EPICYCLES AND SYMMETRY

Is torment so fixed in our way of thinking?
Certainly, for those who, using numbers, vex the Inexpressibles by expressing them.
—Johannes Kepler, *Harmonice Mundi*

Contents

1. The inverse square law

Let the sun lie at the origin of two-dimensional space. Let $\boldsymbol{x}(t)$ be the trajectory of a planet, and assume that its mass is $m = 1$. Newton's law of gravity says that

$$\boldsymbol{x}'' = -\frac{k}{r^2}\boldsymbol{u},$$

where $r := |\boldsymbol{x}|$ is the distance from the origin, $\boldsymbol{u} := \frac{1}{r}\boldsymbol{x}$ is the unit vector pointing toward the planet, and k is a constant. In other words, the planet is attracted toward the sun by a force which varies inversely with the square of the distance.

The velocity is $\boldsymbol{v} := \boldsymbol{x}'$. Define the energy

$$\begin{aligned} E &:= (\text{potential energy}) + (\text{kinetic energy}) \\ &= -\frac{k}{r} + \frac{1}{2}|\boldsymbol{v}|^2 \end{aligned}$$

and the angular momentum

$$\begin{aligned} L &:= (\text{area of parallelogram spanned by } \boldsymbol{x} \text{ and } \boldsymbol{v}) \\ &= x_1 v_2 - x_2 v_1. \end{aligned}$$

Proposition 1.1. *E and L are constant with respect to time.*

Proof. Write $E = -\frac{k}{\sqrt{\boldsymbol{x}\cdot\boldsymbol{x}}} + \frac{1}{2}(\boldsymbol{v}\cdot\boldsymbol{v})^2$. Differentiating gives

$$\begin{aligned} E' &= \frac{1}{2}\,\frac{k}{r^3}(2\,\boldsymbol{v}\cdot\boldsymbol{x}) + \boldsymbol{v}\cdot\boldsymbol{a} \\ &= \boldsymbol{v}\cdot\underbrace{\left(\frac{k}{r^3}\boldsymbol{x} + \boldsymbol{a}\right)}_{0}, \end{aligned}$$

where $\boldsymbol{a}$ is the acceleration. The last term zero by the law of gravity.

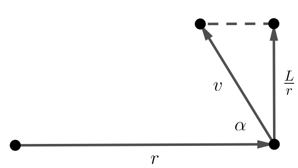
Write $L = \boldsymbol{x}\times\boldsymbol{v}$. Differentiating gives

$$\begin{aligned} L' &= \boldsymbol{x}'\times\boldsymbol{v} + \boldsymbol{x}\times\boldsymbol{v}' \\ &= \boldsymbol{v}\times\boldsymbol{v} + \boldsymbol{x}\times\boldsymbol{a}. \end{aligned}$$

The last term is zero because $\boldsymbol{a}$ and $\boldsymbol{x}$ are collinear. Thus, we have shown that L is constant for any centrally-directed force. □

If the quantities E and L are known, then we can determine the velocity based on the position. Since the area of the parallelogram spanned by $\boldsymbol{x}$ and $\boldsymbol{v}$ is L, the component of $\boldsymbol{v}$ which is perpendicular to $\boldsymbol{x}$ must have length $\frac{L}{r}$. On the other hand, solving the energy equation for $v := |\boldsymbol{v}|$ gives

$$v^2 = \frac{2k}{r} + 2E.$$



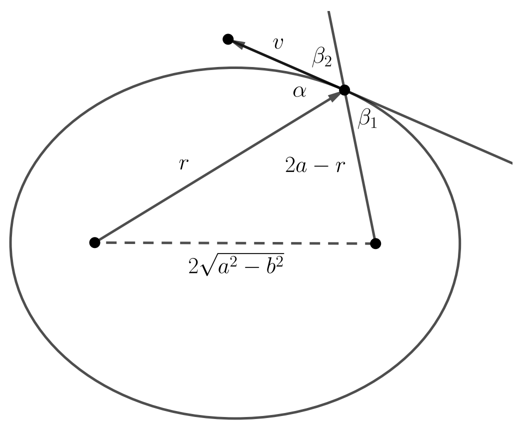
Therefore, the angle α spanned by $\boldsymbol{x}$ and $\boldsymbol{v}$ satisfies

$$\begin{aligned}\sin^2\alpha &= \frac{(\frac{L}{r})^2}{\frac{2k}{r}+2E}\\ &= \frac{L^2}{2kr+2Er^2}\end{aligned}$$

which determines $\sin\alpha$ up to a plus or minus sign.

Proposition 1.2. *The planet moves on an ellipse focused on the sun.*

Proof. Consider a hypothetical planet moving on an ellipse with major radius a and minor radius b. If we can find a and b such that the motion of the hypothetical planet is consistent with the law of gravity, then the true planet must move along the same curve.



For any point on the ellipse, the sum of distance to the two foci is $2a$. Since the distance to the sun is r by definition, the distance to the other focus must be $2a-r$. The distance between the foci is $2\sqrt{a^2-b^2}$ for any ellipse.

The velocity of the hypothetical planet is tangent to the ellipse. Let α be the angle spanned by the position and velocity vectors. We claim that $\beta_1=\alpha$. If not, then moving the hypothetical planet along the tangent line toward the smaller angle would decrease the sum of the distances to the foci. This is impossible because such a point would lie strictly inside the ellipse, while the tangent line lies outside.

We also have $\beta_2=\beta_1$ by vertical angles. Therefore $\alpha=\beta_2$.

Now apply the law of cosines to the triangle:

$$4(a^2-b^2)=r^2+(2a-r)^2+2r(2a-r)\cos(2\alpha).$$

Write $\cos(2\alpha)=1-2\sin^2\alpha$, expand, and cancel:

$$\begin{aligned}-4b^2 &= -2(4ar-2r^2)\sin^2\alpha\\ \sin^2\alpha &= \frac{b^2}{2ar-r^2}.\end{aligned}$$

But the true planet satisfies

$$\sin^2 \alpha = \frac{\frac{L^2}{-2E}}{\left(\frac{k}{-E}\right) - r^2}.$$

The hypothetical planet will match the true planet if we take $a = \frac{k}{-2E}$ and $b = \frac{L}{\sqrt{-2E}}$. Therefore, the true planet moves in an ellipse. $\square$

The proof allows us to deduce the following:

- The energy E is negative. This ensures the planet does not escape.
- Since $a \geq b$, we have $k \geq L\sqrt{-2E}$. If the energy is fixed, then then maximum possible angular momentum is $\frac{k}{\sqrt{-2E}}$.
- The eccentricity of the ellipse is

  $$e = \sqrt{1 - \frac{b^2}{a^2}} = k^{-1}\sqrt{k^2 - L^2(-2E)}.$$

  This rewrites more nicely as

  $$\sqrt{1 - e^2} = k^{-1} L \sqrt{-2E}.$$

  If the energy is fixed, the maximum possible angular momentum corresponds to a circular orbit ($e = 0$).

2. The LRL vector

The quantities E and L determine the shape of the ellipse, but they do not determine its orientation. Thus, the orientation of the ellipse is a third conserved quantity.

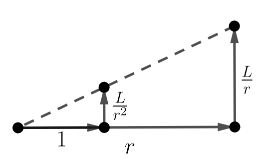
To determine a formula for this quantity, first write the law of gravity like this:

$$\boldsymbol{v}' = -\frac{k}{r^2}\boldsymbol{u}.$$

This says that the change of $\boldsymbol{v}$ depends on $\boldsymbol{u}$. To get a conserved quantity, we will find something else whose change depends on $\boldsymbol{u}$ and then cancel it with $\boldsymbol{v}$. In fact, the derivative of a unit vector is always perpendicular to it, so

$$\boldsymbol{u}' = (\text{scalar})\, \boldsymbol{u}^{\mathsf{rot}}$$

where $(\ )^{\mathsf{rot}}$ is counterclockwise rotation by 90 degrees. To find the scalar, recall from Section 1 that the component of velocity which is perpendicular to the position has length $\frac{L}{r}$.



Scaling by $\frac{1}{r}$ sends $\boldsymbol{x}$ to $\boldsymbol{u}$ and shrinks the perpendicular velocity to $\frac{L}{r^2}$. Therefore

$$\boldsymbol{u}' = \frac{L}{r^2}\,\boldsymbol{u}^{\mathsf{rot}}.$$

Canceling this with $\boldsymbol{v}'$, we find that the vector

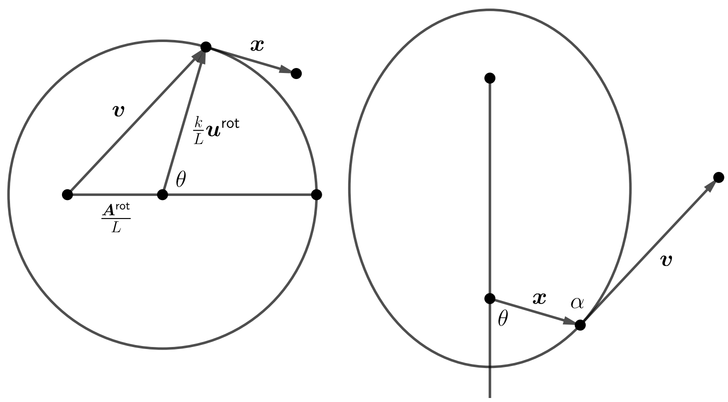
$$\boldsymbol{A} := L\boldsymbol{v}^{-\mathsf{rot}} - k\boldsymbol{u}$$

is conserved, i.e. $\boldsymbol{A}' = 0$. Here $(\ )^{-\mathsf{rot}}$ is clockwise rotation by 90 degrees. This vector is called the *Laplace–Runge–Lenz vector*, and it always points from the focus toward the closest point of the ellipse.

Solving for $\boldsymbol{v}$ gives

$$\boldsymbol{v} = \frac{\boldsymbol{A}^{\mathsf{rot}}}{L} + \frac{k}{L}\boldsymbol{u}^{\mathsf{rot}},$$

so $\boldsymbol{v}$ travels on a circle centered at $\frac{\boldsymbol{A}^{\mathsf{rot}}}{L}$ with radius $\frac{k}{L}$. The velocity circle is called a *hodograph*, and its existence is due to Hamilton.



We now use the LRL vector to give a second proof of the elliptical orbit:

Proposition 2.1. *The planet moves on an ellipse focused on the sun.*

Proof. By rotating the picture, we may assume that $\boldsymbol{x}$ points to the right, so $\boldsymbol{u}^{\mathsf{rot}}$ points upward. Then

$$\begin{aligned}
\boldsymbol{x} &= (r, 0) \\
\boldsymbol{v} &= \tfrac{\boldsymbol{A}^{\mathsf{rot}}}{L} + \tfrac{k}{L}\boldsymbol{u}^{\mathsf{rot}} \\
&= \left(-\tfrac{A}{L}\sin\theta,\ \tfrac{A}{L}\cos\theta + \tfrac{k}{L}\right) \\
L &= x_1 v_2 - x_2 v_1 \\
&= r\left(\tfrac{A}{L}\cos\theta + \tfrac{k}{L}\right)
\end{aligned}$$

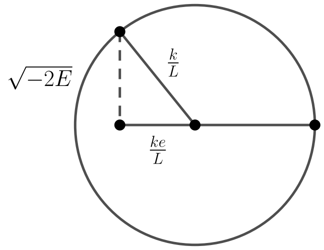
where $A := |\boldsymbol{A}|$. The last equation rewrites as

$$r = \frac{\frac{L^2}{k}}{1 + \frac{A}{k}\cos\theta},$$

This is the polar equation for an ellipse with eccentricity $e = \frac{A}{k}$. $\square$

This proof implies that the LRL vector has length $A = ke$. Therefore the distance from the origin to the circle center is $\frac{A}{L} = \frac{ke}{L}$.

Now draw a point on the circle which forms a right triangle as shown:



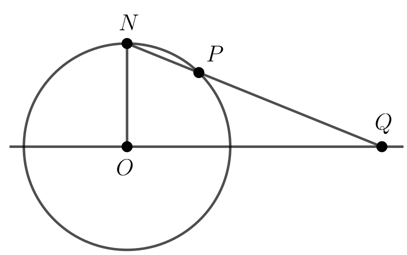
The Pythagorean theorem implies that the dotted length is $\frac{k}{L}\sqrt{1-e^2}$, and we saw in the previous section that $\sqrt{1-e^2} = k^{-1}L\sqrt{-2E}$. Therefore, the dotted length is $\sqrt{-2E}$.

Let us call the dotted length the *lateral distance* from the origin to the circle. We have found that the lateral distance of a hodograph depends only on the energy E. In other words, the set of all hodographs for a fixed energy level E equals the set of all circles whose lateral distance is $\varrho := \sqrt{-2E}$. This set of circles admits a more elegant description which is found in the next section.

3. The hidden sphere

In this section, we describe a transformation which sends points in the plane to points on a sphere which sits in three-dimensional space above and below the plane. Then we will apply this transformation to the set of all circles at a fixed lateral distance ϱ.

The transformation, called *stereographic projection*, is defined as follows. Suppose we are given a plane, a sphere whose center O lies on the plane, and a point N which is the endpoint of a radius perpendicular to the plane. The transformation sends each point P on the sphere to the point Q obtained by intersecting the line NP with the plane. We can also do stereographic projection from the plane to the sphere.

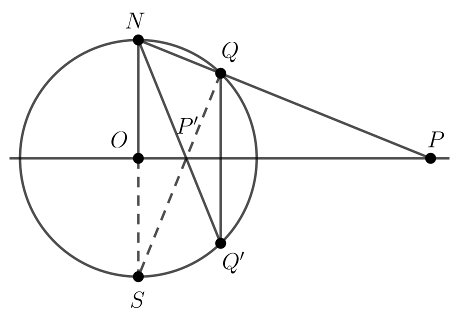


To develop the properties of stereographic projection, we need to consider a closely related notion. *Inversion* is the transformation from the plane to itself which is defined by stereographic projection from the plane to the sphere, reflection across the plane, and stereographic projection from the sphere to the plane. This has a lower-dimensional characterization:

Proposition 3.1. *If inversion sends a point P to another point P', then the points O, P, P' are collinear, and $OP \cdot OP' = ON^2$.*

This says that inversion exchanges points which are far from the origin with points that are closer to it. In particular, it sends the inside of the circle to the outside and *vice versa*, which explains the name 'inversion.'

Proof. The points O, P, P' are collinear because all of the intermediate points in the transformation lie in the plane spanned by N, O, P. Let the intermediate points be as shown:

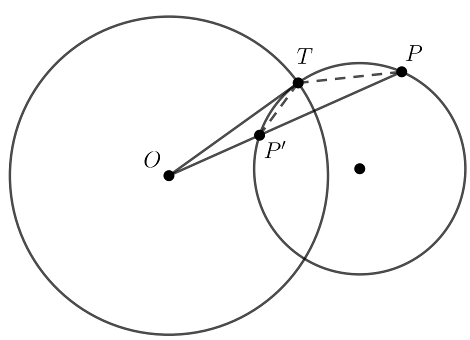


We have reflected N about O to obtain point S. Since N, P', Q' are collinear, reflection implies that Q, P', S are also collinear. The angles $\angle OP'N$ and $\angle ONP$ are congruent because they both subtend a 90-degree arc plus half of arc QQ'. Thus the triangles $OP'N$ and ONP are similar, because they share these angles and also the right angle at O. Equating the ratios of corresponding sides gives

$$\frac{OP'}{ON} = \frac{ON}{OP}. \qquad \square$$

Proposition 3.2. *Inversion sends circles to circles.*

Proof. Let a circle Γ be given. Let T be the intersection of Γ with a tangent drawn from O. Scaling the sphere of inversion does not affect whether circles are sent to circles, so we may scale so that $ON = OT$ without affecting the truth of the proposition.



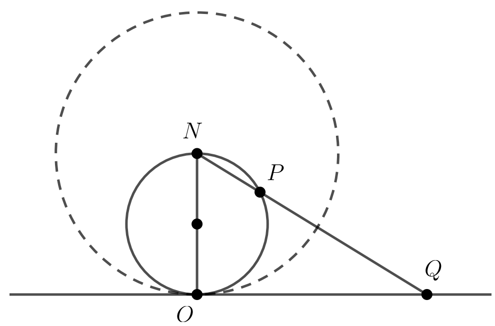
Let P be any point on the circle Γ, and let P' be the second intersection of OP with Γ. Since OT is a tangent line, angles $\angle OTP'$ and OPT subtend equal arcs on circle Γ, so they are congruent. Thus the triangles OTP' and OPT are similar, because they share these angles and also the angle at O. Equating the ratios of corresponding sides gives

$$\frac{OP'}{OT} = \frac{OT}{OP}.$$

But this is the same relation which characterizes the image of P under inversion. Therefore, inversion sends P to P'. Since this works for every point on Γ, we conclude that inversion preserves the circle Γ and therefore sends circles to circles. $\square$

Proposition 3.3. *Stereographic projection sends circles to circles.*

Proof. Scaling the sphere about the point N does not affect whether circles are sent to circles, so we may scale by a factor of $\frac{1}{2}$. In other words, we consider the stereographic projection from the plane to a sphere which contains N and is tangent to the plane at O.



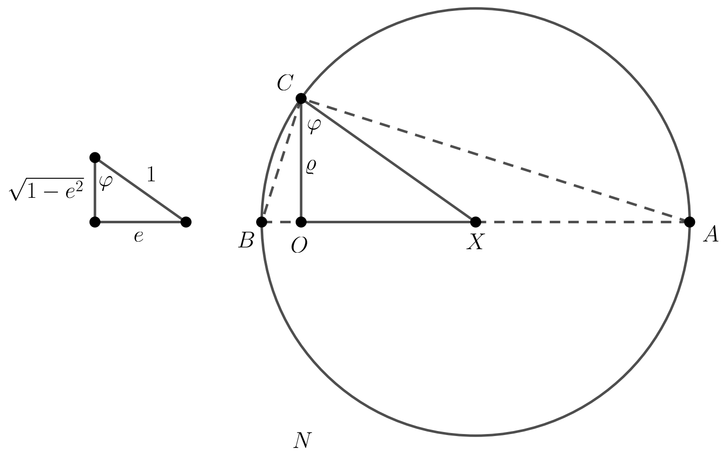
Construct a larger sphere centered at N with radius NO. So far, we have defined inversion as a transformation from the plane to itself, but we can go one dimension higher by analogy: inversion about this larger sphere sends the 3-dimensional space into itself.

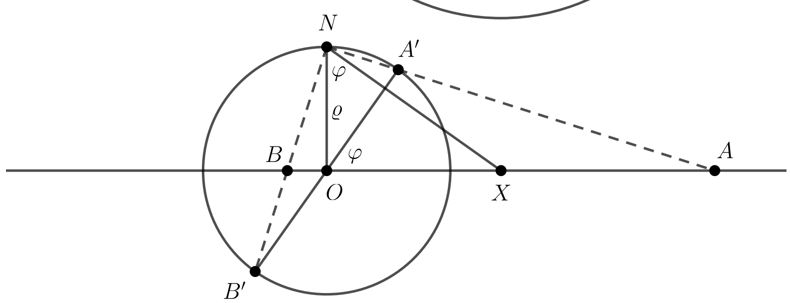
Since triangles NPO and NOQ are right triangles which share an angle at N, they are similar, which implies that inversion sends P to Q as in the previous two proofs. Thus, inversion sends the smaller sphere to the plane, and it agrees with stereographic projection when viewed in this restricted way. Because inversion sends circles to circles, the same must be true of stereographic projection. $\square$

The proof also implies the following:

Corollary 3.4. *Stereographic projection onto the sphere with radius ON can be realized via inversion across the sphere of radius $\sqrt{2} \cdot ON$ centered at N.*

Now let us go back to considering a circle with lateral distance ϱ from the origin. In the diagram below, O is the origin, X is the circle center, and C makes a right triangle COX. Let φ be angle $\angle OCX$. If we identify this circle with the hodograph of the previous section, then the sides of triangle COX are in the ratio $1 : e : \sqrt{1-e^2}$ as shown, so $e = \sin\varphi$.





Construct a sphere with radius ϱ centered at the origin O, and let N be the endpoint of a radius perpendicular to the original plane. Stereographic projection from the plane to the sphere sends the diameter AB of the old circle to the diameter $A'B'$ of a new circle lying on the sphere. Because stereographic projection can be realized via inversion, we may also think of the new circle as the image of the old circle under inversion about a sphere centered at N.

Proposition 3.5. *We have the following:*

- *The new circle is a great circle.*[1]
- $A'B'$ *is perpendicular to* NX.
- *The tilt angle* $\angle XOA'$ *equals* φ.

Proof. Since CO and NO both equal ϱ, the right triangles COX and NOX are congruent. This implies that the triangles BCA and BNA are congruent as well. Therefore, $\angle BNA$ is a right angle. Since it subtends the arc $A'B'$, it follows that the chord $A'B'$ is a diameter, so it contains O. This proves that the new circle is a great circle.

Triangle NXA is congruent to triangle CXA, which is isosceles because CX and CA are radii of the old circle. Therefore $\angle XNA = \angle NAB$. Also, the angles $\angle NAB$ and $\angle NB'A'$ subtend equal arcs in the second diagram, so they are equal. These two equalities imply $\angle XNA = \angle NB'A'$. In other words, the lines $B'A'$ and NX make equal angles with the

[1]This means that its center is the center of the sphere.

lines NB' and NA, respectively. But the latter pair of lines are perpendicular, so the same is true of the former.

The congruence of the right triangles COX and NOX implies that $\angle ONX = \angle OCX = \varphi$. The lines XO and OA' are perpendicular to ON and NX, respectively. Since the latter pair of lines make the angle φ, the same is true of the former pair. Therefore $\angle XOA' = \varphi$. □

We have obtained the elegant description of hodographs at energy $\varrho = \sqrt{-2E}$ which was promised in the previous section:

Corollary 3.6. *Stereographic projection sends all circles with lateral distance ϱ from the origin to great circles on the sphere with radius ϱ centered at the origin.*

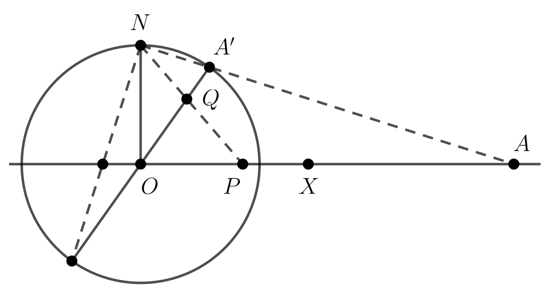
Let us call each such great circle the *lifted circle* of the corresponding hodograph.

We can also determine how the points of the old and new circles match up under stereographic projection. Let P be a point on the old circle, and let stereographic projection send it to Q on the new circle. The locations of these points are determined by the angles

$$\theta = \angle AOP$$
$$\gamma := \angle A'OQ.$$

Note that the angle θ was already defined in the previous section.

Proposition 3.7. *We have* $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$.



Proof. Scaling everything by a common factor does not affect the truth of the proposition, so we may assume that $\varrho = 1$.

Since stereographic projection can be realized via inversion across the sphere of radius $\sqrt{2} \cdot ON$ centered at N, the lower-dimensional characterization of inversion implies that $NQ \cdot NP = 2$. Squaring gives

$$(NQ)^2 \cdot (NP)^2 = 4.$$

To find (NQ), rotate the picture clockwise by angle φ so that OA' is horizontal. Then

$$N = (\sin\varphi, \cos\varphi, 0)$$
$$Q = (\cos\gamma, 0, \sin\gamma),$$

where the z-axis points out of the page. Therefore

$$\begin{aligned}(NQ)^2 &= (\sin\varphi - \cos\gamma)^2 + \cos^2\varphi + \sin^2\gamma \\ &= 2(1 - \sin\varphi\cos\gamma).\end{aligned}$$

A similar computation shows that

$$(NP)^2 = \frac{1}{\cos^2\varphi} \cdot 2(1 + \sin\varphi\cos\theta),$$

where the extra factor of $\frac{1}{\cos^2\varphi}$ comes from $(NX)^2$. The original equation becomes

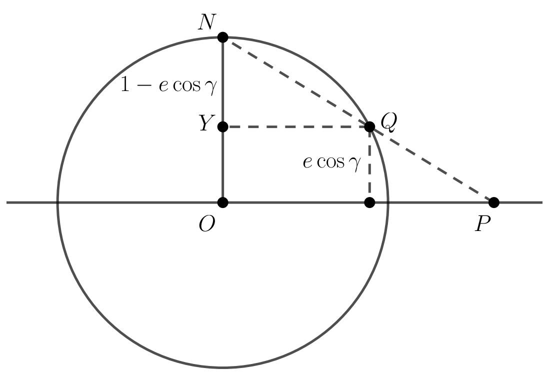
$$2(1 - \sin\varphi\cos\gamma) \cdot 2(1 + \sin\varphi\cos\theta) = 4\,\cos^2\varphi.$$

The result follows from $\sin\varphi = e$ and $\cos\varphi = \sqrt{1-e^2}$. $\square$

Proposition 3.8. *We have* $|\frac{dQ}{dP}| = 1 - e\cos\gamma$.

Proof. As before, we scale so that $\varrho = 1$.

Stereographic projection induces a derivative map from the tangent space to the sphere at Q to the tangent space to the plane at P. Since stereographic projection preserves circles, this derivative map is a rotation with a dilation by some factor. The content of the proposition is that the dilation factor is $1 - e\cos\gamma$.



To determine the dilation factor, consider what happens when Q rotates around line NO. During this motion, the arc swept out by Q is equal to $\frac{NY}{NO}$ times the arc swept out by P, because the triangles NYQ and NOP are similar. Therefore, the dilation factor equals $\frac{NY}{NO}$. Since $\varrho = 1$, this is just NY.

To compute NY, we first compute the distance from Q to the horizontal plane. We have already found that $Q = (\cos\gamma, 0, \sin\gamma)$ in a coordinate system which was rotated clockwise by angle φ. Therefore, in the original coordinate system,

$$Q = (\cos\varphi\cos\gamma, \sin\varphi\cos\gamma, \sin\gamma).$$

Therefore, the distance from Q to the plane is $\sin\varphi\cos\gamma = e\cos\gamma$. It follows that $NY = 1 - e\cos\gamma$, as desired. $\square$

4. Kepler's equation

We have twice proven that planets move along ellipses, but we have not determined where they are at any given time. To answer this, we transfer the movement of the planet to its hodograph and then to its lifted circle.

The speed along the hodograph is $|\frac{d\boldsymbol{v}}{dt}| = |\boldsymbol{a}|$, which equals $-\frac{k}{r^2}$ by the law of gravity. In our second proof of the elliptical orbit, we found that

$$r = \frac{\frac{L^2}{k}}{1 + \frac{A}{k}\cos\theta}$$

which rewrites as

$$\frac{k}{r} = \frac{k^2}{L^2}\left(1 + \frac{A}{k}\cos\theta\right).$$

We also saw that $\frac{A}{k} = e$. The radius of the hodograph was $\frac{k}{L}$, but the right triangle for the lateral distance implies that it is also equal to $\frac{\varrho}{\sqrt{1-e^2}}$. These facts allow us to rewrite the previous equation as

$$\frac{k}{r} = \varrho^2 \cdot \frac{1 + e\cos\theta}{1 - e^2}.$$

Therefore

$$\begin{aligned}\left|\frac{d\boldsymbol{v}}{dt}\right| &= \frac{\varrho^4}{k} \cdot \left(\frac{1 + e\cos\theta}{1 - e^2}\right)^2 \\ &= \frac{\varrho^4}{k} \cdot \frac{1}{(1 - e\cos\gamma)^2},\end{aligned}$$

because we have proved that $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$ in the previous section.

Now let Q be the stereographic projection of $\boldsymbol{v}$. We have $|\frac{dQ}{d\boldsymbol{v}}| = 1 - e\cos\gamma$ from the previous section, so

$$\left|\frac{dQ}{dt}\right| = \frac{\varrho^4}{k} \cdot \frac{1}{1 - e\cos\gamma}.$$

Since Q lies at angle γ on a circle of radius ϱ, we have $|\frac{dQ}{d\gamma}| = \varrho$, so

$$\frac{d\gamma}{dt} = \frac{\varrho^3}{k} \cdot \frac{1}{1 - e\cos\gamma}.$$

This shows that the motion of Q on the sphere is generally not uniform. When $e > 0$, it moves faster when it lies above the plane and slower when it lies below. When $e = 0$, however, it does move uniformly because the planetary orbit is circular in this case.

Let μ be the angle of an ideal planet with a circular orbit. This means that $e = 0$, so

$$\frac{d\mu}{dt} = \frac{\varrho^3}{k},$$

which implies $\mu = \frac{\varrho^3}{k}t$. We can compare the actual and ideal angles as follows:

$$\frac{d\gamma}{d\mu} = \frac{1}{1 - e\cos\gamma}.$$

Proposition 4.1 (Kepler's equation)**.** *We have* $\mu = \gamma - e\sin\gamma$*.*

Proof. Taking reciprocals in the previous equation gives

$$\frac{d\mu}{d\gamma} = 1 - e\cos\gamma.$$

If we view μ as a function of γ, then integration gives

$$\mu = \gamma - e\sin\gamma.$$

The constant of integration is zero if we orient things so that $\mu = 0$ when $\gamma = 0$. □

To summarize, here is how to calculate the planet's position at any given time:

- Use $\mu = \frac{\varrho^3}{k}t$ to find the ideal planet's position.
- Use Kepler's equation to find γ, which specifies a point on the lifted circle.
- Use $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$ to find θ, which specifies the planet's position on its elliptical orbit.

Let us remark that, in astronomy, the important angles θ, γ, and μ are denoted by other symbols and called by other names. They are called *anomalies* because they express the nonuniform motion of the planet along its orbit. (In the rest of the article, I have used 'nonuniformity' in place of 'anomaly' to avoid the heretical suggestion that there is any flaw in the heavens.) The angle θ is the *true anomaly* because it is the angle from which the planet is actually observed. The angle γ is the *eccentric anomaly* for a reason which will be explained in Section 6, and it is typically denoted E. The angle μ is the *mean anomaly* because it corresponds to uniform circular motion, and it is typically denoted M. None of these terms or alternative symbols will be used in the rest of the article.

5. Perturbation theory

The miraculous existence of the LRL vector is due to the fact that planetary orbits are closed curves. In general, a central force (which is not the inverse square law or the harmonic oscillator) will produce orbits that never close up.[2] Conversely, however, if a central force does produce closed curves, then we can define an analogue of the LRL vector by arbitrarily assigning a vector to each possible orbit, as long as we ensure that rotating the orbit also rotates the vector by the same amount.

Therefore, let us temporarily forget all of the theory which lies downstream of this miracle and try to determine from first principles whether the orbits are closed curves. First, we determine the circular orbits. For convenience, we identify the plane with the complex numbers, so a circular solution looks like

$$\boldsymbol{x}(t) = re^{i\omega t}$$

where ω is the angular velocity. The law of gravity becomes

$$-r\omega^2 e^{i\omega t} = -\frac{k}{r^2}e^{i\omega t},$$

so $\omega = k^{\frac{1}{2}}r^{-\frac{3}{2}}$. This is the required angular velocity for a planet to orbit at a radius r.

[2] This is Bertrand's theorem. In real life, the central force does deviate from the inverse square law due to the gravity of other planets and effects from general relativity. Therefore, real planetary orbits are not closed curves. They can be viewed as slowly-rotating ellipses, and this 'rotation' is called apsidal precession.

Next, we test whether there are closed orbits which are near a given circular orbit. Here 'near' means that we look for a solution of the form

$$\boldsymbol{x}(t)\,(1+z(t)),$$

where $z(t)$ is a small time-varying complex number. We will carry out computations to first order in $z(t)$. The left side of the law of gravity is

$$\Big(\boldsymbol{x}(t)\,(1+z(t))\Big)'' = \boldsymbol{x}''(1+z) + 2\boldsymbol{x}'z' + z''.$$

The right side of the law of gravity is

$$\begin{aligned}-k|\boldsymbol{x}(1+z)|^{-3}\boldsymbol{x}(1+z) &= -k\,|\boldsymbol{x}|^{-3}\,|1-3z|\,\boldsymbol{x}(1+z)\\ &\approx -k\,|\boldsymbol{x}|^{-3}\,(1-3\,\mathrm{Re}(z))\,\boldsymbol{x}(1+z)\end{aligned}$$

Therefore, we need to solve

$$\boldsymbol{x}''(1+z) + 2\boldsymbol{x}'z' + z'' = -k\,|\boldsymbol{x}|^{-3}\,(1-3\,\mathrm{Re}(z))\,\boldsymbol{x}(1+z).$$

Since $\boldsymbol{x}$ satisfies the law of gravity by itself, $\boldsymbol{x}'' = -k\,|\boldsymbol{x}|^{-3}\boldsymbol{x}$, so the equation cancels to

$$2\boldsymbol{x}'z' + z'' = 3k\,|\boldsymbol{x}|^{-3}\,\mathrm{Re}(z)\,\boldsymbol{x}(1+z).$$

Note that the right hand side equals $3k\,|\boldsymbol{x}|^{-3}\,\mathrm{Re}(z)\,\boldsymbol{x}$ to first order in z. Next, substituting $\boldsymbol{x} = re^{i\omega t}$ and dividing by $\boldsymbol{x}$ turns this equation into

$$2\omega z' + z'' = 3kr^{-3}\,\mathrm{Re}(z).$$

Since $\omega = k^{\frac{1}{2}}r^{-\frac{3}{2}}$, we can rewrite this more uniformly as

$$z'' = -2\omega z' + 3\omega^2\,\mathrm{Re}(z).$$

This is a linear differential equation in z. To solve it, we write it as a differential equation on the time-varying vector $\tilde{z} := (a,b,c,d)$ where $z = a+bi$ and $z' = c+di$, namely:

$$\tilde{z}' = \begin{pmatrix} 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 3\omega^2 & 0 & 0 & 2\omega\\ 0 & 0 & -2\omega & 0\end{pmatrix}\tilde{z}.$$

The generalized eigenvectors of this matrix are as follows.

- For $\lambda = 0$, there are two generalized eigenvectors: $(0,1,0,0)$ and $(1,0,0,-\frac{3}{2}\omega)$.

  The first corresponds to a solution $z = i$, hence a gravity solution

  $$\begin{aligned}\boldsymbol{x}(t) &= re^{i\omega t}(1+\varepsilon i)\\ &\approx re^{i\omega(t+\varepsilon)}\end{aligned}$$

  for small $\varepsilon > 0$. This is just the circular orbit with a shifted time coordinate.

  The second corresponds to a solution $z = 1 - i\frac{3}{2}\omega t$, hence a gravity solution

  $$\begin{aligned}\boldsymbol{x}(t) &= re^{i\omega t}(1+\varepsilon - i\tfrac{3}{2}\varepsilon\omega t)\\ &\approx r(1+\varepsilon)e^{i\omega(1-\frac{3}{2}\varepsilon)t}.\end{aligned}$$

  This is a larger circular orbit with a slowed-down time coordinate. Neither of these give genuinely new orbits.

- Let j be an imaginary unit. For $\lambda = j\omega$, there is an eigenvector $(1, 2j, j\omega, -2\omega)$. There is also an eigenvector for $\lambda = -j\omega$ obtained by conjugation.

  This corresponds to a solution
  $$\begin{aligned} z(t) &= \mathrm{Re}(e^{j\omega}) + i\,\mathrm{Re}(2je^{j\omega}) \\ &= \cos(\omega t) - 2i\sin(\omega t) \\ &= \tfrac{1}{2}(e^{i\omega t} + e^{-i\omega t}) - (e^{i\omega t} - e^{-i\omega t}) \\ &= \tfrac{3}{2}e^{-i\omega t} - \tfrac{1}{2}e^{i\omega t} \end{aligned}$$
  and hence a gravity solution
  $$\begin{aligned} \boldsymbol{x}(t) &= re^{i\omega t}\left(1 + \varepsilon\left(\tfrac{3}{2}e^{-i\omega t} - \tfrac{1}{2}e^{i\omega t}\right)\right) \\ &= re^{i\omega t} + \tfrac{3}{2}r\varepsilon - \tfrac{1}{2}r\varepsilon e^{2i\omega t}. \end{aligned}$$
  This is a genuinely new solution. It suggests the existence of noncircular closed orbits, but it does not prove this because it is valid only to first order in ε.

Which features of this first-order solution are expected and which are surprising? The solution is a sum of rotating exponentials because the eigenvalue $\lambda = j\omega$ is purely imaginary. This is expected, because if λ had a nonzero real part, say $\lambda = a + bj$, then the perturbed solution would have a factor of $e^{\lambda t} = e^{at}\,e^{ibt}$, which diverges to infinity for future times (if $a > 0$) or past times (if $a < 0$), violating conservation of energy. What is surprising, however, is that the imaginary part of λ is an integer multiple of the old angular velocity ω. It could have turned out that $\lambda = jc\omega$ for some irrational real number c, in which case the perturbed solution would not be closed and periodic, but merely quasiperiodic. In fact, this happens for every central force except for the inverse square law and the harmonic oscillator, as we have already mentioned.

It should also be surprising that the solution is a sum of exponentials in the first place. While it is true that complex exponentials (or sines and cosines) are well-suited to representing periodic functions, the premise of our analysis was to avoid assuming that the orbit is periodic. And indeed, the same analysis applied to any central force would produce sums of exponentials, even though those orbits are not periodic. The actual explanation is that perturbation theory inherently produces sums of exponentials. This is because it linearizes the differential equation, and solutions to linear differential equations with constant coefficients are exponentials (generically). Sines and cosines are distinguished not because they are eigenfunctions of the periodicity operator $f(t) \mapsto f(t + 2\pi)$ but because they are eigenfunctions of the differential operator $\frac{d}{dt}$. If some other operator like $t\frac{d}{dt}$ showed up, which would happen if the laws of physics were not constant in time, then perturbation theory would produce non-exponential functions.

But what is most surprising of all is that the ancients knew this solution. It was found in the 13th century by Mu‘ayyad al-Dīn al-‘Urḍī, who phrased it as a sequence of epicycles roughly like this:[3]

[3]This summary is from Victor Roberts, *The Planetary Theory of Ibn al-Shāṭir: Latitudes of the Planets*, Isis, vol. 57, no. 2 (Summer, 1966).

Al-mā'il ('the inclined') is the radius of the deferent circle and rotates at the planet's mean angular speed.

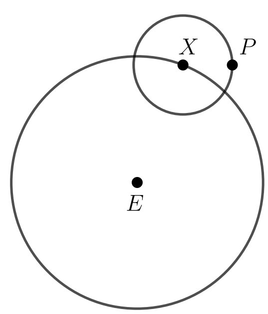
Al-ḥāmil ('the carrier') is $\frac{3}{2}$ times the Ptolemaic eccentricity and does not rotate at all.

Al-mudīr ('the rotator') is $\frac{1}{2}$ times the Ptolemaic eccentricity and rotates at twice the planet's mean angular speed.

We will explain how the ancients derived the correct first-order perturbation using synthetic geometry and how Kepler's derivation of the equation of motion $\mu = \gamma - e \sin \gamma$ relates to the geometry of the hidden sphere.

6. The Ptolemaic equant

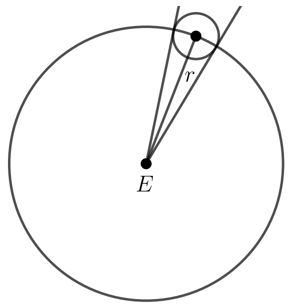
The ancient Greeks observed that the planets' revolutions around the Earth are punctuated by brief intervals of backward motion, lasting weeks to months, which they called *retrograde arcs*.[4] Apollonius explained this via a theory of epicycles, where the planet P revolves around a point X which in turn revolves around the earth E. The smaller circle is called the *epicycle*, and the larger circle is called the *deferent* (which means 'carrier'). If P and X both rotate the same way (let us say counterclockwise), and P rotates sufficiently fast compared to X, then P will have a retrograde arc (clockwise motion with respect to E) whenever it passes between E and X.



This circular symmetry makes the theory philosophically appealing as well as mechanically plausible. For it is easy to produce circular motion via a wheel or a sphere, while other nonlinear periodic motions are much harder to build. The ancient Greeks observed, however, that the circular symmetry is not supported by observations. Neither the locations of the retrograde arcs nor their sizes are perfectly uniform when viewed from the Earth.

To simplify our discussion, let us assume that the epicycle is very small and very fast.

[4]This entire section follows James Evans, *On the function and the probable origin of Ptolemy's equant*, Am. J. Phys. **52**(12), December 1984.



Then the occurrence of retrograde arcs is uniform with respect to time because the rotation of P is like the ticking of a clock:

$$d(\text{retrograde arc count}) = (\text{constant})\, dt.$$

The apparent size of the retrograde arc is determined by the angle between the tangents (shown above), which approximately satisfies

$$\begin{aligned}(\text{retrograde arc angle}) &\approx \sin(\text{retrograde arc angle})\\ &\approx \frac{2(\text{epicycle radius})}{r},\end{aligned}$$

where we have defined $r := EX$. Thus, for our discussion:

- We will interpret the 'retrograde arc occurrence count' as a proxy for time (t or dt).
- We will interpret the 'retrograde arc size' as a proxy for the distance $r = EX$.

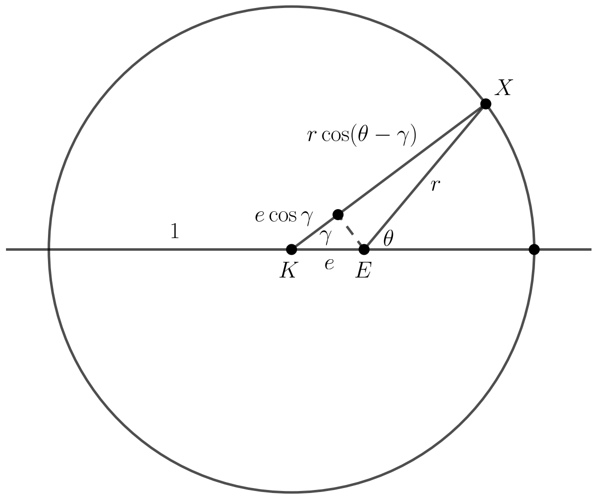
The nonuniformities observed by the ancient Greeks are the following:

- Locations of retrograde arcs: $\frac{d\theta}{dt}$ is not constant, where θ is the angle of EX.
- Sizes of retrograde arcs: r is not constant.

Keep in mind that it was impossible to directly measure r using the technology of their time.

Apollonius suggested that both nonuniformities can be explained by shifting the deferent circle to have some new center K.[5] The ratio $e := \frac{KE}{KX}$ is called the *eccentricity* because it measures how much the new center K lies outside of the Earth. For convenience, scale so that the radius equals 1.

[5]It was Hipparchus, however, who actually applied this model to real data.



The angular position of X on the circle is the angle $\gamma := \angle EKX$, which is different from θ. Since these two angles are not proportional, the rate $\frac{d\theta}{d\gamma}$ is not constant. Since X moves uniformly on the circle, the angular speed $\frac{d\gamma}{dt}$ is constant, which implies that $\frac{d\theta}{dt} = \frac{d\theta}{d\gamma} \cdot \frac{d\gamma}{dt}$ is not constant. The distance r is also not constant. Thus, the model explains the two nonuniformities. Let us state its predictions more precisely:

Proposition 6.1. *The following holds to first order in e:*

$$\frac{d\theta}{d\gamma} = 1 + e\cos\gamma$$
$$r = 1 - e\cos\gamma$$

Proof. Draw the perpendicular line from E to KX. It divides KX into two parts, $e\cos\gamma$ and $r\cos(\theta-\gamma)$, because $\angle X = \theta - \gamma$, so we have

$$1 = e\cos\gamma + r\cos(\theta-\gamma).$$

Since the difference $\theta - \gamma$ is of comparable size to e, we have $\cos(\theta-\gamma) \approx 1$ to first order in e, so the above equation becomes $r = 1 - e\cos\gamma$.

The definition of radians ensures $|dX| = d\gamma$. But, as X rotates, the component of velocity which is perpendicular to EX has length $|dX|\cos(\theta-\gamma)$, and this must equal $r\,d\theta$. Therefore

$$\cos(\theta-\gamma)\,d\gamma = r\,d\theta.$$

Again, $\cos(\theta-\gamma) \approx 1$, so this equation becomes $\frac{d\theta}{d\gamma} = \frac{1}{r} \approx 1 + e\cos\gamma$. □

Therefore, we should find the value of e such that $\frac{d\theta}{d\gamma}$ and r match their observed behavior. The paradox which Ptolemy confronted is the following:

> *For every planet, the value of e which explains the $\frac{d\theta}{d\gamma}$ nonuniformity is about twice the value of e which explains the r nonuniformity.*

In other words, the observational data require that

$$\frac{d\theta}{d\gamma} = 1 + 2e\cos\gamma$$
$$r = 1 - e\cos\gamma,$$

which is not consistent with Apollonius' model. This is the *law of bisected eccentricity.*

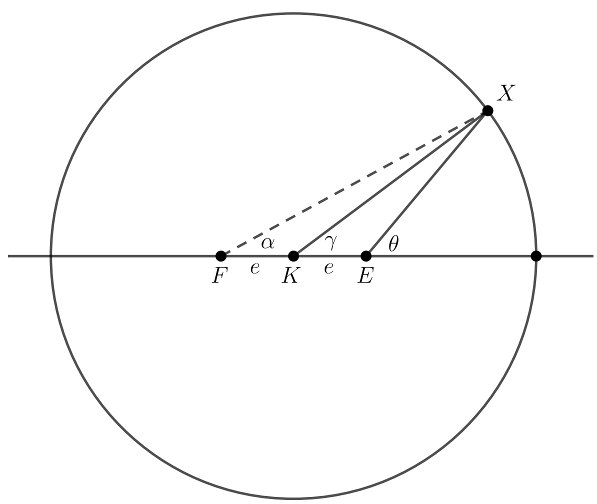
Ptolemy's idea is to keep the circle fixed but look for a different angle α which has the desired rate

$$\frac{d\theta}{d\alpha} = 1 + 2e\cos\gamma.$$

Dividing the already-established $\frac{d\theta}{d\gamma} = 1 + e\cos\gamma$ by this desired equation gives

$$\frac{d\alpha}{d\gamma} = \frac{1 + e\cos\gamma}{1 + 2e\cos\gamma} \approx 1 - e\cos\gamma.$$

This resembles the equation $\frac{d\theta}{d\gamma} = 1 + e\cos\gamma$ except for the negative sign. The $\cos\gamma$ term expresses how close X is to the right side of the circle, so a negative $\cos\gamma$ term would express how close X is to the left side of the circle. Ptolemy therefore constructs the symmetrical point F (the *equant*) and defines $\alpha := \angle KFX$.



The reasoning thus far implies the following:

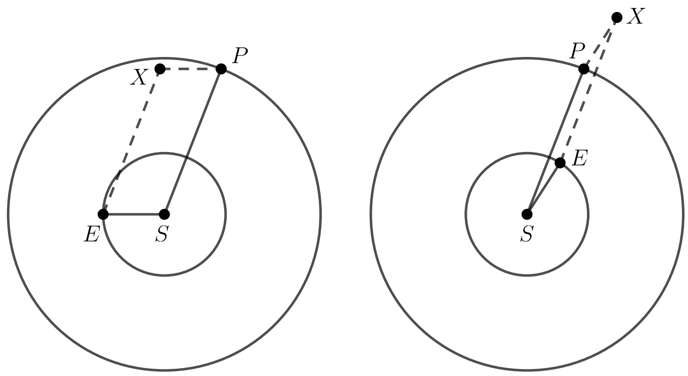
Proposition 6.2. *The following holds to first order in e:*

$$\frac{d\theta}{d\alpha} = 1 + 2e\cos\alpha$$
$$r = 1 - e\cos\alpha$$

Proof. The only remaining thing to prove is that the $\cos\gamma$ terms in the previous proposition can be replaced by $\cos\alpha$ terms. Since $\gamma - \alpha$ is of comparable size to e, the difference $\cos\gamma - \cos\alpha$ is also of comparable size to e, so $e\cos\gamma \approx e\cos\alpha$ to first order in e. ☐

In summary, Ptolemy proposes a model where the planet P rotates on its epicycle at a constant rate, while X rotates on the deferent in a nonuniform way such that $\frac{d\alpha}{dt}$ is constant. This model was unsurpassed for 1500 years and was the basis for the investigations of Copernicus and Kepler. In fact, the model is a first-order solution to the gravity equation, as we shall see in the next section.

Let us remark, in closing, that the model informs us about the orbits of planets around the sun even though it was formulated in a geocentric context. Indeed, if E and P are correctly placed in orbit around the sun S such that P is farther out than E (see below), then we may construct X by forming a parallelogram. The vectors EX and SP are identical, so the orbit of X around E is equivalent to the orbit of P around S. Knowledge of the former can be immediately transferred to knowledge of the latter, which was a part of Copernicus' work but not the whole.[6] Likewise, the epicycle is equivalent to the true orbit of the Earth.



The ancient Greeks understood that the sun has a special significance for the motions of planets, even though they did not place it at the center of the universe. The retrograde arcs of P occur when it passes between E and X, and the second diagram shows that this is equivalent to E passing between P and S. In other words, for a planet which lies farther out than the Earth, retrograde arcs occur when it is in *opposition* to the sun from the Earth's point of view. For an 'inner' planet, however, retrograde arcs occur when it is in *conjunction* with the sun. And the ancients recognized the inner planets not through their solar distances but by the simple fact that they never appear in opposition to the sun.

7. 'Urḍī's rotator

Although Ptolemy's model was known to make imperfect predictions, it was criticized mostly on philosophical rather than empirical grounds. The shifting of the deferent circle

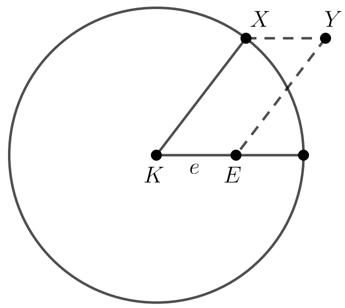
[6]One significant difficulty which Copernicus had to overcome was the fact that ancient models did not scale EX to equal SP. This is because the scaling of the Ptolemaic model is arbitrary if the raw planetary distance r cannot be measured. The main advantages of Copernicus' heliocentric model were that it had no arbitrary scaling and used much smaller epicycles, not that it made better predictions. See Owen Gingerich, *"Crisis" versus Aesthetic in the Copernican Revolution*, Vistas in Astronomy 17(1): 85–95. Another difficulty which Copernicus had to overcome was that, contrary to our simplified discussion, the model for XP differed from planet to planet, with some planets requiring many epicycles. See Figure 101 in `https://farside.ph.utexas.edu/teaching/301/lectures/node151.html`

undermines the idea that the Earth is the center of everything, and the use of the equant destroys uniform circular motion.

In fact, Apollonius had already rebutted the first criticism:

Proposition 7.1. *A shifted deferent model with eccentricity e is equivalent to an unshifted deferent model plus an epicycle of radius e which does not rotate.*

Proof. Let E be the Earth, let K be the center of the shifted deferent, and let X be a point which moves on the deferent.



Draw a new point Y by completing KEX to a parallelogram. Since $EY = KX$, the point Y moves in a circle centered at E. Since $XY = KE = e$, the point X can be obtained from Y by drawing a segment of length e which does not rotate. ☐

The idea of Apollonius' construction is that the segment KE, which is philosophically problematic, can be 'transferred' using a parallelogram. 'Urḍī realized that the parallelogram is not the only shape which accomplishes this:[7]

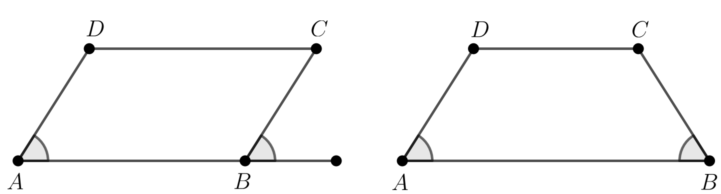
Proposition 7.2 ('Urḍī's lemma)**.** *Let $ABCD$ be a convex quadrilateral. Any two of the three following statements implies the third:*

(1) AB is parallel to CD.

(2) $AD = BC$.

(3) The interior angle $\angle A$ equals the interior ***or*** *exterior angle at B.*

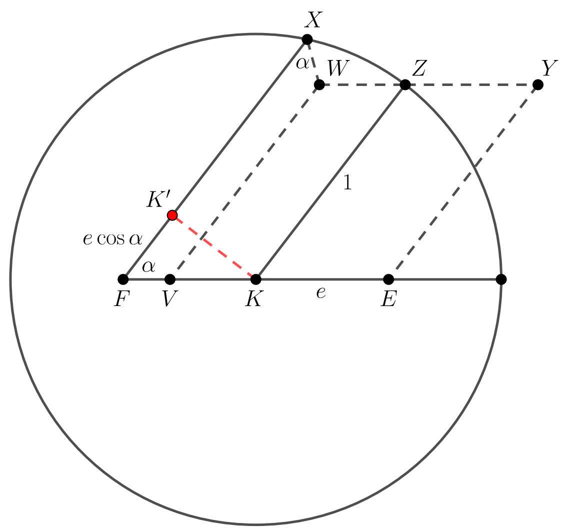
If all three statements hold, then $ABCD$ is a parallelogram or an isosceles trapezoid.



[7]The treatment of 'Urḍī's work follows G. Saliba, *Islamic science and the making of the European renaissance*, Transformations: Studies in the History of Science and Technology, MIT Press, 2007. This book also gives convincing evidence that Copernicus drew upon the work of 'Urḍī and other Islamic astronomers.

Proof. Drop perpendiculars from C and D to line AB to obtain points C' and D', respectively. Statement (1) is equivalent to saying that $DD' = CC'$. The triangles ADD' and BCC' have equal angles at C' and D', so any two of the three statements implies that the two triangles are congruent, which implies the third statement. □

'Urḍī applies this to the equant construction as follows. By definition, X moves on the deferent circle such that $\frac{d\alpha}{dt}$ is constant. Draw a radius KZ parallel to the line FX. Apollonius' construction creates a point Y such that $KEYZ$ is a parallelogram. 'Urḍī extends YZ to a point W such that $\angle FXW = \alpha$. Draw a line through W parallel to FX and let it intersect FK at a point V. This forms an isosceles trapezoid $FVWX$ (by 'Urḍī's lemma) and a parallelogram $VKZW$. Thus, 'Urḍī has 'transferred' the segment FK to the segments XW and WZ using these two special shapes.



In order to get epicycles, he needs to determine the lengths XW and WZ.

Proposition 7.3. *To first order in e, we have $XW = WZ = \frac{e}{2}$.*

Proof. By the reasoning of the previous section, $FX = 1 + e\cos\alpha$ to first order in e.

Drop a perpendicular from K to FX to obtain point K'. The right triangle FKK' implies that $FK' = e\cos\alpha$, and subtraction gives $K'X = 1$ to first order. Since $KZ = 1$, the shape $KZXK'$ is a rectangle. Therefore XZ is perpendicular to KZ and hence to VW.

On the other hand, every acute angle in the picture is equal to α by construction. Therefore, line VW makes equal angles with XW and WZ, so VW is an angle bisector of triangle XWZ. Since an angle bisector of this triangle is perpendicular to the opposite side XZ by the previous paragraph, we conclude that the triangle is isosceles. Therefore $XW = WZ$.

But we have already shown that $XW = FV$ and $WZ = VK$. Since these two add to $FK = e$, we conclude that XW and WZ both equal $\frac{e}{2}$. □

In this way, 'Urḍī arrives at the epicycle construction mentioned in Section 5.

- Y moves in a circle centered at E with radius $YE = KZ = 1$.
- The segment WY has length $\frac{3}{2}e$ and does not rotate at all.

- The segment XW has length $\frac{1}{2}e$ and rotates at twice the speed of Y.

The last bullet point follows from $\angle ZWX = 2\alpha$.

Corollary 7.4. *Ptolemy's equant construction is a first-order solution to the gravity equation.*

Proof. We have just shown that Ptolemy and ʻUrḍī agree to first order. But we already know that ʻUrḍī's construction is a first-order solution because we obtained it from perturbation theory in Section 5. □

Modern commentators have criticized the post-classical astronomers for piling epicycles upon the Ptolemaic model. They conjure up the image of astronomers fitting sine waves to prediction errors using a sort of primitive Fourier analysis with no unifying logic. While this may be true for certain planets which posed special difficulties,[8] it is decidedly false for the three circles of ʻUrḍī, which apply to every planet and were adopted by al-Shāṭir and in turn by Copernicus. These circles were derived from Ptolemy's model using math alone, as we have just done, without any observational data. And the unifying logic which ʻUrḍī sought by enforcing uniform circular motion is also what makes his solution (and not Ptolemy's) canonical from a modern mathematician's point of view: it is the unique periodic solution obtained by first-order perturbation of the circular orbit.

[8]The moon, Mercury, and Mars.

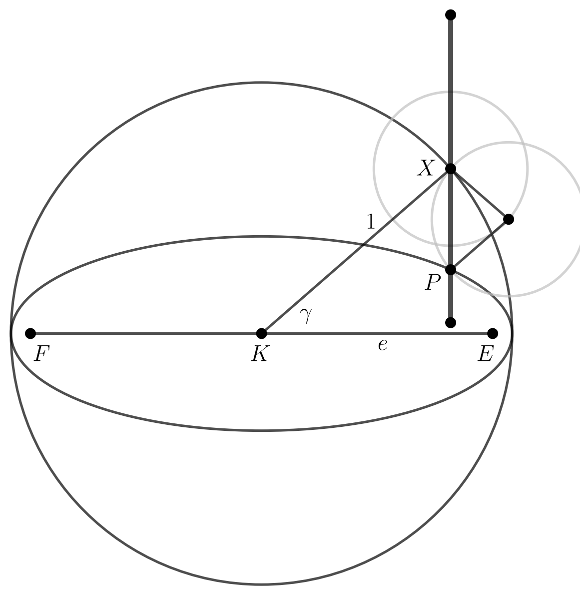
8. Kepler's law of equal areas

Let us begin by remarking that the ellipse is not the hard part. In fact, very early in his investigations, Kepler had already done computations for an epicycle model which produces an ellipse, though he did not realize this at the time.

Proposition 8.1. *Let X move in a circle centered at K with radius 1, and let P oscillate on a segment, which is centered at X and remains orthogonal to EF, so that*

$$XP = (1 - \sqrt{1-e^2}) \sin\gamma.$$

Then P moves on an ellipse with eccentricity e.



Proof. Let K be the origin. We have $X = (\cos\gamma, \sin\gamma)$, so

$$\begin{aligned} P &= \Big(\cos\gamma,\ \sin\gamma - (1-\sqrt{1-e^2})\sin\gamma\Big) \\ &= \Big(\cos\gamma,\ (\sqrt{1-e^2})\sin\gamma\Big). \end{aligned}$$

This shows that the locus of P is obtained from the unit circle by squishing the y-coordinate by a factor of $\sqrt{1-e^2}$. This produces an ellipse with eccentricity e. □

Corollary 8.2. *An ellipse can be traced out by a deferent and two epicycles.*

Proof. The oscillating motion of P can be realized by two epicycles because

$$\sin\gamma = \frac{e^{i\gamma} - e^{-i\gamma}}{2i}.$$

This trick, called the Ṭūsī couple, was discovered even before 'Urḍī's construction. □

The hard part is Kepler's law of equal areas, which we recognize today as the law of conservation of angular momentum. Kepler arrived at it by trying to replace Ptolemy's equant construction. But why did it need to be replaced? The usual account attributes this to empirical analysis of Tycho Brahe's observations of Mars, which showed that Ptolemy's model was wrong. But post-classical astronomers had long known that Ptolemy's model makes prediction errors, so this cannot be the full explanation. The true vehicle was geometrical and philosophical thinking. Something like this:

- Argument 1: Comparison of orbital periods shows that a planet's linear speed is inversely related to its distance from the Sun.

Indeed, in Chapter XX of *Mysterium Cosmographicum*, Kepler writes "everybody wants each planet to proceed with a slower motion the further its distance from the center" and traces the idea back to Aristotle.[9] He clearly distinguishes between angular speed and linear speed, observing that even a hypothesis of equal linear speeds would not be enough to explain the low angular speeds of the outer planets. This argument is quantitatively flawed because a planet on a circular orbit travels at a linear speed proportional to $r^{-\frac{1}{2}}$, not r^{-1}, as Kepler himself would later show.

- Argument 2: The equant construction implies that the (rotational component of the) planet's linear speed is inversely related to its distance from the sun.

This is true to first order in e. It follows from Proposition 6.2 because we have

$$\begin{aligned}(\text{linear speed}) &= (\text{distance})\,(\text{angular speed})\\ &= r\,\frac{d\theta}{d\alpha}\\ &= (1 - e\cos\alpha)(1 + 2e\cos\alpha)\\ &= 1 + e\cos\alpha,\end{aligned}$$

so

$$\begin{aligned}(\text{distance})\,(\text{linear speed}) &= (1 - e\cos\alpha)\,(1 + e\cos\alpha)\\ &= 1.\end{aligned}$$

Something morally equivalent to this appears in Chapter 32 of *Astronomia Nova.*[10]

- Argument 3: The planet's linear speed is caused by distance, so the endpoint of the distance interval should be identifiable as a cause of motion.

Indeed, in Chapter 33 of *Astronomia Nova*, Kepler argues that distance is prior to motion:[11]

> *For distance from the center is prior both in thought and in nature to motion over an interval. Indeed, motion over an interval is never independent of distance from the center, since it requires a space in which to be performed, while distance from the center can be conceived without motion.*

[9]Translation by A. M. Duncan.
[10]Translation by W. H. Donahue.
[11]*Ibid.*

This is surprisingly modern. Indeed, we say today that velocity is the 'derivative' of position, where the choice of terminology conceives of position as being prior to velocity.

He continues by arguing that the cause of motion must reside at an endpoint of the distance interval:

> *And since distance is a kind of relation whose essence resides in end points, while of relation itself, without respect to end points, there can be no efficient cause, it therefore follows (as has been said) that the cause of the variation of intensity of motion inheres in one or the other of the end points.*

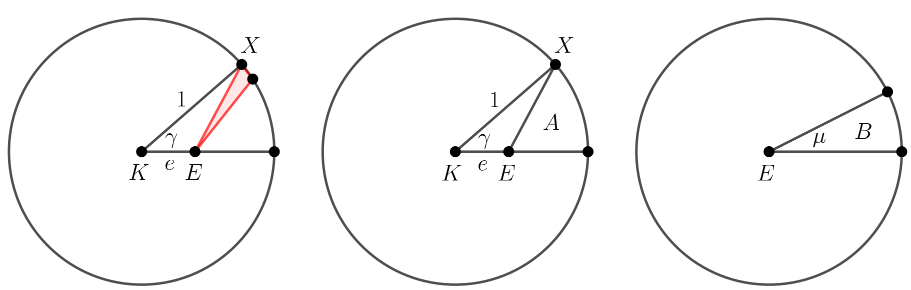
This is a good example of the physical thinking for which Kepler is rightly celebrated. The absurdity of the equant construction is not that it gives up uniform circular motion but rather that it defines the planet's motion by reference to an equant point F which does not host any physical object. In this argument, Kepler is also inspired by contemporary studies on magnetism and luminance, both of which are 'forces' which emanate from physical objects, grow weaker at large distances, and do not require circular mechanisms.

For these reasons, Kepler replaces the equant construction with the law that the (rotational component of the) planet's linear speed is inversely related to its distance from the sun. The rest of the work is purely mathematical:

Proposition 8.3 (Kepler's equation)**.** *Let a planet X travel on a shifted circle with radius 1 and eccentricity e while obeying the aforementioned law. Let its angle with respect to the circle center be γ, and let μ be the angle of an ideal planet which travels at constant speed along an orbit of the same size. Then*

$$\mu = \gamma - e \sin \gamma.$$

Proof. In order to integrate the motion of the planet around the circle, Kepler uses the only calculus available to him, the calculus of Archimedes who computed the area of the circle by assembling infinitely many triangles. Likewise, Kepler considers the motion of the planet to sweep out infinitely many triangles, one for each infinitesimal tick of time.



Each triangle has a base given by the distance EX and a height given by the perpendicular component of the planet's infinitesimal movement, i.e. the rotational component of the planet's velocity. Kepler's law says that these two are inversely related, so their product is constant. Therefore, the area of the triangle for each tick of time is the same.

If the planet begins moving at $\gamma = 0$, then the sum of these triangles up to a certain moment in time is equal to the area A, which can be computed as follows:

$$\begin{aligned}(\text{area } A) &= (\text{area of circular sector}) - (\text{area of triangle}) \\ &= \tfrac{1}{2}\gamma - \tfrac{1}{2}e\sin\gamma.\end{aligned}$$

The analogue for an ideal planet traveling at constant speed is

$$(\text{area } B) = \tfrac{1}{2}\mu.$$

Since the total area in both cases depends only on how many ticks of time have passed, we must have

$$\tfrac{1}{2}\gamma - \tfrac{1}{2}e\sin\gamma = \tfrac{1}{2}\mu$$

at each moment in time. This gives Kepler's equation. □

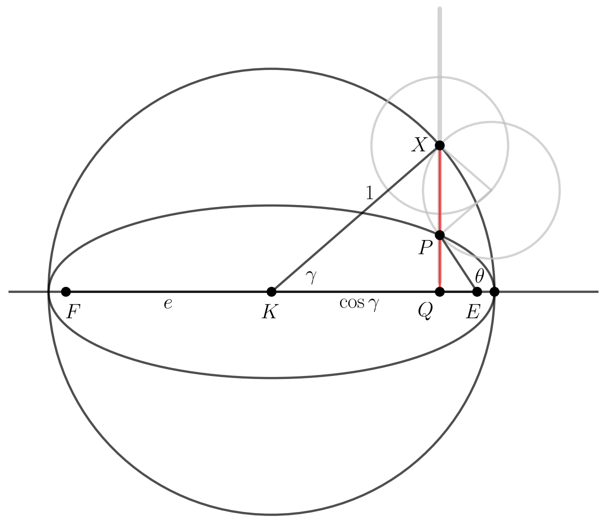
Kepler notices that the predictive errors of this model match with the epicycle model from the beginning of this section, so he combines them to get a perfectly accurate model. To calculate the planet's location at any given time:

- Use $\mu = \frac{2\pi}{(\text{orbital period})}\, t$ to find the ideal planet's position.
- Use Kepler's equation to find γ, which specifies a point X on the deferent circle.
- Shift X by $(1 - \sqrt{1-e^2})\sin\gamma$ to get the planet location P.

But the most important observable is θ, the angle of the line EP which joins the planet to the sun. Kepler computes it as follows:

Proposition 8.4. *We have* $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$.

Proof. Draw the line EP. Also, extend XP to intersect EK at Q. Since XP is perpendicular to EK by construction, the triangle EQP has a right angle at Q. Since it has the exterior angle $\angle E = \theta$, it will help us compute $\cos\theta$.



We have seen that, if K is the origin, then $P = (\cos\gamma, (\sqrt{1-e^2})\sin\gamma)$. This implies

$$QE = e - \cos\gamma$$
$$PQ = (\sqrt{1-e^2})\sin\gamma.$$

The Pythagorean theorem gives

$$\begin{aligned} PE^2 &= (e-\cos\gamma)^2 + (1-e^2)\sin^2\gamma \\ &= e^2 - 2e\cos\gamma + \cos^2\gamma + (1-e^2)(1-\cos^2\gamma) \\ &= (1-e\cos\gamma)^2. \end{aligned}$$

Therefore, the right triangle EQP implies that

$$\begin{aligned} \cos\theta &= \frac{-QE}{PE} \\ &= \frac{\cos\gamma - e}{1-e\cos\gamma}. \end{aligned}$$

Multiply by e, add one, and clear denominators to get the result. $\square$

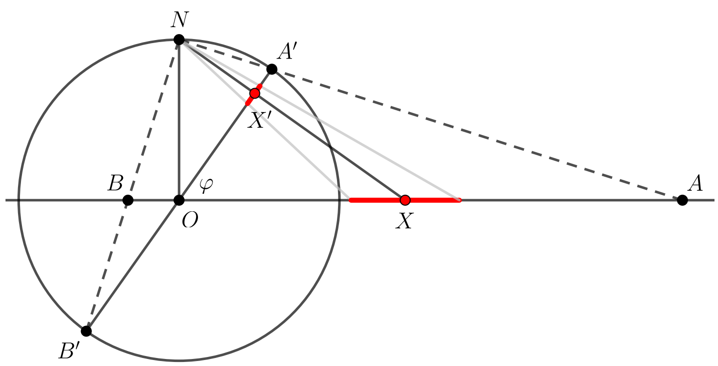
It may seem, then, that Kepler could have phrased his model as "eccentric deferent plus two epicycles plus Kepler's equation" if he were ideologically so inclined. But this is not true. Even if he had wanted to cozy up to the epicyclists, they would have rejected him because his equation violates uniform circular motion. And *vice versa*: it is the planet's progression with respect to time (via Kepler's equation), not the elliptical shape of the orbit, that prevents a finite solution by epicycles. What about us? When we solve Kepler's equation using power series, or Newton–Raphson iteration, or numerical evaluation of contour integrals, are we not just unraveling our own epicycles? The difference is that we (i.e. Kepler) have named the monster that can never be caught, pointed at the star that can never be grasped. Its name is a transcendental equation.

9. Reconciliation of geometries

We have proven the equation $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$ in two settings now, once in the lifted circle via stereographic projection and once in the circumcircle of the original elliptical orbit following in Kepler's steps. This is not a coincidence. We will give a direct geometric proof of the following:

Proposition 9.1. *For any fixed value of θ, the angle γ in the lifted circle (Section 3) equals the angle γ in the circumcircle of the ellipse defined by Kepler (Section 8).*

First, recall this picture from Section 3 which shows the hidden sphere centered at O with radius ON perpendicular to the velocity plane. The hodograph (velocity circle) has diameter AB and center X, while the lifted circle has diameter $A'B'$ and center O.



Consider the projection centered at N from the velocity plane to the plane of the lifted circle, sending X to a new point X'. This projection does not preserve angles, unlike stereographic projection onto the sphere. To visualize the distortion of angles, let us draw a very small red circle centered at X in the velocity plane.[12]

Proposition 9.2. *Projection sends the circle to an ellipse with center X' and eccentricity e.*

Proof. To show that the eccentricity is e, it is equivalent to show that the ratio of the minor to major axes is $\sqrt{1-e^2} = \cos\varphi$.

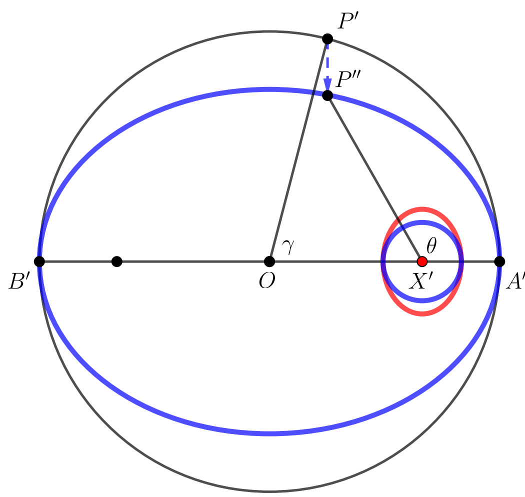
Consider two diameters of the red circle: the one which sticks out of the page, and the one which lies on line AB. Under projection, the former shrinks by a factor of $\frac{NX'}{NX}$. The latter also shrinks by this factor, but it additionally shrinks by $\sin(\frac{\pi}{2} - \varphi) = \cos\varphi$ because the velocity plane AB makes an angle $\frac{\pi}{2} - \varphi$ with the line of projection NX, while the plane of the lifted circle $A'B'$ makes a right angle with it. The projected diameters are the major and minor axes of the ellipse, respectively, and we have shown that their ratio is $\cos\varphi$. □

The failure to preserve angles can be fixed using another transformation. Namely, within the plane of the lifted circle $A'B'$, we compress the direction which sticks out of the page by a factor of $\sqrt{1-e^2}$. This shrinks the major axis of the ellipse to equal its minor axis, so the ellipse becomes a circle. Since the combined two-step transformation sends a small circle around X to a small circle around X', it preserves angles around X and X'.

---

[12]The phrase 'very small' indicates that we will make statements which are true only up to first order in the radius of the red circle.

In the next figure, which shows the plane $A'B'$, the ellipse and resulting circle are colored red and blue, respectively.



Proposition 9.3. *This transformation also turns the lifted circle $A'B'$ into an ellipse with center O, eccentricity e, and one focus at X'.*

Proof. Again, to show that the eccentricity is e, it is equivalent to show that the ratio of the minor and major axes is $\sqrt{1-e^2}$. This is true because the transformation shrinks one direction by $\sqrt{1-e^2}$.

Since the ellipse has eccentricity e, the ratio of its focus to its semi-major axis is equal to e. This implies that X' is a focus, because

$$\begin{aligned}\frac{OX'}{OA'} &= \frac{OX'}{ON} \\ &= \sin\varphi \\ &= e,\end{aligned}$$

where $OA' = ON$ because both are radii of the hidden sphere, and we have used that ONX' is a right triangle with angle $\angle N = \varphi$. □

Now we can finish the story.

Proof of Proposition 9.1. A fixed value of θ determines a point P on the velocity circle such that $\angle AXP = \theta$. Stereographic projection sends this to a point P' on the lifted circle, and we defined $\gamma = \angle A'OP'$.

On the other hand, the projection onto the plane of the lifted circle also sends P to P'. Next, the compression by a factor of $\sqrt{1-e^2}$ sends P' to the point P'' shown above, such that $P'P''$ is orthogonal to $A'B'$. Since the combined two-step transformation preserves angles around X and X', we have $\angle A'X'P'' = \angle AXP = \theta$. This picture is now identical to Kepler's diagram from Section 8, so our angle γ coincides with his. □

We have shown that the geometry of the hidden sphere is in some sense equivalent to Kepler's deconstruction of the ellipse into a circle and two epicycles. The same circle (the lifted circle) can be projected to become the velocity circle and can be compressed to become the position ellipse. The center and the focus are joined by a line to the point of projection, the fictitious north pole, where we may sit for a moment at the place of infinite velocity, at the center of the sun, where the planets arrange themselves for God in perfect nonuniform circles, fast and slow, high and low, intransigent through all that flare and void and cold, ringing out the chorus of the heavens.

10. Concluding remarks

We have explained the coincidence once with our eyes open; now let us explain it again with our eyes closed. Forget geometry and think algebraically. The Kepler problem has two stages: a rational part (θ to γ) and a transcendental part (γ to μ). I mean 'rational' in the sense that all the constructions of Euclidean geometry are limited to solving quadratic equations, and all trigonometric expressions can be written as rational functions of $\tan \frac{\alpha}{2}$. The latter implies that Kepler's expression $\gamma - e \sin \gamma$, which we obtained by integration, cannot arise from pure trigonometry. It is the transcendental core of the problem. Since it is impossible to reduce, it must be common to all solutions. And the leftovers must be rearrangements of the same rational stuff.

The torment of the epicyclists came from trying to wrangle the planets in time and not merely in space. For we have seen that an ellipse can be traced with a circle and two epicycles, but not at the same rate as a planet travels. But even this is anachronistic: the ancients could not observe planetary positions, only planetary angles, which intertwine time and space. They had no choice but to bite the unchewable.

And yet they found a strange kind of poetry, there at the boundary between the known and the unknowable. Not a scrapyard of approximations but a poetry of ideas, derivations, intelligible relations between things. Kepler calls it *proportion*: the perfect harmony in a net of $(d-1)$-dimensional shapes that lets them define the boundary of something d-dimensional.[13] For Kepler, as for the Neoplatonists and Pythagoreans, proportion is the integrability constraint that opens the hierarchy of higher dimensions.

This is precisely the program of Book X of Euclid, which Kepler says "can unfold the secrets of philosophy." The notoriously incomprehensible tome is, from a modern perspective, a classification of identities such as

$$\sqrt{3 \pm \sqrt{5}} = \sqrt{\tfrac{5}{2}} \pm \sqrt{\tfrac{1}{2}}$$
$$\sqrt{\sqrt{12} \pm 3} = \sqrt{\sqrt{27}} \pm \sqrt{\sqrt{3}}$$
$$\sqrt{10 \pm \sqrt{20}} = \sqrt{5 + \sqrt{20}} \pm \sqrt{5 - \sqrt{20}}$$

into thirteen classes based on how the formula

$$\sqrt{x \pm y} = \sqrt{\tfrac{1}{2}x + \tfrac{1}{2}\sqrt{x^2 - y^2}} \pm \sqrt{\tfrac{1}{2}x - \tfrac{1}{2}\sqrt{x^2 - y^2}}$$

simplifies for various choices of x and y.[14] What is the point of this? The world of an ancient Greek geometer was something like the following:

[13] Preface to Book I of *Harmonice Mundi*, translated by Syvia Brewda and Christopher White.

[14] This is taken from D. H. Fowler, *An Invitation to Read Book X of Euclid's Elements*, Historia Mathematica 19 (1992), though he disagrees with this framing.

- The universe consists of quantities which can be built from addition, subtraction, integer division, and square roots.
- The nice quantities are rational numbers, their square roots, and sums of these.
- Applying a square root to something nice may produce something nice, something sort-of-nice, or something *outside*.

Book X was a probe at the boundary of knowledge and an admission that there are things that lead outside, which Euclid called *alogoi* and Kepler translated as *ineffābilēs*. The punchline in Book XIII is that the golden ratio, $\frac{1+\sqrt{5}}{2}$, the emblem of divinity, is *alogos*. Divinity is not nice. As Kepler writes in Book I of *Harmonice Mundi*:[15]

> *Therefore, as much as is taken away from this wedding of the dodecahedron on account of its employing an ineffābilis proportion, is added to it conversely, because its ineffability approaches the divine.*

In Euclid's hands, segments fold into polygons, and polygons fold into the Platonic solids, which Kepler called *Archetypes*. The most irrational is also the most divine, the dodecahedron, which Kepler identifies with the Earth and the faraway stars, viewed in twelve houses, the lowest world and highest aether.

And what about epicycles?

- The universe consists of smooth functions.
- The nice functions are algebraic (rational, trigonometric, finite epicycles).
- An integral or differential equation may produce something nice, something sort-of-nice, or something *outside*.

Like the Pythagoreans who believed the world revolves around distant fire, like Euclid and Theaetetus who reckoned with the *alogous*, Kepler left the sanctum of algebraic motion and found divinity outside. In *Harmonice Mundi*, he has the gall to put ellipses into his celestial spheres, which become cradles of symmetry to a kind of sacred imperfection, echoing with a higher music that flouts the rules of harmony. From Book V:[16]

> *Accordingly the movements of the heavens are nothing except a certain everlasting polyphony (intelligible, not audible) with dissonant tunings, like certain syncopations or cadences (wherewith men imitate these natural dissonances), which tends towards fixed and prescribed clauses—the single clauses having six terms (like voices)— and which marks out and distinguishes the immensity of time with those notes. Hence it is no longer a surprise that man, the ape of his Creator, should finally have discovered the art of singing polyphonically [per concentum], which was unknown to the ancients, namely in order that he might play the everlastingness of all created time in some short part of an hour by means of an artistic concord of many voices and that he might to some extent taste the satisfaction of God the Workman with His own works, in that very sweet sense of delight elicited from this music which imitates God.*

[15] Translated by Charles Glenn Wallis.

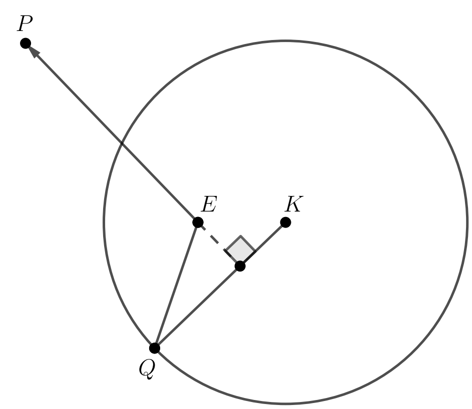
[16] *Ibid.*

We live in a time of many epicycles and much scouring of the circles for divinity. How long will we have to wait for the Kepler to our Ptolemy, for a whisper of inexpressible form in the face of unerring prediction? The anomie inflicted upon the soul of a geometer could not be better expressed than by Copernicus:[17]

> *Nor could they elicit or deduce . . . the principal consideration, that is, the structure of the universe and the true symmetry of its parts. On the contrary, their experience was just like some one taking from various places hands, feet, a head, and other pieces, very well depicted, it may be, but not for the representation of a single person; since these fragments would not belong to one another at all, a monster rather than a man would be put together from them.*

In this age so far from home, where truth was beauty and beauty, truth, I still have faith in the tripartite pattern. The inside, the outside, and the threshold. The world, the firmament, and the veil of heaven. The answer which exceeds the terms of its question. The face at the edge of the forest, unsmoothed by all our babble. I am still haunted by that invisible star in the dark behind your eyes, at the window with your back turned, hidden in plain sight, which chills our blood with dissonance and bewitches us in siren song.

Since I don't know how to conclude, I will end with an offering, which is all that synthetic geometry is still good for. So. An offering to you, hekēbóle Ápollon, and to the wise men who carried your name. Here is a vision of an alternate timeline. Let it be one more proof on the dusty pile, a candle in a fluorescent room, a prayer at my firelink shrine.



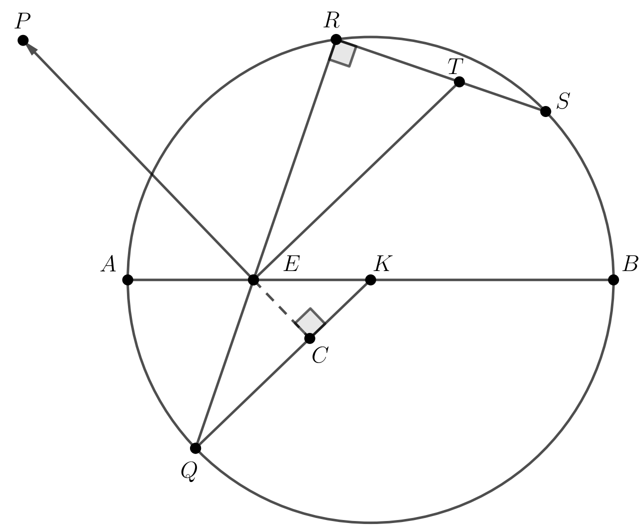
[17] Preface to *De Revolutionibus Orbium Celestium, Libri VI*, translated by Edward Rosen.

Proposition 10.1. *Suppose that a planet P and a carrier Q satisfy these laws of motion:*

- Apollonius' law: *Q moves on a circle centered at K, and $EP \perp KQ$.*
- Kepler's law: *The area of EPQ is constant.*

Then P traces out an ellipse with one focus at E.

Proof. Extend EK to a diameter AB. Let PE and KQ intersect at C. Extend QE to meet the circle at R, and let the line through R perpendicular to QR meet the circle at S. Let the line through E parallel to KQ intersect RS at T.



Since angles $\angle RBA$ and $\angle RAQ$ subtend the same arc, they are equal. Thus the triangles REB and AEQ are similar, because they share these angles and also the angle at E. The proportion $\frac{RE}{BE} = \frac{AE}{QE}$ multiplies to

$$RE \cdot QE = AE \cdot BE.$$

Since the points A, B, E are fixed, this product is constant.

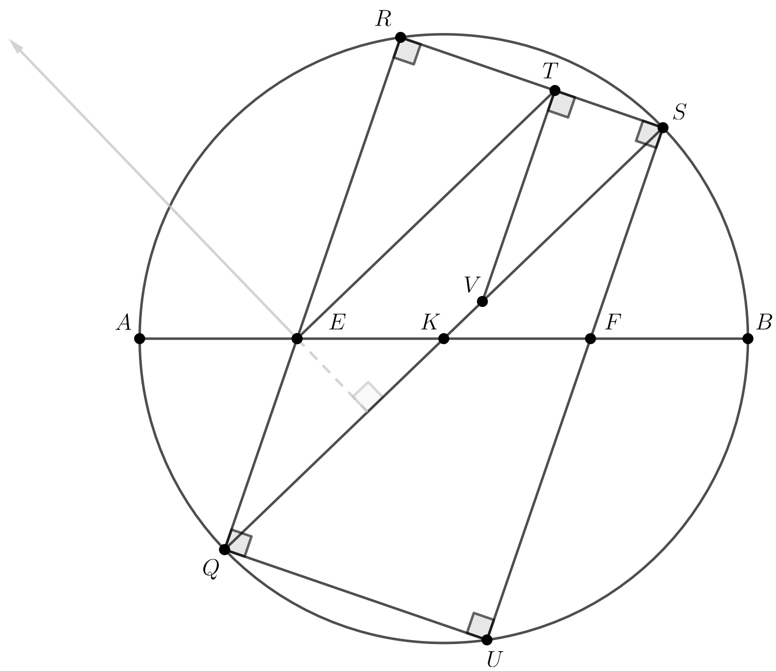
Since ET and QC are parallel, $\angle TER = \angle CQE$. Thus the triangles TER and EQC are similar, because they share these angles and also the right angles at R and C. The proportion $\frac{TE}{RE} = \frac{QE}{QC}$ multiplies to

$$TE \cdot QC = RE \cdot QE.$$

We have already shown that the right hand side is constant.

The area of EPQ is $\frac{1}{2}\, PE \cdot QC$. Dividing this by the constant quantity $TE \cdot QC$, we find that $\frac{PE}{TE}$ is constant. Also, $\angle TEP = 90°$ because ET is parallel to QC which is perpendicular to PE. These two facts imply that the locus of P is obtained from the locus of T by a 90-degree counterclockwise rotation followed by a dilation at E. Therefore, it suffices to prove that T traces out an ellipse with one focus at E.

Reflect R and E through K to get U and F, respectively. Since QRS is a right triangle and SUQ is obtained from it by reflection, $QRSU$ is a rectangle.



Let the line through T parallel to QR intersect QS at V. Since TE is parallel to VQ, the shape $TVQE$ is a parallelogram, so $TV = EQ$. Reflection through K gives $EQ = SF$. Therefore $TV = SF$. Since TV and SF are both perpendicular to TS, we conclude that $TVFS$ is a rectangle.

Since opposite sides of a parallelogram are equal, we have

$$TE = QV.$$

Since diagonals of a rectangle are equal, we have

$$TF = VS.$$

Adding these gives

$$\begin{aligned} TE + TF &= QV + VS \\ &= QS. \end{aligned}$$

Since QS is a diameter of the circle, we conclude that $TE + TF$ is constant. Therefore, T lies on an ellipse with foci at E and F, as was desired to show. $\square$

Appendix A. Isomorphism of Hamiltonian systems

This article began as an attempt to learn about the SO(4) symmetry of the Kepler problem. I eventually understood that the equivalence with geodesic motion on the sphere, which is the strongest statement of the symmetry, is best viewed as a consequence of group theory and not stereographic projection. To explain this, I will prove the equivalence using only the definition of the LRL vector (Section 2) in modern language. The relationship with stereographic projection will be discussed at the end.

I will continue to work one dimension lower, so the position space is $\mathbb{R}^2$ and the symmetry is an action of SO(3). The story works in any dimension, and all verifications can be reduced to the two-dimensional case, so we are not eliding any important difficulty. In particular, we do not use the exceptional covering map $\mathrm{SU}(2) \times \mathrm{SU}(2) \to \mathrm{SO}(4)$.

Let us first set up the Kepler problem as a Hamiltonian system. Start with the symplectic manifold $\mathcal{T}^*(\mathbb{R}^2 \smallsetminus \{0\})$ with coordinates x_1, x_2, p_1, p_2. Setting $k = 1$ for convenience, the Hamiltonian function is

$$E := -|\boldsymbol{x}|^{-1} + \tfrac{1}{2}|\boldsymbol{p}|^2.$$

Since we are only interested in the periodic orbits, we will restrict to the open subset where $E < 0$, which we denote $\mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})$.

The angular momentum $L := \boldsymbol{x} \wedge \boldsymbol{p}$ is a conserved scalar-valued function[18] whose Hamiltonian vector field generates the rotations of $\mathbb{R}^2$ about the origin. The LRL vector $\boldsymbol{A}$ is a conserved vector-valued function. Since $k = 1$, its length is $|\boldsymbol{A}| = e = \sqrt{1 - L^2(-2E)}$, so

$$L^2(-2E) + |\boldsymbol{A}|^2 = 1.$$

Let us define another conserved vector-valued function $\boldsymbol{D}$ by $\boldsymbol{A} = \sqrt{-2E}\,\boldsymbol{D}^{\mathsf{rot}}$, where $\sqrt{-2E}$ is real because we have restricted to $E < 0$. Then

$$L^2 + |\boldsymbol{D}|^2 = \frac{1}{-2E}.$$

This looks like "squared norm of angular momentum equals energy," so we are motivated to make the following definitions:

- Let $\boldsymbol{\Lambda} := (L, D_1, D_2)$ be a function whose values are three-dimensional vectors. Here D_1 and D_2 are the two coordinates of the vector $\boldsymbol{D}$.
- Let $H := \frac{1}{-4E}$ be an alternative Hamiltonian function.

The equation becomes $\frac{1}{2}|\boldsymbol{\Lambda}|^2 = H$. Furthermore, since H is a function of E, the chain rule implies that their Hamiltonian flows have the same orbits, and the corresponding trajectories differ by a rescaling of time which is *constant* along each orbit. We define the *Kepler system* to have the Hamiltonian function H instead of E because this will match with geodesic motion on the sphere.

Proposition A.1 (Hidden symmetry)**.** *The coordinates of $\boldsymbol{\Lambda}$ are scalar-valued functions on $\mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})$ which generate the Lie algebra $\mathfrak{so}(3)$ under the Poisson bracket.*

Proof. Remarkably, this does not require any nontrivial calculations. The first two relations $\{L, D_1\} = D_2$ and $\{L, D_2\} = -D_1$ follow from the fact that $\boldsymbol{D}$ is naturally defined as a

[18]Technically a bivector.

vector. The last relation $\{D_1, D_2\} = L$ can be proved by writing

$$D_2 = \sqrt{2H - L^2 - D_1^2}$$

and differentiating using the chain rule:

$$\begin{aligned}
\{D_1, D_2\} &= \left\{D_1, \sqrt{2H - L^2 - D_1^2}\right\} \\
&= \frac{1}{2\sqrt{2H - L^2 - D_1^2}} \{D_1, 2H - L^2 - D_1^2\} \\
&= \frac{1}{2D_2} \{D_1, -L^2\} \\
&= \frac{1}{2D_2} (-2L) \{D_1, L\} \\
&= \frac{1}{2D_2} (-2L) (-D_2) \\
&= L.
\end{aligned}$$

□

We want to compare this with the *geodesic system* which describes a particle moving freely on the unit sphere $S^2 \subset \mathbb{R}^3$. This is the symplectic manifold $\mathcal{T}^* S^2$ equipped with the Hamiltonian function

$$G := \tfrac{1}{2}|\boldsymbol{p}|^2.$$

Each fiber $\mathcal{T}^*_{\boldsymbol{x}} S^2$ identifies with the subspace of $\boldsymbol{p} \in \mathbb{R}^3$ satisfying $\boldsymbol{x} \cdot \boldsymbol{p} = 0$. This lets us define the angular momentum map $\ell : (\boldsymbol{x}, \boldsymbol{p}) \mapsto \boldsymbol{x} \wedge \boldsymbol{p}$ which takes values in $\mathbb{R}^3$. Its three coordinates generate the Lie algebra $\mathfrak{so}(3)$ under the Poisson bracket, and the resulting $\mathfrak{so}(3)$-action integrates to the SO(3)-action induced by rotations of S^2. This makes ℓ the moment map of a Hamiltonian group action:

$$\begin{tikzcd}
\mathrm{SO}(3) \arrow[r, phantom, "\curvearrowright"] & \mathcal{T}^* S^2 \arrow[r, "\ell"] \arrow[rr, bend right, "G"'] & \mathfrak{so}(3)^* \arrow[r, "\mathsf{invar}"] & \mathbb{R}_{>0}
\end{tikzcd}$$

Also, G factors as ℓ composed with an invariant quadratic form on $\mathfrak{so}(3)$, as shown.

Next, let us delete the zero section of $\mathcal{T}^* S^2$ and the point $0 \in \mathfrak{so}(3)^*$ to get

$$\mathrm{SO}(3) \quad \curvearrowright \quad \mathcal{T}^* S^2 \smallsetminus \mathbf{0} \xrightarrow{\ \ell\ } \mathfrak{so}(3)^* \smallsetminus \{0\} \xrightarrow{\mathsf{invar}} \mathbb{R}_{>0}$$

Then G expresses $\mathcal{T}^* S^2 \smallsetminus \mathbf{0}$ as a principal SO(3)-bundle over $\mathbb{R}_{>0}$. In fact, this group-theoretic property uniquely characterizes the geodesic system:

Proposition A.2. *Let P be a symplectic manifold equipped with a Hamiltonian group action*

$$\mathrm{SO}(3) \quad \curvearrowright \quad P \xrightarrow{\ \mu\ } \mathfrak{so}(3)^* \smallsetminus \{0\} \xrightarrow{\mathsf{invar}} \mathbb{R}_{>0}$$

If the composed map is a principal SO(3)*-bundle over $\mathbb{R}_{>0}$, then there is a (nonunique) symplectomorphism between P and $\mathcal{T}^* S^2 \smallsetminus \mathbf{0}$ which respects the moment maps.*

Proof. Let $Q := \mathcal{T}^* S^2 \smallsetminus \mathbf{0}$ for brevity.

The sheaf of SO(3)-bundle isomorphisms $\mathcal{I} := \mathcal{H}om_{\mathbb{R}_{>0}}(P, Q)$ is a principal $\mathrm{SO}(3)^{\mathsf{op}}$-bundle over $\mathbb{R}_{>0}$. Since the moment maps are principal U(1)-bundles, the requirement that the isomorphism intertwine the moment maps determines a principal U(1)-bundle over $\mathbb{R}_{>0}$, denoted $\mathscr{J} \subset \mathcal{I}$. We claim that any section of $\mathscr{J}$ works.

Choose a section of $\mathscr{J}$, hence an isomorphism of bundles φ. Let $a \in P$ be any point, and let it map to b and c as shown. It suffices to check that φ intertwines the symplectic forms on the tangent spaces at a and $\varphi(a)$.

$$\begin{tikzcd}
a & \in & P \arrow[rr, "\varphi"] \arrow[dr, "\mu"'] & & Q \arrow[dl, "\ell"] \\
b & & & \mathfrak{so}(3)^* \setminus \{0\} \arrow[d, "\text{invariant}"] & \\
c & & & \mathbb{R}_{>0} &
\end{tikzcd}$$

By rotation, we can find standard generators $f_0, f_1, f_2 \in \mathfrak{so}(3)$ such that, when viewed as functions on $\mathfrak{so}(3)^* \setminus \{0\}$, they satisfy $f_1(b) = f_2(b) = 0$. This forces $f_0 = \sqrt{2c}$.

The idea is to use the knowledge that φ is SO(3)-equivariant (by construction) and μ and ℓ are moment maps to get information about the symplectic forms at a and $\varphi(a)$. To this end, let us consider the 1-forms $df_0, df_1, df_2 \in \mathcal{T}_a^* P$ and their corresponding Hamiltonian vectors $v_0, v_1, v_2 \in \mathcal{T}_a P$ which are defined using the symplectic form $\omega_{P,a}$ on $\mathcal{T}_a^* P$.

$$\begin{tikzcd}
\text{in } \mathcal{T}_a^* P : & df_0 \arrow[d, "\omega_{P,a}"] & df_1 \arrow[d, "\omega_{P,a}"] & df_2 \arrow[d, "\omega_{P,a}"] & \\
\text{in } \mathcal{T}_a P : & v_0 \arrow[r, dash, "0"] \arrow[rr, dash, bend right, "0"'] \arrow[rrr, dash, bend right=50, "1"'] & v_1 \arrow[r, dash, "\sqrt{2c}"] \arrow[rr, dash, bend right, "0"'] & v_2 \arrow[r, dash, "0"] & v_3
\end{tikzcd}$$

The definition of the Poisson bracket implies that $\omega_{P,a}(v_i, v_j) = \{f_i, f_j\}(a)$, and the generating relations $\{f_0, f_1\} = f_2$, $\{f_1, f_2\} = f_0$, $\{f_2, f_0\} = f_1$ imply that the values $\omega_{P,a}(v_i, v_j)$ for $i, j \in \{0, 1, 2\}$ are as shown in grey.

Let us now take the unique tangent vector w at b satisfying $df_0(w) = 1$ and $df_1(w) = df_2(w) = 0$ (so w is radial) and lift it to a tangent vector v_3 at a. The definition of w implies that the values $\omega_{P,a}(v_i, v_3)$ are as shown in red. Let us examine the nonuniqueness of the lift. Any other lift would be of the form

$$v_3^{\text{other}} = v_3 + a_0 v_0 + a_1 v_1 + a_2 v_2 + a_3 v_3.$$

Since df_0 kills v_0, v_1, v_2, the condition $df_0(w) = 1$ implies that $a_3 = 0$. Next, since df_1 kills v_0, v_1, the condition $df_1(w) = 0$ implies that $a_2 = 0$. Similarly, $df_2(w) = 0$ implies that $a_1 = 0$. Therefore, the lift is well-defined up to a multiple of v_0:

$$v_3^{\text{other}} = v_3 + a_0 v_0.$$

We can play the same game with Q in place of P and obtain tangent vectors $v_i^{(Q)}$ at $\varphi(a)$. Since the original vectors v_0, v_1, v_2 came from the action of SO(3), and the same is true of the new vectors $v_0^{(Q)}, v_1^{(Q)}, v_2^{(Q)}$, we have

$$d\varphi(v_i) = v_i^{(Q)} \quad i = 0, 1, 2$$

because φ is SO(3)-equivariant. Since φ commutes with the moment maps μ and ℓ, the vector $d\varphi(v_3)$ is also a lift of w to $\varphi(a)$, so the previous paragraph implies that

$$d\varphi(v_3) = v_3^{(Q)} + a_0^{(Q)} v_0^{(Q)}.$$

Using the established properties of v_i and $v_i^{(Q)}$, we can now check that

$$\omega_{P,a}(v_i, v_j) = \omega_{Q,\varphi(a)}(d\varphi(v_i), d\varphi(v_j)) \quad i, j \in \{0, 1, 2, 3\},$$

so φ is a symplectomorphism, as desired. $\square$

A few remarks about the proposition:

(i) The main idea is that the SO(3)-action is so large that it completely determines the symplectic structure by the fact that it is a Hamiltonian group action.

(ii) The nonuniqueness of the isomorphism means that the energy hypersurfaces are good at slipping against each other.[19]

In fact, this is a general phenomenon. Given any Hamiltonian system (M, H), where M is a symplectic manifold and H is the Hamiltonian function, the flow of H is a symplectomorphism $M \to M$ which maps each energy hypersurface into itself. More generally, for any function $f : \mathbb{R} \to \mathbb{R}$, the Hamiltonian vector field $X_{f \circ H}$ equals $f' \circ H$ times X_H, so its flow also maps each energy hypersurface into itself. Thus, it is possible to 'rotate' each energy hypersurface by an arbitrary (smoothly-varying) multiple of the Hamiltonian flow.

In other words, every Hamiltonian system has many natural endomorphisms, a trait shared by all dynamical systems. One should therefore never try to prove that one dynamical system is *uniquely* isomorphic to another.

It is easy to check that every map $Q \to Q$ which respects the Hamiltonian group action arises in this way. Indeed, on each fiber over $\mathbb{R}_{>0}$, the moment map looks like the quotient map $q : \mathrm{SO}(3) \to \mathrm{SO}(3)/\mathrm{U}(1)$, and the only automorphism of SO(3) which respects q and the left SO(3)-actions is the right multiplication by U(1), which corresponds to the (geodesic) Hamiltonian flow. This completely describes the nonuniqueness in the proposition.

(iii) The simplest example of this nonuniqueness is the fact that shears preserve the volume form on $\mathbb{R}^2$. This is relevant for the proof because $v_3 \mapsto v_3 + a_0 v_0$ is a shear.

(iv) In the proposition, there is no natural way to define a function on P which is nonconstant along the Hamiltonian flow, i.e. which measures where a planet lies on its orbit. If f were such a function, then its Hamiltonian vector field X_f would be transverse to the energy hypersurfaces, so we would get a natural energy-increasing vector v_3, which is impossible due to the aforementioned nonuniqueness.

Unfortunately, we cannot apply this proposition to the Kepler system because the action of $\mathfrak{so}(3)$ does not integrate to an action of SO(3). To explain this, we first prove

[19] The celestial spheres are well-lubricated?

Proposition A.3. *Let P be a symplectic manifold with a Hamiltonian Lie-algebra action*

$$\mathfrak{so}(3) \quad \curvearrowright \quad P \xrightarrow{\mu} \mathfrak{so}(3)^* \smallsetminus \{0\} \xrightarrow{\mathsf{invar}} \mathbb{R}_{>0}$$

Every trajectory for the Hamiltonian flow of the composed map $\mathsf{invar} \circ \mu$ *is also a trajectory for the action of some element of* $\mathfrak{so}(3)$.

Proof. Since the composed map is invariant under the $\mathfrak{so}(3)$-action, the trajectory must map to a single point under μ, say $\xi \in \mathfrak{so}(3)^* \smallsetminus \{0\}$. The desired element of $\mathfrak{so}(3)$ is given by the unique linear function on $\mathfrak{so}(3)^*$ whose differential at ξ equals that of invar at the same point. Indeed, this equality of differentials implies that the Hamiltonian vector fields of the two functions agree along the fiber $\mu^{-1}(\xi)$, so their trajectories agree as well. ☐

For the Kepler problem, we have $P = \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})$, and every trajectory has the planet moving along an ellipse or an interval. The latter are called *collision trajectories* because the planet collides with the sun, and they have zero angular momentum ($L = 0$). Since every Hamiltonian trajectory is also a trajectory for the $\mathfrak{so}(3)$-action, the collision trajectories show that the $\mathfrak{so}(3)$-action does not integrate to an action of $\mathrm{SO}(3)$.

However, we can still apply the method of the proposition to show the following equivalence. The existence is due to Ligon and Schaaf.[20]

Proposition A.4 (Regularization)**.** *There exists an unique symplectomorphism*

$$\begin{tikzcd}
\mathfrak{so}(3) & \curvearrowright & \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\}) \arrow[r, "\mathbf{\Lambda}"] \arrow[d, "\mathsf{reg}"', "\wr"] & \mathfrak{so}(3)^* \smallsetminus \{0\} \arrow[r, "\mathsf{invar}"] \arrow[d, equal] & \mathbb{R}_{>0} \\
\mathfrak{so}(3) & \curvearrowright & \mathcal{T}^*\Big(S^2 \smallsetminus \{\mathsf{north\ pole}\}\Big) \smallsetminus \mathbf{0} \arrow[r, "\ell"] & \mathfrak{so}(3)^* \smallsetminus \{0\} \arrow[r, "\mathsf{invar}"] & \mathbb{R}_{>0}
\end{tikzcd}$$

which intertwines the moment maps. Hence, it also intertwines the Hamiltonians H and G and the partially-defined $\mathrm{SO}(3)$*-actions on both sides.*

Proof. We first show uniqueness. Proposition A.3 says that movement along any trajectory for the Hamiltonian flow can be specified by an angle of rotation for the $\mathfrak{so}(3)$-action. Consider the subsets

$$A := \{(\boldsymbol{x}, \boldsymbol{p}) \,|\, \boldsymbol{p} = 0\} \subset \mathcal{T}^*_-\big(\mathbb{R}^2 \smallsetminus \{0\}\big)$$
$$B := \{(\boldsymbol{x}, \boldsymbol{p}) \,|\, \boldsymbol{x} = \mathsf{south\ pole}\} \subset \mathcal{T}^*\big(S^2 \smallsetminus \{\mathsf{north\ pole}\}\big) \smallsetminus \mathbf{0}$$

Each one can be characterized as the set of all points for which the Hamiltonian flow is defined for any angle $< \pi$ but not for π.[21] Therefore, reg must send A to B. Furthermore, the moment maps give bijections from A and B to the subset of $\mathfrak{so}(3)^* \smallsetminus \{0\}$ defined by $L = 0$ (zero angular momentum). This fixes the map $\mathsf{reg} : A \to B$.

Fix $a \in A$, which goes to some $b := \mathsf{reg}(a) \in B$. The orbit of b under the partially-defined $\mathrm{SO}(3)$-action equals the energy hypersurface containing b. If reg exists, then the same must be true of a. Thus, reg is uniquely determined on the energy hypersurface containing a by the requirement that it respect the partially-defined $\mathrm{SO}(3)$-actions. Since A intersects every energy hypersurface, we conclude that reg is unique.

[20] *On the global symmetry of the classical Kepler problem*, Rep. Math. Phys. 9 (1976).

[21] The intuition is that the missing points ($0 \in \mathbb{R}^2$ and $\mathsf{north\ pole} \in S^2$) naturally distinguish their antipodal points ($\mathsf{south\ pole} \in S^2$) in terms of the $\mathfrak{so}(3)$-action.

Next we prove existence. The crux is to show that the orbit of any $a \in A$ under the partially-defined SO(3)-action equals the energy hypersurface containing a. This determines reg on the energy hypersurface of a, as we have already noted. Furthermore, since all points of A lying at the same energy level are related by a rotation of $\mathbb{R}^2$, the previous construction does not depend on the choice of a. So reg is well-defined on each energy hypersurface, and these maps assemble into a diffeomorphism which intertwines the $\mathfrak{so}(3)$-actions. The proof of Proposition A.2 then implies that reg is a symplectomorphism, as desired.

Now let us prove the crux statement. Fix $a \in A$. By rotation of $\mathbb{R}^2$, we may assume without loss of generality that a corresponds to a position on the positive x-axis with zero velocity. By comparing with the geodesic system, we expect the partially-defined SO(3)-action to be defined on the open subset $\mathrm{SO}(3)^\circ \subset \mathrm{SO}(3)$ consisting of rotations which do not send the south pole of S^2 to the north pole. Thus, we need to show that there is a (necessarily unique) diffeomorphism

$$\mathrm{SO}(3)^\circ \longrightarrow \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})|_{H(a)}$$

which sends $1 \mapsto a$ and intertwines the $\mathfrak{so}(3)$-actions. Here $\mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})|_{H(a)}$ is the energy hypersurface containing a.

To this end, consider the multiplication map

$$\mathsf{mult} : \mathrm{U}(1) \times [-\tfrac{\pi}{2}, \tfrac{\pi}{2}] \times \mathrm{U}(1) \to \mathrm{SO}(3)$$

where the factors are generated by the Hamiltonian actions of L, D_1, and D_2, respectively. Let M be the open subset of the domain which excludes $\mathrm{U}(1) \times \{0\} \times \{\pi\}$. Then mult restricts to a surjective map

$$\mathsf{mult}^\circ : M \to \mathrm{SO}(3)^\circ.$$

The purpose of this map is to give us elements of $\mathrm{SO}(3)^\circ$ whose action on a can be described explicitly. For this, we associate to each element $(\alpha, \psi, \mu) \in M$ the piecewise-linear path

$$(0, 0, 0) \to (0, \psi, 0) \to (0, \psi, \mu) \to (\alpha, \psi, \mu).$$

Integrating the $\mathfrak{so}(3)$-action along this path tells us what $\mathsf{mult}(\alpha, \psi, \mu) \cdot a$ should be. Concretely, it is the point obtained as follows:

(1) Use the LRL vector to change the eccentricity of the orbit from 1 to $\cos\psi$ without changing the direction of the major axis. Reflection symmetry shows that this must send a to the aphelion[22] of the new orbit. The sign of ψ specifies the orientation of the new orbit.

(2) Advance a along its Hamiltonian trajectory by the angle μ.

(3) Rotate $\mathbb{R}^2$ by the angle α.

This gives the action map $\mathsf{act} : M \to \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})|_{H(a)}$, where the target is the energy hypersurface containing a. The concrete description shows that the ambiguity of the paths for α and μ (since the U(1) factors are circles) does not cause a problem.

[22]The point farthest from the sun, one of the endpoints of the major axis.

Because the moment map $\mathbf{\Lambda}$ is $\mathfrak{so}(3)$-equivariant, we get a commutative diagram

$$\begin{tikzcd} M \arrow[r, "\mathsf{act}"] \arrow[d, "\mathsf{mult}"] & \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})|_{H(a)} \arrow[d, "\mathbf{\Lambda}"] \\ \mathrm{SO}(3)^\circ \arrow[r, "\mathsf{act}"] & \mathfrak{so}(3)^*|_{H(a)} \end{tikzcd}$$

where the bottom arrow is the usual coadjoint action. We need to show that this diagram has a (necessarily unique) lift.

We use the open cover of $\mathfrak{so}(3)^*|_{H(a)}$ consisting of $\{\boldsymbol{D} \neq 0\}$ and $\{L \neq 0\}$.[23]

- Over $\{\boldsymbol{D} \neq 0\}$, the map mult is a diffeomorphism, so the lift exists.
- Over $\{L \neq 0\}$, the map $\mathbf{\Lambda}$ is proper. Using a relative version of the well-known theorem that vector fields on compact manifolds are complete, we can construct the lift by directly integrating the action of $\mathfrak{so}(3)$.

This completes the proof that the lift $\mathrm{SO}(3)^\circ \to \mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})|_{H(a)}$ exists.

We also need to show that it intertwines the $\mathfrak{so}(3)$-actions. This follows from the fact that the lift is defined, in each case, by integrating the $\mathfrak{so}(3)$-action along a smoothly-varying family of paths. To see this, take any element $g \in \mathrm{SO}(3)^\circ$, and let $c_g : [0,1] \to \mathrm{SO}(3)^\circ$ be the chosen path. Intuitively, an element $\xi \in \mathfrak{so}(3)$ acts on g by sending it to some infinitesimally nearby element $\tilde{g}$. We thus get two paths from 1 to $\tilde{g}$:

- The concatenation of c_g (as path from 1 to g) and ξ (as path from g to $\tilde{g}$).
- The chosen path $c_{\tilde{g}}$ for $\tilde{g}$.

We need to show agreement between the integrals of the $\mathfrak{so}(3)$-action along two paths with the same endpoints. In general, the natural idea is to choose a smooth homotopy between the paths and try to extend the integrated $\mathfrak{so}(3)$-action to it, which may not be possible. But it is always possible in our case because the two paths are infinitesimally close to each other. (This can be rephrased as a computation with tangent vectors along the path c_g, which is rigorous but unenlightening.) We have now proven the crux statement. □

This proposition is the "equivalence with geodesic motion on the sphere" which we wanted at the beginning. It is called 'regularization' because it shows that the Kepler system has a smooth compactification which is equivalent to the geodesic system. The badly-behaved collision trajectories each acquire one new point which can be pictured as an infinitely-tight gravitational slingshot around the sun, and even the well-behaved elliptical trajectories are 'improved' into uniform great circles. The dream of the epicyclists, therefore, is finally realized as the strongest manifestation of hidden symmetry, hence the title of this article. But their tormentor, the Kepler equation, has not truly disappeared: it remains inside the symmetry itself, as Proposition A.3 shows.

The main idea of the proof is that integrating an action of $\mathfrak{so}(3)$ to an open subset $U \subseteq \mathrm{SO}(3)$ can be done in the following way:

---

[23] If we view $\mathfrak{so}(3)^*|_{H(a)}$ as a sphere, then $\{\boldsymbol{D} \neq 0\}$ deletes a pair of antipodal points, while $\{L \neq 0\}$ deletes the corresponding equator. In terms of Kepler orbits, $\boldsymbol{D} = 0$ corresponds to the two circular orbits (clockwise and counterclockwise), while $L = 0$ corresponds to collision orbits.

- Choose a smoothly-varying family of paths whose endpoints include U.
- Integrate the $\mathfrak{so}(3)$-action along these paths.
- Show that the action map descends from the family of paths to U itself.

  It suffices to do this at the level of smooth manifolds because, as we have seen, the compatibility with $\mathfrak{so}(3)$-actions comes for free.

In our situation, $\mathrm{SO}(3)^\circ$ deformation retracts onto the $\mathrm{U}(1)$ generated by L (rotations of $\mathbb{R}^2$), so it is not simply-connected. Because there is no way to specify a family of paths which hits each element of $\mathrm{SO}(3)^\circ$ exactly once, the descent problem in the third bullet must be nontrivial. The proof works because the locus where it is nontrivial happens to lie in an open subset where the action exists for a different reason (properness of $\mathbf{\Lambda}$).

Our choice of paths also enables an explicit description of reg. Let us first observe that the phase space of the Kepler system, $\mathcal{T}^*_-(\mathbb{R}^2 \smallsetminus \{0\})$, can be equivalently described as the set of positions with compatible orbits, i.e. pairs $(p, \mathcal{O})$ where $p \in \mathbb{R}^2 \smallsetminus \{0\}$ and $\mathcal{O}$ is an oriented Kepler orbit containing p. Similarly, the phase space of the geodesic system, $\mathcal{T}^*S^2 \smallsetminus \mathbf{0}$, can be described as the set of triples (s, Γ, h), where $s \in S^2$ is a position, Γ is an oriented orbit (great circle), and $h > 0$ is an energy level. The proof shows that the action of $\mathsf{mult}(\alpha, \psi, \mu) \in \mathrm{SO}(3)^\circ$ on the point $a = (\mathsf{aphelion}, \mathsf{collision\,trajectory})$ sends it to a Kepler pair $(p, \mathcal{O})$ such that

- α is the angle of the major axis of $\mathcal{O}$ (ellipse or line segment).
- $\cos\psi$ is the eccentricity of $\mathcal{O}$, and the sign of ψ determines its orientation.
- μ expresses how far p lies from the aphelion of $\mathcal{O}$, measured in units of time. Here μ is scaled so that an orbital period corresponds to $\mu = 2\pi$.

Under reg, this corresponds to the geodesic triple (s, Γ, h) where

- α specifies how the great circle Γ is rotated horizontally.
- ψ specifies how Γ is tilted vertically. Here $\psi = 0$ means that Γ lies in a vertical plane, while $\psi = \frac{\pi}{2}$ means that Γ is horizontal and hence lies in $\mathbb{R}^2$.
- μ gives the angular distance between s and the lowest point of Γ.
- The energy h is directly computed from the Kepler orbit $\mathcal{O}$.

Now we discuss stereographic projection. For a fixed energy $H = h$, which corresponds to $E = -\frac{1}{4h}$ for the original Kepler Hamiltonian, we can stereographically project both the velocity vector v and the velocity circle (hodograph) of its orbit $\mathcal{O}$ onto the sphere of radius $\varrho = \sqrt{-2E}$ as in Section 3. The resulting point and great circle $(s_{\mathsf{stereo}}, \Gamma_{\mathsf{stereo}})$ are related to the aforementioned (s, Γ) as follows:

Proposition A.5.

(i) The great circles agree: $\Gamma = \Gamma_{\mathsf{stereo}}$.

(ii) Let θ *be the angle between* p *and the aphelion when viewed from the focus* $0 \in \mathbb{R}^2$*, and define* γ *by* $(1 - e\cos\theta)(1 + e\cos\gamma) = 1 - e^2$*. Then* γ *is the angular distance between* s_{stereo} *and the lowest point of* Γ*.*

Therefore, s_{stereo} *lies below* s *and differs by an angle of* $\mu - \gamma = e\sin\gamma$*.*

Proof. The great circles agree because they are determined by the eccentricity of $\mathcal{O}$ and the direction of its major axis in the same way. Statement (ii) follows from Section 3, but there is a tricky notational discrepancy. In Section 3, we defined θ to be the angle between the planet and the perihelion (the point closest to the sun), as is standard in astronomy, while here we define θ with respect to the aphelion because the aphelion was convenient for analyzing the SO(3)-action.[24] Similarly, there we defined γ with respect to the highest point of the lifted circle Γ, while here we use the lowest point. Since the new θ and γ are the complements of the old angles, the old equations $(1 + e\cos\theta)(1 - e\cos\gamma) = 1 - e^2$ and $\mu = \gamma - e\sin\gamma$ need to have their signs switched. □

Appendix B. Literature review

Ligon and Schaaf's discovery of the regularization map was opposite to our presentation. They began with stereographic projection. It had been shown by Moser[25] that, for a fixed energy E, the stereographic projection of the velocity vector onto the sphere of radius $\varrho = \sqrt{-2E}$ gives the desired regularization map except for the discrepancy $\mu \neq \gamma$, which he resolved by reparameterizing the time domain of the Hamiltonian flow. Ligon and Schaaf resolved it in a different way by explicitly rotating each point by the appropriate angle $\mu - \gamma = e\sin\gamma$. This modified map sends Kepler trajectories to uniform great circles on S^2. Moreover, they verified that it satisfies all the conditions of Proposition A.4. We can therefore argue in two ways that their map agrees with ours: we can appeal to the uniqueness in Proposition A.4 or to the explicit description in Proposition A.5.

This account of their thinking is confirmed in the introduction to a more-recent paper by Ligon.[26] I also want to mention the paper[27] by Heckman and de Laat from which I learned about the Ligon–Schaaf regularization. They give a simplified construction of the Ligon–Schaaf map (Definition 3.2) which matches our description: it has a rotation by v_{n+1} radians, where v_{n+1} is the vertical coordinate of the unit tangent vector to Γ at s_{stereo}, and in fact $v_{n+1} = e\cos\gamma$. To see this, first observe that $v_{n+1} = \cos\gamma$ when Γ is vertical, and then rotate by ψ which shrinks the vertical coordinate by a factor of $\cos\psi = e$.

On a philosophical note, it is unusual that there should exist a technique (stereographic projection of the velocity vector) which almost solves a problem (regularization) but not

[24] In *Astronomia Nova*, Kepler defines θ with respect to the aphelion, contrary to current astronomy.

[25] J. Moser, *Regularization of Kepler's Problem and the Averaging Method on a Manifold*, Communications on Pure and Applied Mathematics 23(4) (1970). See also Section 7 of Guillemin and Sternberg, *Variations on a Theme by Kepler*, AMS Colloquium Publications 42 (1990).

[26] T. Ligon, *The symmetry of the Kepler problem, the inverse Ligon–Schaaf mapping and the Birkhoff conjecture*, PLoS ONE 13(9): e0203821. Available at `https://arxiv.org/pdf/1804.03844`

[27] G. Heckman and T. de Laat, *On the regularization of the Kepler problem*, Journal of Symplectic Geometry 10(3) (2012).

quite, and the issue ($\mu \neq \gamma$) can be fixed by an explicit modification but not by a reworking of the technique itself. It is rare that a flaw can be excised but not dissolved. And the only strategically coherent solution which I can see is one which abandons stereographic projection altogether (the proof of Proposition A.4).

An explanation for this state of affairs is suggested in Section 10: the decomposition into 'technique' and 'modification' is analogous to a decomposition into 'algebraic' and 'minimal transcendental' parts, and any two such decompositions should be related in a purely algebraic way. To show this in action, we presented an alternative decomposition in Section 8 which could be conceived like this:

- Since the desired map $(p, \mathcal{O}) \mapsto (s_{\mathsf{stereo}}, \Gamma, h)$ must respect the moment map $\mathbf{\Lambda}$ (the angular momentum and LRL vector), the incline and rotation of Γ are determined by the major axis and eccentricity of the Kepler orbit $\mathcal{O}$.
- It remains to map the points of $\mathcal{O}$ to the points of Γ in an algebraic way. Reflection symmetry implies that the endpoints of the major axis of $\mathcal{O}$ should map to the highest and lowest points of Γ. Thus, we essentially want an algebraic isomorphism $\mathbb{P}^1 \to \mathbb{P}^1$ which preserves two points, say $0, \infty \in \mathbb{P}^1$. The space of such maps is (a torsor of) the multiplicative group $\mathbb{G}_m = \mathbb{R}^\times$.
- Among the maps $\mathcal{O} \to \Gamma$ considered in the previous point, the one which causes the least distortion of arc length is that which was employed by Kepler: identify Γ with the circumcircle of $\mathcal{O}$ and map it onto $\mathcal{O}$ via an affine transformation.

The resulting map from the Kepler system to the geodesic system coincides with that defined by stereographic projection. To prove this "in a purely algebraic way," we used synthetic geometry in Section 9 to relate the two transformations directly.

Once again, however, history runs opposite to our presentation. The first 'regularization map' from the Kepler system to the geodesic system was constructed by Fock[28] in the quantum setting, where stereographic projection is the only available method. The quantum analogue of the Kepler problem is the Schrödinger equation for a wavefunction $\psi(\boldsymbol{x})$ defined on position space $\mathbb{R}^2$ with a potential $V(\boldsymbol{x}) = -\frac{1}{r}$.[29] Fock applies the Fourier transform to get an equation for a wavefunction $\hat{\psi}(\boldsymbol{p})$ on momentum space $\mathbb{R}^2$, then stereographically projects onto the unit sphere and rescales as follows:

$$\Psi(\boldsymbol{p}) := \frac{\pi}{\sqrt{8}}\, \varrho^{-5/2}\, (\varrho^2 + |\boldsymbol{p}|^2)^2\, \hat{\psi}(\boldsymbol{p}) \qquad (\varrho = \sqrt{-2E}).$$

The resulting equation for Ψ is given by an integral operator of the form

$$e^{(\text{constant})\,(\text{Laplace operator on } S^2)},$$

[28]V. Fock, *Zur Theorie des Wasserstoffatoms*, Zeitschrift für Physik 98 (1935). The slides `https://www.physics.smu.edu/scalise/P5382fa15/fock.pdf` were helpful to me.

[29]Remark for clarity: there is no relation between this ψ and the angle ψ in Appendix A.

so the solutions for Ψ are spherical harmonics.[30] The same solutions appear if you quantize the geodesic system $\mathcal{T}^*S^2$ to get the Schrödinger equation for a wave function on S^2 with zero potential, which is the desired regularization.

This appears to be regularization via stereographic projection with no reparameterization of time. However, the reparameterization is there, disguised as the rescaling from $\hat{\psi}$ to Ψ. I claim that this rescaling is the quantum analogue of the reparameterization of a classical Kepler trajectory. Since I do not know how to say this mathematically, let me attempt a physical argument. The function $|\boldsymbol{p}|^2$ is larger in the north of the sphere and smaller in the south. Therefore, the norm of $\hat{\psi}$ is smaller in the north and larger in the south relative to the norm of Ψ. The norm of a wavefunction gives the probability density of the quantum particle, and Ψ is a quantization of geodesic motion. Therefore, relative to geodesic motion, the particle corresponding to $\hat{\psi}$ is less likely to be in the north and more likely to be in the south. This corresponds to a classical particle which spends less time in the north and more time in the south, which is exactly what is accomplished by reparameterizing the classical Kepler trajectory.

Moser[31] transferred these ideas to the classical setting in the most natural possible way. He considers the composite map

$$\mathcal{T}^*\big(\mathbb{R}^2 \smallsetminus \{0\}\big) \xrightarrow{\mathsf{Fourier}} \mathcal{T}^*\mathbb{R}^2 \smallsetminus \mathbf{0} \xrightarrow{\mathsf{stereo}} \mathcal{T}^*\big(S^2 \smallsetminus \{\mathsf{north\ pole}\}\big) \smallsetminus \mathbf{0}$$

where the map Fourier exchanges position and velocity, and the map stereo is induced by stereographic projection $\mathbb{R}^2 \to S^2$ onto the unit sphere. This is a symplectomorphism, so Hamiltonian functions can be freely transferred between the domain and target. Moser defines the following ones:[32]

name	definition	key energy level
geodesic	$F = \frac{1}{2}\lvert(\text{cotangent vector on } S^2)\rvert^2$	$\frac{1}{2}$
rescaled geodesic	$G = \sqrt{2F} - 1$	0
Kepler	$H = \lvert y\rvert^{-1}G - \frac{1}{2}$	$-\frac{1}{2}$

The motivation for focusing on the energy level $H = -\frac{1}{2}$ is that this corresponds to the choice of the unit sphere, since $\varrho = \sqrt{-2H} = 1$. The term $|y|$ is just the familiar $1 - e\cos\gamma$, as Moser remarks in an appendix. It is a height coordinate on the sphere which equals zero at the north pole.

The trajectories of these Hamiltonians are related as follows:

- For any energy level, the Hamiltonian G has the same orbits as F but with speed scaled by a factor which is *constant* along each orbit. This is because G can be differentiated using the chain rule.

[30] Two remarks for clarity. First, the exponential of a differential operator is typically expressed using Green's functions. Second, we have stayed in the two-dimensional setting for consistency, but Fock worked in the three-dimensional setting. Thus, he obtained spherical harmonics on S^3, and these are rich enough to specify both the angular component (spherical harmonics on S^2) and the radial component of an eigenfunction of the hydrogen atom in $\mathbb{R}^3$.

[31] The paper is cited in footnote 25.

[32] Note that our notation in Appendix A does not match Moser's notation: his F is our G, and his H is our E. We will use Moser's notation here.

- For the energy hypersurface $G = 0$, the orbits of H are the same as G, but with speed scaled by the nonconstant factor $|y|^{-1} = \frac{1}{1-e\cos\gamma}$. This follows from differentiating H using the product rule.

  Moser observes that this rescaling of the speed corresponds to "changing the independent [time] variable s to $t = \int |y|\, ds$."

It follows that the composite map $\mathsf{stereo} \circ \mathsf{Fourier}$ induces a diffeomorphism of the (geodesic) $F = \frac{1}{2}$ energy hypersurface with the (Kepler) $H = -\frac{1}{2}$ energy hypersurface sending orbits to orbits, but not in a way which respects the time evolution. Moser applies this to showing that sufficiently small perturbations of the Kepler problem admit a certain number of periodic solutions. The reparameterization of time is not an issue because the application depends only on the topology of the (compactified) energy hypersurface.

So far, this only handles the energy hypersurface $H = -\frac{1}{2}$. To handle the other (negative) energy hypersurfaces, Moser observes that they can be mapped to the $H = -\frac{1}{2}$ hypersurface using certain uniform scalings of time, position, and velocity.[33] Therefore, all (negative) energy hypersurfaces have the same topology and description of orbits.

In closing, we remark that Moser's map $\mathsf{stereo} \circ \mathsf{Fourier}$ is not the same as the 'stereographic projection map' $(p, \mathcal{O}) \mapsto (s_{\mathsf{stereo}}, \Gamma, h)$ which we defined earlier in this appendix. Here are the essential differences:

- Moser's map preserves the symplectic form, but ours does not.
- Moser's map requires the choice of an energy level, but ours does not.
- Moser's map sends Kepler orbits at the chosen energy level to great circles, but Kepler orbits at other energy levels do not go to great circles. In contrast, our map sends all Kepler orbits to great circles.

[33] Wikipedia attributes this scaling transformation to Lie, see `https://en.wikipedia.org/wiki/Laplace%E2%80%93Runge%E2%80%93Lenz_vector#Lie_transformation`.