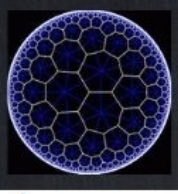
Blackboard

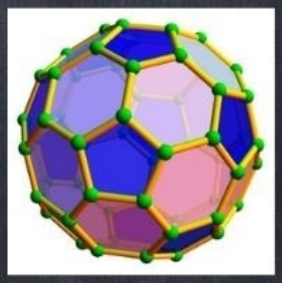


Issue 4

MTA (I)

$$\sum_n \frac{(-1)^n}{n^n} = \int_0^1 x^x\,dx$$

$$\sum_n \frac{1}{n^n} = \int_0^1 \frac{dx}{x^x}$$



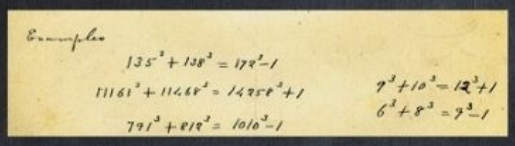
Here is the Ramanujan-Hardy formula for the calculation of the number of partitions:

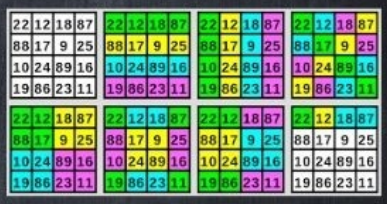
$$p(n) = \frac{1}{2\sqrt{2}} \sum_{k=1}^{v} \sqrt{k}\, A_k(n) \frac{d}{dn} \exp\left(\pi\sqrt{\frac{2}{3}}\frac{\sqrt{n-\frac{1}{24}}}{k}\right)$$

where

$$A_k(n) = \sum_{0 \le m < k;\ (m,k)=1} e^{\pi i\left[s(m,k) - \frac{1}{k}2nm\right]}$$

$$1/2+1/3+1/7+1/43+1/1807+\ \cdots = 1$$





$$e^{\pi\sqrt{163}} = 262537412640768743.99999999999992\ldots$$

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{8 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{10 + \cdots}}}}}}}}}}}}}$$

$$\begin{array}{ccccccc} 0 & \longrightarrow & \mathbb{Z}^2 & \xrightarrow{\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}} & 3\mathbb{Z} \oplus 2\mathbb{Z} & \longrightarrow & \\ \downarrow & & \downarrow{\scriptstyle \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 2 \end{pmatrix}} & & \downarrow & & \\ \mathbb{B}_d^a & \longrightarrow & \mathbb{Z}^4 & \xrightarrow{\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}} & \mathbb{Z}^2 & \longrightarrow & \end{array}$$

The aim of *Blackboard*, the Bulletin of the Mathematics Teachers' Association (India), is to promote interest in mathematics at various levels and to facilitate teachers in providing a well-rounded mathematical education to their students, in curricular as well as extra-curricular aspects. The Bulletin also serves as an interface between MTA (I) and the broad mathematical community.

Registered Office

Homi Bhabha Centre for Science Education
Tata Institute of Fundamental Research
V. N. Purav Marg, Mankhurd
Mumbai, 400088 INDIA

https://www.mtai.org.in/bulletin

Blackboard

Bulletin of the Mathematics Teachers' Association (India)
Issue 4

January 2022

Contents

Editorial

After the previous issue of Blackboard appeared, the Editorial Board took a few decisions to improve the visibility and the engagability of our Bulletin. One of these decisions was to invite to the editorial board some teachers who are associated with students and teachers at the school level also. In particular, I am glad to share with you the information that we have four new editors in Shailesh Shirali, Sangeeta Gulati, Aekta Agarwal and Anisa Chorwadwala. The other important decision was to bring out an issue more frequently - once in six months.

In the present issue, we have an assortment of articles that would be of interest to a diverse group of readers. A write-up by Raghavendra Kulkarni is on a novel method to solve cubic equations and would be accessible even to students at the high school level. For lovers of historical aspects of mathematics, Shailesh Shirali tells us about the historical roots of calculus - this is the first in a series of articles to be written by him. Kesavan writes on the notion of reflexivity of normed spaces. This article involves linear algebra and analysis at the college level but starts with simple notions, and gradually grows more advanced until it reaches the notion of proximality in Banach spaces. This is expected to be of interest to teachers at the postgraduate level. There is an absorbing account on alternating sign matrices by Satyanarayana Reddy and a student Mallika Muralidharan; this contains a description of the Dodgson condensation method to evaluate determinants. This is the same Dodgson who wrote under the pen name of Lewis Caroll.

The issue also contains an exposition on gaps between prime numbers and connections with the cyclotomic polynomials - the topic is advanced but the skilled exposition is sufficiently elementary for every one to know the state of the art about these deep aspects of prime number distribution. In a lovely article, Shantha Bhushan, Divakaran and Tulsi Srinivasan share their experiences with using Inquiry-Bsed learning in their undergraduate teaching. This is hoped to convince undergraduate teachers that using IBL can be an enriching experience. Finally, we have a beautiful introduction to the tidbits around the Euclid-method of proving the infinitude of primes. This has a companion article on how an analogue of the method carries over to prove infinitude of primes whose digits end in 1 for instance. The last piece in this issue is a purported conversation between a professor and a talented student which tries to unravel the differences between a polynomial function of 2 variables and a function that is separately a polynomial function of each variable but may not be a polynomial function of 2 variables.

Last, but not least, we would like to thank Professor A. Raghuram who served as an editor and played an important role in bringing the three previous issues into fruition. We hope to continue receiving his support in the forthcoming issues of Blackboard also albeit from outside the editorial board.

Here is wishing all of the readers a very Happy Year with beautiful Mathematical pursuits that elevate, inform and amuse at the same time!

— *B. Sury, Indian Statistical Institute Bangalore.*

1 Use of Möbius transformation for solving cubic equations

Raghavendra G. Kulkarni

Department of Electronics & Communication Engineering
PES University
100 Feet Ring Road, BSK III Stage
Bengaluru 560085

Email: raghavendrakulkarni@pes.edu, dr_rgkulkarni@yahoo.com

Abstract: In this paper we make use of the Möbius transformation for solving cubic equations. The proposed method has the advantage of obtaining only true solutions, in contrast to the method that employs the Tschirnhaus transformation which yields true as well as false solutions. We solve one numerical example using the proposed method.

It is well known that the solutions of cubic equations obtained from the Tschirnhaus transformation are a mixture of true and false solutions, and one has to employ trial and error method to sort out the true solutions [1, 2]. In this note, we make use of the Möbius transformation [3] to solve the cubic equation. All the solutions obtained using the proposed method are true solutions only, in contrast to that obtained from the Tschirnhaus transformation. Consider the following depressed cubic equation,

$$x^3 + ax + b = 0, \tag{1}$$

where a and b are coefficients in (1). Let Möbius transformation be defined as,

$$x = (cy + d)/(y + 1), \tag{2}$$

where the two variables, x and y, are connected through two unknown numbers, c and d. Using the transformation (2) we eliminate x from (1) to get the cubic equation in y as shown below.

$$y^3 + \left(\frac{3c^2d + 2ac + ad + 3b}{c^3 + ac + b}\right) y^2 + \left(\frac{3cd^2 + ac + 2ad + 3b}{c^3 + ac + b}\right) y + \frac{d^3 + ad + b}{c^3 + ac + b} = 0 \tag{3}$$

Equating the coefficients of y and y^2 in (3) to zero transforms (3) into a binomial cubic equation as shown below,

$$y^3 + \frac{d^3 + ad + b}{c^3 + ac + b} = 0, \tag{4}$$

as well as yields the following two expressions,

$$3cd^2 + ac + 2ad + 3b = 0, \tag{5}$$

$$3c^2d + 2ac + ad + 3b = 0, \tag{6}$$

which are used to determine the two unknowns, c and d. Subtracting (5) from (6) results in,

$$(c - d)(3cd + a) = 0.$$

Note that if the factor $c - d$ is equated to zero, the transformation (2) vanishes; hence, we equate the other factor, $3cd + a$, to zero, which yields,

$$cd = -a/3. \tag{7}$$

Adding (5) and (6) yields,

$$(c + d)(cd + a) = -2b. \tag{8}$$

Using (7) cd is eliminated from (8) leading to,

$$c + d = -3b/a. \tag{9}$$

Notice that (7) and (9) represent product and sum of roots (c and d) of a quadratic equation, say $z^2 + (3b/a)z - (a/3) = 0$, and therefore c and d are determined as:

$$c,\ d = \frac{3b}{2a}\left(-1 \pm \sqrt{1 + \frac{4a^3}{27b^2}}\right). \tag{10}$$

Now, the binomial cubic equation (4) can be solved; first, it is rearranged as, $y^3 = k^3$, where k is given by,

$$k = \left(-\frac{d^3 + ad + b}{c^3 + ac + b}\right)^{1/3}. \tag{11}$$

Hence the three solutions of $y^3 = k^3$ are:

$$y_1 = k, \quad y_2,\ y_3 = \frac{k}{2}(-1 \pm \sqrt{3}\, i), \tag{12}$$

where $i = \sqrt{-1}$. Subsequently from the transformation (2) the three solutions of given cubic equation (1) are obtained.

Let us solve one numerical example using the proposed method. Consider the following cubic equation,

$$x^3 - 6x - 9 = 0,$$

for solving by the proposed method. First, c and d are determined from (10) as: -0.5 and -4; using these values, we determine k from (11) as: -2. Using (12) the three solutions in y are obtained as: -2, $1-\sqrt{3}\,i$ and $1+\sqrt{3}\,i$. Using (2) we obtain the three solutions in x as: 3, $(-3-\sqrt{3}\,i)/2$, and $(-3+\sqrt{3}\,i)/2$. Note that even if the values of c and d are interchanged, we still get the same solutions in x as before. Interested readers may verify this using $c=-4$ and $d=-0.5$.

Acknowledgments

The author thanks the administration of PES University for supporting this work.

Bibliography

[1] Ehrenfried W. Tschirnhaus, "A method for removing all intermediate terms from a given equation", Acta Eruditorum, May 1683, pp. 204 - 207. Translated by R. F. Green in ACM SIGSAM Bulletin, Vol. 37, No. 1, March 2003, pp. 1 - 3.

[2] Victor S. Adamchik and David J. Jeffrey, "Polynomial transformations of Tschirnhaus, Bring and Jerrard" ACM SIGSAM Bulletin, Vol. 37, No. 3, September 2003, pp. 90 - 94.

[3] Douglas N. Arnold and Jonathan Rogness, "Möbius transformations revealed", Notices of the American Mathematical Society, Vol. 55, No. 10, November 2008, pp. 1226 - 1231.

2 Historical roots of calculus – 1

Shailesh Shirali

Sahyadri School KFI
Rajgurunagar, Khed
Pune – 410513

Email: shailesh.shirali@gmail.com

Abstract: Everyone knows that Newton and Leibniz independently invented the calculus. When we study calculus for the first time, we marvel at its beauty, just as we do at its power and versatility. We wonder at the same time: from where did such beautiful ideas emerge? Did they simply come out of nowhere? The answer is No. In this series of articles, we shall touch upon some strands that led ultimately to the calculus.

Interest in curves grew steadily over the 17th century. In earlier eras, interest had been confined to the conic sections; now it included curves such as the cycloid and the tractrix. There were many who wondered how the slopes of such curves could be ascertained. In Part 1 of this series of articles, we look at approaches pioneered by Roberval, Descartes and Fermat to draw tangents to curves. Though unrealised at the time, it is Fermat's approach that is closest in spirit to the modern idea of the derivative as a limit. (During Fermat's lifetime, ironically, his approach was regarded by his contemporaries as problematic and inherently flawed.)

Gilles de Roberval (1602–1675), Pierre de Fermat (1601–1665) and René Descartes (1596–1650) were contemporaries of one another; all belong to the era just preceding Isaac Newton.

Roberval's approach

A curve can be thought of as a set of points determined by an equation, which makes it a static object, or as the path traced by a moving point, which brings in a dynamic element. Roberval's viewpoint is the latter one. He sees that if he can identify the vector

components of the motion of the point, then he can find the tangent by adding the two vectors, for that would yield the direction of instantaneous motion at that point. (He does not use this terminology, which came later, but it is clear that this is what he is doing.) Let us see how this idea works for a parabola. (For ease of comprehension, we use modern terminology and modern symbols in our description, but obviously this is not the way that Roberval would have presented it.)

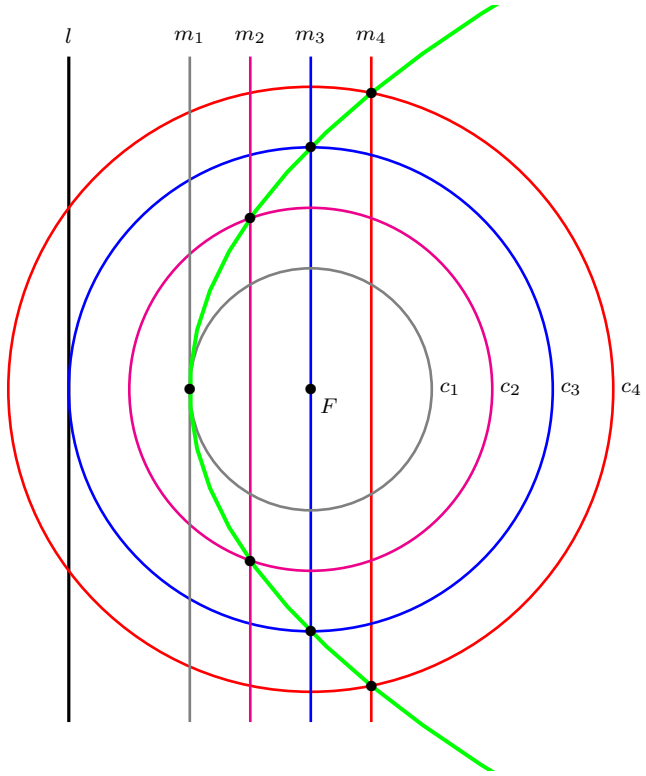


Figure 1: Dynamically generating a parabola with directrix l and focus F

Consider how a parabola is generated. Let l be a fixed line and F a fixed point not on l. Imagine a second line m, initially coincident with l, moving away from l at a uniform speed v, towards the side where F is located (Figure 1). Imagine also a circle c centred at F, initially with radius 0, expanding at the same uniform speed v. If the two movements start at the same instant, then each point of intersection of m and c will be equidistant from l and F. So the moving point describes a parabola. In Figure 1, corresponding pairs of circle and line have the same colour.

Now consider the component vectors (Figure 2), shown in red. One component v_1 is directed away from l, while the other component v_2 is directed away from F; moreover, the two have the same length as the two velocities are equal. Hence the vector sum $v_1 + v_2$ is equally inclined to the two component vectors. Letting A denote the foot of the perpendicular from P to l, we see that the direction of instantaneous motion at

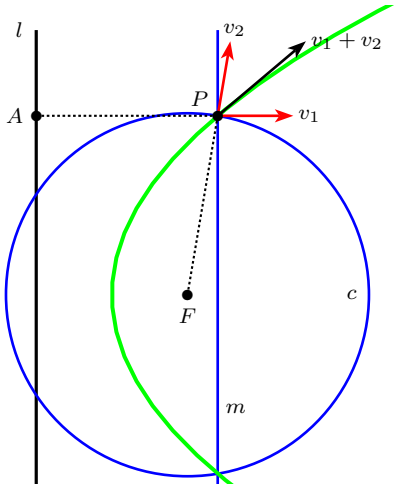


Figure 2: Using the component vectors to find the direction of instantaneous motion

P bisects $\angle APF$. In other words, the tangent to the parabola at P bisects $\angle APF$. We thus have a construction for the tangent to the curve at P. (*Note.* The result just derived implies the well known "perfect-focus" property of a parabolic mirror.)

The case of the ellipse may be handled in the same manner; here we use the two-foci definition of the ellipse. Consider an ellipse $\mathcal{E}$ with a pair of foci A, B, with distance sum d. It may be generated by imagining circles C_A and C_B centred at A and B respectively, the sum of their radii being d (see Figure 3). Let C_A expand at a uniform rate; let C_B shrink at the same rate. At each instant, the two circles give rise to two points of intersection P, Q (possibly non-existent, if one circle is too big and the other is too small). These moving points generate the ellipse $\mathcal{E}$.

Consider the component vectors of the motion of P. One vector (v_1) is directed away from A, while the other vector (v_2) is directed towards B; the two vectors have equal length. Consequently, the direction of their sum $v_1 + v_2$ bisects the angle between them. This gives an easy way of drawing the tangent to the ellipse at any given point P on it. It also demonstrates why a ray of light proceeding from one focus is reflected by the ellipse to the other focus.

The case of the hyperbola may be handled in exactly the same manner. Observe that numerous extensions are possible using simple tweaks. For example, in the case of the parabola, we need not insist on the vectors having the same length; we could require only that their lengths maintain a constant ratio. This model allows us to consider the other conic sections (with eccentricities not equal to 1). Similar tweaks are possible in the other model considered.

Attractive as Roberval's method is, we must note that it is also quite limited. It depends critically on our being able to find a mechanical model for generating the curve. In

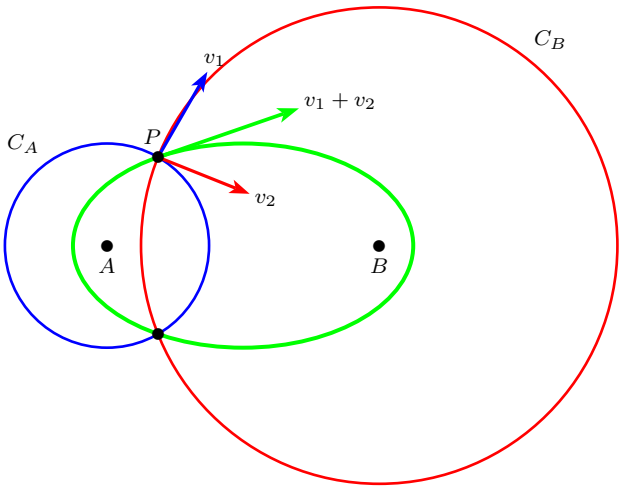


Figure 3: Dynamically generating the ellipse and using the component vectors to find the direction of instantaneous motion

situations where this does not work out in a simple way, the method ends up being quite contrived. Probably this is the reason why Roberval's method was not pursued in later decades.

Descartes' approach

The approach suggested by Descartes is conceptually simpler than Roberval's; it will readily appeal to today's students. Let a curve be given, with defining equation $y = f(x)$. Let P be a point on the curve at which we need to draw the tangent. Let C be a variable point on the x-axis. Consider the circle centred at C and passing through P. It will likely intersect the curve again at least one other point, say Q. For some choice of C it may happen that Q coincides with P. In this situation, the circle has double contact with the curve, so the tangent to the curve at P will coincide with the tangent to the circle at P. But the tangent to the circle at P is perpendicular to PC, whose slope is known as its endpoints are known. This enables us to draw the tangent to the given curve at P. The task now reduces to finding the point C such that the curve and the circle have double contact at P (which means that the associated equation has a pair of equal roots). We illustrate the procedure with an example.

Example. Let us find the slope of the curve $y = x^2/4$ at the point $P = (2, 1)$. See Figure 4.

As earlier, we show the calculations using modern notation. Let $C = (c, 0)$. Then we

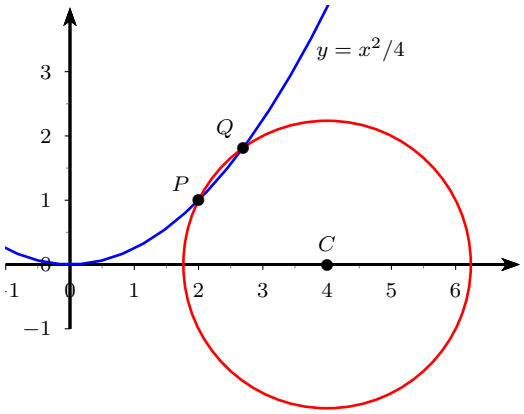


Figure 4: Finding the slope of the curve $y = x^2/4$ at the point $P = (2, 1)$ using Descartes's method

have,

$$CP^2 = (c-2)^2 + (0-1)^2 = c^2 - 4c + 5,$$

so the equation of the circle with centre C, passing through P, is $(x-c)^2+y^2 = c^2-4c+5$, which simplifies to

$$x^2 - 2cx + y^2 + 4c - 5 = 0.$$

We now substitute $y = x^2/4$ to find the points of intersection of the circle and the curve. We get the equation

$$x^4 + 16x^2 - 32cx + 64c - 80 = 0.$$

Note that $x = 2$ is a root of the equation (as it must be, since P is one of the points of intersection). Therefore $x - 2$ must be a factor of the polynomial on the left side. Dividing through by $x - 2$, we obtain the cubic equation

$$x^3 + 2x^2 + 20x - 32c + 40 = 0.$$

If P is to be a point of double contact of the curve and the circle, then $x = 2$ must be a root of this equation as well. Substituting $x = 2$ we get

$$96 - 32c = 0, \quad \therefore c = 3, \quad \therefore C = (3, 0)$$

For this choice of C, the slope of PC is -1, hence the slope of the tangent at P is 1.

This approach will work for any rational function f. The algebra may get messy, but the approach remains simple conceptually. All we need to do is use the factor theorem repeatedly. This is all familiar territory for today's 11-12 mathematics students.

But it will not work for many familiar functions; in particular, it will not work for trigonometric functions, exponential functions and logarithmic functions. This naturally limits its usage hugely.

Fermat's approach

We close with a discussion of the approach used by Fermat. As noted earlier, this comes closest in spirit to the approach we follow today. But during its time, it was regarded as suspect (indeed, his contemporaries found the approach mysterious), and Fermat was the target of a great deal of criticism because of this.

Today, we first learn how to find the derivative of a function and then apply the theory to finding the extreme points of the function. But Fermat proceeded in the reverse direction: he worked out a way of finding the extreme points of a function, and then applied the same logic to finding the slope of the function. His insight came from the observation that at the extreme point of a function $f(x)$, a tiny change in x appears to have "almost no effect" on $f(x)$, but this is not so at points that are not extreme points.

For example, consider the function $f(x) = x^2$; it has an extreme point at $x = 0$. Let x change from 0 to 0.01; then the function value changes from 0 to 0.0001. Observe that the change in the function value is small in comparison with the change in x. In contrast, if x changes from 1 to 1.01 (the same incremental change as earlier), then the function value changes from 1 to 1.0201; the change in the function value (0.0201) is roughly comparable to the change in x. What Fermat has noted is of great significance. Using today's language, his observation amounts to saying that the slope of the function at an extreme point is 0, and it explains why extreme points are also called "stationary points". (*Comment.* It is of great interest to learn that in the 12th century, a very similar observation was made by Bhaskara II, in India. We shall have more to say about this later in this series.)

Let us see how Fermat applies this observation to find the extreme points of $f(x) = x^3 - 3x$. Let the argument of the function change from x to $x + E$, where E is tiny. Then the function value changes from $x^3 - 3x$ to $(x + E)^3 - 3(x + E)$. The change in function value is therefore

$$\begin{aligned} f(x + E) - f(x) &= \big((x + E)^3 - 3(x + E)\big) - \big(x^3 - 3x\big) \\ &= 3Ex^2 + 3E^2x + E^3 - 3E. \end{aligned}$$

Fermat's next step is the one which looked mysterious to his associates. He divides the difference $f(x + E) - f(x)$ by E, obtaining the expression

$$\begin{aligned} \frac{f(x + E) - f(x)}{E} &= \frac{3Ex^2 + 3E^2x + E^3 - 3E}{E} \\ &= 3x^2 + 3Ex + E^2 - 3, \end{aligned}$$

then puts $E = 0$ in the final expression, thereby getting $3x^2 - 3$. He then equates this to 0 and solves the resulting equation for x, getting $x = \pm 1$. (Fermat discards the negative

value as irrelevant. We need not worry about this point here.) Observe that he has correctly identified the extreme points of the function.

From our vantage point, i.e., with our modern understanding of infinitesimals and limits, we can see what Fermat is doing. But to his contemporaries, his action of dividing by E and then setting $E = 0$ seemed to be essentially division by zero (a suspect action then as now) and therefore a piece of trickery.

Now let us see how Fermat applies this idea to finding the slope of a curve $y = f(x)$ at an arbitrary point P on the curve.

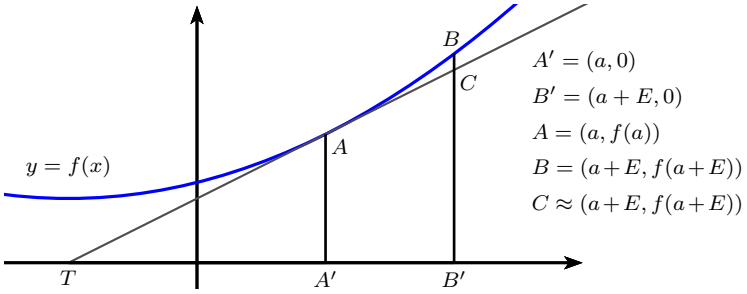


Figure 5:

In Figure 5, $A = (a, f(a))$ is an arbitrary point on the curve $y = f(x)$, and $A' = (a, 0)$ is the foot of the perpendicular from A to the x-axis. The tangent to the curve at A meets the x-axis at T. Let $B' = (a+E, 0)$, where E plays the same role as earlier (it represents a 'tiny increment'). Let $B = (a + E, f(a + E))$ be the point on the curve corresponding to $x = a + E$, and let C be the point where BB' intersects line AD. As E is small, we regard the distance between B and C as negligible, so we take the coordinates of C to be $(a + E, f(a + E))$. Now consider the pair of similar triangles ATA' and CTB'. By similarity,

$$\frac{AA'}{TA'} = \frac{CB'}{TB'}.$$

The quantity AA'/TA' is, of course, the slope of the curve at A. From the above relation we get

$$\text{Slope of curve at } A = \frac{CB' - AA'}{TB' - TA'} = \frac{f(a + E) - f(a)}{E}.$$

Observe that we have obtained an expression that is suspiciously familiar! This will explain the comment made earlier, that Fermat's approach is very close to the approach we follow even today.

Fermat's prescription to find the slope of the curve at $P = (a, f(a))$ is thus:

Simplify the expression

$$\frac{f(a + E) - f(a)}{E}$$

and then put $E = 0$. The answer will be the slope of the curve at P.

Examples

1. If the given curve is $f(x) = x^2$, then we have:

$$\frac{f(a+E) - f(a)}{E} = \frac{(a+E)^2 - a^2}{E}$$
$$= \frac{2aE + E^2}{E} = 2a + E.$$

   Putting $E = 0$, we get $2a$. Hence the slope of the given curve at (a, a^2) is $2a$.

2. If the given curve is $f(x) = x^3$, then we have:

$$\frac{f(a+E) - f(a)}{E} = \frac{(a+E)^3 - a^3}{E}$$
$$= \frac{3a^2E + 3aE^2 + E^3}{E} = 3a^2 + 3aE + E^2.$$

   Putting $E = 0$, we get $3a^2$. Hence the slope of the given curve at (a, a^3) is $3a^2$.

One can sympathise with Fermat's contemporaries. They must have observed that the procedure gave correct results, and yet something seemed wrong! How baffling and how frustrating!

We had noted above that Descartes' method will not work for trigonometric functions, exponential functions and logarithmic functions. Where does Fermat's method stand in this regard? The same criticism will evidently hold here as well. For example, if we had to apply this method to the sine function, we would find ourselves trying to simplify the expression

$$\frac{\sin(x+E) - \sin x}{E}.$$

Unfortunately, this expression does not yield to algebraic simplification. It is not just a different path that is required to find the answer in such a case: a different conceptual basis is required. But that development lay several decades in the future.

It is of interest to read the reaction of Descartes to Fermat's method. As already noted, Descartes' method is conceptually simple (though it can get messy algebraically). Descartes was aware of this and therefore proud of his invention, which he regarded as much superior to Fermat's (here, he shared the suspicions of his contemporaries). But the verdict of history has gone against him.

Pedagogically, however, Descartes' method has instructional value. It will surely appeal to the student, as its conceptual basis is quite transparent. Moreover, it is very easy to write a GeoGebra applet to illustrate the working of the method. This greatly enhances its appeal and value.

Closing remarks

We can see from the above that the accomplishments of Newton and Leibniz, astonishing as they were, did not happen in a vacuum. Interest in such matters was in the air at the time (the middle decades of the 17th century), and it remained for two people of superlative ability to make the next crucial conceptual leap.

In the next article of this series, we shall explore the historical roots of integration. We shall find that Fermat's name comes up again (as it does in so many parts of mathematics). Later, we shall touch upon related work of Indian mathematicians of earlier centuries.

Bibliography

[1] "The History of the Calculus and the Development of Computer Algebra Systems," from http://www.math.wpi.edu/IQP/BVCalcHist/calctoc.html

[2] "History of the Integral from the 17th Century," from http://www.math.wpi.edu/IQP/BVCalcHist/calc1.html

[3] "History of the Differential from the 17th Century," from http://www.math.wpi.edu/IQP/BVCalcHist/calc2.html

[4] MAA, "Historical Activities for Calculus, Module 1", https://www.maa.org/press/periodicals/convergence/historical-activities-for-calculus-module-1-curve-drawing-then-and-now

[5] MAA, "Historical Activities for Calculus, Module 2", https://www.maa.org/press/periodicals/convergence/historical-activities-for-calculus-module-2-tangent-lines-then-and-now

3 Reflexivity and its avatars

S Kesavan

Formerly Professor, Institute of Mathematical Sciences, Chennai

Email: kesh@imsc.res.in

Abstract: We will study various equivalent forms of the notion of reflexivity of a Banach space via continuous linear functionals which do not attain their norm, weak topologies, and existence of solutions to some optimization problems.

1 Preliminaries

Let V be a vector space. For purposes of exposition, we will assume that the scalar field is that of the real numbers, $\mathbb{R}$, though it will not be difficult to extend the arguments to the case when the base field is $\mathbb{C}$.

Recall that a norm on a vector space (over $\mathbb{R}$ or $\mathbb{C}$) is a mapping $\|\cdot\| : V \to [0, \infty)$ such that (i) $\|v\| = 0$ if, and only if, $v = 0$, (ii) $\|\alpha v\| = |\alpha| \, \|v\|$, for every scalar α and for every $v \in V$, and (iii) $\|v + w\| \leq \|v\| + \|w\|$, for every v and w in V. The last inequality is called the *triangle inequality*.

It follows from the properties of the norm that the function, d, defined on $V \times V$ by $d(x, y) = \|x - y\|$ defines a metric on V and the corresponding metric topology on V is called its *norm topology*. A vector space equipped with a norm, is called a *normed linear space*. If the space is complete with respect to the induced metric topology, we say that V is a **Banach space**. For examples of normed linear spaces, see Kesavan [2].

Given two normed linear spaces V and W and a linear map $T : V \to W$, we say that T is a *continuous linear transformation* if it is a continuous map with respect to the norm topologies of V and W. The collection of all continuous linear transformations from V into W, denoted $\mathcal{L}(V, W)$, forms a vector space with respect to pointwise addition and pointwise scalar multiplication, i.e., if T and S belong to $\mathcal{L}(V, W)$, and if α is a scalar,

the mappings $T+S$ and αT defined by

$$(T+S)v \;=\; Tv+Sv, \text{ and } (\alpha T)v \;=\; \alpha Tv,$$

for every $v \in V$, also belong to $\mathcal{L}(V,W)$. In fact, this space is also a normed linear space for the norm defined by

$$\|T\| \;=\; \sup_{v\in V,\|v\|_V\leq 1}\|Tv\|_W \;=\; \sup_{v\in V,\|v\|_V=1}\|Tv\|_W \;=\; \sup_{v\in V,v\neq 0}\frac{\|Tv\|_W}{\|v\|_V}. \tag{1.1}$$

(In the above relations, we have denoted the norms in V and in W by $\|\cdot\|_V$ and $\|\cdot\|_W$ respectively.) If W is Banach, then so is $\mathcal{L}(V,W)$.

Proofs of these assertions and examples of continuous linear transformations can be found in any textbook on Functional Analysis. See, for instance, Kesavan [2].

In particular, the scalar field itself is a (one dimensional) Banach space over itself. Thus $\mathcal{L}(V,\mathbb{R})$ (or $\mathcal{L}(V,\mathbb{C})$, in the case of complex vector spaces) is a Banach space. Its elements are called *continuous linear functionals* and the space itself is called the *dual space* of V and is denoted by V^*.

If $V=W$, then $\mathcal{L}(V,V)$ is denoted by $\mathcal{L}(V)$, and, usually, it is common to refer to its elements as *continuous linear operators*.

Remark 1.1 Continuous linear transformations (respectively, operators, functionals) are synonymously referred to as *bounded* linear transformations (respectively, operators, functionals). ■

2 Reflexivity

One of the grand theorems of functional analysis is the Hahn-Banach theorem which says that if W is a subspace of a normed linear space V, and if f is a continuous linear functional on W (which inherits the same norm from V and hence can be considered as a normed linear space in its own right), then f can be extended as a continuous linear functional on all of V, *preserving the norm*.

An immediate corollary of the Hahn-Banach theorem is the following result.

Proposition 2.1. *Let V be a normed linear space and let $x_0 \in V$ be a non-zero vector. Then, there exists $f \in V^*$, the dual space of V, such that $\|f\| = 1$ and such that $f(x_0) = \|x_0\|$.*

Proof. Consider the one-dimensional space spanned by x_0 and define $g(tx_0) = t\|x_0\|$. The result now follows from the Hahn-Banach theorem. ■

Corollary 2.1. *Let V be a normed linear space and let V^* denote its dual. Let $x \in V$. Then*

$$\|x\| \;=\; \sup_{\substack{f\in V^*\\ \|f\|=1}} |f(x)| \;=\; \max_{\substack{f\in V^*\\ \|f\|=1}} |f(x)|. \tag{2.1}$$

Proof. Indeed, if $f \in V^*$ and $\|f\| = 1$, then $|f(x)| \leq \|x\|$. The preceding proposition assures us that this supremum is attained. ■

The relation (2.1) tells us that if we define a linear functional $J(x)$ on V^*, for $x \in V$, by

$$J(x)(f) \;=\; f(x),$$

for every $f \in V^*$, then $J(x) \in V^{**}$ and also that $\|J(x)\| = \|x\|$. Thus J defines a canonical isometry from V into V^{**}. This is the starting point of the notion of relexivity.

Definition 2.1. *A normed linear space V is said to be reflexive if the canonical isometry J is surjective.* ■

In other words, we may identify the spaces V and its bi-dual, V^{**}. Since any dual space is complete, a reflexive space is necessarily a Banach space.

Example 2.1. If V were finite dimensional, then the dimension of V, V^* and V^{**} are all the same. Since an isometry is injective, in this case J is automatically surjective as well. Thus, every finite dimensional normed linear space is reflexive. ■

Example 2.2. Let $x = (x_1, x_2, \cdots, x_k, \cdots)$ be a real (or, complex) sequence. For $1 \leq p < \infty$, it is said to be p-summable if

$$\sum_{k=1}^{\infty} |x_k|^p \;<\; +\infty.$$

Define ℓ_p to be the collection of all p-summable sequences. It can be shown that this is a vector space and that it becomes a Banach space for the norm given by

$$\|x\|_p \;=\; \left(\sum_{k=1}^{\infty} |x_k|^p\right)^{\frac{1}{p}}, \; x \in \ell_p.$$

The collection of all bounded sequences is denoted ℓ_∞ and it is a Banach space when equipped with the norm given by

$$\|x\|_\infty \;=\; \sup_k |x_k|, x \in \ell_\infty.$$

If $p = 1$, we define $p^* = \infty$ and if $1 < p < \infty$, we define p^* via the relation

$$\frac{1}{p} + \frac{1}{p^*} \;=\; 1.$$

The quantity p^* is called the *conjugate exponent* of the exponent p. For $1 \leq p < \infty$, it can be shown that the dual of ℓ_p is ℓ_{p^*}, i.e., $\ell_p^* = \ell_{p^*}$. If $x = (x_1, \cdots, x_k, \cdots) \in \ell_p$ and if $y = (y_1, \cdots, y_k, \cdots) \in \ell_{p^*}$, the action of the linear functional generated by y on x (in the real case) is given by

$$< y, x > \;=\; \sum_{k=1}^{\infty} x_k y_k. \tag{2.2}$$

(In the complex case, y_k in the above summation is replaced by its complex conjugate.) From this it is easy to see that ℓ_p is reflexive when $1 < p < \infty$. ■

It is interesting to compare the relation (2.1) with the following one, which is a *definition* (cf. (1.1)), while (2.1) is a result of the theory.

$$\|f\| \;=\; \sup_{\substack{x \in V \\ \|x\| \leq 1}} |f(x)| \;=\; \sup_{\substack{x \in V \\ \|x\| = 1}} |f(x)|. \tag{2.3}$$

Remark 2.1. An important result in functional analysis is that that the closed unit ball, i.e., the set of all vectors with norm less than or equal to unity, is compact if, and only if, the space is finite dimensional. If the closed unit ball is compact, then it is easy to see that there will exist a vector $v \in V$ such that $\|v\| = 1$ and such that $|f(v)| = \|f\|$. (In the real case, by considering v and $-v$, we can also assume that we have a unit vector v such that $f(v) = \|f\|$). In the infinite dimensional case, such a vector may, or may not, exist for a given continuous linear functional. ■

While the theory states that the supremum in (2.1) is always attained, the supremum in (2.3) may not be attained, as remarked above. Now, if V is reflexive, and we apply the above corollary to the space V^*, we get that (since every element of V^{**} is of the form $J(x)$ for some $x \in V$ and since J is an isometry)

$$\|f\| \;=\; \max_{\substack{x \in V \\ \|x\| = 1}} |f(x)|.$$

That is, for reflexive spaces, the supremum in (2.3) is also attained, for every $f \in V^*$.

The existence of a continuous linear functional on a Banach space V for which the supremum in (2.3) is not attained, gives a proof of the non-reflexive nature of the space. We give some examples below.

Example 2.3. Consider the space ℓ_1. Let $y = (y_1, \cdots, y_k, \cdots) \in \ell_\infty$ where $y_k = 1 - \frac{1}{k}$. Clearly, $\|y\|_\infty = 1$. If $x \in \ell_1$ is such that $| < y, x > | = 1 = \|y\|_\infty$, then, since $|y_k x_k| < |x_k|$ for each positive integer k, it follows that

$$1 \;=\; | < y, x > | \;<\; \|x\|_1.$$

Thus it follows that the supremum in (2.3) cannot be attained, in the closed unit ball, for this functional. Thus, it follows that ℓ_1 is not reflexive. ■

Example 2.4. Consider the space c of all real sequences which are convergent. This is a closed subspace of ℓ_∞. Let $y = (y_1, y_2, \cdots, y_k, \cdots) \in \ell_1$. Then, if $x = (x_1, x_2, \cdots, x_k, \cdots) \in c$, we have that y defines a continuous linear functional on c, via the action defined by (2.2) and it is easy to see that the norm of this functional is given by $\|y\|_1$. Assume that the supremum in (2.3) is attained on the unit sphere of c. Without loss of generality, we may assume that there exists $x \in c$, with $\|x\|_\infty = 1$, such that $< y, x >= \|y\|_1$. Let $\|y\|_1 = 1$. Thus,

$$1 \;=\; \|y\|_1 \;=\; \sum_{k=1}^{\infty} y_k x_k.$$

Since, $\|x\|_\infty = 1$, it follows that, for each k, $y_k x_k \leq |y_k x_k| \leq |y_k|$. Then it follows from the preceding equation that, for each k, we have

$$|y_k| \;=\; y_k x_k.$$

Now assume that, for each k, $y_k \neq 0$ and that $y_k = (-1)^k |y_k|$. (Example: $y_k = (-1)^k (\frac{1}{2})^k$.) Then it follows that $x_k = (-1)^k$, which is a contradiction since $x \notin c$ in this case. Thus, for all such $y \in \ell_1$, the supremum is not attained in (2.3) and so c is not reflexive. ■

Example 2.5. Consider the space c_0 of all real sequences which converge to zero. This is a closed subspace of c. One can prove that $c_0^* = \ell_1$. If $y \in \ell_1$ and if $x \in c_0$, again, the action of the functional defined by y on x is given by (2.2), where, as in the preceding example, x_k and y_k are the components of x and y respectively. If $\|y\|_1 = 1$ and if $\|x\|_\infty = 1$, we have that $|x_k| < 1$ for all $k \geq N$, for some positive integer N. Then it is clear that $| < y, x > | < \|y\|_1 = 1$. Thus, for no continuous linear functional on c_0 we have that the supremum in (2.3) is attained. Thus, c_0 is not reflexive. ■

Example 2.6. Let $V = C[0, 1]$, the space of continuous real-valued functions defined on the interval $[0, 1]$, equipped with the usual 'sup-norm', denoted $\|\cdot\|_\infty$. Consider the linear functional φ defined on V by

$$\varphi(f) \;=\; \int_0^{\frac{1}{2}} f(t)\, dt - \int_{\frac{1}{2}}^{1} f(t)\, dt,$$

for every $f \in V$. Clearly $|\varphi(f)| \leq \|f\|_\infty$ and so $\varphi \in V^*$ and $\|\varphi\| \leq 1$. Now consider the sequence of functions $\{f_n\}$ in V, where, for each sufficiently large positve integer n, we

have

$$f_n(t) \;=\; \begin{cases} +1, & \text{if } t \in \left[0, \frac{1}{2} - \frac{1}{n}\right], \\ 1 + n\left(\frac{1}{2} - \frac{1}{n} - x\right), & \text{if } t \in \left[\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}\right], \\ -1, & \text{if } t \in \left[\frac{1}{2} + \frac{1}{n}, 1\right]. \end{cases}$$

The graph of f_n is given in the figure below.

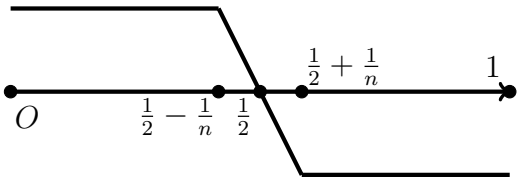


Figure 1: The function f_n

We see from this picture (by computing the relevant areas) that $\varphi(f_n) = 1 - \frac{1}{n}$. Since $\|f_n\|_\infty = 1$ for all n, it follows that $\|\varphi\| = 1$.

We now show that there is no function $f \in V$ such that $\|f\|_\infty = 1$ and such that $\varphi(f) = \|\varphi\| = 1$. Indeed if there were such a function, then consider the function g defined on $(0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$ by

$$g(t) \;=\; +1, \text{ if } t \in \left(0, \frac{1}{2}\right), \text{ and } g(t) \;=\; -1 \text{ if } t \in \left(\frac{1}{2}, 1\right).$$

Then

$$1 \;=\; \int_0^{\frac{1}{2}} g(t)\, dt - \int_{\frac{1}{2}}^1 g(t)\, dt \;=\; \int_0^{\frac{1}{2}} f(t)\, dt - \int_{\frac{1}{2}}^1 f(t)\, dt.$$

Then

$$\int_0^{\frac{1}{2}} (g(t) - f(t))\, dt \;=\; \int_{\frac{1}{2}}^1 (g(t) - f(t))\, dt. \tag{2.4}$$

But $|f(t)| \leq 1$, i.e., $-1 \leq f(t) \leq +1$ for all t and so the integrand on the left-hand side of (2.4) is non-negative, while that on the right-hand side is non-positive. Thus each of the integrals in (2.4) is zero. But then, again, since those integrands are of constant sign, it follows that $f \equiv +1$ on $(0, \frac{1}{2})$ and that $f \equiv -1$ on $(\frac{1}{2}, 1)$, which contradicts the continuity of f.

Thus, φ is a continuous linear functional on $C[0,1]$ for which $\|\varphi\|$ is not attained on the unit sphere and we conclude that $C[0,1]$ is not reflexive. ■

We have shown the non-reflexivity of a Banach space by exhibiting continuous linear functionals which do not attain their norm on the unit sphere, since, in a reflexive space,

every linear functional must attain its norm. A deep and famous theorem of James [1] states that if a Banach space is such that every continuous linear functional attains its norm on the unit sphere, then the space must be reflexive. We will see applications of this to some optimization problems in the sequel. It will also give other characterizations of reflexive spaces.

3 Weak topologies

Let V be a Banach space. The *weak topology* on V is the smallest topology on V such that every element of V^* is still continuous. The open (respectively, closed, compact, ...) sets of this topology are said to be weakly open (respectively, weakly closed, weakly compact, ...). A sequence in V is said to be *weakly convergent* if it converges with respect to the weak topology, and is *norm convergent* if it is convergent in the usual sense, i.e., with respect to the norm topology). A linear map between two Banach spaces is said to be *weakly continuous* if it is continuous as a map between these two spaces when both of them are provided with the weak topology.

If a Banach space V is finite dimensional, then the weak and norm topologies coincide. In the case of infinite dimensional spaces, the weak topology is strictly smaller than the norm topology, i.e., every weakly open set is norm open but the converse is not necessarily true. In fact, the open unit ball is not weakly open! Again, every weakly closed set is norm closed, but the converse need not hold. However, using the Hahn-Banach theorem, we can show that every (norm) closed and convex set in V is weakly closed. Thus, the closed unit ball is weakly closed. But the closed unit sphere is not weakly closed. In fact, the weak closure of the unit sphere is the closed unit ball!

We state three important facts:

- The weak topology is Hausdorff.
- A sequence $\{x_n\}$ in a Banach space V converges weakly to $x \in V$ if, and only if, $f(x_n) \to f(x)$ for every $f \in V^*$.
- If V and W are Banach spaces and if $T : V \to W$ is a linear map, then T is weakly continuous if, and only if, $T \in \mathcal{L}(V, W)$.

The proofs of all results connected with weak topologies which are stated in this article can be found, for instance, in Kesavan [2].

Now consider the dual space, V^*, of a Banach space V. Here we have, again, two topologies. There is the norm topology and the weak topology, viz. the smallest topology

on V^* such that every element of V^{**} is continuous. We also have a third topology, viz. the smallest topology such that every element of $J(V)$, where $J : V \to V^{**}$ is the canonical isometry, is continuous. This is called the weak* topology on V^*. This is the smallest of the three topologies and is also Hausdorff.

Every weak* open (respectively, closed) set in V^{**} is weakly open (respectively, closed) and hence open (respectively, closed) in the norm topology. If the space V is finite dimensional, then the three topologies on V^* are the same. If V is an infinite dimensional reflexive space, then, since $J(V) = V^{**}$, we have that the weak and weak* topologies are the same, but strictly smaller than the norm topology. If V is non-reflexive then the weak* topology is strictly smaller than the weak topology on V^*. For instance, we saw that a closed convex set is weakly closed. It need not be weak* closed. In fact, if B is the closed unit ball in V and B^{**} is the closed unit ball in V^{**}, then $J(B)$ is a closed convex set in V^{**} which is strictly contained in B^{**}, if V is non-reflexive. Now, V^{**}, being a dual space, we can look at its weak* topology. It turns out that the weak* closure of $J(B)$ is B^{**}. Thus, if V is non-reflexive, closed convex sets in V^{**}, which are weakly closed, need not be weak* closed.

The same is true in V^* as well. Let $\varphi \in V^{**}$. Consider the set

$$H \;=\; \{f \in V^* \mid \varphi(f) = \alpha\},$$

where α is a fixed scalar. This set is clearly closed and convex and hence weakly closed in V^*. It will be weak* closed if, and only if, $\varphi = J(v)$ for some $v \in V$.

At this juncture, the reader may wonder what is the point of impoverishing the topologies like this. The reason is the following. The open sets in the weak topology are not only fewer, but they are bigger as well. An open neighbourhood in the weak topology can contain one dimensional affine subspaces. Thus with fewer and larger open sets, we can hope that finding a finite subcover of an open cover will be more feasible, i.e., we can expect sets which are not compact in the norm topology become compact in one of the weaker topologies. This hope is justified by the following theorem.

Theorem 3.1. *(Banach-Alaoglu) Let V be a Banach space. The closed unit ball, B^* of V^* is weak* compact.* ■

If V is reflexive, then the above result states that B^* is weakly compact. Recall that if V is infinite dimensional, B,B^* and B^{**} are all non-compact sets in their respective norm topologies.

4 Reflexivity and the weak topology

Notation: Given a Banach space V, we will denote the closed unit balls in V, V^* and V^{**} by B, B^* and B^{**}, respectively.

Theorem 4.1. *A Banach space V is reflexive if, and only if, B is weakly compact.*

Proof. Assume that B is weakly compact. Since $J : V \to V^{**}$ is an isometry, it is continuous and hence weakly continuous as well and so $J(B)$ is weakly compact. Hence it is weak* compact as well. The weak* topology being Hausdorff, it follows that $J(B)$ is weak* closed. But then it follows that $J(B) = B^{**}$. This immediately implies that J is surjective, i.e., V is reflexive.

Conversely, let V be reflexive. Then the weak and weak* topologies on V^* coincide. Hence, by the Banach-Alaoglu theorem, B^* is weakly compact. Then, by the preceding arguments, it follows that V^* is reflexive. Then, just as we saw earlier, it follows that B^{**} is weakly compact. Since V is reflexive, we have $B = J^{-1}(B^{**})$. Also, since $J^{-1} : V^{**} \to V$ is continuous, it is weakly continuous as well and so B is weakly compact. ■

Remark 4.1. Of course, the above result implies immediately that any closed ball in a reflexive space is weakly compact. ■

Corollary 4.1. *Let V and W be Banach spaces and let $T : V \to W$ be an isometric isomorphism. Then, if V is reflexive, so is W.*

Proof. Let B_V and B_W be the closed unit balls in V and W, respectively. Since T is an isometric isomorphism, we have that $T(B_V) = B_W$. Now, T being continuous, it is weakly continuous as well. Since V is reflexive, we have that B_V is weakly compact and so $B_W = T(B_V)$ is also weakly compact, which implies that W is reflexive. ■

Corollary 4.2. *Let V be a reflexive Banach space and let W be a* closed *subspace of V. Then W is also reflexive.*

Proof. It is easy to see that the weak topology on W is none other than the topology induced on W by the weak topology of V. Since V is reflexive, it follows that B is weakly compact. The unit ball in W is none other than $W \cap B$. But W being a closed subspace, it is weakly closed and since B is weakly compact, it follows that $W \cap B$ is weakly compact as well. Thus, it follows that W is reflexive. ■

Corollary 4.3. *Let V be a Banach space. Then, V is reflexive if, and only if, V^* is reflexive.*

Proof. We already saw in the proof of Theorem 3.1 that if V is reflexive, then V^* is reflexive.

Conversely, let V^* be reflexive. Then, as before, V^{**} is reflexive. Now, $J(V)$ is a closed subspace of V^{**} and so, by the preceding corollary, it follows that $J(V)$ is reflexive. But then $J^{-1} : J(V) \to V$ is an isometric isomorphism and so V is reflexive by Corollary 4.1. ■

Example 4.1. We already saw that ℓ_1 is not reflexive. Thus, since $c_0^* = \ell_1$ and $\ell_1^* = \ell_\infty$, it follows that c_0 and ℓ_∞ are not reflexive. Further, since c_0 is a non-reflexive closed subspace of c, it follows that c is also not reflexive. We already looked at these examples in Section 2. ■

The proof of the theorem of James stated in Section 2 uses Theorem 4.1. In fact, James [1] proves that a weakly closed set A in a Banach space is weakly compact if, and only if, every continuous linear functional attains its supremum in A. Thus, if every continuous linear functional attains its supremum, i.e., its norm, in the closed unit ball, it follows that the closed unit ball is weakly compact and so it follows that the space is reflexive, by Theorem 3.1.

Remark 4.2. When we stated James' theorem, we said that every continuous linear functional should attain its norm on the unit sphere. Now the unit sphere is not a weakly closed set, as mentioned earlier, but the closed unit ball is weakly closed and we can apply the James' compactness criterion mentioned above. But the unit sphere is a subset of the closed unit ball. Now, if a continuous linear functional, f, attains its norm at v, where $\|v\| \leq 1$, then

$$\|f\| \;=\; |f(v)| \;\leq\; \|f\|\,\|v\| \;\leq\; \|f\|,$$

and so we have that $\|f\| = \|f\|\,\|v\|$, whence we deduce that, in fact $\|v\| = 1$. Thus, the supremum of a continuous linear functional over the closed unit ball can only be attained on the unit sphere. ■

5 Reflexivity and bounded sequences

In a metric space, a sequence in a compact set will have a convergent subsequence. This is not true in a general topological space. However, this is true for the weak topology. We will prove a related result for reflexive spaces.

Recall that a topological space is said to be *separable* if it contains a countable dense set. It can be shown that the spaces ℓ_p are all separable for $1 \leq p < \infty$ and that ℓ_∞ is not separable. Thanks to the Weierstrass approximation theorem, we can easily show that $C[0,1]$ is separable.

We now will state some important facts about separability in Banach spaces. For proofs, see, for instance, Kesavan [2].

- Let V be a Banach space. If V^* is separable, then V is separable. The converse is not true: for example, ℓ_1 is separable, but $\ell_1^* = \ell_\infty$ is not.

- V is reflexive and separable if, and only if, V^* is reflexive and separable.

Let V be separable and let $\{v_k\}_{k=1}^\infty$ be a countable dense set in V. Without loss of generality, we can assume that all the v_k are non-zero (why?). Define, for f and g in B^*,

$$d(f,g) \;=\; \sum_{k=1}^{\infty} \frac{1}{2^k \|x_k\|} |(f-g)(x_k)|.$$

Then, one can show that this defines a metric on B^* and that the corresponding metric topology on B^* is the same as the topology induced on B^* by the weak* topology on V^*. In other words, if V is separable, then the weak* topology restricted to B^* is metrizable. Since B^* is weak* compact, it follows then that every sequence in B^* will have a weak* convergent subsequence. More generally, any bounded sequence in V^* will have a weak* convergent subsequence.

We will exploit this result in the following manner. Let V be reflexive and let $\{v_n\}$ be a bounded sequence in V. Set

$$W = \overline{\text{span}\{\{v_n\}_{n=1}^\infty\}},$$

i.e., W is the smallest closed subspace of V containing all the v_n. Then, as we saw in the previous section, W is also reflexive. By construction, it is separable as well (why?). Again, it follows from the facts we have stated above that W^* is also reflexive and separable. Then, every bounded sequence in W^{**} will have a weak* convergent subsequence. But the weak* and weak topologies on W^{**} are the same, since W^* is reflexive. Thus $\{J(v_n)\}$ has a weakly convergent subsequence and, since J^{-1} is an isometry, it is continuous and hence weakly continuous, and so $\{v_n\}$ has a weakly convergent subsequence.

Let us now assume that V is a Banach space such that every bounded sequence admits a weakly convergent subsequence. Let $f \in V^*$. Then, there exists a sequence $\{v_n\}$ in B such that $|f(v_n)| \to \|f\|$, by definition of the norm. Thus, there exists a weakly convergent subsequence, say, $\{v_{n_k}\}$. Let the weak limit be v. Since B is weakly closed, we have that $v \in B$. Further, by the weak convergence, it follows that $f(v_{n_k}) \to f(v)$.

Consequently, $|f(v)| = \|f\|$. Since we arbitrarily chose $f \in V^*$, we see that every continuous linear functional attains its norm in B and so, by James' theorem, V is reflexive.

The above arguments can be summarized in the following statement.

Theorem 5.1. *(Eberlein-Šmulian) A Banach space V is reflexive if, and only if, every bounded sequence admits a weakly convergent subsequence.* ■

Example 5.1. Let $V = C[0,1]$ and let $W = C^1[0,1]$ (equipped with their usual norms). Consider $T \in \mathcal{L}(V,W)$ defined by

$$T(f)(t) \;=\; \int_0^t f(s)\, ds,\; t \in [0,1], f \in V,$$

and $S \in \mathcal{L}(W,V)$ defined by $S(f) = f',\ f \in W$. Then $S \circ T$ is the identity map in V. Assume that W is reflexive. Let $\{f_n\}$ be a bounded sequence in V. Then $\{T(f_n)\}$ will be a bounded sequence in W and so, it will admit a weakly convergent subsequence, say, $\{T(f_{n_k})\}$. Since S is continuous, it is weakly continuous as well and so $\{S(T(f_{n_k}))\}$ will be weakly convergent in V. In other words, $\{f_n\}$ admits a weakly convergent subsequence $\{f_{n_k}\}$. But this implies that V is reflexive, which we know is false. Thus $W = C^1[0,1]$ is not reflexive. ■

We will apply these results to certain optimization problems and obtain a further characterization of reflexive spaces.

6 Proximal points in closed convex sets

Let V be a Banach space and let $K \subset V$ be a closed convex set. Let $x \in V$. A proximal point of x in K is a vector $y \in K$ which is closest to x, i.e., $y \in K$ (if it exists) satisfies the relation

$$\|x - y\| \;=\; \min_{z \in K} \|x - z\|. \tag{6.1}$$

One of the first results one proves in this context is that in a Hilbert space (i.e., a Banach space whose norm comes from an inner-product), given any closed convex set, there exists a unique proximal point for each vector. The existence of a proximal point can be proved for reflexive spaces, though we may not have uniqueness. We now do this.

Definition 6.1. *Let V be a Banach space. Let $\varphi : V \to \mathbb{R}$ be a given mapping. We say that φ is coercive if*

$$\lim_{\|x\| \to +\infty} \varphi(x) \ = \ +\infty.$$

We say that φ is weakly sequentially lower semi-continuous if whenever we have a sequence $\{x_n\}$ converging weakly in V to x, then

$$\varphi(x) \ \leq \ \liminf_{n \to \infty} \varphi(x_n). \ \blacksquare$$

Notation: If a sequence $\{x_n\}$ converges to x weakly in a Banach space V, we write $x_n \rightharpoonup x$.

For example the mapping $x \mapsto \|x\|$ is clearly coercive. It is also weakly sequentially lower semi-continuous. Indeed, if $x_n \rightharpoonup x$ in V, then, for every $f \in V^*$, we have $f(x_n) \to f(x)$. In other words, if $J : V \to V^{**}$ is the canonical imbedding, be have $J(x_n)(f) \to J(x)(f)$ for every $f \in V^*$. Now, $|J(x_n)(f)| \leq \|J(x_n)\|\|f\| = \|x_n\|\|f\|$. Passing to the limit, we deduce that

$$|J(x)(f)| \ \leq \ \liminf_{n \to \infty} \|x_n\| \ \|f\|,$$

from which we deduce that

$$\|x\| = \|J(x)\| \leq \liminf_{n \to \infty} \|x_n\|.$$

Theorem 6.1. *Let V be a reflexive Banach space and let $K \subset V$ be a closed convex subset. Let $x \in V$. Then, there exists $y \in K$ such that (6.1) holds.*

Proof. We may assume that $x \notin K$ (for, otherwise, we have, trivially, $y = x$). In that case,

$$0 \ < d \ = \ \inf_{z \in K} \|x - z\| \ < \ +\infty.$$

The mapping $z \mapsto \|x - z\|$ is coercive and weakly sequentially lower semi-continuous, as we can easily prove by the the arguments given previously. Let $\{y_n\}$ be a minimizing sequence in K, i.e.,

$$\|x - y_n\| \ \to d, y_n \in K.$$

By the coercivity of the norm, it follows that $\{y_n\}$ is bounded in V. Since V is reflexive, there exists a weakly convergent subsequence $\{y_{n_k}\}$. Let y be the weak limit of this subsequence. Then, on one hand, since K is closed and convex, it is weakly closed (this is a consequence of the Hahn-Banach theorem, as already observed in Section 3) and so $y \in K$. On the other hand, by the weak sequential lower semi-continuity, we have that

$$\inf_{z \in K} \|x - z\| \ \leq \ \|x - y\| \ \leq \ \liminf_{k \to \infty} \|x - y_{n_k}\| \ = \ \inf_{z \in \widetilde{K}} \|x - z\|.$$

Thus, $y \in K$ and

$$\|x - y\| \;=\; \inf_{z \in K} \|x - z\|.$$

This completes the proof. ■

Remark 6.1. The same proof can easily be modified to show that if $\varphi : V \to \mathbb{R}$ is a coercive and weakly sequentially lower semi-continuous mapping, then it attains a minimim on every closed convex subset K of V (cf. Kesavan [2]). ■

Definition 6.2. *A Banach space is uniformly convex if for every $\varepsilon > 0$, there exists $\delta > 0$, such that if $\|x\| = \|y\| = 1$ and if $\|x - y\| > \varepsilon$, then*

$$\left\| \frac{x+y}{2} \right\| \;<\; 1 - \delta.$$

The above definition quantifies the fact that the unit ball 'bulges uniformly in all directions'. In particular there are no 'flat portions' on the unit sphere. The spaces ℓ_1 and ℓ_∞, and their finite-dimensional counterparts, are not uniformly convex. Every Hilbert space is uniformly convex as can be easily deduced from the paralellogram identity:

$$\left\| \frac{x+y}{2} \right\|^2 + \left\| \frac{x-y}{2} \right\|^2 \;=\; \frac{1}{2}(\|x\|^2 + \|y\|^2).$$

The space ℓ_2 is a Hilbert space. If $2 \leq p < \infty$, then we have Clarkson's inequality (cf. Kesavan [2]) which generalizes the parallelogram identity: for every x and y in ℓ_p, we have

$$\left\| \frac{x+y}{2} \right\|_p^p + \left\| \frac{x-y}{2} \right\|_p^p \;\leq\; \frac{1}{2}(\|x\|_p^p + \|y\|_p^p).$$

Thus, the spaces $\ell_p, 2 \leq p < \infty$ are all uniformly convex.

It is known that every uniformly convex space is reflexive (see, for example, Kesavan [2]). The converse, of course, is not true: $\mathbb{R}^N$ with the norm $\|\cdot\|_1$ or $\|\cdot\|_\infty$ is not uniformly convex, though it is reflexive, since all finite dimensional spaces are reflexive.

Theorem 6.2. *Let V be a uniformly convex Banach space and let $K \subset V$ be a closed convex set. Then, for every $x \in V$, there is a unique proximal point of x in K.*

Proof. The existence follows from the reflexivity of uniformly convex spaces. Let, if possible, there exist $y_1, y_2 \in K$ such that

$$\alpha \;=\; \min_{z \in K} \|x - z\| \;= \|x - y_1\| \;=\; \|x - y_2\|.$$

let $\|y_1 - y_2\| > \varepsilon > 0$. Then, by uniform convexity, there exists $\delta > 0$ such that

$$\left\| x - \frac{y_1 + y_2}{2} \right\| \;=\; \left\| \frac{(x - y_1) + (x - y_2)}{2} \right\| \;<\; \alpha(1 - \delta).$$

But since K is convex, $(y_1+y_2)/2 \in K$ and the above relation contradicts the minimality of α. Thus $y_1 = y_2$. ■

In case the proximal point is unique for every x, it is also referred to as the *projection* of x onto K.

Example 6.1. We cannot guarantee uniqueness if the space is not uniformly convex. Consider $\mathbb{R}^2$ equipped with the norm $\|\cdot\|_1$. Then, we saw that it is not uniformly convex but, being finite-dimensional, it is reflexive. Let K be the closed unit ball in this space. Consider the point $x = (1, 1)$. Then, if $z = (a, b) \in K$, we have $|a| + |b| \leq 1$ and

$$\|x - z\|_1 \;=\; |1 - a| + |1 - b| \;\geq\; 1 - |a| + 1 - |b| \;\geq 1.$$

Now if we consider points on the boundary of the unit ball of the form $z = (a, b)$ where $a \geq 0, b \geq 0$ and $a + b = 1$, we get that

$$\|x - z\| \;= 1 - a + 1 - b \;= 1.$$

Thus we have an uncountable number of points which minimize the distance of x to K. ■

If V is not reflexive we cannot even guarantee existence. We can sharpen this statement further in the result below. This is where we need James' theorem.

Theorem 6.3. *If V is a non-reflexive Banach space, then there always exists a point $x \in V$ and a closed convex set K such that there is no minimizer in K to the function $z \mapsto \|x - z\|$, where $z \in K$.*

Proof. By James' theorem, there exists $f \in V^*, f \neq 0$ such that $\|f\|$ is not attained on the unit sphere. Without loss of generality, we can assume that $\|f\| = 1$. Now define

$$K \;=\; \{z \in V \mid f(z) = 1\}.$$

By the continuity and linearity of f, it follows immediately that K is a closed convex set. Let us take $x = 0$. If $z \in K$, then

$$1 \;=\; f(z) \;\leq \|f\|\, \|z\|.$$

Thus $\|z\| \geq 1$ for all $z \in K$. Now, there exists a sequence $\{x_n\}$ such that $\|x_n\| = 1$ for all n and such that $f(x_n) \to \|f\| = 1$. Thus $f(x_n)$ is non-zero for large enough n. Define $z_n = x_n/f(x_n)$ so that $z_n \in K$. Further $\|z_n\| \to 1$. Thus

$$\inf_{z \in K} \|z\| \;=\; 1.$$

But there is no $z \in K$ such that $\|z\| = 1$ since $\|f\|$ is not attained on the unit sphere. ■

Thus, a Banach space is reflexive if, and only if, every closed convex set admits a proximal point for every point in the space.

7 Quotient space

Let V be a normed linear space and let W be a closed subspace. Then, the quotient space V/W is the collection of all cosets $x + W$ where $x \in V$ and

$$x + W \;=\; \{x + w \mid w \in W\}.$$

This is a vector space with addition and scalar multiplication defined by

$$(x + W) + (y + W) \;=\; (x + y) + W, \text{ and } \alpha(x + W) \;=\; \alpha x + W,$$

where $x, y \in V$ and α is a scalar. This space is endowed with the norm defined by

$$\|x + W\|_{V/W} \;=\; \inf\{\|x + w\| \mid w \in W\}. \tag{7.1}$$

In other words, the norm of $x + W$ in the quotient space is just the distance of x from the subspace W. A natural question to ask is that if there exists a vector $w \in W$ such that $\|x + W\|_{V/W} = \|x + w\|$.

Since a closed subspace is automatically a closed convex set, the answer to this question is affirmative if the space V is reflexive. If V is any normed linear space and if W is a finite dimensional subspace (which is then automatically closed), again the answer is affirmative. For, if $\{w_n\}$ is a sequence in W such that

$$\|x + w_n\| \;\to\; \|x + W\|_{V/W},$$

then, the sequence being bounded, admits a convergent subsequence, since W is finite dimensional. Let $w_{n_k} \to w$ in W. Then $\|x + W\|_{V/W} = \|x + w\|$.

Thus, if we wish to give an example where the infimum is not attained in (7.1), we need to look for an infinite dimensional subspace W of a non-reflexive space V. To do this we appeal to the following result (cf. Nair [3], for instance).

Theorem 7.1. *Let V be a normed linear space and let f be a non-zero linear functional defined on V. If the null-space of f,*

$$N(f) \;=\; \{x \in V \mid f(x) = 0\},$$

is closed in V, then f is continuous. Further, in this case, if $x_0 \notin N(f)$, we have

$$\|f\| \;=\; \frac{|f(x_0)|}{\|x_0 + N(f)\|_{V/N(f)}}. \tag{7.2}$$

Proof. First of all, since $N(f)$ is closed, if $x_0 \notin N(f)$, the distance of x_0 from $N(f)$ is strictly positive, i.e., $\|x_0 + N(f)\|_{V/N(f)} > 0$. Since $f(x_0) \neq 0$, we have that, for any $x \in V$,

$$x - \frac{f(x)}{f(x_0)}x_0 \;\in\; N(f).$$

Consequently,

$$\|x + N(f)\|_{V/N(f)} \;=\; \left\|\frac{f(x)}{f(x_0)}x_0 + N(f)\right\|_{V/N(f)} \;=\; \frac{|f(x)|}{|f(x_0)|}\|x_0 + N(f)\|_{V/N(f)}. \tag{7.3}$$

By definition of the quotient norm (cf. (7.1)), we have that $\|x + N(f)\|_{V/N(f)} \leq \|x\|$. Thus, it follows from (7.3) that

$$|f(x)| \;\leq\; \frac{|f(x_0)|}{\|x_0 + N(f)\|_{V/N(f)}}\|x\|,$$

from which we see immediately that f is continuous and that

$$\|f\| \;\leq\; \frac{|f(x_0)|}{\|x_0 + N(f)\|_{V/N(f)}}.$$

Now, since $f(x_0) = f(x_0 + w)$ for every $w \in N(f)$, we have, by the continuity of f, $|f(x_0)| \leq \|f\|\ \|x_0 + w\|$, whence it follows, on taking the infimum over all $w \in N(f)$, that

$$|f(x_0)| \;\leq\; \|f\|\ \|x_0 + N(f)\|_{V/N(f)}.$$

This gives the reverse inequality and establishes (7.2). ∎

Corollary 7.1. *Let V be a normed linear space and let f be a non-zero continuous linear functional on V. Let $W = N(f)$. If the infimum is attained in (7.1) for some $x_0 \in V \backslash N(f)$, then f attains its norm in the unit sphere.*

Proof. Let $x_0 \notin N(f)$. Let $w \in N(f)$ be such that $\|x_0 + w\| = \|x_0 + N(f)\|_{V/N(f)}$. Then by (7.2), we get

$$\|f\| \;=\; \frac{|f(x_0)|}{\|x_0 + N(f)\|_{V/N(f)}} \;=\; \frac{|f(x_0 + w)|}{\|x_0 + w\|},$$

which shows that f attains its norm for the unit vector $(x_0 + w)/\|x_0 + w\|$. ■

This gives another proof of Theorem 6.3: we consider a functional f which does not realise its norm and then if $x_0 \notin N(f)$, then there is no proximal point of x_0 in the closed convex set $N(f)$. The same gives us an example of the non-existence of a minimum in (7.1). Thus, we can consider any of the examples in Section 2, to produce a non-reflexive space, and an infinite dimensional subspace thereof, so that for any point not in that subspace, the infimum in (7.1) will not be attained.

To summarize, the following statements are equivalent.

- A Banach space is reflexive.
- (James) Every continuous linear functional attains its norm in the closed unit ball.
- The closed unit ball is weakly compact.
- (Eberlein-Šmulian) Every bounded sequence has a weakly convergent subsequence.
- Every point has a proximal point in every closed convex set.

Acknowledgements

The author thanks Prof. M. T. Nair for useful discussions and for drawing his attention to Theorem 7.1 and its corollary.

Bibliography

[1] James, R. C. Reflexivity and the supremum of linear functionals, Ann. Math., **66**(1), 1957, pp. 159-169.

[2] Kesavan, S, Functional Analysis, TRIM Series No. 52, Hindustan Book Agency, New Delhi, 2009.

[3] Nair, M. T. Functional Analysis: A First Course, Second Edition, PHI Learning, 2021.

4 Alternating Sign Matrices

Mallika Muralidharan and A. Satyanarayana Reddy

Department of Mathematics
Shiv Nadar University
Gautam Buddha Nagar – 201314

Email: mm238@snu.edu.in, satya.a@snu.edu.in

1 Introduction

A matrix is a rectangular array of numbers, symbols or expressions arranged in rows and columns. In this article, we will study a very special kind of matrix, known as the *alternating sign matrix*. An alternating sign matrix (ASM) is a square matrix with elements from $\{-1, 0, 1\}$ with the following special properties:

1. The elements of each row sum up to one.
2. The elements of each column sum up to one.
3. The non-zero entries in each row/column alternate in sign.

An example of an ASM is

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

They are often considered as a generalization of the well-known *permutation matrices*. A permutation matrix has elements from $\{0, 1\}$ such that each row and each column has only one 1. There are exactly $n!$ permutation matrices of order n. It is easy to see that the first row and first column of every ASM matrix contain exactly one nonzero entry 1. Further, the difference of number of 1's and number of -1's in each row and each

column is 1. The only 2×2 ASM matrices are

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Thus for $n = 2$, ASM matrices and permutation matrices coincide. All seven ASM matrices of order 3 are given below.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

These matrices currently might seem like they were defined for no particular reason; however, ASMs arise naturally from an algorithm used to calculate determinants, known as the *Dodgson condensation algorithm*.

2 Dodgson Condensation Algorithm

Given a matrix, how can one calculate its determinant? There is one method you may already be familiar with: the Laplace formula, which defines the determinant of a 2×2 matrix as

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

and of a 3×3 matrix as

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix},$$

and so on. However, there is another method, one that is almost never taught at the school level but is remarkably efficient for large matrices, known as the *Dodgson condensation algorithm*, which was devised in 1866 by the Reverend C.L. Dodgson, better known by his pen-name *Lewis Carroll*.

2.1 The algorithm

Begin with an $n \times n$ matrix, M.

1. Make sure that the interior (obtained by deleting the first and last row as well as the first and last column) has no zeros in it. If it does, then perform appropriate row and column operations such that the value of $\det M$ doesn't change.

2. Create an $(n-1)\times(n-1)$ matrix, N, where

$$n_{ij} = \begin{vmatrix} m_{ij} & m_{ij+1} \\ m_{i+1\,j} & m_{i+1\,j+1} \end{vmatrix}.$$

3. Use this to create an $(n-2)\times(n-2)$ matrix, P, such that

$$p_{ij} = \frac{\begin{vmatrix} n_{ij} & n_{ij+1} \\ n_{i+1\,j} & n_{i+1\,j+1} \end{vmatrix}}{m_{i+1\,j+1}}.$$

4. Let $M = N$, and let $N = P$. Repeat step 3 until a 1×1 matrix is obtained, whose only entry would be $\det M$.

Let's demonstrate this method using an example. Let

$$M = \begin{bmatrix} 3 & 5 & 2 \\ 1 & 8 & 9 \\ 2 & 3 & 6 \end{bmatrix}.$$

Then

$$\begin{vmatrix} 3 & 5 \\ 1 & 8 \end{vmatrix} = 24-5 = 19; \begin{vmatrix} 5 & 2 \\ 8 & 9 \end{vmatrix} = 45-16 = 29; \begin{vmatrix} 1 & 8 \\ 2 & 3 \end{vmatrix} = 3-16 = -13; \begin{vmatrix} 8 & 9 \\ 3 & 6 \end{vmatrix} = 48-27 = 21.$$

Using this, we obtain a new 2×2 matrix

$$N = \begin{bmatrix} 19 & 29 \\ -13 & 21 \end{bmatrix}.$$

Its determinant is 776. However, at this stage, we need to divide 776 by the element in the second row and second column of M, i.e., 8, which gives us a final answer of $\Delta = 97$. This is the determinant of M. One can verify this by using the Laplace method. Note that the algorithm does not work if, at any stage, $m_{i+1,j+1} = 0$. This is a drawback of the algorithm. Further, from the algorithm it is by no means obvious that the determinant of a matrix of integers is an integer. Recently, a few papers have tried to overcome these shortcomings (see [3]).

2.2 The link to alternating sign matrices

Now, let's use this algorithm on a symbolic 3×3 matrix,

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}.$$

Assuming $e \neq 0$, we see that

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \to \begin{bmatrix} ae - bd & bf - ce \\ dh - eg & ei - fh \end{bmatrix} \to \left[\frac{(ae - bd)(ei - fh) - (bf - ce)(dh - eg)}{e} \right].$$

Expand the value of the determinant obtained at the last stage, taking care *not* to cancel any elements out. You can verify that the following expression will be obtained:

$$\frac{ae^2i - aefh - bdei + bdfh - bdfh + befg + cdeh - ce^2g}{e}$$
$$= (1)aei + (-1)afh + (-1)bdi + (0)bde^{-1}fh + (1)bfg + (1)cdh + (-1)ceg.$$

Now, here comes the interesting part. What if we were to take each term in this seven-term expansion and create a new matrix out of it? Take each term and substitute the power of each variable in the original 3×3 matrix: for instance, $(1)aei = (1)a^1b^0c^0d^0e^1f^0g^0h^0i^1$, so it corresponds to the 3×3 identity matrix. We do the same for the five terms $(-1)afh, (-1)bdi, (1)bfg, (-1)cdh, (1)ceg$. What we get are the six 3×3 permutation matrices (mentioned in the first part), and the values of their determinants correspond to the coefficients of their terms!

The same works for $(0)bde^{-1}fh$. If you work out its determinant, it will come to 0 – the value of the coefficient.

If you notice, all seven matrices thus obtained are ASMs. In fact, these are all of the 3×3 ASMs!

ASMs have been recast into different ways, such as domino tilings, descending plane partitions, the arrangement of unit cubes in a corner, etc. To know these connections and more about this topic, refer to [1]. We illustrate one such connection.

For an initial example, let us take the 3×3 alternating sign matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Taking the partial sums of the rows, $(R_1, R_1 + R_2, R_1 + R_2 + R_3)$, we get

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

This matrix corresponds to the *monotone triangle* below, obtained by noting the positions of the non-zero (unit) entries. In the first row, there is only one 1, in the second

position. So the top row of the triangle is just 2.

$$\begin{array}{ccccc} & & 2 & & \\ & 1 & & 3 & \\ 1 & & 2 & & 3 \end{array}$$

In fact, the seven monotone triangles corresponding to each alternating sign matrix of order 3×3 are

$$\begin{array}{ccccc} & & 1 & & \\ & 1 & & 2 & \\ 1 & & 2 & & 3 \end{array} \qquad \begin{array}{ccccc} & & 1 & & \\ & 1 & & 3 & \\ 1 & & 2 & & 3 \end{array} \qquad \begin{array}{ccccc} & & 2 & & \\ & 1 & & 2 & \\ 1 & & 2 & & 3 \end{array} \qquad \begin{array}{ccccc} & & 2 & & \\ & 1 & & 3 & \\ 1 & & 2 & & 3 \end{array}$$

$$\begin{array}{ccccc} & & 2 & & \\ & 2 & & 3 & \\ 1 & & 2 & & 3 \end{array} \qquad \begin{array}{ccccc} & & 3 & & \\ & 1 & & 3 & \\ 1 & & 2 & & 3 \end{array} \qquad \begin{array}{ccccc} & & 3 & & \\ & 2 & & 3 & \\ 1 & & 2 & & 3 \end{array}$$

If one observes the triangles closely, a few general properties are seen:

1. The bottom row of a monotone triangle created from an $n \times n$ matrix (henceforth, this will be a *monotone triangle of order n*) is always $1 \quad 2 \quad 3 \quad \ldots \quad n$.
2. There is a strict increase across the rows.
3. There is a weak increase up or down to the right.

Definition 2.1. *A triangular array of numbers with n entries on each edge and which satisfies properties 1 to 3 as above is known as a monotone triangle of order n.*

3 Counting ASMs

In this section, we address an interesting question which has historical importance: For a given $n \in \mathbb{N}$, how many ASMs of order n are there? Let A_n denote the number of ASMs of order n. Since all permutation matrices are ASMs, $n! \leq A_n$. The following table suggests that A_n grows much faster than $n!$.

n	1	2	3	4	5	6	7	8	9
$n!$	1	2	6	24	120	720	5040	40320	362880
A_n	1	2	7	42	429	7436	218348	10850216	911835460

But can we find a simple formula that we can use to count ASMs? Let us try to refine the counting. As we noticed earlier, the first row of any ASM matrix contains exactly one 1. This property greatly simplifies the way in which we can count ASMs. Let us define $A_{n,k}$ as the number of $n \times n$ ASMs with a 1 in column k, $1 \leq k \leq n$, of the first row. Then $A_n = \sum_{k=1}^{n} A_{n,k}$.

3.1 Mills, Robbins and Rumsey: The ASM Conjecture [4]

The mathematicians William Mills, David Robbins and Howard Rumsey thought of putting the values of $A_{n,k}$ in a Pascal's triangle of sorts. This seems like a natural place to go, since the symbol $A_{n,k}$ does suggest some sort of triangular array.

n = 1						1					
n = 2					1		1				
n = 3				2		3		2			
n = 4			7		14		14		7		
n = 5		42		105		135		105		42	
n = 6	429		1287		2002		2002		1287		429

With some observation and tinkering, you'll see that this triangle is symmetric: $A_{n,k} = A_{n,n-k+1}$, and that $A_{n+1} = \sum_{k=1}^{n} A_{n,k}$, and $A_n = A_{n+1,1}$. (Try and see it for yourself!)

Robbins and Rumsey then took ratios of adjacent terms in order to look for some pattern.

n = 1						1					
n = 2					1	$\frac{2}{2}$	1				
n = 3				2	$\frac{2}{3}$	3	$\frac{3}{2}$	2			
n = 4			7	$\frac{2}{4}$	14	$\frac{5}{5}$	14	$\frac{4}{2}$	7		
n = 5		42	$\frac{2}{5}$	105	$\frac{7}{9}$	135	$\frac{9}{7}$	105	$\frac{5}{2}$	42	
n = 6	429	$\frac{2}{6}$	1287	$\frac{9}{14}$	2002	$\frac{16}{16}$	2002	$\frac{14}{9}$	1287	$\frac{6}{2}$	429

Do you see a pattern here? If you look closely, you'll see that the numerators come from

a Pascal's triangle with the second line being $(2, 1)$, and the denominators come from a similar Pascal's triangle with the second line being $(1, 2)$, the elements of which aren't 1.

So the numerators decompose as

$$\begin{array}{ccccccccccccc}
 & & & & & & 1{+}1 & & & & & & \\
 & & & & 1{+}1 & & & & 1{+}2 & & & & \\
 & & & 1{+}1 & & & 2{+}3 & & & & 1{+}3 & & \\
 & & 1{+}1 & & & 3{+}4 & & & 3{+}6 & & & 1{+}4 & \\
 & 1{+}1 & & & 4{+}5 & & & 6{+}10 & & & 4{+}10 & & 1{+}5 \\
1{+}1 & & & 5{+}6 & & & 10{+}15 & & & 10{+}20 & & & 5{+}15 & & 1{+}6
\end{array}$$

while the denominators decompose as

$$\begin{array}{ccccccccccccc}
 & & & & & & 1{+}1 & & & & & & \\
 & & & & 1{+}2 & & & & 1{+}1 & & & & \\
 & & & 1{+}3 & & & 2{+}3 & & & & 1{+}1 & & \\
 & & 1{+}4 & & & 3{+}6 & & & 3{+}4 & & & 1{+}1 & \\
 & 1{+}5 & & & 4{+}10 & & & 6{+}10 & & & 4{+}5 & & 1{+}1 \\
1{+}6 & & & 5{+}15 & & & 10{+}20 & & & 10{+}15 & & & 5{+}6 & & 1{+}1.
\end{array}$$

Exercise 3.1. *Try and figure out what the ratio* $\frac{A_{n,k}}{A_{n,k+1}}$ *might look like. For instance, try and calculate what* $\frac{A_{6,2}}{A_{6,3}}$ *and what* $\frac{A_{5,1}}{A_{5,2}}$ *are.*

If you use the above calculations, hopefully you can see that $\dfrac{A_{n,k}}{A_{n,k+1}} = \frac{\binom{n-2}{k-1}+\binom{n-1}{k-1}}{\binom{n-2}{k-1}+\binom{n-1}{k}}$.

This was part of the puzzle known as the **ASM Conjecture** till 1992, which is when the mathematician Doron Zeilberger [5] proved it. In fact, the above equation of the ratio is known as the **Refined ASM Conjecture**. The ASM theorem states that the number of $n \times n$ alternating sign matrices is

$$A_n = A_{n+1,1} = \prod_{j=0}^{n-1} \frac{(3j+1)!}{(n+j)!}. \tag{1}$$

What seems to be an entirely simple conjecture actually remained unproved for several years.

It was after Zeilberger's proof that other proofs began to show up, demonstrating some remarkable places where ASMs appear: Greg Kuperberg [2] presented a proof using statistical mechanics, when he learnt that physicists had also been studying ASMs,

but in connection with the structure of water ('square ice'). By using Kuperberg's observations, Doron Zeilberger [6] proved the Refined ASM Conjecture.

We hope this excursion lent an insight into a very fascinating mathematical object. We also hope this lent a new insight into mathematics beyond school, and piqued your interest in mathematics for the future.

Acknowledgements

We would like to thank the referee for valuable comments.

Bibliography

[1] D.M. Bressoud, *Proofs and Confirmations: The Story of the Alternating Sign Matrix Conjecture,* Cambridge University Press, 1999.

[2] Greg Kuperberg, *Another proof of the alternative-sign matrix conjecture*, International Mathematics Research Notices, Volume 1996, Issue 3, (1996), 139–150.

[3] Hou-biao Li, Hong Li, Ting-zhu Huang, *A note on Dodgson's determinant condensation algorithm*, https://arxiv.org/abs/1907.12010.

[4] W. H. Mills, D. P. Robbins, and H. Rumsey, *Alternating sign matrices and descending plane partitions*, J. Combin. Theory Ser. A 34 (1983), 340–359.

[5] D. Zeilberger, *Proof of the Alternating Sign Matrix Conjecture,* Electronic J. Combinatorics 3, No. 2, R13, (1996a), 1–84.

[6] D. Zeilberger, *Proof of the Refined Alternating Sign Matrix Conjecture,* New York J. Math. 2 (1996), 59–68.

5 Prime gaps and cyclotomic polynomials[1]

Pieter Moree

Max Planck Institute for Mathematics
Bonn

Abstract: We investigate which numbers occur as the maximum coefficient of some cyclotomic polynomial and relate this to large gaps between consecutive prime numbers. We also make a connection with variants of the problem whether there exist infinitely many Sophie Germain primes (primes p such that $2p+1$ is also prime). This is an informal account of a very recent joint research paper with Kosyak, Sofos and Zhang [26], where deep methods and results of prime number theory are used to make progress on the cyclotomic problem.

1 Cyclotomic polynomials: basics

It is clear that $X^2-1=(X-1)(X+1), X^3-1=(X-1)(X^2+X+1)$ and $X^4-1=(X-1)(X+1)(X^2+1)$. Over the rationals none of the factors can be factorized further and the expressions give the factorization into *irreducibles*. However, it is not so obvious how to factorize X^n-1 for an *arbitrary* integer $n\geq 1$ into *irreducibles* over the rationals in a systematic way.

Over the *complex numbers* the answer is easy:

$$X^n-1=\prod_{m=1}^{n}(X-e^{\frac{2\pi im}{n}}). \tag{1.1}$$

The roots are the n-th roots of unity and these divide the circle into equal parts.

[1]Republished with permission from the Nieuw Archief Voor Wiskunde.

The word *cyclotomy* comes from ancient Greek and literally means circle-cutting. A root of unity ζ is said to be a *primitive* n-th root of unity if it satisfies $\zeta^n = 1$, but not $\zeta^d = 1$ for any $1 \leq d < n$. For any two integers n and d by the Euclidean algorithm we can find integers a and b such that $an + bd = \gcd(n, d)$, where gcd is a shorthand for *greatest common divisor*. Thus if $\zeta^n = 1$ and $\zeta^d = 1$, it follows that $\zeta^{\gcd(n,d)} = 1$. Therefore, in order to check that ζ is a primitive n-th root of unity, it suffices to check that $\zeta^n = 1$ and $\zeta^d \neq 1$ for every proper divisor d of n. By a similar argument one deduces that if ζ is a primitive n-th root of unity, then ζ^j is of order $n/\gcd(j, n)$. It follows that all the primitive n-th roots of unity are of the form ζ^j, with $1 \leq j \leq n$ and $\gcd(j, n) = 1$. There are precisely $\varphi(n)$ primitive n-th roots of unity, where φ is the *Euler totient function*, which is defined as

$$\varphi(n) = \sum_{\substack{j=1 \\ (j,n)=1}}^{n} 1.$$

An obvious primitive n-th root of unity is $e^{2\pi i/n}$.

The n-th *cyclotomic polynomial* can be defined as

$$\Phi_n(X) = \prod_{\substack{j=1 \\ (j,n)=1}}^{n} (X - e^{\frac{2\pi ij}{n}}). \tag{1.2}$$

It thus has precisely the n-th order primitive roots of unity as its simple roots. (Note that of all Greek letters Φ looks the most like a cut circle.) The degree of $\Phi_n(X)$ is $\varphi(n)$ and we have $\Phi_n(x) = X^{\varphi(n)} + \ldots$.

By reducing the fractions m/n in (1.1) (e.g., $4/6 = 2/3$), we see that for each divisor d of n there are $\varphi(d)$ reduced fractions with denominator d. These correspond to roots of unity of order d. We thus infer from (1.1) and (1.2) that

$$X^n - 1 = \prod_{d|n} \Phi_d(X). \tag{1.3}$$

Setting $n = 1$ we get $\Phi_1(X) = X - 1$. In case $n = p$ is a prime, we obtain

$$\Phi_p(X) = X^{p-1} + X^{p-2} + \ldots + X + 1.$$

It can be shown that all cyclotomic polynomials have integer coefficients and are irreducible, and so (1.3) gives the factorization of $X^n - 1$ into irreducibles over the rationals. Indeed, many famous mathematicians gave proofs of the irreducibility of the cyclotomic polynomials (Gauss, Kronecker, Eisenstein, Dedekind, Landau, Schur, …). For some of these proofs, see Weintraub [46]. The (very short) proof of Schur was even set to rhyme! (Cremer [14, p. 39-41]).

n	$\Phi_n(x)$
5	$x^4+x^3+x^2+x+1$
12	x^4-x^2+1
15	$x^8-x^7+x^5-x^4+x^3-x+1$
16	x^8+1
60	$x^{16}+x^{14}-x^{10}-x^8-x^6+x^2+1$
105	$x^{48}+x^{47}+x^{46}-x^{43}-x^{42}-2x^{41}-x^{40}-x^{39}+\ldots+1$
210	$x^{48}-x^{47}+x^{46}+x^{43}-x^{42}+2x^{41}-x^{40}+x^{39}+\ldots+1$
240	$x^{64}+x^{56}-x^{40}-x^{32}-x^{24}+x^8+1$

Table 1: Some cyclotomic polynomials

Write

$$\Phi_n(x)=\sum_{j=0}^{\varphi(n)} a_n(j)x^j. \tag{1.4}$$

For $j>\varphi(n)$ we put $a_n(j)=0$. We define

$$A(n)=\max_{k\geq 0}|a_n(k)|, \quad A\{n\}=\{a_n(k):k\geq 0\},$$

and call $A(n)$ the *height* of Φ_n. Note that, for example, $A\{105\}=\{-2,-1,0,1\}$, see Table 1. Our interest is in the possible heights $A(n)$ and extrema of $A\{n\}$ as n runs over the integers.

The cyclotomic coefficients $a_n(j)$ are usually very small. Indeed, in the 19-th century mathematicians even thought that they are always 0 or ± 1. The first counterexample to this claim occurs at $n=105$; we have $a_{105}(41)=a_{105}(7)=-2$. Issai Schur in a letter to Edmund Landau proved that every negative even number occurs as a coefficient of some cyclotomic polynomial. Emma Lehmer [29] reproduced Schur's argument, which is easily adapted to show that *every* integer is assumed as value of a cyclotomic coefficient [44]. For the best result to date in this direction, see Fintzen [17].

Nowadays computations can be extended enormously far beyond $n=105$, cf. Figure 1. These and analytic number theoretical considerations show clearly that the complexity of the coefficients is a function of the number of distinct odd prime factors of n, much rather than the size of n. Complex patterns arise (see Figure 1) and a lot of mysteries remain.

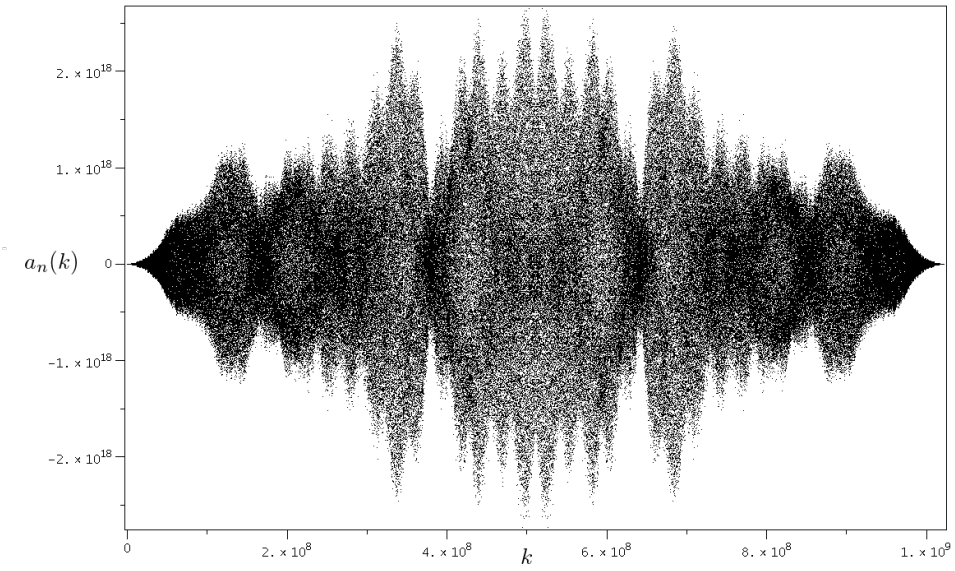


Figure 1: Coefficients of the n-th cyclotomic polynomial for $n = 3234846615 = 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 29$, cf. [2].

2 Which maximum coefficients of cyclotomic polynomials do occur?

The very innocent looking question we consider here is the following.

Question 2.1. *Which integers occur as a maximum coefficient of some cyclotomic polynomial?*

For example, Φ_{210} has 2 as a maximum coefficient. We propose the following conjecture.

Conjecture 2.2. *Every natural number occurs as the maximum coefficient of some cyclotomic polynomial.*

The rest of the paper discusses the progress we made on establishing this conjecture. Surprisingly, a big role in this is played by deep work done by many number theorists on the distribution of gaps between primes. Last but not least, everything hinges on a construction found by Eugenia Roşu [38] improving on an earlier construction due to Yves Gallot and myself [21].

3 Prime gaps

3.1 Elementary material, generalities

For millenia now (some!) humans have been fascinated by prime numbers and their distribution. Recall that *prime numbers* are numbers > 1 only divisible by themselves and 1 (it turns out that it is much better to consider 1 itself not as a prime number). It is usually attributed to Euclid (circa 300 BCE) that he proved there are infinitely many primes. Several formulas producing infinitely many primes are known, but they turn out to be practically useless. A famous example is a result of Mills, which asserts the existence of a real number $A > 1$ with the property that A^{3^n} rounded down to the nearest integer is prime for each natural number n. This first "defeat" forces us to take a step back and ask less precise questions such as to estimate the *prime counting function* $\pi(x)$, which counts the number of primes p not exceeding x; that is $\pi(x) = \sum_{p \leq x} 1$. In the course of answering this, the *stochastic* nature of the prime numbers will become apparent. The notion of an error term will also be involved. If $|f(x)| \leq Bg(x)$, for some positive constant B and all values of $x \geq 1$, we write this compactly as $f(x) = O(g(x))$. This notation was introduced by Bachmann in 1894 and popularized by Landau and is generally named *Landau's Big O notation*. Edmund Landau (1877–1938) was the first to put prime number theory as a separate field on the mathematical map and wrote a bulky standard work [28] on it. Two non-Germans mathematicians, who studied the original German version, were surprised to learn about a very strong mathematician called Verfasser they had never heard of (Verfasser means author...).

The first mathematicians to investigate the growth of $\pi(x)$ had of course to start with collecting data to get some intuition for what is going on. They did this by painfully setting up tables of consecutive prime numbers. The most famous of these computers was Carl-Friedrich Gauss. In 1791, when he was 14 years old, he noticed that as one gets to larger and larger numbers the primes thin out, but that locally their distribution appears to be quite erratic. He based himself on a prime number table contained in a booklet with tables of logarithms he had received as a prize, and went on to conjecture that the "probability that an arbitrary integer n is actually a prime number should equal $1/\log n$". Thus Gauss conjectured the following approximations:

$$\pi(x) \approx \sum_{2 \leq n \leq x} \frac{1}{\log n} \approx \mathrm{Li}(x),$$

where

$$\mathrm{Li}(x) = \int_2^x \frac{dt}{\log t},$$

denotes the *logarithmic integral*. By partial integration one sees that $\mathrm{Li}(x) \sim x/\log x$, where by $A(x) \sim B(x)$ we mean that $\lim_{x \to \infty} A(x)/B(x) = 1$. Thus Gauss's heuristic

leads to the conjecture that

$$\pi(x) \sim \frac{x}{\log x}.$$

This was proved much later, in 1896, by Hadamard and independently by de la Vallée-Poussin and is now called the *Prime Number Theorem* (PNT). Both of them were divinely rewarded for doing so and became immortal. Well, almost – they lived to be near centenarians…

If the *Riemann Hypothesis* (RH) were true, it would imply that

$$\pi(x) = \mathrm{Li}(x) + O(\sqrt{x}\log x). \tag{3.1}$$

The RH is one of the Millennium Problems and will not be discussed further here. Its intimate connection with the distribution of prime numbers is discussed in an introductory way in [37].

Prime number questions fall into two main categories: *global problems* and *local problems*. The former concerns asymptotic formulae, sums, estimations and the like of $\pi(x)$ and related functions (of which the PNT is an example), while local problems involve questions dealing with the individual primes. Our focus here will be on large differences between primes (a local property) and their applications.

We let p_n denote the n-th prime number and put $d_n := p_{n+1} - p_n$. For example, the first few prime numbers are $p_1 = 2, p_2 = 3, p_3 = 5, p_4 = 7$, which means that the first few prime gaps are $d_1 = 1$, $d_2 = 2$, and $d_3 = 2$. Note that $\sum_{k=1}^{n}(p_{k+1} - p_k) = p_{n+1} - 2$. By an equivalent form of the PNT the n-th prime number p_n asymptotically grows as $n \log n$. (This is plausible as by the PNT the number of primes not exceeding $n \log n$ is asymptotically equal to $n \log n/(\log(n \log n)$, that is to n.) Thus on average the prime gap is $\log n$, which behaves as $\log p_n$. A natural question is then how often d_n is behaving far from average. E.g., looking at the d_n one might suspect that infinitely often $d_n = 2$. This happens when both p_n and p_{n+1} are primes (they then form a *twin prime pair*) and the Twin Prime Conjecture states that there are infinitely many twin prime pairs. Similarly it is suspected that, given any even number $2k$, infinitely often $d_n = 2k$. Proving results in this direction is extremely hard. If one focuses on rather bigger gaps, life is a bit easier. For example, Helmut Maier [31] showed that $p_{n+1} - p_n \leq (\log p_n)/4$ for infinitely many n. There are a lot of interesting things to say further on small gaps and some spectacular recent developments to report on, see, e.g., the recent book by Broughan [10]. However, our focus will be on large prime gaps. One does not need the PNT to see that there are arbitrarily large prime gaps, i.e. arbitrarily large stretches of composite integers. Namely, for every $N > 1$ there exists a string of at least N consecutive composite integers. An example is given by the string $(N+1)!+2, (N+1)!+3, \ldots, (N+1)!+N+1$. Experimentally gaps of size N have been found between numbers much smaller than $(N+1)!+N+1$. Rankin [40] proved in 1938 that there exists a positive constant c such

that, for infinitely many n, we have

$$p_{n+1} - p_n \geq c \log p_n \frac{(\log\log p_n)(\log\log\log\log p_n)}{(\log\log\log p_n)^2}.$$

This improved on work of Westzynthius (1931) who showed that the sequence $(p_{n+1} - p_n)/\log p_n$ is unbounded. In his final paper on this topic Rankin showed that one can take c to be any number smaller than e^γ, where $\gamma = 0,5772156649\ldots$ is *Euler's constant*. This had been shown already in 1935 by Pál Erdős [16]. Indeed, Erdős who had the habit of offering prizes for solving various open problems, offered 10.000 dollar to anyone who could prove that c can be replaced by any arbitrarily large constant. In 2016, twenty years after Erdős passed away, this conjecture was independently established by Ford, Green, Konyagin and Tao [18] and Maynard [33]. The group of four authors and Maynard received each 5,000 dollars from Ron Graham, a close friend of Erdős.

The function $\log\log x$ walks off to infinity in such a gentle way that one does not notice it. For example, the reciprocal prime sum $\sum_{p\leq x} 1/p$ behaves in that way. It comes perhaps as a surprise (or shock!) to the reader that if we sum the reciprocals of all different primes any human eye has ever looked at, the number comes to be out less than ... 4! The fact that making conjectures in analytic prime number theory is a notoriously dangerous endeavour is related to this. The danger lies in the fact that computers can barely spot $\log\log$ terms and are certainly blind to the $\log\log\log$ terms that frequently occur. It is there that the log log log devil is in his element. The presence of such terms can result in the conjecture being false on very thin subsequences. A famous example is the conjecture that $\pi(x) < \mathrm{Li}(x)$. It is false, but true up to gigantic values of x. Littlewood proved that $\pi(x)$ and $\mathrm{Li}(x)$ carry out an eternal dance around each other. This is now a classic result, but falls a bit short of proving RH (on the suggestion of his tutor Littlewood tried to prove RH during his postdoctoral studies!). Further examples of log log log devil teases are discussed in my article [36].

3.2 Large prime gaps

There is a whole range of conjectures on gaps between consecutive primes; from more careful to high-risk. The most famous one is Legendre's and claims that there is a prime in $(m^2, (m+1)^2)$ for every natural number m. This is a conjecture that is on the safer side, but for example Firoozbakht's conjecture that $p_n^{1/n}$ is a strictly decreasing function of n is "trés risqué". It implies that $d_n < (\log p_n)^2 - \log p_n + 1$ for all n sufficiently large (see Sun [43]), contradicting a heuristic model suggesting that, given any $\epsilon > 0$, there are infinitely many n such that $d_n > (2e^{-\gamma} - \epsilon)(\log p_n)^2$; see Banks, Ford and Tao [4]. Cramér in 1936 conjectured that $d_n = O\left((\log p_n)^2\right)$. Piltz in 1884 conjectured more modestly that $d_n = O(p_n^\epsilon)$ for every $\epsilon > 0$. The first to *prove* that $d_n = O(p_n^\theta)$ for some $\theta < 1$ was Hoheisel in 1930. He took $\theta = 1 - \frac{1}{33000} + \epsilon$. Well-known to number theorists

is Huxley's [24] result from 1972 showing that one can take $\theta = 7/12 + \epsilon$. Baker et al. [3] showed that $d_n = O(p_n^{0.525})$, which is not much weaker than what one can prove assuming RH. Under RH it is an easy consequence of (3.1) that $d_n = O(\sqrt{p_n}(\log p_n)^2)$. Cramér [13] improved on this by showing in 1920 that $d_n = O(\sqrt{p_n}\log p_n)$ under RH. More explicitly, Carneiro et al. [11] established under RH that $d_n \leq \frac{22}{25}\sqrt{p_n}\log p_n$ for every $p_n > 3$.

We will be especially interested in the following conjecture, which is in the same league as Legendre's conjecture.

Conjecture 3.1 (Andrica's conjecture)**.** *For $n \geq 1$, $p_{n+1} - p_n < \sqrt{p_n} + \sqrt{p_{n+1}}$, or equivalently $\sqrt{p_{n+1}} - \sqrt{p_n} < 1$, or equivalently $p_{n+1} - p_n < 2\sqrt{p_n} + 1$.*

Andrica's Conjecture is currently out of reach as we have just seen (even under RH). The next best thing one can then hope for is to prove that there are not too many n for which the inequality fails (more on that later).

Many mathematicians take it that an unproven assertion can only be called conjecture if there are overwhelming reasons for its truth. From this perspective it seems fair to say that this does not apply to any of the conjectures in this section. Some log log log devil (or any of its kin) might well be lurking somewhere…

3.3 The size of large prime gaps

Estimating the size of large prime gaps by establishing a small exponent α in

$$\sum_{\substack{p_n \leq x \\ p_{n+1}-p_n \geq \sqrt{p_n}}} (p_{n+1} - p_n) = O(x^\alpha) \tag{3.2}$$

is a sport. The current record is due to Heath-Brown [23], who established $\alpha = 3/5 + \epsilon$, with ϵ any positive number. This result is very relevant for us, as we will see in the sequel. I include the table with "exponent hunters", as it strongly suggests how much effort it often takes in prime number theory to achieve seemingly small improvements.

exponent	author	year
0.9666	D. Wolke	1975
0.8674	R.J. Cook	1979
0.8243	M.N. Huxley	1980
0.8083	A. Ivić	1981
0.8055	R.J. Cook	1981
0.7501	D.R. Heath-Brown	1979
0.6944	A.S. Peck	1998
0.6666	K. Matomäki	2007
0.6001	D.R. Heath-Brown	2019

Table 2: Record exponents α in (3.2) over time

4 More on cyclotomic polynomials

From (1.3) it can be deduced by so-called Möbius inversion that

$$\Phi_n(X) = \prod_{d\,|n} (X^d - 1)^{\mu(n/d)}, \tag{4.1}$$

where the product is over all positive divisors d of n and μ is the *Möbius function* defined by $\mu(n) = (-1)^t$ if n is a square-free positive integer having t prime factors, and $\mu(n) = 0$ if n has a repeated prime factor.

Let p be a prime and n a positive integer. Then from (4.1) the following properties are easily deduced

1. $\Phi_{pn}(X) = \Phi_n(X^p)$ if p divides n;

2. $\Phi_{2n}(X) = (-1)^{\varphi(n)}\Phi_n(-X)$ if n is odd;

3. $\Phi_n(X) = X^{\varphi(n)}\Phi_n(1/X)$, that is, Φ_n is *self-reciprocal* if $n > 1$.

For example, using the first property we infer that $\Phi_{16}(X) = \Phi_2(X^8) = X^8 + 1$.

It is a classical result that if n has at most two distinct odd prime factors, then $A(n) = 1$, cf. Lam and Leung [27]. The first non-trivial case arises where n has precisely three distinct odd prime divisors and thus is of the form $n = p^e q^f r^g$, with $2 < p < q < r$ prime numbers. By repeatedly invoking the first property above we have $A\{p^e q^f r^g\} = A\{pqr\}$, and hence it suffices to consider only the case where $e = f = g = 1$ and so $n = pqr$. This motivates the following definition.

p	3	5	7	11	13	17	19	23	29	31	37	41
$(p+1)/2$	2	3	4	6	7	9	10	12	15	16	18	21
$M(p) \geq$	2	3	4	7	8	10	12	14	18	19	22	26
$\lfloor 2p/3 \rfloor$	2	3	4	7	8	11	12	15	19	20	24	27

Table 3: Some numerical evidence for the corrected Sister Beiter conjecture

Definition 4.1. A cyclotomic polynomial Φ_n is said to be *ternary* if $n = pqr$, with $2 < p < q < r$ primes. In this case we call the integer n *ternary*.

An important subclass of these polynomials where we have even more control are the *optimal* ternary cyclotomic polynomials.

Definition 4.2. A ternary cyclotomic polynomial Φ_{pqr} is said to be *optimal* if its coefficients assume $p+1$ different values, that is $A\{pqr\}$ has cardinality $p+1$.

The usage of the word optimal comes from the fact that $p+1$ is the maximum number of distinct coefficients that can occur.

A special property of ternary cyclotomic polynomials is that consecutive coefficients differ by at most one (proven in [20]). Here an example:

$$\Phi_{11\cdot 13\cdot 17}(X) = ... - X^{672} - 2X^{673} - 2X^{674} - 2X^{675} - 3X^{676} - 4X^{677} - 3X^{678} \ldots$$

It follows that $A\{n\}$ consists of consecutive integers if n is ternary (this is not true in general!). For example, $A\{11 \cdot 13 \cdot 17\} = \{-4, -3, \ldots, 1, 2, 3\}$, as can be read off from Table 5. In the ternary case the behaviour of the coefficients is both non-trivial, but also understood so well, that we can use this to our benefit. This is not the case if n has four or more distinct odd prime factors. For optimal ternary cyclotomic polynomials the situation is even more under control, since if we know that $a_{pqr}(k_1) = b$ and $a_{pqr}(k_2) = a$, with $b - a = p$, then b must be the maximal coefficient and a the minimal one.

4.1 The family Φ_{pqr} with p fixed

In this subsection we briefly discuss other research on ternary coefficients.

The height $A(n)$ is unbounded if n ranges over the ternary integers. However, if we restrict to ternary n having a prescribed smallest prime factor $P(n) = p$, we get a bounded quantity $M(p)$. The definition of $M(p)$ can be stated more explicitly as

$$M(p) = \max\{A(pqr) : 2 < p < q < r\},$$

where p is a fixed odd prime and q, r range over the primes satisfying $r > q > p$. As the definition of $M(p)$ involves infinitely many cyclotomic polynomials, it is not clear whether there exists a finite procedure to determine it. Duda [15] provided such a procedure. It reduces the computation of $M(p)$ to the determination of the maximum value of $A(n)$, with n running through a *finite* set of ternary integers pqr. As the n involved are huge, the procedure is unfortunately not practical. It is a major open problem to find a *practical* procedure leading to explicit values of $M(p)$.

In 1971, Möller [35] gave a construction showing that $M(p) \geq (p+1)/2$ for $p > 5$. On the other hand, in 1968, Sister Marion Beiter [5] had conjectured that $M(p) \leq (p+1)/2$ and shown that $M(3) = 2$ [7], which on combining leads to the conjecture that $M(p) = (p+1)/2$ for $p > 2$. The bound of Möller together with Beiter's [6] bound $M(5) \leq 3$ shows that $M(5) = 3$. Zhao and Zhang [47] showed that $M(7) = 4$. Thus Beiter's conjecture holds true for $p \leq 7$. Gallot and Moree [21] showed that Beiter's conjecture is false for every $p \geq 11$. Moreover, they showed that for every $\epsilon > 0$ we have $M(p) \geq (2/3 - \epsilon)p$ and conjectured that always $M(p) \leq 2p/3$, dubbing this conjecture the "corrected Sister Beiter conjecture".

The true behavior of $M(p)$ is much more complicated than suggested by Beiter's conjecture. For one, it is related to the distribution of inverses modulo primes p. Given any integer a coprime to p, any integer b with $ab \equiv 1 (\mathrm{mod}\ p)$ is its *modular inverse*. The collection of points (a, b) with $0 < a, b < p$ is called the *modular hyperbola*; for a survey see Shparlinski [42]. The distribution of points on the modular hyperbola is traditionally investigated using the *Kloosterman sum* $K(a, b; p)$, which is defined as

$$K(a, b; p) = \sum_{1 \leq x \leq p-1} e^{2\pi i (ax + b\overline{x})/p},$$

with $\overline{x}$ any modular inverse of x modulo p. (As an aside we note that the Dutch word kloosterman means "cloister man" and thus the cloister man sums can be used to investigate a conjecture of a nun. Honi soit qui mal y pense! Reader beware: too intense study of these sums and their applications can lead to "Kloostermania" [34].) By a fundamental result of Weil we have $|K(a, b; p)| \leq 2\sqrt{p}$, which can be used to show that $M(p) > 2p/3 - 3\, p^{3/4} \log p$ (see Cobeli et al. [12]).

In Figure 2 we display part of the modular hyperbola mod 241 that is relevant in constructing a sharp lower bound for $M(241)$ in the work of Gallot and myself. It gives integer pairs (a, b) with $1 \leq a, b \leq 240$ in certain triangles with $ab \equiv 1 (\mathrm{mod}\ 241)$. For a detailed analysis of this construction, see Cobeli et al. [12].

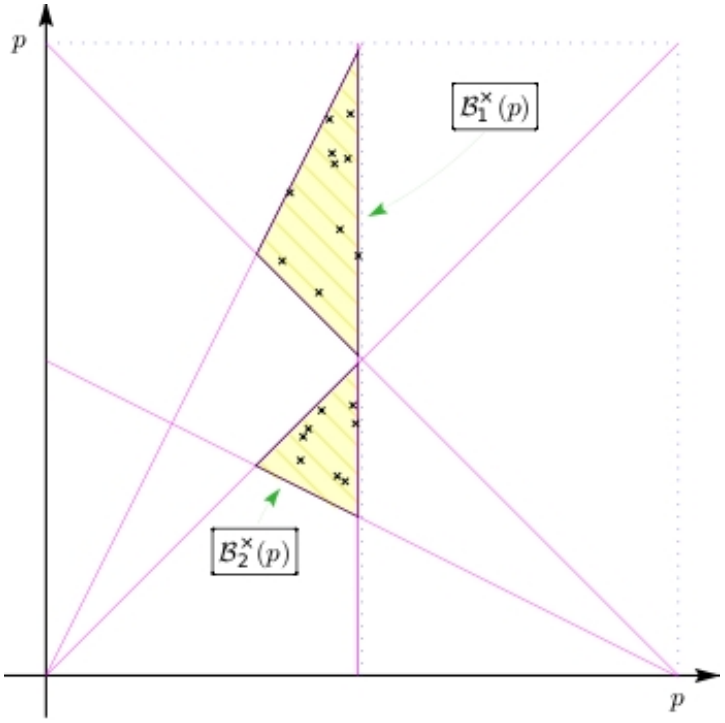


Figure 2: $M(241)$ estimation relevant part of modular hyperbola mod 241.

5 Our results on the possible maximum coefficient

In this section I finally return to Question 2.1 and discuss the recent progress made on it in my paper with Kosyak, Sofos and Zhang [26]. It relies on a construction found by Eugenia Roşu improving on an earlier construction by Gallot and myself. The original formulation is quite lengthy, however for us the following watered down version will do.

Theorem 5.1 (Moree and Roşu [38]). *Let $m \geq 0$ be an arbitrary integer and $p \geq 4m^2+2m+3$ be any prime. Then there exist primes q_1, r_1, q_2, r_2 such that $\Phi_{pq_1r_1}$ and $\Phi_{pq_1r_1}$ have maximum coefficient $(p-1)/2-m$, respectively $(p+1)/2+m$.*

This shows that the set of cyclotomic maximum coefficients we can obtain certainly contains

$$\begin{aligned}\mathcal{R} : \; = \; & \left\{\frac{p-1}{2} - m : p \text{ is a prime}, \, m \geq 0, \, 4m^2+2m+3 \leq p\right\} \\ & \cup \left\{\frac{p-1}{2} + m : p \text{ is a prime}, \, m \geq 0, \, 4m^2+2m+3 \leq p\right\}.\end{aligned}$$

We conjecture that this set equals the set of all natural numbers, thus implying that each natural number can arise as maximum coefficient of some cyclotomic polynomial.

h	p	q
3	5	11
5	13	53
55	139	7507
117	263	30509
219	449	97883

Table 4: Smallest choice of $p \geq 2h-1$ with $q := 1+(h-1)p$ prime

Roughly speaking $\mathcal{R}$ is a union of integers in intervals of the form $((p-1)/2-\sqrt{p}/2, (p-1)/2+\sqrt{p}/2)$, and thus if the gaps between successive primes are *always* sufficiently small, all natural integers will be covered. Working out the technicalities one arrives at the following result.

Theorem 5.2. *If $p_{n+1}-p_n < \sqrt{p_n}+\sqrt{p_{n+1}}$ holds for $p_n \leq 2h$, then the integers $1, 2, \ldots, h$ are in $\mathcal{R}$. Andrica's conjecture, Conjecture 3.1, implies that every natural number occurs as the maximum coefficient of some ternary cyclotomic polynomial.*

A lot of numerical work on large gaps has been done (see the website [39]). This can be used to infer that the inequality in Theorem 5.2 holds for $p_n \leq 2 \cdot 2^{63} \approx 1.8 \cdot 10^{19}$, leading to the following corollary.

Corollary 5.1. *Every integer up to $9 \cdot 10^{18}$ occurs as the maximal coefficient of some ternary cyclotomic polynomial.*

If holes in the set $\mathcal{R}$ appear, it is when $p_{n+1}-p_n \geq \sqrt{p_n}+\sqrt{p_{n+1}}$. The number of natural numbers up to x that are not in $\mathcal{R}$ (if any), is close to

$$\sum_{\substack{p_n \leq 2x \\ d_n \geq \sqrt{p_n}+\sqrt{p_{n+1}}}} (d_n - \sqrt{p_n} - \sqrt{p_{n+1}}) \leq \sum_{\substack{p_n \leq 2x \\ d_n \geq \sqrt{p_n}+\sqrt{p_{n+1}}}} d_n \leq \sum_{\substack{p_n \leq 2x \\ d_n \geq \sqrt{p_n}}} d_n.$$

Now the reader might be reminded of (3.2). An easy climb on the shoulders of giants in analytic number theory then leads to the following result.

Theorem 5.3. *For any fixed $\epsilon > 0$, there exists a constant C_ϵ such that the number of positive integers $\leq x$ that do not occur as a height of a ternary cyclotomic polynomial is at most $C_\epsilon x^{3/5+\epsilon}$. Under the Riemann Hypothesis this number is at most $C_\epsilon x^{1/2+\epsilon}$.*

5.1 A different approach

Let $h > 1$ be odd. If there exists a prime $p \geq 2h - 1$ such that $q := 1 + (h - 1)p$ is a prime too, then for some prime $r > q$ it can be shown that Φ_{pqr} has maximum coefficient h. This is a consequence of work of Gallot, Moree and Wilms [22] and involves ternary cyclotomic polynomials that are not optimal.

For some choices of h, p and q see Table 4.

Conjecture 5.4. *Let $h > 1$ be any odd integer. There exists a prime $p \geq 2h - 1$, such that $1 + (h - 1)p$ is a prime too.*

This conjecture is a consequence of the widely believed Bateman–Horn conjecture [1], which implies that, given an arbitrary odd integer $h > 1$, there are *infinitely* many primes p such that $1 + (h - 1)p$ is a prime too.

Theorem 5.5. *If Conjecture 5.4 holds true, then every positive odd natural number occurs as maximal coefficient of some ternary cyclotomic polynomial. Unconditionally a positive fraction of all odd natural numbers occur as maxima.*

Our proof of the second assertion makes use of deep work of Bombieri, Friedlander and Iwaniec [8] on the level of distribution of primes in arithmetic progressions with fixed residue and varying moduli. Although the unconditional statement in Theorem 5.5 is surpassed by the unconditional statement in Theorem 5.3, the proof of Theorem 5.5 is, in a way, 'orthogonal' to the one of Theorem 5.3; it thus has the potential of working for variations of the problem where the method behind Theorem 5.3 would fail. Interestingly, like our prime gap criterion, it rests on a variation of a certain very well studied problem involving prime numbers. Both prime number questions are, however, quite different.

6 Concluding remarks

In [26] we also obtain the same type of results as described in the previous section for the minimum coefficient and for the height. In case of the height a conjecture slightly stronger than Andrica's enters the game.

Conjecture 6.1. *Every natural number occurs as the height of some cyclotomic polynomial.*

height	p	q	r	k	sign	diff.
1	3	7	11	0	+	2
2	3	5	7	7	−	**3**
3	5	7	11	119	−	**5**
4	11	13	17	677	−	7
5	11	13	19	1008	−	9
6	13	23	29	2499	−	10
7	17	19	53	6013	+	14
8	17	31	37	5596	−	14
9	17	47	53	14538	−	**17**
10	17	29	41	4801	−	**17**

Table 5: Minimal ternary examples with prescribed height

We demonstrate this in Table 5, which gives the minimum ternary integer $n = pqr$ with $p < q < r$ such that Φ_n has height m for the numbers $m = 1, \ldots, 10$. The integer k has the property that $a_{pqr}(k) = \pm m$, with the sign coming from the sixth column. The seventh column records the difference between the largest and smallest coefficient and is in bold if this is optimal, that is, if the difference equals p (compare Definition 4.2). See [26] for the continuation of the table up to $m = 40$.

Prime differences make their appearance since in our approach we work with ternary cyclotomic polynomials. One would want to work with Φ_n with n having at least four prime factors; however, this leads to a loss of control over the behaviour of the coefficients in general and the maximum, minimum and height in particular. Prime number properties play a true role if one asks for the possible heights $A(n)$ and extrema of $A\{n\}$ with n restricted to ternary integers.

7 Further reading

Ribenboim's book [41] gives a wealth of results on prime numbers and their distribution. It can be thought of as a number-theoretical version of the Guinness Book of Records. Also some of the underlying mathematics is explained. For a computational history of prime numbers and Riemann zeros see [37]. The truly courageous might have a go at the monumental book of Landau [28].

Acknowledgments

I would like to thank Alexandru Ciolan, Kate Kattegat, Alexandre Kosyak and Lola Thompson for proofreading earlier versions, and Igor Shparlinski for a mathematical comment. Figure 2 was provided by Cristian Cobeli. The part of the Table 2 up to 1981 is taken from Ivić [25, p. 350], who also gives a proof of the 1979 result of Heath-Brown. Table 3 was computed by Yves Gallot, and Table 5 by Bin Zhang.

This article is a slightly modified version of an article that has been published in the Dutch journal Nieuw Archief voor Wiskunde. We thank the editors for their permission to republish it.

Bibliography

[1] S.L. Aletheia-Zomlefer, L. Fukshansky and S.R. Garcia, The Bateman-Horn conjecture: heuristics, history and applications, *Expos. Math.* **38** (2020), 430–479.

[2] A. Arnold and M. Monagan, http://www.cecm.sfu.ca/~ada26/cyclotomic/.

[3] R.C. Baker, G. Harman and J. Pintz, The difference between consecutive primes. II, *Proc. London Math. Soc.* (3) **83** (2001), 532–562.

[4] W. Banks, K. Ford and T. Tao, Large prime gaps and probabilistic models, arXiv:1908.08613.

[5] M. Beiter, Magnitude of the coefficients of the cyclotomic polynomial $F_{pqr}(x)$, *Amer. Math. Monthly* **75** (1968), 370–372.

[6] M. Beiter, Magnitude of the coefficients of the cyclotomic polynomial F_{pqr}. II, *Duke Math. J.* **38** (1971), 591–594.

[7] M. Beiter, Coefficients of the cyclotomic polynomial $F_{3qr}(x)$, *Fibonacci Quart.* **16** (1978), 302–306.

[8] E. Bombieri, J.B. Friedlander and H. Iwaniec, Primes in arithmetic progressions to large moduli. II, *J. Math. Ann.* **277** (1987), 361–393.

[9] K. Broughan, *Equivalents of the Riemann Hypothesis*. Vol. 2. Analytic equivalents, Encyclopedia of Mathematics and its Applications **165**, Cambridge University Press, Cambridge, 2017.

[10] K. Broughan, *Bounded gaps between primes: the epic breakthroughs of the early* 21*st century*, Cambridge University Press, Cambridge, 2021.

[11] E. Carneiro, M.B. Milinovich and K. Soundararajan, Fourier optimization and prime gaps, *Comment. Math. Helv.* **94** (2019), 533–568.

[12] C. Cobeli, Y. Gallot, P. Moree and A. Zaharescu, Sister Beiter and Kloosterman: a

tale of cyclotomic coefficients and modular inverses, *Indag. Math.* **24** (2013), 915–929.

[13] H. Cramér, Some theorems concerning prime numbers, *Arkiv f. Math. Astr. Fys.* **15** (1920), 1–33. [Collected Works **1**, 85–91, Springer, Berlin-Heidelberg, 1994.]

[14] H. Cremer, *Carmina mathematica und andere poetische Jugendsünden*, 7. Aufl., Aachen: Verlag J. A. Mayer, 1982.

[15] D. Duda, The maximal coefficient of ternary cyclotomic polynomials with one free prime, *Int. J. Number Theory* **10** (2014), 1067–1080.

[16] P. Erdős, On the difference of consecutive primes, *Quart. Journ. of Math.* **6** (1935), 124–128.

[17] J. Fintzen, Cyclotomic polynomial coefficients $a(n, k)$ with n and k in prescribed residue classes, *J. Number Theory* **131** (2011), 1852–1863.

[18] K. Ford, B. Green, S. Konyagin and T. Tao, Large gaps between consecutive prime numbers, *Ann. of Math.* (2) **183** (2016), 935–974.

[19] K. Ford, B. Green, S. Konyagin, J. Maynard and T. Tao, Long gaps between primes, *J. Amer. Math. Soc.* **31** (2018), 65–105.

[20] Y. Gallot and P. Moree, Neighboring ternary cyclotomic coefficients differ by at most one, *J. Ramanujan Math. Soc.* **24** (2009), 235–248.

[21] Y. Gallot and P. Moree, Ternary cyclotomic polynomials having a large coefficient, *J. Reine Angew. Math.* **632** (2009), 105–125.

[22] Y. Gallot, P. Moree and R. Wilms, The family of ternary cyclotomic polynomials with one free prime, *Involve* **4** (2011), 317–341.

[23] D.R. Heath-Brown, The differences between consecutive primes. V, *Int. Math. Res. Not. IMRN*, to appear, https://academic.oup.com/imrn/article-abstract/doi/10.1093/imrn/rnz295/5676434.

[24] M.N. Huxley, On the difference between consecutive primes, *Invent. Math.* **15** (1972), 164–170.

[25] A. Ivić, *The Riemann zeta-function*, John Wiley & Sons, Inc., New York, 1985.

[26] A. Kosyak, P. Moree, E. Sofos and B. Zhang, Cyclotomic polynomials with prescribed height and prime number theory, *Mathematika* **67** (2021), 214–234.

[27] T.Y. Lam and K.H. Leung, On the cyclotomic polynomial $\Phi_{pq}(X)$, *Amer. Math. Monthly* **103** (1996), 562–564.

[28] E. Landau, *Handbuch der Lehre von der Verteilung der Primzahlen*, 2 Bände, second ed. With an appendix by P.T. Bateman, Chelsea Publishing Co., New York, 1953.

[29] E. Lehmer, On the magnitude of the coefficients of the cyclotomic polynomials, *Bull. Amer. Math. Soc.* **42** (1936), 389–392.

[30] H. Maier, Primes in short intervals, *Michigan Math. J.* **32** (1985), 221–225.

[31] H. Maier, Small differences between prime numbers, *Michigan Math. J.* **35** (1988), 323–344.

[32] J. Maynard, Small gaps between primes, *Ann. of Math.* **181** (2015), 1–31.

[33] J. Maynard, Large gaps between primes, *Ann. of Math.* **183** (2016), 915–933.

[34] P. Michel, Some recent applications of Kloostermania, *Physics and number theory*, IRMA Lect. Math. Theor. Phys. **10**, Eur. Math. Soc., Zürich (2006), 225–251.

[35] H. Möller, Über die Koeffizienten des $n$ten Kreisteilungspolynoms, *Math. Z.* **119** (1971), 33–40.

[36] P. Moree, Irregular behaviour of class numbers and Euler-Kronecker constants of cyclotomic fields: the log log log devil at play, in *'Irregularities in the Distribution of Prime Numbers - Research Inspired by Maier's Matrix Method'*, Eds. J. Pintz and M.Th. Rassias, Springer, 2018, 143–163.

[37] P. Moree, I. Petrykiewicz and A. Sedunova, A computational history of prime numbers and Riemann zeros, https://arxiv.org/abs/1810.05244.

[38] P. Moree and E. Roşu, Non-Beiter ternary cyclotomic polynomials with an optimally large set of coefficients, *Int. J. Number Theory* **8** (2012), 1883–1902.

[39] T.R. Nicely, First occurrence prime gaps, web page http://www.trnicely.net/gaps/gaplist.html.

[40] R.A. Rankin, The difference between consecutive prime numbers, *J. London Math. Soc.* **11** (1936), 242–245.

[41] P. Ribenboim, *The book of prime number records*, second edition, Springer-Verlag, New York, 1989.

[42] I. Shparlinski, Modular hyperbolas, *Jpn. J. Math.* **7** (2012), 235–294.

[43] Z.-W. Sun, On a sequence involving sums of primes, *Bull. Aust. Math. Soc.* **88** (2013), 197–205.

[44] J. Suzuki, On coefficients of cyclotomic polynomials, *Proc. Japan Acad. Ser. A Math. Sci.* **63** (1987), 279–280.

[45] R. Thangadurai, On the coefficients of cyclotomic polynomials, *Cyclotomic fields and related topics* (Pune, 1999), 311–322.

[46] S. Weintraub, Several proofs of the irreducibility of the cyclotomic polynomials, *Amer. Math. Monthly* **120** (2013), 537–545.

[47] J. Zhao and X. Zhang, Coefficients of ternary cyclotomic polynomials, *J. Number Theory* **130** (2010), 2223–2237.

6 Inquiry-based learning in an Indian context

Shantha Bhushan, Divakaran D, and Tulsi Srinivasan

Azim Premji University
Survey No 66, Burugunte Village
Bikkanahalli Main Road
Sarjapura, Bengaluru 562125

Abstract: In this article, we reflect on our experiences in using Inquiry-Based Learning (IBL) to teach undergraduate mathematics in an Indian classroom. This pedagogy was implemented in first year courses, to ease the transition from school to college mathematics. We discuss how we implemented IBL in our courses and some of our observations, especially issues that we believe are more specific to Indian classrooms.

1 Introduction

Inquiry-based learning (IBL) is a form of active learning, based on the Moore method. The topologist RL Moore famously conducted courses without lectures or textbooks, giving students a set of axioms and definitions, and then asking them to provide proofs for various results without consulting any other resource, and present them to the class for verification. Many credit this method for Moore's success in identifying and guiding fifty doctoral students and creating a flourishing school of topology, and his students have described the strong positive impact the method had on their learning, their interest in the subject and their self-confidence [9,10]. However, others warn of the difficulty implementing the method in undergraduate classes where students cannot be "hand-picked", causing many to fall behind, as well as the tendency for a hostile or overly competitive atmosphere to develop [2].

Over the years, mathematics teachers have used variations of this method to include

students in the creation of mathematics, tweaking the method to suit their classrooms and requirements. IBL grew (in North America) out of the legacy of RL Moore, but many involved in its development brought in aspects of collaboration (forbidden by Moore) and classroom inclusivity [1, 4].

Broadly, IBL courses are designed so that students mostly learn through activities, exercises, presentations and discussions, rather than lectures. The instructor's role is to create material that challenges students at the right level, and to guide them through the process of discovery. The instructor also gets an insight into how students learn mathematics [6].

The undergraduate programme in mathematics at Azim Premji University uses IBL for introductory first year courses. There were two broad reasons for this – to ease the transition into college mathematics, and to break down various barriers between students. Many students who join our programme have primarily learnt mathematics through rote learning and memorisation of algorithms, and also join with different ideas about what mathematics means. We therefore wanted the introductory courses to emphasise the practice of doing mathematics – playing with concepts and ideas, asking questions, looking for examples, connecting ideas from different areas, justifying solutions, and then communicating ideas to peers for feedback. These courses are also meant to prepare students for second-year courses in linear algebra, abstract algebra, real and multivariable analysis, and probability.

Additionally, we have a mix of students from different social, economic and academic backgrounds, and wanted a classroom in which all students get to speak up without fear, and experience the pleasure and power of collaborative work in maths. Existing literature on IBL (mostly from North America) has indicated the positive impact of the pedagogy on women students, and, to a lesser extent, on minority students, [7, 8] and we hoped to see similar effects in our classrooms. We know of only a few other uses of IBL by colleagues at other universities in India (like RIE Mysuru and Ashoka University), and are not aware of any literature on the impacts.

This article is a collection of reflections from our experiences using IBL to teach first year maths undergraduates at Azim Premji University over the last two years. The number of students is too small for any kind of formal study. Further, the second year saw a switch to online classes due to the pandemic, which lead to minor modifications. Instead, we informally discuss the broad structure of the courses, some of the challenges in implementation, and the impact it has had in our classrooms.

2 Implementation

We primarily use IBL in two first-year courses – Introduction to Mathematical Thinking 1 and Calculus 1. The first course introduces some topics from combinatorics and number theory. The second mainly deals with the structure of the real line, motivated by its role in results from single-variable calculus that most students encounter in school. In both courses, some notions of sets, functions, cardinality, and propositional logic are introduced informally (and covered more rigorously in a later course).

Guided worksheets form the backbone of these two courses. For Calculus 1, we drew heavily upon existing IBL courses by CA Coppin [3] and WT Ingram [5], which are closer to Moore's approach of starting with definitions and axioms. For the other course, we created worksheets from scratch, and the problems started by asking students to play around with questions about counting, divisibility etc. and tried to lead them to some foundational notions in combinatorics and number theory, and to see the importance of the language of sets and functions and the axiomatic method.

In both courses, students are not allowed to look up any books or resources. During class, students present their solutions to their classmates. While a student is presenting, others can ask doubts or express disagreement in a polite and constructive manner. If a flaw is observed in the argument, the student presenting can decide if the mistake is small and can be fixed in class, or to work on it and present in next class, or to pass the question to another student. Other students are asked not to offer suggestions or alternative solutions during a presentation. In one iteration, students were asked to keep a weekly journal to log their progress and indicate which problems they were prepared to present each class. When the courses were taught online, the class was split into smaller sections of 12-15 students for greater individual attention.

Another significant difference between the two courses is in the way students solve the worksheets. In Introduction to Mathematical Thinking 1, students work as groups, while in Calculus 1 they are expected to work independently. These two courses therefore complement each another in content (discrete and continuous structures) and mathematical approach (building towards formality and starting from axioms), as well as approach to collaboration.

Assessment was based on presentations, regular assignments, and tests. When assessing presentations, instructors tried to value the development of ideas and sincere attempts in wrong directions as well as correct proofs. We also found that weekly assignments with detailed feedback on mathematics and writing were useful for students and gave the instructor a better sense of how concepts being discussed in class were absorbed by individuals. In some iterations, assignments consisted of writing up solutions for only problems discussed in class, but we later found it necessary to include additional problems of varying difficulty as well.

2.1 The instructor's role

After the problem sets are created, the instructor has several closely intertwined roles: providing "big-picture" ideas, tweaking the material each week to suit the class, providing academic guidance for each student, and monitoring the well-being of each student and the general mood of the class.

Sometimes the whole class gets stuck, and the instructor needs to take a call on whether to push the class to make progress, to put that particular topic or question aside for a while, or to cover some material as a short lecture so the class can continue as planned. In this, our aim of preparing students for second year courses in algebra, analysis, linear algebra, probability etc. sometimes added constraints. A question that we grappled with through these courses was about how open-ended inquiry can be when covering core content. Our courses were leading students through a pre-planned narrative, and the space available for diversions was limited. There were instances where students wanted to take the course in a particular direction, and the instructor had to take a call on how much that would disrupt the remainder of the course.

It is a foregone conclusion that we need to keep content to a minimum if we want to focus on the process. It is also usually the case that the class has a natural pace, slowing down when encountering a hard concept, and then suddenly speeding up when that concept makes sense. However, there are some strategies through which the instructor can keep the pace relatively steady – homework involving many examples when a difficult concept is first encountered, one-on-one or small group tutorials when the class is stuck, putting aside a very hard question for a week or two to build up momentum, and providing perspectives on the concepts being discussed to keep the class motivated.

Instructors also assist students in carving out a path for their progress each week. While all students are encouraged to work on all problems, the instructor may sometimes want to ask a student to present in a particular class if they have not presented for a while, or to nudge a student who has made many presentations in the direction of a hard problem. In some sense IBL allows for multiple classes to be run together, with some students mostly focusing on understanding concepts through examples, some also proving short results that follow from definitions, and others finding creative proofs to more difficult statements. A student may switch tracks for different courses or topics, and both they and the instructor need to ensure the student is channeling their efforts in the right direction each week. As students mature and learn more mathematics, they are able to connect the explorations of the IBL course and their own learning.

Finally, all this is only possible when the class is committed to the rules of engagement – everyone works hard at the problems, each solution is closely examined for correctness, no one looks up solutions, no one feels intimidated about asking questions, and no question is dismissed as trivial. Reflection exercises on the content and learning styles, sessions on

diversity and inclusivity, conversations outside class about academic integrity, and so on, help maintain such at atmosphere, but these ultimately only work when the instructor can persuade students of the value of the pedagogy.

The tasks of the students and the instructor inside and outside class for an IBL course involving group work are described below:

	Inside class	Outside class
Student	1. Discuss new problems in small groups 2. Present solutions of the small group to the whole class 3. Ask questions and critique solutions presented by others 4. Ensure participation by peers	1. Work on new problems individually or with the whole small group 2. Review solutions to previously presented problems 3. Work on weekly written assignments (problems presented in class and new problems) 4. Attend office hours or tutorials to clarify concepts 5. Reflect on learning through journals and discussions
Instructor	1. Observe small groups and clarify their questions, check everyone is participating, note ideas, solutions, and common difficulties 2. Monitor presentations and direct discussions, ensure all questions are answered and errors are pointed out (even if not resolved), note down solutions 3. Ensure a non-judgmental atmosphere, resolve conflicts, find solutions to lethargy or loss of motivation/confidence 4. Provide summaries, big-picture ideas and short lectures as needed	1. Design problem sheets and assessments 2. Conduct additional tutorials and office hours, advise individual students about their direction 3. Provide detailed feedback on written work 4. Arrange discussion and feedback sessions on class dynamics, inclusivity, academic integrity, and other issues that come up

3 What does an IBL class look like?

Ideally each worksheet should be designed so that every student struggles with the material, but can make progress in at least one or two problems each week. The questions and concepts that students seem to remember and appreciate best are those for which multiple incorrect solutions are given and the discussion involves questions at various levels. In this section, we describe a few examples to illustrate the process.

Counting the number of onto functions between two finite sets

The question was introduced as a word problem in both iterations of the course. In the first iteration, students were asked in how many ways six houses can be painted using three colours such that each house is painted using exactly one colour and all colours are used. When the groups found this difficult, the number of colours was reduced to two. Once students had the solution, they proceeded to solve the original problem by making partitions. Students found it hard to generalise the solution, but were able to frame it in terms of counting onto functions. In a subsequent worksheet on recursion, we counted the number of k-partitions of a set with n elements. They were able to use this to calculate the Stirling number of the second kind and use that to find the number of onto functions. The class got used to the idea of using easy cases to frame the question for the general case. By the end of the course, they were juggling between special and general cases, and also using concepts from a different worksheet or course.

In a subsequent course, the question was framed as a teacher trying to assign three questions to nine students so that each student is given exactly one problem, and no question is left out. A few groups came up with the following argument: first assign each problem to some student. This can be done in $9 \times 8 \times 7$ ways. The remaining problems can now be distributed in any way, bringing the total to $9 \times 8 \times 7 \times 3^6$. Another kind of reasoning was that of the total of 3^9 possibilities, various cases need to be deleted – those where one question is assigned to all students, those where of two questions, one is assigned to one student, and the other to eight, those where of two question one is assigned to two students and the other to seven, and so on. The mismatch in the final numerical answers sparked further discussion.

As before, the groups were asked to look to modify the question with a smaller number of students and questions, where the number of ways could be easily verified by hand, to see if their reasoning held. This exercise clarified that the first argument was over-counting, and also convinced students of the validity of the second. During the presentation of the second, certain simplifications could also be made in the calculations. When we moved on to the question of how to count the number of onto functions from one set to another, some students also tried to return to the first technique and modify it, although this

did not work out. The exercise served as a reminder of why one asks for rigorous proofs and peer feedback even for small counting problems, and also how one can get useful insights from incorrect arguments. The class then moved on to discussing the Stirling number of the second kind, as above.

Coming up with a definition of convergence for sequences of real numbers.

This was an ad hoc exercise given in one iteration of the calculus course, as students were very keen on coming up with a definition after seeing that of a limit point of a set. The exercise was drafted in a hurry, and therefore not very good: students were given a set of sequences, told which ones diverged and which ones converged, with the limit specified for converging sequences. They needed to find a definition that worked for all these examples.

The initial attempts used ideas similar to monotonicity, but the class quickly agreed that oscillating sequences can also converge. The next obstacle was that the class was not convinced that constant sequences should "tend to" anything at all, after which the instructor tried to motivate the definition through some uses of sequences. Following this, the class produced various definitions inspired by that of a limit point of a set, such as the range having exactly one limit point, and the range being bounded and having exactly one limit point. Despite a valiant effort by the class, no satisfactory definition could be come up with in the short time we had.

While this was an enjoyable exercise for students, it was a reminder for the instructor about the importance of motivating ideas and designing problems carefully around them. For example, by motivating sequences through integrals or derivatives right at the beginning, it might have been clear why we want constant sequences to converge. Also, since coming up with such a definition is hard, it would also have been better to have more scaffolding around the question.

GCD through paper folding and the jug problem

In this exercise, students were asked to go through the physical and mental exercise of folding a rectangular sheet of paper about its shorter edge to form a square, and then repeating the step for the remaining rectangle (if any) until no further folds are possible. Students were asked to begin with rectangles with specific dimensions, and then asked questions about whether the process always terminates, how many steps it takes, the connection between the sides of the rectangle and the dimensions of the last square, and so on.

Nearly all students observed (without proof) that the process terminates if the sides are natural numbers, and that the side of the final square is then the GCD of the initial sides of the rectangle. A few could make rough arguments about the process terminating when the sides are rational, and give some examples where it does not. A few students, especially those who had studied some programming in school, were able to write the steps as an algorithm, though some missed some minor points like what happens if rectangle is a square to begin with. A further subset could also merge together consecutive foldings along the same edge into one step to get the Euclidean algorithm familiar to them from school. A large part of the class found the final exercise hard, and no one was able to prove why the algorithm produced the GCD or why it terminated. Many students talked about being able to "see" the idea but not prove it, which lead the instructors to reflect on the distinction between proving something, seeing roughly why something works without a proof, and believing something to be true through many examples without having any idea about why it is true.

To aid this transition in understanding, we created a homework exercise asking students to prove that $(b, a) = (b - na, a)$ whenever n is a natural number for which $b - na \geq a$, and to connect this claim with the process of paper folding. One observation we could make is that the many students who struggled to prove this could suddenly do so when reminded about the definition of the GCD. Our guess is that their previous attempts involved writing what they knew in the hope that something would come to light, rather than directly applying the definition.

Proving that the process should terminate was even harder. Some were able to argue that the process has to end as the sides of the rectangle are natural numbers and keeps decreasing, and a few more could point out (the perhaps obvious fact) that this is because 1 is the smallest natural number. Eventually only one or two could give a complete and rigorous proof using the well-ordering principle or induction. We had a similar experience with a problem about measuring out different quantities with two jugs of different integer volumes. Seeing the notion of GCD in an unexpected context was a good motivator, but mathematically framing and proving findings was hard for most. In both cases, we saw an example of multiple classrooms within the same class, with students wrestling with questions and sub-questions at various levels, and gaining understanding in different ways.

4 Impact on students

We asked students to reflect on their learning at various points during the course. These questions were about their comfort with different topics in maths, and about their learning and relationship with maths.

For the first batch of fifteen students, anonymous surveys were conducted right before and right after the semester, after students had completed two IBL courses. The surveys indicated a small positive shift in the number of students who believe they can understand difficult concepts with enough time. This was accompanied by a small decrease in those who said they feel embarrassed or demotivated by mistakes. When asked about this in an anonymous survey more than a year after the IBL courses ended, one student said

> I feel like I am a "mathematician" instead of feeling like I am solely trying to learn mathematics from outside.

Another said

> At university, I learned that being good at Math could mean much more than snagging the first rank. The warmth of our classrooms here and the candid conversations have slowly helped rebuild my confidence in the subject. I truly value the openness with which we learn, furthermore, I appreciate the encouragement we receive when we encounter new knowledge.

This person also talked a bit about women students sometimes being left out of discussions in IBL classes, and being assigned "mindless work" like writing up solutions after group-work. They concluded

> IBL necessitates everyone's participation, lecture-based classes do not. In both class formats, I did experience gender-related discrimination, but I felt like we could do something about it, make a positive change in the IBL setup.

There was a significant increase in those who find one-on-one tutorials and activities, demos, experiments useful. More worrying, there was a decrease in those who find reading a textbook useful. In the anonymous survey mentioned above, one student pointed out

> I think IBL allowed me to follow lectures better but made it harder to read textbooks and hence the second year courses became a bit harder because we were not used to reading textbooks.

Another said "I feel like if we didn't have IBL in first year, we would be more into reading mathematical texts." This was an unexpected consequence, that we are trying to tackle through more readings.

These students also reported a higher level of comfort with topics from college than those from school. Many more students said they could explain concepts in words, rather than

just solve problems in them. Interestingly, those saying they are more comfortable with complex numbers went up although these were not discussed in the courses.

For the second batch of thirty students, we conducted learning surveys three times during the term, and these surveys were not anonymous. This was for an online version of Introduction to Mathematical Thinking 1, and the main aim was for us to stay abreast of students' difficulties.

The surveys indicate a mostly positive impact of IBL on students' learning and relationship with maths. On the other hand, there were a few students negatively impacted by the course or pedagogy, and we discuss possible reasons for this below. We also observed an increase in self-reported sensitivity to others' learning, as well as slight increases in self-reported resilience, interest in maths and the ability to critique an argument. There was also a slight decrease in confidence in mathematical abilities and grasp over basics. In an anonymous survey conducted a few months after the two IBL courses ended, one student said

> I felt far more confident and certain of my journey in Mathematics after IBL, particularly because I recognised that Maths is 'messy' so to speak. This gave me the confidence to make mistakes, and learn from them, which in turn helped me uncover some very cool things.

A particularly poignant comment was

> I always had a fear with mathematics. So, I think it [hasn't] increased my confidence in maths for now. But it increased my interest towards mathematics.

In informal discussions outside class, a few students also spoke about the positive impact of having their ideas be acknowledged and listened to. One said

> It made me independent while having fun. It made me realise that mathematics is not about computation or calculation. The best part was that the class and the teacher valued our ideas. We were happy when we had good ideas and were happy when our classmates had new ideas.

We also tried to make informal observations about students' progress in future non-IBL courses, but this was partly hampered by the switch to online classes. One broad observation was that IBL is helpful in motivating the language of sets and functions and the need for rules of logic. It is, however, not a replacement for a introductory course on sets, functions, logic and proof techniques.

5 Language and logic

The one observation we made that did not seem to occur the literature we could find on IBL is the impact of language. Our classes have students with various levels of familiarity with English, and the problem sheets we shared did not fully take these disparities into account. The small differences between spoken and mathematical language can throw off even students who speak English as a first language (for example, a common attempt at defining a function being one-to-one is to say each point maps to a unique point).

One of many aims of these courses was to motivate the rules of propositional and predicate logic. It seemed, however, that students who are more comfortable with English have quite an advantage here. As an example, many students observe soon that negation switches around "for all" and "there exists". Those comfortable in English will also immediately identify the many phrases synonymous with "there exists" – not every, there is at least one, in some instance etc. When they mechanically negate "For all x, there exists y such that..." they know that "There exists x, for all y such that..." is grammatically problematic and make the necessary tweaks. Therefore, again, people with similar mathematical understanding face different levels of difficulty because of their comfort with English.

This seemed a greater problem in the calculus course, which begins formally with a set of axioms and definitions. We would therefore like to try changing the direction of that course, so that students begin with concrete questions, and move towards abstraction. It might also be that such a course works better in the second semester, when students have spent more time in college where all classes and many conversations are in English.

When asked about the question of language in the surveys mentioned earlier, students did not talk about the language of the worksheets, but a few did talk about challenges faced during discussions. One person said

> Yes, I felt that the background affected our learning. Some who knew better English, or had a better school education was able to answer the questions quicker, and that got my confidence a bit down. I was truly interested in solving through IBL but the rest of the students' pace seemed to affect some of us. But even in lecture based courses, I felt that the same issue still prevailed. It was not more in IBL than otherwise.

Another said

> In the lecture based classes they used to teach or explain problems on [the] board, so we can understand by seeing the mathematical terms. So, here don't need to know much in language. In IBL we need to present problems, so here language matters.

Discrepancies may arise in any classroom due to comfort with English, but there are ways in which it can be more stark in IBL classes. This was an observation we mostly made when courses were online, so that may also have exacerbated the problem. Nevertheless, we intend to modify the worksheets considerably in the future based on these experiences.

6 Instructor experience and learning

As we wrote a first draft of this, each of us, independent of the other two, described the instructor experience as "frustrating but rewarding". While this might be true of any class, the sentiment was stronger in IBL courses. The main sources of frustration were a lack of control over pace, and having to restrain ourselves from interfering instead of allowing the class make mistakes and discover them slowly.

The most rewarding aspect was witnessing the close bond that students form with one another through the process. This pedagogy demands a lot of intellectual and emotional maturity from students – presenting a solution that that they have worked on for a week, which is very likely to be wrong, sitting still and listening when they think they have a much better argument, finding constructive and supportive ways of highlighting a mistake, taking feedback and working and reworking the same solution. It is a privilege to watch these processes take place. Further, we are asking students to be creative and self-driven, so we need to find ways to critique the ideas they present while ensuring that the person remains motivated enough to go back and work on fixing any errors. We believe this has made us more patient and mindful of how we give feedback.

Additionally, as others like Rasmussen and Kwon [6] have observed, these courses have also given us some insights into how students learn: which are the hard concepts, what are the likely sources of confusion. We have described some instances in sections 4 and 6 above, but there were several others that stood out. For example, an exercise counting the number of paths in a grid highlighted the problem of some standard formulas in permutations and combinations being memorised without questioning their meaning. Another exercise in calculus about defining the dictionary order on the plane with polar coordinates, was a reminder to the instructor and students about how there is a lot to understand about trigonometric functions and their inverses behind standard conventions and formulas.

Implementing IBL in Indian colleges can be hard for many reasons including large classes, and standardised syllabi and assessment, but, given the positive impacts, it seems a worthwhile exercise to look for spaces in which it can be implemented. We would like to continue the practice of IBL and understand it better through documentation, discussion and dissemination of IBL material.

Bibliography

[1] H Bennet, *Inquiry based learning*, Notices of the American Mathematical Society, Vol 66, No 7, 2019.

[2] DW Cohen, *A modified Moore Method for teaching undergraduate mathematics*, The American Mathematical Monthly, Vol 89, No 7, 1982.

[3] CA Coppin, *Linear point set theory*, Journal of Inquiry-Based Learning, No 14, 2009.

[4] Z Haberler, SL Laursen and CN Hayward, *What's in a name? Framing struggles of a mathematics education reform community*, International Journal of Research in Undergraduate Mathematics Education, 2018.

[5] WT Ingram, *Foundations of Calculus: Properties of the Real Numbers, Functions and Continuity*, Journal of Inquiry-Based Learning, No 11, 2009.

[6] C Rasmussen and ON Kwon, *An inquiry-oriented approach to undergraduate mathematics*, The Journal of Mathematical Behavior, 26(3), 189-194, 2007.

[7] SL Laursen, M-L Hassi, M Kogan, A Hunter, and T Weston, *Evaluation of the IBL Mathematics Project: Student and instructor outcomes of inquiry-based learning in college mathematics*, Colorado University, 2011.

[8] SL Laursen, M-L Hassi, M Kogan, and T Weston, *Benefits for women and men of inquiry-based learning in college mathematics: a multi-institution study*, Journal for Research in Mathematics Education, Vol 45, No 4, 2014.

[9] LS Whyburn, *Student oriented teaching – the Moore method*, The American Mathematical Monthly, Vol 77, No 4, 1970.

[10] DE Zitarelli, *The origin and early impact of the Moore nethod*, The American Mathematical Monthly, Vol 111, No. 6, 2004.

7 Euclid's proof of the infinitude of primes

Variations on the theme

Shailesh Shirali

Sahyadri School KFI
Rajgurunagar, Khed
Pune – 410513

Email: shailesh.shirali@gmail.com

Euclid's proof of the statement that there exist infinitely many prime numbers is extremely well known these days, even at the school level. Though it is most often presented as a proof by contradiction, there are other ways of writing it. Here is one such.

Theorem 1. *There exist infinitely many prime numbers.*

Euclid's proof. We re-cast the theorem as follows:

> *Given any finite list of prime numbers, one may extend the list by including more primes in it.*

Given any finite list $S = \{q_1, q_2, \ldots, q_n\}$ of prime numbers (note that they do not have to be the first n primes), Euclid's idea is to consider the number A defined as follows,

$$A = q_1 q_2 \cdots q_n + 1. \tag{1}$$

Observe that A exceeds all the q_i's, and also that A is not divisible by any of the q_i $(i = 1, 2, \ldots, n)$. We now ask: *What type of number is A? Is it prime? Is it composite?* If A is prime, then we replace S by $S \cup \{A\}$, and we have accomplished our task. If not, then A has some prime factor p, and this prime number cannot be any of the q_i. In this case, we replace S by $S \cup \{p\}$, and once again we have accomplished our task. As this step can be repeated as often as we wish, there must exist infinitely many primes. ■

Remark 1. This proof is so well-known that we may not appreciate how beautiful it is! Observe the economy with which it accomplishes what it sets out to do. Faced with the task of proving that there exist infinitely many prime numbers, a more natural strategy would perhaps be: given the first n prime numbers $p_1, p_2, \ldots, p_n$, find the next prime number p_{n+1}. (This is the classic "find the next term" problem.) If Euclid had decided that this was the way to proceed, he would probably not have gotten very far, and the whole history of mathematics might have been very different! As it happens, no one has succeeded in finding a proof along such lines even 23 centuries after Euclid.

In [1], G H Hardy has this classic and beautiful comment on Euclid's theorem (he refers to two theorems in the quote; the other one is the claim that the square root of 2 is irrational):

> *I will state and prove two of the famous theorems of Greek mathematics. They are 'simple' theorems, simple both in idea and in execution, but there is no doubt at all about their being theorems of the highest class. Each is as fresh and significant as when it has discovered—two thousand years have not written a wrinkle on either of them …*

Remark 2. In most presentations of the above proof, S is taken to be the set of the first n primes, i.e., $S = \{p_1, p_2, \ldots, p_n\}$, where p_i is the i-th prime ($p_1 = 2$, $p_2 = 3$, $p_3 = 5$, …). Forming the number $A = p_1 p_2 \cdots p_n + 1$ in the usual manner, let p be the smallest prime factor of A. *It is important here to point out that p is not necessarily the next prime after p_n.* For example:

- For $n = 2$ we get $A = (2 \cdot 3) + 1 = 7$, which is a prime number, so $p = 7$. Note that 7 is not the next prime number after 3.
- For $n = 3$ we get $A = (2 \cdot 3 \cdot 5) + 1 = 31$, which again is prime, so $p = 31$. Note that 31 is not the next prime number after 5.
- For $n = 4$ we get $A = (2 \cdot 3 \cdot 5 \cdot 7) + 1 = 211$, which yet again is prime, so $p = 211$. Note that 211 is not the next prime number after 7.
- For $n = 5$ we get $A = (2 \cdot 3 \cdot 5 \cdot 7 \cdot 11) + 1 = 2311$, which yet again is prime, so $p = 2311$. Note that 2311 is not the next prime number after 11.

Looking at these statements, we may be tempted to suppose that A will always be prime. But this is not the case! Indeed, the first counterexample is found at the very next step. Thus, for $n = 6$ we get $A = (2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13) + 1 = 30031$, and this is not prime: $30031 = 59 \cdot 509$.

The numbers $p_1 p_2 \cdots p_n + 1$ (for $n = 1, 2, 3, \ldots$) are known as the *Euclidean numbers*;

see [2], [3]. Here is the sequence of such numbers:

$$3,\ 7,\ 31,\ 211,\ 2311,\ 30031,\ 510511,\ 9699691,\ 223092871,\ \ldots$$

As of now, it is unknown whether or not this sequence contains infinitely many primes. After 2311, which is the 5-th number in the sequence, the next prime turns out to be the 11-th number, and after that it is the 75-th number! The extreme irregularity of occurrence of primes in this sequence is another illustration of the fact that the primes harbour very deep secrets which are revealed only occasionally (and perhaps reluctantly!) to mathematicians.

Variations on a theme: applications of Euclid's method

The strategy used by Euclid can be put to use to prove more such results, and that is what this article is all about.

Theorem 2. *There exist infinitely many prime numbers of the form* $-1 \pmod 4$.

Proof. Given any finite list $S = \{q_1, q_2, \ldots, q_n\}$ of prime numbers all of the form $-1 \pmod 4$, we show how to extend the list by considering the following number A:

$$A = 4q_1q_2\cdots q_n - 1. \tag{2}$$

Observe: (i) A is not divisible by any of the q_i, and (ii) $A \equiv -1 \pmod 4$. Also observe that the product of primes all of the form $1 \pmod 4$ is again of this form, so A must have at least one prime factor p of the form $-1 \pmod 4$. This prime number cannot be any of the q_i. So we may replace S by $S \cup \{p\}$, and we have accomplished our task. As this step can be repeated as often as we wish, there must exist infinitely many primes of the stated form. ■

Theorem 3. *There exist infinitely many prime numbers of the form* $-1 \pmod 3$.

Proof. Given any finite list $S = \{q_1, q_2, \ldots, q_n\}$ of prime numbers all of the form $-1 \pmod 3$, we consider the following number A:

$$A = 3q_1q_2\cdots q_n - 1. \tag{3}$$

Observe: (i) A is not divisible by any of the q_i, and (ii) $A \equiv -1 \pmod 3$. Also observe that the product of primes all of the form $1 \pmod 3$ is again of this form, so A must have at least one prime factor p of the form $-1 \pmod 3$. This prime number cannot be any of the q_i. So we replace S by $S \cup \{p\}$, and we have accomplished our task. ■

This proof can obviously be mimicked to establish the following result (and others of its kind):

Theorem 4. *There exist infinitely many prime numbers of the form* $-1 \pmod 6$.

Can this style of proof be used to prove that there exist infinitely many primes of the form $1 \pmod 4$? The obvious approach fails; for, if we consider a list $S = \{q_1, q_2, \ldots, q_n\}$ of primes all of the form $1 \pmod 4$, and then construct the following number A,

$$A = 4q_1q_2 \cdots q_n + 1, \tag{4}$$

we are **not** now able to claim that A must have a prime factor of the form $1 \pmod 4$. This is because the product of an *even* number of prime numbers of the form $-1 \pmod 4$ will be of the form $1 \pmod 4$. Therefore, some other strategy is required.

Success is close at hand, however; only, it needs prior knowledge of these facts: (i) -1 is a quadratic residue of all prime numbers of the form $1 \pmod 4$, (ii) -1 is not a quadratic residue of any prime number of the form $-1 \pmod 4$, and therefore, (iii) the prime factors of a number of the form $4x^2 + 1$ are all of the form $1 \pmod 4$.

Theorem 5. *There exist infinitely many prime numbers of the form* $1 \pmod 4$.

Proof. Given any finite list $S = \{q_1, q_2, \ldots, q_n\}$ of prime numbers all of the form $1 \pmod 4$, we consider the following number A:

$$A = 4(q_1q_2 \cdots q_n)^2 + 1. \tag{5}$$

Since A is of the form $4x^2 + 1$, each prime factor p of A must be of the form $1 \pmod 4$. Since, trivially, p cannot be any of the q_i, we replace the set S by $S \cup \{p\}$, and we have accomplished what we need to do. The stated claim follows. ■

How far can we go in this direction? How many such results can we establish using this style of proof? The most general result of this family of results is the famous theorem proved by Dirichlet:

Theorem 6 (Dirichlet)**.** *Given any pair of co-prime integers* a, b*, with* $b > 1$*, there exist infinitely many prime numbers of the form* $a \pmod b$.

Given the extreme generality of this statement, we may expect that it is difficult to prove. And this is indeed the case; see [4], [5]. So there is no question of finding a Euclid-style proof of this theorem!

In the other direction, there is the following very remarkable theorem proved by Murty [7]:

Theorem 7 (Murty)**.** *A "Euclidean proof" exists for the arithmetic progression* $a \pmod b$ *if and only if* $a^2 \equiv 1 \pmod b$.

For more on this theme, we recommend references [6] and [7]. See also [8] for another Euclidean-style result.

Bibliography

[1] G H Hardy, *A Mathematician's Apology.*

[2] Weisstein, Eric W. "Euclid Number." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/EuclidNumber.html

[3] Euclid numbers. From http://oeis.org/A006862

[4] Wikipedia, "Dirichlet's theorem on arithmetic progressions" from https://en.wikipedia.org/wiki/Dirichlet%27s_theorem_on_arithmetic_progressions.

[5] Wikipedia, "Primes in arithmetic progression" from https://en.wikipedia.org/wiki/Primes_in_arithmetic_progression.

[6] Romeo Mestrovic, "Euclid's theorem on the infinitude of primes: a historical survey of its proofs (300 B.C.–2017) and another new proof", https://arxiv.org/abs/1202.3670.

[7] M Ram Murty & Nithum Thain, "Primes in Certain Arithmetic Progressions," Functiones et Approximatio Commentarii Mathematici, Funct. Approx. Comment. Math. 35 (none), 249-259 (January 2006), https://doi.org/10.7169/facm/1229442627.

[8] B Sury, "Euclid-like proofs for primes ending in 1" (published in this issue of *Blackboard*).

8 A Euclid-like Proof for Primes ending in 1

B Sury

Stat-Math Unit
Indian Statistical Institute
8th Mile Mysore Road
Bangalore 560059

Email: surybang@gmail.com

Can the beautiful proof of the infinitude of primes due to Euclid's school be generalized? An article by Shailesh Shirali in this issue discusses Euclid's proof and its generalizations. It also points out the limitations to the cases to which the proof applies. It is natural to wonder if it can be generalized by taking a slightly general point of view. For instance, can we have a generalization that would show there are infinitely many prime numbers such as $11, 31, 41, 61, 71$ whose last digit is 1? Indeed, this can be done. In simple terms, one might say that Euclid's proof considers the values of the polynomial $x + 1$ and the generalization we discuss will look at a more general polynomial. However, it is not completely elementary and needs a 'little' bit of help from Fermat.

Consider the polynomial

$$f(x) = x^4 - x^3 + x^2 - x + 1.$$

The first five primes whose last digit is 1 are $11, 31, 41, 61, 71$. Imitating Euclid's proof, we consider the product of the first $n \geq 5$ primes $p_1, p_2, \cdots, p_n$ with last digit 1. For convenience (later, it will become clearer why), we also multiply by 10 and look at the natural number $a = 10p_1p_2 \cdots p_n$. We will show that the natural number $f(a) = a^4 - a^3 + a^2 - a + 1$ is divisible by a prime whose last digit is 1. Note that $f(a)$ itself has the last digit to be 1. In fact, we will do better and show that ANY prime divisor of $f(a)$ has 1 as its last digit. This is the reason that we multiply by 10 also – we will see why later.

Firstly, it is clear that $f(a) > 1$ and, therefore, does have a prime divisor p say. Now

$$p|(a^4 - a^3 + a^2 - a + 1)|(a^5 + 1)|(a^{10} - 1).$$

We will show that the smallest positive integer d for which p divides $a^d - 1$ must be 10. Let us call this smallest d as the "order of a with respect to the modulus p".

Now, d must divide 10 because if we write $10 = qd + r$ with $0 \leq r < d$, then

$$a^{10} - 1 = a^{qd+r} - a^r + a^r - 1 = a^r(a^{qd} - 1) + (a^r - 1)$$

being a multiple of p implies $a^r - 1$ is a multiple of p - we have used the fact that $a^{qd} - 1$ is a multiple of p. This would contradict the choice of d as the smallest for that property unless $r = 0$. So, $d|10$ and the choices for d are $1, 2, 5, 10$.

If $d = 1$, then p divides $a - 1$ but p also divides $f(a) = a^3(a - 1) + a(a - 1) + 1$, which is impossible.

If $d = 2$, then $p|(a^2 - 1) = (a - 1)(a + 1)$ and hence $p|(a + 1)$. But, as p divides

$$a^4 - a^3 + a^2 - a + 1 = (a^4 + a^3) - (2a^3 + 2a^2) + (3a^2 + 3a) - (4a + 4) + 5$$

$$= (a + 1)(a^3 - 2a^2 + 3a - 4) + 5,$$

p must divide 5. But $p \neq 5$ as $f(a)$ ends in 1.

If $d = 5$, then $p|(a^5 - 1)$ but $p|f(a)|(a^5 + 1)$. Thus, $p = 2$, which is also not possible as $f(a)$ is odd.

Therefore, we have proved that 10 is the order of a for the modulus p.

At this juncture, we require that 'little' help from Fermat - the so-called little theorem of Fermat. It asserts for any prime q and any integer b that is not divisible by q that $b^{q-1} - 1$ is a multiple of q. Incidentally, Fermat's assertion follows immediately from the assertion that q divides $u^p - u$ for ANY positive integer u which, in turn, easily follows by induction on u. Returning to our p and a, note that p is relatively prime to 10 as $f(a) = 1+$ a multiple of 10 – this is the reason we multiplied by 10 earlier. Thus, we have that $a^{p-1} - 1$ is divisible by p. As we know that 10 is the order of a, the argument used earlier above shows that 10 must divide $p - 1$ (indeed, if not, p would divide $a^r - 1$ where r is the reminder when $p - 1$ is divided by 10). Thus, p ends in 1. As p must be different from any of the prime factors of a, it is a new prime ending in 1. The proof is complete.

Now, the moment has arrived when we should reveal the magic – how one thought of the polynomial $f(x)$. Euclid's proof is the case $f(x) = x + 1$. For our situation, we need to look at the polynomial $x^{10} - 1$ and find its largest degree factor which turns out to be our $f(x)$. Thus, the above argument carries over verbatim to show that for every positive integer d, there is a Euclid-like argument to prove there are infinitely many prime numbers which leave reminder 1 when divided by d.

9 Poly-folly

Kanakku Puly

Dikzihw: "Hello Professor, I have a question about polynomials."

Rehcaet: "Yes? What is it?"

Dikzihw: "I thought polynomials are very simple objects. But, I am in a quandary, being unable to decide if a function of two variables which is a polynomial in each variable must itself be a polynomial."

Rehcaet: "That is a nice question, and you know that such questions usually have a negative answer. You may recall the famous example of the function $f(x,y) = xy/(x^2 + y^2)$ for $(x,y) \neq (0,0)$ and $f(0,0) = 0$ that is continuous in each of the real variables x, y but not continuous as a function of two variables. I expect that the polynomial question also has a negative answer."

Dikzihw: "I understand but might these not be very different questions?"

Rehcaet: "That is true. Ok, let me think a little."

(After a very short time) "Ok, I would like to correct myself; the answer for polynomials is yes. Here is an argument."

"For each fixed $b \in \mathbb{R}$, suppose $f(x,b)$ is a polynomial; then, it is a finite $\mathbb{R}$-linear combination of the polyonmials $f_n(x) = x(x-1)\cdots(x-n+1)$. Write $f(x,b) = \sum_n b_n f_n(x)$ where b_n's are real numbers depending on b; they are uniquely determined because $f_n(x)$'s are linearly independent as n varies. Thus, we have functions $b_n : \mathbb{R} \to \mathbb{R}$ such that for all $x, y \in \mathbb{R}$,

$$f(x,y) = \sum_{n=0}^{\infty} f_n(x) b_n(y).$$

Now, for each b, one has $b_n(b) = 0$ for large enough n onwards, as $f(x,b)$ is a polynomial.

But, for each positive integer N, we have then

$$f(N,y) = \sum_{n=0}^{N} f_n(N)b_n(y) = \sum_{n=0}^{N} N(N-1)\cdots(N-n+1)b_n(y).$$

As $f(N,y)$ is a polynomial in y by assumption, it follows by induction on n that $b_n(y)$ is a polynomial. We will see that b_n must be the zero polynomial for all large enough n. If not, there are infinitely many n for which $b_n(y)$ is a non-zero polynomial. But then the zeroes of these polynomials form a countable set. However, we already observed that for each real b, the polynomials $b_n(b) = 0$ for all but finitely many n, thereby implying that we have uncountably many zeroes of these non-zero polynomials b_n. This contradiction shows that the b_n's are zero polynomials from some n onwards; hence $f(x,y) = \sum_n f_n(x)b_n(y)$ is itself a polynomial."

Dikzihw: "That is a nice proof, thank you. But, is this true for polynomials over rational numbers too?"

Rehcaet: (hesitantly) "I expect that would also be true. Why not take it as a homework problem and let me know tomorrow."

(Next day:)
Rehcaet: "Well, were you able to prove it over $\mathbb{Q}$? I expect it is true."

Dikzihw: "Indeed, I was able to answer the question over $\mathbb{Q}$; however, the answer is 'No'!"

"Here is an example."

"Let us enumerate the rational numbers as $t_1, t_2, t_3, \cdots$ Consider the polynomials $p_n(x) = \prod_{i=1}^{n}(x - t_i)$ for $n \geq 1$. Notice that the functions $f(t_r, y) = \sum_{n=1}^{r-1} p_n(t_r)p_n(y)$ and $f(x, t_r) = \sum_{n=1}^{r-1} p_n(x)p_n(t_r)$ are polynomials of degree $r-1$ as $p_k(t_r) = 0$ for all $k \geq r$. I claim that

$$f(x,y) = \sum_{n=1}^{\infty} p_n(x)p_n(y)$$

is not a polynomial. If it were a polynomial of total degree d say, then $f(x, t_{d+2})$ would have degree at the most d but it has degree $d+1$ as we observed above. Therefore, $f(x,y)$ is not a polynomial."

Rehcaet: "Surprise surprise! I am glad that you didn't just meekly accept what I asserted. By the way, going back to the original example of the discontinuous function, here is a nice exercise."

If a, b, c, d are positive integers, consider the function $f(x, y) = \dfrac{x^a y^b}{x^{2c} + y^{2d}}$ for $(x, y) \neq (0, 0)$ and $f(0, 0) = 0$. What are the possible values of a, b, c, d when f is discontinuous?

Dikzihw: "I have a wonderful solution but this space isn't enough to write it down."