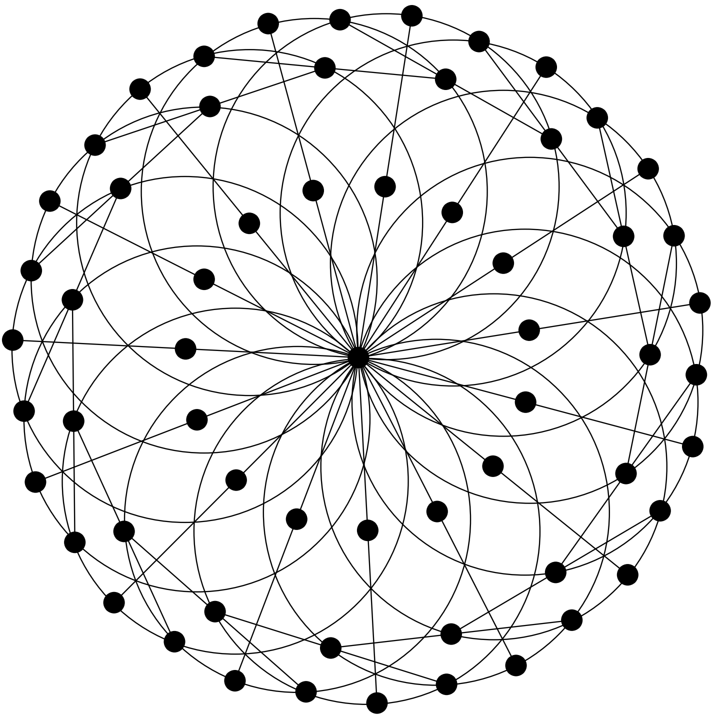
*

ISSN 1350-8539

M500 193



The M500 Society and Officers

The M500 Society is a mathematical society for students, staff and friends of the Open University. By publishing M500 and 'MOUTHS', and by organizing residential weekends, the Society aims to promote a better understanding of mathematics, its applications and its teaching.

The magazine M500 is published by the M500 Society six times a year. It provides a forum for its readers' mathematical interests. Neither the editors nor the Open University necessarily agree with the contents.

MOUTHS is 'Mathematics Open University Telephone Help Scheme', a directory of M500 members who are willing to provide mathematical assistance to other members.

The September Weekend is a residential Friday to Sunday event held each September for revision and exam preparation. Details available from March onwards. Send SAE to Jeremy Humphries, below.

The Winter Weekend is a residential Friday to Sunday event held each January for mathematical recreation. Send SAE for details to Norma Rosier, below.

Advice to authors. We welcome contributions to M500 on virtually anything related to mathematics and at any level from trivia to serious research. Please send material for publication to Tony Forbes, above. We prefer an informal style and we usually edit articles for clarity and mathematical presentation. If you use a computer, please also send the file on a PC diskette or via e-mail. Camera-ready copy can be accepted if it follows the general format of the magazine.

Dot products and determinants – is there a connection?

Robin Marks

First I will try to explain briefly what is meant by a vector, an inner product, a dot product and a determinant.

What is a vector? A position vector in Euclidean space can be thought of as a line with two properties, length and direction. Two lines of the same length which point in the same direction are considered to be the same vector.

Suppose we have a vector f. If this exists in the vector space, 2-dimensional Euclidean space, it can be represented by Cartesian coordinates, which we will call f_0 and f_1. Thus f can be represented as the list (f_0, f_1). This is actually a short way of saying $f = f_0 e_0 + f_1 e_1$, where e_0 and e_1 are vectors of length 1 which are mutually perpendicular (known as orthonormal), and define the coordinate axes. They form a set of basis vectors for the 2-dimensional Euclidean space. Note that we choose the vector f first, then we choose any one set out of the infinite number of sets of orthonormal basis vectors. Other sets can be obtained by rotating the coordinate axes. The particular set of basis vectors chosen will not matter in the following. Vectors can exist in any number of dimensions. For example, in a 5-dimensional space, we can represent a vector f by $(f_0, f_1, f_2, f_3, f_4)$.

What is an inner product? What is a dot product? An inner product is the result of combining two vectors in a particular way. There are different types of inner product for different vector spaces. In Euclidean space we use an inner product known as the dot product. The inner product or the dot product of two vectors f and g may be written $\langle f, g\rangle$. The dot product of vectors that are perpendicular to one another is zero. Therefore the dot product of any two different basis vectors is zero. The dot product of a vector of length 1 with itself is 1. Dot products are linear. This means that for vectors $f = f_0 e_0 + f_1 e_1$ and $g = g_0 e_0 + g_1 e_1$ we have $\langle f, g\rangle = \langle f_0 e_0 + f_1 e_1, g_0 e_0 + g_1 e_1\rangle = \langle f_0 e_0, g_0 e_0\rangle + \langle f_0 e_0, g_1 e_1\rangle + \langle f_1 e_1, g_0 e_0\rangle + \langle f_1 e_1, g_1 e_1\rangle = f_0 g_0 \langle e_0, e_0\rangle + f_0 g_1 \langle e_0, e_1\rangle + f_1 g_0 \langle e_1, e_0\rangle + f_1 g_1 \langle e_1, e_1\rangle = f_0 g_0 + 0 + 0 + f_1 g_1$.

Similarly, in N-dimensional space, $\langle f, g\rangle = \sum_{i=0}^{N-1} f_i g_i$. The square of the length of f is given by $\langle f, f\rangle = \sum_{i=0}^{N-1} f_i f_i$ so that the the length of f is $\langle f, f\rangle^{1/2}$ which can be written $\|f\|$. The dot product is equal to $\|f\|\|g\| \cos\theta$, where θ is the angle between the vectors f and g. In two dimensions we can prove this by putting $f_0 = \|f\| \cos\theta_f$, and $f_1 = \|f\| \sin\theta_f$, with similar expressions for g, from which we get $\sum_{i=0}^{1} f_i g_i = \|f\|\|g\| \cos(\theta_f - \theta_g)$.

What is a determinant? Eliminating x and y from the two equations $f_0x + g_0y = 0, f_1x + g_1y = 0$, gives $f_0g_1 - f_1g_0 = 0$. Now $f_0g_1 - f_1g_0$ is known as the determinant for this system of equations, or the determinant of the matrix $\begin{bmatrix} f_0 & g_0 \\ f_1 & g_1 \end{bmatrix}$, or det $\begin{bmatrix} f_0 & g_0 \\ f_1 & g_1 \end{bmatrix}$.

Determinants of larger matrices can be found in a similar manner. Determinants exist only for square matrices. One useful property is that multiples of rows or columns of a matrix can be added together without changing the determinant's value. Another property is that if we have two square matrices M and N, the determinant of their product equals the product of their determinants:

$$\det MN \;=\; (\det M)(\det N).$$

A third property is that transposing a matrix leaves the determinant unchanged: $\det M = \det M^T$. So an inner product is the result of combining two vectors in a particular way, and a determinant is a property of a square matrix. Why should I think these two might be connected? Well, if we take vectors $f = (f_0, f_1)$ and $g = (g_0, g_1)$, with an angle θ between them, the area of the parallelogram made with 'sides' f and g is equal to $\|f\|\|g\| \cos\theta = \langle f, g\rangle$, the inner product. The area of this parallelogram is also equal to the absolute value of det $\begin{bmatrix} f_0 & g_0 \\ f_1 & g_1 \end{bmatrix}$. (I came across this last statement, with no explanation, in a mathematics book. Problem: Can any M500 reader can prove it?)

So, at least in the 2-dimensional case, the inner product equals the absolute value of the determinant. This set me wondering ... how and ... why?

Take a vector f. For convenience of illustration we will work in three dimensions, so the vector can represented as (f_0, f_1, f_2).

The projection of f onto each of the three axes is shown in Figure 1. What is the length of the projection of the vector f onto the third axis? Let the projection of f onto the axis with unit vector e_2 be αe_2, with α a real number. The inner product of the orthogonal vectors n_2 and αe_2 must be zero:

$$\begin{aligned} &\langle n_2, \alpha e_2\rangle = 0 \;\Leftrightarrow\; \langle f - \alpha e_2, \alpha e_2\rangle = 0 \;\Leftrightarrow\; \langle f, \alpha e_2\rangle - \langle \alpha e_2, \alpha e_2\rangle = 0 \\ &\quad\Leftrightarrow\; \alpha\langle f, e_2\rangle - \alpha^2\langle e_2, e_2\rangle = 0 \;\Leftrightarrow\; \alpha = \langle f, e_2\rangle/\langle e_2, e_2\rangle = f_2. \end{aligned}$$

Thus the length of the projection of f is f_2, hence the projection of f is f_2e_2, which is $(0, 0, f_2)$ in list notation.

Figure 1: Projection of f on to $\{e_i\}$

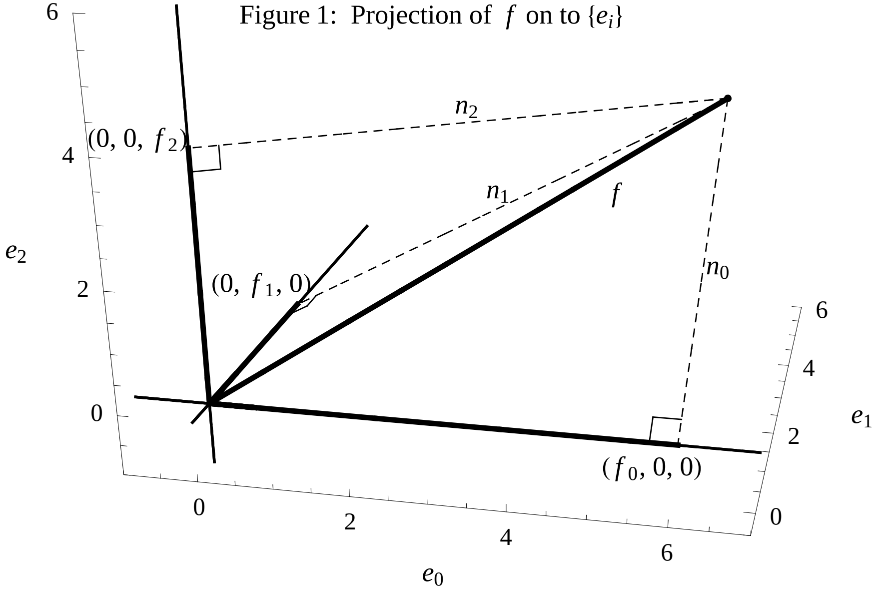
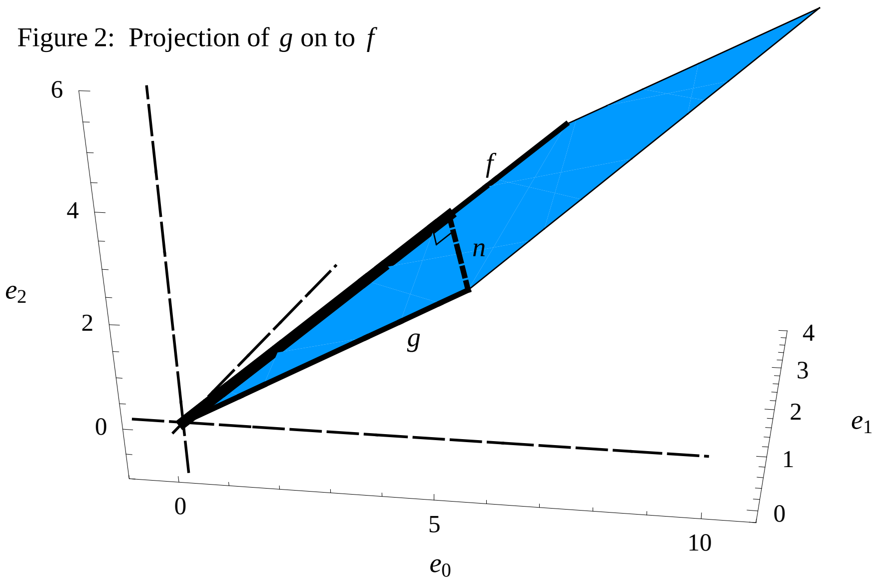


Figure 2: Projection of g on to f



Now let us introduce a second vector g, and 'drop a perpendicular' from g to f (Figure 2). Call the perpendicular n, where n is a vector. Now $n = g - \alpha f$ for some value of α that minimizes the length of n, with $\alpha \in \mathbb{R}$; αf is known as the projection of g onto f. To find the value of α giving a minimum value of $\langle n, n\rangle$, put $d\langle n, n\rangle / d\alpha = 0$. Thus

$$\frac{d\langle g - \alpha f, g - \alpha f\rangle}{d\alpha} = 0$$
$$\Leftrightarrow \frac{d(\langle g, g\rangle - \alpha\langle f, g\rangle - \alpha\langle g, f\rangle + \alpha^2\langle f, f\rangle)}{d\alpha} = 0$$
$$\Leftrightarrow 2\alpha\langle f, f\rangle - 2\langle f, g\rangle = 0 \Leftrightarrow \alpha = \frac{\langle f, g\rangle}{\langle f, f\rangle}.$$

We can get the same answer more easily by arguing that the inner product of orthogonal vectors, $\langle n, f\rangle$, must be zero:

$$\langle g - \alpha f, f\rangle = 0 \Leftrightarrow \langle g, f\rangle - \alpha\langle f, f\rangle = 0 \Leftrightarrow \alpha = \frac{\langle g, f\rangle}{\langle f, f\rangle}.$$

Now we can work out the area of the parallelogram which has sides f and g.

$$\begin{aligned}
(\text{area } fg)^2 &= (\text{height})^2 \cdot (\text{base})^2 \\
&= \langle n, n\rangle\langle f, f\rangle = \langle g - \alpha f, g - \alpha f\rangle\langle f, f\rangle \\
&= (\langle g, g\rangle - 2\alpha\langle f, g\rangle + \alpha^2\langle f, f\rangle)\langle f, f\rangle \qquad \text{(i)} \\
&= \langle g, g\rangle\langle f, f\rangle - 2\frac{\langle g, f\rangle}{\langle f, f\rangle}\langle f, g\rangle\langle f, f\rangle + \left(\frac{\langle g, f\rangle}{\langle f, f\rangle}\right)^2 \langle f, f\rangle^2 \\
&= \langle g, g\rangle\langle f, f\rangle - \langle g, f\rangle\langle f, g\rangle.
\end{aligned}$$

I was very pleased to find this lovely symmetrical expression. Don't you think its nice? This $(\text{area } fg)^2$, as I will call it, has a maximum when $\langle f, g\rangle = 0$; that is, when the parallelogram is a rectangle. The area is zero when $f = kg$, with $k \in \mathbb{R}$, a constant; that is, when f and g are colinear. For example, suppose $f = (1, 1, 0)$ and $g = (0, 1, 1)$ Then

$$\begin{aligned}
(\text{area } fg)^2 &= \langle g, g\rangle\langle f, f\rangle - \langle g, f\rangle\langle f, g\rangle \\
&= \left(\sum_{i=0}^{2} f_i^2\right)\left(\sum_{i=0}^{2} g_i^2\right) - \left(\sum_{i=0}^{2} f_i g_i\right)^2 = 2 \cdot 2 - 1^2 = 3.
\end{aligned}$$

Hence the parallelogram area is $\sqrt{3}$.

What about the determinant $\det \begin{bmatrix} f_0 & g_0 \\ f_1 & g_1 \end{bmatrix}$? Note that f and g now have only two dimensions whereas in the discussion up to now they could have had any number of dimensions. Call the matrix M. One day I suddenly realized that by pre-multiplying M with its transpose we obtain a matrix full of inner products:

$$M^TM \;=\; \begin{bmatrix} f_0 & f_1 \\ g_0 & g_1 \end{bmatrix} \begin{bmatrix} f_0 & g_0 \\ f_1 & g_1 \end{bmatrix} \;=\; \begin{bmatrix} \langle f,f\rangle & \langle f,g\rangle \\ \langle g,f\rangle & \langle g,g\rangle \end{bmatrix}.$$

The determinant of M^TM is

$$\begin{aligned} \det M^TM \;&=\; (\det M^T)(\det M) \;=\; (\det M)(\det M) \;=\; (\det M)^2 \\ &=\; \langle f,f\rangle\langle g,g\rangle - \langle g,f\rangle\langle f,g\rangle \;=\; (\text{area } fg)^2. \end{aligned}$$

So in two dimensions, the relation between between determinants and inner products is

$$(\det M)^2 \;=\; \langle f,f\rangle\langle g,g\rangle - \langle g,f\rangle\langle f,g\rangle \;=\; (\text{area} fg)^2.$$

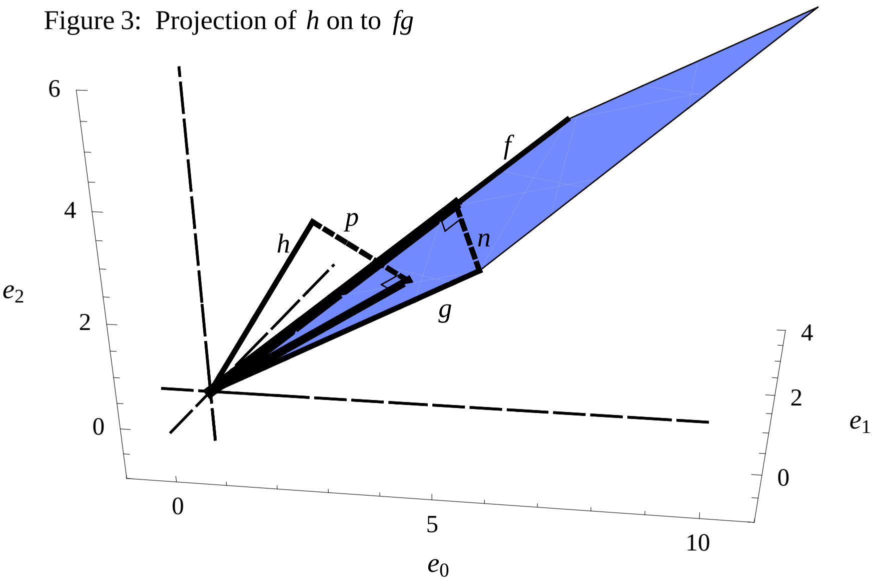
Let us look at integer-valued components of the 2-dimensional vectors f and g, in particular, values giving a parallelogram with $(\text{area } fg)^2 = 1$, that is $(f_0g_1 - f_1g_0)^2 = 1$. It is easy to generate such values. We start with a matrix with a determinant which will square to 1; that is, with determinant 1 or -1. We will choose $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Now add an integer multiple of one column to the other column, or an integer multiple of one row to the other row. For example add 3 times column 1 to column 2: $\begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$, then add 2 times row 1 to row 2: $\begin{bmatrix} 1 & 3 \\ 2 & 7 \end{bmatrix}$. These correspond to parallelograms constructed with the vectors given by the rows or columns of the matrix, each with area 1. There are clearly an infinite number of such parallelograms.

If one of the vectors is very long, the parallelogram will be correspondingly very thin. For example, given the matrix $\begin{bmatrix} 100 & 101 \\ 99 & 100 \end{bmatrix}$, with determinant 1, the parallelogram's width is of the order $1/100$ at its widest point. Problem: Suppose f and g are 3-dimensional, with integer components, and suppose that at least one of f and g has a non-zero component in each of the three dimensions. Can the area of the parallelogram formed by them have $(\text{area } fg)^2 = 1$? If not, can you find the smallest possible non-zero area? What about when f and g are 4-dimensional vectors?

Now we introduce a third vector, h, and drop a perpendicular p to the plane in which f and g lie (Figure 3).

We 'drop a perpendicular' from the tip of h to the subspace generated by a linear combination of f and g. That is, the plane $\alpha f + \beta g$, with $\alpha, \beta \in \mathbb{R}$. We seek to minimize $\langle p,p\rangle$, the (perpendicular distance)^2, with respect to both α and β simultaneously. So, $\partial\langle p,p\rangle/\partial\alpha = 0$ and $\partial\langle p,p\rangle/\partial\beta = 0$:

Figure 3: Projection of h on to fg



$$\frac{\partial\langle h-(\alpha f+\beta g), h-(\alpha f+\beta g)\rangle}{\partial\alpha} = 0$$

and

$$\frac{\partial\langle h-(\alpha f+\beta g), h-(\alpha f+\beta g)\rangle}{\partial\beta} = 0.$$

This is fairly easily solved to yield:

$$\alpha = \frac{\langle g,g\rangle\langle h,f\rangle-\langle f,g\rangle\langle g,h\rangle}{\langle g,g\rangle\langle f,f\rangle-\langle g,f\rangle\langle f,g\rangle}, \quad \beta = \frac{\langle f,f\rangle\langle g,h\rangle-\langle f,g\rangle\langle h,f\rangle}{\langle g,g\rangle\langle f,f\rangle-\langle g,f\rangle\langle f,g\rangle}.$$

The vector $\alpha f+\beta g$ is the projection of h onto the subspace. The parallelepiped defined by the vectors f, g and h has a volume

$$\begin{aligned}(\text{volume } fgh)^2 &= (\text{height})^2\cdot(\text{area } fg)^2\\ &= \langle p,p\rangle(\langle g,g\rangle\langle f,f\rangle-\langle g,f\rangle\langle f,g\rangle)\\ &= \langle h-(\alpha f+\beta g), h-(\alpha f+\beta g)\rangle(\langle g,g\rangle\langle f,f\rangle-\langle g,f\rangle\langle f,g\rangle).\end{aligned}$$

Expanding and simplifying eventually (after considerable effort!) gives

$$\begin{aligned}&(\text{volume } fgh)^2\\ &\quad= \langle f,f\rangle\langle g,g\rangle\langle h,h\rangle-\langle f,f\rangle\langle g,h\rangle^2-\langle g,g\rangle\langle f,h\rangle^2\\ &\qquad-\langle h,h\rangle\langle f,g\rangle^2+2\langle f,g\rangle\langle g,h\rangle\langle h,f\rangle. \end{aligned} \qquad \text{(ii)}$$

This is another very nice expression, don't you think? It holds for vectors f,

g and h of any number of dimensions. For example, suppose $f = (1,0,1,0)$, $g = (0,1,1,1)$ and $h = (1,0,0,1)$. Then

$$\begin{aligned}
&(\text{volume } fgh)^2 \\
&\qquad = \langle f,f\rangle\langle g,g\rangle\langle h,h\rangle - \langle f,f\rangle\langle g,h\rangle^2 - \langle g,g\rangle\langle f,h\rangle^2 \\
&\qquad\quad - \langle h,h\rangle\langle f,g\rangle^2 + 2\langle f,g\rangle\langle g,h\rangle\langle h,f\rangle \\
&\qquad = 2\cdot 3\cdot 2 - 2\cdot 1^2 - 3\cdot 1^2 - 2\cdot 1^2 + 2\cdot 1\cdot 1\cdot 1 \;=\; 7.
\end{aligned}$$

Hence (volume $fgh) = \sqrt{7}$.

This (volume $fgh)^2$ is a maximum when $\langle f,g\rangle = \langle g,h\rangle = \langle h,f\rangle = 0$, that is, when the parallelepiped is a cuboid. The volume is zero when any two of f, g and h are colinear, and when all the vectors are coplanar. For 2-dimensional vectors we get

$$\begin{aligned}
&(\text{volume } fgh)^2 \\
&\quad = (f_0^2 + f_1^2)((g_0^2 + g_1^2)(h_0^2 + h_1^2) - (f_0^2 + f_1^2)(g_0h_0 + g_1h_1)^2 \\
&\qquad - (g_0^2 + g_1^2)(f_0h_0 + f_1h_1)^2 - (h_0^2 + h_1^2)(f_0g_0 + f_1g_1)^2 \\
&\qquad + 2(f_0g_0 + f_1g_1)(f_0h_0 + f_1h_1)(g_0h_0 + g_1h_1) \;=\; 0.
\end{aligned}$$

That is, the formula gives a volume of zero for any three coplanar 2-dimensional vectors.

What about the determinant $\det\begin{bmatrix} f_0 & g_0 & h_0 \\ f_1 & g_1 & h_1 \\ f_2 & g_2 & h_2 \end{bmatrix}$? Note that f, g and h now have only three dimensions whereas in the discussion up to now they could have had any number of dimensions. Call the matrix M. As before, we calculate

$$M^TM \;=\; \begin{bmatrix} f_0 & f_1 & f_2 \\ g_0 & g_1 & g_2 \\ h_0 & h_1 & h_2 \end{bmatrix}\begin{bmatrix} f_0 & g_0 & h_0 \\ f_1 & g_1 & h_1 \\ f_2 & g_2 & h_2 \end{bmatrix} \;=\; \begin{bmatrix} \langle f,f\rangle & \langle f,g\rangle & \langle f,h\rangle \\ \langle g,f\rangle & \langle g,g\rangle & \langle g,h\rangle \\ \langle h,f\rangle & \langle h,g\rangle & \langle h,h\rangle \end{bmatrix}.$$

Hence

$$\begin{aligned}
\det M^TM \;&=\; \langle f,f\rangle\langle g,g\rangle\langle h,h\rangle + \langle f,g\rangle\langle g,h\rangle\langle h,f\rangle \\
&\qquad + \langle f,h\rangle\langle h,g\rangle\langle g,f\rangle - \langle f,f\rangle\langle h,g\rangle\langle g,h\rangle \\
&\qquad - \langle g,g\rangle\langle f,h\rangle\langle h,f\rangle - \langle h,h\rangle\langle f,g\rangle\langle g,f\rangle \\
&=\; (\text{volume } fgh)^2 \;=\; (\det M)^2,
\end{aligned}$$

because, as we showed earlier, $\det M^TM = (\det M)^2$. So in three dimensions, the relation between between determinants and inner products is

$$\begin{aligned}
(\det M)^2 \;&=\; \langle f,f\rangle\langle g,g\rangle\langle h,h\rangle + \langle f,g\rangle\langle g,h\rangle\langle h,f\rangle \\
&\quad + \langle f,h\rangle\langle h,g\rangle\langle g,f\rangle - \langle f,f\rangle\langle h,g\rangle\langle g,h\rangle \\
&\quad - \langle g,g\rangle\langle f,h\rangle\langle h,f\rangle - \langle h,h\rangle\langle f,g\rangle\langle g,f\rangle \;=\; (\text{volume } fgh)^2.
\end{aligned}$$

Again, it is interesting to find sets of integer values f_0, f_1, f_2, g_0, g_1, g_2, h_0, h_1, h_2 such that $(f_0g_1h_2+f_1g_2h_0+f_2g_0h_1-f_0g_2h_1-f_1g_0h_2-f_2g_1h_0)^2 = 1$, thus giving a parallelepiped with volume 1. It is easy to generate such sets. We start with a matrix with determinant 1 or -1. An obvious choice is $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Now add an integer multiple of one column to another column, or an integer multiple of one row to another row. For example add 3 times column 1 to column 2 and add 2 times column 1 to column 3: $\begin{bmatrix} 1 & 3 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, then add 2 times row 1 to row 2 and add 3 times row 1 to row 3: $\begin{bmatrix} 1 & 3 & 2 \\ 2 & 7 & 4 \\ 3 & 9 & 7 \end{bmatrix}$.

There are clearly an infinite number of such parallelepipeds with unit volume. The longer the vectors, the thinner the parallelepiped. For example, given the either the matrix $\begin{bmatrix} 101 & 101 & 102 \\ 100 & 101 & 101 \\ 100 & 100 & 101 \end{bmatrix}$, or the matrix $\begin{bmatrix} 101 & 101 & 102 \\ 100 & 101 & 102 \\ 100 & 100 & 101 \end{bmatrix}$ (spot the difference!), both of which have determinant 1, the corresponding parallelepipeds are extraordinarily spindly. Problem: Suppose f, g and h are 4-dimensional, with integer components, and suppose that at least one of f, g and h has a non-zero component in each of the four dimensions. Can the area of the parallelogram formed by them have $(\text{volume } fgh)^2 = 1$? If not, can you find the smallest possible non-zero volume? What about when f, g and h are 5-dimensional vectors?

Inner products. Here is a brief introduction. A *Banach space* is particular type of vector space that, among other things, admits a norm. The norm defines a *distance*. A *Hilbert space*, denoted by $\mathcal{H}$, is a Banach space with an inner product. The inner product defines *angles*. The inner product of two vectors $\langle f, g\rangle$ is linear with respect to its first argument. For all $\lambda_1, \lambda_2 \in \mathbb{C}$, $\langle \lambda_1 f_1 + \lambda_2 f_2, g\rangle = \langle \lambda_1 f_1, g\rangle + \langle \lambda_2 f_2, g\rangle$. It also has Hermitian symmetry $\langle f, g\rangle = \langle g, f\rangle^*$, where * denotes complex conjugation. Moreover, $\langle f, f\rangle \geq 0$ and also $\langle f, f\rangle = 0 \Leftrightarrow f = 0$. Hence $\langle f, f\rangle^{1/2} = \|f\|$ is a norm; that is, for all f, $g \in \mathcal{H}$ and $\lambda \in \mathbb{C}$, $\|f\| \geq 0$, $\|f\| = 0 \Leftrightarrow f = 0$, $\|\lambda f\| = |\lambda|\|f\|$ and $\|f + g\| \leq \|f\| + \|g\|$.

For more information see http://mathworld.wolfram.com/HilbertSpace.html. *Mathworld* is a wonderful Web site for mathematicians. Inner products involving complex numbers are used extensively in signal processing, for example in image compression and audio compression, and also they are of crucial importance in quantum mechanics. In these applications an inner product looks like $\langle f, g\rangle = \sum_{i=0}^{N-1} f_n g_n^*$, or $\langle f, g\rangle = \int_{-\infty}^{+\infty} f(t)g^*(t)dt$, and f and g are generally signals that can be decomposed into components which are complex sinusoids like $e^{i\omega t}$. For more on this try looking up 'Fourier transforms'. For advanced readers, a very up-to-date and thoroughly mathematical treatment of signal processing techniques, including audio processing and MPEG video compression, is *A Wavelet Tour of Signal Processing* by Stephane Mallat, Academic Press, 1999, http://www.hbuk.co.uk/ap/. This amazing book runs to over 600 pages of brilliant maths, and has many illustrations of images being processed.

Dot Product. In Euclidean space $\mathbb{R}^N$, with N finite, we can define an inner product $\langle f, g\rangle$, as follows: $\langle f, g\rangle = \sum_{i=0}^{N-1} f_i g_i$, where f_i and g_i are the ith components of vectors f and g in some orthogonal basis. A set of vectors $\{e_i\}$, $i = 0, 1, 2, \ldots, N-1$, is said to be an orthogonal basis for $\mathbb{R}^N$. When we take inner products on both sides of the above equation we get

$$\langle f, e_j\rangle = \left\langle \sum_{i=0}^{N-1} f_i e_i, e_j \right\rangle = f_j\langle e_i, e_j\rangle$$

since $\langle e_i, e_j\rangle = 0$ if $e_i \neq e_j$. Thus $f_j = \langle f, e_j\rangle / \langle e_j, e_j\rangle$. Hence f can be decomposed as a sum of orthogonal vectors,

$$f = \sum_{i=0}^{N-1} \frac{\langle f, e_i\rangle}{\langle e_i, e_i\rangle} e_i. \tag{iii}$$

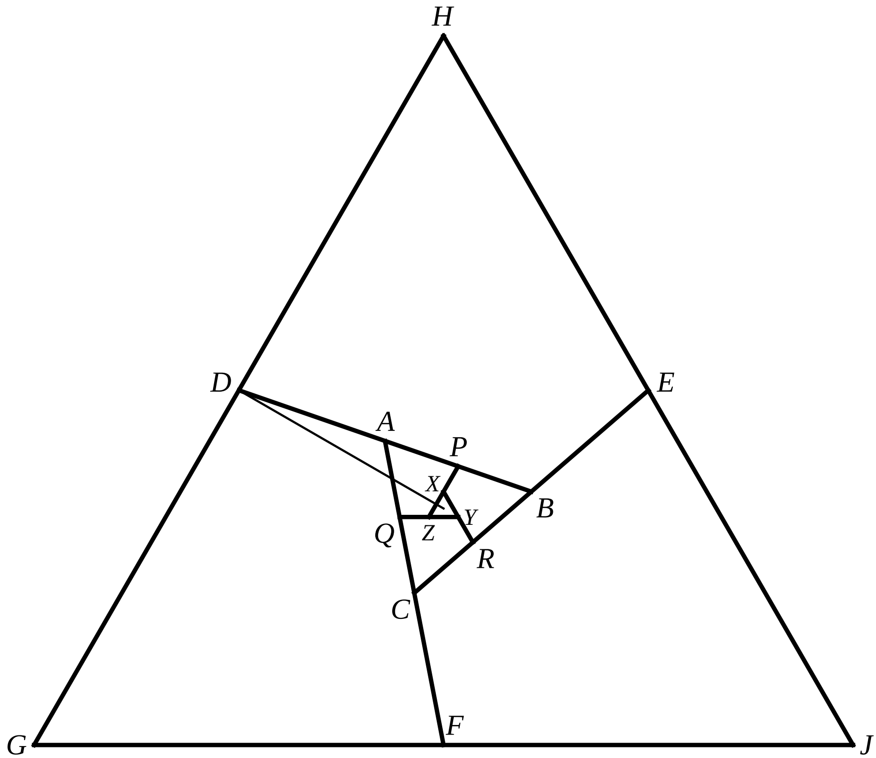
Computing the inner product of each side of (iii) with g yields

$$\begin{aligned}\langle f, g\rangle &= \left\langle \sum_{i=0}^{N-1} \frac{\langle f, e_i\rangle}{\langle e_i, e_i\rangle} e_i, g \right\rangle = \sum_{i=0}^{N-1} \frac{\langle f, e_i\rangle\langle e_i, g\rangle}{\langle e_i, e_i\rangle} \\ &= \sum_{i=0}^{N-1} \frac{f_i\langle e_i, e_i\rangle g_i\langle e_i, e_i\rangle}{\langle e_i, e_i\rangle}.\end{aligned}$$

If we use orthonormal, that is, unit length orthogonal, basis vectors, then $\langle e_i, e_i\rangle = 1$, hence $\langle f, g\rangle = \sum_{i=0}^{N-1} f_i g_i$, which is the dot product formula.

Solution 190.6 – Triangle

The problem asks for a construction of the polygon ABC and everything inside it.



Ken Greatrix

Instead of drawing smaller triangles, draw larger ones. Extend the sides of the triangle alternately clockwise and anticlockwise

In such an infinite cascade, every alternate figure is in the same orientation. It is relatively simple to construct a larger figure, then the smaller ones can be drawn by parallel lines. This proof and the proof for n-sided regular polygons is left for the amusement of the Editor and/or the reader.

Referring to the diagram, above: ABC is a triangle with its sides extended to D, E, F such that $AB = AD$, $BC = BE$, $AC = CF$; P, Q, R are the mid-points of AB, AC, BC with perpendiculars from these points meeting at O (the centre of the figure which I haven't shown because it's a bit cluttered in there already!).

Join OD, OE, OF and then construct triangle GHJ such that GH, HJ, GJ, are perpendicular to OD, OE, OF.

Triangle XYZ is the required triangle inside ABC such that QY is parallel to GJ, ZP with GH, XR with HJ.

I expect a pentagon in the next issue!

ADF

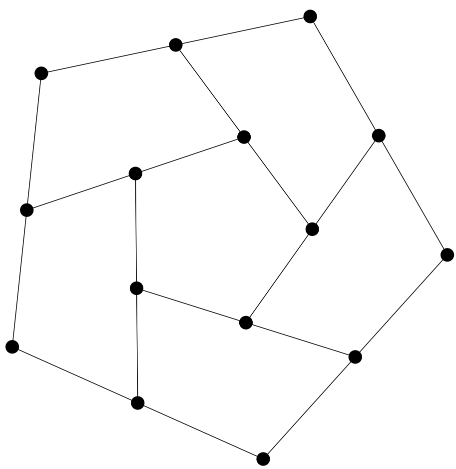
I admit that I didn't see the possibility of starting from the centre and drawing outwards. As Ken has shown, the construction becomes quite easy once you decide to take that sensible viewpoint. In fact, starting with

$$X = (0, \sqrt{3}/3), \;\; Y = (1/2, -\sqrt{3}/6), \;\; Z = (-1/2, -\sqrt{3}/6),$$

one can construct the other points as follows:

$$\begin{gathered} P = 2X - Z, \;\; Q = 2Z - Y, \;\; R = 2Y - X, \\ A = -2R, \;\; B = -2Q, \;\; C = -2P, \\ D = 2A - B, \;\; E = 2B - C, \;\; F = 2C - A, \\ G = -2E, \;\; H = -2F, \;\; I = -2D. \end{gathered}$$

We received similar solutions and observations from **Ted Gore** and **Robin Marks**. Just to prove that it really works, on the right we have drawn the pentagon using MATHEMATICA code supplied by Robin. And mainly because we thought it looked pretty, we have printed part of Robin's construction for the 15-gon on the front cover of this issue of M500.



John Spencer

The Argand diagram provides a general method for solving the problem of finding the coordinates of an inner polygon whose side can be extended to meet the midpoint of the side of a regular n-sided polygon so that the line joining the midpoint to the inner polygon is the length of a side of the inner polygon. From the origin draw n equally spaced 'spokes' of unit length. These represent the original or 'outer' polygon. Label them $\{o_1, o_2, o_3, \ldots, o_n\}$.

Then rotate and shrink the 'outer' polygon to create an 'inner' polygon

$$\{i_1, i_2, i_3, \ldots, i_n\} \;=\; \alpha \exp(i\theta)\{o_1, o_2, o_3, \ldots, o_n\},$$

where $\alpha < 1$ is the length of the 'spokes' of the inner polygon and θ is the angle through which the inner polygon is rotated relative to the original polygon.

For some combination of values of θ and α the line $i_1 i_2$ meets the midpoint of the edge $o_1 o_2$ (i.e. $(o_1 + o_2)/2$) in such a way that

$$i_2 - i_1 \;=\; \frac{o_1 + o_2}{2} - i_2.$$

Substituting for i_1 and i_2,

$$\alpha \exp(i\theta) \;=\; \frac{o_1 + o_2}{2(2o_2 - o_1)},$$

from which the length and orientation of the inner polygon can be read off.

For example, for the square, with $o_1 = 1$ and $o_2 = i$, i.e. in Cartesian coordinates, $(1, 0), (0, 1), (-1, 0), (0, -1)$,

$$\alpha \exp(i\theta) \;=\; \frac{1 + i}{2(2i - 1)} \;=\; \frac{1 - 3i}{10}.$$

So the points of the 'inner' square are located at $(\frac{1}{10}, -\frac{3}{10})$, $(\frac{3}{10}, \frac{1}{10})$, $(-\frac{1}{10}, \frac{3}{10})$, $(-\frac{3}{10}, -\frac{1}{10})$, which is indeed the required figure.

For the pentagon, with $o_1 = 1$ and $o_2 = \exp(\frac{2\pi i}{5}) \approx 0.309 + 0.951i$,

$$\alpha \exp(i\theta) \;=\; \frac{1.309 + 0.951i}{2(-0.382 + 1.902i)} \;=\; 0.174 - 0.379i.$$

So the points of the inner pentagon are at approximately $(0.174, -0.379)$, $(0.414, 0.0483)$, $(0.082, 0.409)$, $(-0.363, 0.204)$ and $(-0.307, -0.283)$.

The number you have dialed is imaginary. Please divide it by i and try again.

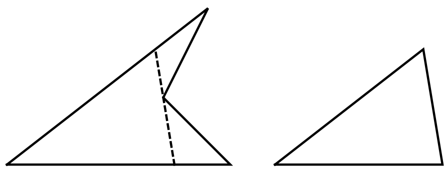
Problem 193.1 – Smallest square

Given a convex quadrilateral Q, with area A and diagonals r, s, show that the smallest square containing Q has area at least

$$\frac{r^2s^2 - 4A^2}{r^2 + s^2 - 4A}.$$

Problem 193.2 – Concave to convex

Start with a non-convex quadrilateral. By removing some bits of it you can end up with a convex polygon. What is the minimum area you have to remove?



Problem 193.3 – Thirteen tarts

Tony Forbes

There are thirteen tarts. All weigh the same, with one exception. Either (i) devise a strategy involving three weighings to determine the odd tart and whether it is lighter or heavier than the others; or (ii) prove that (i) is impossible.

As usual, a *weighing* means deciding whether one set of tarts is heavier than, lighter than, or the same weight as another set of tarts. Recall from M500 **191** that there exists a strategy for twelve tarts. The reason why I ask about 13 is that the number of different sets of observations you can obtain in three weighings is 27, which, unless someone can show otherwise, looks sufficient to resolve the 26 cases pertaining to 13 tarts.

Ron Potkin has provided us with a most interesting solution to the general problem of resolving $(3^n - 3)/2$ tarts in n weighings—see his letter on page 25—but here I am only after a simple proof one way or the other about $(3^n - 1)/2$ tarts in n weighings.

Problem 193.4 – Factorial inequality

Show that for positive integer n,

$$n! \;\le\; \left(\frac{n(n+1)^3}{8}\right)^{n/4}.$$

Finite and discontinuous

Sebastian Hayes

Dick Boardman (M500 **190** p. 26) refers to my article 'Why does calculus work?' and, rather surprisingly, gives substantially the same answer as I arrived at—though I hasten to add that the much more controversial suggestions made in this letter / article are wholly my responsibility.

Some thirty years ago, despite detesting the subject at school, I decided I'd have to study mathematics. My reasons? I wanted to find out about 'reality', what was real and what wasn't, and I reckoned mathematics ought to be able to help me there.

On being exposed to calculus for the first time, I noted at once that the assumptions required for the mathematical treatment—in particular 'infinite divisibility' and 'continuity'—were completely unrealistic. This worried me a good deal—for how could a symbolic system that didn't square with real-life conditions come up with the right answers? But no mathematician I met was even remotely interested in the issue, let alone capable of resolving it; the only answer I ever got from OU tutors or professional mathematicians I wrote to was substantially the answer given by d'Alembert to one of his pupils, 'Allez de l'avant, la foi vous viendra' ('Keep going, you'll end up as a believer').

'La foi' never came. For the greater part of my adult life I have thus been the mathematical equivalent of a churchgoer who mumbles his way through the Apostles' Creed when invited to do so by the vicar but does not in fact believe that Christ was the Son of God—does not because he *cannot*. I did at one time consider trying to invent a 'discontinuous' calculus but eventually gave up the attempt as being both beyond my powers and, in the final analysis, unnecessary; in the vast majority of cases dx is small and so calculus is a good enough approximation and, when it is not, these days we can slog it out with computers.

The conceptual problem remains, however, though few indeed are prepared to confront it. Ask a pure mathematician about any question concerning 'reality' and he or she will refer you at once to the Philosophy Department. But philosophers are a declining species these days and those there are have a severe inferiority complex with regard to mathematics. Far from daring to pronounce on issues mathematical, they take the view that 'if the mathematicians say so, it must be true'. What they (the philosophers) fail to realize is that 'true' for a mathematician these days simply means 'consistent with a particular set of assumptions' whereas 'true' in the philosophic sense means, or should and used to mean, 'is the case'. The precision and cleanliness of modern mathematical systems have been bought at a price—mathematics has turned its back on the real world.

Thus one reads again and again in 'philosophical' works that Zeno's paradoxes of motion have been 'decisively refuted by the methods of modern mathematics'. But the question to be addressed is not whether such and such mathematical assumptions permit, or on the contrary rule out, such and such consequences but whether the assumptions made are plausible and realistic in the first place. Zeno hardly needed the techniques of transfinite set theory to determine that Achilles overtakes the tortoise—simple observation and Greek mathematics sufficed. Zeno's parables were designed to make people think; he wanted people to confront the *philosophic* issues of the nature of space, time and motion—and they are not empty mathematical postulates but postulated realities. Mathematics cannot arrogate to itself complete freedom to invent whatever it wants and at the same time claim to be a guide to what goes on outside the printed page—this is what is commonly known as having your cake and eating it.

Nonetheless, so great is the prestige of mathematics and the corresponding fear that mathematicians inspire amongst laymen that the most extravagant propositions are taken seriously just so long as they are dressed up in symbolic garb whilst inherently plausible ideas never get a hearing if they run counter to ingrained mathematical habits of thought. Cantor's weird cogitations about the nature of the continuum have been the subject of much learned debate but what one might call the *discontinuum hypothesis* is never discussed at all.

I take the commonsense view that the bulk of our knowledge of the world comes from sense-data and, far from being an exception to the rule, mathematics is—or rather *was* up to about the middle of the nineteenth century—a brilliantly successful example of an empirically based discipline. Very early on in life we perceive what is around us as a plurality of discrete items which can be gathered together into groups—*my* toys, *your* brothers—or on the contrary separated out again. Such *experiences*—not *a priori* assumptions or 'intuitions'—gave rise at a certain stage in human history to the appropriately named *natural* numbers and to the basic operations of arithmetic. 'Continuity' is not a numerical notion but a geometric one and it was only at the end of the nineteenth century and with the greatest difficulty that it was given adequate numerical expression. There is no doubt in my eyes as to which of the two disciplines, geometry and arithmetic, is the more fundamental; it is arithmetic, and by arithmetic I mean the manipulation of whole numbers.

As for 'infinity', it is hardly something within our experience and ranks at best as a wholly derived and essentially negative concept. Certainly it is not a number and not a quantity, it is more a direction than anything else meaning roughly 'Keep on going as long as you wish'. We could do mathematics without it—I have actually made the experiment—and just use the arrow sign. When pointing to the right $\rightarrow$ instructs us to make something as large as we see fit, and when pointing to the left $\leftarrow$ instructs us to make something as small as we see fit: $\lim_{n\rightarrow} f(n) = 2$, $\lim_{\leftarrow n} f(n) = 0$. Different colours or other distinguishing features could be used to indicate positive and negative movements to a limit. But of course there is no harm in using the infinity sign as long as one suspends disbelief. The trouble is, in my experience, mathematicians through force of habit do tend to believe that there is such a thing as infinity: I shall never forget the look of pure horror on the face of the Oxford mathematician as he said, pointing at me, 'Good God, this fellow does not even believe in ordinary infinity, let alone the transfinite!'

So, if as I claim, our basic apprehension of life and the world around us is finite and discontinuous, why not state this in so many words and see where it gets us? So here goes, the grand *discontinuity postulate*:

All processes/phenomena are discontinuous.

Because of the sweeping scope of the claim and complications such as the Heisenberg Uncertainty Principle, such a postulate will presumably never be proved in its entirety. However, it is a postulate which is in accord with our instincts and sense-impressions and is supported by an enormous amount of scientific data. All around me I see nothing but collections of discrete entities, stones, trees, humans, &c. Fluids, it is true, appear to be continuous but under the microscope reveal a molecular structure. The emission of 'light', long thought to be continuous, is in fact not so: indeed all energy exchanges are quantized, thus discontinuous, and since most of what goes on in the physical world is some sort of an energy exchange this covers rather a large field.

Furthermore, we have a strong psychological conviction of the postulate's truth with respect to ourselves and our inner processes. If we, as humans, really lived and felt 'continuously' all this anguish about growing old would have no *raison d'être*, there would be no sense of 'time passing' and little fear of personal extinction. The higher religions, Christianity and Buddhism in particular, may in fact be described as largely unsuccessful attempts to shift human sensibility from a discontinuous to a continuous mode. Only the saint feels with and for humanity as a whole, the rest of us are islands and we do ask for whom the bell tolls.*

Not only do we feel in spasms but we think discontinuously as well. The thinking process is *sequential*, that is, we think step by step, not holistically, 'all at once'. It is for this very reason that Benjamin Peirce made the excellent observation that, though we can perhaps envisage a world without 'space', we cannot envisage one without time.

Where does the discontinuity postulate take us? Well, if we are to rule out infinite regress, which the postulate is framed precisely to avoid, there must seemingly be certain 'things' which are entire, not divisible (i.e. if you like 'continuous' though the term is misleading), which is why I used the plural terms *processes* and *phenomena* in formulating the principle. What are these irreducibles? Can anything be said about them? One way of proceeding is to hold on to the notion of the *event* which, at least for the purposes of this brief overview, I take to be 'intuitively clear'. It is possible to view apparently solid objects in terms of 'mini-events' but it is not really possible to view an event—say a 'blow' or a 'bang'—in terms of mini-objects. However, clearly not all events are irreducible so we must introduce the notion of *ultimate event* which is an event which cannot be further decomposed. Also, since an event, ultimate or not, seemingly must take place somewhere, we must needs have a location of some sort, I call it the *Locality*. An ultimate event may be defined as an event which has occurrence *at a single point* (or within a single 'square') of the Locality. Objects can now be defined as self-repeating patterns of ultimate events: they possess *persistence*, a feature they share with human beings.

As stated in my article 'Why does calculus work?' (M500 **185**), the assumption of discontinuity resolves all of Zeno's paradoxes. The door does not have to traverse an infinite number of positions before it slams shut in about two seconds, and Achilles is always able to overtake the tortoise since the latter is only so many grid-points ahead of him. The arrow moves by jerks towards its target even though we do not see the jerks: it is *here*, then it is *there*. Motion turns out to be nothing more than 'being at different places at different times' (Russell) and a collision is an attempt by two competing event-clusters to share the same grid-points on the Locality.

How does the postulate tally with the assumptions of modern physics? Within the limits of the terms used so far, there exist only sequences and clusters of events spread out over the Locality, some self-repeating, some not. There is no need to believe in wave–particle duality as such because neither waves nor particles are absolutely fundamental items. An energy disturbance is not propagated continuously across the Locality any more than a stream of particles is projected continuously through a narrow slit, but in the case of so-called wave energy propagation we are dealing with a much finer mesh. Conceptually, I do not think this makes too much difference—the real problem with quantum mechanics is whether it violates causality, that is, whether there can be 'space–time hopping' which seemingly there can be. The discontinuity postulate does not have anything specifically to say on this point—further postulates are required—but it is surely easier to conceive that 'effects' can be widely separated from their 'causes' in a discontinuous model. (Causally related events can, in quantum mechanics, be light years apart without there necessarily being any intermediate events connecting them.) Indeed, one could argue that the very language of causality implies a discontinuous model since otherwise it would not be meaningful to distinguish a particular event, named the 'cause', from a different event, named the 'effect'. If processes were truly continuous there would be an imperceptible grading from one state to another—something that is rarely if ever observed and is, for me at any rate, difficult to even conceive of. For me change, like enlightenment, either occurs at once or does not occur at all.

So far, so good. Although the assumption of discontinuity fits tolerably well with quantum mechanics, at least on a cursory view, it clashes head on with relativity. Why is this? According to the discontinuity postulate and the innocuous sounding assumptions made about ultimate events, it follows that every physical process is composed of a specific (thus finite) number of ultimate events, not a variable (infinite) number. There is thus in principle an *event-number* associated with every terminating physical process though naturally this number (an integer) will not normally be knowable. This is the *event-chain number hypothesis*. Now, I take it as absurd to suppose that a *bona fide event*, say an explosion, takes place in one person's coordinate frame and not in that of someone else—a supernova explosion or a heart beat either has occurrence on the Locality or it doesn't, full stop. Coordinate systems are mathematical conveniences, not things that 'really exist'—indeed, it was precisely this consideration which led Einstein to attempt a generalization of his original theory of relativity.

But the event-chain number hypothesis knocks out the so-called *twins paradox*. If Jack the Nimble, travelling close to the speed of light, ever returned to Earth—a big 'if'—his heart would by that time (*sic*) have thumped a specific number of times and this 'heart-beat number' would have nothing whatever to do with his state of motion or that of his brother. And all his other physical processes, being a discontinuous stream of physical events, would have event-chain numbers attached to them. So, if aging depends on how many times your heart has beaten and suchlike matters, Jack the Nimble wouldn't be a day younger—contrary to what special relativity states and certain experiments apparently confirm.

The above does not in itself dispose of the discontinuity postulate in my eyes but the discrepancy is certainly troublesome. Indeed, I must confess that, contrary to the majority of people, I find relativity more difficult to swallow conceptually than quantum mechanics because it is more of a threat to the idea of an objective reality.

Can I suggest any experiments that would confirm the discontinuity postulate? Well, I cannot be expected to predict everything that is observed on the basis of just a single postulate and one or two related assumptions, but having adopted at the outset an empiricist approach I don't intend to duck this question completely. If motion is in reality a succession of stills, rapid motion presumably means that a greater number of lattice-points are missed out—this is at any rate one way of envisaging rapid motion as compared with slow motion. This in turn means that a rapidly moving projectile should be able to pass clean through an obstruction that is not too thick—for the two event-chains would not need to clash in a desperate attempt to share the same intermediary 'space'. Has such a phenomenon been observed? I have read that a neutrino can pass through the Earth without leaving much trace if any. On the other hand, according to my principles, a ray of light, simply because it travels so quickly, should, just like a neutrino, be able to pass through any reasonably thin obstacle while this is by no means the case.

Sheldon Attridge (M500 **188**, Letters, page 25) makes what is perhaps a more serious, because more basic, objection to the discontinuity postulate. He writes

> If there really are 'holes' in space which a closing door does not traverse—then whereabouts are these holes located? To paraphrase Kant: it is impossible for the human mind to conceive of 'no-space', it is an a priori concept.

Well, for me there is no great conceptual problem. I envisage the Locality as a sort of vast Solitaire board where each depression can receive an ultimate event. In between depressions there is not enough room (i.e. space) for an event, that's all—what's difficult about that? Physicists today do seemingly manage to conceive of something which is 'no-space' or at least 'pre-space' since this is what the quantum vacuum is. According to certain versions of the 'great beginning', e.g. that of the so-called Brussels School, the universe itself including space–time is a runaway disturbance of this much more basic entity, the quantum vacuum. So, to take up Sheldon Attridge's query, 'Where is the quantum vacuum situated?'

Is there any way of conceiving reality as both continuous and discontinuous without contradicting oneself? Perhaps there is on the analogy with quantum mechanics. According to the usual interpretation of quantum mechanics, reality is two-tiered: there is a qualitative difference between what pertains *prior* to an act of measurement and what happens subsequently. Thus, if we have a 'photon' confined to a box, it is, prior to an act of measurement 'all over the place' and does not have a precise location. However, an act of measurement which inevitably involves interaction with the object of the experiment pins the latter down to a specific position.

It would be possible to envisage the Locality, empty of ultimate events, as being 'all of a piece', i.e. continuous if you like. Nonetheless, an ultimate event, when and if it has occurrence, is precisely located: it carves out its own spatial position as it takes place. This approach is very similar to that taken by Bohm who makes a qualitative distinction between what he calls the Implicate Order and the Explicate Order. According to Bohm, science and mathematics deal only with what is measurable, and is therefore part of the Explicate Order. But Bohm, originally a more or less orthodox quantum mechanical physicist, became at one point in his life profoundly impressed by the views of the Indian mystic Krishnamurti and considered that we have an intuitive awareness of the more basic underlying reality which he called the Implicate Order and which really is continuous.

This is a feasible approach but it must be stressed that it is not the world-view that is implicit in calculus and analysis generally. As I understand quantum mechanics, a 'photon' inside a box does not have an 'infinity' of positions which it occupies simultaneously prior to being observed: it has strictly *no* fixed position, or none smaller than the total area of the box. Now in analysis the definition of continuity** depends on the limit concept and on the surely rather overdone image of 'a line made up of points', the only difference being that we now have an unlimited amount of points to any line. In fact what we obtain is not continuity as the word is generally understood but rather an 'infinite discontinuity'—for if a line or surface really were entire, of a piece, i.e. continuous, it would not be composed of points at all, it would be a single indissoluble entity. But then there would be very little that could be said about it. For mathematical purposes we need to have it both ways but that should not lead us to assume that this is how things really are.

---

* 'No man is an island intire of itselfe, every man is a piece of the Continent, a part of the maine; if a clod bee washed away by the Sea, Europe is lesse, as well as if a promontorie were, as well as if a mannor of thy friends or of thine owne were; any man's death diminishes me, because I am involved in Mankinde; and therefore never send to know for whom the bell tolls; it tolls for thee.' — John Donne, *Devotions XVII*

** 'A function is continuous at a point ξ if and only if $f(x) \to f(\xi)$ as $x \to \xi$. A function is continuous on an open interval I if and only if it is continuous at each point of I.' — K. Binmore, *Mathematical Analysis*

Solution 190.7 – Four roots

Show that if $a^3 > 4b > 0$, the polynomial $x^4 - ax^3 + bx - b^2/a^2$ has four real roots which are in a harmonic ratio.

David Porter

I first tried to do this assuming that the four roots being in harmonic ratio meant that they formed a harmonic sequence; i.e. their reciprocals formed an arithmetic sequence. Eventually this lead me to the conclusion that the proposition was false so I belatedly looked up the definition in my *Dictionary of Mathematics* and discovered that it really meant that the product of two of the roots was minus the product of the other two.

The result then falls out because this particular form of quartic can be factorized into two quadratics thus:

$$x^4 - ax^3 + bx - b^2/a^2 \;=\; (x^2 - b/a)(x^2 - ax + b/a).$$

So the four roots are

$$\sqrt{\frac{b}{a}},\ -\sqrt{\frac{b}{a}},\ \frac{a - \sqrt{a^2 - 4b/a}}{2} \text{ and } \frac{a + \sqrt{a^2 - 4b/a}}{2},$$

and since a and b are real (implied by $a^3 > 0$ and $4b > 0$) and $a^3 > 4b > 0$, all the roots are real. Furthermore,

$$(+\sqrt{b/a})(-\sqrt{b/a}) \;=\; -\frac{b}{a}$$

and

$$\frac{a - \sqrt{a^2 - 4b/a}}{2} \cdot \frac{a + \sqrt{a^2 - 4b/a}}{2} \;=\; \frac{a^2 - (a^2 - 4b/a)}{4} \;=\; \frac{b}{a};$$

so these roots are in an harmonic ratio.

Dick Boardman

First the harmonic ratio. Choose four points on a line in order A, C, B, D, where C is between A and B and is said to divide AB internally in the ratio AC/BC, and D is not between A and B and is said to divide AB externally in the ratio AD/BD. The ratio of these ratios is called the *cross ratio*, $\{A, B, C, D\}$. Let A be a distance a from some arbitrary origin, and similarly for B, C and D. Then the cross ratio is $\dfrac{(c-a)/(c-b)}{(d-a)/(d-b)}$. Note that where C is between A and B and D is not, the number is negative. In the special case where the cross ratio is -1, $\{A, B, C, D\}$ are said to be in harmonic ratio.

To find out the importance of all this we have to go back more than 500 years. In the 15th century, artists could not draw realistic three-dimensional pictures. They knew that things looked smaller, the further they were away, but they did not know by how much. Then someone invented a device which was equivalent to viewing the scene through a fixed eyehole and through a sheet of glass and painting on the glass what he saw through it. In the hands of a few artistic geniuses, this technique produced pictures that were stunning in realism and in detail. To the modern eye, they look like high quality colour photographs, but of course they pre-date colour photography by 400 years. These pictures attracted mathematicians and the collaboration between them was very fruitful. Artist got the rules of perspective that they have used ever since and mathematics got a new form of geometry, projective geometry. The well-known artist David Hockney thinks that, before photography, many artists used similar optical devices to keep their pictures in proportion. (Reference: D. Hockney, *Secret Knowledge.*)

Suppose in the scene being viewed there were four points in a straight line. These four points would be painted on the glass as four points in a line. One of the key theorems in projective geometry is that the cross ratios of the two sets of points are the same; that is, cross ratio is invariant under a projection.

Now for the problem itself. The polynomial can be factored into two quadratics and the roots of these quadratics are the roots of the quartic. They are

$$\frac{a^{3/2}-\sqrt{a^3-4b}}{2\sqrt{a}},\ \frac{a^{3/2}+\sqrt{a^3-4b}}{2\sqrt{a}},\ -\sqrt{\frac{b}{a}} \text{ and } \sqrt{\frac{b}{a}}.$$

Label these values c, d, a, b and put them into the formula for cross ratio. This gives (after some algebra) a value -1 so that the roots are in a harmonic ratio. This seems easy enough here but I would hate to be given that question in an exam.

Problem 193.5 – Dissect a triangle

Dick Boardman

Dissect an equilateral triangle into three triangles such that

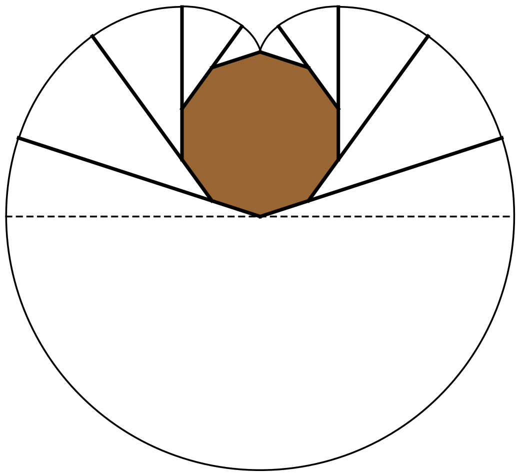
(1) Their areas are in the ratio 3:3:2.

(2) All the sides of all the triangles are integers.

'Damien Hirst tends to use everyday objects such as a shark in formaldehyde.'—Fashion commentator, R4 [EK]

Solution 190.3 – Goat

A field contains a barn occupying a space in the form of a regular polygon with $2n$ sides of length 1 metre. A goat is tethered to a corner of the barn by a rope of length n metres. What is the area of grass that the goat can reach?



Basil Thompson

It is possible to calculate individual cases, $7\pi/2$ for four sides, $23\pi/3$ for six sides, $27\pi/2$ for eight sides, 21π for ten sides, ..., but what we want is a general formula.

The goat's rope wraps around the polygon on both sides to the point opposite to where it is tethered. The area of the sector grazed with radius n is given by

$$\pi n^2 \frac{\pi + \pi/n}{2\pi} = \frac{\pi}{2} n^2 \left(1 + \frac{1}{n}\right).$$

The area of two sectors with radius $n-1$ is

$$2\pi(n-1)^2 \frac{\pi/n}{2\pi} = \frac{\pi}{n}(n-1)^2,$$

the area of two sectors with radius $n-2$ is

$$2\pi(n-2)^2 \frac{\pi/n}{2\pi} = \frac{\pi}{n}(n-2)^2,$$

and so on. Hence the total area is

$$\frac{\pi}{2} n^2 \left(1 + \frac{1}{n}\right) + \frac{\pi}{n}(n-1)^2 + \frac{\pi}{n}(n-2)^2 + \cdots + \frac{\pi}{n}$$

$$= \frac{\pi}{2} n(n+1) + \frac{\pi}{n} \sum_{k=1}^{n-1} k^2.$$

Using the formula $m(m+1)(2m+1)/6$ for the sum of the first m squares, this simplifies to

$$\frac{\pi}{2}n(n+1) + \frac{\pi}{6}(n-1)(2n-1) \;=\; \frac{\pi}{6}(5n^2+1).$$

Checking this formula against the values quoted at the start shows full agreement. However, I am not happy with the statement that the ratio (goat area) / (barn radius) tends to $5\pi^3/6$ as $n \to \infty$. The ratio must have an n in it and it will go to infinity with n.

A more interesting ratio is (goat area) / (barn area). It is easily seen that for large n the area of the polygon is of order n^2/π. Hence as $n \to \infty$ the ratio tends to $5\pi^2/6$.

ADF writes—It looks like ADF has goofed. What he possibly meant was that the goat area divided by the *square of the* barn radius tends to $5\pi^3/6$. This is indeed the case, as you can verify from Basil's analysis.

Also I remember asking for a simple proof that the area is $5\pi^3/6$ when the barn is a circle of radius 1 metre. **John Spencer** showed that it can be tackled as a traditional A-level integration problem. You let the angle of the rope increase by $d\theta$, work out the corresponding increase in the area grazed, and integrate. The required answer comes out quite easily. However, this is not necessary; as we have seen, we can actually dispense with calculus altogether!

Problem 193.6 – Fair coin

Tony Forbes

I have always been puzzled as to why authors of probability / statistics text books often and persistently insist that the coins in their examples be fair. Then one day I actually looked at a typical British 2p piece and it occurred to me that if the fairness is supposed to be strict, this attribute cannot possibly apply to any of the coins that make up the world's currencies.

I understand that casino operators are obliged by law to have their various randomizing devices tested for bias. So I ask: How would you estimate the bias to heads or tails in a freshly minted 2p coin? Approximately how many tosses would you need? My gut feeling is that the answer is many—10^{12}, 10^{15}, whatever—but I might be wrong.

'Pi, the number which represents the circumference of a circle divided by its radius.'—A History of Human Folly, BBC R4, 4 June 2003. [**JRH**]

Letters to the Editor

Cyclones

Dear Tony,

Re: M500 **188**, page 18, in which you ask why dust at the bottom of a tall cylindrical container goes into a rapid swirling motion whilst it is being vacuum-cleaned.

An interesting experiment; ask a fluid dynamics expert about turbulent flow. Had you done this experiment a few years ago you would have been rich and James Dyson could now be editing M500!

Ken Greatrix

ADF replies—Of course! As owner of one of his *root cyclone* devices I should have realized immediately what was going on. In fact it is even more amazing. Something which starts as a whirlwind inside the Dyson becomes a linear flow through the length of the hose and, in turn, induces another whirlwind inside the spaghetti jar. *Interesting Question*: Are the angular velocities of the two cyclones related?

Re: Amazing object

Dear Tony,

I cannot agree with the sentiments in the letter of Ralph Hancock in M500 **192** (page 27), where he says, '... the problem becomes trivial because an ordinary sphere will do the job and any [fool] can calculate its volume.'

The first person to calculate the volume of a sphere in terms of simpler solids was Archimedes. The people before him were not fools, merely not as good as one of the greatest geniuses of all time.

Dick Boardman

Roots

Dear Tony,

Just a comment on Problem 190.2 – Nested roots. You can use the well known result that $+\sqrt{x} = -\sqrt{x}$, proved as follows:

$$y = \sqrt{x} = \sqrt{(-1)^2 x} = (-1)\sqrt{x} = -\sqrt{x},$$

which gives the (wrong) answer of 3.

John Bull

Twelve tarts

Dear Tony,

I thought you might be interested in the following. I read about it in *The Woodworker* magazine in around 1960 but it first appeared in 1945. Martin Gardner's *Sixth Book of Mathematical Games* contains a good description. Warning: I have described this to friends on a number of occasions but few managed to stay awake until the end. But here goes anyway.

Row 1: Write down the numbers 1 to 12 in a row. Row 2: Underneath write the ternary value of row 1. Row 4: Enter the value of 222 less the value in row 2. Imagine the face of a clock with the digits 0, 1, 2. Going clockwise, we have 01, 12 and 20; anticlockwise, we have 02, 21, 10. Row 3: Look for the first change in digit in row 2. If it's clockwise, write C; otherwise write A. For example, 010 is C and 110 is A. Row 5: Repeat the procedure in row 3, using row 4. You should obtain the following table:

Row 1	1	2	3	4	5	6	7	8	9	10	11	12
Row 2	001	002	010	011	012	020	021	022	100	101	102	110
Row 3	C	A	C	C	C	A	A	A	A	A	A	A
Row 4	221	220	212	211	210	202	201	200	122	121	120	112
Row 5	A	C	A	A	A	C	C	C	C	C	C	C

Using Rows 2 and 4, write down the numbers containing 0 in the first digit of the number and separate them into clockwise and anticlockwise. Repeat for the second and third digits:

	Clockwise	Anticlockwise
First weighing	1, 3, 4, 5	2, 6, 7, 8
Second weighing	1, 6, 7, 8	2, 9, 10, 11
Third weighing	2, 3, 8, 11	5, 6, 9, 12

Using your description of the left-hand pan, 0 equates to down, 2 is up and 1 is balanced. Also, we will recognize C as heavy and A as light. So, if the three weighings are carried out and you obtain 010, then 3 is heavy and if you obtain 101 then 10 is light.

It can be seen that the maximum number of tarts can be determined by $(3^W - 3)/2$, where W is the number of weighings. So the heavy or light tart among 39 could be found with 4 weighings or among 120 with 5 weighings and so on.

With practice the table is unnecessary. Thus, if the result is 201, this must be tart 7 and we can see that 201 is C, so 7 is heavy.

Ron Potkin

Message on a jar of mincemeat: 'The contents are sufficient for a pie for six persons or 12 small tarts.' [EK]

Special offer

Dear Tony,

Whilst on holiday in the far northern reaches of Europe, I happened to come across the following sign:

Special offer: Knitwear from
Dale of Norway ÷20 − 40% off

Being puzzled by the apparent mathematical nonsense in the last line, I showed the advertisement to a Norwegian lady whom I know locally. She said that in Norway the sign that the English use for division means subtraction, and the sign that the Norwegians use for division is ':'; that is, two dots with no line between. A horizontal line with no dots, '−', is a hyphen; '+' for addition, and '×' for multiplication are the same in Norway and England.

I said that I found that very hard to believe. Surely the four principal arithmetical signs must be international. However, she insisted that she had been to school in Norway until she was 18, and that was how things were there. She said that she was not saying that it was like that in Sweden or Finland or anywhere else, but that was definitely how it is in Norway.

So to her the advertisement makes total sense. Read it as 'Minus 20 to 40 per cent off', though she commented that the word 'off' was grammatically wrong if not a tautology. The sign should read, she says, 'Special offer: Knitwear from Dale of Norway ÷20% − 40%'.

I know that the Europeans swap commas and full stops for marking thousands and decimal points, but this division/subtraction confusion is quite new to me. I don't know any more Norwegians to ask, but I think someone in M500 must be able to comment.

Colin Davies

What's next?

Dear Tony,

I haven't had a go at the problem [191.4 – What's next?] as such but I would like to make the following point.

Wittgenstein once poured scorn on the concept of there being a 'right answer' to problems involving guessing the next number in a sequence of numbers, and indeed on the whole concept of there being a 'correct answer' in any endeavour.

For instance, take the sequence 1, 4, 9, 16, ...

One would hope that the next term would be 25, using the familiar squares. However, an equally valid answer would be '32', where the formula for the nth term is given by $n^2 + [n!/2^{9-n}]$, where [ ] means take the integer part (my example). He then argued that the same sort of problem could occur with a sequence of any length.

Out of interest, how many different answers (with formula) can people find for this problem (or the harder sequence given in problem 191.4)? I would be particularly interested in finding a simpler alternative, as I can't remember Wittgenstein's example, which was remarkably simple.

Cheers,

Sheldon Attridge

Forty-two revisited

Eddie Kent

In late May Tony Forbes and I enjoyed a meeting of the Lewis Carroll Society, where David Singmaster gave an entertaining talk about 'some of the 42 ways in which Carroll was involved in recreational mathematics', and described topics 'where his work was or may have been original, where the history is exceptionally interesting, or where there are unsolved problems'.

This reminded me that I once wrote about Lewis Carroll and the number 42 in M500. After some searching I found it, in no. **131**.

The impetus was a short piece in *The Times* in 1992 announcing that the dimensions for a boat to take part in the America's Cup had been changed, from the old 12 metre rule that had been standard since World War II, to a new, tighter formula. This had become necessary because boatmakers were beginning to exploit loopholes in the old conditions and producing some curious structures.

In fact *The Times* got the formula wrong and a student had to correct it for us. It now looks like

$$\frac{L + 1.25 + \sqrt{S} - 9.8 \times \sqrt[3]{D}}{0.388} \;=\; 42,$$

where L is the length of the boat at the waterline in feet, S is the sail area in square feet and D is displacement in pounds (constant times volume displaced in cubic feet). The potential speed of a boat is proportional to the square root of its waterline length, while its ability to attain that speed depends on its area of sail. So the longer you can make the deck of the boat in relation to its length in the water, the more sail you can cram on without increasing L.

The Times went on 'If you wonder where the 42 came from, it is rumoured that one of the committee members, Derek Clark, is a fan of *The Hitch Hiker's Guide*'. Our correspondent pointed out that 42 is arbitrary, and was probably chosen as about one metre more than the old 12 metre rule when measured in feet.

It took very little effort to discover that Douglas Adams, the author of *The Hitch Hiker's Guide to the Galaxy* in which 42 plays an important part, was an admirer of Lewis Carroll. And Carroll was strangely fascinated by the number. In *The Hunting of the Snark* the Baker arrived with 42 boxes, which were all left behind on the beach. Henry Holiday's illustration for the first edition shows the pathetic pile of boxes, some clearly numbered. The box next to box 42 has the number 27, which is 3 cubed and thus the number of cells in a magic cube. The magic constant of the smallest such cube is 42.

Nor is this all. Lewis Carroll was 42 when he began writing *The Snark*. The poem *Phantasmagoria* was written while he was in his thirties yet he claimed to be 42. The Baker skipped forty years in telling his story, having described his parents as honest, though poor, so he was probably 42. The King says, in Chapter 12 of *Alice's Adventure in Wonderland*: 'Rule Forty-two. All persons more than a mile high to leave the court'. The preface to *The Snark* mentions yet another Rule 42.

This rule is of some importance. The Bellman had the bowsprit removed for revarnishing once or twice a week and when it came back the crew couldn't remember which end of the ship it belonged to. There was no point in asking the helmsman because of Rule 42 of the Naval Code, 'No one shall speak to the Man at the Helm'. This rule had been completed by the Bellman with the words 'and the Man at the Helm shall speak to no one'. So he was unable to tell them when they got it wrong, as they did from time to time, and the bowsprit got mixed with the rudder sometimes, as well it might, and no further sailing could be done till next varnishing day.

Where does 42 come from? Carroll was born in 1832 so that doesn't help. There might be a clue in the Admiralty Code: did he know of Rule 42, or did his liking for the number cause him to look it up? Or did he just make it all up? These are questions that might repay some serious research so I leave them for the consideration of the committed reader.

As a final thought, a sailor once told me that one of the more useful naval rules is Rule 13a, which insists that when overtaking a slower vessel you should avoid running into it. One hopes that the author of this rule received his just promotion.

References, in addition to those cited in the text: *The Times* 2.5.1992; EK M500 **129** 16; Stuart Cresswell M500 **131** 12; EK M500 **131** 13; His Honour Commander Derek Inman, RN Rtd, QC (personal communication).

Jäpperivokki

Colin Davies

Rillikki oli, ja lipiäset toopeet
Pyörivät ja kaksoisivat vaapeessa.
Ihan mimsiä olivat porokroopit
Ja muumiraatit rotkosesta pois.

"Varo Jäpperivokki, poikani!
Purevat leuat, tarttuvat kynnet!
Varo Jupjuplintu ja karta
Rumionen Panterisieppaja!"

Hän otti vorpaalen miekansa käteen.
Kauan etsii mankkomaa vihollista.
Siis lepäsi Tumtumpuun alla
Ja hetkeks seisoi miettien.

Ja uppilaisess tuumimassaan
Jäpperivokki, liekehtivine silmineen,
Tuli vippiläen tulken metsän läpi
Ja pulppui tulessaan!

Ykskaks ykskaks! Ja ihan läpi
Vorpaali veitsi puukkoili.
Jätti sen kuolleena, ja päänsä kanssa
Hän paloi kallumpain.

"Ja oletko tappanut Jäpperivoken?"
"Tule syyliini säteilevä poikani!"
Hei räpjyysi päivä! Kaluu! Kalei!
Hihitti ilossaan

Rillikki oli, ja lipiäset toopeet
Pyörivät ja kaksoisivat vaapeessa.
Ihan mimsiä olivat porokroopit
Ja muumiraatit rotkosesta pois.

Problem 193.7 – Binomial coefficients

ADF

Which binomial coefficients $\binom{2n}{n} = \dfrac{(2n)!}{(n!)^2}$ are not divisible by the square of an odd prime?

Contents

M500 193 – August 2003