The
Mathematical Gazette

A JOURNAL OF THE MATHEMATICAL ASSOCIATION

Vol. 102 March 2018 No. 553

Is a straight line the shortest path?

JESSICA E. BANKS

Is the shortest path from A to B the straight line between them? Your first response might be to think it's obviously so. But in fact you know that it's not quite that straightforward. Your sat-nav knows it's not that straightforward. It asks whether you would like it to find the shortest route or the fastest route, because finding the best path depends on knowing what exactly you mean by 'long'. Likewise, if you're on a walk in the mountains, there's a good chance you'd rather follow the path around the head of the valley, rather than heading down the steep slope and up the other side.

The same sorts of considerations apply in mathematical worlds. I use the mountainside image because it is my preferred way of thinking of a Riemannian metric. Pick an abstract surface S. A Riemannian metric on S gives a well-behaved distance function. By force of habit I tend to picture S as sitting somehow within the physical world. Probably, I'm looking at it from the outside. But if I change viewpoint, so that I am walking around on S, I can picture how the topography affects the idea of the 'shortest path'.

A path that is locally the shortest path between any two points on it is called a *geodesic*. Given an embedded path $p : [0, 1] \to S$ on S, we can choose a Riemannian metric for which p is a geodesic: put p at the bottom of a very steep valley. So, if you generalise far enough, 'short' paths can be very much not straight. That's perhaps not very surprising; many things become possible if you relax the rules enough.

Here I'd like to focus on a much more restricted situation: the case where all points on the surface are 'the same'. If you're standing on the surface S, and every direction you look looks exactly the same, intuition says that the best way to reach a given point is to head straight for it. My aim here is to use that sameness to show that this intuition is correct.

Euclidean space

First consider the Euclidean plane $\mathbb{R}^2$, with its usual metric

$$d\left((x_1, y_1), (x_2, y_2)\right) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

Our aim is to show that the unique geodesic between two points is the straight line between them. To do so we'll use the following result.

Proposition 1: (Strict triangle inequality for $\mathbb{R}^2$)

Given three points in $\mathbb{R}^2$ that do not lie on a line, the sum of the lengths of any two sides of the triangle is strictly greater than the length of the third side. That is,

$$|(x_1, y_1) - (x_2, y_2)| + |(x_2, y_2) - (x_3, y_3)| > |(x_1, y_1) - (x_3, y_3)|.$$

Proof: We could show this by calculation as follows. To make this as simple as possible, assume $x_1 = y_1 = y_2 = 0$. If

$$\sqrt{x_2^2} + \sqrt{(x_2 - x_3)^2 + y_3^2} \leqslant \sqrt{x_3^2 + y_3^2}$$

then

$$x_2^2 y_3^2 \leqslant 0.$$

This cannot be the case, since the hypothesis that the three points are not collinear means both $x_2 \neq 0$ and $y_3 \neq 0$.

On the other hand, we know this result intuitively: if you bend a straw in the middle, the ends get closer together, never further apart.

Alternatively, we can prove the proposition with isosceles triangles.

Proof: Clearly we only need to worry about the longest edge of the triangle. Let A, B and C be the vertices of a triangle, where AC is the longest side. Suppose the angle at A is α, the angle at B is β and the angle at C is γ. Then $\alpha + \beta + \gamma = 180°$.

Extend the edge AB past B to a point D so that AD is the same length as AC. Then $\angle ADC = \frac{1}{2}(180° - \alpha)$. See Figure 1.

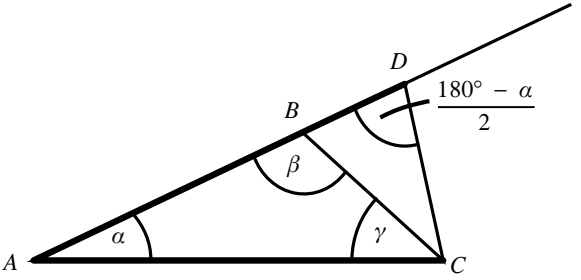

FIGURE 1

Again extend the edge AB past B, this time to a point E such that BE is the same length as BC. Then $\angle AEC = \frac{1}{2}\beta$. See Figure 2.

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

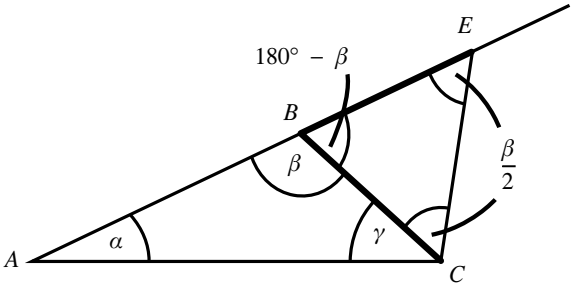

FIGURE 2

Now compare the angles $\angle ACD$ and $\angle ACE$. The first is $180° - \alpha - \frac{1}{2}(180° - \alpha)$, while the second is $\gamma + \frac{1}{2}\beta$. Notice that

$$\begin{aligned} 180° - \alpha - \frac{180° - \alpha}{2} &= \frac{180° - \alpha}{2} \\ &= \frac{\beta + \gamma}{2} \\ &< \gamma + \frac{\beta}{2}. \end{aligned}$$

This tells us that E is further from A than D is.

Armed with this fact, we can now show our main result.

Proposition 2: Geodesics are all straight lines.

Proof: Fix two points A and B, and join them by a geodesic g. Also draw the line l_∞ that passes through A and B, and name the segment of this line between A and B as l. Using translation and rotation if necessary, we can safely picture l as being horizontal, with A on the left and B on the right. See Figure 3.

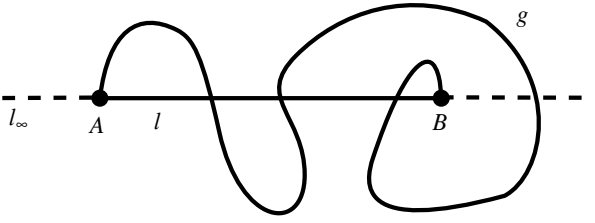

FIGURE 3

We want to show that all points of g lie on l_∞. The proof will be by contradiction, so we shall assume this is not the case. Then there is a point C on g that is not on l_∞. By using reflection if needed, we can picture C as above l_∞. The point C divides the geodesic g into two pieces, g_A and g_B, each

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

of which is also geodesic. Draw two straight lines, l_A and l_B, joining A and B respectively to C. See Figure 4.

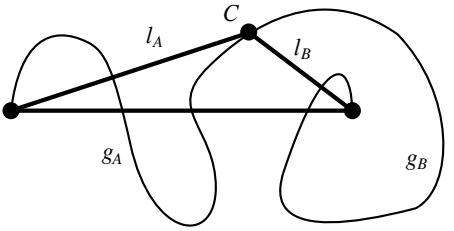

FIGURE 4

Our next step is to make use of the existence of a rotation by any angle about any point. How we proceed depends on the relative lengths of l, l_A and l_B. Since we have done nothing to distinguish A from B, we can assume that l_A is at least as long as l_B.

We can construct a rotation around A that moves l_A to a segment l'_A of l_∞ that overlaps with l. This rotation also takes C to a point D, and g_A to a geodesic g'_A between A and D. See Figure 5. Notice that l'_A is the same length as l_A, and g'_A is the same length as g_A. Since g_A is strictly shorter that g, this means D cannot coincide with B. Accordingly, either l'_A is strictly shorter than l, in which case D lies on l between A and B, or else l'_A is strictly longer than l, in which case B lies on l'_A between A and D.

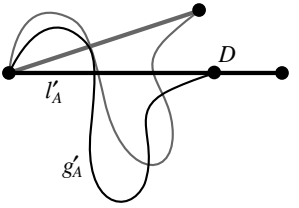

FIGURE 5

Likewise, we can rotate about B to take l_B to a segment l'_B of l_∞, C to a point E, and g_B to a geodesic g'_B from B to E. There are two possible positions for the point E, determined by whether l'_B overlaps with l or not. Our aim is to arrange that the points A and D interleave with B and E along l_∞. Here we consider two cases.

If l'_A is strictly shorter than l, then l'_B is also shorter than l. On the other hand, by the strict triangle inequality, l is strictly shorter than the sum of the lengths of l'_A and l'_B. In this case we arrange that l'_B overlaps with l. The point E must then lie between A and D.

If l'_A is strictly longer than l, we arrange that l'_B does not overlap with l. Again using the strict triangle inequality, we know that l'_A is strictly shorter than the sum of the lengths of l and l'_B. This tells us that E must lie on the other side of D to B.

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

In either case, we have a pair of geodesics, g'_A joining A to D and g'_B joining B to E, and the points A and D interleave with the points B and E along l_∞.

Using the reflection in l_∞, we can create from g'_A a geodesic h_A from A to D that lies entirely on or above l_∞. See Figure 6. The path h_A has the same length as g'_A, and so also as g_A. Similarly, we can create a geodesic h_B from E to B that lies entirely on or above l_∞ and has the same length as g_B.

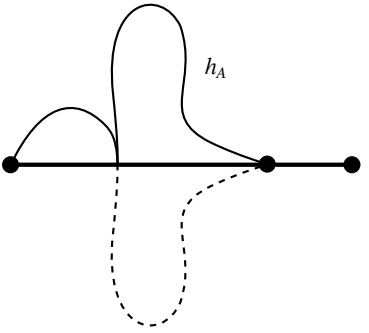

FIGURE 6

Since their endpoints interleave, h_A and h_B must meet at a point F. See Figure 7. Following h_A from A to F, and then following h_B from F to B gives a path from A to B that is strictly shorter than g. That is not possible, since we chose g to be a geodesic between A and B. This contradiction shows that our original assumption was wrong, so every point of g must lie in l_∞.

Since a geodesic cannot pass through the same point twice, the path g must coincide with l.

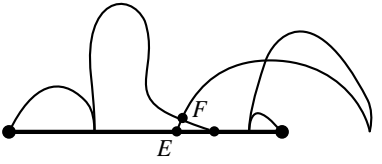

FIGURE 7

This proof doesn't make much use of the Euclidean metric. It uses the strict triangle inequality. It uses the fact that we can rotate by any angle about any point. It uses the fact that we can reflect in any line. That's all. If we can find other situations where the same three things are true, then the same proof will apply.

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

Spherical geometry

Another form of geometry we find in everyday life is spherical geometry. When walking, or even driving, we can safely pretend that we live on a flat Earth, but once you start trying to run an airline it becomes important to remember that the surface of the Earth is actually (roughly) spherical.

Mathematically, we picture 'the' sphere as the set of points at distance 1 from the origin in three-space. That is, the equation of the 2-sphere $\mathbb{S}^2$ is $x^2 + y^2 + z^2 = 1$.

In this context, the term 'line' refers to a great circle. A *great circle* is the intersection of $\mathbb{S}^2$ with a plane through the origin in $\mathbb{R}^3$. It passes through pairs of antipodal points on $\mathbb{S}^2$. Examples of great circles on Earth include all the lines of longitude and the Equator (but not the other lines of latitude).

As in Euclidean 2-space, in spherical 2-space we can reflect in any line, and we can rotate by any angle about any point (the rotation will also be a rotation about the antipodal point). To prove Proposition 2, we therefore only need to establish that the strict triangle inequality also holds in spherical geometry.

The catch is that it doesn't. One big difference between Euclidean and spherical geometry is that if you keep walking in the same direction for a long time, in Euclidean space you continue getting farther from where you started, but in spherical space you eventually get back to where you started. If two sides of a triangle together reach more than halfway round $\mathbb{S}^2$, the third side can be shorter because it 'goes round the other way'.

Nevertheless, we can still use the proof of Proposition 2 for spherical geometry. The reason for this is that the property of being a geodesic is a local property. That is, to check if a path is a geodesic, you only need to look at a small part of it at a time. We will see that the strict triangle inequality holds in $\mathbb{S}^2$ if we restrict it to sufficiently small scales. The proof of Proposition 2 will then hold under the additional assumption that the points A and B chosen are suitably close together.

Before we can actually prove what we want, we need to consider how to find the lengths of sections of great circles. If we take a great circle in $\mathbb{S}^2$, and consider it as a circle in the plane through the origin used to define it, we can use the standard formula to relate the length of an arc to the angle subtended at the centre of the circle. Given a sector with angle θ of a circle with radius r (see Figure 8), the length l of the corresponding arc of the circle is proportional to $r\theta$. If we measure θ in radians then in fact $l = r\theta$. If we measure in degrees instead, the constant of proportionality serves to convert the units. This relationship between lengths and angles is used in the following proof to convert one problem into another (easier) one.

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

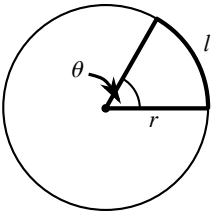

FIGURE 8

Proposition 3: (Strict triangle inequality for $\mathbb{S}^2$)

Given three points close together in $\mathbb{S}^2$ that do not lie on a line, join the three points by short arcs of great circles. Then the sum of the lengths of any two sides of the resulting triangle is strictly greater than the length of the third side.

Proof: Call the three points A, B and C. Label the arc opposite A as α, the arc opposite B as β and the arc opposite C as γ. Also join each vertex to the origin O by a (Euclidean) straight line. See Figure 9. The three lengths α, β and γ correspond to the three angles α', β' and γ' made at O by these lines. See Figure 10. This means that proving the triangle inequality for the lengths is the same problem as proving it for the angles. However, when we think about the angles, this is again a situation we know about from real life. Imagine making the sector AOC out of paper, and using a second piece of paper, folded, to make the two other sectors AOB and BOC. See Figure 11. Once you unfold the second piece of paper to lie it flat, it will open out to have a bigger angle than the first piece of paper.

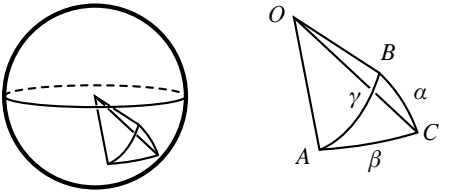

FIGURE 9

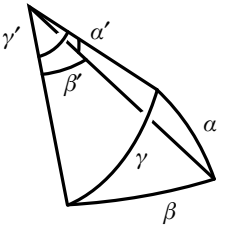

FIGURE 10

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

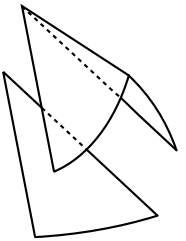

FIGURE 11

Alternatively, add in the three Euclidean straight lines AB, BC and AC. Focusing on these lines, instead of the arcs, gives a Euclidean triangle ABC with side lengths α'', β'' and γ''. See Figure 12. As the three edges OA, OB and OC all have the same length, the triangle inequality holds for α', β' and γ' if, and only if, it holds for α'', β'' and γ''. That is, we can deduce the result in the spherical case from the result in the Euclidean case.

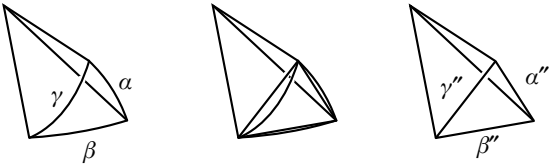

FIGURE 12

Hyperbolic geometry

The third form of geometry we will consider is hyperbolic geometry. This is harder to visualise than Euclidean or spherical geometry, as we do not really encounter it in everyday life. Hilbert's Theorem (see [1, 5.11] or [2, 5.12]) says that there is no way to fit the whole of the hyperbolic plane $\mathbb{H}^2$ smoothly inside $\mathbb{R}^3$ so that all distances are correct. We can, however, make small pieces of the hyperbolic plane. The pseudosphere (see Figure 13 and [3, 20.7]) is a geometrical shape that locally models $\mathbb{H}^2$. For a more intuitive, if less mathematically precise, understanding of hyperbolic geometry, you could crochet some hyperbolic coral (originally designed by Daina Taimina, see [4, 5, 6] or sew yourself a hyperbolic blanket (instructions by Jeff Weeks are available at [7], with credit for the design given to Helaman Ferguson).

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

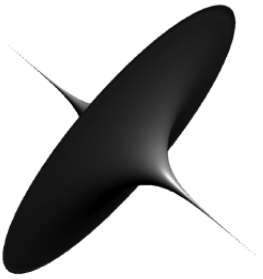

FIGURE 13

To work mathematically with $\mathbb{H}^2$, we use different models that fit more happily in $\mathbb{R}^3$ (in fact in $\mathbb{R}^2$) by changing the metric. For our purposes now, we will use the Poincaré disc model $\mathbb{D}$ of $\mathbb{H}^2$. In this model, the points of $\mathbb{H}^2$ are identified with the open unit disc in the complex plane $\mathbb{C}$:

$$\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}.$$

The metric at each point is scaled by a factor of $\dfrac{2}{1 - |z|^2}$. The bounding circle $|z| = 1$ is called the *boundary* of $\mathbb{H}^2$.

The 'straight lines' in this case are given by circles perpendicular to the boundary. This includes the lines through the origin, which can be thought of as 'circles with infinite radius'. These are often the easiest hyperbolic lines to calculate with. For example, consider the line segment along the real axis from the origin to the point r, where $0 < r < 1$. The hyperbolic length of this line segment is given by the integral

$$\begin{aligned}\int_0^r \frac{2}{1 - t^2}\, dt &= \int_0^r \frac{1}{1 - t} + \frac{1}{1 + t}\, dt \\ &= \Big[-\ln|1 - t| + \ln|1 + t|\Big]_0^r \\ &= \ln\left(\frac{1 + r}{1 - r}\right).\end{aligned}$$

This value becomes arbitrarily large as r approaches 1; that is, the boundary is infinitely far away.

To apply the proof of Proposition 2 to $\mathbb{H}^2$, we need to know about isometries. One type of isometry that is easy to picture is a rotation by any angle around the origin. This is an isometry because the metric is scaled by a factor that only depends on the distance from the origin. Another type of isometry is a reflection in a line through the origin. Moreover, although we will not prove it here, any function of the complex numbers of the form

$$f(z) = \frac{az + b}{\bar{b}z + \bar{a}},$$

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

for complex numbers a and b, is an isometry of $\mathbb{D}$. Given a point $\omega \in \mathbb{D}$, taking $a = 1$ and $b = -\omega$ gives a map f_ω that takes ω to the origin, while the inverse map

$$f_\omega^{-1}(z) = \frac{z + \omega}{\overline{\omega} z + 1}$$

takes 0 to ω. By combining these three types of isometry, it is possible to rotate by any angle about any point, and to reflect in any hyperbolic line. The following result, which makes use of these ideas, is therefore the final step in proving Proposition 2 for $\mathbb{H}^2$.

Proposition 4: (Strict triangle inequality for the Poincaré disc)

Let A, B and C be three points in $\mathbb{D}$ that do not lie on a hyperbolic line (that is, on a single circle perpendicular to the boundary). Join the three points by hyperbolic line segments to form a triangle. Then the length of the line segment from B to C is strictly less than the sum of the lengths of the other two sides.

Proof: In $\mathbb{D}$, we can move any point to any other by an isometry, and we can rotate about any point by any angle. This means we can assume the three vertices are $A = 0$, $B = r$ and $C = se^{i\theta}$ for positive real values of r and s. Since the points are not collinear, we know $0 < r < 1$, $0 < s < 1$ and $-1 < \cos\theta < 1$.

We have already calculated that the distance from A to B along the line joining them is given by

$$\gamma = \ln\left(\frac{1 + r}{1 - r}\right)$$

and the distance from A to C is given by

$$\beta = \ln\left(\frac{1 + s}{1 - s}\right).$$

Denote by α the distance from B to C. Rather than trying to calculate the line connecting B to C directly, we can use another isometry to change the line into one we already know about. Define a map $f_r : \mathbb{C} \to \mathbb{C}$ by

$$f_r(z) = \frac{z - r}{1 - rz}.$$

This map is chosen because $f_r(r) = 0$; we can find the distance of C from B by calculating the Euclidean distance t of $f_r\left(se^{i\theta}\right)$ from 0. To make things simpler, we will focus on the square of the distance, which is given by multiplying by the complex conjugate of the complex number $f_r\left(se^{i\theta}\right)$:

$$\begin{aligned} t^2 = \left(\frac{se^{i\theta} - r}{1 - rse^{i\theta}}\right)\overline{\left(\frac{se^{i\theta} - r}{1 - rse^{i\theta}}\right)} &= \left(\frac{se^{i\theta} - r}{1 - rse^{i\theta}}\right)\left(\frac{se^{-i\theta} - r}{1 - rse^{-i\theta}}\right) \\ &= \frac{r^2 + s^2 - rs\left(e^{i\theta} + e^{-i\theta}\right)}{1 + r^2s^2 - rs\left(e^{i\theta} + e^{-i\theta}\right)} \\ &= \frac{r^2 + s^2 - 2rs\cos\theta}{1 + r^2s^2 - 2rs\cos\theta}. \end{aligned}$$

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

Our aim is to prove that $\alpha < \beta + \gamma$. Assume otherwise (that is, that $\alpha \geqslant \beta + \gamma$). Then

$$\ln\left(\frac{1+s}{1-s}\right) + \ln\left(\frac{1+r}{1-r}\right) \leqslant \ln\left(\frac{1+t}{1-t}\right)$$
$$\Rightarrow \left(\frac{1+s}{1-s}\right)\left(\frac{1+r}{1-r}\right) \leqslant \left(\frac{1+t}{1-t}\right)$$
$$\Rightarrow t(2+2rs) \geqslant 2r + 2s$$
$$\Rightarrow t^2(1+rs)^2 \geqslant (r+s)^2.$$

Substituting in the formula for t^2 shows that

$$(r^2+s^2-2rs\cos\theta)(1+rs)^2 \geqslant (r+s)^2\left(1+r^2s^2-2rs\cos\theta\right)$$
$$\Rightarrow \left((r+s)^2-2rs(1+\cos\theta)\right)(1+rs)^2 \geqslant (r+s)^2\left((1+rs)^2-2rs(1+\cos\theta)\right)$$
$$\Rightarrow -2rs(1+\cos\theta)(1+rs)^2 \geqslant -2rs(1+\cos\theta)(r+s)^2.$$

Recalling that $r > 0$, $s > 0$ and $1+\cos\theta > 0$, this simplifies to

$$(1+rs)^2 \leqslant (r+s)^2$$
$$\Rightarrow 1 + r^2s^2 - r^2 - s^2 \leqslant 0$$
$$\Rightarrow \left(1-r^2\right)\left(1-s^2\right) \leqslant 0.$$

This is not possible, since $r^2 < 1$ and $s^2 < 1$. We therefore conclude that $\alpha < \beta + \gamma$, as required.

Ideas for further reading

Euclidean, spherical and hyperbolic geometry are closely related. They each satisfy the first four of Euclid's postulates, and are distinguished by satisfying distinct variants of the fifth ([8] is an online edition of *Euclid's Elements*). The history of the study of these different cases is widely documented (for example see [9], [10] or [11]).

Coxeter [3] discusses Euclidean and hyperbolic geometry, as well as projective geometry (which is also known as elliptic geometry, and is closely related to spherical geometry). Some of the results are proved in the setting of absolute geometry (the case where the fifth postulate is not used at all, giving proofs that are equally applicable to Euclidean, hyperbolic and spherical geometry).

An exploration of other geometric ideas taking an intuitive approach is given in [12]. If, on the other hand, you prefer a more calculation-based approach, other techniques for proving Proposition 2 can be found at [13].

Acknowledgement

I am grateful to Trevor Jarvis, who read a draft and spotted a detail I had overlooked.

https://doi.org/10.1017/mag.2018.2 Published online by Cambridge University Press

References

1. M. P. do Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall (1976).
2. M. Spivak, *A comprehensive introduction to differential geometry*. Vol. III. (2nd edn.), Publish or Perish (1979).
3. H. S. M. Coxeter, *Introduction to geometry* (2nd edn.) Wiley (1989).
4. D. W. Henderson and D. Taimina, Crocheting the hyperbolic plane, *The Mathematical Intelligencer*, **23**(2) (2001) pp. 17-28.
5. D. Taimina, Hyperbolic crochet – some fiber for thoughts about art, math, crochet, and all the various threads in our lives (2016), available at http://hyperbolic-crochet.blogspot.co.uk
6. *Institute For Figuring*, Hyperbolic space, accessed August 2017 at http://crochetcoralreef.org/about/hyperbolic_space.php
7. J. Weeks, How to sew a hyperbolic blanket, accessed August 2017 at http://www.geometrygames.org/HyperbolicBlanket/index.html
8. D. E. Joyce, *Euclid's Elements*, Clark University, Massachusetts (1998), available at http://aleph0.clarku.edu/~djoyce/java/elements/
9. F. P. Lewis, History of the parallel postulate, *Amer. Math. Monthly*, **27**(1) pp. 16-23 (1920).
10. R. Osserman, *Poetry of the Universe*, Anchor Books (1996).
11. J. J. O'Connor and E. F. Robertson, Non-Euclidean geometry, MacTutor History of Mathematics archive (accessed 11 Jan 2017).
12. D. Hilbert and S. Cohn-Vossen, *Geometry and the imagination*, Chelsea Publishing Company (1952).
13. Anirbit Dhar, Geodesics on spheres are great circles (2010) available at http://mathoverflow.net/q/12200

10.1017/mag.2018.2

JESSICA E. BANKS

School of Mathematics, University Walk, Bristol BS8 1TW

e-mails: *jessica.banks@lmh.oxon.org, jessica.banks@bristol.a.c.uk*

Appeal for more referees

The *Gazette* relies on good peer-refereeing to maintain the standard of the articles accepted. We are always seeking new referees, and are currently very short in some areas. In particular, we do not have enough referees for submissions on Euclidean geometry, mechanics, number theory and analysis, although fresh volunteers are welcome for any subject

I would be delighted to hear from any readers who would be willing to join the refereeing team. If you are not sure what is involved, I can send you details of what the work entails and a form so that you can specify particular areas of expertise. Please email me or write to me at the address on the last page of the journal if you are interested in supporting the journal in this vital way.

Gerry Leversha

g.leversha@btinternet.com

Editor, *The Mathematical Gazette*