Is a 2000-Year-Old Formula Still Keeping Some Secrets?

Keith Kendig

A triangle is determined by its three sides, so there should be a formula for its area in terms of those sides. There is:

$$\text{Area} = \sqrt{s(s-a)(s-b)(s-c)}\,;$$

a, b, and c are the triangle's sides, and $s = (a+b+c)/2$ is its semi-perimeter. The oldest known proof of this formula dates back to Heron, who lived in Alexandria during the first century AD [**10**, p. 178]. Heron's Formula can also be written directly in terms of a, b, c:

$$\text{Area} = \frac{1}{4}\sqrt{(a+b+c)(-a+b+c)(a-b+c)(a+b-c)} \qquad (1)$$

or, multiplying out the four factors, as

$$\text{Area} = \frac{1}{4}\sqrt{2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4}\,. \qquad (2)$$

For example, $a = b = c = 1$ defines an equilateral triangle and the formula gives Area $= \sqrt{3}/4$. However $a = 3$, $b = c = 1$ does not define a triangle, and the formula produces $\frac{1}{4}\sqrt{-45}$ This would have made perfect sense to the ancients—there is no triangle, and there is no number.

Today we know more: $\frac{1}{4}\sqrt{-45} \approx 1.677i$ is indeed a number, and that creates a problem. Does it mean that $a = 3$, $b = c = 1$ really does define some triangle? The formula tantalizes us: for example, $a = 1.99$, $b = c = 1$ just barely defines a triangle; the formula correspondingly gives a very small real area, which becomes zero as a increases to 2. As a increases beyond 2, the sides at first just miss forming a triangle, and the formula produces small imaginary numbers; as a increases, the sides get further and further away from forming a triangle, and the imaginary numbers grow. It seems that the formula is tracking triangles we don't see, and is reporting their areas to us. What's going on? After two millennia, is Heron's Formula still keeping some secrets?

1. AN EXAMPLE. Heron's Formula does indeed track triangles. *Where do they go?* In Example 1, we follow an isosceles triangle as its two equal sides get shorter and shorter. What we learn suggests what happens more generally.

Example 1. Consider the triple of side lengths $\{2, r, r\}$. To construct the triangle when, for example, $r = 3$, choose a side (we use $a = 2$) and construct a circle of radius 3 centered at each of its endpoints. Either of the two points where the circles intersect determines a $\{2, 3, 3\}$-triangle. From the ancient Greek perspective, the thing that goes wrong with this construction applied to, say, $\{2, \frac{1}{2}, \frac{1}{2}\}$ is that the circles don't intersect. Today, however, we have something they didn't: coordi-

nate equations. They had nothing like $x^2 + y^2 = r^2$ to describe a circle—plane geometry was pretty much limited to the straightedge and compass. Today, we can simply write the equations for the two circles and solve for their points of intersection; the Fundamental Theorem of Algebra ensures that they have a solution. So for the triangle $\{2, r, r\}$, take two of its vertices to be $(-1, 0)$ and $(1, 0)$. Then a third vertex is found by solving the two circle equations

$$(x + 1)^2 + y^2 = r^2 \quad \text{and} \quad (x - 1)^2 + y^2 = r^2. \tag{3}$$

The solutions are $x = 0$, $y = \pm\sqrt{r^2 - 1}$, so the third vertex can be either $(0, \sqrt{r^2 - 1})$ or $(0, -\sqrt{r^2 - 1})$. Let's agree to write $x = x_1 + ix_2$, $y = y_1 + iy_2$. Then for $r > 1$, these vertices are on the y_1-axis; for $r = 1$, they've coalesced to $(0, 0)$, and the triangle's area is zero. As r decreases below 1, the coalesced vertices split up and move apart on the iy_2-axis. The triangles have thus moved from the usual Euclidean (x_1, y_1)-plane into the (x_1, iy_2)-plane, which is an example of two-dimensional "space-time"—that is, a Lorentz or Minkowski plane. The x_1-axis is space-like: $(x_1)^2 > 0$.The iy_2-axis is time-like: $(iy_2)^2 < 0$.

Being able to follow these vertices means that we can now draw triangles for any r with $0 \le r < 1$. Figure 1 shows one of the two triangles for $r = \frac{1}{2}$:

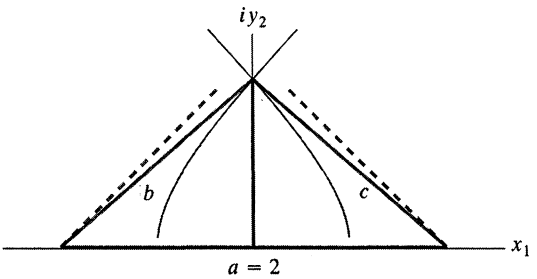

Figure 1. An isosceles triangle $a = 2$, $b = c = \frac{1}{2}$ in two-dimensional space-time. The top vertex is the intersection of two complex circles of radii $\frac{1}{2}$ centered at $x = \pm 1$. The restrictions of these circles to this Minkowski plane are hyperbolas; we see parts of them intersecting in the top vertex. Sides b and c have directions close to those of the dashed null lines; their metrics are weak compared to the horizontal and vertical directions, where they're the strongest. Though on Euclidean paper, b and c look like $\sqrt{7}/2 \approx 1.32$, they're really only $\frac{1}{2}$ in this Minkowski plane. All sides of the triangle are real and thus "space-like"; the altitude is imaginary and is "time-like". All areas in this (and any Minkowski) plane are pure imaginary.

It has base 2 and altitude $\sqrt{\left(\frac{1}{2}\right)^2 - 1} = \sqrt{3}\,i/2$. If the area formula $\frac{1}{2}$Base $\times$ Height is valid here, then the area should be $\sqrt{3}\,i/2$. Substituting $a = 2$, $b = c = \frac{1}{2}$ into Heron's Formula does indeed give $\sqrt{3}\,i/2$.

This provocative result raises some questions. For example, the sides $b = c = \frac{1}{2}$ certainly don't *look* $\frac{1}{2}$ long! Yet the Pythagorean Theorem says they are: $1^2 + (\sqrt{3}\,i/2)^2 = \left(\frac{1}{2}\right)^2$. Of course, the third vertex comes from circle equations, and the equation of a circle is basically a restatement of the Pythagorean Theorem. Very importantly, so also is the definition of "metric". Heron's Formula seems to suggest that if we use the Pythagorean Theorem to define a metric, we get consistent results.

2. A BILINEAR FORM ON $\mathbb{C}^2$. We found the area of the triangle $\{2, \frac{1}{2}, \frac{1}{2}\}$ two ways: using Heron's Formula, and using $\frac{1}{2}Bh$. Were we just lucky, or can we really measure triangles and, perhaps, more generally do analytic geometry in the planes

they move into? The answer to both is yes; the reason rests on the fact that $\mathbb{C}^2$ has a natural symmetric bilinear form on it: for points (x, y) and (x', y') in $\mathbb{C}^2$, let $(x, y) \cdot (x', y') = xx' + yy'$. This scalar product of course defines the quadratic form $(x, y) \cdot (x, y) = x^2 + y^2$. When this is real, we call it the *separation-squared* of (x, y)—intuitively, it's the distance-squared from the origin. $\sqrt{(b-a) \cdot (b-a)}$ is then the separation between points a and b in $\mathbb{C}^2$. This "distance" has a different appearance from the usual hermitian one: for example, setting our quadratic form equal to 1 gives, for the real and imaginary parts of $x^2 + y^2 = 1$,

$$x_1^2 - y_1^2 + x_2^2 - y_2^2 = 1 \quad \text{and} \quad x_1 y_1 + x_2 y_2 = 0.$$

This defines a real, two-dimensional complex circle in $\mathbb{C}^2$, which, to our Euclidean eyes, is unbounded [**5**, Chapter I]. In contrast, the hermitian counterpart $x\bar{x} + y\bar{y} = 1$ has real and imaginary parts

$$x_1^2 + y_1^2 + x_2^2 + y_2^2 = 1 \quad \text{and} \quad 0 = 0,$$

and this defines a real, three-dimensional sphere $S^3 \subset \mathbb{R}^4$.

Our bilinear form has the advantage that it leads to the Pythagorean Theorem and the Law of Cosines [**1**, p. 112] and this, together with the usual field properties, means we can do analytic geometry in $\mathbb{C}^2$. In fact, because of the "persistence of functional identities", any purely algebraic argument in $\mathbb{R}^2$, being a chain of algebraic expressions, remains valid when $\mathbb{R}^2$ is extended to $\mathbb{C}^2$. From $\mathbb{C}^2$, one can then restrict to various other subplanes and get valid theorems there. For example, a plane on which the quadratic form is real is Euclidean, Minkowski or "anti-Euclidean" when the quadratic form there is positive definite, indefinite, or negative definite, respectively, and we get theorems valid in all these planes. (All such planes in $\mathbb{C}^2$ are depicted in Figure 3.) As an application, the following proof of Heron's Formula is purely algebraic, so each side of the formula has meaning and is valid in $\mathbb{R}^2$ (where it was designed), in $\mathbb{C}^2$, and, of importance to us, in any Euclidean, Minkowski, or anti-Euclidean plane in $\mathbb{C}^2$.

Proof of Heron's Formula. Let the vertices of a triangle be (a_1, a_2), (b_1, b_2), (c_1, c_2), and let the respective sides opposite them be a, b, c. Then

$$a^2 = (b_1 - c_1)^2 + (b_2 - c_2)^2$$

$$b^2 = (a_1 - c_1)^2 + (a_2 - c_2)^2$$

$$c^2 = (a_1 - b_1)^2 + (a_2 - b_2)^2.$$

After solving for the altitude's base point on side a, one can apply the Pythagorean distance formula to get that the altitude-squared is

$$h_a^2 = \frac{(2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4)}{4a^2}. \tag{4}$$

Substituting (4) into $(\text{Area})^2 = \left(\frac{1}{2}ah_a\right)^2$ gives

$$(\text{Area})^2 = [2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4]/16. \tag{5}$$

This is just the square of Heron's Formula in (2). ∎

In this proof, choosing a as base is of course arbitrary; this is reflected in the fact that (5) is symmetric in a, b, and c.

3. BACK TO EXAMPLE 1. Though we can formally do analytic geometry in all the cases mentioned in Section 2, things of course can look different there, so it's instructive to work out examples in concrete cases. Let's therefore continue exploring the triangles in Example 1. For instance, notice that the altitude $(0, \sqrt{3}\,i/2)$ and the base $(2, 0)$ have scalar product zero. For each of the other bases, we can find the corresponding altitude by requiring it to have zero scalar product with the base, and then check that Area $= \frac{1}{2}$Base $\times$ Height holds in each case, as guaranteed by our proof of Heron's Formula. For example, the equation of the line containing side b is $2y_2 = \sqrt{3}\,i(x_1 + 1)$, and the altitude line through $(1, 0)$ is $\sqrt{3}\,iy_2 = -2(x_1 - 1)$. They intersect at their common solution, which is the point $(7, 4\sqrt{3}\,i)$. The Pythagorean Theorem then gives the altitude to side b: $\sqrt{(7-1)^2 + (4\sqrt{3}\,i)^2} = 2\sqrt{3}\,i$. Therefore, as expected, $\frac{1}{2}$Base $\times$ Height $= \frac{1}{2} \cdot \frac{1}{2} \cdot 2\sqrt{3}\,i = \sqrt{3}\,i/2 =$ Area.

The Law of Cosines is useful, too. It provides, for example, a way to find each angle of a triangle of known sides. Here's how this works in Figure 1: let α, β, γ be the angles opposite a, b, c. To get α, for instance, apply the Law of Cosines to evaluate $\cos\alpha$: $a^2 = b^2 + c^2 - 2bc \cdot \cos\alpha$ becomes $2^2 = \left(\frac{1}{2}^2\right) + \left(\frac{1}{2}\right)^2 - 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \cos\alpha$, giving $\cos\alpha = -7$. This defines two candidates for α within the period strip $0 \le Re(\alpha) < 2\pi$; with a little help from Mathematica or Maple, we find them to be $\pi \pm (2.6339\ldots)i$. To narrow this to a single candidate, obtain $\tan\alpha$, working from Figure 1; it turns out to be $-4\sqrt{3}\,i/7$, and this produces the two values $-(2.6339\ldots)i$ and $\pi - (2.6339\ldots)i$. Our angle is the single common value, $\alpha = \pi - (2.6339\ldots)i$. A similar approach gives $\beta = \gamma = (1.3169\ldots)i$. Note that $\alpha + \beta + \gamma = \pi$.

There are some other interesting features of Figure 1. The two arcs through the vertex $(0, \sqrt{3}\,i/2)$ come from the two equations in (3). These equations, when restricted to the Minkowski plane of Figure 1, are

$$(x_1 + 1)^2 - y_2^2 = r^2 \quad \text{and} \quad (x_1 - 1)^2 - y_2^2 = r^2.$$

We can put this into perspective: each equation in (3) actually defines a complex circle in $\mathbb{C}^2$ (which is topologically a real 2-manifold [**5**, Chapter I]). The real and imaginary parts of the first equation in (3) are

$$(x_1 + 1)^2 - x_2^2 + y_1^2 - y_2^2 = r^2 \quad \text{and} \quad (x_1 + 1)x_2 + y_1 y_2 = 0;$$

setting $x_2 = 0$ in these implies $y_1 y_2 = 0$, that is, $y_2 = 0$ or $y_1 = 0$. When $y_2 = 0$, the part of the complex circle in (x_1, y_1, iy_2)-space lies entirely in the (x_1, y_1)-plane; when $y_1 = 0$, the part lies in the (x_1, iy_2)-plane. In the first of these planes we see a real circle, and in the second, a real hyperbola; both appear lightly drawn in Figure 2. The pair defined by the second equation in (3) is drawn there more heavily. The top, middle, and bottom sketches correspond to $r > 1$, $r = 1$, and $r < 1$; for clarity, we have not drawn the y_1-axis. The two intersecting branches in the bottom sketch extend the two arcs appearing in Figure 1; see [**6**] or [**7**]. For a more detailed account and further examples, see [**5**, Chapter I].

Figure 2 shows what happens to the points of intersection as r passes through $r = 1$. In the top sketch ($r > 1$), the circles in the (x_1, y_1)-plane intersect in two distinct points, and the two hyperbolas in the (x_1, iy_2)-plane don't intersect. In the middle sketch, r has decreased to $r = 1$; the two circles intersect tangentially at the origin, and so do the two hyperbolas. In the bottom sketch, $r < 1$; the circles have separated and it is now the hyperbolas that intersect in two distinct points.

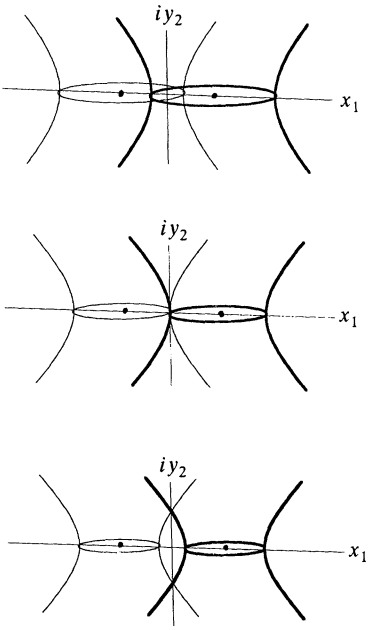

Figure 2. In each sketch, the light circle/hyperbola is the restriction of a complex circle centered at the left dot $x_1 = -1$; the heavily-drawn circle/hyperbola is centered at the right dot $x_1 = +1$. From top to bottom, the circles' radii are decreasing.

Going from top to bottom in the figure, we see that the two intersection points coalesce along the y_1-axis, then make an abrupt turn, separating along the iy_2-axis. Since the other two triangle vertices are fixed at $x_1 = \pm 1$, one can visualize the triangles as they start in the Euclidean plane, collapse to a line segment, and emerge in the Minkowski plane.

4. A TORUS OF 1-SPACES. The altitude in Figure 1 cuts the triangle into two congruent pieces. The smaller pieces introduce a new feature: an imaginary side. Substituting the new sides into Heron's Formula nevertheless correctly gives half the area of the original triangle, as our proof of the formula ensures. Actually, the (x_1, iy_2)-plane is quite rich, and one can find non-degenerate triangles with 3, 2, 1, or 0 real sides. This is easy to do: since the separation from the origin to (x_1, iy_2) is $s = \sqrt{x_1^2 - y_2^2}$, s is real when $|y_2/x_1| < 1$, zero when $|y_2/x_1| = 1$, and imaginary when $|y_2/x_1| > 1$. The line segments are respectively space-like, null, and time-like. Since separation is invariant under translation and since any three distinct slopes can be used to define a triangle, we can choose appropriate slopes to design a triangle having any of the four desired combinations of real and imaginary separations.

This raises an interesting question: by shortening two sides of a triangle in the (x_1, y_1)-plane (which is Euclidean), we gained entrance to the (x_1, iy_2)-plane (which is Minkowski), and the triangles that greeted us there still had three real sides. Since there are triangles in the (x_1, iy_2)-plane having three *imaginary* sides, does this Minkowski plane have, symmetrically, a back door? If it does, where does it lead? If one reflects the triangle $\{2, \frac{1}{2}, \frac{1}{2}\}$ in Figure 1 about a null line there, the reflected triangle has three imaginary sides, $\{2i, \frac{1}{2}i, \frac{1}{2}i\}$. If we increase the short sides $r = \frac{1}{2}i$ to $r > i$, the scenario mimics Figure 2, going from bottom to top. One

can easily check that as the triangle collapses and we go out this back door, we enter the (ix_2, iy_2)-plane. This "anti-Euclidean" plane is a negative-definite version of the ordinary Euclidean (x_1, y_1)-plane: our bilinear form on $\mathbb{C}^2$ reduces in this anti-Euclidean plane to $(ix_2, iy_2) \cdot (ix_2, iy_2) = -x_2^2 - y_2^2$. The (x_1, y_1)-plane is isomorphic to the (ix_2, iy_2)-plane; if one places one on top of the other, any separation in the anti-Euclidean plane is i times the corresponding separation in the Euclidean plane.

To keep things simple, we've mostly considered isosceles triangles. In general, one can increase or decrease the separation-squared of any side(s) and this, together with congruence, gives triangles considerable freedom to wander throughout various 2-spaces in $\mathbb{C}^2$. The reader may by now be wondering just how many different spaces triangles can travel into. There's a simple answer, and it's encapsulated in the torus of Figure 3. The idea is this: lines in $\mathbb{C}^2$ on which s^2 is real are of

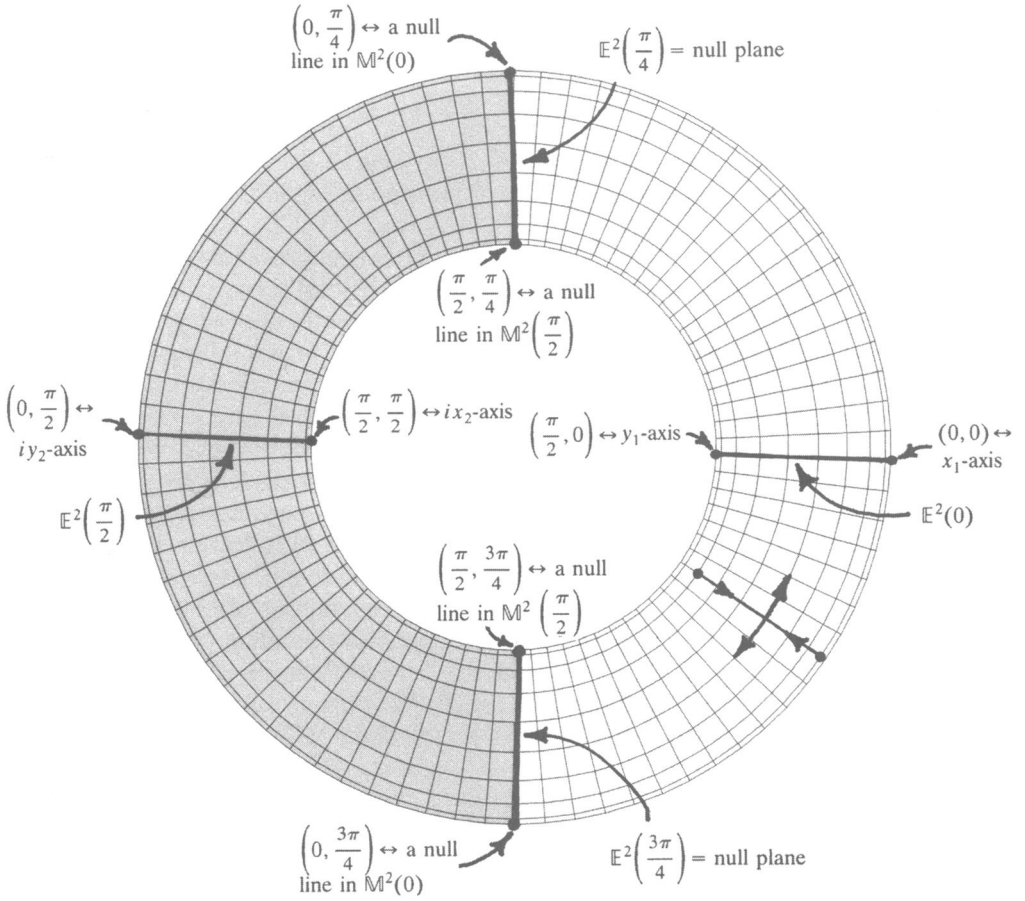

Figure 3. Each point of the torus represents a 1-space consisting of all real scalar multiples of a point in $\mathbb{C}^2$. The points in the right half of the torus represent space-like 1-spaces; the points in the left half represent time-like 1-spaces; the points on the two meridian circles between the halves represent the null lines. The torus parallels appear here as concentric circles; they are all the different Minkowski 2-spaces in $\mathbb{C}^2$, and are denoted by $\mathbb{M}^2(\phi)$. The meridian circles orthogonal to them appear edge-on, and are denoted by $\mathbb{E}^2(\theta)$. The plane corresponding to any particular meridian circle $\theta = \theta_0$ is "isotropic"—the metric is independent of direction. As θ varies from 0 to π, however, those isotropic planes vary smoothly from Euclidean to one null plane, to anti-Euclidean, to the other null plane, and back to Euclidean.

three kinds: $s^2 > 0$ (space-like), $s^2 = 0$ (null), and $s^2 < 0$ (time-like). We'll set up a parametrization so that the points in the right half of the torus represent the space-like lines; the points in the left half of the torus represent the time-like lines; and the points on the two circles between these halves represent the null lines. The torus, as with any surface of revolution, has parallels and meridians. The parallels are the circles on the surface that are perpendicular to the axis of revolution, and the meridians are the curves on the surface that lie on planes containing the axis of revolution. On the torus, the parallels and meridians are just the orthogonal sets of coordinate circles. We'll see that *the planes in which our triangles live are represented by precisely the parallels and meridians of the torus*. Some specific lines and planes that we have met are indicated in Figure 3.

Before turning to specifics, it's hard to resist making a sales pitch: this torus not only answers an immediate question about triangles, but spotlights an often-neglected side of $\mathbb{C}^2$. The phrase "planes in $\mathbb{C}^2$" usually brings to mind complex 1-spaces, and this already points us in a direction that, when taken far enough, leads to matrix (quantum) mechanics. But there are other planes in $\mathbb{C}^2$ that connect with another part of the real world: space, time, and space-time. This, taken far enough, puts us in special relativity. (For example, rotations in a Minkowski plane are Lorentz contractions.) It is amazing that $\mathbb{C}^2$ should contain the seeds of these two great physical theories. For some decades, the outstanding problem in physics has been to unify them.

Let's start by parametrizing the space-like lines. Since separation is unchanged by translation, we assume that our lines pass through the origin. To parametrize these 1-spaces of $\mathbb{C}^2$, we mimic their familiar parametrization in the (x_1, y_1)-plane. There, the unit circle $x_1^2 + y_1^2 = 1$ is parametrized by $x_1 = \cos\phi$, $y_1 = \sin\phi$ $(0 \le \phi < 2\pi)$. Any 1-space intersects this circle in two diametrically opposite points, so ϕ parametrizes the 1-spaces when ϕ is restricted to $0 \le \phi < \pi$. More generally, in $\mathbb{C}^2$, each point of any space-like 1-space in $\mathbb{C}^2$ has positive separation-squared. Intersecting the 1-space with the complex circle $x^2 + y^2 = 1$ $(x = x_1 + ix_2, y = y_1 + iy_2)$ selects those two points on the 1-space having separation-squared equal to 1. For more on complex circles, see [**5**, Chapter I]. This complex circle is parametrized by $x = \cos(\phi + i\vartheta)$, $y = \sin(\phi + i\vartheta)$ $(0 \le \phi < 2\pi, -\infty < \vartheta < +\infty)$. Restricting ϕ to $0 \le \phi < \pi$ then parametrizes the space-like lines of $\mathbb{C}^2$. These lines can therefore be identified with the points in the infinite vertical strip $0 \le \phi < \pi$ in $\mathbb{C}^1$. We can shrink the vertical dimension so the strip becomes a rectangle; a little algebra shows that the change of variable $\vartheta = \operatorname{arctanh}(\tan\theta)$ accomplishes this. In the variables ϕ, θ the parametrization then reads

$$\mathbf{V}(\phi, \theta) = (\cos\phi\cos\theta - i\sin\phi\sin\theta, \sin\phi\cos\theta + i\cos\phi\sin\theta). \qquad (6)$$

By choosing $0 \le \phi < \pi$, $0 \le \theta < \pi$, (6) parametrizes the space-like lines $(0 \le \theta < \pi/4, 3\pi/4 < \theta < \pi)$, the time-like lines $(\pi/4 < \theta < 3\pi/4)$, and the two null lines $(\theta = \pm\pi/4)$. One can now paste together opposite sides of the square to obtain the torus in Figure 3. The parallels appear in Figure 3 as concentric circles; note that we go once around a parallel as θ increases from 0 to only π (rather than 2π). In Figure 3, the meridian circles are of course viewed edge-on, and once again we go once around any one of these circles as ϕ increases from 0 to π.

5. THE 2-SPACES. Since the torus accounts for all the space-like, null, and time-like 1-spaces of $\mathbb{C}^2$, any Euclidean, null, anti-Euclidean, or Minkowski 2-space in $\mathbb{C}^2$ must correspond to a subset of our torus. Since *any* real 2-space of

$\mathbb{C}^2$ is determined by two distinct real 1-spaces of $\mathbb{C}^2$, two distinct points on the torus thus define a 2-space of $\mathbb{C}^2$. Let's choose two points in our torus coming from, say, points $\phi + i\theta$ and $\phi' + i\theta'$ in the square. These two points correspond to two vectors in $\mathbb{C}^2$; let an arbitrary linear combination of them be $\mathbf{W} = c_1\mathbf{V}(\phi, \theta) + c_2\mathbf{U}(\phi', \theta')$. A little algebraic manipulation shows that the imaginary part of $\mathbf{W} \cdot \mathbf{W}$ is $2c_1c_2 \cdot \sin(\phi - \phi') \cdot \sin(\theta - \theta')$; this must be zero if $\mathbf{W}$ is to have real separation-squared. Therefore in our square, $\phi = \phi'$ or $\theta = \theta'$, meaning that the two points must lie on a coordinate line there—that is, on a coordinate circle of the torus. All the points on any one coordinate circle correspond to 1-spaces in a 2-space of $\mathbb{C}^2$. The coordinate circles corresponding to $\phi = \phi'$ are the parallels of the torus; we denote the corresponding 2-spaces by $\mathbb{E}^2(\theta)$. The coordinate circles corresponding to $\theta = \theta'$ are the meridians; we denote the corresponding 2-spaces by $\mathbb{M}^2(\phi)$.

If we extend θ and ϕ to the period rectangle $0 \le \phi < 2\pi$, $0 \le \theta < 2\pi$, then for fixed θ_0 or ϕ_0, the images $\mathbf{V}(\phi_0, \theta)$ and $\mathbf{V}(\phi, \theta_0)$ are genuine circles in $\mathbb{R}^4$; this follows from the fact that if $\mathbf{v}$ and $\mathbf{w}$ are orthonormal vectors in $\mathbb{R}^n$, then

$$\mathbf{v} \cdot \cos(t) + \mathbf{w} \cdot \sin(t) \quad (0 \le t < 2\pi) \tag{7}$$

is a unit circle there. In our case, the vectors $\mathbf{v} = (\cos\phi_0, 0, \sin\phi_0, 0)$ and $\mathbf{w} = (0, -\sin\phi_0, 0, \cos\phi_0)$ are orthonormal in $\mathbb{R}^4$, so our parametrization (6) extends to an instance of (7), and each $\mathbf{V}(\phi_0, \theta)$ is a unit circle. Choosing $\mathbf{v} = (\cos\theta_0, 0, 0, \sin\theta_0)$, and $\mathbf{w} = (0, -\sin\theta_0, \cos\theta_0, 0)$ similarly shows each $\mathbf{V}(\phi, \theta_0)$ is a unit circle. *Note that all these coordinate circles are centered at the origin*! This seems at odds with our everyday view of a torus, but the natural environment for the product of two circles is $\mathbb{R}^4$. We're lucky to be able to see a torus in $\mathbb{R}^3$ as well as we do. Our "normal" view, however, is actually a contorted one. (The necessary contortion is even worse for its non-orientable counterpart, the Klein bottle, where one has to introduce self-intersections to make it fit in $\mathbb{R}^3$.) From the torus' perspective, its most natural portrait is as the image of $\mathbf{V}(\phi, \theta)$. Because the hermitian inner product $\mathbf{V}(\phi, \theta) \cdot \overline{\mathbf{V}(\phi, \theta)}$ is 1, this image is contained in the real unit sphere $S^3 \subset \mathbb{R}^4$.

A little algebra reveals that our (non-hermitian) separation-squared of a typical point $\mathbf{V}(\phi, \theta)$ in the torus is $\mathbf{V}(\phi, \theta) \cdot \mathbf{V}(\phi, \theta) = \cos(2\theta)$. Obligingly, this is independent of ϕ, so for any θ, all the points on the unit circle in $\mathbb{E}^2(\theta)$ have the same separation-squared from the origin of $\mathbb{E}^2(\theta)$. So although all the circles $\mathbf{V}(\phi, \theta_0)$ look like unit circles to our Euclidean eyes, with respect to the metric defined by our bilinear form, the "unit circle" in $\mathbb{E}^2(\theta)$ has radius-squared $\cos(2\theta)$. This is indeed its maximum of 1 when $\theta = 0$, but for $\theta = 0.1$, say, it's only 0.98. It is natural to describe this visual disparity by saying that the metric is strongest, or most concentrated, on the plane $\mathbb{E}^2(0)$, and a little weaker (or less concentrated) on the plane $\mathbb{E}^2(0.1)$. More generally, for $-\pi/4 < \theta < \pi/4$, each $\mathbb{E}^2(\theta)$ is Euclidean, but the strength of the metric decreases symmetrically to 0 at $\pm\pi/4$. The planes $\mathbb{E}^2(\pm\pi/4)$ are null—in either plane, the separation between any two points is 0. As θ continues to grow in magnitude, the planes become anti-Euclidean, their negative-definite metrics symmetrically becoming stronger or more concentrated, up to $\theta = \pm\pi/2$.

Notice in Figure 3 that any particular parallel $\mathbb{M}^2(\phi)$ intersects each meridian $\mathbb{E}^2(\theta)$ in one point. Geometrically, this means that $\mathbb{M}^2(\phi)$ is the union of 1-subspaces, one from each $\mathbb{E}^2(\theta)$. Since the metric in $\mathbb{E}^2(\theta)$ varies with θ, so does the metric in $\mathbb{M}^2(\phi)$; θ parametrizes the 1-subspaces in any $\mathbb{M}^2(\phi)$. This fits in with a familiar illustration one often sees in expositions of space-time: the points of

separation-squared 1 lie on the two hyperbolas $\pm x^2 \mp y^2 = 1$. The ordinary Euclidean distance from the origin to the hyperbola is great in directions near that of a null line, because the metric is weak in those directions. In a nutshell, all the $\mathbb{M}^2(\phi)$'s are metrically identical, and are "non-isotropic"; each $\mathbb{E}^2(\theta)$ is isotropic, but its metric strength varies with θ.

Some additional facts can be read from Figure 3. The rightmost meridian $\mathbb{E}^2(0)$ is the (x_1, y_1)-plane; the two antipodal points shown on it correspond to its x_1- and y_1-axes. The leftmost meridian $\mathbb{E}^2(\pi/2)$ is the (ix_2, iy_2)-plane, which is anti-Euclidean; the two antipodal points shown on it correspond to its ix_2- and iy_2-axes. Every torus parallel $\mathbb{M}^2(\phi)$ intersects each of the two null meridians in one point; these two points correspond to the two null lines through the origin in $\mathbb{M}^2(\phi)$. (One can call the two null lines through any point in a Minkowski plane, the "light cone" through that point.) Figure 3 also shows three points on a meridian $\theta =$ constant; these three points correspond to a triangle in the Euclidean plane $\mathbb{E}^2(\theta)$. The two arrows pointing toward each other indicate two triangle sides approaching a third side. After coalescing, the points separate along a parallel $\phi =$ constant; the triangle still has three real sides, but it is now in $\mathbb{M}^2(\phi)$. This plus-shaped picture can be translated around on the torus (and one can also reverse the arrow directions) to depict various other scenarios of a triangle disappearing from one space and reappearing in another.

6. THE SPHERE. The torus of Section 5 provides an overview of the various spaces in which our triangles live. There is also a sphere that gives an overview of the triangles themselves, and we can use it to read off further basic information. For example, where should one look to find all the triangles of real area? zero area? imaginary area? maximum area? minimum area? And although in Example 1 we followed triangles from a Euclidean plane into a Minkowski plane, we did so along one particular path. What about along other paths? Is there a path leading directly from a Euclidean to an anti-Euclidean plane, without going through Minkowski space? The sphere model makes answers to this and other questions obvious.

Our model is essentially a parametrization of the triangles, but we don't parametrize them directly. This is because point pairs typically make an abrupt "turn" as they go from real to pure imaginary separation, as Figure 2 illustrates. This behavior disappears if we use separation-squared instead of the separation itself; the critical transition is then simply from positive to negative instead of from real to pure imaginary.

We therefore begin by squaring both sides of (2), giving

$$(\text{Area})^2 = [2a^2b^2 + 2a^2c^2 + 2b^2c^2 - a^4 - b^4 - c^4]/16. \tag{8}$$

Fortuitously, this involves only the squares of separations; let's rename these squares:

$$u = a^2, \qquad v = b^2, \qquad w = c^2.$$

The bracketed expression in (8) now becomes a quadratic form:

$$2uv + 2uw + 2vw - u^2 - v^2 - w^2. \tag{9}$$

Setting this equal to a constant gives all triangles having a particular $(\text{Area})^2$, and the equation defines a quadric surface in (u, v, w)-space. We can sketch this

surface using standard eigen-arguments: in matrix form (9) is

$$(u \quad v \quad w)\begin{pmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}\begin{pmatrix} u \\ v \\ w \end{pmatrix}. \tag{10}$$

Since the 3×3 matrix in (10) is real-symmetric, it has real eigenvalues and three mutually orthogonal real eigenvectors. The eigenvalues are $\lambda = 1, -2, -2$. An eigenvector associated with $\lambda = 1$ is $(1, 1, 1)$; two eigenvectors for $\lambda = -2$ can be chosen in the two-space orthogonal to the vector $(1, 1, 1)$—that is, in the plane $u + v + w = 0$. Now generally, for any quadratic form $Q = Q(u, v, w)$, the surface $Q = 1$ intersects a principal axis (that is, the set of real scalar multiples of an eigenvector) if and only if that axis' eigenvalue is positive. We can apply this to our situation: if A denotes Area, then for $A^2 > 0$, the surface $Q = A^2$ is a hyperboloid of two sheets; it intersects the axis through $(1, 1, 1)$. If $A^2 < 0$, then $Q = A^2$ is a hyperboloid of one sheet and intersects the plane $u + v + w = 0$ in a circle. $A^2 = 0$ defines a circular cone whose axis of symmetry is the 1-space through $(1, 1, 1)$. Therefore as A^2 takes on all real values, we get a family of hyperboloids, with the cone $A^2 = 0$ separating the one-sheeted from the two-sheeted. This is shown in Figure 4:

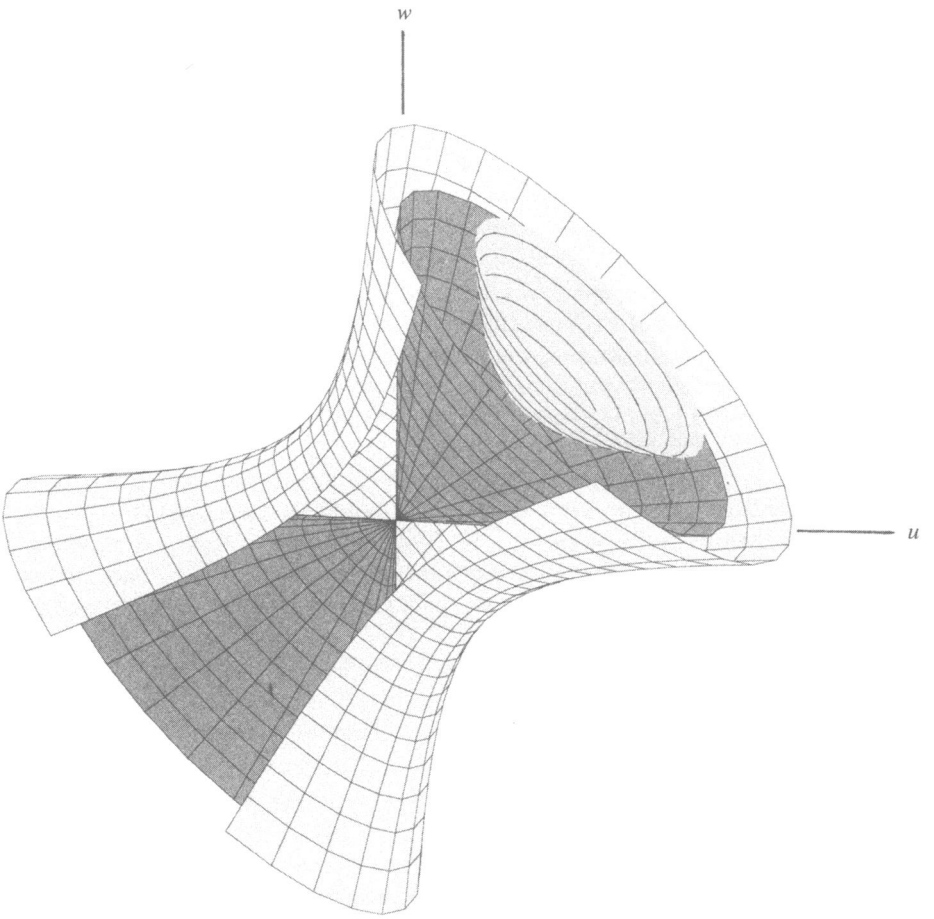

Figure 4

The cone is very natural: besides being symmetric with respect to the 1-space through $(1, 1, 1)$, it is tangent to each of the three coordinate planes. (This is because the cone is tangent to a plane if and only if it intersects it in a single line. For example, the (u, v)-plane has equation $w = 0$; when substituted into the cone's equation, this gives $u^2 - 2uv + v^2 = 0$, that is, $(u - v)^2 = 0$. Therefore the cone intersects the plane in the line $\{u = v, w = 0\}$. Similarly for the other two coordinate planes.) The repeated eigenvalue implies that all these surfaces are surfaces of revolution about the 1-space through $(1, 1, 1)$; their intersections with planes orthogonal to this line are therefore circles.

Since any point (u, v, w) represents a triangle of separations-squared u, v, and w, the nonzero points of the ray through a particular (u, v, w) represent, in the obvious sense, triangles that are all mutually similar. We may make this representation more efficient by choosing just one point on each ray—say, the one intersecting the unit sphere centered at the origin in (u, v, w)-space. The various hyperboloids intersect the sphere in a family of "circles of latitude"; every point on the sphere is on some latitude, and the points on any latitude represent triangles all having the same area. Figure 5 depicts the situation.

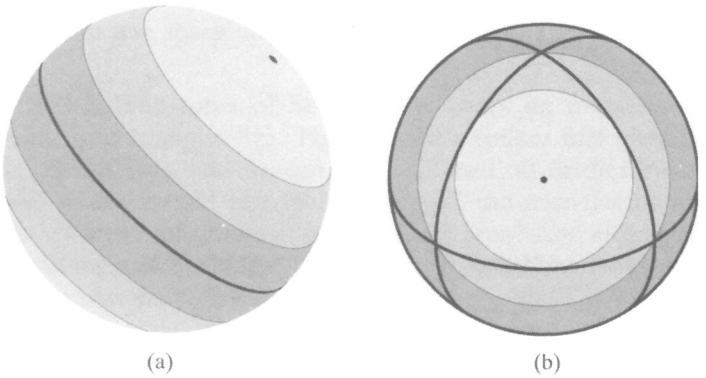

Figure 5. View (a) shows naturally-defined zones on the parameter sphere. The polar cap represents all the triangles the ancients ever knew, occupying less than 10% of the sphere's surface! View (b) is a polar view of the parameter sphere, showing zone boundaries and basic spherical triangles. There is an interesting interplay between them.

The heavily-drawn single point (the "north pole") in each picture corresponds to the ray of symmetry through $(1, 1, 1)$. Since $u = v = w$ there, it represents a real, equilateral triangle. The polar cap boundaries (the "arctic" and "antarctic" circles) are where the cone intersects the sphere; these circles therefore represent all triangles of zero area. Now each of the three coordinate planes in our (u, v, w)-space intersects the sphere in a great circle. The northern halves of them appear, heavily drawn, in Figure 5b. These great circles divide the sphere into eight congruent equilateral spherical triangles, corresponding to the eight octants of (u, v, w)-space . In Figure 5b we see the entire "north triangle" and partial views of six others. The north triangle corresponds to the $(+ + +)$-octant; points in the north triangle therefore correspond to triangles with three real sides. Points in the diametrically opposite "south triangle" correspond to the $(- - -)$-octant and, correspondingly, triangles there have three imaginary sides.

There is an interesting interplay between these spherical triangles and the sphere's zones, shaded light, medium, and dark gray in Figure 5. Since all triangles

in the northern hemisphere that have real area occur in the polar cap, and since the polar cap doesn't occupy the entire north triangle, it is natural to ask, what do the left-over points represent? They're just the triangles in a Minkowski space with three real sides. If, for example, we start at the north pole and travel due south toward a vertex of the north triangle; we essentially replay the isosceles triangle scenario of Example 1—as we cross the arctic circle, we enter a Minkowski space, but all sides of the triangle are still real. Notice that every time we cross a great circle, the number of real sides changes by 1; that corresponds to leaving one octant of (u, v, w)-space and entering another one. Approaching the vertex on our journey means that the separations of the two changing sides are getting smaller and smaller; they simultaneously reach zero separation at the vertex. (This says that the sides lie on the null lines of a Minkowski space.) When we pass through the vertex, we cross two great circles; the number of real sides thus decreases by two, so the isosceles triangle now consists of one real and two imaginary sides. In some $\mathbb{M}^2(\theta)$, these two sides are time-like. Continuing our trip south, we approach a side of the south triangle. Any corresponding triangle has two long imaginary sides; its base is real. Since we're on the unit sphere in (u, v, w)-space, the representative triangles are scaled so that $u^2 + v^2 + w^2 = 1$, which means the base is becoming smaller. It goes to zero separation when we meet a side of the south triangle. After crossing it, the base becomes imaginary so all three sides of the isosceles triangle are imaginary.

Let's take another stroll south from the north pole, this time observing some of the landscape at various latitudes On the polar latitudes, we of course see only triangles with three real sides. But on a circle in the north temperate (medium gray) zone just below the arctic circle, we see the first examples of triangles with one imaginary side and two real sides. Their proportion increases as we continue south; at the boundary with the equatorial zone, they all have one imaginary side and two real sides. As we first enter the equatorial (dark gray) zone, we encounter triangles with two imaginary sides. The proportion of these increases as we continue south, reaching equal parts at the equator itself. By the time we reach the boundary between the equatorial and the south-temperate zones, all triangles have two imaginary sides. Crossing this boundary leads to our first glimpse of triangles with three imaginary sides—at first rare, and rising to the entire triangle population on the antarctic circle. Inside this circle, all triangles have three imaginary sides.

Here are a few additional facts about the parametric sphere. Consider the points

$$\pm(1,1,1)/\sqrt{3}\,, \quad \pm(1,1,0)/\sqrt{2} \quad \pm(1,0,0)/\sqrt{1}\,, \quad \pm(1,-1,0)/\sqrt{2}\,. \tag{11}$$

The first $\pm$-pair are the north and south poles, the other three pairs are points on the polar, temperate/equatorial, and equatorial circles, respectively. Substituting these into (9) gives values for A^2 of $1/16$ times:

$+1$ at the poles. This is the maximum value of A^2;
0 on the arctic and antarctic circles;
-1 on the temperate/equatorial circles;
-2 on the equator. This is the minimum value of A^2.

The inner product of each of the vectors in (11) with the north pole vector $(1,1,1)/\sqrt{3}$ gives the projection on the north-south axis; these are, respectively: ± 1, $\pm 2/\sqrt{6}$, $\pm 1/\sqrt{3}$, 0. Now the area of a spherical zone varies linearly with its

altitude (height measured along the axis of symmetry); from these facts, one can check that the north polar cap occupies $\frac{1}{2}(1 - 2/\sqrt{6}) \approx 9.17\%$ of the sphere's total surface area; the northern temperate zone occupies $\frac{1}{2}(2/\sqrt{6} - 1/\sqrt{3}) \approx 11.94\%$, and the northern equatorial zone $\frac{1}{2}(1/\sqrt{3}) \approx 28.87\%$. The southern figures are of course the same.

What has Heron's Formula taught us? A physicist once described empty space as a wild party in a large building; we're on the outside, and are able to see only an occasional party hat being tossed out of a window. In a sense, Heron's Formula provides a way to pry into a similar geometric party. How successful have we been? As measured on the sphere, our view has broadened from the polar cap to the entire sphere; as measured on the torus, it has broadened from one meridian to the whole torus. In addition, the torus, which arose from following where triangles go, not only parametrizes an important family of 2-spaces in $\mathbb{C}^2$—generally overshadowed by the complex subspaces of $\mathbb{C}^2$—but it also puts space-time in perspective, as part of the story of looking at all planes in $\mathbb{C}^2$ having real separation-squared.

6. HERON OF ALEXANDRIA: A SKETCH. Heron of Alexandria lived in Egypt during a long period of prosperity and technological progress. Ancient Alexandria was a coastal, walled metropolis of some one million people, with lavish parks and buildings laid out in a large grid. From the founding of the Alexandrian school of engineering around 250 BC to Heron's time in the first century AD, Alexandria had evolved to everyday use of pumps, pulleys, and all sorts of pneumatic and geared devices—even odometers on vehicles were common. Many devices were powered by steam, water, or compressed air. During annual religious parades, awestruck citizens saw mechanical doves rise into the air and slowly descend, buoyed up by invisible steam. In public gardens, water pressure powered moving statues and impressive fountains [**8**, p. 62].

Heron thrived in this milieu. He was an applied mathematician and a highly ingenious engineer who helped contribute to Egypt's technological advancement. He stood out from other engineers of the day in that he wrote voluminously—some 14 books. They were very practical in spirit, often featuring worked-out numerical examples; the reader would then substitute in his own numbers. Heron's books became very much in vogue; he was doubtless helped along because papyrus growing on the Nile was cheap and plentiful, and Alexandria became the book-copying center of the world [**8**, p. 63] He worked and wrote extensively on a broad range of applied subjects and was especially interested in mensuration of all sorts. He wrote an ambitious three-volume work (*Metrica*), which provided an encyclopedic set of examples illustrating how to calculate a wide range of plane and surface areas, as well as volumes. He invented the "diopter", a forerunner of today's surveyor's transit, and wrote a two-volume work (*Dioptra*) about it, and about surveying in general. In this work he shows how to determine directions to dig through a mountain so that workers will meet in the middle [**9**, p. 50]. He made several important improvements to water clocks, and wrote a four-book work on them, of which only a few scattered fragments survive. Heron's Formula appears and is proved in both *Dioptra* and *Metrica*. The formula is actually due to Archimedes some four centuries before him, but Heron's is the oldest extant proof we have, so his name is attached to it; see [**4**, volume ii, p. 322] and [**2**, p. 172]. There is a sketch of Heron's proof in [**3**, p. 147] and a longer account in [**4**, volume ii, pp. 321–2]. In practical use, the square root in the formula generally must be evaluated and, true to character, Heron supplies a method. It goes back to an

ancient Babylonian algorithm, and its simplicity and power still make it a favorite among computer programmers today: if a is smaller (larger) than $\sqrt{n}$, then $b = n/a$ is larger (smaller) than $\sqrt{n}$. Their average $(a + b)/2$ is closer to $\sqrt{n}$, and becomes the new a for the next approximation cycle. Amazingly, this is the same as Newton's successive tangents method for quadratics! In *Metrica*, Heron presents his area formula using triangle sides 7, 8, and 9. The nearest-integer approximation to the area $\sqrt{720}$, is $a = 27$. Only two passes yields 9 correct decimal places.

Some of his other ideas are also used today: He invented what is essentially a steam-driven version of a rotating garden sprinkler, making it the first steam engine. In those days, no one understood Newton's third law (action/reaction), but this whirling "aeolipile" also represented the first reaction (jet) engine. He invented a forerunner of the thermometer, as well as the coin-operated dispenser; inserting a coin would open doors, sprinkle holy water, or operate a musical organ (which could be powered by an air-pump connected to a windmill). He even made the wine and olive oil industries more profitable with his invention of the screw press, where high pressures could squeeze out substantially more juice or oil than before.

Remarkably, for many years historians had known of his most important geometrical work, *Metrica*, only through a single allusion to the square-root algorithm in it. Then in 1896, R. Schöne, working in Constantinople, chanced upon an 800–900-year-old copy of *Metrica*. This find turned out to be all the more significant because, more than any of his other works, this manuscript appears to be preserved in its original form [**4**, volume ii, pp. 309, 317].

REFERENCES

1. E. Artin, *Geometric Algebra*, Interscience Publishers, Inc., New York, 1957.
2. C. B. Boyer (U. Merzbach, rev. ed.), *A History of Mathematics*, John Wiley and Sons, Inc., New York, 1991.
3. H. Eves, *An Introduction to the History of Mathematics*, 5th ed., Saunders College Publishing, Philadelphia, 1982.
4. T. L. Heath, *A History of Greek Mathematics*, Dover Publications, Inc., New York, 1981.
5. K. Kendig, *Elementary Algebraic Geometry*, Springer-Verlag, New York, 1977.
- K. Kendig, Algebra, Geometry, and Algebraic Geometry: Some Interconnections, *Amer. Math. Monthly* **90** (1983) 161–173.
- K. Kendig, Stalking the Wild Ellipse, *Amer. Math. Monthly* **102** (1995) 782–787.
8. M. Klein, *Mathematics in Western Culture*, Oxford University Press, New York, 1953.
9. S. Mason, *A History of the Sciences*, The Macmillan Company, New York, 1962.
10. O. Neugebauer, *The Exact Sciences in Antiquity*, Dover Publications, Inc., New York, 1969.

KEITH KENDIG received his Ph.D. from UCLA under Basil Gordon and subsequently spent two years at the Institute for Advanced Study working with Hassler Whitney. Keith is a firm believer in exploring concrete examples, both in teaching and in discovering, and more often than not, uses the computer to enhance this process. He is a semi-professional cellist, an avid chamber music player, sporadic composer, and regular jogger. He's writing a book on how mathematicians make discoveries.

Cleveland State University, Cleveland, OH 44114

kendig@math.csuohio.edu