), 0, 0)$, where $r(t) \geq 0$ satisfies the differential equation $(dr/dt)^2/2 = E + k/r$. As an example, if $E = 0$, it follows that r is proportional to $(t - t_0)^{2/3}$. If $E < 0$, then the solution curve $r = r(t)$ is a cycloid, periodic in t, and smooth except for a cusp wherever $r = 0$. (See Appendix 1.) Note in particular that the derivative dr/dt tends to minus infinity as r decreases to zero, and then jumps to plus infinity. *If $\mathbf{x}(t)$ bounces off the origin, then the velocity vector $\mathbf{v}(t)$, moving along a straight line, sweeps through the "point at infinity" in 3-space.* This velocity line corresponds, under stereographic projection, to a great circle which sweeps through the north pole of the unit 3-sphere.

Thus, if we adopt this reflection convention, then there is a precise one-to-one correspondence between orbits of energy $-1/2$ and great circles on the unit 3-sphere. The situation for other negative values of E is the same, except for a scale change in 3-space.

According to Osipov [17], [18], and Belbruno [2], [3], there is an analogous description for parabolic or hyperbolic orbits. First consider a parabolic orbit, with energy $E = 0$. Recall that the velocity vector $\mathbf{v}(t)$ associated with such an orbit moves along a circle C which passes through the origin. *Let us transform this problem by applying the inversion operation* $\mathbf{v} \mapsto \mathbf{w} = \mathbf{v}/|\mathbf{v}|^2$. Then a circle C through the origin corresponds to a straight line in the space of vectors $\mathbf{w}$. (See Appendix 2.) *In this way we obtain a precise one-to-one correspondence between orbits of energy zero and straight lines in the Euclidean space consisting of all inverted velocity vectors* $\mathbf{w}$. In this picture, the orbits which bounce off the origin correspond to straight lines through the origin in inverted velocity space.

Finally let us look at the positive energy case, say $E = +1/2$. Since $\mathbf{v} \cdot \mathbf{v} > 2E = 1$, it follows that the inverted velocity vector $\mathbf{w} = \mathbf{v}/|\mathbf{v}|^2$ varies over the unit ball $|\mathbf{w}| < 1$. Using Lemma 1, we see that the corresponding velocity circles C are invariant under inversion, and hence intersect the boundary of the unit ball orthogonally. *Thus each orbit of energy* $+1/2$ *corresponds to a circle arc* $t \mapsto \mathbf{w}(t)$ *which spans the unit ball* $|\mathbf{w}| < 1$, *and intersects its boundary 2-sphere orthogonally.* Orbits which bounce off the origin correspond to diameters spanning this unit 3-ball.

It is natural to compare this picture with the *conformal unit ball model* for the "hyperbolic" non-Euclidean geometry of Lobachevsky. (See Appendix 2.) In this model, discovered by Beltrami and later by Poincaré, points of hyperbolic space correspond to points in the Euclidean unit ball, and hyperbolic straight lines correspond to circle arcs or diameters which span the unit ball, meeting its boundary orthogonally. *Thus the orbits of energy* $+1/2$ *correspond precisely to "straight lines" in this model for hyperbolic 3-space.* There is an analogous picture for any positive value of E.

In terms of Riemannian geometry, we can put these three different constructions into one common framework as follows. Levi-Civita pointed out that it is possible to simplify solutions to Equation (1) by introducing a *fictitious time parameter* $s = \int dt/r$ along any orbit, where $r = |\mathbf{x}|$. (Compare Appendix 1.) Using this parameter, we will prove the following.

THEOREM 2 (Osipov and Belbruno). *Fixing some constant energy* E, *consider the space* M_E *consisting of all velocity vectors* $\mathbf{v}$ *for which* $\mathbf{v} \cdot \mathbf{v} > 2E$, *together with a single improper point* $\mathbf{v} = \infty$. *This space possesses one and only one Riemannian metric* ds^2 *so that the arc-length parameter* $\int ds$ *along any velocity circle* $t \mapsto \mathbf{v}(t)$ *is precisely equal to the Levi-Civita parameter* $\int dt/|x(t)|$. *This metric is smooth and complete, with constant curvature* $-2E$, *and its geodesics are precisely the circles or lines* $t \mapsto \mathbf{v}(t)$ *associated with Kepler orbits.*

In other words, there is a unique way of defining the "length" of a curve Γ in this space M_E of all compatible velocity vectors so as to coincide with Levi-Civita's integral $\int_\Gamma dt/r$. If $K = -2E$ is positive, then $M_{E'}$ with this definition of length, is isometric to a 3-sphere of radius $1/\sqrt{K}$ in Euclidean 4-space; and the great circles on this sphere correspond to velocity circles. If $K = 0$, then M_E is isometric to Euclidean space; and if $K < 0$, it is isometric to hyperbolic space, with the unit of distance chosen appropriately. In particular, in all three cases, M_E can be made into a smooth manifold even in a neighborhood of the special point $\mathbf{v} = \infty$. Geodesics which pass through this special point correspond to Kepler orbits which bounce off the origin.

Proof. The Newton equation $d\mathbf{v}/dt = -k\mathbf{x}/r^3$ implies that $|d\mathbf{v}/dt| = k/r^2$. Dividing this equation by the definition $ds = dt/r$, and recalling the definition of E, we obtain $|d\mathbf{v}/ds| = k/r = \mathbf{v} \cdot \mathbf{v}/2 - E$. Therefore $|ds| = 2|d\mathbf{v}|/(\mathbf{v} \cdot \mathbf{v} - 2E)$, or in other words

$$ds^2 = 4\, d\mathbf{v} \cdot d\mathbf{v}/(\mathbf{v} \cdot \mathbf{v} - 2E)^2, \tag{7}$$

Thus there is one and only one Riemannian metric on M_E *which satisfies our condition, and it is given by this formula* (7). To describe what happens in a neighborhood of infinity, we work with the inverted velocity coordinate $\mathbf{w} = \mathbf{v}/|\mathbf{v}|^2$. Note that $2E\mathbf{w} \cdot \mathbf{w} < 1$. Computation shows that $d\mathbf{w} \cdot d\mathbf{w} = d\mathbf{v} \cdot d\mathbf{v}/(\mathbf{v} \cdot \mathbf{v})^2$. (See Formula (10) of Appendix 2.) Hence $ds^2 = 4\, d\mathbf{w} \cdot d\mathbf{w}/(1 - 2E\mathbf{w} \cdot \mathbf{w})^2$, where the denominator is always strictly positive. If we set $K = -2E$, this becomes

$$ds^2 = 4\, d\mathbf{w} \cdot d\mathbf{w}/(1 + K\mathbf{w} \cdot w)^2, \tag{7'}$$

Except for a scale change, this is just the form given by Riemann, in his inaugural dissertation, for a metric of constant curvature K. In the case $K = 0$, this metric is obviously flat, with straight lines as geodesics. In the positive curvature case, we can compare it with the standard metric on a 3-sphere of radius $1/\sqrt{K}$. It is not difficult to check that these two metrics correspond isometrically, via stereographic projection from the 3-sphere to a 3-plane passing through its center; and that great circles correspond precisely to our (inverted) velocity circles. Further details of the proof, particularly in the case $K < 0$, may be found in Appendix 2. ■

COROLLARY. *The functions* $\mathbf{x}(t)$ *and* $\mathbf{v}(t)$ *depend smoothly on the time t and the initial conditions* $\mathbf{x}(0)$ *and* $\mathbf{v}(0)$, *even in the neighborhood of an orbit which bounces off the origin, so long as the vectors* $\mathbf{x}(0)$ *and* $\mathbf{x}(t)$ *themselves are not zero.*

Here is an explicitly worked out example, to illustrate the behavior of these functions. Consider the family of parallel straight lines $s \mapsto \mathbf{w} = (-s, \alpha, \beta)/2$ in inverted velocity space, depending on two real parameters α and β. These correspond to the family of parabolic orbits

$$\mathbf{x} = k(\alpha^2 + \beta^2 - s^2, 2\alpha s, 2\beta s)/2.$$

Not only this orbital position $\mathbf{x}$, but also the distance

$$r = |\mathbf{x}| = k(\alpha^2 + \beta^2 + s^2)/2$$

and the time

$$t = \int_0 r\,ds = ks(\alpha^2 + \beta^2 + s^2/3)/2$$

can be expressed as smooth functions of the three parameters α, β, s. Using the implicit function theorem, we can solve for s, and hence $\mathbf{x}$, as functions of α, β, t. These functions are smooth except precisely at those points where $\partial t/\partial s = |\mathbf{x}|$ is zero.

Similarly, we can look at the six parameter family of orbits depending on the initial position vector $\mathbf{x}(0) \neq \mathbf{0}$ and the initial velocity vector $\mathbf{v}(0)$. Using the identity

$$r = 2k/(\mathbf{v} \cdot \mathbf{v} - 2E) = 2k\mathbf{w} \cdot \mathbf{w}/(1 - 2E\mathbf{w} \cdot \mathbf{w}),$$

where $E = \mathbf{v}(0) \cdot \mathbf{v}(0) - k/|\mathbf{x}(0)|$, we see that r and $t = \int_0 r\,ds$ can be expressed as smooth functions of the parameters $\mathbf{x}(0) \neq \mathbf{0}$, $\mathbf{v}(0)$ and s. In fact it is not difficult to verify the identity

$$\mathbf{x} = 4k(2(\mathbf{w} \cdot \mathbf{w}')\mathbf{w} - (\mathbf{w} \cdot \mathbf{w})\mathbf{w}')/(1 - 2E\mathbf{w} \cdot \mathbf{w}),$$

where $\mathbf{w}' = d\mathbf{w}/ds$, which shows that $\mathbf{x}$ can also be expressed as a smooth function of these parameters. Again, if we use t in place of s as parameter, then we lose the smoothness of the resulting functions only at those points where $\mathbf{x}$ is precisely equal to the zero vector. ■

Appendix 1. Conic sections.

This will be a brief review of some classical constructions. An *ellipse*, with foci $\mathbf{f}_1$ and $\mathbf{f}_2$, can be defined as the set of all vectors $\mathbf{x}$ in the Euclidean plane which satisfy the equation

$$|\mathbf{x} - \mathbf{f}_1| + |\mathbf{x} - \mathbf{f}_2| = 2a.$$

Here $2a > |\mathbf{f}_1 - \mathbf{f}_2|$ is a constant equal to the *major axis* of the ellipse. The ratio $\varepsilon = |\mathbf{f}_1 - \mathbf{f}_2|/2a$

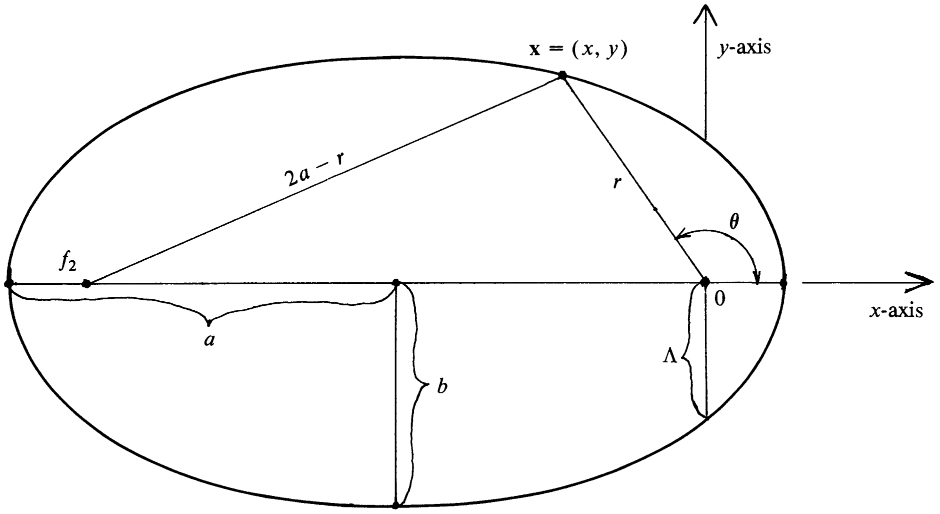

FIG. 2

of the distance between foci to the major axis is called the *eccentricity* of the ellipse. Note that $0 \leq \varepsilon < 1$, where $\varepsilon = 0$ only for a circle. If we choose cartesian coordinates x, y so that $\mathbf{f}_1$ is the origin and $\mathbf{f}_2$ is the point $(-2\varepsilon a, 0)$, then the distances r and $2a - r$ of (x, y) from $\mathbf{f}_1$ and $\mathbf{f}_2$ are given by the equations

$$x^2 + y^2 = r^2, \quad (x + 2\varepsilon a)^2 + y^2 = (2a - r)^2.$$

(Fig. 2.) Subtracting one equation from the other and dividing by $4a$, we obtain

$$(8) \qquad r = \sqrt{x^2 + y^2} = \Lambda - \varepsilon x,$$

where Λ denotes the constant $(1 - \varepsilon^2)a$, known as the "semi-latus-rectum."

More generally, if we fix any constants $\varepsilon \geq 0$ and $\Lambda > 0$, then the locus described by Equation (8) is called a *conic section* with focus at the origin, and with eccentricity ε. Squaring both sides of (8), we obtain

$$(1 - \varepsilon^2)x^2 + y^2 = \Lambda^2 - 2\Lambda\varepsilon x.$$

Evidently this equation describes a *parabola* if $\varepsilon = 1$, or a *hyperbola* if $\varepsilon > 1$. Substituting $x = r\cos\theta$, $y = r\sin\theta$ in (8), we obtain the equation

$$(3) \qquad r = \Lambda/(1 + \varepsilon\cos\theta)$$

which defines such a conic section, with focus at the origin, in polar coordinates. In the elliptic case $\varepsilon < 1$, note that the quadratic equation relating x and y can be put in the form $(x + \varepsilon a)^2/a^2 + y^2/b^2 = 1$, where $b = a\sqrt{1 - \varepsilon^2}$. Thus this ellipse is centered at the point $(-\varepsilon a, 0)$, and has principal axes $2a \geq 2b$.

Given the constant $k > 0$, we can introduce the time t into this geometrical picture by means of Kepler's Second Law, $d\theta/dt = h/r^2$, or in other words $t = \int r^2\, d\theta/h$, where $h = \sqrt{k\Lambda}$. Similarly, Levi-Civita's fictitious time can be introduced as the integral $s = \int dt/r = \int r\, d\theta/h$ along an orbit. In practice, it turns out to be easiest to express both time and position as functions of s. Substituting (3) in the equations $x = r\cos\theta$, $y = r\sin\theta$ and differentiating, we find that

$$dx/ds = -Ry \quad dy/ds = h\varepsilon + (1 - \varepsilon^2)Rx,$$

and therefore $d^2y/ds^2 = 2Ey$, where $R = h/\Lambda$, and $2E = -(1 - \varepsilon^2)R^2$. From this, it is not difficult to compute y, x, and $t = \int(\Lambda - \varepsilon x)\, ds$ as functions of s. As an example, in the negative energy case $2E = -\alpha^2 < 0$, if we normalize so that $s = 0$ at the point of closest approach, we find that

$$x = -\varepsilon a + a\cos\alpha s, \quad y = b\sin\alpha s,$$

with

$$(9) \qquad r = a(1 - \varepsilon\cos\alpha s), \quad t = a(s - \varepsilon\alpha^{-1}\sin\alpha s).$$

(Compare KEPLER.) This form of the equations again makes the elliptical shape of the orbit quite clear. In Kepler's terminology, the angle αs is known as the "eccentric anomaly," and the angle $\alpha t/a = 2\pi t/T$ is called the "mean anomaly." If the eccentricity ε tends to 1, keeping α and $a = k/\alpha^2$ fixed, so that b and Λ tend to zero, note that these expressions tend to well behaved limits. The resulting curve in the t, x plane is known as a *cycloid*. In the positive energy case, there are completely analogous formulas involving the hyperbolic sine and cosine functions. In the zero energy case, evidently y is a linear function of s, hence x is a quadratic function, and t is a cubic function of s.

Appendix 2. Inversive geometry, stereographic projection, and hyperbolic geometry.

This will be a quick review of well-known material. Further details may be found for example in Coxeter [4] or in Hilbert and Cohn-Vossen [**10**].

Let **p** be a base point in the n-dimensional Euclidean space E. The operation of *inversion*, in the unit sphere centered at **p**, is a smooth mapping from $E - \mathbf{p}$ to itself, defined as follows. *Each point* $\mathbf{x} \neq \mathbf{p}$ *of Euclidean space maps to the unique point* **y** *which lies on the ray* (= *half-line*) *which starts at* **p** *and passes through* **x**, *such that the Euclidean distance* $|\mathbf{y} - \mathbf{p}|$ *is equal to the reciprocal* $1/|\mathbf{x} - \mathbf{p}|$. Thus each point of the unit sphere $|\mathbf{x} - \mathbf{p}| = 1$ is fixed by this inversion map, but the inside and outside of the sphere are interchanged. More generally, given any constant $r > 0$, we can invert in the sphere of radius r centered at **p**. The definition is the same, except that we set $|\mathbf{y} - \mathbf{p}|$ equal to $r^2/|\mathbf{x} - \mathbf{p}|$.

It is often convenient to extend this construction by adjoining a formal *point at infinity* to Euclidean space. Then inversion maps $E \cup \infty$ to itself, with the understanding that the base point **p** maps to the point ∞ and that ∞ maps to **p**.

Here is an example to illustrate the inversion map. *Consider an* $(n - 1)$*-dimensional sphere* S *in Euclidean space, with two diametrically opposite points* **p** *and* **q**, *and let* P *be an* $(n - 1)$*-dimensional hyperplane, situated as in Figure* 3, *so that the half-line from* **p** *through* **q** *meets* P *orthogonally at a point* $\mathbf{z} \neq \mathbf{p}$. Then for any **x** of S and **y** of P which lie on a common line through **p**, the right triangle **p**, **x**, **q** is similar to the right triangle **p**, **z**, **y**. Therefore the ratio $|\mathbf{x} - \mathbf{p}|/|\mathbf{q} - \mathbf{p}|$ is equal to the ratio $|\mathbf{z} - \mathbf{p}|/|\mathbf{y} - \mathbf{p}|$. In other words, the distance from **x** to **p** is inversely proportional to the distance from y to p. *This proves that the sphere* S *maps onto the hyperplane* $P \cup \infty$ *under inversion from* **p**, *provided that the constant* r *is chosen correctly*; *and conversely it proves that* $P \cup \infty$ *maps onto* S.

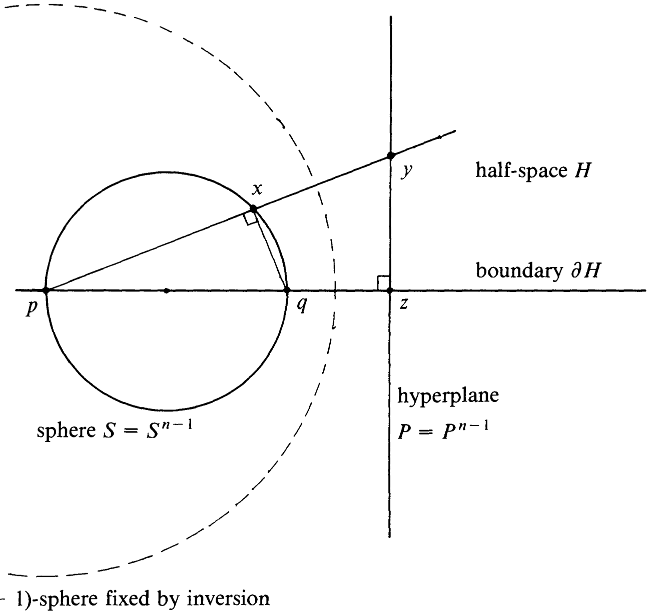

FIG. 3

Definition. The one-to-one correspondence between the sphere S and the hyperplane $P \cup \infty$ which is constructed in this way is called *stereographic projection* from **p**. Geometrically, stereographic projection maps each point $\mathbf{x} \neq \mathbf{p}$ of S to the unique point **y** of P such that **x**, **y**, and **p** all lie on a common straight line.

If P^k is a plane of dimension k in Euclidean n-space, where k can be any number between 1 and $n - 1$, then a completely analogous argument shows that the image of $P^k \cup \infty$ under inversion is either a sphere S^k (if P^k does not contain the base point **p**), or is equal to $P^k \cup \infty$ itself if P^k does contain **p**. More generally, we will prove the following.

LEMMA 2 (Steiner, 1824). *The operation of inversion carries every straight line or circle in $E \cup \infty$ to a straight line or circle. Similarly it carries every k-dimensional sphere or plane to a k-dimensional sphere or plane. Furthermore it preserves angles. That is, if two lines or circles intersect at angle θ (at a point other than* $\mathbf{p}$), *then their images under inversion also intersect at angle θ.*

To make sense of this statement, we must adopt the convention that the point at infinity belongs to every straight line or hyperplane, but that it does not belong to any circle.

As an immediate corollary of this lemma: *Stereographic projection maps every circle in the sphere S to a circle or straight line in the hyperplane $P \cup \infty$. Furthermore, stereographic projection also preserves angles.*

Proof of Lemma 2. First consider a sphere of dimension $n - 1$. By the remarks above, we need only consider the case where this sphere S^{n-1} does not contain the base point $\mathbf{p}$. It will be convenient to choose cartesian coordinates so that $\mathbf{p}$ is the origin. If $\mathbf{x}$ and $\mathbf{x}'$ are two points of S^{n-1} which lie on a common line through the origin $\mathbf{p}$, then, as in the proof of Lemma 1, the scalar product $\mathbf{x} \cdot \mathbf{x}'$ is equal to a nonzero constant, which we can write as r^2/a. It follows that the image of $\mathbf{x}'$ under inversion is equal to $a\mathbf{x}$. *Therefore, the image of S^{n-1} under inversion is precisely equal to the sphere consisting of all points $a\mathbf{x}$ with $\mathbf{x}$ in S^{n-1}.*

The proof for a sphere S^k of arbitrary dimension now follows easily. For S^k can be described as the intersection of the sphere S^{n-1} which has the same center and the same radius with a plane P^{k+1}. Since inversion maps both S^{n-1} and P^{k+1} to a sphere or plane of the same dimension, it follows that it maps their intersection to a sphere or plane.

Next consider two smooth curves Γ_1 and Γ_2 in Euclidean space which intersect at a point $\mathbf{x} \neq \mathbf{p}$. Then their images $\Gamma_1{}^*$ and $\Gamma_2{}^*$ under inversion intersect at a corresponding point $\mathbf{x}^*$. If θ is the angle between the tangent vectors of the two curves at $\mathbf{x}$, and θ^* is the angle between the corresponding tangent vectors at $\mathbf{x}^*$, we must prove that $\theta = \theta^*$. Let L_i be the straight line which is tangent to Γ_i at $\mathbf{x}$, let L_i' be the parallel straight line through $\mathbf{p}$, and let C_i be the image of L_i under inversion. Then L_i' is tangent to C_i at $\mathbf{p}$, so the angle between C_1 and C_2 at $\mathbf{p}$ is equal to θ. Evidently the angle between these two circles at their two points of intersection must be equal. Since C_i is tangent to $\Gamma_i{}^*$ at $\mathbf{x}^*$, this completes the proof. ∎

Now let us describe a very different kind of geometry. *Non-Euclidean geometry*, also called *hyperbolic geometry*, was first introduced by Lobachevsky and Bolyai by means of a collection of geometric axioms. These were identical to the axioms of Euclidean geometry, except for the hypothesis that, within a plane, it is possible to construct more than one parallel to a given line through a given point. There was no proof that these axioms were consistent for many years, until Beltrami showed that hyperbolic geometry could be modeled within Euclidean geometry. (See Milnor **[15]** for references.) One of Beltrami's Euclidean realizations of hyperbolic geometry was the *upper half-space model*, later utilized by Poincaré. It can be described as follows.

Let H be an open half-space, with boundary ∂H, in the n-dimensional Euclidean space E. It will be convenient to make use of Cartesian coordinates $\mathbf{x} = (x_1, \cdots, x_n)$, chosen so that the half-space H is defined by the inequality $x_n > 0$. Thus ∂H is the hyperplane $x_n = 0$. If Γ is a smooth curve, described parametrically by the equation $\mathbf{x} = \mathbf{x}(t)$, then we will use the notation $\int_\Gamma |d\mathbf{x}|$ for the Euclidean length $\int |d\mathbf{x}/dt|\, dt$.

Definition. By the *hyperbolic length* of a smooth curve Γ in the half-space H will be meant the integral $\int_\Gamma |d\mathbf{x}|/x_n$, along Γ, of the Euclidean length element $|d\mathbf{x}|$ divided by the Euclidean distance from ∂H.

A curve Γ in H will be called a *hyperbolic line* if it provides the hyperbolically-shortest possible path between any two of its points. In other words, there must be no curve in H which joins two points of Γ and has hyperbolic length strictly less than the hyperbolic length of the segment Γ_0 of Γ between these points. (In Riemannian geometry, such a curve Γ is called a *minimal geodesic.*)

LEMMA 3 (Beltrami, 1868). *A curve in the half-space H is a hyperbolic line if and only if it is* *either a Euclidean half-line which meets the boundary of H orthogonally, or a Euclidean semi-circle which meets ∂H orthogonally.*

Proof. First suppose that Γ is a half-line

$$x_1 = \text{constant}, \ldots, x_{n-1} = \text{constant}, x_n > 0.$$

If Δ is any other curve segment joining two points of Γ, then it is easy to check that

$$\int_\Delta |d\mathbf{x}|/x_n \geq \int_\Delta |dx_n|/x_n \geq \int_{\Gamma_0} |dx_n|/x_n.$$

In fact there is a strict inequality unless Δ is also a vertical line segment. *This proves that Γ is a hyperbolic line; and furthermore that it provides the unique hyperbolically-shortest path between any two of its points.*

Now consider an inversion mapping $\mathbf{x} \mapsto \mathbf{y}$, using any base point $\mathbf{p}$ on ∂H. To simplify the computation, let us first suppose that $\mathbf{p} = \mathbf{0}$. Differentiating the equation $\mathbf{y} = r^2\mathbf{x}/\mathbf{x} \cdot \mathbf{x}$, we obtain

$$d\mathbf{y} = r^2((\mathbf{x} \cdot \mathbf{x})\, d\mathbf{x} - 2(\mathbf{x} \cdot d\mathbf{x})\mathbf{x})/(\mathbf{x} \cdot \mathbf{x})^2.$$

If we take the dot product of this equation with itself, the terms involving $\mathbf{x} \cdot d\mathbf{x}$ cancel, so that we obtain

$$(10) \qquad d\mathbf{y} \cdot d\mathbf{y} = r^4\, d\mathbf{x} \cdot d\mathbf{x}/(\mathbf{x} \cdot \mathbf{x})^2.$$

More generally, for any choice of base point p, the appropriate formula is

$$d\mathbf{y} \cdot d\mathbf{y} = d\mathbf{x} \cdot d\mathbf{x}(r/|\mathbf{x} - \mathbf{p}|)^4.$$

It follows easily that

$$|d\mathbf{y}|/|\mathbf{y} - \mathbf{p}| = |d\mathbf{x}|/|\mathbf{x} - \mathbf{p}|.$$

since the vector $\mathbf{y} - \mathbf{p}$ is a positive multiple of $\mathbf{x} - \mathbf{p}$, this implies that

$$|d\mathbf{y}|/y_n = |d\mathbf{x}|/x_n,$$

whenever the base point p belongs to the boundary of H. *Thus the inversion mapping* $\mathbf{x} \mapsto \mathbf{y}$, *with any choice of base point in ∂H, maps the half-space H into itself so as to preserve hyperbolic length.* Such a length preserving mapping from H to itself is called a *hyperbolic isometry.*

It follows from Lemma 2 that inversion carries any half-line meeting ∂H orthogonally to a semicircle meeting ∂H orthogonally. Since we have shown that these Euclidean half-lines are hyperbolic lines, it follows that *these semi-circles must also be hyperbolic lines.*

To see that there are no other hyperbolic lines, suppose that we start with two arbitrary points $\mathbf{x}$ and $\mathbf{y}$ of the half-space H. Then it is not difficult to check that $\mathbf{x}$ and $\mathbf{y}$ lie on either a Euclidean half-line or a Euclidean semi-circle which meets ∂H orthogonally. Thus we have constructed enough hyperbolic lines to join any two points of H; and it follows that we have constructed *all* hyperbolic lines. ■

More generally, a k-dimensional subset of H is called a *hyperbolic k-plane* if it contains the hyperbolic line joining any two of its points. A similar argument shows that a subset of H is a hyperbolic k-plane if and only if it is either a Euclidean half-plane or a Euclidean hemisphere, meeting the boundary of H orthogonally, and having dimension k.

Note that the group consisting of all hyperbolic isometries from H to itself is quite large. The proof above shows that any inversion map, with base point in ∂H, belongs to this group of isometries. As a further example, it is not difficult to check that the mapping

$$(x_1, \cdots, x_n) \mapsto c_n(x_1, \cdots, x_n) + (c_1, \cdots, c_{n-1}, 0)$$

is a hyperbolic isometry which carries the point $\mathbf{u} = (0, \cdots, 0, 1)$ to a completely arbitrary point $\mathbf{c} = (c_1, \cdots, c_n)$ of H.

One characteristic property of this geometry is the following. *In any hyperbolic plane, consider a triangle Δ of hyperbolic area A, bounded by three hyperbolic line segments. The sum of the interior angles of Δ is equal to $\pi - A$.* Here we use the usual Euclidean definition of angle in the upper half-space model, but A must be defined as the integral $\int_\Delta (dA)_{\text{Euclidean}}/(x_n)^2$ so as to be invariant under hyperbolic isometries; where $(dA)_{\text{Euclidean}}$ stands for the usual Euclidean area element.

Proof. We may assume that Δ lies in the plane $x_2 = \cdots = x_{n-1} = 0$. Setting $x = x_1, y = x_n$, and integrating $A = \int\int_\Delta dx\, dy/y^2$ with respect to the y variable, we obtain $A = \int_{\partial\Delta} dx/y$, to be integrated around the boundary of Δ. Each edge of Δ is either a vertical line segment, on which dx/y is zero, or a circle arc

$$x = x_0 + a\cos\theta, \quad y = a\sin\theta,$$

on which $dx/y = -d\theta$. Thus $A = -\int_{\partial\Delta} d\theta$. An elementary argument then shows that this expression is equal to π minus the sum of the angles. ∎

By way of contrast, for any triangle on the unit 2-sphere the sum of the interior angles is equal to π *plus* the area. (Compare Coxeter [**4**], pp. 95, 297.) *On a sphere of radius r, the corresponding formula would be*

$$\Sigma(\text{interior angles}) = \pi + KA, \tag{11}$$

where $K = 1/r^2$. Let us take this last formula as a definition of the *curvature* of a space of constant curvature K. Then evidently Euclidean space has curvature $K = 0$, and hyperbolic space has curvature $K = -1$. More generally, if we change the definition of "length," multiplying all hyperbolic lengths by r and all areas by r^2, then we obtain a space of constant curvature $K = -1/r^2$.

Section 2 makes use of the *conformal unit ball model* for hyperbolic space. This is a slightly different Euclidean model which can be constructed as follows. Let us invert the upper half-space H, consisting of all points $\mathbf{x}$ with $x_n > 0$, with respect to a sphere of radius $r = \sqrt{2}$ centered at the point $\mathbf{p} = (0, \ldots, 0, -1)$. Then it is not difficult to check that the image of H under this inversion is precisely the open unit ball B, consisting of all vectors $\mathbf{y}$ with $|\mathbf{y}| < 1$. Let $\mathbf{u}$ be the upward unit vector $-\mathbf{p} = (0, \cdots, 0, 1)$. A brief computation, based on the formulas $\mathbf{x} + \mathbf{u} = 2(\mathbf{y} + \mathbf{u})/|\mathbf{y} + \mathbf{u}|^2$ and $|d\mathbf{x}| = 2|d\mathbf{y}|/|\mathbf{y} + \mathbf{u}|^2$, shows that the hyperbolic length element $|d\mathbf{x}|/x_n = |d\mathbf{x}|/\mathbf{x} \cdot \mathbf{u}$ in the half-space H corresponds to the length element

$$2|d\mathbf{y}|/(1 - \mathbf{y} \cdot \mathbf{y}) \tag{12}$$

in the ball B. *This last expression is the Riemann-Beltrami formula for the hyperbolic length element in the unit ball* $\mathbf{y} \cdot \mathbf{y} < 1$.

We can take this ball B, with the length element (12), as an alternative model for hyperbolic space. It follows from Lemmas 2 and 3 that the *hyperbolic lines* in this model are just the circle arcs or diameters which span the ball B, meeting its boundary sphere orthogonally.

This model gives us a further understanding of the richness of the group of isometries of hyperbolic space, for the subgroup consisting of all hyperbolic isometries of B which fix the origin can evidently be identified with the orthogonal group, consisting of all isometries of Euclidean space fixing the origin. The existence of such a large group of isometries would provide a key step in a proof that hyperbolic geometry, constructed in this way, satisfies all of the axioms of Euclidean geometry with the exception of the parallel axiom.

References

1. V. I. Arnold, Mathematical Methods of Classical Mechanics, Graduate Texts in Math, 60, Springer, 1978.

2. E. A. Belbruno, Two body motion under the inverse square central force and equivalent geodesic flows, Celest. Mech., 15 (1977) 467–476.

3. ______, Regularizations and geodesic flows, pp. 1–11 of Classical Mechanics and Dynamical Systems, R. L. Devaney and Z. H. Nitecki, Editors, Dekker, 1981.

4. H. S. M. Coxeter, Introduction to Geometry, 2nd ed., Wiley, 1969.

5. R. L. Devaney, Blowing up singularities in classical mechanical systems, this MONTHLY, 89 (1982) 535–552.

6. W. B. Gordon, On the relation between period and energy in periodic dynamical systems, J. Math. Mech., 19 (1969) 111–114.

7. G. Györgyi, Kepler's equation, Fock variables, Bacry's generators and Dirac brackets, Nuovo Cimento 53A, (1968) 717–736.

8. W. R. Hamilton, The hodograph or a new method of expressing in symbolic language the Newtonian law of attraction, Proc. Roy. Irish Acad., 3 (1846) 344–353 (Math. Papers v. 2, 287–294, Camb. U. Press, 1940).

9. G. Herglotz, Bemerkungen zum dritten Keplerschen Gesetz, Probleme der Astronomie: Festschrift für Hugo v. Seeliger, Springer Berlin, 1924, 197–199.

10. D. Hilbert and S. Cohn-Vossen, Geometry and the Imagination, Chelsea, 1952.

11. J. Kepler, Astronomia Nova, Prague 1609, §59,60 (cf. "Neue Astronomie," Munich-Berlin 1929, 412–413, or Gesam, Werke 3, Munich, 1937, 480–482).

12. M. Kummer, On the regularization of the Kepler problem, Comm. Math. Phys., 84 (1982) 133–152.

13. T. Levi-Civita, Fragen der klassischen und relativistischen Mechanik, Springer, 1924.

14. R. McGehee, Singularities in classical celestial mechanics, Proc. Int. Congr. Math., Helsinki, 1978, 827–834.

15. J. Milnor, Hyperbolic geometry: the first 150 years, Bull. Amer. Math. Soc., 6 (1982) 9–24.

16. J. Moser, Regularization of Kepler's problem and the averaging method on a manifold, Comm. Pure App. Math., 23 (1970) 609–636.

17. Yu. S. Osipov, Geometrical interpretation of Kepler's problem (Russian), Uspehi Mat. Nauk, 27 #2 (1972) p. 161.

18. ______, The Kepler problem and geodesic flows in spaces of constant curvature, Celest. Mech., 16 (1977) 191–208.

19. B. Riemann, Ueber die Hypothesen welche der Geometrie zu Grunde liegen, Abh. K. G. Wiss. Göttingen, 13 (1868).

20. E. L. Stiefel and G. Scheifele, Linear and Regular Celestial Mechanics, Springer, 1971.

21. J. L. Synge and B. A. Griffith, Principles of Mechanics, McGraw Hill, 2nd ed., 1949.

22. A. Wintner, The Analytical Foundations of Celestial Mechanics, Princeton U. Press, 1941.

SHORT THEOREMS WITH LONG PROOFS

JOEL SPENCER

Department of Mathematics, State University of New York, Stony Brook, NY 11794

Long proofs are an anathema to mathematicians. Part of mathematics' uniqueness is the verifiability of an argument. This is being called into question with the existence of proofs of inordinate length. A folk theorem, surely known in the 1930's, is particularly germane today. A recent note in which F. H. Norwood [1] appears unaware of this result provided the motivation for these remarks. We give three variants, with increasing formality.

(i) There exist short theorems with long proofs.

(ii) There exist theorems T whose shortest proofs have length at least $10^{100}2^{2^n}$ where n is the length of T.

(iii) For all recursive functions F there exist theorems T whose shortest proofs have length at least $F(n)$ where n is the length of T.

Joel Spencer received his Ph.D. degree from Harvard University in 1970 under the direction of Andrew Gleason. His research interests are in combinatorial analysis, where he considers himself a disciple of Paul Erdős. He has taught at UCLA, MIT, and, since 1975, at SUNY at Stony Brook. He spent the academic years 1976-77 in Budapest and 1980-81 in Rehovot (Israel) and Reading (England). He enjoys coaching Putnam teams for his students and soccer teams for his children.