Eureka Digital Archive

archim.org.uk/eureka

Eureka Editor
The Archimedeans
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WA
United Kingdom

archim-eureka@srcf.net

Published by The Archimedeans, the mathematics student society of the University of Cambridge

Thanks to the Betty & Gordon Moore Library, Cambridge

Editorial

Eureka 63

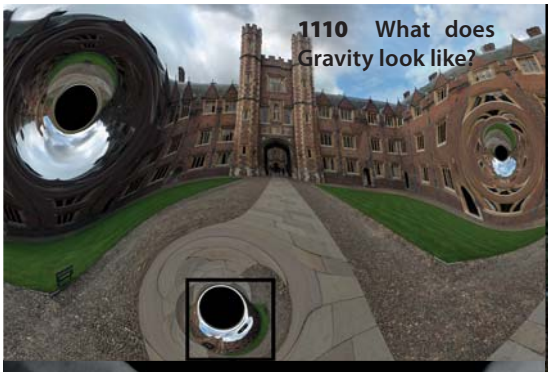
In the long and esteemed tradition of Eureka, I would like to start by apologising for the lateness of publication this year. Various organisational issues outside of our control (acts of God, one might even say) have set us back a few months. Despite this, it has been a real pleasure to edit Eureka for the first time, and I hope that I have been able to do the fantastic journal justice.

This year, our focus on the editorial committee has been to increase the quantity of original articles from well-known professional mathematicians, to add to the usual set of brilliant articles from students here in Cambridge. This has been largely successful, and this issue has contributions from many 'big names'. Throughout the editing process, we've also made efforts to prioritise the accessibility of articles, and to keep the mathematics readable. With that being said, however, the wide variety of topics and approaches should mean that there is something for everyone, and we've marked more technical articles with stars in the contents.

Over the next year, we plan to resume the publication of Qarch, our problems journal. We intend to publish it more frequently than has previously been the case, although the exact frequency and medium remain under discussion. We would be very interested to hear of any potential contributors.

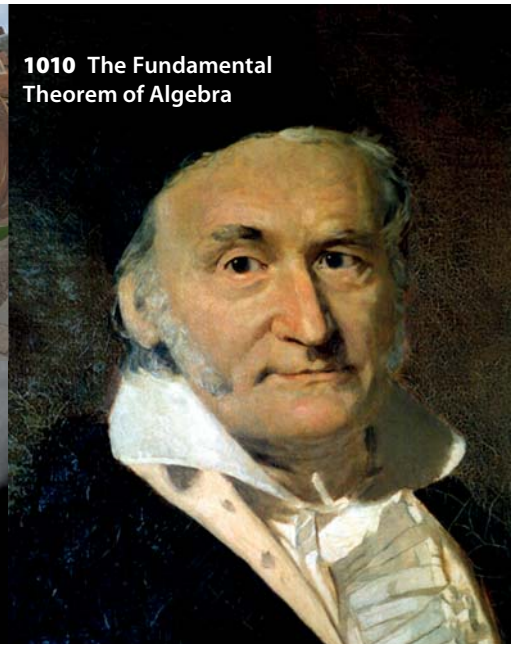
I've had the privilege of working with an excellent editorial team, who I would like to thank for all of their hard work. I would also like to thank former editor Philipp Legner for his invaluable advice and support, as well as our writers, our sponsors, the Archimedeans, and our readers. To run the risk of sounding cliché - without you none of this would have been possible. I hope you enjoy reading Eureka 63!

Jasper Bird
Editor, 2013



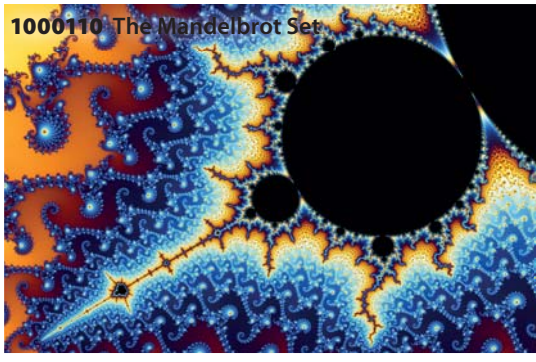
1110 What does Gravity look like?

100000 Chains Between Prisoners



1010 The Fundamental Theorem of Algebra

CONTENTS



The Archimedeans

James Bell, President 2013 – 2014

The Archimedeans have delivered another year of social and mathematical events to our members. We have welcomed almost 200 new members and held a variety of events. Our speakers this year have included Sir Michael Atiyah, Simon Singh and Sir Roger Penrose, amongst many others, providing a talk for almost every Friday of Michaelmas and Lent. Topics ranged from numerical analysis to number theory and from geometry to The Simpsons!

We have had another successful annual dinner. This year, for the first time, the venue was Double Tree by Hilton on the bank of the Cam, and also for the first time we invited a fair number of members of the faculty to dine with us.

The Archimedeans' Problems Drive went ahead as ever with many teams and a set of questions which can be found in this journal. The competition was very tight and in the end went down to a tie breaking game of rock paper scissors!

We once again participated in the Science societies' garden party, and again provided an abundance of cheese to a party of various foods, drinks and jazz.

In Michaelmas Term we had our usual Freshers' squash with plenty of free pizza and a talk from Prof David Tong. We also extended our use of pizza to our board games night, which was plenty of fun for the large number of people who turned up.

I hope that this edition of Eureka is of great interest to you and that our body of members and the Archimedeans can continue to thrive for many years yet.

The Committee 2013 – 2014

President
James Bell (Gonville and Caius)

Vice-President
Dana Ma (Newnham)

Corporate Officer
Maithra Raghu (Trinity)

Secretary
Daochen Wang (Sidney Sussex)

Treasurer
Rowan Purvis (Jesus)

Events Managers
Lukasz Segiet (St Catherine's)

Publicity Officer
Emily Bain (Emmanuel)

Webmaster
Katarzyna Kowal (Churchill)

The Archimedeans' Annual Dinner

The Audience at an Archimedeans Talk

The Archimedeans Committee

Prof Körner addresses the Archimedeans

On Mathematical Method and Mathematical Proof

Prof Reuben Hersh
Emeritus Professor of Mathematics and Statistics, University of New Mexico

Everybody knows that the sum of the first n integers is $n(n + 1)/2$. There's an old anecdote about Carl Friedrich Gauss as a little boy astonishing his schoolmaster by summing from 1 to 100 and getting 5,050.

What about the sum of the first n squares, or the first n cubes? You probably met these in high-school, required to discover the formulas by intelligent guessing, and then to prove them by mathematical induction. The sum of the cubes is easy, because you get 1, 9, 36, 100 as the sum of the first, the first two, the first three, and the first four cubes. You can't help noticing that these numbers are the squares of the sums of the first, the first two, the first three and the first four integers. You easily write down the general case and prove it by induction. The sum of the first n squares is a bit more trouble, but you can do it. In less than half an hour. Work it out, just for fun. It comes out as $n(n + 1)(2n + 1)/6$. Not as beautiful as the sum of n cubes.

Somehow, nobody ever bothers with the sums of fourth, or fifth, or sixth powers. It doesn't seem that awfully interesting to just keep going. It would be worth the trouble for you to figure out how, if you wanted to, you could actually obtain formulas for the sums of higher powers, one at a time. The higher you go, the more trouble it would be; you'd find that you were solving systems of $n + 1$ linear equations. You could write a computer program to print out all the sum formulas up to $n = 500$, $p = 500$. Forget that, we're not getting into that sort of thing in this article.

What you really want is a general formula for the sum of the pth powers of the first n integers, for any positive integers n and p. Knowing it for $p = 1$, 2 and 3, or even for $p = 1, 2, 3, 4, 5$ doesn't seem to get you close to a general formula!

Experimentation

Let's lay our information out in a little table. As we move to the right, we add one more term to the sum. The first row is the sum of the first powers, the second row is the sum of the squares, and so on. Each row starts with 0, because you start with nothing, and then add terms one at a time.

n / p	0	1	2	3	4	...
1	0	1	3	6	10	...
2	0	1	5	14	30	...
3	0	1	9	36	100	...
4	0	1	17	...	...	...
...	...	...	...	...	...	...

Figure 1 Table showing sums of p^{th} powers up to n^p

The table can go on and on, to the right and downwards, but what would we learn from that? At this point there are two options. You can just quit. The hell with it, this isn't getting anywhere!

The other option is better. It seems going to the right any further is a waste of time. What else can you do? Go to the left! Instead of adding 1 and moving right in the first row, subtract 1 and move left! Instead of adding successive integers and moving right in the second row, subtract them and move left! That will mean subtracting negative numbers, once you go to the left of zero. Instead of adding squares of integers in the third row, subtract them! "What for?" you say. "Why bother?" That's not thinking like a mathematician. A mathematician is curious. He/she wants to know more, to understand more. Do it, just to see what happens! I am resisting the temptation to do it for you. It's easy enough, you must have a pencil and paper close by.

Figure 2 Draw your own!

The expanded rows now have zeroes in the middle, and stretch out in both directions. The result turns out to be surprisingly simple! There are two opposite cases - the even powers and the odd powers. In each case, there is symmetry around a midpoint. The midpoint is halfway between 0 and -1. The midpoint is at -1/2. Bit of a shock, that! And around that midpoint, the odd powers have even symmetry (they are equal on the right and the left). The even powers have odd symmetry (they are negatives of each other on the right and the left). All this is apparent, as soon as you extend the rows to the left by subtraction.

Now a little baby algebra comes in handy. By shifting the variable n, it is evident that the sums of odd powers are even functions of $(n + 1/2)$, and the sums of even powers are odd functions of $(n + 1/2)$. Moreover, the sum of the p^{th} powers, p being even or odd, is a polynomial of degree $p + 1$. Therefore, these sums are respectively, sums of just even or odd powers of $(n + 1/2)$.

Does that work for the examples we started with, for $p = 1$, 2 and 3? You can check it in a few minutes, that last statement does hold for the expressions we obtained in the first few lines of this piece. Just completing the square is all it takes.

It is fascinating to learn that back in 1615, in the time of Pascal, before Newton and Leibniz, an obscure German city official liked playing with this kind of algebra. Johannes Faulhaber was his name. He published a long-forgotten little pamphlet where he actually represented the sums of powers of integers as sums of even or odd powers of $(n + 1/2)$. How did he know?

There's more. Once you've gone this far, you can easily conclude that the sum of the p^{th} powers, for all p odd (not just $p = 3$) is a polynomial in that first little expression, $n(n + 1)/2$. And for all p even (not just $p = 2$) it is equal to $2n + 1$ times a polynomial in that first little expression.

Rigorization

But my title promised you some methodology. Let's get to that. We have our results. But where are the proofs? Where are the axioms, that are supposed to be the first line of the proof? We never wrote a single axiom! Does that mean we never proved anything? And if we never proved anything, does that mean we don't know anything, that all this is just a waste of time, not mathematics at all?

Or does it mean that there is something misguided, or wrong-headed, in the notion that mathematics is all about logical deductions from axioms?

When I wrote up this bit of elementary algebra for publication, I included two little algebraic identities, stating the oddness and evenness of the sums of powers, a fact which can be read off directly from the enlarged tables. I asserted that these two identities can easily be proved by induction. That task was left for the skeptical reader to confirm, and for the less skeptical to accept. Other than that, the claims I made here were all stated without proof. The proofs are too simple and

elementary to justify taking up valuable journal space.

Does that mean that the whole article was too simple and elementary to get published? No, not at all! It contains a new idea, a new approach, to a very old topic. The fact that the new idea is as simple as possible makes it more interesting, not less.

The idea, the device, of extending the sum functions to negative *n*, is what's interesting. The proofs of the consequences of this simple idea are obvious, mere elementary exercises. Not worth publishing.

So, OK, the proofs aren't published, but they are easily supplied. But how can there be proofs without axioms? How do you even start a proof, without having the starting point, the axioms or assumptions or hypotheses? In fact, it is not hard to figure out that the starting point here, and in fact in nearly all math proofs, is simply all that we need to know and do know from established mathematics. We are entitled to use all of that, without apology. That is how mathematics is done. The mathematician asks him/herself, "What do I know that would help solve this problem?" Meaning, mostly, what do I know from established math, that every mathematician knows, or could know by looking it up. If what I already know is inadequate, what does my office mate know? What could I find in the right book, article, or on-line source? I solve the problem by thinking about the things involved in the problem – mathematical entities – as I have them stored or represented in my mind/brain, and then also "in the literature". You might say, if you wanted to stretch it, that in all mathematical work, there is already stated, as a hypothesis, all of established mathematics.

The End Result

The end result of a successful mathematical investigation is to add something to, or to make some improvement in, the body of established mathematics. To achieve this, the mathematician plays around with his/her concepts, her mental models, turning them upside down and inside out, trying to get to where he/she wants to go. And he/she calls in, as needed, anything and everything already established in mathematics. Guessing by analogy, piling up examples, drawing pictures and erasing them and making new pictures, till something clicks, something makes sense, some new understanding is achieved.

Then, as a final chore, arranging it logically so that it will convince other interested mathematicians. The proof need only dwell on anything unfamiliar, anything non-routine. What's interesting is the new idea. If there is no new idea, there is probably no reason for anybody to read about the work. The proof is only what's needed to explain and convince the appropriate reader – the mathematician or maths student who has the appropriate background and preparation. So doing math requires, of course, the socialization, the indoctrination if you will, to enable the student to know what is required to make his/her new result convincing to the intended audience, of qualified readers. Just as in any conversation, you don't bore people with a lot of old trivia that they already know all about.

References

[1] Reuben Hersh; 2012; *Why the Faulhaber polynomials are sums of even or odd powers of (n + 1/2)*; College Mathematics Journal; MAA; (43) 4, 322-324; http://eric.ed.gov/?id=EJ985833.
[2] Mohammed T. Dashti; 2011; *Faulhaber's Triangle*; College Mathematics Journal; MAA; (42) 2, 96-97.
[3] Alan F. Beardon; 1996; *Sums of powers of integers*; American Mathematical Monthly; MAA; (103) 3, 201-213.

About the Author

Reuben Hersh received a BA in English literature from Harvard University, then decided to study Mathematics at the Courant Institute of Mathematical Sciences. In 1962 he was awarded a PhD in Mathematics from New York University and he is currently a professor emeritus at the University of New Mexico. He is best known for his writings on the nature, practice and social impact of mathematics.

Allegory of Mathematics, by Bernardo Strozzi

The Fundamental Theorem of Algebra

Prof Ian Stewart FRS
Emeritus Professor of Mathematics, University of Warwick

The Fundamental Theorem of Algebra states that any nonconstant polynomial $p(z)$ over $\mathbb{C}$ has at least one zero; that is, $p(z_0) = 0$ for some $z_0 \in \mathbb{C}$. This easily implies that if the degree of p is n then there are n zeros, provided multiple zeros are counted correctly. Indeed, $z - z_0$ divides $p(z)$, and the quotient has degree $n - 1$, so we can proceed by induction.

This result was widely used by Euler, Lagrange, and others, who offered various handwaving 'proofs'. The first rigorous proof was given by Gauss in his doctoral thesis. It involved the manipulation of complicated trigonometric series to derive a contradiction, and was far from transparent. The underlying idea can be reformulated in topological terms, involving the winding number of a curve about a point, see [1].

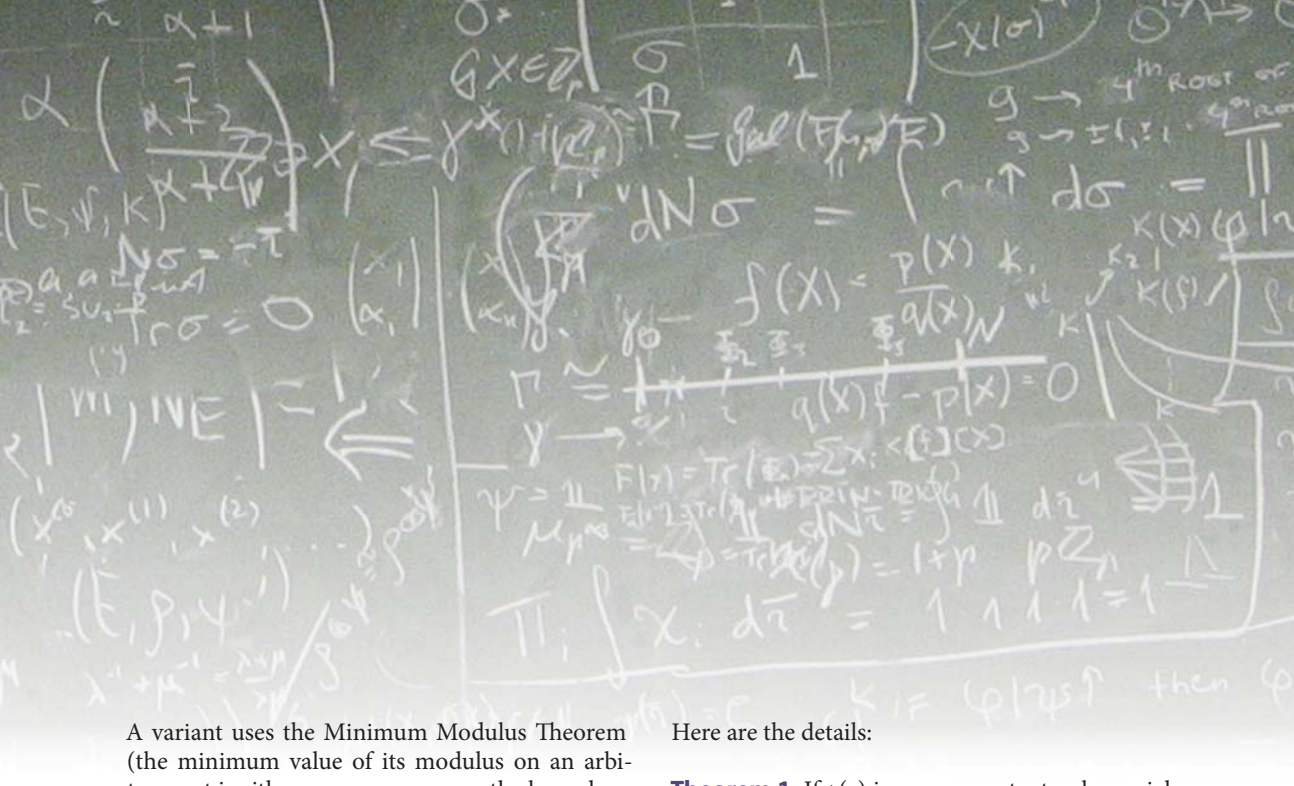
Modern Proofs

Other classical proofs use deep results in complex analysis, such as Liouville's Theorem: a bounded function which is analytic on the whole of the complex plane is constant. This depends on Cauchy's Integral Formula and takes most of a course in complex analysis to prove. See [3], 2.51. An alternative approach uses Rouché's Theorem, see e.g. [3], 3.44. Another proof – the first one I was shown as a student - uses the Maximum Modulus Theorem: if an analytic function is not constant, then the maximum value of its modulus on an arbitrary set occurs on the boundary of that set.

Carl Friedrich Gauss (1777-1855) a German mathematician and scientist. Sometimes referred to as ***Princeps mathematicorum*** (Latin, "the Prince of Mathematicians"), he is often considered one of history's greatest mathematicians.

"EYPHKA! num = Δ + Δ + Δ"

(A famous note written in Gauss' diary after he proved that every positive integer could be expressed as the sum of three triangular numbers)



A variant uses the Minimum Modulus Theorem (the minimum value of its modulus on an arbitrary set is either zero or occurs on the boundary of that set). See [2], Theorems 10.14, 10.15. Euler's approach, which sets the real and imaginary parts of $p(z)$ to zero and proves that the resulting curves in the plane must intersect, can be made rigorous. Clifford gave a proof based on induction on the power of 2 that divides the degree n, which is most easily explained using Galois theory.

An Elementary Proof

All of these proofs are quite sophisticated. But there's an easier way. Some years ago I found a simple proof using a few ideas from elementary point-set topology and estimates of the kind we encounter early on in any course on real analysis. I quickly discovered that it was already known to experts: you can find it on Wikipedia, for example. But it deserves to be more widely known, because it is simple and cuts straight to the heart of the issue. The necessary facts can be proved directly by elementary means, and would have been considered obvious before mathematicians started worrying about rigour in analysis, in around 1850.

The idea can be summarised in a few lines. Assume for a contradiction that $p(z)$ is never zero. Then $|p(z)|^2$ has a nonzero minimum value and attains that minimum at some point $w \in \mathbb{C}$. Consider points v on a small circle centred at w, and show that $|p(v)|^2$ must be less than $|p(w)|^2$ for some v. Contradiction.

Here are the details:

Theorem 1 If $p(z)$ is a non-constant polynomial over $\mathbb{C}$, then there exists $z_0 \in \mathbb{C}$ such that $p(z_0) = 0$.

Proof Suppose for a contradiction that no such z_0 exists. For some $R > 0$ the set $D = \{z : |p(z)|^2 \le R\}$ is non-empty. The map $\psi : \mathbb{C} \to \mathbb{R}^+$ defined by $\psi(z) = |p(z)|^2$ is continuous, so $D = \psi([0, R])$ is compact. For a subset of $\mathbb{C}$ this is equivalent to being closed and bounded. It follows that $|p(z)|^2$ attains its minimum value on D. By the definition of D this is also its minimum value on $\mathbb{C}$.

Assume this minimum is attained at $w \in \mathbb{C}$. Then $|p(z)|^2 \ge |p(w)|^2 \quad \forall z \in \mathbb{C}$ and by assumption $p(w) \ne 0$.

We now consider $|p(z)|^2$ as z runs round a small circle centred at w, and derive a contradiction.

Let $h \in \mathbb{C}$. Expand $p(w + h)$ in powers of h to get

$$p(w + h) = p_0 + p_1 h + p_2 h^2 + \cdots + p_n h^n \qquad (1)$$

where n is the degree of p. Here the p_j are specific complex numbers. They are in fact the Taylor series coefficients $p_j = p^{(j)}(w)/j!$ but we don't actually need to use this, and (1) can be proved algebraically without difficulty.

Clearly $p_0 = p(w)$, and we are assuming this is

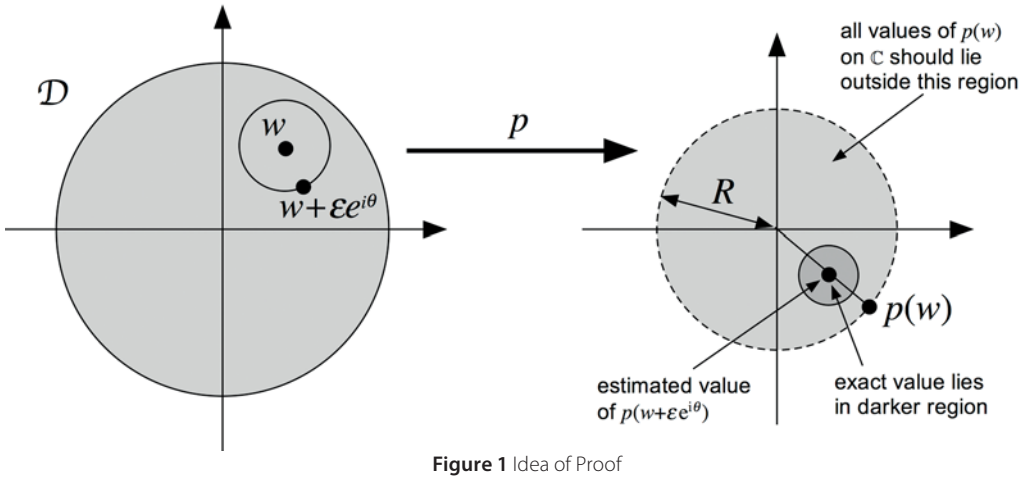


Figure 1 Idea of Proof

nonzero, so $p_0 \neq 0$. If $p_1 = p_2 = \ldots = p_n = 0$ then $p(z) = p_0$ is constant, contrary to hypothesis. So some $p_j \neq 0$. Let m be the smallest integer ≥ 1 for which $p_m \neq 0$. In (1), let $h = \varepsilon e^{i\theta}$ for small $\varepsilon > 0$.

$$p(w + \varepsilon e^{i\theta}) = p_0 + p_m \varepsilon^m e^{mi\theta} + O(\varepsilon^{m+1})$$

Therefore

$$|p(w + \varepsilon e^{i\theta})|^2 = |p_0 + p_m \varepsilon^m e^{mi\theta}|^2 + O(\varepsilon^{m+1})$$

$$= p_0\overline{p}_0 + \overline{p}_0 p_m \varepsilon^m e^{mi\theta} + p_0 \overline{p}_m \varepsilon^m e^{-mi\theta} + O(\varepsilon^{m+1})$$

Let $p_0\overline{p}_m = re^{i\phi}$ for $r \geq 0$. Since $p_0 \neq 0$ and $p_m \neq 0$ we have $r > 0$. Setting $h = 0$ we see that $p_0\overline{p}_0 = |p(w)|^2$. Now

$$|p(w + \varepsilon e^{i\theta})|^2$$
$$= p_0\overline{p}_0 + re^{i\phi}\varepsilon^m e^{mi\theta} + re^{-i\phi}\varepsilon^m e^{-mi\theta} + O(\varepsilon^{m+1})$$
$$= |p(w)|^2 + 2\varepsilon^m r\cos(m\theta + \phi) + O(\varepsilon^{m+1})$$

Set $\theta = \frac{1}{m}(\phi - \pi)$, so that $\phi = \pi - m\theta$. Then $\cos(m\theta + \phi) = \cos(\pi) = -1$, and

$$|p(w + \varepsilon e^{i\theta})|^2 = |p(w)|^2 - 2\varepsilon^m r + O(\varepsilon^{m+1})$$

But $\varepsilon, r > 0$, so for sufficiently small ε we have

$$|p(w + \varepsilon e^{i\theta})|^2 < |p(w)|^2$$

contradicting the definition of w. Therefore there exists $z_0 \in \mathbb{C}$ such that $p(z_0) = 0$.

The same idea can be adapted to give an equally simple proof of Liouville's Theorem:

Theorem 2 If $f(z)$ is analytic on the entire complex plane, and is not constant, then $f(z_0) = 0$ for some $z_0 \in \mathbb{C}$.

The only new feature in the proof is that the polynomial in (1) becomes a power series, and now we really do need Taylor's theorem.

References

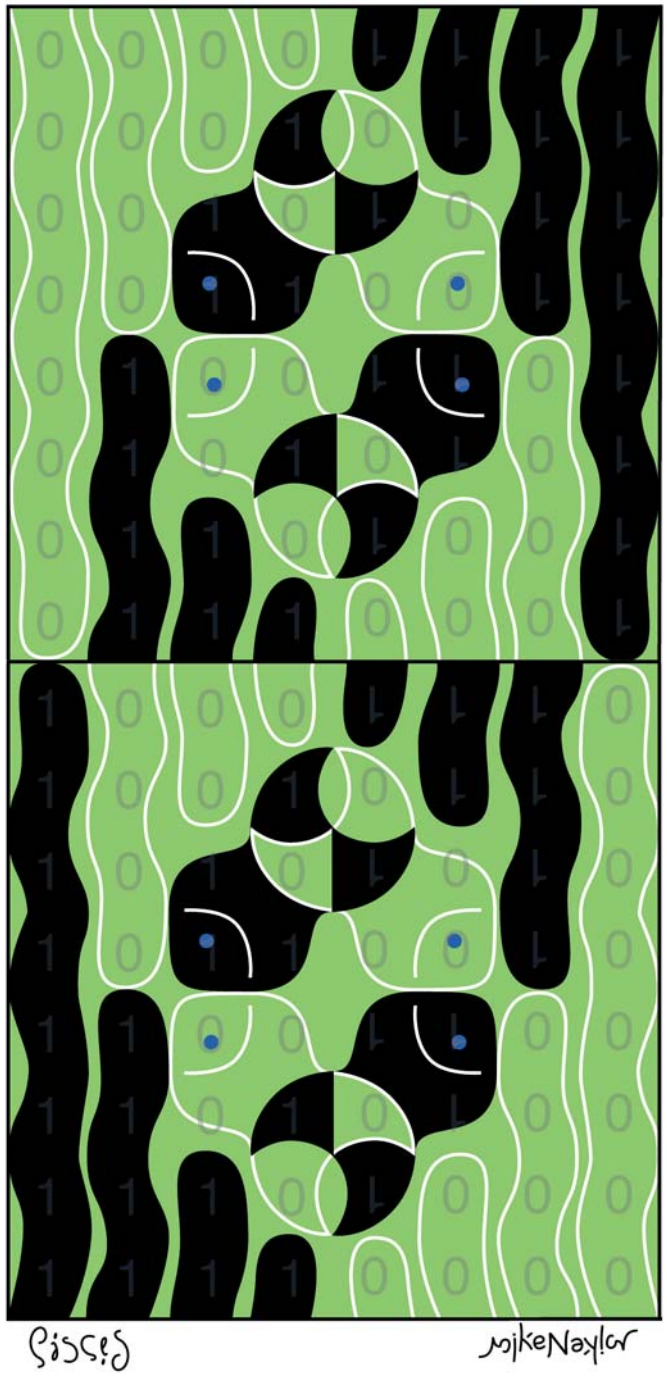
[1] Ian Stewart; 1977; *Gauss*; Scientific American 237 122-131.

[2] Ian Stewart and David O. Tall; 1983; *Complex Analysis*; Cambridge University Press; Cambridge.

[3] Edward C. Titchmarsh; 1939; *The Theory of Functions*; Clarendon press; Oxford.

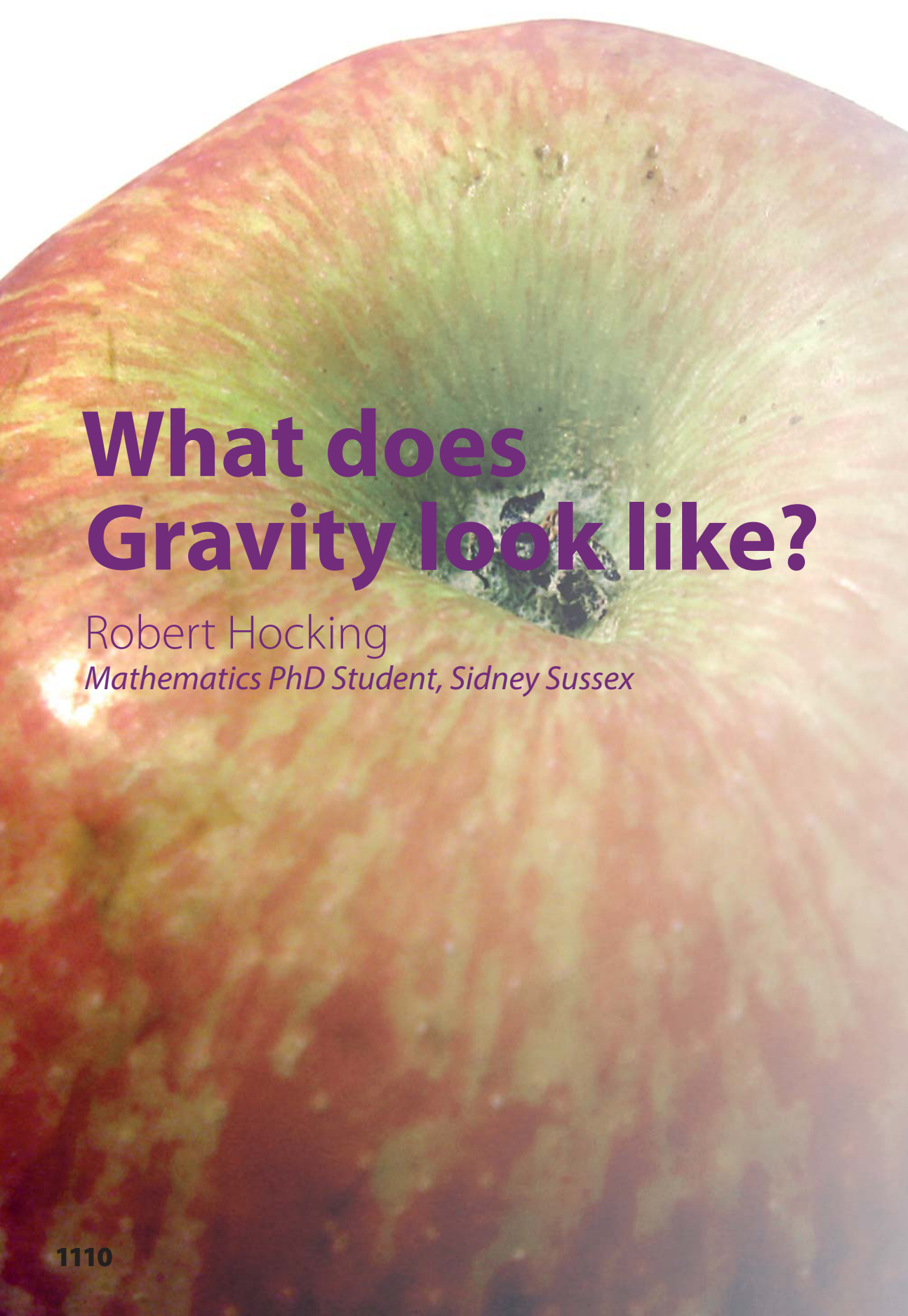
About the Author

Ian Stewart FRS is a professor of mathematics at the University of Warwick, and a widely known author of science fiction and popular science books. His research interests include dynamical systems, bifurcation theory and pattern formation. He lists his recreational interests as painting, guitar, Egyptology and snorkelling.



Pisces, by Mike Naylor

This image is a visualisation of the sequence of binary numbers from 0000 to 1111. The digits determine the colour of each section, forming a pattern of seaweed and fish!



What does Gravity look like?

Robert Hocking
Mathematics PhD Student, Sidney Sussex

One of the more striking predictions of general relativity is the formation of black holes via the gravitational collapse of sufficiently concentrated mass and/or energy. This prediction is also one of the most famous, in part due to popular science books like Stephen Hawking's A Brief History of Time bringing the concept into mainstream culture. Black holes have even made it to Hollywood – for example, in the 2009 film Star Trek, the villain Romulan uses a black hole to consume the Vulcan homeworld. More recently, a "singularity" was created to stop the armies of General Zod in 2013's Man of Steel.

The appearance of black holes in major blockbuster films means that special effects teams have had to take a stab at depicting what one should look like. Disappointingly, the best Hollywood could come up with is not only wrong but downright unimaginative. Star Trek's planet Vulcan turned black hole, arguably the worst, was nothing more than a circle shaped black silhouette. For the same companies that succeeded in creating such convincing CG fire and water in movies like Quantum of Solace and Ice Age 2 to fail completely when it comes to black holes suggests that there might be something intrinsically harder about the latter. However, this is not case. In this article I hope to convince the reader that realistic computer renditions of black holes are not only straightforward to generate, but are also far more compelling and beautiful than anything produced by Hollywood.

The Theory

A natural place to begin is to first address the question of why a black hole is even visible at all. Strictly speaking, the answer here is the same as for regular matter like rocks and trees – black holes are visible because of the effect that they have on passing light. However, whereas the former affect light primarily through reflection and absorption – that is through the electromagnetic force and quantum effects acting on a very small scale – in the case of a black hole it is the long range action of gravity that is important.

General relativity predicts that light rays travel along geodesics in curved spacetime, meaning that light is both bent and focused by a massive body – an effect sometimes called gravitational lensing. This means that the presence of a massive body in (for example) the night sky will affect both the apparent position and brightness of the stars in its vicinity. This effect was first confirmed by Sir Arthur Eddington in 1919 when he measured the (slight) deflection of starlight by the sun during a solar eclipse.

Unlike the sun, which distorts its environment slightly but is visible mainly for other reasons, a black hole appears as a pure distortion in whatever happens to be behind it. Therefore, a black hole's environment is important to its appearance. Although space might seem to be the most natural setting, in this article we instead go with a theme of famous places – rendering locations like the great wall of China and the Eiffel tower as they would appear with a black hole in the vicinity. For the sake of simplicity we only consider distortion due to the deflection of light, neglecting changes in brightness due to focusing, and also colour due to gravitational redshift. For this, it suffices to compute (numerically) the set of light-like geodesics (i.e. light ray paths) starting from a fixed point in spacetime and expanding outwards in all directions. These rays are lines of sight connecting the observer and observed, determining what the human eye (or a camera) sees when looking into a black hole. Modifying the celebrated ray tracing algorithm from computer graphics to make use of these geodesics allows us to simulate gravitational lensing on the computer.

First we need to settle on a suitable spacetime. The most obvious choice is probably the famous Schwarzschild solution – discovered in 1916 only about a month after Einstein published his general theory of relativity – describing a single non-rotating, uncharged black hole which has existed since the beginning of time. However, in this article we will be working with a slightly more exotic creature – charged black holes. The addition of electric charge has (in the author's experience) no effect on the qualitative appearance of a black hole. At the same time, it comes with the important advantage of allowing us to visualize spacetimes containing more than one of them.

In general, multi black hole spacetimes are highly complex, dynamic objects. The nonlinear interaction of the black holes means that the metric cannot be written in closed form, and one must resort to the numerical solution of the Einstein equations. Numerically solving for geodesics on a manifold that is itself constructed numerically is a challenging task, and beyond the scope of this article. However, multiple charged black holes

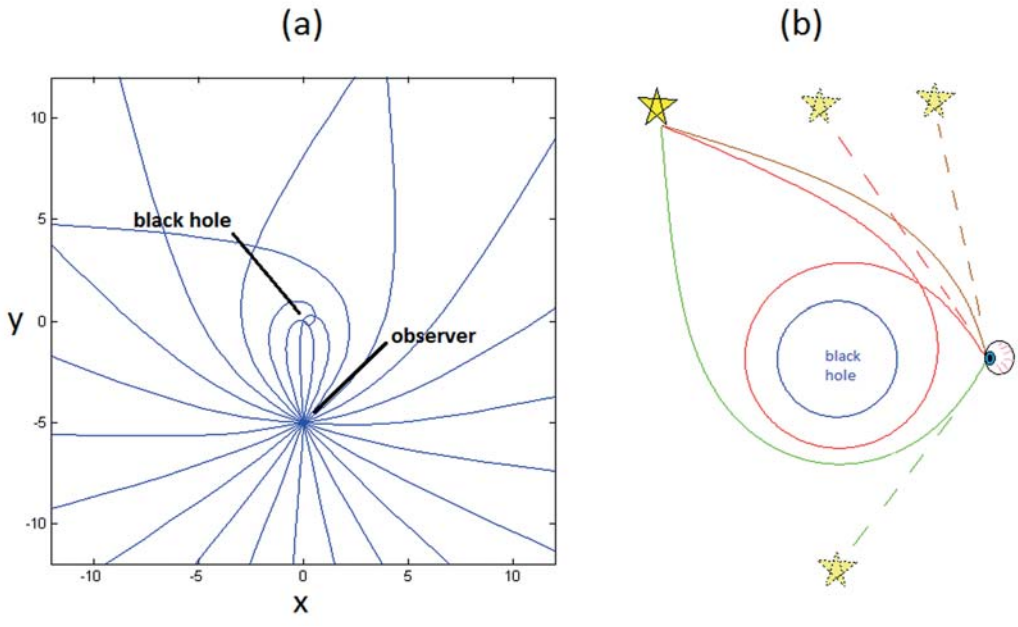


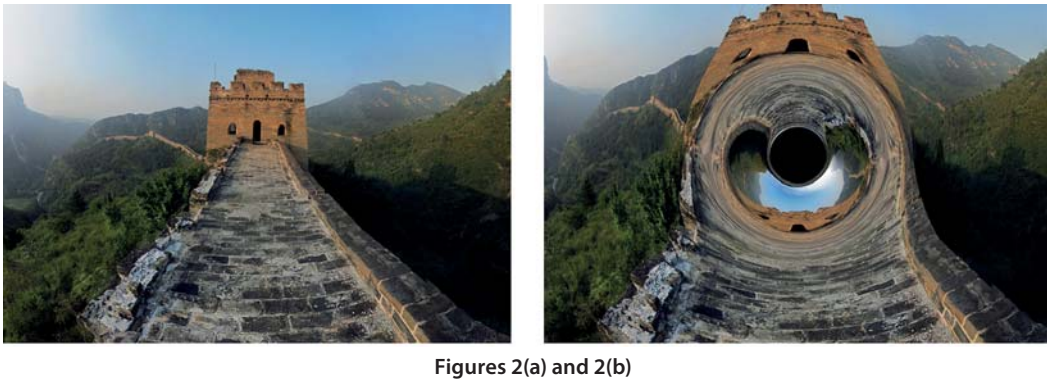
Figure 1

provide a way out – when the amount of charge is just enough to exactly cancel their mutual gravitational attraction, the resulting spacetime is static and can be written in closed form. Such spacetimes belong to the Majumdar-Papapetrou family of solutions discovered in 1947 (see [2]), from which all pictures in this article have been created.

Figure 1(a) shows an example where a set of test rays are fired backwards in time from an observer situated in a Majumdar-Papapetrou spacetime containing a single black hole at the origin. The paths of the rays are calculated by numerically integrating the geodesic equations. Notice that they appear to be divided into two categories – those that are deflected as they pass the black hole but ultimately escape to infinity, and those that are "captured" and fall into the origin.

Here appearances are a little misleading – the origin is actually the black hole's event horizon in our current choice of coordinates, and time (as measured by a distant, stationary observer) is diverging to minus infinity as the rays approach. Thus the apparently captured rays never actually make it inside the black hole (it's a good thing too because if they did, then their time reversals would be rays climbing out from inside the event horizon, contradicting the definition of the latter). Nevertheless for convenience we continue to refer to these rays as captured.

Bearing this in mind, let us try to imagine what our (single) black hole might look like. For the



Figures 2(a) and 2(b)

Figure 3 *(Thanks to Alexandre Duret-Lutz for the original picture)*

purposes of discussion it is conceptually easier if we assume all light sources are far away and point-like, and hence for the time being we assume a backdrop of stars. However, this makes for less interesting pictures so we will drop this assumption when it comes time to show some results.

The apparent position of a star is determined by the tangent vector to the geodesic(s) joining the eye/camera with it – see Figure 1(b). As is often the case for boundary value problems, this geodesic is not unique – hence each star has multiple images. In fact, it turns out that a light ray may circle the black hole any number of times, either clockwise or counterclockwise, en route from a distant star to the observer. Hence each (physical) star maps to infinitely many star images.

The collection of captured rays correspond to a set of directions from which no light can reach the observer. This is called the black hole shadow, and appears as a great dark void – circle shaped, in the present case. Surrounding the shadow is a halo of stars – a result of the accumulation of infinitely many star images into a finite region of image space.

Examples

When the background consists not of point light sources but rather extended bodies, the situation is more complicated, but not a lot more. Figure 2(b) shows a first example with a single black hole in front of the great wall of China (for reference, an undistorted photo is provided in figure 2(a)). As in the discussion, we see a dark shadow region surrounded by duplicates of the objects in the scene. However, since these objects now occupy a finite region in image space, it is possible for their various copies to partially merge, resulting in a kind of Siamese twin effect. This is illustrated in Figure 3, where two black holes hover in front of the Eiffel tower.

It is worth mentioning a kind of gravitational fractal effect that only occurs when the number of black holes is two or more. In this case, the black holes have a lensing effect on *each other*, meaning that around each black hole there are multiple apparent copies of the other black holes. But then these copies are in turn surrounded by more apparent copies, and so on, forever. This is illustrated (to a recursion depth of one) in Figure 4, using St. John's College, Cambridge as a backdrop.

Wormholes

Either a tunnel connecting two otherwise separate universes or else two locations within the same universe, the concept of a wormhole is almost as famous as that of a black hole. However, whereas most physicists and astronomers today accept the reality of black holes as real astronomical objects, the existence of wormholes is far more suspect. While one can construct wormhole spacetimes which technically do solve the Einstein equations, the caveat is that the local energy density (the technical term here is stress-energy tensor) has to take on values which many experts believe to be physically impossible (see [1] pp. 151). This detail has not stopped a parade of science fiction films from featuring wormholes. It also doesn't stop us from computing what a wormhole would look like if it did exist (it goes without saying that

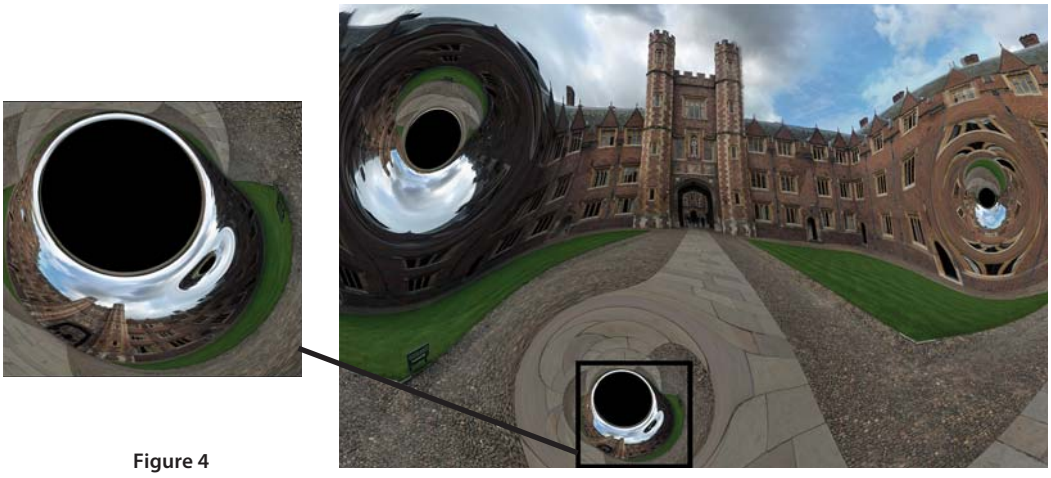


Figure 4

Figure 5 *(Thanks to Alexandre Duret-Lutz for the beach picture and Marco Reinhardt for the monument picture)*

the movies all got it wrong). An example is shown in Figure 5 of a wormhole connecting the Berlin Holocaust monument to a beach in the Caribbean.

Things start to get interesting when the number of wormholes is greater than one. Suppose we take our spacetime with a single wormhole connecting two universes A and B, and add to it a second wormhole also linking A and B. An observer O in universe A can see universe B through wormhole 1. But since universe B contains wormhole 2 leading back to universe A, O can also see a miniature copy of their own universe (including a copy of themselves) by looking first through wormhole 1 and then wormhole 2. The result is an infinite cascade of wormholes within wormholes, similar to the effect of standing between two large mirrors on opposite sides of a hallway.

One can take this game further and construct spacetimes with increasingly complex topologies by adding more wormholes and/or universes, resulting in fascinating pictures. To be appreciated properly these kind of images need to viewed at a higher resolution than is possible in a magazine – therefore I refer the interested reader to view them on my website, www.maths.cam.ac.uk/postgrad/cca/people/lrh30.html.

References

[1] James B. Hartle; 2003; *Gravity: An Introduction to Einstein's General Relativity*; Benjamin Cummings, illustrate edition.

[2] Sudhansu D. Majumdar; 1947; *A class of exact solutions of einstein's field equations*; Phys. Rev. 72:390-398.

See every side of HP

Graduate and internship programmes

We're HP. You probably know we make printers, PCs and laptops – that's a major side to our business – but it's not the only side. We're a multi-faceted organisation whose technologies touch millions of lives across the world every day.

Join one of our graduate or internship schemes and you'll see them all. We're looking for bright minded self starters who could get to grips with our cloud technology, play a part in business strategy or even come up with the security and information solutions to fire the future of our business.

If this sounds like you, come and meet us at one of our campus events.

For dates and times, go to **www.hp.co.uk/jobs**

Sums, Products and Sums-and-Products

Prof Imre Leader
Professor of Pure Mathematics, DPMMS

Sums

Suppose that we partition the natural numbers into finitely many classes. Can we always find x and y such that all of x, y and $x + y$ are in the same class? Equivalently, if we 'finitely colour' the naturals, can we find x, y, z of the same colour with $x + y = z$? This is a typical question in Ramsey theory, which seeks to answer questions of the form 'can we find some order in enough disorder?' In this case, the disorder is the unknown colouring, and the little patch of order is the $x, y, x + y$. (It is worth mentioning in passing that the reader can deduce from the asking of the question that we are not considering 0 as a natural number. More interesting would be to ask 'are we allowed $x = y$', but in fact it is quite easy to find colourings that ensure that we cannot have $x = y$, so in fact it makes no difference whether or not we insert the word 'distinct'.)

For example, if we colour every even number red and every odd number blue then we could take $x = 4$ and $y = 6$. If we colour every square red and every non-square blue we could take $x = 5$ and $y = 6$ (or indeed $x = 9$ and $y = 16$). In any particular example, it may be quite easy to find a suitable x and y, but is this always the case? It turns out that the answer is yes. This is called Schur's theorem, dating from 1916. Schur's theorem is not too hard to deduce from Ramsey's theorem – indeed, it often appears on the examples sheet for the Part II Graph Theory course, in the section on Ramsey Theory.

What about sums of more terms? How about all of the seven non-zero sums from x, y, z? And in general, how about $FS(x_1, \ldots, x_n)$, meaning the set of all sums $\sum_I x_i$, where I is a non-empty subset of the index set? This is much harder: it is not on any Part II sheet that I am aware of. It is true, though, that (for any n) whenever the naturals are finitely coloured there exist $x_1, \ldots, x_n$ such that $FS(x_1, \ldots, x_n)$ is all one colour. This is called the Finite Sums theorem, and it is a special case of a famous theorem of Rado from 1933.

Frank Ramsey (1903-1930) was a brilliant mathematician, philosopher and economist. Ramsey's Theorem was in fact proved by Ramsey as a minor lemma on the way to a result in logic. He suffered chronic liver problems, dying from jaundice at the age of just 26. (*Photo courtesy of Stephen Burch, grandson of Frank Ramsey.*)

Products

What about products? Can we always find, in a finite colouring of the naturals, three numbers x, y, z of the same colour with $xy = z$ (apart from $x = y = z = 1$, of course)? Actually, it turns out that this is a silly question: it follows immediately from Schur's theorem, just by restricting our attention to the powers of 2. Indeed, given a finite colouring of the naturals, consider a 'new' colouring where the new colour of x is the original colour of 2^x. For this new colouring, Schur's theorem tells us that we can find $x, y, x + y$ of the same colour, and this translates to $2^x, 2^y, 2^{x+y}$ having the same colour in the original colouring, as required.

And similarly, of course, for products of more terms. We can always find a set of the form $FP(x_1, \ldots, x_n)$ (meaning all non-empty products) that is just one colour, by exactly the same argument.

Sums and Products

So far so good. We have dealt with sums, and with products. How about combining them? So it is true that whenever the naturals are finitely coloured we can always find x and y such that the four numbers $x, y, x + y, xy$ are the same colour? More generally, of course, we would like to find (for any given n) numbers $x_1, \ldots, x_n$ such that $FS(x_1, \ldots, x_n) \cup FP(x_1, \ldots, x_n)$ (in other words: all sums and all products) is one colour. Actually, even more generally one would like to be able to iterate the summing and the producting (without repeating a term): so for example for three terms one would *really* like to have $x, y, z, x + y, x + z, y + z, xy, xz, yz, x + y + z, xyz$ and also $xy + z, xz + y, yz + x$ and $x(y + z), y(x + z), z(x + y)$.

The answer is: nobody knows. This is an open problem. It has been thought about a lot, but with no success. Remarkably, leaving aside all this 'what we would really want' stuff, it is even an open problem for the case of just two terms! In other words, it is unknown whether or not whenever the naturals are finitely coloured there exist x and y such that all of $x, y, x + y, xy$ are the same colour. It seems utterly scandalous that this very small starter case should remain unsolved. Actually no-one even has any ideas or hunches. No-one has ever put forward a 'proof idea' or 'proof scheme' that even looks for a moment like it might have any chance of working. In the other direction, no-one has ever come up with a colouring that might, even with 5 seconds of thought, have the potential to be a counterexample. To put it another way, for any actual colouring you are told it is invariably very easy indeed to find such x and y.

Part of the reason for the lack of progress on a proof might be what one could call a 'lack of portability'. If we prove some statement about sums (like for example that if we colour the numbers from 1 to 100 with 2 colours then we can find a solution to $x + y = z$ in one colour class), then we can repeat it: if we look at the even numbers from 2 to 200 we will find the same object (in pretentious language, there are plenty of homomorphisms from the group $\mathbb{Z}$ to itself). This 'repeating' feature turns out to be of great importance in most proofs of results about sums. And the same for products: if we know something about products inside the numbers from 1 to 100 then we know the same statement about the set of square numbers from 1^2 to 100^2 (again, there is a pretentious way to say this). But, sadly, there is no way to transfer results about sums-and-products around the place – there are not too many ring homomorphisms from $\mathbb{Z}$ to itself!

It is also possible to give a reason why bad colourings are so hard to think of. To find a colouring without these sum-and-product structures, one would need a colouring that 'meshes well' with addition and multiplication. Now, there are lots of colourings that mesh well with addition (for example, colouring by value modulo something, or colouring by least significant nonzero digit in base something), and similarly for multiplication – but nobody has come up with a colouring that meshes with both.

I have left the most embarrassing 'lack-of-knowledge' until last. We have said that, even in the simplest possible case of 'two terms', i.e. $x, y, x + y, xy$, the answer is not known. But what if we go even further, and ignore the colour of x and y themselves? So the question becomes: is it true that, whenever the naturals are finitely coloured, there exist x and y such that the two numbers $x + y$ and xy have the same colour as each other? (We exclude the case of $x = y = 2$ in this, or equivalently we insist that x and y are distinct.)

Even this super-special small case is unknown. It can be checked on computer for a small number of colours (I think it is known for up to 5 colours), but that is not much evidence in its favour. This is one of the most maddening, and most tantalising, problems in the whole of Ramsey theory!

The Axiom of Choice

Robin Elliott
Second Year Mathematics Undergraduate, Clare College

The axiom of choice (AC) states that for any collection X of non-empty sets, there exists a function from X to the union of all sets in X such that $f(A) \in A \ \forall \ A \in X$. The slogan is "you can choose an element of each set in any collection of non-empty sets". This may seem like an axiom that states the bleeding obvious, but this is only because any intuition we may have of a "choice function" will appeal to finite cases, or sets with "small" infinite cardinality, such as the naturals or reals. The power of AC is that it applies to an arbitrarily large collection of non-empty sets. In the words of Douglas Adams, "Sets can be big. Really big". Given an arbitrarily large collection of non-empty sets U_i, each of which themselves can be arbitrarily large, we have no hope of explicitly constructing a function that chooses an element in each of the U_i, especially if we have little to no information about each of the U_i. Indeed, in general we can't explicitly construct such a choice function, since AC is independent of the rest of standard (ZF) set theory. AC, therefore, steps in to guarantee a non-constructive existence of such a choice function.

The Prisoners and Hat Puzzle

Suppose n prisoners are positioned on the integers 1, 2, ..., n on the number line, and all are facing in a positive direction: prisoners can see other prisoners standing on larger integers than they are at, but not lower integers. An executioner places a black or white hat on each of the prisoners' heads and stipulates that, in some order, each prisoner must guess the colour of his hat. If he guesses correctly he lives and if he guesses incorrectly he dies. The prisoners are not allowed to communicate except for exclaiming their guesses of either black or white, although they are allowed to agree on a strategy prior to being lined up. What strategy should the prisoners employ to minimise the number of deaths, and how many prisoners are going to die?

At first thought, it seems like prisoners may as well guess randomly, since although they can see some other prisoners' hats, this gives them no information about the colour of their own hat. This leads to half of the prisoners dying, and half going free. You may be surprised, therefore, to learn that there is a strategy in which all but one prisoner is assured freedom.

The strategy a mathematically-minded set of prisoners would devise is as follows. Each prisoner counts the number of black hats he can see in front of him, and remembers the parity of this count. Then the prisoner standing at 1 exclaims "black" if he sees an odd number of black hats, and "white" if he sees an even number of black hats. Although he may die as a result of this, the prisoner at 2 now knows the colour of his own hat: if 1 called "black" and 2 sees an odd number of black hats, he knows he must have a white hat. Similarly if 1 called "white" and 2 sees an even number of black hats, 2 knows he is wearing a black hat. For

the other two cases, 2 knows he is wearing a white hat.

But then the prisoner standing at 3 can also deduce the colour of his hat in a similar manner: he knows the colour of the hats of everyone 1 could see apart from his own, and from this he can deduce the colour of his own hat. Continuing inductively, all subsequent prisoners can correctly identify the colour of their own hats.

More Prisoners, with Earmuffs

Let's have the same setup as before, except this time we have infinitely many prisoners standing on the naturals 1, 2, ... instead of finitely many prisoners. We further stipulate that each prisoner wears earmuffs! That is, each prisoner can now no longer hear anything any other prisoner says. The above strategy, cunning as it may have been, will clearly no longer work. What strategy should the prisoners employ now?

More Prisoners, More Hats, with Earmuffs

Let's up the ante further. We have the same setup as before, (i.e. infinitely many prisoners, who cannot hear each other) but instead of two possible colours for the hats, we have uncountably many colours. To give a concrete example, the executioner could assign a real number to each prisoner, with the extra condition that the prisoner knows the real numbers assigned to all prisoners higher up than him on the number line. If the prisoners were to guess randomly, they would all die with probability 1: if I tell you I'm thinking of a real number (it follows a normal distribution, say), you have probability 0 of guessing it first time round.

Remarkably, in this set-up – and indeed the previous one with infinitely prisoners and two hat colours – we can ensure that all but finitely many prisoners live. That is, past some prisoner standing at N_0, we have that everyone remaining correctly guesses the real number that has been assigned to them. The strategy relies heavily on AC and is therefore, like all proofs relying on AC, non-constructive.

Strategy with Choice

We present here the strategy that works for an

infinite number of earmuffed prisoners, and the 'colours' of the hats members of any set X. Consider the set $X^{\mathbb{N}}$ of sequences in X, that is, sequences $(x_1, x_2, ...)$ such that $x_i \in X \; \forall \; i$. Define an equivalence relation $\sim$ on $X^{\mathbb{N}}$ by $A, B \in X^{\mathbb{N}}$ having $A \sim B$ if and only if A and B differ only in finitely many terms. That is, $A \sim B$ if there exists some N_0 such that, for all $n \geq N_0$, $A_n = B_n$. Checking $\sim$ is an equivalence relation is not too strenuous, and once this is established we can consider the equivalence classes $X^{\mathbb{N}}/\sim$. Each equivalence class consists of "sequences in X that eventually end up the same", so for example if X = {black, white} then one equivalence class of X would be "all sequences which eventually end up all black". Another would be "all sequences which eventually have black in even positions and white in odd positions".

Now, for each equivalence class C in $X^{\mathbb{N}}/\sim$, pick an element in C. This seemingly simple statement is where we have invoked the axiom of choice: $X^{\mathbb{N}}/\sim$ is our collection of non-empty sets, and by choosing an element in C for all C in $X^{\mathbb{N}}/\sim$, we are assuming the existence of a choice function on the elements of $X^{\mathbb{N}}/\sim$. So we have a set $\mathcal{F}$ such that for each C in $X^{\mathbb{N}}$, there exists an S $\in \mathcal{F}$ such that $C \sim S$. In more informal terms, we have a set $\mathcal{F}$ which contains examples of all possible ways a sequence in X can eventually end up.

We are nearly there. The prisoners agree beforehand on the set $\mathcal{F}$, which they memorise. Then the strategy is simple: each prisoner inspects all those in front of him, and so can tell which equivalence class of $X^{\mathbb{N}}/\sim$ he is in. He recalls the element s of $\mathcal{F}$ corresponding to that equivalence class, and if the prisoner is at position n then he exclaims the nth element of s.

Why does this work? All prisoners, no matter what position they are in, will recall the same element s of $\mathcal{F}$ given their viewpoint of the rest of the sequence. If the actual sequence of hat colours does not correspond to the sequence s, it must at least differ from s only in finitely many terms, since it is equivalent under the relation $\sim$ to s. So there is a point at which the actual sequence of prisoners' hats will always agree with s, and any prisoners past this point will correctly identify their own hat colour.

Further Variations, Additional Remarks

1. It is interesting to consider further the case without earmuffs, and also without using the axiom of choice. We may have a greater cardinality of prisoners (in some totally ordered set), or of hat sizes. The prisoners can only transmit information from a set of the same cardinality as the set of hat sizes (e.g. in the black/white case, no prisoner could say a real number; they would have to say black or white). For example, in the case with real-numbered hats between 0 and 1, the reader may like to find a convergent subsequence the first prisoner can sum to guarantee the lives of all of the other prisoners. Indeed, the reader is invited to think about which sets X have the same cardinality as $X^{\mathbb{N}}$, and why for these X the prisoners can ensure that all but the first survive.

2. If the warden knows the quotient set $\mathcal{F}$ that the prisoners have chosen, he can ensure that any prisoner of his choosing will die, by choosing the element s $\in \mathcal{F}$ and then giving said prisoner a hat colour not equal to his corresponding hat colour in s. Moreover, if the warden knows the quotient set $\mathcal{F}$ he can choose any finite set of prisoners to die.

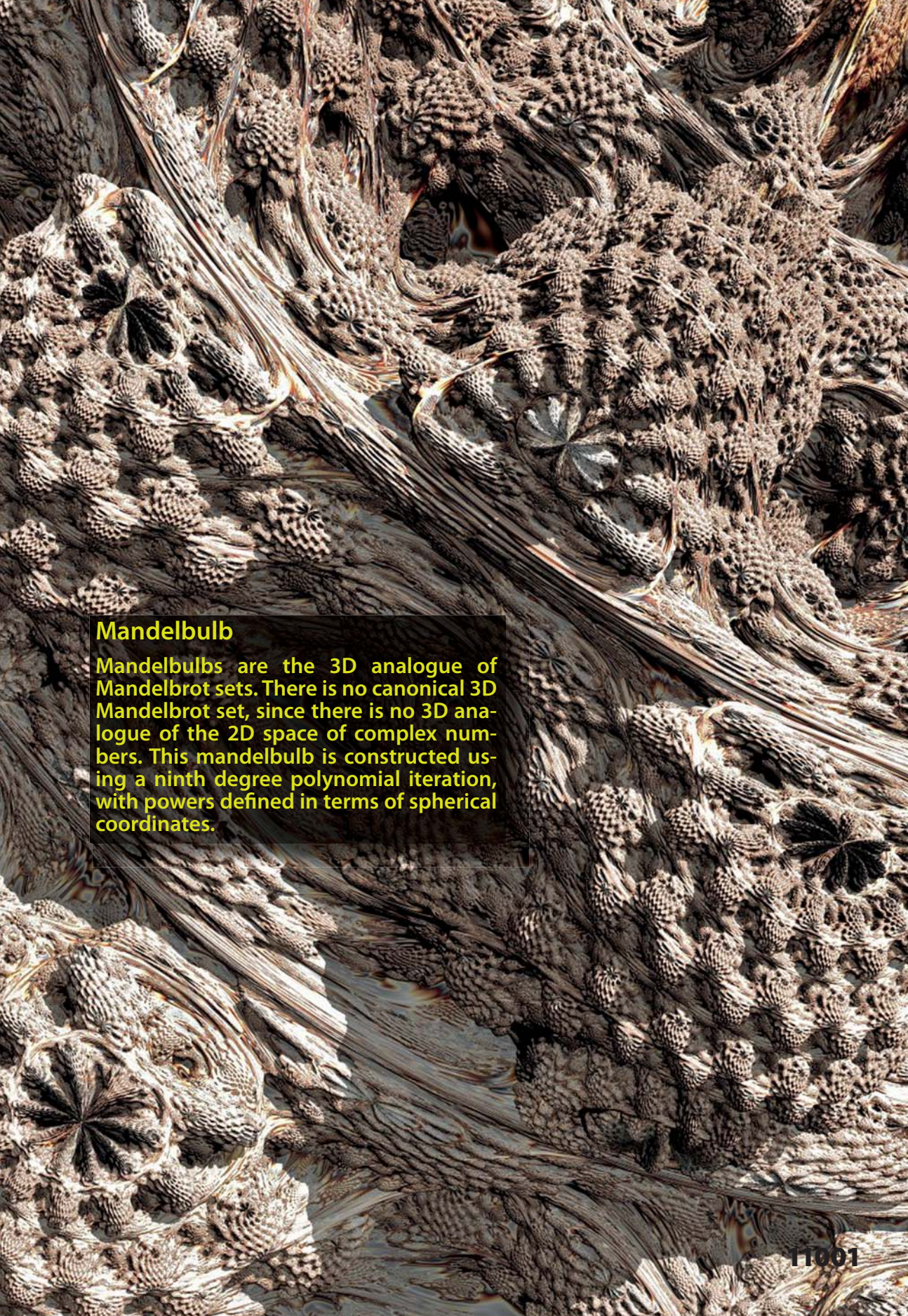
3. Editor's Note: Extending this idea, the reader may like to consider what happens when the warden chooses each hat colour according to some continuous distribution. What is the probability that a given prisoner dies? What is the probability that the first N all die? He/she may then recall the continuity property of probability. (The interested reader may wish to consult a text on measure theory, such as [3].)

References and Further Reading

[1] *Prisoners and hats puzzle*; Wikipedia; http://en.wikipedia.org/wiki/Prisoners_and_hats_puzzle.

[2] Thomas J. Jech; 1973; *The Axiom of Choice*; Dover.

[3] Marek Capiński, Ekkehard Kopp; 2004; *Measure, Integral and Probability*; Second Edition; Springer.



Mandelbulb

Mandelbulbs are the 3D analogue of Mandelbrot sets. There is no canonical 3D Mandelbrot set, since there is no 3D analogue of the 2D space of complex numbers. This mandelbulb is constructed using a ninth degree polynomial iteration, with powers defined in terms of spherical coordinates.

Spot It!® Solitaire

Dr Donna Dietz
Professorial Lecturer of Mathematics and Statisics, American University

As I was shopping online one day, an advertisement for Spot it!® caught my eye. This game is played with 55 circular cards. Each card has several images, and each pair of cards has exactly one common image (see Figure 1). Several games can be played with the deck, all involving multiple players trying to be the first to spot the matching image between two cards. With my curiosity piqued, I purchased the deck, which is made by *Blue Orange Games*. I quickly discovered that the deck is two cards shy of fully representing an order 7 finite projective plane. It seemed a natural course of action to create the two missing cards and then proceed to arrange the cards into a configuration which would make it easy to demonstrate the order 7 finite projective plane. I didn't realize how fun and challenging this would be. I'm hoping the rules (and solution) of this single-player challenge will be entertaining to mathematicians and game-lovers alike. For those who would like to play, but who don't have a Spot it!® deck, interactive games are available on my web page: http://www.donnadietz.com/Projective.html.

Figure 1 These two cards have an image in common

(Thanks to Theirry Denoual, co-founder of Blue Orange Games, for permission to use this artwork.)

Background

Before discussing the higher order finite projective planes and affine planes, let us address the order 3 case (see Figure 2). In both the projective plane and the affine plane, points are connected by "lines", and any lines not sharing a point are parallel. In the affine plane, there are four sets of three parallel lines (which share a colour), each with three points. However, in the projective plane, there are four additional points and no parallel lines. An affine plane of prime order n contains n^2 points and has $n + 1$ sets of n parallel lines, each with n points. The associated projective plane contains $n^2 + n + 1$ points and $n^2 + n + 1$ lines, each pair of points sharing a line and each pair of lines sharing a point. From the projective plane, *any* $n + 1$ (here, four) collinear points may be removed, along with all of their incident lines, and an affine plane results.

[1] is a great overview of the finite geometry behind Spot it!®. For those whose interest in finite geometry is beyond the scope of this discussion, I recommend [2] for those who have not yet mastered abstract algebra, and [3] and [4] for those who have. This technique generalizes for other orders, although its utility as a game diminishes due to the quadratic growth of the set size. This discussion is arranged so that those wishing fewer clues can read fewer sections, thus leaving more of the fun for themselves.

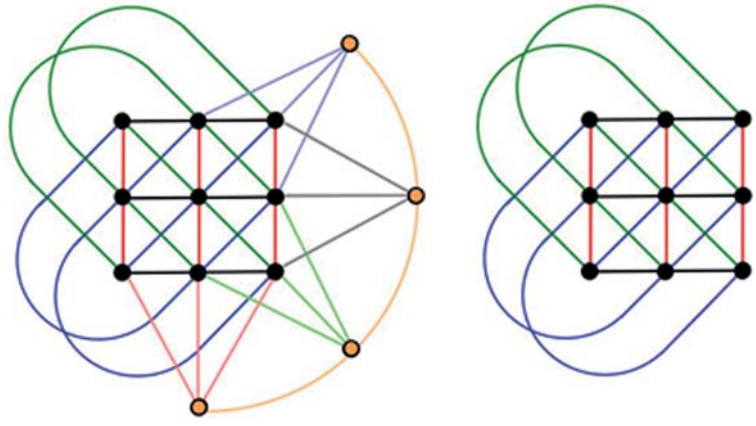


Figure 2 The order 3 projective (left) and affine (right) planes

Finding the Missing Cards

I will presume for the moment that the reader wishes to find the two "missing" cards in a Spot it!® deck. First, list the images in the deck. A quick way to do this is to locate an image which occurs eight times (that is, $n + 1$), and pull out all those cards. For example, if the deck has eight spiders, pull out the spider "set". There can be no other images common to any pair out of those eight cards. Thus, there must be $8 \times 7 + 1$ or all 57 images on those eight cards. Next, tabulate the frequencies for each image. One image is present only six times, while 14 images are present seven times. All other images should be present eight times, as they are not missing from any card. The image which is missing twice must be assigned to both missing cards. Without loss of generality, assign one additional missing image to one of the missing cards. Call this image the "reference image". For the remaining 13 images, search the entire deck to see if it occurs with the reference image or not. If it does, it cannot do so again, so it must be assigned to the other missing card. If, however, it does not occur with the reference image, it should. So it should go on the missing card which has the reference image.

The Rules

Begin by removing all cards having a specific common image which we will call the "infinity image" The remaining cards form an affine plane of the order 7 (or n). (In my Spot it!® deck, the two missing cards both contain a snowman. So, if I simply remove all the snowman cards at this step, I do not actually need to find the two missing cards in order to proceed.) The infinity cards are analogous to the orange points in Figure 3.

The ultimate goal is to lay out the remaining 49 (or n^2) cards in a 7×7 (or $n \times n$) grid so that this one rule is satisfied: Given any two cards in the grid, with positions given by (x, y) and $(x + h, y + k)$ (with x and y numbered between zero and six inclusive), the common image of the two cards must also be present at position ($x + 2h \pmod 7$, $y + 2k \pmod 7$) (or ($x + 2h \pmod n$, $y + 2k \pmod n$)). For example, in Figure 3, the solved $n = 5$ case is shown. Consider the (row, column) positions (1, 1) and (2, 4). The red coloured circle (with a plus symbol) is common to these cards, so at (3, 2), we expect to find this symbol again, and indeed it is there. Since seven (or n) is prime, and all elements are generators in $\mathbb{Z}_7$ (or $\mathbb{Z}_n$) there will be seven (or n) such images in a set, within the 7×7 ($n \times n$) grid. This also implies that each row (and each column) will have a common image. (The symbols in Figure 3 give the same information as the colours. For example, all the red symbols have a "plus" sign on them.)

The families of parallel images in the $n = 5$ affine plane are analogous to the parallel lines in Figure 2. For one parallel family, all images lie on lines having slope 1. In Figure 3, this is the coloured triangle family. The purple coloured triangles lie on the line $r = c + 1 \pmod 5$, while the red ones lie on the line $r = c + 4 \pmod 5$. The blue coloured squares are from another parallel family, and they lie on the line $r = -2c + 2 \pmod$

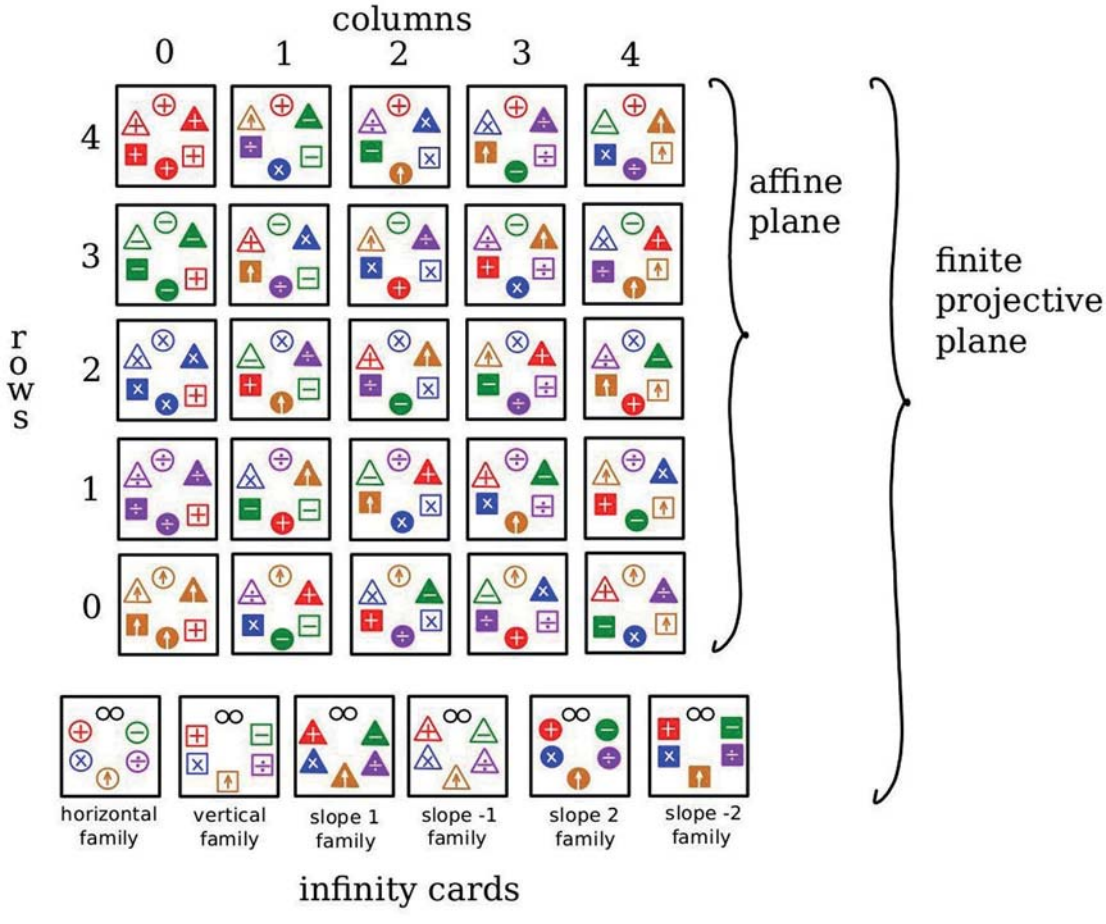


Figure 3 The solved challenge for the $n = 5$ case

5). (Another equivalent equation for this line is $c = 1 + 2r \pmod 5$).

The method for creating such a deck of cards should be obvious now. First, generate the affine grid using parallel lines. Then, each parallel image family is placed on a new card, together with the infinity image. However, the goal here is not to create such a deck but rather to properly arrange an already existing deck. Those readers desiring the maximal fun should now attempt to solve this inverse problem without reading further. One more warning will be given after the easy clues are presented.

Initial Setup of the Grid

First, pull out any set of eight cards which share an image, to be used as the infinity cards. (Or, if the missing cards have not been created, the six cards sharing the twice-missing image should be pulled aside.) Next, select one of the infinity cards to represent your row family and one to represent your column family, and keep them in view. Arrange your grid so that each column contains a common image and each row contains a common image. (Note that there are $57 \times 8 \times 7 \times 7! \times 7!$ ways to make these choices. There are 57 images to pick as the infinity image, then eight cards from that infinity set which can be used to define the rows. Once the rows are chosen, seven cards remain to define the columns. There are 7! ways to order rows and 7! ways to order columns. Also, note that the $n = 3$ case is fully solved at this stage.) Next, by swapping rows and/or columns of cards, place a common image on the line which runs from the lower left corner to the upper right corner ($r = c$), which we'll call the first diagonal from now on. All "moves" henceforth will consist of swapping two rows or swapping two columns. We know from abstract algebra that all permutations can

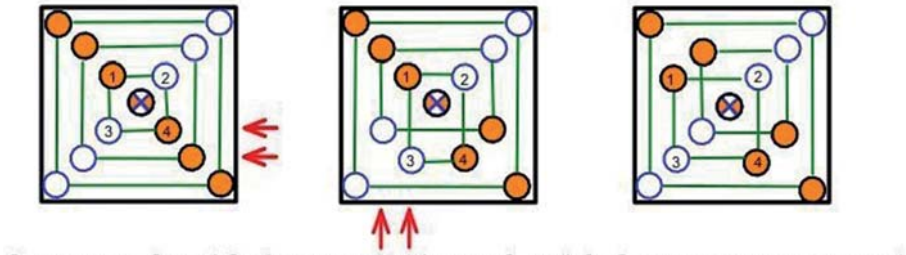


Figure 4 Squares on the diagonal deforming and returning again

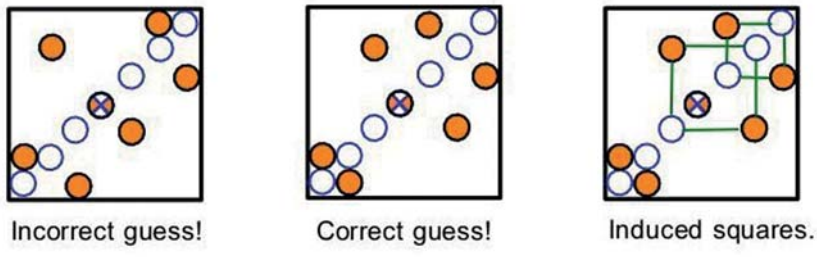


Figure 5 Choosing the correct image for the second diagonal

be formed by swaps, so swapping rows/columns is sufficient to solve this puzzle.

The next objective is to get a common image on the second diagonal, which we define to run from the upper left corner to the lower right corner ($r = -c - 1$). Swapping rows will preserve the common image along the first diagonal as long as the columns with corresponding indices are also swapped. For example, if rows 2 and 3 are swapped, then columns 2 and 3, the net effect on the line $r = c$ is to swap the card at (2, 2) with the card at (3, 3). This technique allows us to maintain the $r = c$ diagonal while still giving enough freedom to finish the puzzle. Without loss of generality, freeze the middle card. Thus, you will not move the middle row or middle column. (The fact that we are working on a torus allows us to freeze any one card, but the choice of the middle card makes the solution easier to execute, due to symmetry.) This gives 6! = 720 remaining grid arrangements, six of which are valid solutions.

Finding the Second Diagonal and Finishing the Puzzle

At this point, if you wish to have any fun with the puzzle, you should stop reading. This is your last warning!

The somewhat surprising fact is that once you have chosen the image for the line $r = c$ and have frozen the middle card, the image for the line $r = -c - 1$ is already determined. It has to be one of the images on your middle card, but it cannot be the image of that row or of that column or of the line $r = c$. So there appear to be five options remaining. But that is not so. Only one will work, and any attempt to set the incorrect image will end in frustration! So how do we figure out which image will work?

Since all moves are reversible, we may simply track the scrambling process to see where images on the second diagonal $r = -c - 1$ may go when we allow paired swaps of rows/columns. We imagine three (that is, $\frac{n-1}{2}$) concentric squares around the frozen middle card to help us keep track of where the second diagonal set can go. Figure 4 demonstrates one of the 15 possible deformations of the second diagonal and also demonstrates how, as promised, the cards along $r = c$ are, as a set, invariant. The white circles represent the cards on the line $r = c$, while the orange circles represent those which will ultimately be on the line $r = -c - 1$. The (group) actions of swapping matched rows/columns maintains these three sets of four cards as corners of squares which are symmetric about the line $r = c$, though they are not concentric ex-

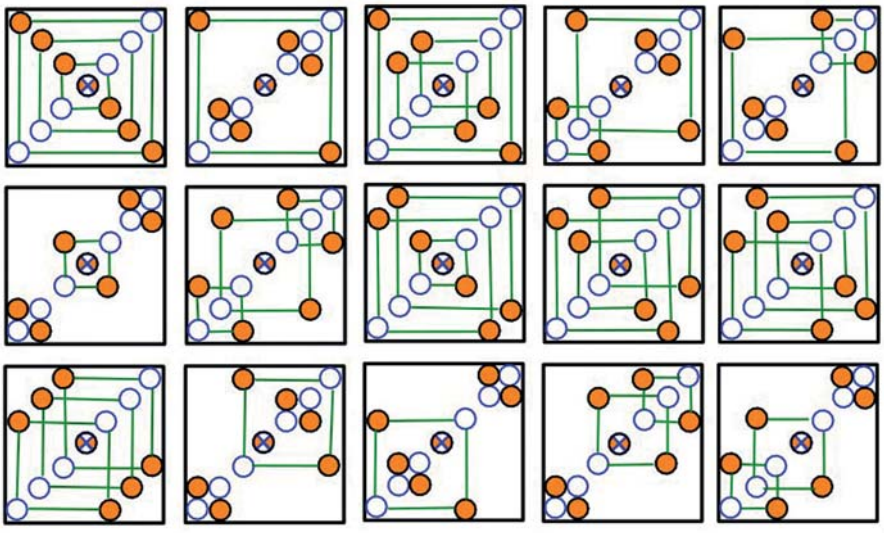


Figure 6 The 15 ways the cards of the second diagonal may initially appear

cept when the two diagonals are both correctly set. To determine which of the five candidate images should be used as the image for the line $r = -c - 1$, we simply search for the image which is already symmetric with respect to the line $r = c$ (see Figure 5). We can think of the six non-fixed cards along $r = c$ as being in three pairs, thus inducing the squares. Combinatorially, this can occur 15 ways, as shown in Figure 6. Swap rows and corresponding columns until the images on the second diagonal are set. Note that, for the order 5 case, even the incorrect candidate images for the second diagonal are arranged symmetrically with respect to $r = c$. Look for four misplaced candidate images instead of two.

Once the two diagonals are set, remaining moves must not only have paired row/column moves, but must also maintain symmetry between right and left (as well as up and down). For example, if rows/columns two and four are swapped, this is already balanced and will not disrupt either diagonal line. However, if columns zero and one are swapped, columns five and six must also be swapped (as well as rows zero and one, and also rows five and six). By tightening these orbits, we close in on one of the six solutions.

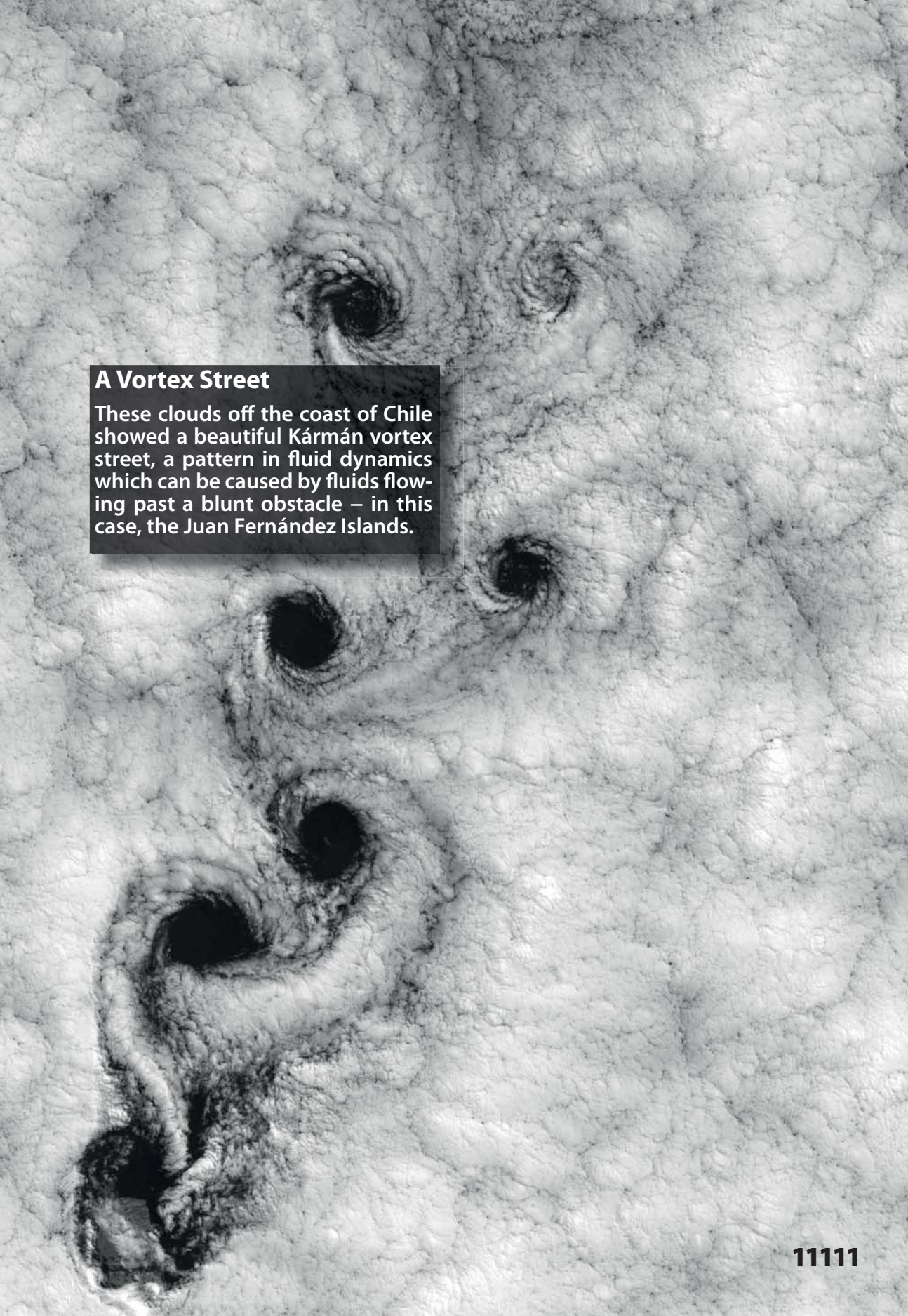
Final Moves

Once the two diagonals are established, there are still 48 possible card arrangements. Each "square" may be in each of the three locations (3! = 6), and each has two legal orientations as it is legal to rotate it 180°. Since $2^3 = 8$ and $6 \times 8 = 48$, there are 48 possible arrangements of the cards, six of which are solutions. For example, you have the freedom to choose any one square's location and orientation, but the rest is then predetermined. For simplicity, we presume that the innermost square is set properly, i.e. the nine cards in the middle of the grid are now fixed.

Now, using what is known about the families of parallel images, move the cards into their final positions. For example, since we now know which family of images has a slope of –1, we can deduce which of those images should appear on the cards in positions (0, 4) and (1, 3) just by looking at the card in position (2, 2). If an image is not in the desired location, look for it on the opposite side of the affine grid, relative to a 180° rotation about the centre. You might also need to swap the middle and outermost squares. In a few moves you will see before you a perfectly arranged affine plane!

References

[1] Maxime Bourrigan; 2011 *Dobble et la geometrie finie. Images des Mathematiques*; CNRS.
[2] Harold L. Dorwart; 1966; *The Geometry of Incidence*; Prentice-Hall, Inc, Englewood Clis, N.J.
[3] James W. P. Hirschfeld; 1998; *Projective Geometries over Finite Fields*; Second Edition; Oxford Science Publications, Clarendon Press; Oxford.
[4] Daniel R. Hughes, Fred C. Piper; 1973; *Projective Planes*; Springer-Verlag, New York.



A Vortex Street

These clouds off the coast of Chile showed a beautiful Kármán vortex street, a pattern in fluid dynamics which can be caused by fluids flowing past a blunt obstacle – in this case, the Juan Fernández Islands.

Chains Between Prisoners

Prof Karl Sigmund and Prof Christian Hilbe
University of Vienna and Max Planck Institute, Germany

The Prisoner's Dilemma game is the mathematicians' gift to moral philosophy. The odd name of the game comes from the story of two prisoners accused of a joint crime. They are kept in separate cells. The state attorney visits them in turn, and offers each a deal. If one of them confesses, he will become a witness for the prosecution, and immediately released; the other one will be sentenced to ten years. If both confess, however, there is no need for a witness for the prosecution, and both will face seven years in jail. Of course, both could refuse to rat on the other. In that case, the attorney promises to keep them in custody for the maximum period legally allowed, one year. No matter what the other is going to decide, the best response for either of them is to confess. Hence both confess. Each will have seven long years to ponder the diabolic nature of the attorney's offer!

A prisoner with a gift for abstraction may come up with a matrix resuming the conundrum with a chain of inequalities. In the tradition-bound jargon of game theory (the mathematics of conflicts of interest), two players 1 and 2 each have to decide between two strategies, namely C (to cooperate) and D (to defect). This yields four possible outcomes for player 1, given by R, S, T or P (see Figure 1). Player 1 prefers T to R, R to P, and P to S; if the outcomes correspond to numbers (so-called *utilities* or *payoffs*), $T > R > P > S$. If player 2 chooses C, player 1 is better off playing D, since $T > R$; if player 2 chooses D, player 1 is better off playing D, since $P > S$. Hence, no matter what player 2 chooses, player 1 should opt for D. But since the same goes for 2, who is in exactly the same situation, both end up with P. And the reason that they have forfeited the better outcome R is, ironically, self-interest!

The Iterated Prisoner's Dilemma

The situation changes considerably if the two players interact not only once (the so-called *one-shot* Prisoner's Dilemma) but repeatedly (an *iterated* Prisoner's Dilemma). In daily life, repeated interactions are the rule rather than the exception. One can rarely be sure of seeing the last of another person. If the likelihood of another encounter is high, players can hope to influence their co-player's behaviour. It is possible to react. In the gift-giving game, for instance, players will anticipate that if they fail to give in one round, they will fail to receive in the following round. And our 'prisoner' may thus be swayed by the prospect that

the co-player will have an opportunity for 'getting even'.

In the iterated Prisoner's Dilemma game, the strategies must specify what to do in each round. For instance: defect in each round (this is the strategy *AllD*). Or else: play *C* in round *n* iff *n* is prime. Or else: play *C* with probability 0.5. More interesting opportunities are provided by strategies conditioned on the co-player's actions. For instance: play *C* in the first round, and from then on, play *C* iff the co-player used *C* in the previous round (this is the strategy *Tit For Tat, TFT*). Or else: play *TFT* in the first hundred rounds, then *AllD* for the remainder of the game, etc.

Memory-One Strategies

Strategies for the iterated Prisoner's Dilemma can be extremely complex; the only restriction is that they cannot depend on the future. A particularly simple class of strategies – the *memory-one* strategies – have the property that the player's next move depends only on the current *state of the game*. The state in round n is determined by the payoff $k \in \{R, S, T, P\}$ obtained by player 1 in that round. Thus a memory-one strategy is given by a quintuple $(p_R, p_S, p_T, p_P, p_0)$, where p_0 is the probability to play C in the initial round ($n = 0$) and the vector $\boldsymbol{p} = (p_R, p_S, p_T, p_P)$ denotes the conditional probabilities to play *C* in the next round, when the current state is $k \in \{R, S, T, P\}$. If the probability δ for another round is fixed (with $0 < \delta < 1$), then the number of rounds is a random variable with expectation $\frac{1}{1-\delta}$. If $\pi_1(n)$ is player 1's payoff in round *n*, then the total payoff is $\sum_{n=1}^{\infty} \pi_1(n)$ and the payoff per round is defined as the Abelian mean:

$$\pi_1 := (1-\delta)\sum_{n=0}^{\infty} \delta^n \pi_1(n)$$

The same holds for player 2's payoff π_2. In the limiting case $\delta = 1$, the payoff per round is given by the Cesaro mean:

$$\pi_1 := \lim_{N\to\infty} (N+1)^{-1} \sum_{n=0}^{N} \pi_1(n)$$

This limit need not always exist: for instance, if the co-player plays *C* in the first ten rounds, *D* in the next hundred rounds, *C* in the following thousand rounds, *D* in the next ten-thousand rounds etc, then the mean payoff is likely to oscillate endlessly. But if the Cesaro mean exists, it is the limit of the Abelian mean, for $\delta \to 1$. The pair (π_1, π_2) is in the quadrangle *Q* spanned by the four points (R, R), (S, T), (T, S) and (P, P) (Figure 2). If player 1 adopts a specific strategy, then the pair (π_1, π_2) depends on player 2's choice, and will typically range over a two-dimensional subset of *Q*. But when player I adopts a so-called *zero determinant (ZD) strategy*, then the pairs are restricted to a line. The construction of such strategies is not too tricky, and further details are provided in the appendix.

If player 1 adopts a ZD strategy, then

$$\pi_2 - \kappa = \chi(\pi_1 - \kappa)$$

irrespective of the strategy adopted by player 2. This means that player 1 can unilaterally enforce a linear relation between the two payoff values – the two prisoners are 'chained together'. Thus the payoff pair (π_1, π_2) is restricted to the line with slope χ intersecting the diagonal in (κ, κ). The value of κ corresponds to the payoffs of the two players when both use the same strategy. If the parameters of the Prisoner's Dilemma game satisfy $P \leq \frac{T+X}{2} \leq R$, then one can easily show that κ is between P and R, whereas χ is between –1 and 1. Let us consider a few instances. By choosing $\chi = 0$, player 1 can assign payoff κ to the co-player. Player 2 need not be restricted to memory-one strategies, but can switch capriciously between *C*s and *D*s. Player 2's actions can affect only the payoff for player 1, but not his own payoff π_2. The converse aim (to assigns a specific value to his own payoff) cannot be realized. Indeed, since the gradient of the payoffs line is at most one, it cannot be vertical.

Some Classes of Strategies

By choosing $\kappa = P$ and $\chi = 1/2$, player 1 can act as *extortioner*. Whenever player 2 attempts to gain a surplus over and above the maximin value *P*, player 1's surplus will be twice as large. When two extortioners meet, their surplus will be zero, of course.

Possibly the most surprising class of ZD strategies is obtained for $\kappa = R$ and $0 < \chi < 1$. Players using such a strategy provide payoff R to their co-player, as long as they themselves receive *R*. If their payoff is less than *R*, however, then the co-players' payoff is also reduced. This seems to be just the type of strategy that conditional cooperators

have in mind. It is a strategy of 'live and let live', but with a barb: 'if you push me down I will take you down'! No self-interested co-player will have reason to cheat on such a player, or both payoffs will be reduced (that of the co-player is actually reduced by less, since $0 < \chi < 1$). Such strategies fare remarkably well, despite their readiness to shoulder the larger part of an eventual loss (with respect to R). They seem to embody the spirit of partnership.

If player 1 sees the co-player not as a partner, but as a rival, then 1's main aim will be to ensure that $\pi_1 \geq \pi_2$. The only ZD strategies which achieve this are strategies with $\kappa = P$, such as the extortioner strategies. (Aiming for the strict inequality $\pi_1 > \pi_2$ is infeasible, of course, since the co-player can play *AllD*).

The Evolution of ZD Strategies

Let us consider a very large population of players, all using the same pair of values (χ, κ). If a small dissident minority using a slightly changed pair shows up, it may do better or worse than the resident population. If it does worse, it will vanish. If it does better, it will spread, and eventually take over. It can easily be shown that it is most advantageous for the dissident to keep the slope χ unchanged, to increase κ if the slope is positive, and to decrease κ if it is negative. This means that whenever the payoff values are positively related (i.e., when $\chi > 0$), then adaptation, in the sense above, drives κ from P to R. Evolution leads from extortion to generosity. This is a surprising result: whereas extortion strategies are those that guarantee that the player's own payoff is not less than the co-player's, generous strategies ensure that his payoff is never above the co-player's. Does this suggest, then, that the meek will inherit the Earth?

References

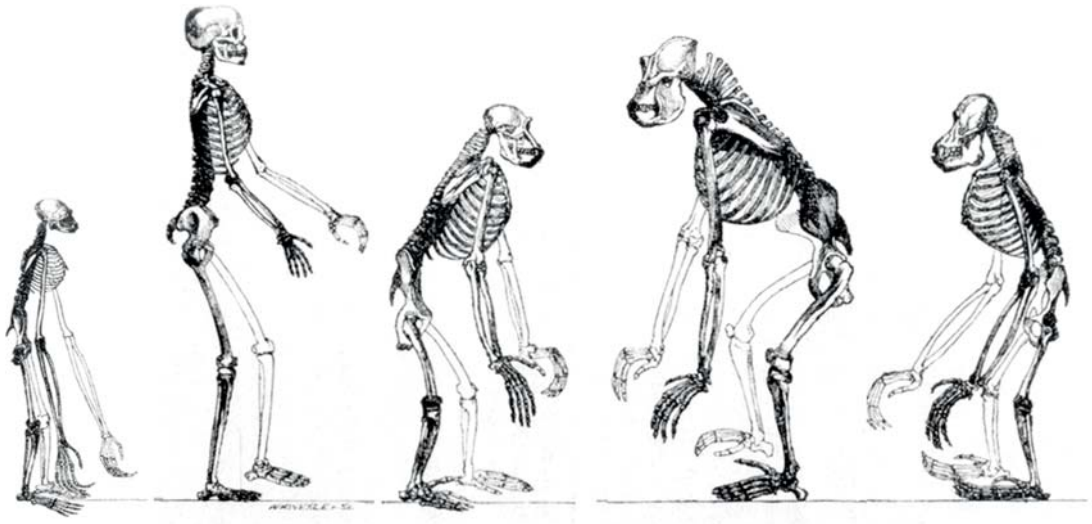
[1] Christian Hilbe, Martin A. Nowak, Arne Traulsen; 2013; *Adaptive dynamics of extortion and compliance*; Working paper.

[2] William H. Press, Freeman J. Dyson; 2012; *Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent*; ProcNat Acad Sci USA 109: 10409–10413.

[3] Karl Sigmund; 2010; *The Calculus of Selfishness*; Princeton University Press.

[4] William Poundstone; 1992; *Prisoner's Dilemma*; Doubleday.

[5] Alexander J. Stewart, Joshua B. Plotkin; 2013; *From extortion to generosity, evolution in the Iterated Prisoner's Dilemma;* Proc Nat Acad Sci USA, in press.



Solving, not Selling

The markets in which we trade change rapidly, but our intellectual approach changes faster still. Every day, we have new problems to solve and new theories to test. We use innovative technology, a scientific approach, and a deep understanding of markets to stay successful. With over 380 employees in our New York, London, and Hong Kong offices, that's a lot of ideas. Our next great idea could come from you; what will you come up with?

Curious? Learn more at www.janestreet.com

LEARN • TRADE • CODE • TEACH

Jane Street

NEW YORK • LONDON • HONG KONG

Tom Anthony, Yanqing Cheng

Archimedeans Annual

Problems Drive

1 Infinite Sequences

Is it possible to arrange an infinite number of ones and zeroes such that no pattern (of any given length) is repeated three times in a row (for example 111, or 111011101110).

2 Isometries

Find a bounded subset X of the real plane and a distance preserving function f on X which takes X to a strict subset of itself.

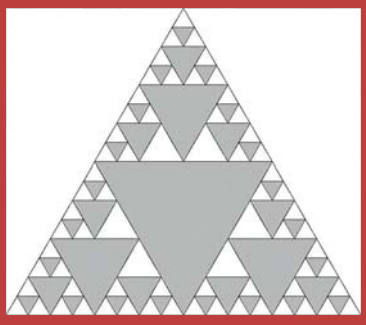
3 Recursively Coloured Triangles

Consider the following construction:

Begin with a white triangle of area 1

At each step, join the midpoints of the edges of every white triangle that has an edge contained in an edge of the original triangle, and shade the resulting triangles.

After 4 steps, the diagram looks like this:



How much area is eventually shaded?

4 Primes

A deletable prime is a number such that it is a prime number, and digits can be deleted one by one from its base 10 expansion such that a prime is obtained at each step. Which of a) 395247625, b) 410139761, c) 410256793 and d) 357019711 is a deletable prime?

5 Subgroups

Consider the group of possible permutations of a Rubik's Cube®. What is the order of the subgroup generated by rotating the middle 'slices' by half turns?

6 Random Numbers

Four of these are random numbers, and one is an encrypted message. Which is the encrypted message?

a) 0011000111011000101010111110001010100010011110110100111010000101010111000101110001011100010100001110110100000000011000100

b) 1110110101011000011001110110001110101101110110100110100001011111100111001001111110110101010110000110011100110011011001011100101111

c) 1010101110111000000000100101100011111000110000000001000100111001010111011011011110001111111100010000101001001000011110010011

d) 0000000111011110111100000000010101110111011110111010001100101000001011010000010011110110100000011011011000110001000011101000001001111011010000001101101100011000100001111

e) 1111101101000110110000000010101101011000111000000101010110000011010111000110110001111101100010111111001000100101

8 Spirals

Now, each of the 6 robots flies directly towards the robot nearest to it clockwise, at a speed of one unit per second, where a unit is the side length of the hexagon.

Assuming the size of the robots is negligible, how long (in seconds) will it take for the robots to collide?

9 Paths

Suppose we have six dots arranged in a line. Find the number of paths through some or all (but not none) of these points, subject to the limitations:

a) No point can be visited twice,

b) If a point obstructs another point, the obstructing point should be visited first.

(e.g. If the points are numbered 1-6 from left to right, then the paths "1 - 2 - 3", "2 - 3 - 1", "3 - 4 - 5 - 2" are permitted, but "1 - 3", "3 - 4 - 1" are not.)

10 Binary Numbers

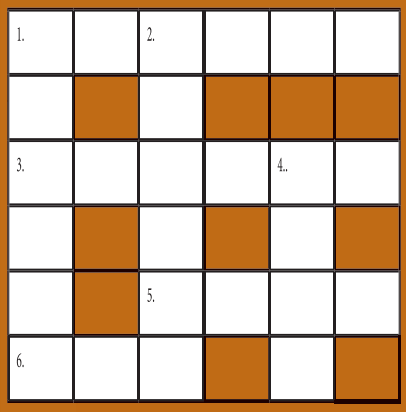
Find (in decimal) the smallest natural number n such that n is odd, the number of 1s in its binary expansion, n_1, is even, the number of 1s (n_2) in n_1's binary expansion is odd, the number of 1s (n_3) in n_2's binary expansion is even, and the number of 1s in n_3's binary expansion is odd.

11 Probabilty

You have volunteered to undergo the following experiment: On Sunday you will be put to sleep. Once or twice, during the experiment, you will be wakened, interviewed, and put back to sleep with an amnesia-inducing drug that makes you forget that awakening. A fair coin will be tossed to determine which experimental procedure to undertake: if the coin comes up heads, you will be wakened and interviewed on Monday only. If the coin comes up tails, you will be wakened and interviewed on Monday and Tuesday. In either case, you will be wakened on Wednesday without interview and the experiment ends.

You have been woken up and are being interviewed. What is the probabilty that the coin landed heads?

12 Time for a Crossword



Across

1 This answers everything

3 $6 \times 10 - 1$ written backwards

5 An abundant number

6 Size of smallest non abelian group

Down

1 A triangle number

2 Sum of two squares in two ways

4 $a^b + 1 = n = b^a$

13 Some Cake

What's the maximum number of pieces you can make with 5 straight cuts (without rearranging the pieces) on a:

a) 2D circular cake?

b) 2D crescent cake?

c) 3D spherical cake?

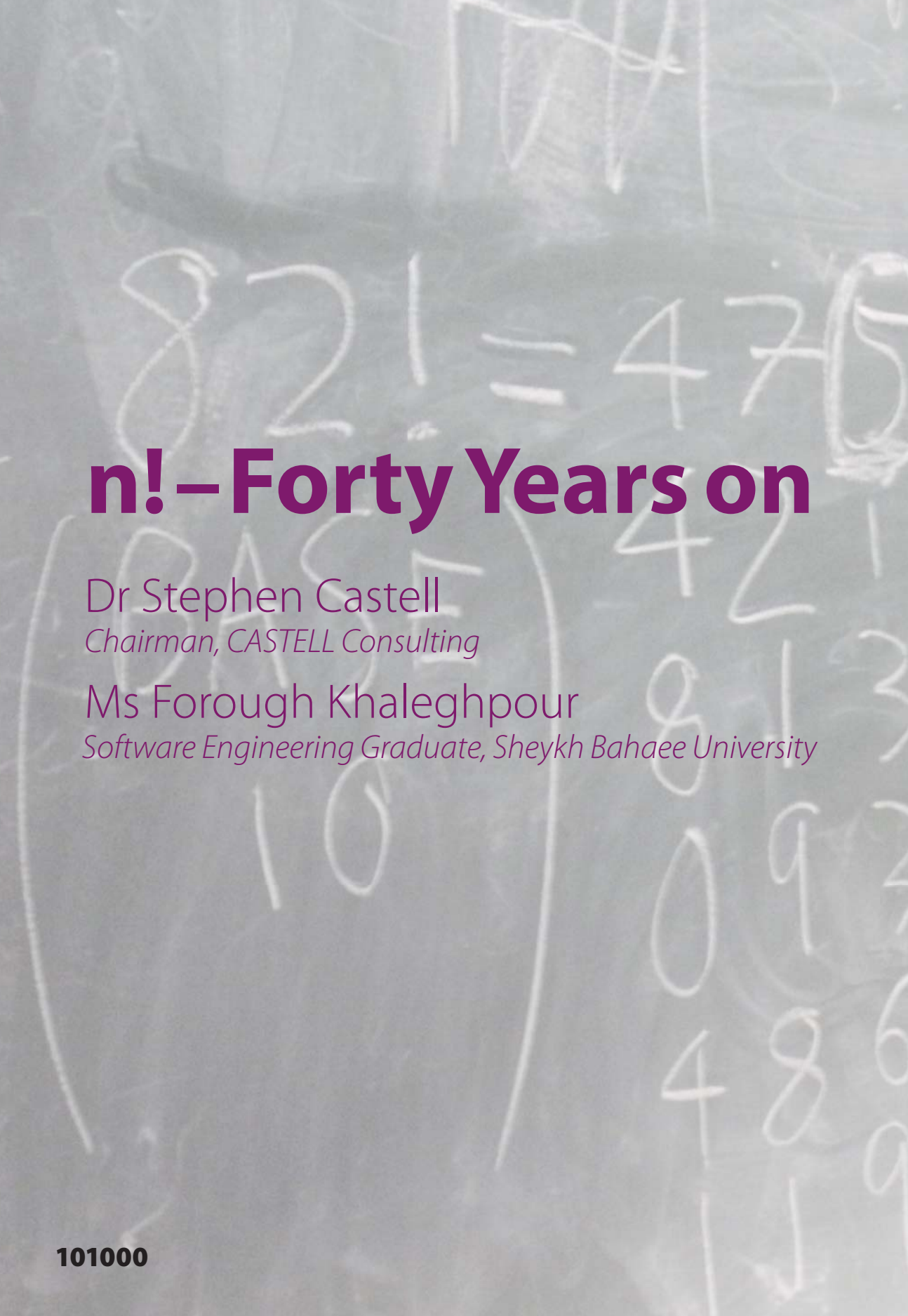
14 And Some More Cake

There are 9 plates in a line. The four plates on the left contain chocolate cakes. The four plates on the right contain banana cakes. There is an empty plate in the middle.

A move consists of either moving any cake onto an adjacent empty plate, or moving a cake past one other cake onto an empty plate.

What's the minimum possible number of moves to reverse the initial arrangement of the cakes?

Not all of the questions from the 2013 Problems Drive are contained here. Some of the problems here were not present in the problems drive, or are paraphrased. The intersection between these two sets, however, is indeed non-empty! Solutions can be found on page 1110110.



n! – Forty Years on

Dr Stephen Castell
Chairman, CASTELL Consulting

Ms Forough Khaleghpour
Software Engineering Graduate, Sheykh Bahaee University

The Original Conjecture

In an article published in Eureka 36 (October 1973), the unique factorial discovery that there is only one digit 5 in 82! was reported (see [1]). The following was also conjectured:

Conjecture (Castell, 1973)

Let y be the percentage of zeroes in the $n!$ number string S_n, x the percentage of any other digit. For large n, x and y satisfy: $y \simeq 2x$; $(r-1)x + y \simeq 100$, where r is the base of the number system.

For example, for decimal factorials ($r = 10$) this conjecture gave: $x = 9.09\%$ (1/11) and $y = 18.18\%$ (2/11). As was reported in the 1973 Eureka paper, this result did reasonably accurately apply for factorials up to 800!, where $n = 800$ was the computational limit imposed by the technology available at that time.

Castell's Original (1973) Conjecture was derived essentially using the following intuitive reasoning: the string *Sn* is in total not a random number, and there is a definite 'extra doubling' effect of the occurrences of 'trailing zeroes' whenever a multiplication by a 5 or a 10 occurs (for $r = 10$); however, intuitively, all digits other than a zero in *Sn* are equally probable. The computational data available at the time seemed to indicate that zeroes did indeed occur twice as often as any other of the digits 1 – 9.

New Developments

The following new standout results have recently been discovered by the present authors:

1. There is no digit B (i.e. '11' in decimal) in 75! (base 16), a 160 digit number.

2. The original conjecture does not fit well with data for base 16 or base 2, perhaps due to the number of factors of each base. In addition, for larger factorials than were computationally available in 1973, the original conjecture is inaccurate for base 10.

3. A new conjecture:

Conjecture (Castell and Khaleghpour, 2010)

Let y be the percentage of zeroes in the $n!$ number string S_n, x the percentage of any other digit. For large n, x and y satisfy: $y \simeq (0.075r + 1.05)x$; $(r-1)x + y \simeq 100$, where r is the base of the number system.

This gives the following table, a very good fit to the $n!$ data we have obtained:

Base r	x	y
2	54.55	45.45
10	16.66	9.26
16	13.00	5.80

A New Challenge

We challenge the interested reader to prove or disprove the above revised conjecture. In particular, we are interested in whether the non-zero digits, for large n, do generally occur in equal proportion within each factorial number string (see [2] for contrast, which considers the pattern of leading digits of factorials). It may additionally be of interest for the reader to consider our discovered 'anomalous' factorials (82!, base 10 – only one digit 5; and 75!, base 16 – no digit B) from a formal 'statistical likelihood' perspective; and also to investigate whether there are any further factorials that stand out for any reason.

References and Notes

[1] Stephen P. Castell; 1973; *On the Distribution of Decimal Digits in n!*; Eureka, 36, 45-47; http://www.archim.org.uk/archives/eureka/#36.
(This discovery that there is only one digit 5 in 82! was subsequently designated by Computer Bulletin as being 'the most useless fact discovered by a computer'.)
[2] John D. Cook; *Leading digits of factorials*; http://www.johndcook.com/blog/2011/10/19/leading-digits-of-factorials/.

The authors are happy to be contacted with any thoughts on this topic at, respectively:
cstll01@attglobal.net and f.khp01@gmail.com.


Figure 3 "Ecce Homo" (left) and "Ecce Mono" (right)

(a) Mask for restoration

(b) Initialisation of the algorithm with random colours

(c) Restored image with local inpainting

(d) Restored image with global inpainting

Figure 4 Mathematical image restoration of "Ecce homo". (d) is courtesy of Rob Hocking, using the algorithm described in [1] and [2].

Image Restoration

Dr Carola-Bibiane Schönlieb
Lecturer in Applied and Computational Mathematics, DAMTP

In our modern society we encounter digital images in a lot of different situations: from everyday life, where analogue cameras have long been replaced by digital ones, to their professional use in medicine, earth sciences, arts, and security applications. The images produced in these situations usually have to be organized and possibly postprocessed. The organization and processing of digital images is known under the name of image processing or computer vision.

We often have to deal with the processing of images, e.g., the restoration of images corrupted by noise, blur, or intentional scratching. The idea behind image processing is to provide methods that improve the quality of these images by postprocessing them. For an introduction to digital image processing we refer to [3] or [8]. Virtual image restoration or image interpolation (also referred to as "inpainting") denotes the methodology whereby missing parts of damaged images are filled in, based on the information obtained from the intact part of the image and a priori assumptions made on the missing image structures. Virtual image restoration is an important challenge in our modern computerized society: From the reconstruction of crucial information in satellite images of our earth to the renovation of digital photographs and ancient artwork, virtual image restoration is ubiquitous. Considering this huge – but by no means complete – range of image processing applications and the fact that there are still problems in this area which have not been satisfactorily solved, it is not surprising that this is a very active and broad field of research. From mathematicians, to engineers and computer scientists, a large group of people have been and are still working in this area.

The Digital Image: a Mathematical Object?

In order to appreciate the following theory and the image processing applications, we first need to understand what a digital image really is. Roughly speaking, a digital image is obtained from an analogue image (representing the continuous world) by sampling and quantization. Basically this means that the digital camera superimposes a regular grid on an analogue image and assigns a value, e.g., the mean brightness in this field, to each grid element. In the terminology of digital images these grid elements are called pixels. The image content is then described by grey values or colour values prescribed in each pixel. The grey values are scalar values ranging between 0 (black) and 255 (white). The colour values are vector values, e.g., (r, g, b), where each channel r, g and b represents the red, green, and blue component of the colour and ranges, as the grey values, from 0 to 255.

The mathematical representation of a digital image is a so-called image function u defined on a two dimensional (in general rectangular) image domain, the grid. Indeed, in some applications, images are three dimensional (e.g. videos, 3D medical imaging) or even four dimensional (involving three spatial dimensions and time) objects, but for simplicity we focus on the

two dimensional case for the following conceptual presentation. The image function is either scalar valued in the case of a grey value image, or vector valued in the case of a colour image. Here the function value $u(x, y)$ denotes the grey value, i.e., colourvalue, of the image in the pixel (x, y) of the image domain. Figure 1 visualizes the connection between the digital image and its image function.

Typical sizes of digital images range from 2000×2000 pixels in images taken with a simple digital camera, to $10\,000 \times 10\,000$ pixels in images taken with high-resolution cameras used by professional photographers. The size of images in medical imaging applications depends on the task at hand. PET for example produces three dimensional image data, where a full-length body scan has a typical size of $175 \times 175 \times 500$ pixels.

Now, since the image function is a mathematical object we can treat it as such and apply mathematical operations to it. These mathematical operations are summarized by the term *image processing techniques*, and range from statistical methods, morphological operations, to solving a partial differential equation for the image function. We are especially interested in the last, i.e., PDE – and variational methods used in virtual restoration.

We have introduced a *digital* image as a sampled and quantised version of an *analogue* (also called physical or real) image. The higher the resolution of a digital image, the closer it is to the analogue image in the real-world. While digital image processing is indeed concerned with digital images, the methods used are often motivated from considerations in the continuum, that is methods are formulated for the analogue image. In this article we take up this mathematically more challenging and analytically more beautiful position, and let our image u be a continuous object defined on a rectangular domain $\Omega = (a, b) \times (c, d)$. Within this framework, there are many possibilities for how images can be modelled, compare [8], Chapter 3. For our purposes we will focus on the representation of images as elements in a function space such as the Lebesgue space $L^2(\Omega)$, Sobolev spaces such as $H^1(\Omega)$ and the space of functions of bounded variation $BV(\Omega)$. The latter space is especially suited for images since an element in BV can be discontinuous and hence the representation of image edges is possible.

Local and Global Features: What is Important in Image Restoration?

An important task in image processing is the process of filling in missing parts of damaged or occluded images based on the information obtained from the intact parts in the image. It is essentially a type of interpolation and we will refer to it as virtual image restoration or inpainting (various terminologies are used for image interpolation depending on the application).

Let f represent some given image defined on an image domain Ω. Loosely speaking, the problem is to reconstruct the original image u in the (damaged) domain D of Ω, called inpainting domain or a hole/gap (cf. Figure 2).

Virtual image restoration methods can be roughly divided into two groups: 1) local inpainting, and 2) global inpainting methods. The main difference between these two classes lies in the type of image information used from the intact part of the image, as well as the different kind of inpainting processes with which this information is propagated into the missing domain.

A method is local if the information used to fill in D is only taken from of the boundary ∂D (or a small neighbourhood of D). In a local inpainting method the restored image u can be formalised as a solution of either a variational problem or a partial differential equation (PDE). The easiest example is harmonic inpainting, where:

$$u \in \operatorname{argmin}_v \left\{ \|\nabla v\|^2_{L^2(\Omega)} : \text{ such that } v = f \text{ on } \Omega \setminus D \right\} \quad (1)$$

Equivalently, the first-order optimality condition for the above variational problem (Euler-Lagrange equation) states that the restored image u solves the Laplace equation

$$\begin{cases} \Delta u = 0 & \text{in } D \\ u = f & \text{on } \partial D. \end{cases} \quad (2)$$

As such, u can be seen as the harmonic extension of f from ∂D into D. Of course, any image structures such as image edges are not preserved by the harmonic extension (rather diffused into D). More sophisticated local inpainting methods have been proposed in the community during the last fifteen years that are able to propagate geometric image information such as object edges,

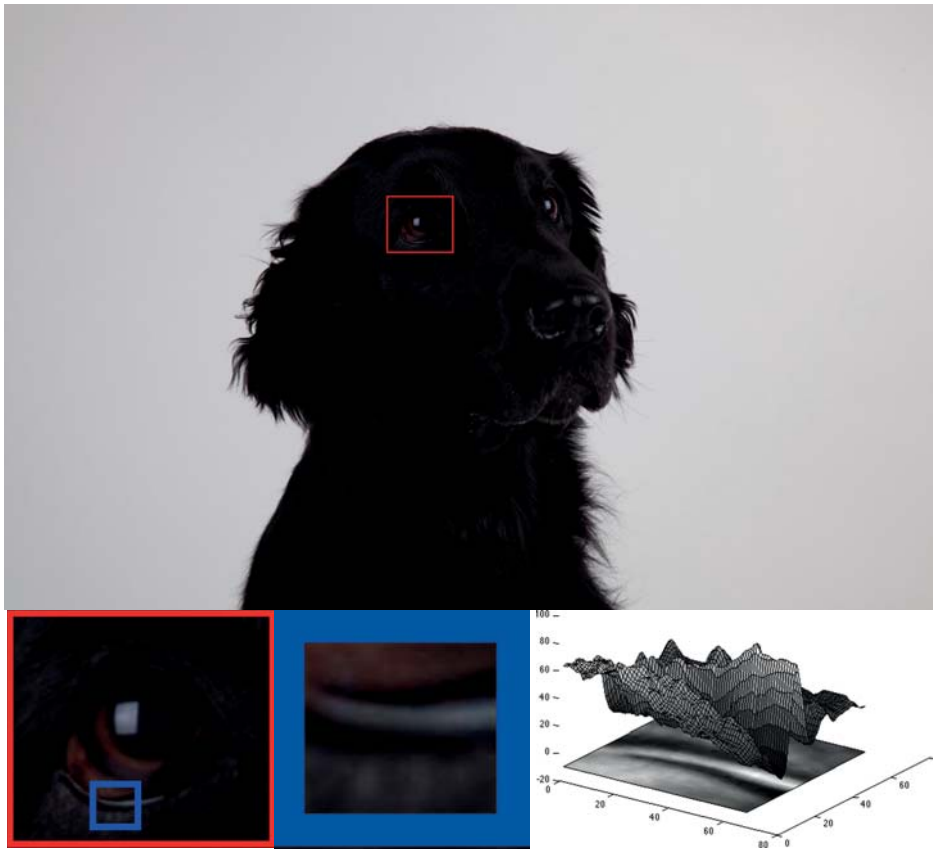


Figure 1 Digital image versus image function: Gradually zooming in to the level where the image pixels are visible (blue framed detail), the image function of the red channel $u(x, y, r)$ of the digital photograph is plotted as the height (the value for red) over the (x, y)-plane.

their orientation and curvature. These approaches are mainly based on extensions of (1) and (2) to non-smooth variational problems and nonlinear (and often higher-order) PDEs, respectively. For different types of local inpainting methods we refer the reader for instance to [4], [5], [12] (inpainting via transport), [14], [6] (TV inpainting), [9] (curvature driven diffusion inpainting), [11] (Mumford-Shah based inpainting) and [13], [15] (Euler's elastica inpainting).

What all of these methods have in common is that they can only reproduce local image features (encoded in the value of the image function and its derivatives in a pixel) but are not able to pick up image patterns or texture which are non-local image features. Global (or non-local) inpainting methods take into account all the information from the known part of the image, usually weighted by its distance and similarity (measured in a certain way) to a neighbourhood of the point that is to be filled in. Such methods usually work on image patches rather than on single image pixels. They are mathematically formalised as non-local variational problems and engineering-type discrete algorithms. This class of methods is very powerful, able to fill in structures and textures almost perfectly. However, they still have some disadvantages. One major one is the high computational cost involved in their solution. For some of these methods another disadvantage is their dependence on an initial guess for the restored image in D. Local methods are sometimes more desirable, especially when the inpainting domain is relatively small. If D is large a local method can serve as a good initialisation for the global inpainting method. For more discussion on global methods the reader is referred to [7], [10], [1], [2].

Mathematical Algorithms Versus an Amateur's Attempt

In August 2012 Cecilia Giménez, an eighty year old amateur artist from a small village near Zaragoza (Spain) gained fame by an attempt to restorate a wall painting in a local church. She produced the by now famous painting dubbed "Ecce Mono" (Behold the Monkey) when aiming to restore the wall painting "Ecce Homo" (Behold the Man) by the spanish painter Elías García Martínez (see the comparison in Figure 3).

Let's see what virtual image restoration methods make of this. In Figure 4 a local and a global inpainting result for the head of the Jesus figure are shown. For the local inpainting we used higher-order total variation inpainting (see [6]) and for the global inpainting method a variational exemplar-based method with the L^1-norm as similarity measure between image patches (see [2]). Local inpainting is doing pretty well in recovering the main structures in the painting but smoothing out small-scale features and texture. Being initialised with the local inpainting result, the global inpainting method performs reasonably well. We leave it to the reader to decide which restoration is more realistic: Cecilia's "Ecce mono" in Figure 3 or the mathematically formalised inpainting in Figure 4.

Conclusion

Mathematical concepts such as nonlinear PDEs and variational calculus offer a beautiful and rich framework for formalising and solving real world problems in imaging. Of course, virtual image restoration cannot (yet or never?) replace human expertise. In fact, virtual image restoration algorithms have been very much influenced by the experience and guidelines of art restorers, aiming to formalise what art restorers do mathematically (see [4]). Virtual image restoration can also help art restorers by producing digital templates for damaged art pieces.

Acknowledgements

This article is dedicated to Vicent Caselles (August 10, 1960 – August 14, 2013) who was one of the founding fathers and research guides of virtual image restoration and of mathematical imaging in general (see [17]). My sincerest thanks go to Sue Hickman Pinder from the Millennium Mathematics Project and to Rebecca Paul for proofreading.

References

Due to the large number of references, they have been listed on the Archimedeans' website at archim.org.uk/Eureka/Supplements.

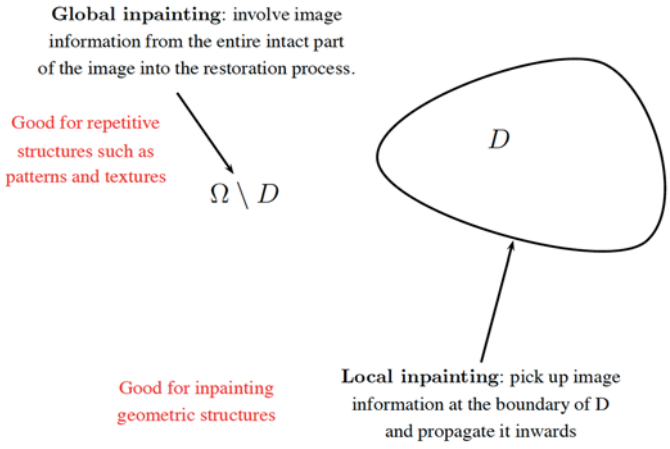
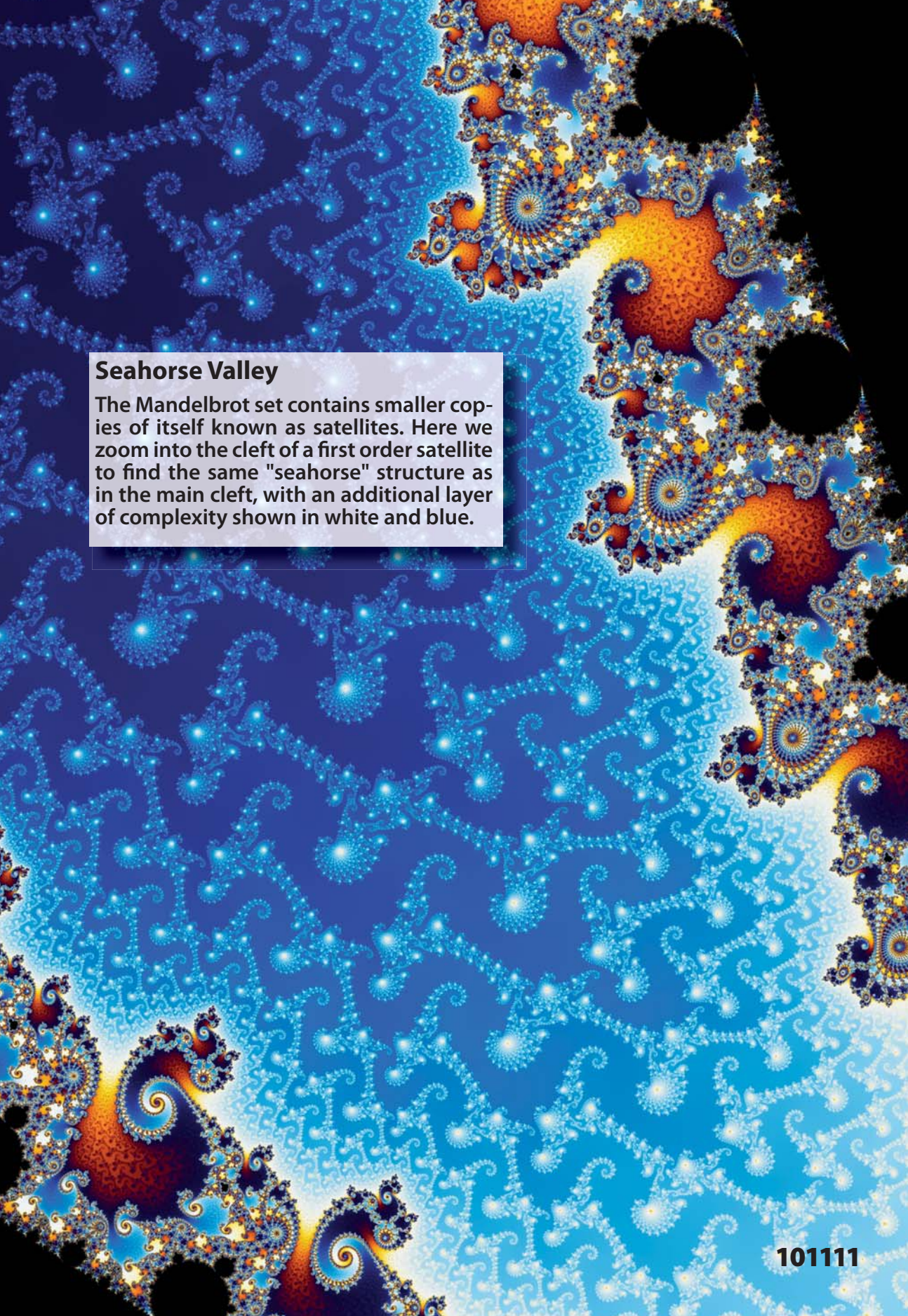


Figure 2 Virtual image restoration: based on the intact image information f inside $\Omega\backslash D$, one seeks the inpainted image u that extends f into the inpainting domain D. The difference between local and global inpainting lies in its conceptually different method of recovering u from f.



Seahorse Valley

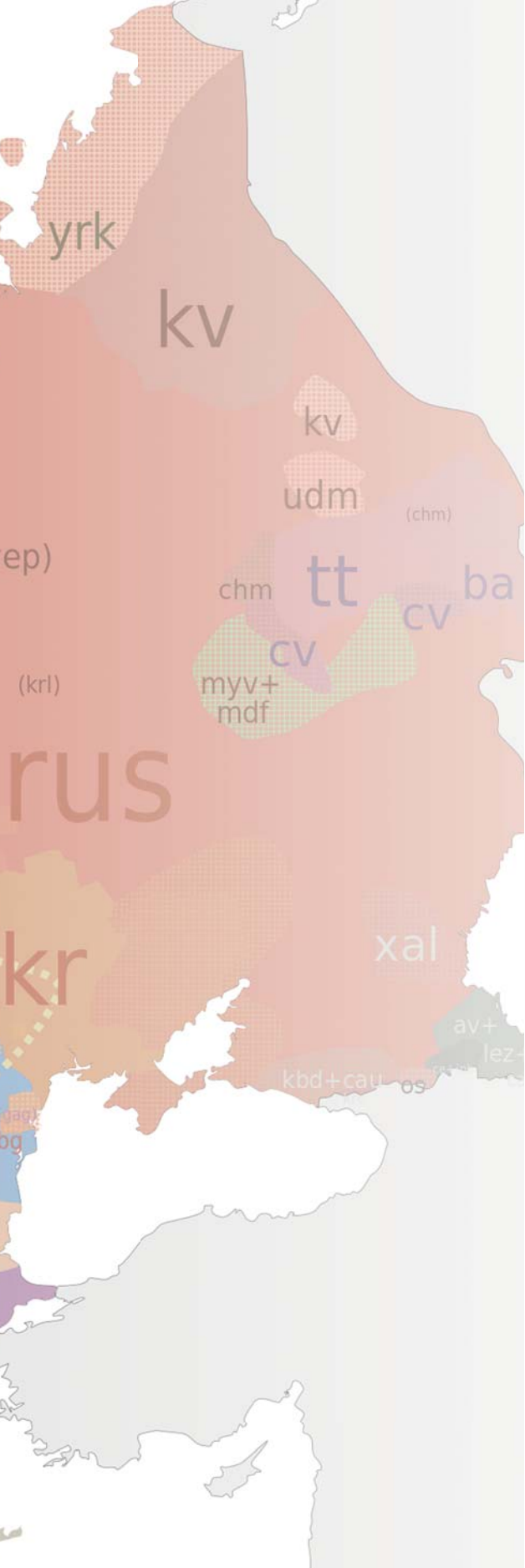
The Mandelbrot set contains smaller copies of itself known as satellites. Here we zoom into the cleft of a first order satellite to find the same "seahorse" structure as in the main cleft, with an additional layer of complexity shown in white and blue.

Statistically Speaking

Prof John Aston

Professor of Statistics, Statslab

110000



It might seem somewhat implausible that functional analysis, non-Euclidean geometry and statistical shape analysis have much to tell us about the spread of European languages. Historical linguistics has traditionally been something of a qualitative discipline, but recently there has been considerable interest in taking a more quantitative approach to the subject, through textual analysis but also through the analysis of acoustic recordings. It is this latter data that has allowed some more unusual links between mathematics and phonetics to be made.

Acoustic recordings yield considerable quantities of data which to all intents and purposes can be seen as continuous over time. For example, in Figure 1 below, a two dimensional surface (spectrogram) can be seen, where the first axis represents time while the second represents the frequency of the sound wave being recorded. This spectrogram not only conveys all the time and frequency information contained in the word being said, but can be treated (when suitably normalised) as a random element, say X, $X \in L^2$.

Functional Data Analysis

The relatively new field of functional data analysis (FDA) (see [2], [4]) is something of a cross between functional analysis and classical statistics. Unlike the usual univariate or multivariate analysis undertaken in most statistics, FDA is the branch of statistics that concerns data where the fundamental unit of that data is a function in some suitable (often infinite dimensional Hilbert) space. The main idea is to use the properties of smoothness and regularity in the functions to allow statistical analysis to be carried out, even though the functions are only ever discretely observed with noise.

One of the most important quantities in FDA is

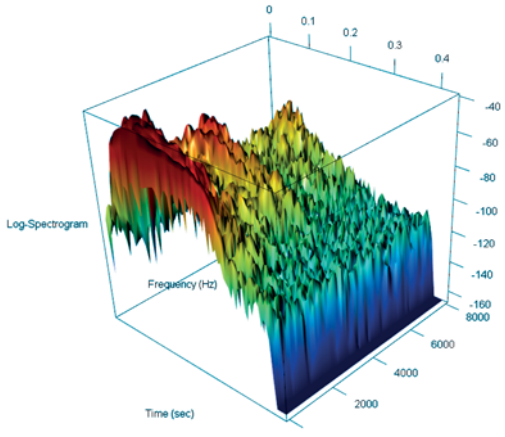


Figure 1 An example of a raw spectrogram (in logarithmic scale) as obtained by taking a windowed discrete fourier transform of a 22kHz sound sample of a single syllable (the word one ("un") in French). The fourier transform was computed every 10 ms to yield the discretised version of the function.

the covariance operator. For a random square integrable function X, with $\mathbb{E}(X)=0$, the operator $C(y)=\mathbb{E}(\langle X,y\rangle X), y\in L^2$ is defined as the covariance operator, where $\langle\cdot,\cdot\rangle$ is the usual inner product in L^2 (see [2] for more details). It is, by definition, non-negative definite, and in many data analysis situations is assumed to be a trace class operator, i.e. $\Sigma_j\lambda_j<\infty$, where λ_j are the eigenvalues of the spectral decomposition of the operator. In many situations, FDA proceeds by using one of the fundamental theorems of Stochastic Processes, the Karhunen-Loeve decomposition of the operator, to provide a basis for expansion of the data. This allows a possible dimension reduction on the data to be performed, something that has been common in multivariate statistics since the early 20th century, in the finite dimensional setting. However, in the case of examining differences between languages, it is the operator itself which will be of interest.

Assume that we are interested in understanding the relationship between languages through their acoustic properties. Given a set of recordings for a particular language, spectrograms can be produced and an estimate of the associated covariance operators obtained. Languages, of course, have many characteristics, but it has been shown that one characteristic of interest is the variational patterns that are present in the sounds. These differences are captured exactly by the covariance operators. Therefore by comparing covariance operators we can provide one particular comparison of the languages themselves.

Statistics in Non-Euclidean Spaces

However, covariance operators are not the usual type of data that statistical analysis is designed for. They are non-negative definite trace class operators, so do not lie in a standard "Euclidean" space. The usual Euclidean metrics used in statistical analysis, extended to FDA, are not valid given the restricted space. This requires a new type of metric to be used, one with its roots in statistical shape analysis (see [1]), where non-Euclidean geometry is commonplace.

Let us start by considering a closely related finite dimensional problem, defining a distance between two positive definite matrices. Possibly, the simplest approach to take would be to take the matrix logarithm and compute the usual Frobenius norm between the matrix logarithms. This is indeed a Riemannian distance on the space of positive definite matrices, and as such allows statistical analysis to be developed. However, even if our covariance operators were positive definite, their trace class nature implies that their eigenvalues tend to zero, and hence the equivalent of the matrix logarithm is unbounded. However, this is not the case for all transformations. For example, the square-root transformation is well defined and the resulting operator, while not guaranteed to be trace-class, is still a Hilbert-Schmidt operator, and as such the distances are still well defined.

The square-root of a matrix, or operator, is, however, not uniquely defined. It would be somewhat more elegant if the distance between two languages was independent of the choice of square-root. This is a well studied problem in statistical shape analysis, where the equivalent problem is that of how to match shapes that are subject to rotation and translation. The shape of dog is still a dog, whether it is standing with its head to the left or to the right. Equivalently the uniqueness of the square-root is defined up to its rotation, and as such by quotienting out the rotation group we obtain a unique distance. These ideas yield the following metric to measure the distances between our languages. For two covariances C_1 and C_2, the Procrustes metric is defined as

$$d_P(C_1,C_2)^2 = \inf_{R\in O\{L^2(\Omega)\}} \| L_1 - L_2R \|^2_{\mathrm{HS}}$$
$$= \inf_{R\in O\{L^2(\Omega)\}} \mathrm{tr}\{(L_1-L_2R)^*(L_1-L_2R)\},$$

where L_i are such that $C_i = L_i L_i^*$, for $i = 1, 2$, and $O\{L^2(\Omega)\}$ is the space of unitary operators on L^2. Procrustes was the Greek innkeeper of myth who fitted everyone to his iron bed by either stretching or chopping them to size, and as such this metric equivalently gives a distance that disregards the orientations of the initial estimates of the covariance operator. This distance, although somewhat complex, has a simple closed form solution, where, for any choice of L_i satisfying the above,

$$d_P(C_1, C_2)^2 = \|L_1\|_{\text{HS}}^2 + \|L_2\|_{\text{HS}}^2 - 2\sum_{k=1}^{\infty}\sigma_k.$$

where σ_k are the singular values of the compact operator $L_2^* L_1$. It can be shown that, even when there are only finite amounts of discretised data present, the estimates of the distance converge asymptotically.

Investigating the Relationships in Romance Languages

The statistical analysis of non-Euclidean and functional data are of interest in and of themselves, and are some of the fastest growing areas of modern statistics. However, this is in many ways because of their ability to be used to give insights into other areas such as historical linguistics. In a recent study, recordings of the pronunciation of the numbers one to ten were taken from four different romance languages (French, Italian, Spanish and Portuguese) with one language having two different dialects present (Iberian Spanish and American Spanish). 219 spectrograms (there were several repetitions of each word in each language) were generated from the sound samples, and preprocessed to form aligned functions from which covariances were formed. The distances between these covariances were then examined.

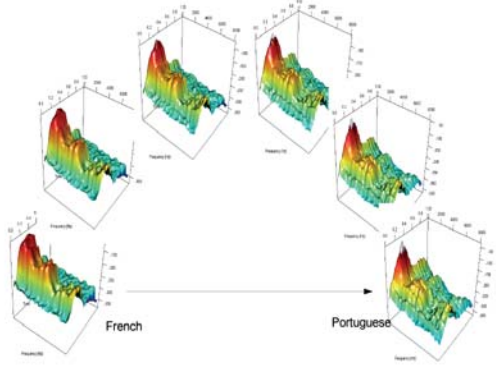


Figure 2 Representation of the geodesic taking a speaker saying the French word "un" and turning it into the Portuguese word "um". The geodesic is based on the Procrustes metric in the space of covariance functions.

It is possible to use the Procrustes metric to not only define distances between covariances but also by extension to define geodesics within the space of covariance functions (see [3]). These can then be used to define covariances for languages "between" any two of the observed languages or even to predict how one speaker might sound when speaking another language. Figure 2 shows one such predicted path. Here a speaker saying the word "un" (one in French) is mutated along a geodesic path into saying the word "um" (one in Portuguese). The speaker characteristics are retained but the variations attributed to the languages are captured via the geodesic path. These spectrograms can then be transformed back into audio to hear the results. This opens up a world of possibilities of discovering how one language might be related to another, or how historical language groups might have evolved into modern day languages.

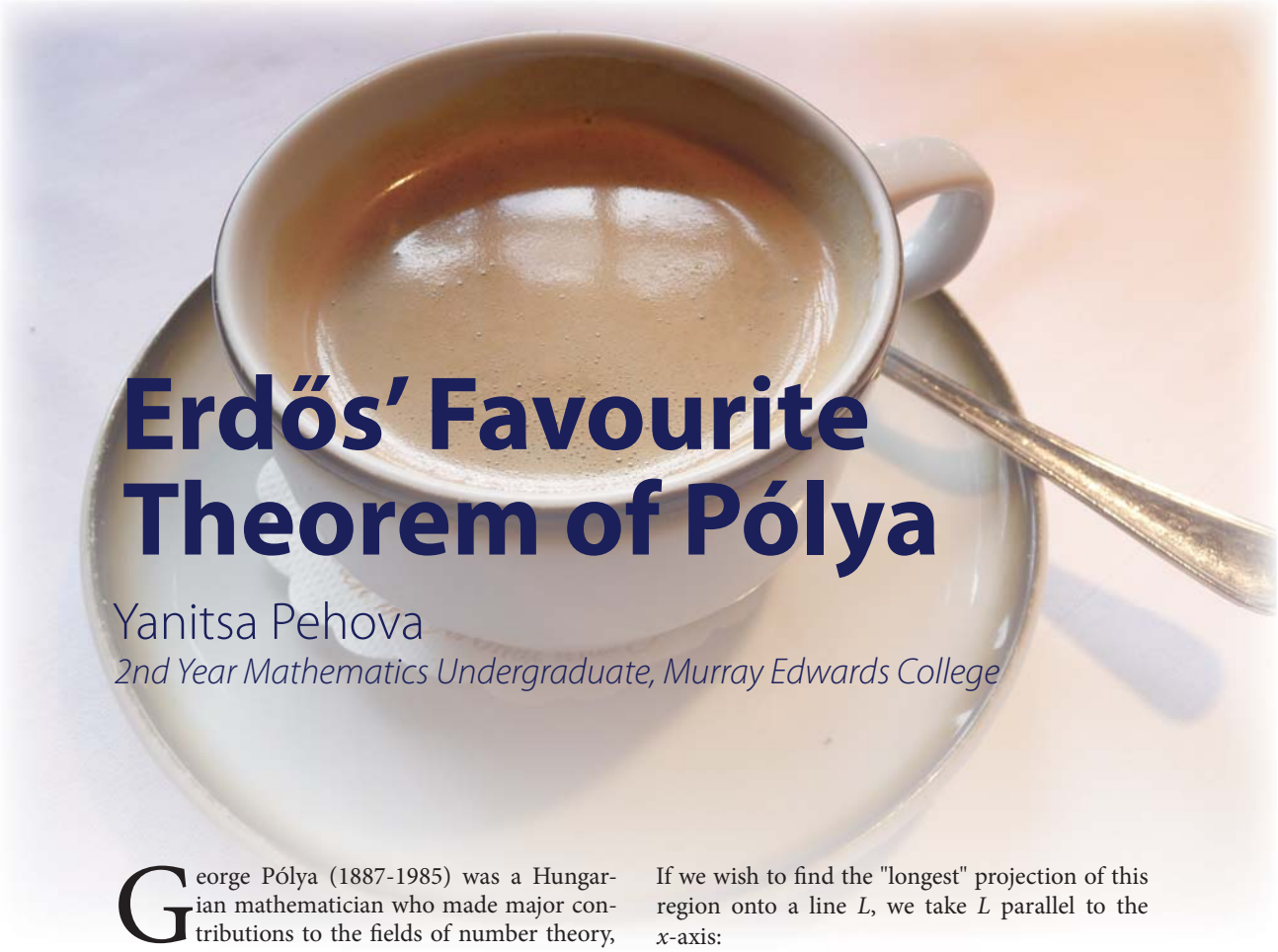
The integration of concepts from geometry, analysis and other areas of mathematics into data analysis through statistics has a long history. However, modern data sources are constantly raising new challenges and areas such as non-Euclidean FDA are being developed for applications as diverse as brain imaging to those seen here in linguistics.

Acknowledgements

Considerable thanks are due to my collaborators John Coleman (Oxford), Ian Dryden (Nottingham), Pantelis Hadjipantelis (Warwick), Davide Pigoli (Cambridge) and Piercesare Secchi (Milan), for all their work upon which this article is based.

References

[1] Ian L. Dryden, Kanti V. Mardia; 1998; *Statistical Shape Analysis*; Wiley; Chichester.
[2] Horvath L, Kokoszka P; 2012; *Inference for Functional Data with Applications*; Springer, New York.
[3] Davide Pigoli, John A.D. Aston, Ian L. Dryden, Piercesare Secchi; 2014; *Distances and inference for covariance operators*; Biometrika; 101:409-422.
[4] James O Ramsay, Bernard W. Silverman; 2005; *Functional Data Analysis (2nd Ed)*; Springer, New York.



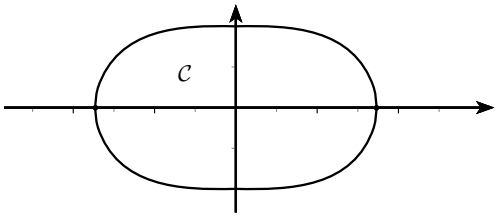
Erdős' Favourite Theorem of Pólya

Yanitsa Pehova
2nd Year Mathematics Undergraduate, Murray Edwards College

George Pólya (1887-1985) was a Hungarian mathematician who made major contributions to the fields of number theory, combinatorics, probability, analysis and many more (see [3]). The following theorem is said to have been Erdős's favourite result proved by Pólya:

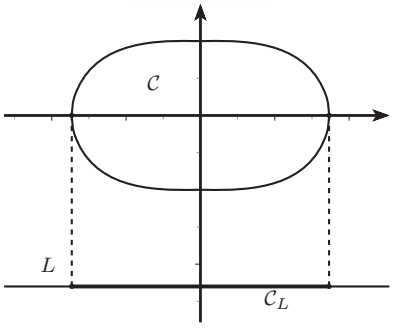
Theorem 1 Let $f(z) = z^n + a_{n-1}z^{n-1} + \ldots + a_0$ be a complex monic polynomial of degree n and let $C = \{z \in \mathbb{C} : |f(z)| \leq 2\}$ be the set of points mapped within the circle of radius 2 with centre the origin. Then the total length of the projection C_L of $\mathbb{C}$ onto any line L in the complex plane never exceeds 4.

Let's look at an example. Take the complex function $f(z) = z^2 - 1$. We solve $|f(z)| \leq 2$ to obtain the following region of the complex plane:



The boundary of the region has equation $y^2 = 2\sqrt{1+x^2} - 1 - x^2$.

If we wish to find the "longest" projection of this region onto a line L, we take L parallel to the x-axis:



We then get that the length of the projection is $C_L = 2\sqrt{3} < 4$, as expected. We will now start proving this by taking simpler cases and then generalising. We will restrict ourselves to real polynomials and we will only take the projection of $\mathbb{C}$ onto the x-axis, i.e. we will prove the following theorem:

Theorem 2 Let $f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0$ be a real monic polynomial of degree n with n real roots and let $C = \{x \in \mathbb{R} : |f(x)| \leq 2\}$ be the set of points mapped within the interval $[-2, 2]$. Then

C can be covered by intervals with total length at most 4. The reader can easily see how this relates to Pólya's theorem. We will see that most of the work is in proving this special case and the transition from real to complex numbers is natural and easy.

Chebyshev's theorem

We will use the following theorem by Chebyshev (see [2], pp.14):

Theorem 3 (Chebyshev) (proof omitted) Let $P(x)$ be a real monic polynomial. Then

$$\max_{a\le x\le b} P(x) \ge \frac{1}{2^{n-1}},$$

where n is the degree of $P(x)$.

This is a fact that seems completely unrelated to Pólya's theorem but it leads us to the following corollary:

Corollary If $|P(x)| \le 2$ for all $x \in [a, b]$, then $b - a \le 4$.

Proof of Corollary This can be obtained by mapping the interval $[a, b]$ onto $[-1, 1]$ by substituting $y = \frac{2}{b-a}(x-a) - 1$ and applying Chebyshev's theorem to $P(x)$ as a polynomial $Q(y)$ in y (Q isn't monic but we can scale it to make it monic). Then since

$$\max_{a\le x\le b} P(x) = \max_{-1\le y\le 1} Q(y)$$

we obtain

$$2 \ge \max_{a\le x\le b} P(x) \ge \underbrace{\left(\frac{2}{b-a}\right)^n}_{\text{scaling}} \frac{1}{2^{n-1}}$$

which yields $b - a \le 4$, as required.

Now this looks more like it is going somewhere towards our goal! It shows that Theorem 2 is true if C_1 is an interval (instead of the more general union of disjoint intervals $I_1 \cup I_2 \cup \ldots \cup I_k = [a_1, b_1] \cup [a_2, b_2] \cup \ldots [a_k, b_k]$, with total length $l(I_1) + \ldots l(I_k)$).

Pólya's Idea

Now what Pólya did was to try and construct another real monic polynomial $\hat{P}(x)$ of degree n such that the projection onto the x-axis is an interval of length at least $l(I_1) + \ldots + l(I_k)$. Here is how we do this: We have a polynomial $P(x) = (x - x_1)(x - x_2) \ldots (x - x_n)$ with $C = \{x \in \mathbb{R} : |P(x)| \le 2\} = I_1 \cup I_2 \cup \ldots I_k$, where the intervals are arranged in ascending order. After some elementary (but not necessarily easy) observations, we claim that the endpoints of the intervals have functional values +2 and −2 and that all of these intervals contain a root of the polynomial (see [1], pp. 141-142). We prove this by assuming $P(x) = 2$ at both endpoints and looking at the first and second derivatives at a critical point (which exists by Rolle's theorem). Then we use $p'(x)^2 \ge p(x)p''(x)$.

Now suppose I_k contains m roots of our polynomial (with their multiplicities), namely x_1', x_2', ... , x_m', and let the rest of the roots be y_1', y_2', ... , y_{n-m}'. Of course we assume that $m < n$, otherwise C_L would be a single interval and we would be done. Let d be the distance between I_k and I_{k-1}, i.e. $d = a_k - b_{k-1}$.

Since the x's and the y's are roots, we can write $P(x) = Q(x)R(x)$ where $Q(x) = (x - x_1') \ldots (x - x_m')$ and $R(x) = (x - y_1') \ldots (x - y_{n-m}')$ (we are "splitting" $P(x)$). Now we construct the following polynomial (new line because it is *important*):

$$P_1(x) = Q(x + d)R(x)$$

We can see that $P_1(x)$ has roots $x_1' - d$, $x_2' - d$, ... , $x_m' - d$, y_1', y_2', ... , y_{n-m}'. Now $C_1 = \{x : |P_1(x)| \le 2\}$ contains the intervals $I_1, I_2, \ldots, I_{k-1}$ and I_{k-d}. That's because given a point $x \in I_1 \cup \ldots \cup I_{k-1}$, we have $|x - x_i' + d| = -x + x_i' - d < -x + x_i' = |x - x_i'|$ since $x < x_i' - d$ for x in the intervals under consideration, so $|Q(x + d)| < |Q(x)|$ and then again $|P_1(x)| \le |P(x)| \le 2$. Similarly if $x \in I_k$ we look at $R(x)$: $|R(x - d)| \le |R(x)|$ and so $|P_1(x - d)| = |Q(x)||R(x - d)| \le |Q(x)||R(x)| = |\mathrm{P(x)}| \le 2$. Here the last two intervals get "glued together" forming a single interval, so now we have $P_1(x)$ with only $k - 1$ intervals. By induction we can reach the desired $\hat{P}(x)$. Note that the set of values of x such that $\hat{P}(x) \le 2$ is not the same as the set for our original polynomial, but instead has total length at least the total length of C for $P(x)$. Thus Theorem 2 is proved. QED.

To help understand the process, we

George Pólya (1887 – 1985) was a Hungarian mathematician who made fundamental contributions to a wide range of topics, including series, number theory, mathematical analysis and geometry. He is also widely know for his work in heuristic reasoning, writing many books on the subject, including the famous *How to Solve It*.

"Beauty in mathematics is seeing the truth without effort."

Paul Erdős (1913-1996) was among the most prolific mathematicians of all time, working with hundreds of collaborators. He was also known for his eccentric personality. He spent much of his life as a vagabond, often turning up at colleagues' homes and announcing that his 'brain was open', and staying for a few days to collaborate, before moving on.

"Another roof, another proof."

will look at another example. Let's take $P(x) = x^3 - 5x^2$. We have two intervals $I_1 \approx [-0.6, 0.68]$ and $I_2 \approx [4.92, 5.08]$, with $d \approx 4.24$, which satisfy $|P(x)| \leq 2$. The second interval contains the root $x = 5$ while the first contains the double root $x = 0$. Therefore we split $P(x)$ into $Q(x) = x - 5$ and $R(x) = x^2$. We then form $P_1(x) = Q(x + d)R(x) = (x + 4.24 - 5)x^2$. Now we can check that $|P_1(x)| \leq 2$ only has one interval as a solution, namely $[-1.05, 1.57]$, which contains the intervals $I_1 \approx [-0.6, 0.68]$ and $I_2 - d \approx [0.68, 0.84]$.

From Real Line to Complex Plane

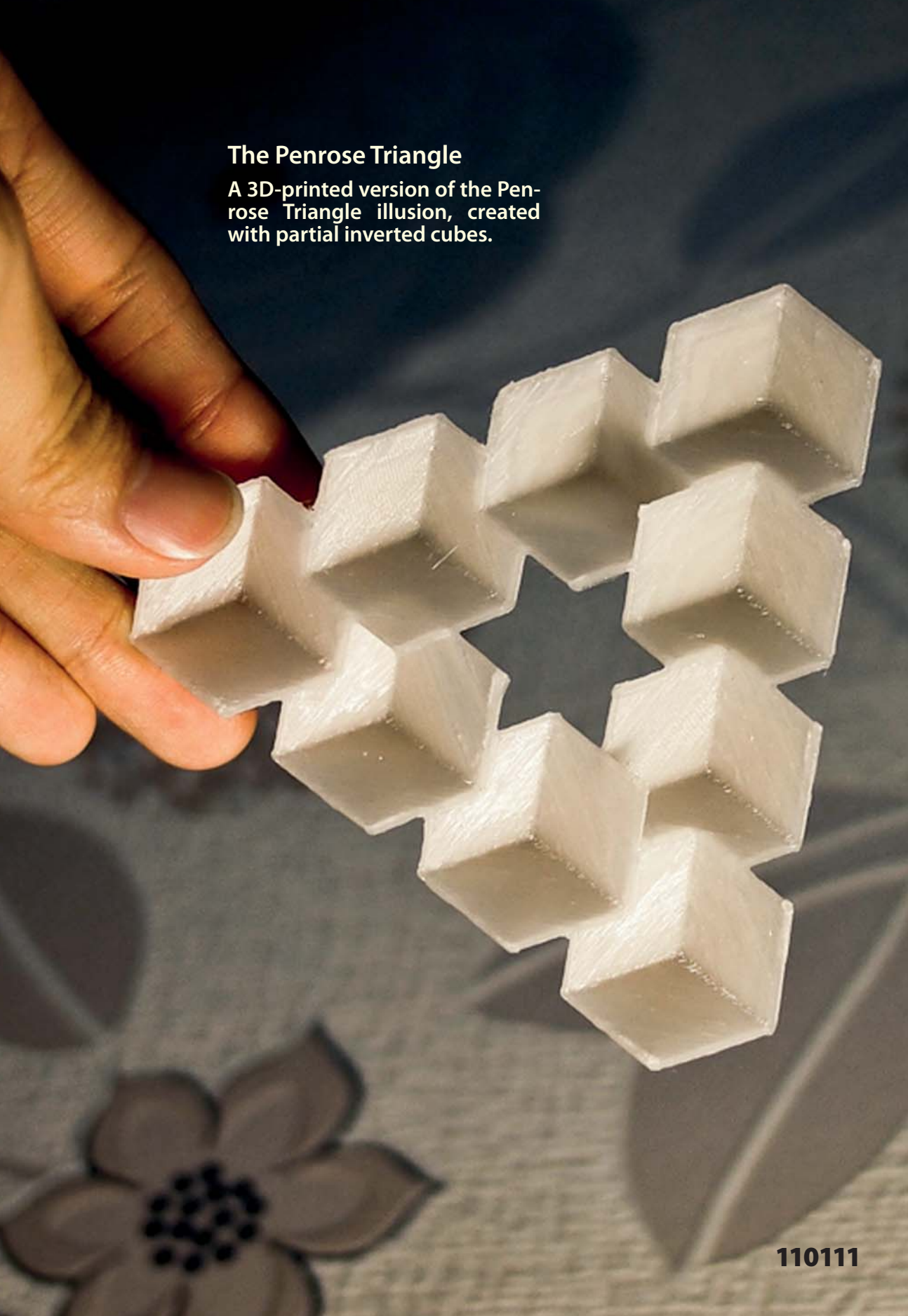
First of all, let's crack the scary every line L in the complex plane bit. It takes very little thought to figure out that we only need to prove the theorem for L being the real line and all the other lines can be obtained by rotating the plane. Thus we conclude that this part isn't as impressive as it sounds and we don't even need to worry about it. Now let's take our complex polynomial $f(z)$ from Theorem 1. Let $C_R = \{x \in \mathbb{R} : x = Re(z) \text{ and } |f(z)| \leq 2\}$. We can write $f(z) = (z - z_1)(z - z_2) \ldots (z - z_n)$ where $z_i = a_i + ib_i$ are the roots of $f(z)$. Now if we consider the "real" version $g(x) = (x - a_1)(x - a_2) \ldots (x - a_n)$ of $f(z)$, we have $|x - a_i|^2 + |y - b_i|^2 = |z - z_i|^2$ by Pythagoras' theorem. Hence $|x - a_i| \leq |z - z_i|$ for all $1 \leq i \leq n$. This means that $|g(Re(z))| \leq |f(z)| \leq 2$ for $z \in \mathbb{C}$. Or, after taking a few seconds to think and let the above sink in, $C_R \subseteq \{x \in R : |g(x)| \leq 2\}$. Thus Theorem 2 actually implies Theorem 1 (Pólya's result).

References

[1] Martin Aigner, Günter M. Ziegler; 2010; *Proofs from THE BOOK*; Springer.

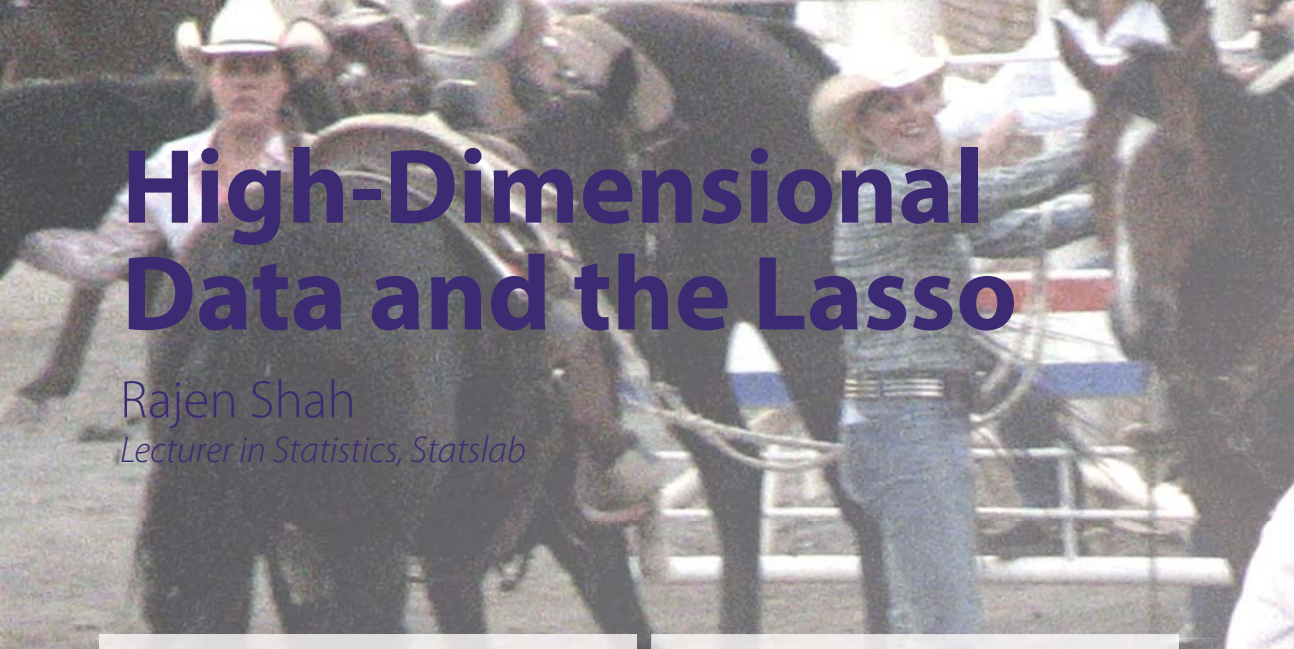
[2] Guiseppe Mastroianni, Gradimir V. Milovanović; 2008; *Interpolation Processes: Basic Theory and Applications*; Springer.

[3] Gerald L. Alexanderson; 2000; *The Random Walks of George Pólya*; The Mathematical Association of America.



The Penrose Triangle

A 3D-printed version of the Penrose Triangle illusion, created with partial inverted cubes.



High-Dimensional Data and the Lasso

Rajen Shah
Lecturer in Statistics, Statslab

How would you try to solve a linear system of equations with more unknowns than equations? Of course, there are infinitely many solutions, and yet this is the sort of problem statisticians face with many modern datasets, arising in genetics, imaging, finance and many other fields. What's worse, our equations are often corrupted by noisy measurements! In this article we will introduce a statistical method that has been at the centre of the huge amount of research that has gone into solving these problems. We'll begin by reviewing the classical version of the problems, before moving on to the more modern setting hinted at above.

Regression Analysis

Imagine data are available in the form of observations $(Y_i, x_i) \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$, and the aim is to infer a simple *regression function* relating the average value of a *response*, Y_i, and a collection of *predictors* or *variables*, x_i. This is an example of regression analysis, one of the most important tasks in statistics.

Often, we may assume that the unknown regression function is linear in the predictors, giving the following mathematical formulation of the problem:

$$Y = X\beta + \varepsilon \qquad (1)$$

where $Y \in \mathbb{R}^n$ is the vector of responses; $X \in \mathbb{R}^{n \times p}$ is the predictor matrix with i^{th} row x_i^T; $\varepsilon \in \mathbb{R}^n$ represents random error; and $\beta \in \mathbb{R}^p$ is the unknown vector of coefficients that determines the regression function and is to be estimated using the data.

A traditional application of the model (1) may have the responses as blood pressure measurements for $n = 100$ patients and the predictors could include height, weight, age and daily calorie intake, for example. In this case, one might estimate β by ordinary least squares (OLS), a technique dating back to Gauss (1795). This yields an estimator $\hat{\beta}^{\text{OLS}}$ with

$$\hat{\beta}^{\text{OLS}} := \underset{b \in \mathbb{R}^p}{\arg\min} \|Y - Xb\|_2^2 = (X^T X)^{-1} X^T Y,$$

provided X has full column rank. Here $||\cdot||_2$ denotes the Euclidean norm. We can analyse the quality of the estimate of the regression function by calculating its *mean-squared prediction error* (MSPE). Under the assumptions that (i) $\mathbb{E}(\varepsilon_i) = 0$ and (ii) $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \mathbb{I}_{\{i=j\}}$, it holds that

$$\text{MSPE}(\beta^{\text{OLS}}) := \mathbb{E}\left\{\frac{1}{n}\left\|X(\beta - \beta^{\text{OLS}})\right\|_2^2\right\} = \frac{p}{n}\sigma^2.$$

We see that provided p/n is small, the MSPE is small. When this is true, and under the assumptions given above, OLS is a very reasonable choice of estimator. Indeed, the Gauss-Markov theorem shows that it has the minimum MSPE among all linear unbiased estimators of the regression function, i.e. among all estimators $\hat{F} := AY$ of $X\beta$, for some $n \times n$ matrix A such that $\mathbb{E}(AY) = X\beta$.

High-dimensional Data

One might think that OLS essentially solves the problem of linear regression, at least under assumptions (i) and (ii). However, the field of statistics must constantly adapt and innovate to develop methods that accommodate the data it is tasked with to study, and today, much of that data is *high-dimensional*: p is very large and often greatly exceeds n. Where in the past only a few carefully chosen variables were measured for each observation, nowadays any variable that might conceivably have an effect on the response tends to be recorded, leading to ratios of $p/n > 1000$ being common in some areas. Out of these many variables, it may be only a few that are really important for predicting the response, but of course which these are would not be known in advance. In the context of our model (1), this would translate as β being sparse, i.e. many of its components being exactly 0.

How are we to proceed with analysis given such datasets? Clearly OLS is unhelpful when $p \geq n$ if X has full column rank, as then predictions are simply the original responses themselves. Ideally, we would like a sparse estimator that first attempted to detect the relevant variables, and then estimated just *their* coefficients. Even when β is not truly sparse, it can make sense to produce a sparse estimate for it. A final estimate with only a few non-zero coefficients may be much easier to interpret, and given a new observation $x \in \mathbb{R}^p$, computing the estimate of the regression function at x would be much faster.

In view of this, one might consider trying to estimate β by $\hat{\beta}^{\mathrm{BS}}$ (Best Subsets) defined as the minimiser of a penalised least squares objective:

$$\hat{\beta}^{\mathrm{BS}} := \underset{b \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \|Y - Xb\|_2^2 + \lambda \|b\|_0 \right\}, \quad (2)$$

where $\|b\|_0 := \sum_{k=1}^{p} \mathbb{I}_{\{b_k \neq 0\}}$. Large values of the *regularisation parameter*, λ, will cause $\hat{\beta}$ to have very few non-zero components, and lower values will produce less sparse models. Several methods are available for choosing λ; we will not go into the details here.

A major problem with the estimator $\hat{\beta}^{\mathrm{BS}}$ is that the optimisation in (2) is in general computationally infeasible as one would essentially need to evaluate the objective at all 2^p possible subsets of variables in order to guarantee finding the optimiser. As p runs into the hundreds, the number of computations required to perform the optimisation quickly surpasses the number of atoms in the observable universe!

The Lasso

The key property of the objective in (2) which makes the optimisation intractable is that it is non-convex. Convex problems are much easier to solve, one reason for this being that any local optimum is also a global optimum. It is thus sensible to consider convex approximations to the objective in (2). One such approximation results from replacing $\|b\|_0$ with $\|b\|_1 := \sum_{k=1}^{p} |b_k|$, so our estimator $\hat{\beta}$ is given by

$$\hat{\beta} := \underset{b \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right\}, \quad (3)$$

This is the Lasso (Least Absolute Shrinkage and Selection Operator) estimator (see [2]): one of the most popular methods in high-dimensional data analysis. Applications of the Lasso and related methods range from identifying which of our thousands of genes are related to particular diseases, to the click-through rate prediction task that optimises web advertising for search engines.

The optimisation in (3) can be solved even for very large problems where p is hundreds of thousands. Yet importantly, sparsity of the estimator is retained. This is essentially because the set $\{b : ||b||_1 \leq r\}$ for $r > 0$ has corners; see [1] for details.

Sparsity and computational feasibility are attractive properties, but what really makes the Lasso appealing is its remarkable performance as both a selector of important variables, and as a prediction engine. This is perhaps surprising given that the Lasso optimisation only approximates (2). A vast amount of work has gone into trying to understand why the Lasso works so well, and also into developing improvements and adapting the method to suit many other problems. We will fin-

ish with a theorem that forms part of that work: we show that provided $||\beta||_1$ is small, as would be the case when β is sparse, p can be almost exponential in n, and the Lasso can still estimate the regression function well. This is a really rather striking result, considering that OLS requires p to be much smaller than n. In fact much more is true; the interested reader is directed to [1].

Prediction Error of the Lasso

Theorem Let $\hat{\beta}$ be the Lasso estimator (3) with $\lambda = \sigma\sqrt{2(\log p + t^2)/n}$. Assume that the errors ε_i are independent and normally distributed. Then with probability at least $1 - e^{-t^2}$, we have

$$\frac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le 2\sigma\|\beta\|_1\sqrt{\frac{2(\log p + t^2)}{n}}.$$

Proof By the definition of β we have

$$\frac{1}{n}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \le \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Recalling that $Y = X\beta + \varepsilon$ and rearranging, we get

$$\frac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le \frac{1}{n}2\varepsilon^T X(\hat{\beta} - \beta) + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1$$

Denote the k^{th} column of X by $X_k \in \mathbb{R}^n$. Now

$$\frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta) \le \frac{1}{n}\|\beta - \hat{\beta}\|_1 \max_{1\le k\le p}\left|X_k^T\varepsilon\right|,$$

and as $\varepsilon \sim N_n(0, \sigma^2 I)$ and $||X_k||_2^2 = n$, we have $X_k^T\varepsilon/n \sim N(0, \sigma^2/n)$. Now let $Z \sim N(0,1)$ so $\sigma Z/\sqrt{n}$ has the same distribution as $X_k^T\varepsilon/n$ for each k. We argue

$$\begin{aligned}\mathbb{P}\left(\max_{1\le k\le p}\left|X_k^T\varepsilon\right|/n \ge \lambda\right) &= \mathbb{P}\left(\bigcup_{k=1}^{p}\left\{\left|X_k^T\varepsilon\right|/n \ge \lambda\right\}\right)\\ &\le \sum_{k=1}^{p}\mathbb{P}\left(\left|X_k^T\varepsilon\right|/n \ge \lambda\right)\\ &= 2p\mathbb{P}\left\{Z \ge \lambda\sqrt{n}/\sigma\right\}.\end{aligned}$$

A standard tail bound for normal random variables (proved below) gives us that for $\zeta \ge 0$, $1 - \Phi(\zeta) \le e^{-\zeta^2/2}/2$. Applying this to the above, we see that the final term in the last display is bounded above by $p\exp\left\{-n\lambda^2/(2\sigma^2)\right\} = e^{-t^2/2}$. Now working on the event $\left\{\max_{1\le k\le p}\left|X_k^T\varepsilon/n\right| < \lambda\right\}$, which we have shown has probability at least $1 - e^{-t^2}$, we have from (3) that

$$\frac{1}{n}\|X(\beta - \hat{\beta})\|_2^2 \le \lambda\|\beta - \hat{\beta}\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.$$

Noting that $\|\beta - \hat{\beta}\|_1 - \|\hat{\beta}\|_1 \le \|\beta\|_1$ by the triangle inequality completes the proof.

Standard Normal Tail Bound

Theorem Let $Z \sim N(0,1)$. Then

$$1 - \Phi(\zeta) := \mathbb{P}(Z \ge \zeta) \le \tfrac{1}{2}e^{-\zeta^2/2} \text{ when } \zeta \ge 0.$$

Proof Let $f(\zeta) := 1 - \Phi(\zeta) - \frac{1}{2}e^{-\zeta^2/2}$. Now

$$\begin{aligned}\mathbb{P}(Z \ge \zeta) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_\zeta^\infty e^{-z^2/2}dz\\ &\le \frac{1}{\zeta\sqrt{2\pi\sigma^2}}\int_\zeta^\infty ze^{-z^2/2}dz\\ &= \frac{1}{\zeta\sqrt{2\pi\sigma^2}}e^{-\zeta^2/2}.\end{aligned}$$

Thus if $\zeta \ge \sqrt{2/(\pi\sigma^2)}$, then $f(\zeta) \le 0$. Also, $f(0) = 0$. Finally observe that

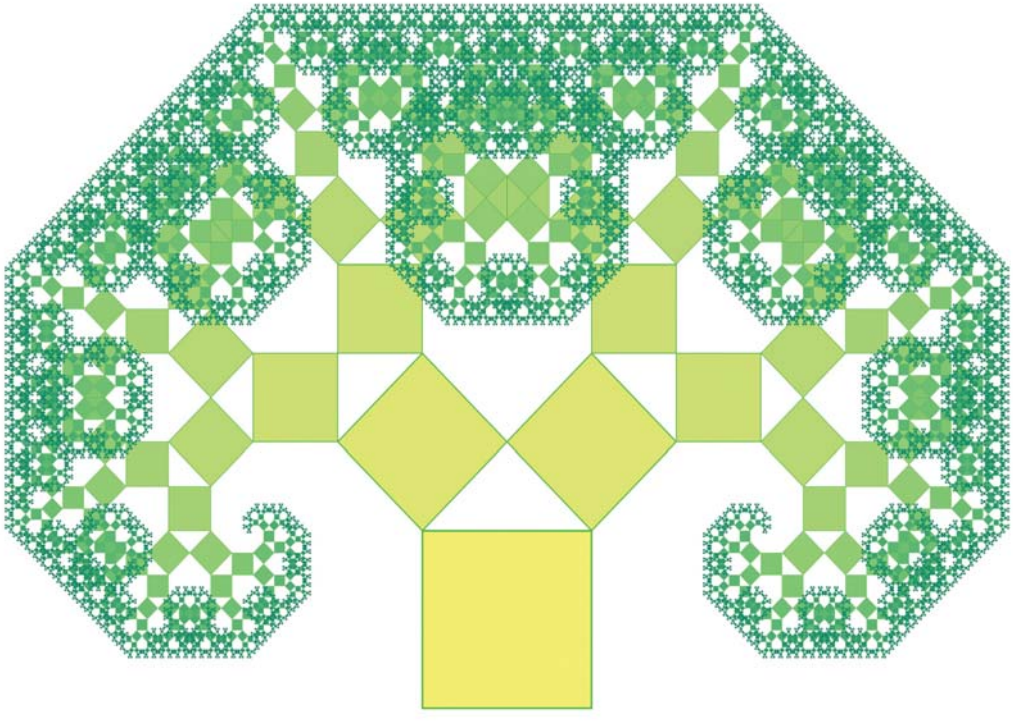
$$f'(\zeta) = -\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\zeta^2/2} + \frac{\zeta}{2}e^{-\zeta^2/2},$$

so $f'(\zeta) \le 0$ for $\zeta \le \sqrt{2/(\pi\sigma^2)}$. Conclude that $f(\zeta) \le 0$ when $\zeta \ge 0$.

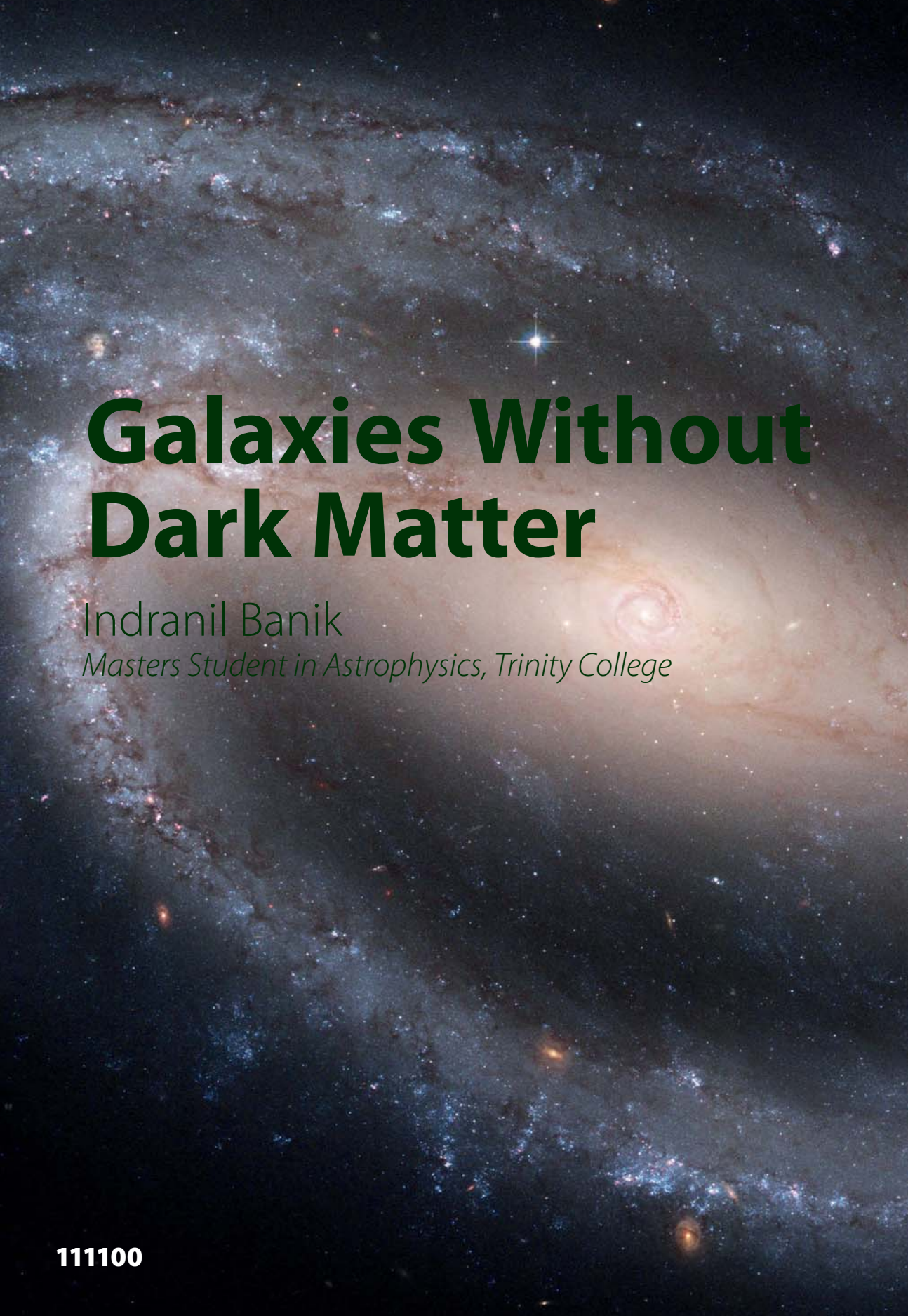
References

[1] Peter Bühlmann, Sara van de Geer; 2011; *Statistics for High-Dimensional Data: Methods, Theory and Algorithms*; Springer.

[2] Robert Tibshirani; 1996; *Regression shrinkage and selection via the Lasso*; J. Roy. Statist. Soc., Ser. B, 58, 267-288.



The Pythagoras trees are fractals constructed from squares. Each triple of touching squares encloses a right-angled triangle, thus demonstrating Pythagoras' Theorem infinitely many times.



Galaxies Without Dark Matter

Indranil Banik
Masters Student in Astrophysics, Trinity College

111100

Recent observations indicate that the Universe as a whole is dominated by unknown substances. Dark energy is needed to make the expansion rate increase with time. The precise way in which this happens indicates that it makes up ~68% of the total. It is thought to be spread completely uniformly and may be a property of spacetime itself. Other arguments indicate that baryons (protons and neutrons) can only make up 5%. The strongest argument comes from nuclear reactions between protons and neutrons shortly after the Big Bang. These reactions generated a large amount of helium and traces of deuterium. The amounts produced would be altered if there were more baryons. Today, therefore, the existence of dark matter seems indisputable.

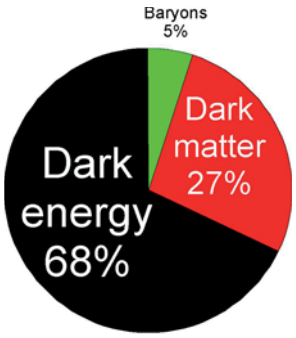


Figure 1 The total mass-energy content of the Universe

Where Dark Matter Might Be

Dark matter is thought to be an undiscovered fundamental particle which can't radiate, preventing it forming small objects like planets. If it tried, it would get hot and the pressure would stop further collapse. But do galaxies contain large amounts of dark matter, or is it spread even more thinly? Answers may be provided by galactic rotation curves (Figure 2). For circular orbits, the speed should be given approximately by

$$\frac{v^2(r)}{r} = \frac{GM_b(< r)}{r^2} \quad .$$

Note M_b means baryonic mass only, as one might at first expect. Only the mass at radii smaller than r contributes to the force because of the Shell Theorem. At large radii, this includes the bulk of the mass. Thus, M_b hardly varies with r, making for a so-called Keplerian rotation curve. This drops to 0 at large radii. It does not go flat at a non-zero

value. Real rotation curves, however, usually do.

Taking a hint from Figure 1, most scientists believe that adding dark matter can resolve the problem. The trick is to fudge the mass distribution so that $M(<r) \propto r$, even outside the bulk of the visible mass. The stars and gas are not distributed like this, but the dark matter might be. This explanation seemed to resolve other problems too, the main one being that a self-gravitating disk (like a spiral galaxy) is unstable. Adding a huge halo of dark matter provides an additional restoring force.

The dark matter and the baryons would be subject to very different effects. For example, radiation from supernovae can heat and eject large amounts of gas but little affect dark matter. Gas can also be accreted from surrounding regions more easily than dark matter. Thus, the ratio of the two should be unlikely to remain fixed. Observing the baryons would shed little light on how much dark matter there should be, just as measuring a star's properties can't tell us what sort of planets orbit it. Galaxies are perhaps even more complicated. Because of this, plus the fact that it is the dark matter that dominates their total mass, we wouldn't expect it to be possible to predict the rotation curve based on the observed baryons.

The Baryonic to Dark Matter Ratio in Galaxies

Despite all this complexity, however, the value of v_∞, the velocity at which a rotation curve flatlines (if it does so) can be predicted remarkably accurately based on the baryonic mass alone. This result is called the Baryonic Tully-Fisher Relation (Figure 3). This is surprising, considering that some galaxies have lost more than 95% of the baryons originally present (assuming a 1 : 5 ratio of baryonic : dark matter initially). This loss is often due to supernovae – explosions when massive stars die. In a dwarf galaxy, only a few of them would be necessary to remove most of the baryonic mass. Thus, loss of baryons must be a somewhat random process. Certainly it seems feasible that two dwarf galaxies could have started similarly, with one losing 'only' 90% of its baryons and the other losing 95%. The latter would have nearly the same v_∞ but only half the baryonic mass. Galaxies can also accrete gas from their surroundings, with the amount depending on their environment and merger history (whether other galaxies collided with them).

The ratio of baryonic to dark matter would therefore seem unpredictable. But it should certainly be very small, especially in dwarf galaxies. Baryons should hardly matter. This, together with their

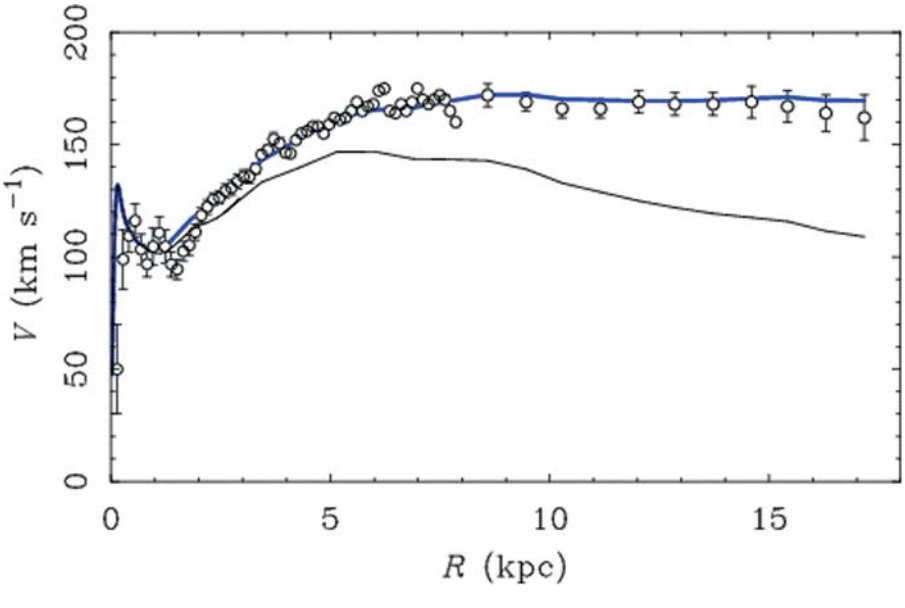


Figure 2 Rotation speed as a function of radius in a disk-like galaxy. The dashed line is the prediction of Newtonian dynamics, based on observed baryonic mass. The solid line is the prediction of MOND.

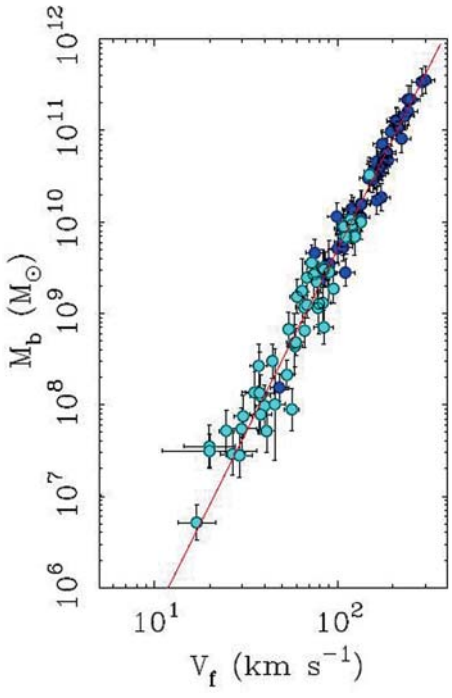


Figure 3 This shows v_∞ against baryonic mass for a large number of galaxies. Different colours indicate gas dominated (usually less massive) or star dominated, quite different types of galaxy. The MOND prediction is the thin red line, assuming $a_0 = 1.2 \times 10^{-10}$ m/s^2.

complexity, would make them a poor guide to the total mass. Yet accounting for the baryons lets one *predict* the dynamics fairly straightforwardly and with very high accuracy. The most natural explanation is that baryons are all there is.

MOND

This elegant solution suffers a serious problem: observed baryons don't exert enough force. Or do they? If one is willing to modify Newton's laws, then masses may exert more gravity than he assumed. Or perhaps Newton's second law does not work, though we won't consider that here. Thinking along these lines, Mordehai Milgrom proposed in 1983 a theory known as Modified Newtonian Dynamics (MOND). The essential thing is for gravity to behave as $1/r$ rather than $1/r^2$ at large distances from a point mass, thereby leading to a flat rotation curve. Looking at the data, Milgrom realised that distance was not the crucial parameter, rather acceleration was. Newtonian gravity needed to break down below a threshold acceleration, a_0.

The governing equation for gravitational fields (the Modified Poisson Equation) reads

$$\nabla \cdot \left(\mu\left(\frac{|\vec{g}|}{a_0}\right) \vec{g} \right) = -4\pi G\rho \,.$$

$\vec{g}$ is the gravitational field (force per unit mass) and μ is a function. For spherically symmetric situations, the end result is that

$$\mu\left(\frac{|\vec{g}|}{a_0}\right) \vec{g} = \vec{g}_n,$$

where $\vec{g}_n$ is the acceleration if Newtonian gravity were to hold exactly. If we set $\mu(x) = 1 \ \forall x$, then we recover the usual Poisson equation (leading to Newtonian gravity). Although everyday experience clearly shows that μ must be 1 for everyday accelerations, there is no good reason why this should be the case for very low accelerations. In fact, Milgrom suggested that a_0 was a tiny 1.2×10^{-10} m/s^2.

The Modification comes from setting $\mu(x) = x$ when $x << 1$. This way, we get $g = (g_n a_0)^{1/2}$. Putting in $g_n = GM/r^2$ for a point mass, one sees that the force eventually becomes

$$g = \frac{\sqrt{GMa_0}}{r}.$$

Note that the force due to a combination of masses is not the sum of the forces due to each considered individually. Using $g = v^2/r$ to find the speed of particles (e.g. stars) on circular orbits, one obtains two important results: the orbital speed does not vary with r (by construction) and

$$v_\infty = \sqrt[4]{GMa_0}.$$

Predictive Power

Not only can MOND predict the value of v_∞, it can also fit individual rotation curves in detail (Figure 4), including bumps and wiggles. These are due to similar features in the underlying mass distribution.

Without MOND, however, it would be a remarkable coincidence if a bump in the baryonic mass density were matched by a similar (but much larger) bump in the dark matter density at the same position. After all, baryons can form clumpy structures because they can radiate and cool. Dark matter cannot radiate, so it cannot be expected to have such small-scale features. Moreover, it is in a

nearly spherical halo whereas the baryons are in a disc. So it is very difficult to see how a theory like MOND that just boosts the gravity of the baryons by a fixed factor (depending on the acceleration) can ever match reality.

There are dozens more galactic rotation curves well fitted by MOND. By and large, without adding arbitrary and very large amounts of dark matter, it seems to predict rotation curves well (using $\mu(x)=x/(1+x)$). However, as is inevitable, there are a small fraction of cases (about 10%) in which the observations are less accurate and the orbits of stars aren't purely circular etc.

Figure 5 summarises the results from rotation curve studies of nearly 100 galaxies. Amazingly, the baryonic mass distribution is sufficient to predict the acceleration. This is despite galaxies supposedly having lost most of their baryons in highly complicated processes and accreted variable amounts depending on environment, leading inevitably to different relative proportions of baryonic and dark matter. This proportion should not be calculable based on observing the baryons alone. Yet Figure 5 essentially provides a formula for doing so, even when the true acceleration is tens of times larger than g_n (see [1]).

Further Evidence for MOND

Galaxies may appear isolated, but they often interact. Standard simulations of this indicate gas and dark matter end up separated due to their very different initial distributions and physics. Gas is drawn into long tidal tails, which may later form into dwarf galaxies. Crucially, these galaxies should be devoid of dark matter. They are also too puny to accrete much of it. Therefore, their rotation curves should follow from applying Newton's laws to the observed baryons. Figure 6 shows the results of such an attempt.

Mergers between spiral galaxies can easily destroy their disks, forming an elliptical galaxy. Yet disks are hardly rare: the majority of nearby heavy galaxies are rotating disks, with little evidence of disruption. One reason why mergers are believed to be common is because galaxies are supposedly surrounded by huge dark matter halos, making them easy targets. Another is that galaxies empty their surroundings fairly slowly, leading to collisions at late times. By then, a lot of the gas has been converted into stars. This means there is little gas drag to prevent disruption of the disk in a merger.

On a larger scale, the distribution of galaxies is unusual in the standard picture. This involves starting with a slightly inhomogeneous mixture of dark and baryonic matter. Overdense regions have stronger gravity and become even denser, some eventually forming galaxies. The Local Void is a nearby large underdense region. Observing a portion of the Local Void (within 25MLy of Earth), one finds three large galaxies. Yet simulations suggest we should observe around 19.

One possibility is that galaxies exert stronger gravity than in the model, letting them empty their surroundings faster and more thoroughly. This might also explain the prevalence of disks. MOND would indeed provide stronger gravity. It can also solve several other problems not discussed here (see [2]).

Conclusion

The Universe probably has large amounts of dark matter. It is tempting to use this to explain motions within galaxies. But doing so leads to amazing coincidences and major problems. Moreover, a theory that does not invent vast and arbitrary amounts of invisible matter actually performs better. Galaxies can and should be understood using only actually observed mass. What you see is all there is. But using Newtonian dynamics will force you to invent dark matter.

Real dark matter might well exist, but only on larger scales. Perhaps it resides in galaxy clusters, where it is needed to explain dynamics, even in MOND (the required ratio of baryonic to dark matter is, on average, the same as in Figure 1). This would allow it to slow down the expansion of the Universe without affecting internal galactic dynamics. If this were so, then General Relativity (which reduces to Newtonian dynamics in galaxies) cannot be the whole story. Considering that this theory has never been directly tested at accelerations as low as a_0, perhaps this is not very surprising. After all, no theory works everywhere.

References

[1] Mordehai Milgrom; 2013; *Testing the MOND Paradigm of Modified Dynamics with Galaxy-Galaxy Gravitational Lensing*; http://arxiv.org/pdf/1305.3516v2.pdf.

[2] Benoit Famaey; 2012; *Modified Newtonian Dynamics (MOND): Observational Phenomenology and Relativistic Extensions*; http://arxiv.org/pdf/1112.3960v2.pdf.

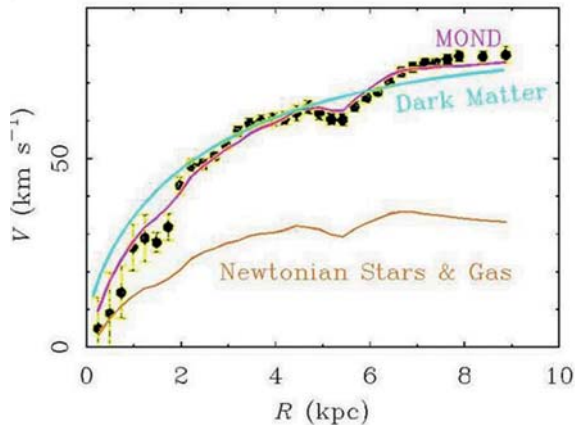


Figure 4 Using the same value of a_0, it is also possible to fit in detail the rotation curve of NGC1560, a dwarf galaxy. It is clear that the true rotation curve is best explained by multiplying the Newtonian rotation curve (without dark matter) by some factor which increases with r. This is precisely what MOND does. If instead we add a smooth halo of dark matter, the bump disappears. Note the MOND calculation is based on actually observed mass, unlike the dark matter curve (the total halo mass and size are adjusted to match the data best).

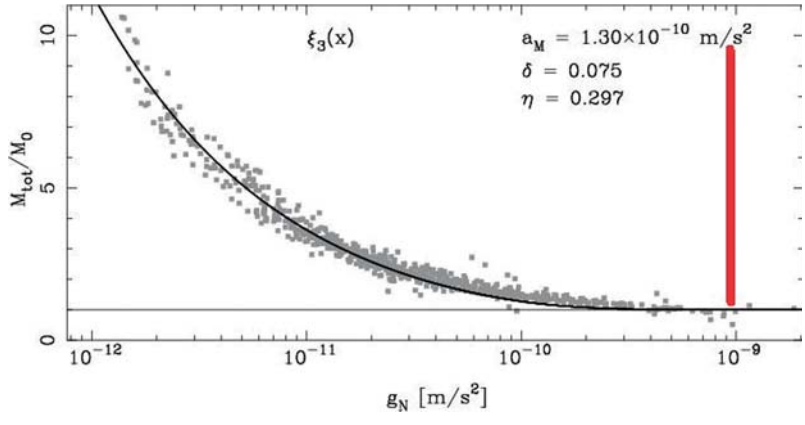


Figure 5 The ratio between the true acceleration and that predicted by Newtonian dynamics from the baryonic mass, as a function of the latter. MOND requires a unique relation between the two, unexpected with dark matter. Quantum gravity effects should be important at ultralow accelerations (left of red line) because the energy density in the gravitational field is smaller than that in the vacuum due to quantum fluctuations (the zero point, or dark, energy).

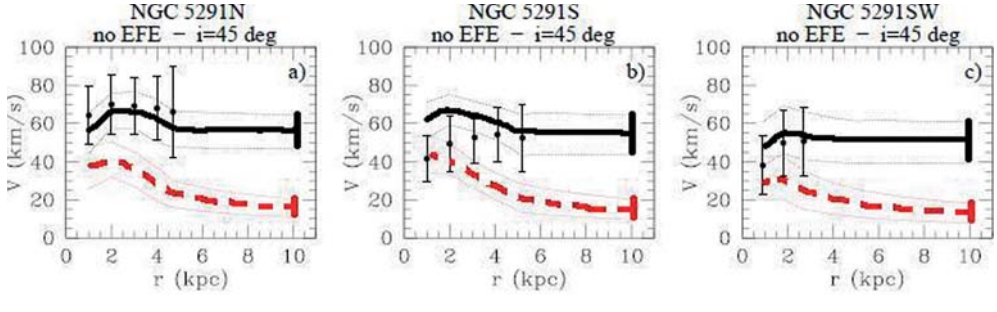


Figure 6 Rotation curves of 3 galaxies born from debris of a galactic close encounter. The dashed curves are predictions of Newtonian dynamics without dark matter, which should not be present in these galaxies. The solid curves are predictions of MOND. Both predictions and observations have errors shown. The inclination of the dwarfs to the plane of the sky is fixed at the most likely value, based on the orbital geometry of the interacting progenitor galaxies.

The Death of a Mathematician

Dr Mario Livio
Senior Astrophysicist, Hubble Space Telescope Science Institute

In the morning hours of May 30, 1832, a single shot fired from 25 paces hit Évariste Galois in the stomach. Although fatally wounded, Galois did not die on the spot. He remained lying on the ground until an anonymous good Samaritan, perhaps a former army officer, perhaps a peasant passing by, picked him up and brought him to the Cochin hospital in Paris. The following day, with his younger brother Alfred at his side, Galois died of peritonitis. His last known words were: "Don't cry, I need all my courage to die at twenty."

This was the gloomy end of the most romantic of all mathematicians – a young man in whose mind the sweeping ideals of the French Revolution were inseparable from the revolutionary new branch of mathematics he had invented. Galois is the originator of Group Theory – the mathematical language that describes symmetry. Just as arithmetic has become the language of accounting, and colour and shape the language of abstract art – mathematicians, physicists, and even economists use group theory to explore the labyrinths of symmetry.

You might have expected that every intimate detail in the life of such a prominent mathematician would be widely known. Yet, Galois's death has remained veiled in mystery for almost two centuries. What is known is that Galois was killed in a duel with pistols on that fateful morning in 1832, but the questions of who killed him and why have been the subject of conspiracy theories galore. Biographers have further been perplexed by the fact that the wounded young man appeared to have been abandoned in the field.

Following three years of intensive research, I proposed in 2005 that the fog surrounding Galois's mysterious death may have finally lifted.

Various Theories

The known facts concerning Galois's activities in the last week of his life are precious few. Even Galois's own three letters, indicating that "two patriots" (meaning active republicans) provoked the duel over "something so contemptible" involving an "infamous coquette," did not shed sufficient light on the identity of his opponents or their motives. The fact that Galois was a revolutionary firebrand inspired many of his early biographers to speculate that political enemies killed Galois. A few have allowed their imaginative plots to take off and include even more intrigue, suggesting that the "coquette" was in fact a police agent masquerading as a prostitute.

The first clue pointing to an unrequited love as the potential cause for the duel came from the work of an unlikely "detective" – a Uruguayan university professor. Using a magnifying glass and special lighting to examine Galois's papers, Carlos Alberto Infantozzi discovered in 1968 the identity of the "infamous coquette" – Stephanie Potterin du Motel. This young woman lived in a building that housed a convalescent home where the troubled Galois was placed on parole after being released from prison. Stephanie was certainly neither a prostitute nor a police provocateur.

Based on Infantozzi's "forensic" work, as well as on an article from the period in the Lyon newspaper Le Précurseur (reproduced by the French author Andre Dalmas in 1956), Dalmas and the American physicist and author Tony Rothman have started to painstakingly put together the pieces of the puzzle. They suggested that the person who shot Galois was Stephanie's presumed lover (and a personal friend of Galois), since they believed that the young seductress was playing a double game with the hearts of the two young men.

This was more or less the accepted consensus until 1996, when a new biography by an Italian historian of mathematics turned the entire story on its ear. Laura Toti Rigatelli suggested that the famous duel wasn't even a duel at all! Rather, this somewhat Machiavellian theory proposed that Galois sacrificed himself for the Republican cause – the republicans needed a corpse to stir up rebellion, and he offered his. While many accepted Toti Rigatelli's story, not all did. The French researcher and author Jean-Paul Auffray, who conducted an extensive study of documents related to Galois, concluded that the duel was real. Auffray reintroduced the theory that the unfortunate love affair with Stephanie provoked the duel, and suggested that one of the opponents was none other than Stephanie's father.

I have always been fascinated by Galois. How can you not be? When you realize that this flamboyant romantic brought about one of the greatest breakthroughs in mathematics, and that he achieved this feat before the age of twenty! When I started to research the life, and especially the death, of this visionary genius, I decided to embark on this task with no prejudices, and to leave no stone unturned. Having had the added advantage of being able to examine critically all the evidence collected by numerous researchers and their conclusions, in three years I was able to develop what at least appears to be an entirely self-consistent picture. While the new theory clearly contains elements of previous scenarios, it combines these elements with new insights that give them, in my humble opinion, an enhanced credibility. I therefore strongly believe that I have come closer to the truth than was ever possible before.

My Conclusion

So, who killed Galois and why? A key point ignored by many biographers is that Galois always talked about two people who provoked the duel. One could, therefore, not expect to find a complete answer without an identification of both opponents. My conclusion is that these two people were Denis Faultrier and Ernest Duchatelet. The former was a close friend of Stephanie's family, and he later married her widowed mother. The latter was Galois's republican friend (and Stephanie's presumed lover), and it was he who shot the fatal bullet. The entire affair was a classical case of cherchez la femme. From Stephanie's two devastating letters to Galois we learn that either by some careless words, or by too impetuous a behaviour, the inexperienced Évariste offended

Galois at 15, as drawn by a classmate

Galois at 17, as drawn by his brother

Galois's Birthplace All images: Municipality of Bour-la-Reine, through the assistance of Philippe Chaplain

Stephanie, who immediately informed her so-called "fiancé" (Duchatelet) and so-called "uncle" (Faultrier; the two descriptions were given by Galois's cousin, Gabriel Demante). When the two men confronted Evariste, the hot-blooded young man added insult to injury, referring to the entire incident as a "miserable piece of gossip." At a time when invitations to duels were issued at the drop of a hat, this was more than enough for the two men to challenge Galois. A seventeen-year-old young woman who did not return his love sealed the fate of one of the most brilliant mathematicians to have ever lived.

Why did Galois appear to have been left wounded on the ground by most, if not all, of the seconds? Galois' autopsy report describes a large bruise on his head that was probably caused when he fell. He might have been knocked unconscious and presumed dead. One of the reports from the period notes that a "former officer" brought Galois to the hospital. This fits Denis Faultrier, a former captain in the national guard, and the second opponent in my scenario, like a glove. In the book "The Equation that Couldn't Be Solved", I presented more details of what had led me to the proposed course of events. Can the two-centuries-old case finally be closed? Hopefully, yes. But with a number of gaps in the hard evidence, uncertainties are likely to remain. What is certain is the fact that Évariste Galois will always be remembered as one of the most creative individuals to have ever lived. The new branch of mathematics that he established has expanded far beyond the boundaries of pure mathematics, into the realms of physics, economics, music, visual arts, and wherever symmetries can be found.

About the Author

Mario Livio is an astrophysicist at the Space Telescope Science Institute in Baltimore, Maryland. He is also the author of several popular science books. His most recent book, "Brilliant Blunders" was published in June 2013, and was a New York Times bestseller. Mario Livio earned a BS degree in Physics and Mathematics at the Hebrew University of Jerusalem, a MS degree in Theoretical Particle Physics at the Weizmann Institute and a PhD in Theoretical Astrophysics at Tel-Aviv University.

References

[1] Mario Livio; 2005; *The Equation That Couldn't Be Solved*; New York; Simon & Schuster.
[2] Jean-Paul Auffray; 2004; *Evariste 1811-1832*; Lyon; Aleas.
[3] Carlos A. Infantozzi; 1968; *Sur la mort d'Evariste Galois*; Revue d'histoire des sciences; 21, 1968.
[4] Tony Rothman; 1982; *Genius and Biographers*; The American Mathematical Monthly; 89, 2, 84.
[5] André Dalmas; 1956; *Evariste Galois: Revolutionnaire et Geometre*; Paris; Fasquelle.
[6] Laura T. Rigatelli; 1996; *Evariste Galois: 1811-1832*; Basel; Birkhauser Verlag.
[7] Alexandre Astruc; 1994; *Evariste Galois*; Paris; Flammarion.

Triangular Wriggle, by Roberto Giardili

The Triangular Wriggle won the prize for 'Most Effective Use of Mathematics' at the 2013 Bridges Conference for Mathematics in Art. The sculpture is based on a Lindenmayer-system, a parallel rewriting system that can be used to generate fractals.

The Mandelbrot Set

Nikolaos Athanasiou
Second Year Mathematics Undergraduate, King's College

The rise and development of the field of complex dynamics (pioneering research in this area has been carried out by A. Douady, P. Fatou , J. Hubbard , G. Julia , C. Mcmullen , J. Milnor , W. Thurston , J. C. Yoccoz *et al.*) revealed a set with properties of such peculiarity and beauty that one cannot help but be amazed at it. It is perhaps the most famous example of how chaos, or to be even more precise, infinite complexity, can be created by iterating a procedure as simple as $z \mapsto z^2 + c$.

Definition and Basic Result

.Let $f(z) = z^2 + c$. The Mandelbrot set is defined as:

$$\mathcal{M} = \{c \in \mathbb{C} \mid \exists\ K \in \mathbb{R}^+ : |f^{(n)}(0)| < K\ ,\ \forall n \in \mathbb{N}\}$$

where $f^{(n)}$ denotes the n-fold application of f. Put simply, $\mathcal{M}$ is the set of all complex numbers c for which the sequence

$$(f^{(n)}(0)), n = 1, 2, ...$$

is bounded in modulus. Strictly, we are interested in the dynamical system

$$\left(\mathbb{C}, z \mapsto z^2 + c\right) = F_c$$

which is the real object we wish to understand. At first sight, the choice of the number 0 in the definition seems arbitrary. However, 0 is the only critical point of $f(z) = z^2 + c$ (the point where its derivative vanishes). Since critical points govern the dynamics of f (see [4]), the use of the number 0 makes sense.

The iteration seems fairly simple, yet the resulting shape is certainly not:

The first reasonable question to pose is what happens if we restrict the domain to the real numbers.

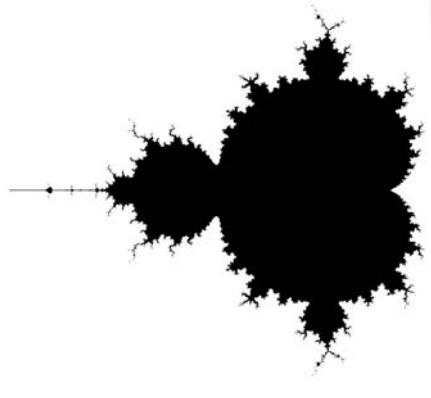


Figure 1 The shape of the Mandelbrot set

In this article we will focus our attention on real parameters. Our task is to find the set $\mathcal{A} = \mathcal{M} \cap \mathbb{R}$, called the antenna.

Theorem $\mathcal{M} \cap \mathbb{R} \equiv [-2, \frac{1}{4}]$

Proof We begin by stating a ratio test for sequences, in the following form:

Ratio test for sequences Let a_n be a sequence of positive reals such that

$$\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \lambda$$

- If $\lambda > 1$ the sequence diverges (escapes to infinity)
- If $\lambda < 1$ the sequence converges.

First of all, the sequence $b_n = f^{(n)}(0)$ satisfies $b_{n+1} = b_n^2 + x$ and thus $b_{n+1} - b_n = b_n^2 - b_n + x$.

Lemma 1 $\mathcal{A} \cap (\frac{1}{4}, \infty) = \varnothing$

Proof For $x > \frac{1}{4}$, notice that $b_i > 0$, i and $b_{n+1} - b_n > (b_n - \frac{1}{2})^2$, which in turn implies that our sequence is increasing. This implies that the limit

$$\lim_{n\to\infty} b_n$$

exists (but is not necessarily finite). Hence $\lim_{n\to\infty} \frac{x}{b_n}$ exists and is finite. In fact,

$$\lim_{n\to\infty} \frac{b_{n+1}}{b_n} = \lim_{n\to\infty} \left(b_n + \tfrac{x}{b_n}\right)$$

exists (but is not necessarily finite).

Now we make use of the inequality $b_n + \frac{x}{b_n} \geq 2\sqrt{x} > 1$ (AM-GM) and we conclude that

$$\lim_{n\to\infty} \frac{b_{n+1}}{b_n} > 1$$

so the sequence diverges for these values of x.

Lemma 2 $[0, \frac{1}{4}] \subset \mathcal{A}$

Proof To show that $\frac{1}{4} \in \mathcal{A}$ we prove by induction that $b_n \leq \frac{1}{2}n$. Indeed, the base case is trivially true: if $b_k \leq \frac{1}{2}$, then $0 < b_k + 1 = b_k^2 + \frac{1}{4} \leq \frac{1}{2}$. For the rest of the proof, observe that each part of the sequence is a polynomial in x with only positive coefficients and hence each term forms a strictly increasing function of x in $(0, \infty)$. Thus, since $x \in [0, \frac{1}{4}]$, we have $b_n = b_n(x) \leq b_n(\frac{1}{4}) \leq \frac{1}{2}$ and the lemma follows.

Let's now move on to the negative reals.

Lemma 3 $\mathcal{A} \cap (-\infty, -2) = \varnothing$

We demonstrate once more that, within the given interval, our sequence is positive and increasing. Note that for $x < -2$ we have $b_2 = x^2 + x > |x| > 2$. urthermore, $b_2^2 - b_1^2 = x^3(x + 2) > 0$. Now, observe that $b_{n+1} - b_n = b_n^2 - b_{n-1}^2$ and an easy induction shows the sequence is increasing. For every number x in our interval, the same proof as in lemma 1 implies that the limit

$$\lim_{n\to\infty} \left(b_n + \tfrac{x}{b_n}\right) = \lim_{n\to\infty} \frac{b_{n+1}}{b_n} = A$$

exists (though it may be $+\infty$) and is equal to

$$A = \lim_{n\to\infty} \left(b_n + \tfrac{x}{b_n}\right) > \lim_{n\to\infty} \left(2 + \tfrac{x}{b_n}\right) \geq 1$$

since $b_n \geq x$. Again, by the ratio test for sequences, the claim follows.

Lemma 4 $[-2, 0] \subset \mathcal{A}$

Proof If $x = -2$ then, $b_n(x) = 2\ \forall n \geq 2$, so $(-2) \in \mathcal{A}$. If $x \in (-2, 0]$, then $b_{n+1} = b_n^2 + x > b_n^2 - 2 \geq -2$ and the sequence is bounded below. Finally, $|b_2| = |x^2 + x| \leq |x|$. By induction if $|b_k| \leq |x|$ then $b_{k+1} = b_k^2 + x \leq x^2 + x \leq |x| \leq 2$ and the sequence is also bounded above, which proves the lemma. □

Lemmas 1-4 combined finish the proof of the theorem. □

Some Further Remarkable Properties

Infinite complexity No matter how much you zoom in on a point close to the boundary of the Mandelbrot set, there is going to be a new geometric shape revealed after any degree of magnification!

Preperiodic (Misiurewicz) points We call a sequence (a_n) pre-periodic if and only if it becomes periodic after a finite number of steps, that is if there exist M, T: $a_{n+T} = a_n$, $\forall n > M$. Accordingly, a point in the Mandelbrot set is called a Misiurewicz point if the resulting sequence from this point is pre-periodic. (If the reader is interested, he/she may search about post-critically finite maps, Thurston's theorem or rigidity.) Misiurewicz points are dense on the boundary of the Mandelbrot set and in fact, the Mandelbrot set is self-similar around such a point (its geometric image when zoomed in on that point resembles the initial shape of Figure 1).

Julia Sets (see [2]) Quadratic Julia sets are generated by the mapping $z \mapsto z^2 + c$, for fixed c. The filled-in Julia set is the set of all complex numbers z for which the sequence defined by the iterations of the mapping does not approach infinity. The definition of a Julia set generalises to rational functions, but in the case in which we are interested, it is a very nice property that the Mandelbrot set is the set of all points c for which the corresponding Julia set is connected.

Relation to the logistic map There exists an amazing correspondence between points of the antenna of the Mandelbrot set (on the boundary of the so-called "Mandelbrot bulbs") and the bifurcation diagram of the logistic map, as seen in Figure 3.

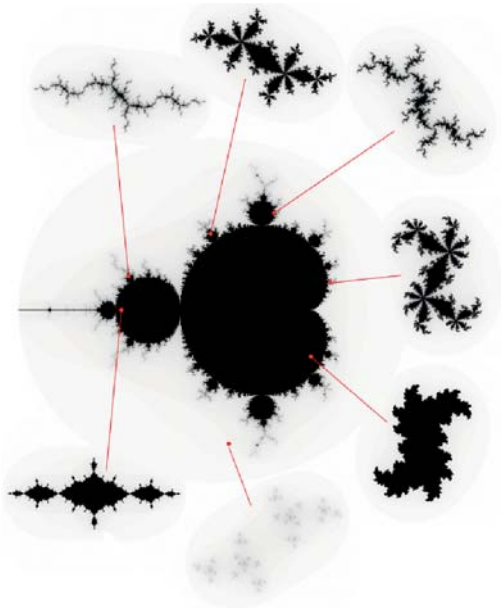


Figure 2 The Julia set

A Challenge: Open Problem

A number c is called hyperbolic if the critical point of the map $f = z^2 + c$ is attracted to a periodic cycle (i.e. the sequence b_n converges to a periodic cycle in our case). Is the number $c = -1.99999999$ hyperbolic (see[4] p.14)? The point is that no computer calculation is reliable and thus an answer to this question carries a high scientific interest.

As an epilogue, it suffices to say that the field of research in fractals – chaotic dynamical systems – is extremely active and we know much less about it than we would like to!

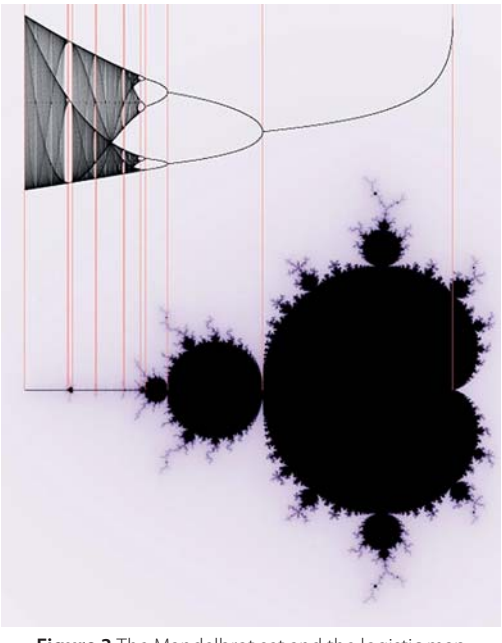


Figure 3 The Mandelbrot set and the logistic map

Acknowedgements

I would like to thank Stergios Antonakoudis, former student of Trinity College, for introducing me to the topic, and both he and Yanitsa Pehova for their helpful comments/corrections.

References

[1] Alan F. Beardon; 2000 *Iteration of Rational functions*, Springer.

[2] John W. Milnor; *Dynamics in one Complex Variable: Introductory Lectures*; Annals of Mathematics; http://arxiv.org/abs/math/9201272.

[3] Curtis T. McMullen; 1994; *Frontiers in Complex Dynamics*; Princeton University Press; http://math.harvard.edu/~ctm/papers/home/text/papers/front/front.pdf.

[4] *Fatou's theorem*; http://www.ibiblio.org/e-notes/MSet/inside.html.

Julia Set

This picture is derived from the Julia set for the function $\exp(z^3) - 0.621$.

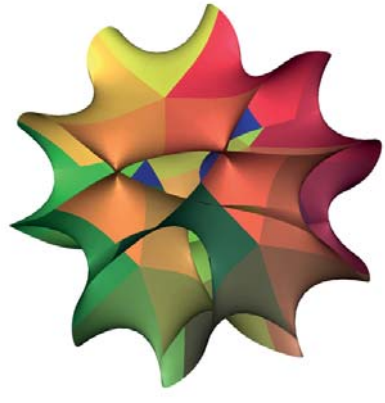
Geometry Through the Eyes of Physics

Prof David Tong
Professor of Theoretical Physics, DAMTP

It's no secret that there is a close connection between geometry and physics. Probably the most famous example is the theory of General Relativity, in which the force of gravity is recast in terms of the geometry of space and time. The purpose of this article, however, is not to wax poetic about geometry in Nature. Instead, I'd like to describe how things work the other way round, when Nature gets into geometry. I will try to explain how we can use ideas from physics to give new insight into mathematics.

To tell the story, we'll need two simple ideas: one from maths and one from physics. From maths, the main character is a *manifold*. If you haven't heard of this before, then you should have in the back of your mind a curved, closed surface, like that of a sphere or a torus. A manifold is a generalisation of this shape to higher dimensions. The purpose of geometry is to understand the properties of different manifolds, the relationships between them and the language we need to describe them. Meanwhile, from physics, the only object that we'll need to begin with is the humble particle. Our plan is as follows: we'll place the particle on the manifold and let it roam around. By understanding the behaviour of the particle, we'll try to infer various properties of the underlying space.

To start, we'll think about a particle obeying the laws of classical mechanics. Here there are few surprises and the particle does exactly what you would expect: it rolls around, guided by the contours of the space. The path it takes has some special mathematical properties and is called a *geodesic*. But the particle is too limited to know anything very deep about the underlying manifold. Its perspective is too parochial; it knows only about the small region in its immediate neighbourhood and has little to tell us about the global properties of the manifold.



A Calabi-Yau manifold
Andrew J Hanson, Indiana University

Geometry and Quantum Mechanics

Things get more interesting when we turn to quantum mechanics. In the quantum world, the particle no longer has a definite position. Instead, things are more uncertain and we have to talk in the language of probabilities. The mathematical description of a quantum particle is in terms of a wavefunction, $\psi(\mathbf{x})$. This is a complex valued function, with $\mathbf{x}$ a set of coordinates which label points on the manifold. The probability of finding a particle at the point $\mathbf{x}$ is proportional to $|\psi(\mathbf{x})|^2$.

The fact that the quantum particle spreads out in a wave of uncertainty gives it more power. It can feel its way all over the manifold. It knows about the global structure of the space. The state of the particle is described by the Schrödinger equation

$$-\nabla^2\psi = E\psi \qquad (1)$$

You've probably seen the symbol ∇^2 before. It's called the Laplacian. Roughly speaking, it means that you should differentiate ψ twice with respect to every coordinate that it depends upon. The first time you see the Laplacian is usually in the context of flat $\mathbb{R}^3$, where $\mathbf{x} = (x,y,z)$ and

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

There is an obvious generalisation of this to different dimensions. But, most importantly, there is also a generalisation of the Laplacian to manifolds that are curved. In this case, the Laplacian depends on the *metric* on the manifold which means that the symbol contains within it information about the distances between different points on the manifold.

The E in equation (1) is just a real number. Physicists would identify it with the energy of the particle. The key idea is that the Schrödinger equation doesn't admit solutions $\psi(\mathbf{x})$ for any value of E. Instead, there are only solutions when the energy E takes certain, discrete values. Moreover, because ∇^2 depends on the underlying space, so too does the list of allowed energies. This provides a very different way of thinking about geometry. You give me a manifold and specify its shape and curvature (or, more precisely, its topology and metric). With that information, I solve the Schrödinger equation and hand you back a list of numbers E. That list of numbers is called the spectrum of the Laplacian and it contains, encoded with it, much of the information about the manifold. This way of thinking is called *spectral geometry*.

There is a more down-to-earth version of spectral geometry, made famous by the mathematician Mark Kac in an article called *"Can One Hear the Shape of a Drum?"*. The frequencies at which a drum beats are again governed by the equation (1), now with particular boundary conditions imposed by the shape of the rim of the drum. The question is: if you know all the frequencies, can you figure out the shape? The answer, it turns out, is no, but you can extract a lot of information. Similarly, it is known in geometry that the spectrum is not necessarily sufficient to determine uniquely the underlying manifold. Nonetheless, the study of spectral geometry is a rich subject, with different properties of the manifold encoded in the spectrum in interesting ways.

It will be useful to work through a (very) simple example of spectral geometry: the one-dimensional circle. We will label the position along the circle by the coordinate x. If the circle has radius R, we should identify $x \equiv x + 2\pi R$. The Schrödinger equation now reads

$$-\frac{d^2\psi}{dx^2} = E\psi.$$

The solutions are simply $\psi = e^{inx/R}$. The information that the space is a circle arises through the requirement that ψ is single valued, so that $\psi(x) = \psi(x + 2\pi R)$. This tells us that we must have $n \in \mathbb{Z}$. The spectrum of the circle is therefore just a tower of numbers

$$E = \frac{n^2}{R^2}, \qquad n \in \mathbb{Z}$$

We'll return to this shortly.

Although I introduced spectral geometry by thinking about quantum physics, the subject wasn't discovered by physicists. Nonetheless, it's pleasing that it sits so naturally in the framework of quantum mechanics and there are many further related connections between the two subjects. For example, a more complicated quantum mechanical Hamiltonian which has a property called *supersymmetry* naturally captures the de Rahm or Dolbeault cohomology of the manifold. In this way, many of the great results from differential geometry can be recast in the language of quantum mechanics. However, rather than exploring these directions here, I would instead like to tell you about something novel and surprising that came out of thinking about geometry in the language of physics.

Geometry and String Theory

String theory is currently the best guess that we have for a unified theory of gravity and quantum mechanics. The basic idea is, on the face of it, slightly daft: string theory postulates that, at the fundamental level, if you look deep inside every particle, you will see a tiny vibrating loop of string. At the moment there is no experimental evidence for string theory. Nonetheless, it is a powerful

mathematical framework. Here we're going to bring that framework to bear on questions in geometry. We use the same strategy that we've seen above and ask: what is the energy spectrum of a string moving on a manifold?

Let's return to our example of the circle. Now there are two different things that the string can do. First, the string can form a little loop which then moves around the circle. Because, from afar, this loop of string looks like a particle, it shouldn't be too much of a surprise to learn that the energy spectrum is identical to that of a particle: $E = n^2/R^2$ with $n \in \mathbb{Z}$. But the string can also do something that the particle can't: it can stretch itself all around the circle. You can think of the string as an elastic band; stretching it costs energy and a string which winds m times around the circle has energy $E = (2\pi mR)^2$, with $m \in \mathbb{Z}$. This means that the energy spectrum of a string moving on a circle consists of two towers of numbers

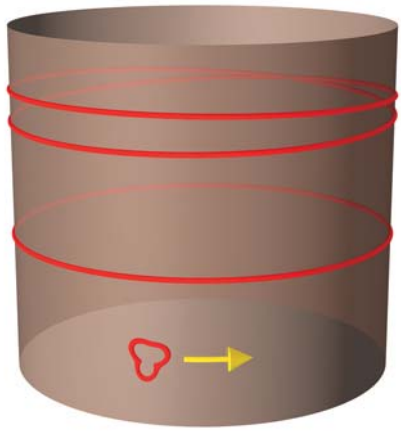
$$E = \frac{n^2}{R^2} + 4\pi^2 m^2 R^2, \quad n, m \in \mathbb{Z}.$$

But there's something interesting here. This set of numbers remains the same if we swap

$$R \longleftrightarrow \frac{1}{2\pi R}. \qquad (2)$$

This means that, if all you're given is this list of numbers, then you can't tell the difference between very big circles of size R and very small circles of size $1/2\pi R$. As far as the string is concerned, these circles look exactly the same! Of course, we've only discussed the energy spectrum of the string but it turns out that all properties of the string remain invariant under the interchange (2). Strings really can't tell the difference between big circles and small circles. This beautiful fact has a rubbish name: it is called *T-duality*.

The confusion of strings extends to other manifolds as well. Roughly speaking, manifolds come in pairs. Although particles view these pairs very differently, to a string they look identical. (This is literally true of a special class of manifolds called *Calabi-Yau* and there is a slightly generalised version of the statement for other manifolds). But these two manifolds are not related in a simple way like the big and small circles. Instead, at first sight, the two manifolds seem to have nothing to do with each other. Typically, they don't even have the same topology (i.e. the same number of holes).



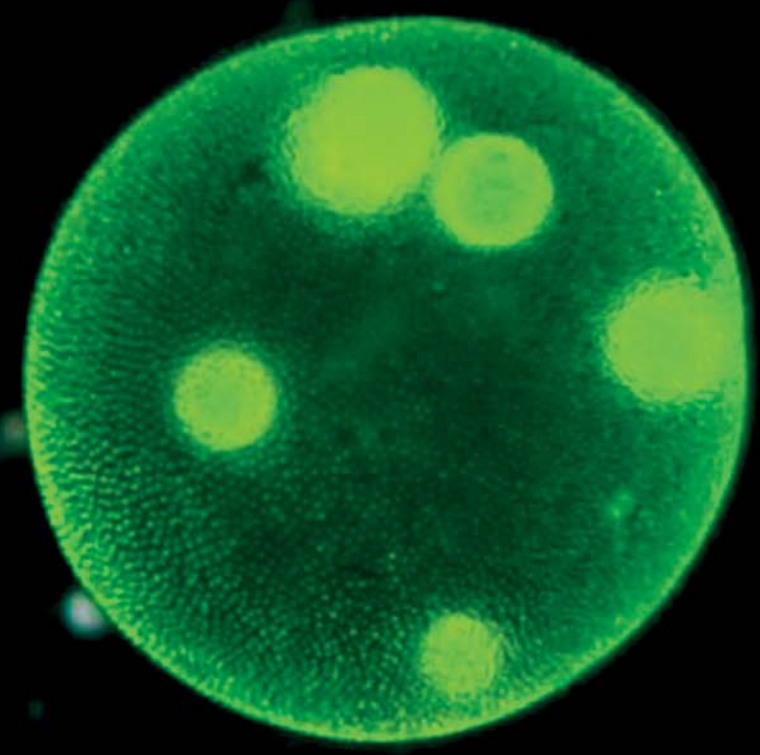
How strings can behave
Steuard Jensen, Alma College

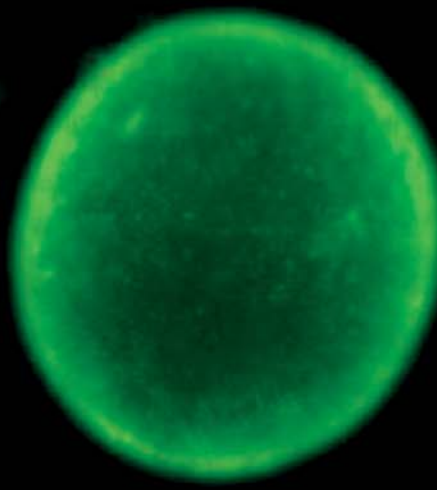
This pairing between manifolds is called *mirror symmetry*. The string's inability to distinguish between these two manifolds turns out to be a great strength. For a start, we learn that there's a very surprising and unexpected relationship between manifolds. Moreover, it turns out that mathematicians were often able to say a lot about one of these manifolds, but almost nothing about the other. Yet, according to string theory, the two manifolds should be identical; you just have to look at them in the right way. Any question that you can answer about the first manifold is telling you something interesting about the other. (Technically, questions in *complex geometry* for the first manifold are turned into questions in *symplectic geometry* for the second). Mirror symmetry then becomes a powerful tool which allows you to reinterpret properties of one manifold to provide answers to previously unsolved questions about the other.

Mirror symmetry was discovered almost 25 years ago. In the intervening time, it has become one of the most vibrant areas of research in geometry, with insight coming from both mathematicians and physicists. There is, admittedly, a difference in the style of research. Physicists tend not to be overly consumed with matters of rigour, relying instead on an intuition for how Nature should work to build conjecture upon conjecture. Mathematicians, of course, are not content until each conjecture becomes a proof. Yet this is one of an increasing number of areas in which mathematicians and physicists find themselves exploring the same questions hand in hand. It is a relationship which has enriched both communities.

Volvox

Volvox is a genus of chlorophytes, a type of green algae. It forms spherical colonies of up to 50000 cells. The colonies contain eyespots, allowing them to swim towards light.





How to Build the Perfect Igloo

Dr Andrzej Odrzywolek
Institute of Physics, Jagiellonian University

Building an igloo, or dome in general, is a task humanity has faced since antiquity. The chord lengths of geodesic domes were considered classified military information in the United States until the sixties, and some believe that the secrets of medieval cathedral dome builders form the origins of Freemasonry. Even now, the construction is not an easy procedure.

The Inuit are known for their ability to build snow domes. They build layers of bricks in a spiral pattern, causing the dome to close in loxodromically (see Figure 1). Due to the multitude of different brick shapes, this method is rather difficult for the amateur to carry out.

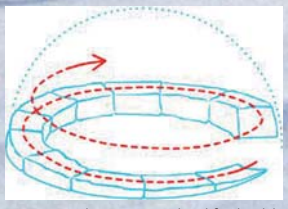


Figure 1 The Inuit method for building an igloo

Mathematical Formulation

In developing an easier process for igloo building, we are interested the following question: *is it possible to split the spherical dome into identical elements*? The answer is yes, of course. For example, we can cut the dome into n slices using lines of longitude, forming spherical triangles with two right angles at the base. Such a form would not be very useful for our purposes though. We must impose additional requirements on our block forms, with the first two considered essential, and the final two ideal:

1. We want to use as few different shapes as possible, ideally just one.

2. The volume and dimensions of the shapes should be small fractions of the total dome volume and radius.

3. The shapes should be roughly polyhedral.

4. The building procedure should be described by a simple algorithm.

Very similar requirements are found in many areas of science, for instance in the construction of grids on spheres in climate research, and in football construction.

It is well known that if three positive integers p, q, r satisfy $1/p + 1/q + 1/r > 1$, then the spherical triangle with angles $A = \pi/p$, $B = \pi/q$, $C = \pi/r$ provides a non-overlapping tiling of the sphere. Since the area of each triangle is $S = \pi(1/p + 1/q + 1/r - 1)$, the half-sphere is divided into $2\pi/S$ segments. The smallest possible such triangle has $p = 2$, $q = 3$, $r = 5$. It is a right angled triangle, which splits the half-sphere into 60 tiles. 30 of them are 'left-handed', and the remaining 30 are the mirrored counterparts of these.

Given a tessellation of the sphere, it is conceptu-

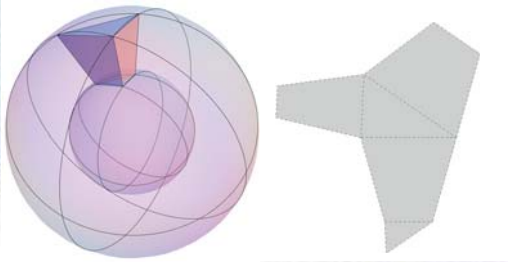


Figure 2 Conversion from spherical triangles into polyhedral dome elements (left) and the net of the dome elements (right)

ally very easy to split a spherical dome of given thickness. We draw aligned spherical triangles (polygons in general) on the inner and outer sphere, and connect them by straight line segments (see Figure 2).

The Construction Procedure

Combining the above gives us a method for constructing an igloo. To start the dome, we begin with two concentric circles (see Figure 3a). To initialise construction, we first place 12 segments in a non-trivial order (Figure 3b). Note that three elements of the same orientation are placed next to each other, on three different triangle sides. The first row has point reflection symmetry with respect to the centre of the circles. Further blocks are simply reflections of those already placed (Figures 2c and 2d). The most difficult operation is the placement of the final four elements (Figures 2e and 2f), which should ideally all be placed at once.

Paper, gypsum, wet snow and ice bricks have been used to test this procedure on small scales. The igloo has some tendency to come apart under its own weight, so a band around the base must be used.

Conclusion

The '2, 3, 5' spherical triangle above provides a working solution to the igloo building problem, requiring only two different brick forms (the two orientations). Another interesting solution is based on geodesic domes (two different equilateral triangles, 90 bricks). It is still not known whether any single small block type is sufficient to tile the hemispherical dome. Possible search areas are exceptional spherical tilings, and nearly spherical polyhedrons similar to the deltoidal icositetrahedron.

References

[1] Douglas Wilkinson; 1949; *Arctic notebook no. 1: How to build an igloo*; http://www.nfb.ca/film/how_to_build_an_igloo.

[2] Robert J. Mac, G. Dawson; 2003; *Tilings of the Sphere with Isosceles Triangles*; Disc. and Comp. Geom. 30, 467-487; http://cs.stmarys.ca/~dawson/images4.html.

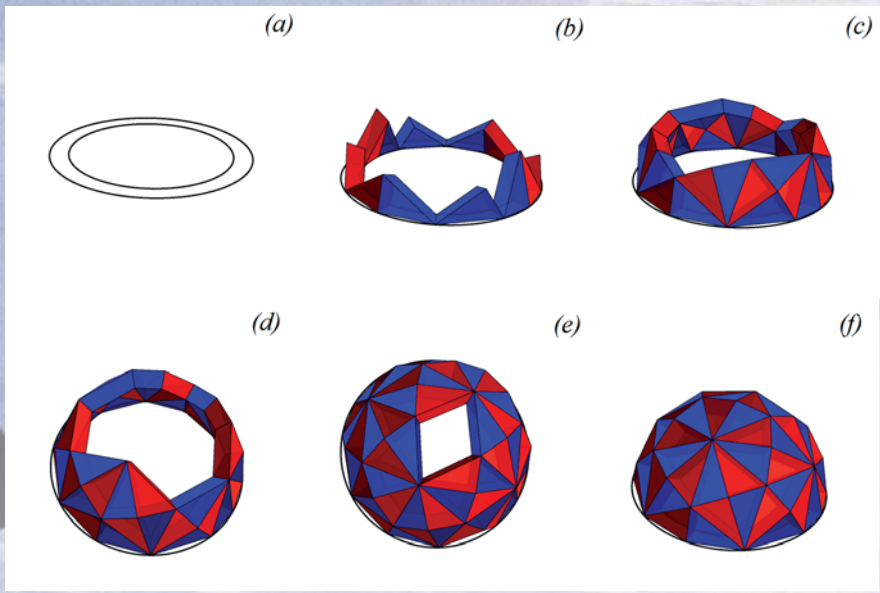


Figure 3 The construction of the igloo, left-handed and right-handed blocks coloured red and blue respectively

500 Years of Mathem

1713: 300 years have now passed since Jacob Bernoulli's book *Ars conjectandi* (The Art of Conjecture). It proves to be an extremely important work on probability. It contains the first discussion of the Bernoulli numbers.


JACOBI BERNOULLI
ARS CONJECTANDI,
TRACTATUS
DE SERIEBUS INFINITIS,
DE LUDO PILÆ RETICULARIS
BASILEÆ,

1913: 100 years ago this year, Hardy received the now infamous letter from Ramanujan, and the prolific mathematician Paul Erdős was born. In addition, Bohr's first quantum model of the atom was written down.

1613 1713 1813 1913

1663: 350 years ago, Isaac Barrow became the first Lucasian Professor of Mathematics at the University of Cambridge.

1738: The 275th anniversary of the publishing of Daniel Bernoulli's *Hydrodynamica* has now passed. It gave for the first time the correct analysis of water flowing from a hole in a container, whilst also providing the basis of the kinetic theory of gases.

1010000

atical Anniversaries

1963: 2013 sees the 50th anniversary of Paul Cohen demonstrating that neither the continuum hypothesis nor the axiom of choice can be proven from the standard axioms of set theory. Furthermore, Edward Norton Lorenz published solutions for a simplified mathematical model of atmospheric turbulence – generally known as the Lorenz Attractor or the Butterfly Effect.

2013: This year we witnessed the Nobel Prize for Physics being awarded to Higgs for his work on the Higgs Boson. Additionally, the whole of 2013 was designated the International Year of the Statistic.

1938 1963 1988 2013

1938: 75 years since, Kolmogorov publishes Analytic Methods in Probability Theory which lays the foundations of the theory of Markov random processes.

2003: 10 years ago, Grigori Perelman proved the Poincaré conjecture. Perelman was later offered a Field's Medal for his proof of this Millennium Maths Problem. However, he chose to decline the award.

1010001

The Mathematics of Pointless

Prof Yigal Gerchak
Professor of Industiral Engineering, Tel Aviv University

One of the most popular BBC TV programs in the UK, which also spawned a popular board game, is called Pointless. In this game, a team is given a category which consists of several items (e.g. "What are the Canadian Provinces?"), and is asked to name one item. If it is correct, they are awarded points, the number of which decreases with the "obscurity" of the item; if it is incorrect they score a large number of points. Each team's aim is to minimise their total score. The level of obscurity is determined by a questionnaire administered to audience members ahead of time, asking to list all items belonging to this category that they know. Items which are listed by x% of the audience are worth x points (or a function thereof) for the team selecting this (single) item later.

The game is therefore based on the trade-off between selecting a high probability item, with a (likely) high score, and selecting a low probability item, with a (likely) low score. Of course, the teams do not know exactly how well known various items are. The game is related to other probability-selecting games of the type discussed in [1].

A Risk-Neutral Team

We shall formulate the problem as maximization (rather than the minimisation format of Pointless). Let c be the penalty for a wrong answer. Let b be the highest number of points possible (corresponding to an item which no-one in the audience listed). Let x be the subjective probability that an item is correct. It becomes the expected number of audience members listing the item.

The expected number of points earned by selecting an item with subjective probability x of being correct is

$$P(x) \equiv \mathbb{E}_Y(x(b - x^a Y)) + (1 - x)(-c). \quad .$$

The term x^a, $a > 0$, embodies the contest's rule about how more and more popular choices will be penalized – for instance, Pointless has $a = 1$. Y is a multiplicative noise term with $\mathbb{E}(Y) = \mu$, reflecting the variation in the audience popularity of an item with its subjective probability. Then

$$P(x) = -c + (b - c)x - \mu x^{a+1}.$$

Note that only the mean of Y matters here. Elementary calculus shows that

$$x^* = [\tfrac{b+c}{\mu(a+1)}]^{1/a},$$

where x^* is the optimal subjective probability item to choose in this model.

In order to have $P(x^*) > 0$, so that playing optimally is beneficial (playing not optimally might not be), we need

$$[\tfrac{b+c}{\mu(a+1)}]^{1/a} > [\tfrac{(a+1)c}{a(b+c)}]^{1/a}$$

(clearly holds for small c). If $c = b$, we need $b > \mu$.

We see that x^* increases in c (a higher error penalty induces more caution, which makes sense). It is increasing in b and decreasing in μ.

Sequential Guessing

If, on the other hand, the teams guess sequentially, and the first teams attained v points, the second (us) wishes to maximise the probability of attaining more than v points.

This is equal to

$$\mathbb{P}_Y[-c + (b-c)x - Y^* x^{a+1} > v]$$
$$= F_Y[\tfrac{(b+c)x-c-v}{x^{a+1}}]$$

Elementary calculus gives

$$x^{**} = \frac{(a+1)(c+v)}{a(b+c)},$$

where x^{**} is the optimal subjective item probability to choose in this model.

Thus $x^{**} > x^*$ iff

$$v > \frac{a(b+c)}{a+1}[\frac{b+c}{\mu(a+1)}]^{1/a} - c.$$

An Alternative Approach

Suppose now that the team selects an item whose audience popularity they estimate as v.

If the item is correct, they will obtain a prize with a utility of $u(1-v)$; otherwise they will obtain $-c$. For such an item, let the probability of being correct be denoted by $p(v)$. We assume that p', $u' > 0$ and p'', $u'' \leq 0$. That implies that the function $u(1-v)$ is concave in v. So our problem is to find

$$\max_v[p(v)u(1-v) + (1-p(v))(-c)].$$

A comparison with the previous model reveals that the current model is more general, allowing $u(1-v)$ to be an arbitrary function, and the probability $p(v)$ to be general.

References and Notes

[1] Yigal Gerchak, Marc Kilgour; 2011; *Does Competition Improve Performance*; to appear in International Game Theory Review.
[2] BBC One; *Pointless*; bbc.co.uk/programmes/boorhg2r.

Thanks to Endemol for permission to use the Pointless concept and logo.

On Computable Functions

Marc Khoury
Computer Science PhD Student, The Ohio State University

What does it mean to be computable? A function is computable if for a given input its output can be calculated by a finite mechanical procedure. But can we pin this idea down with rigorous mathematics?

In 1928, David Hilbert (see [4]) proposed his famous Entscheidungsproblem, which asks if there is a general procedure for showing that a statement is provable from a given set of axioms. To solve this problem mathematicians first needed to define what it meant to be computable. The first attempt was through primitive recursive functions and was a combined effort by many researchers, including Kurt Gödel, Alonzo Church, Stephen Kleene, Wilhelm Ackermann, John Rosser, and Rózsa Péter.

Recursive Functions

Primitive recursive functions are defined as a recursive type, starting with a few functions that we assume are computable, called founders, and operators that construct new functions from the founders, called constructors. The founders are the following three functions:

The constant zero function a function that always returns zero.

The successor function $S(n) = n + 1$.

The projection function proj_n^m is an m-ary function that returns the n^{th} argument

Computability theory wasn't going to get very far if these functions weren't computable. Next, we have two operations for constructing new functions from old: composition and primitive recursion.

Composition Given a primitive recursive m-ary function h and m n-ary functions $g_1, \ldots, g_m$, the function $f(\mathbf{x}) = h(g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))$ is primitive recursive.

Primitive Recursion Given primitive recursive functions g and h, the function $f(\boldsymbol{x}, 0) = g(\boldsymbol{x})$, $f(\boldsymbol{x}, y + 1) = h(\boldsymbol{x}, y, f(\boldsymbol{x}, y))$ is primitive recursive.

The set of primitive recursive functions is the set of functions constructed from our three initial functions and closed under composition and primitive recursion. Many familiar functions are primitive recursive: addition, multiplication, exponentiation, primes, max, min, and the logarithm function all fit the bill.

So are we done? Is every computable function also primitive recursive? Sadly, no: the Ackermann function ($A(m, n)$ below) would be proven in 1928 to be a counterexample.

$$A(m,n) = \begin{cases} n+1 & \text{if } m = 0 \\ A(m-1,1) & \text{if } m > 0 \text{ and } n = 0 \\ A(m-1, A(m,n-1)) & \text{if } m > 0 \text{ and } n > 0 \end{cases}$$

The Ackermann function is a total (defined for all inputs) function that is clearly computable but not primitive recursive. Indeed, in 1928 Ackermann

(see [1]) showed that his function bounds every primitive recursive function – it grows too fast to be primitive recursive.

Something was clearly wrong, but early computability theorists didn't want to abandon primitive recursive functions entirely. What came next was a rather surprising idea at the time: perhaps computable functions need not be total! This was the key that unlocked computability theory: focusing on partial functions, those that may not be defined on all possible inputs.

The reason for focusing on partial functions is to allow an unbounded search operator. That is, we want to be able to search for the least input value that satisfies a condition and simply be undefined if no such input value exists. This operation is captured by Kleene's μ-operator.

μ-operator $f(x) = (\mu y)(g(x, y) = 0)$ returns the least y such that $g(x,y) = 0$ and is undefined if no such y exists. The function $g(x, y')$ must be defined for all $y' < y$.

Taking the closure of the μ-operator with all primitive recursive functions gives a class of μ-recursive functions. In 1943, Kleene (see [5]) used his μ-operator to provide an alternative, but equivalent, definition of general recursive functions. The original definition was given by Gödel in 1934 (see [3]), based on an observation by Jacques Herbrand. It would later be shown that μ-recursive functions are the exact same class of functions defined by two competing approaches (see [6]).

λ-Calculus

Simultaneously, from 1931-1934, Church and Kleene were developing λ-calculus as an approach to computable functions. The syntax of λ-calculus defines certain expressions as valid statements, which are called λ-terms. A λ-term is built up from a collection of variables and two operators: abstraction and application.

Let's start with a collection of variables $x, y, z, \ldots$ and suppose M, N are valid λ-terms. The abstraction operator creates the term $\lambda x.M$, which is a function taking an argument x and returning M with each occurrence of x replaced with the argument. The application operator creates the term $M\ N$, which represents the application of a function M on input N.

The λ-term $\lambda x.M$ represents a function $f(x) = M$ and – like recursive functions – many familiar functions are λ-definable. The α-conversion and β-reduction are classic examples of *reductions*, which describe how λ-terms are evaluated. An α-conversion captures the notion that the name of an argument is usually immaterial. For instance $\lambda x.x$ and $\lambda y.y$ both represent the identity function and are α-equivalent. A β-reduction applies a function to its arguments. Take, as an example, the λ-term $(\lambda x.x)y$, which represents the identity function $(\lambda x.x)$ applied to the input y. Substituting the argument y for the parameter x, the result of the function is y. So we say $(\lambda x.x)y$ β-reduces to y.

In 1934 Church proposed that the term "effectively calculable" be identified with λ-definable. While Church's formalization of computability would later be shown to be equivalent to Turing's, Gödel was dissatisfied with Church's work. In fairness, Gödel was also dissatisfied with his own work! Church would go on to advocate that "effectively calculable" should be identified with general recursive functions (which Gödel still rejected). In 1936 Church (see [2]) published his work proving that the Entscheidungsproblem was undecidable: there is no general procedure for determining if a statement is provable from a given set of axioms.

Turing Machines

Meanwhile, after hearing about Hilbert's Entscheidungsproblem, a 22 year old Cambridge student named Alan Turing began working on his own solution to the problem. Turing was unaware of Church's work at the time, so his approach wasn't influenced by λ-expressions (this wasn't the first time Turing failed to perform a literature review). Instead, he envisioned an idealized human agent performing a computation, which he called a "computer". To avoid confusion with the modern definition of computer, we'll adopt the terminology of Robin Gandy and Wilfried Sieg and use the term "computor" to refer to an idealized human agent. The computor had infinite available memory called a tape, essentially an infinite strip of paper, that was divided into squares. The computor could read and write to a square, as well as move from one square to another.

Turing put several conditions on the computation that the computor could perform. The computor

could only have finitely many states (of mind) and the tape could only hold symbols from a finite alphabet. Only a finite number of squares could be observed at a time and the computor could only move to a new square that was at most some finite distance away from an observed square. He also required that any operation must depend only on the current state and the observed symbols, and that there was at most one operation that could be performed per action (his machines were deterministic).

From this, Turing would go on to define his automatic machines – which would later come to be known as Turing machines – and show the equivalence of the two formalisations. He'd then show that "effectively calculable" implied computable by his idealized human agent, which in turn implied computable by such a machine. Turing then went on to show that the Entscheidungsproblem was undecidable. Shortly before publishing his work, he learned that Church had already shown that the Entscheidungsproblem was undecidable using λ-calculus. Turing quickly submitted his work in 1936 (see [7]) – six months after Church – along with a proof demonstrating the equivalence between his machines and λ-calculus.

After reading Turing's seminal paper, Gödel was finally convinced that the correct notion of computability had been determined. It would later be shown that all three formalisations – Turing machines, μ-recursion, and λ-calculus – actually define the same class of functions. That these three approaches all yielded the same class of functions suggested that mathematicians had captured the correct notion of computation, and supported what would come to be known as the Church-Turing Thesis.

Three years later, in 1939, Turing completed his PhD at Princeton under the supervision of Church. In his thesis he'd state the following (see [8]): "We shall use the expression 'computable function' to mean a function calculable by a machine, and let 'effectively calculable' refer to the intuitive idea without particular identification with any one of these definitions."

Church-Turing Thesis Every effectively calculable function is a computable function.

Church intended for his original thesis to be taken as a definition of what is computable. Likewise, even though he never stated it, Turing had the same intention. In fact, the term "Church's Thesis" was coined by Kleene many years after Church had published his work. These days, many people take the Church-Turing Thesis as a definition of what is computable; less formally stating that a function is computable if and only if it can be computed by a Turing machine.

It's important to stress that the Church-Turing Thesis is not a definition as many believe. It does not refer to any particular formalization that we've discussed and is not a statement that can be formally proven. It is a statement about the nature of computation. Everything that is "effectively calculable", in the vague and intuitive sense, is a computable function.

References:

[1] Wilhelm Ackermann; 1928; *Zum hilbertschen aufbau der reellen zahlen*; Mathematische Annalen, 99(1): 118–133.

[2] Alonzo Church; 1936; *An unsolvable problem of elementary number theory*; American Journal of Mathematics; 58(2): 345–363.

[3] Kurt Gödel; 1934; *On Undecidable Propositions of Formal Mathematics Systems*; Institute for Advanced Study.

[4] David Hilbert; 1900; *Mathematical problems*; International Congress of Mathematicians.

[5] Stephen C. Kleene; 1943; *Recursive predicates and quantifiers*; AMS; 53(1): 41-73; http://www.jstor.org/stable/1990131.

[6] Stephen C. Kleene; 1952; *Introduction to metamathematics*; North-Holland Publishing Company.

[7] Alan M. Turing; *On computable numbers, with an application to the entscheidungsproblem*; Proceedings of the London Mathematical Society; 2(42), 1936.

[8] Alan M. Turing; *Systems of logic based on ordinals*; Proceedings of the London Mathematical Society; 2(1):161–228, 1939.

About the Author

Marc Khoury is a first year Computer Science PhD student at the University of California, Berkeley. He received his Master's degree in Mathematics from the University of Cambridge and his Bachelor's degree in Computer Science and Engineering from the Ohio State University. His research interests are in computational geometry.


This is how we order lunch:
$L^* = \arg\max_{L \in \mathcal{L}} U_{\text{CANTAB}}(X_0(L), \ldots, X_{N-1}(L))$
interested in being the N^th?
cantabcapital.com/yourfuture
CANTAB
CAPITAL PARTNERS

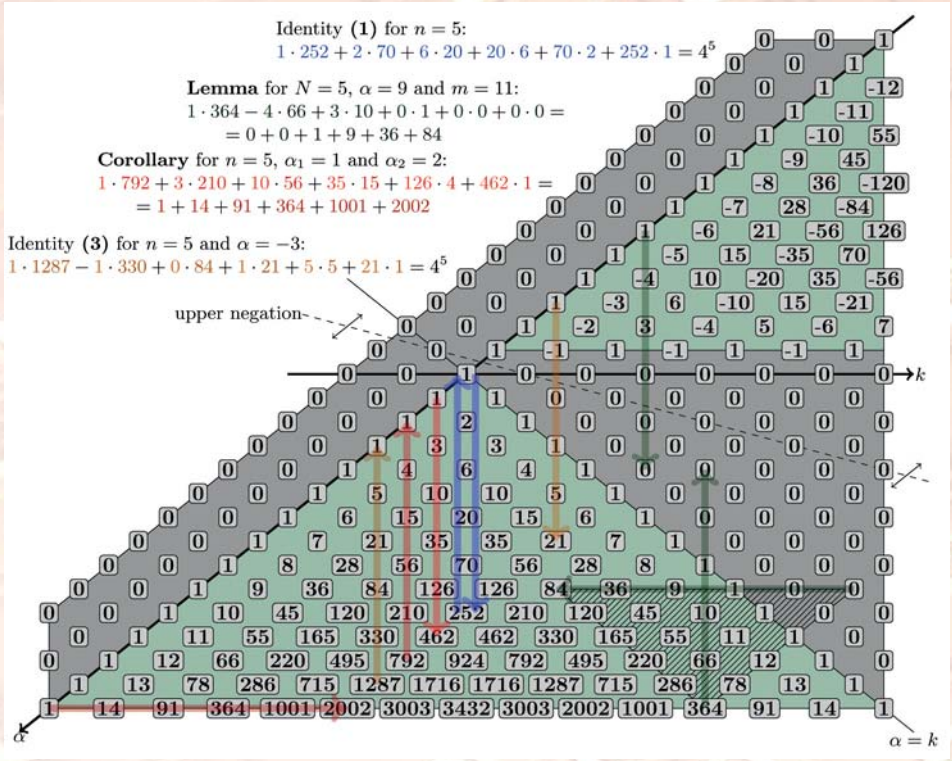


Figure 4 The extended Pascal's triangle at integer gridpoints with some identities illustrated

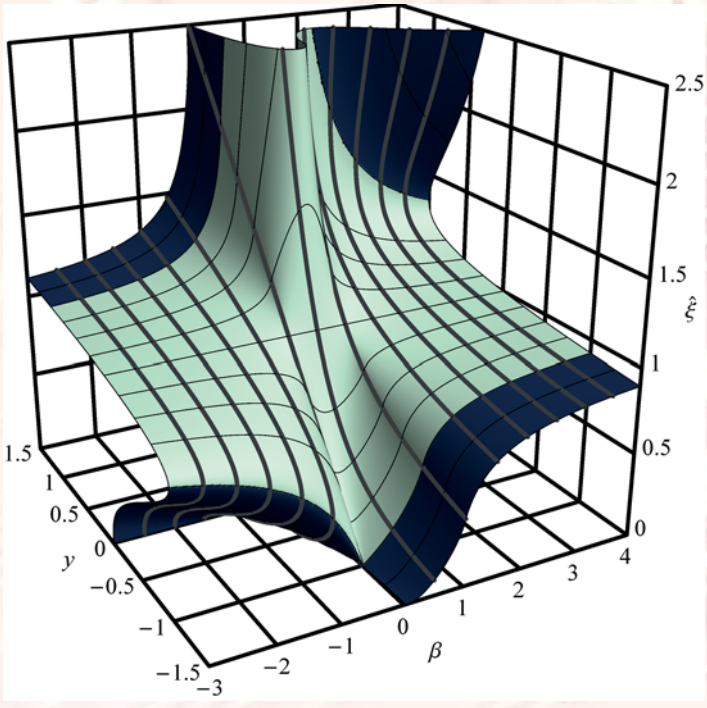


Figure 5 The underlying surface defined by $\{(\beta, y, \xi)$: $R_\beta y \xi^{t\beta} - \xi^t + 1 = 0\}$. The series of $\xi^t{}_{\beta_0}(y)$ converges on the light part and it is given by the intersection of the $\beta = \beta_0$ plane with the surface.

A Binomial Identity

Dávid Szabó
MSc Mathematics Student, Eötvös Loránd University

In 2009 I was given the following identity (cf. Fig. 4) as an exercise in a class.

$$\sum_{k=0}^{n} \binom{2k}{k}\binom{2n-2k}{n-k} = 4^n \qquad (1)$$

Throughout in this paper $n \geq 0$ is an integer. This paper summarises my investigation of this identity of that time. In four independent sections we will see several solutions and generalisations. The developed ideas will lead to a generalisation of the binomial theorem. I advise the reader to try to prove (1) before reading on.

Extending Pascal's Triangle

Considering the Taylor expansion of $(1+x)^\alpha$ about $x = 0$, it is sensible to generalise the *binomial coefficients* as $\binom{\alpha}{k} := \frac{1}{k!}\prod_{i=0}^{k-1}(\alpha - i)$ for $\alpha \in \mathbb{R}$, $k \in \mathbb{Z}_{\geq 0}$, which is a polynomial in α (the empty product and 0! is 1). With this new notation the above Taylor series gives the *binomial theorem* $(1+x)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$ for $|x| < 1$. These coefficients indeed extend the combinatorial definition for $0 \leq k \leq \alpha$ all integers, and they still satisfy $\binom{\alpha+1}{k+1} = \binom{\alpha}{k} + \binom{\alpha}{k+1}$ (*Pascal's recursion*). We extend the definition further for $k \in \mathbb{Z}_{<0}$ using this recursion formula (cf. Fig. 4). After short calculations we see that:

$$\begin{aligned}
&\tbinom{\alpha}{k} = 0 && \text{for } \alpha \in \mathbb{Z}, 0 \leq \alpha < k,\\
&\tbinom{\alpha}{k} = 0 && \text{for } k < 0,\\
&\tbinom{-1}{k} = (-1)^k && \text{for } k \geq 0,\\
&\tbinom{\alpha}{k} = (-1)^k\tbinom{k-\alpha-1}{k} && \text{the } \textit{upper negation}.
\end{aligned}$$

Combinatorial Solutions

Consider the directed graph $\vec{P}$ (cf. Fig. 1) with vertices $\left[{n \atop k}\right]$, edges $\left[{n \atop k}\right] \to \left[{n+1 \atop k}\right]$ and $\left[{n \atop k}\right] \to \left[{n+1 \atop k+1}\right]$, for all $n, k \in \mathbb{Z}$, $0 \leq k \leq n$. Say vertex $\left[{n \atop k}\right]$ is in row n and column k; call $\left[{2n \atop n}\right]$ a *central* vertex for any n.

For a directed path Γ in $\vec{P}$ with start vertex from row s and end vertex from row e we denote its part restricted to rows between and including s' and e' by $\Gamma|_{s'}^{e'}$ for $s \leq s' \leq e' \leq e$.

We denote by $\{S \xrightarrow{\mathcal{O}} E\}$ the set of directed paths with start vertex from the set S and end vertex from E satisfying some optional condition

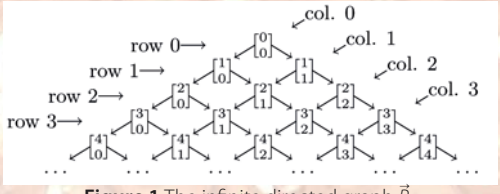


Figure 1 The infinite directed graph $\vec{P}$

$\mathcal{O}$. Here S and E will be of the form $\{\left[{n \atop k'}\right]: k' \mathrel{R} k\}$ for some binary relation R (such as $<$ or $\geq$), and we abbreviate this set as $\left[{n \atop R\,k}\right]$ and omit writing R when it is '='. Finally let $\left[{n \atop *}\right] = \left[{n \atop \geq 0}\right]$, the whole of row n.

Solution I Note that $\#\{\left[{0 \atop 0}\right] \to \left[{n \atop k}\right]\} = \binom{n}{k}$. It is natural to think of 4^n as $2^{2n} = \sum_{k=0}^{2n}\binom{2n}{k}$, i.e. $4^n = \#\{\left[{0 \atop 0}\right] \to \left[{2n \atop *}\right]\}$. Consider an arbitrary $\Gamma \in \{\left[{0 \atop 0}\right] \to \left[{2n \atop *}\right]\}$. The term $\binom{2k}{k}$ suggests that in (1) the set $\{\left[{0 \atop 0}\right] \to \left[{2n \atop *}\right]\}$ is counted in two different ways, with one counting conditioning on a (unique) k such that $\left[{2k \atop k}\right]$ is in Γ, it is natural to choose the biggest such k. Formally, let condition $\mathcal{C}$ mean that the path contains a non-start central vertex. Then Γ has a unique vertex $\left[{2k \atop k}\right]$ for some $0 \leq k \leq n$ such that $\Gamma|_{2n}^{2k}$ does not satisfy $\mathcal{C}$ (written as $\neg\mathcal{C}$), [*i.e.* Γ *meets the central vertices at* $\left[{2k \atop k}\right]$ *for the last time*]. This is well-defined as vertex $\left[{0 \atop 0}\right]$ is in Γ. This gives

$$\sum_{k=0}^{n} \#\left\{\begin{bmatrix}0\\0\end{bmatrix} \to \begin{bmatrix}2k\\k\end{bmatrix}\right\} \cdot \#\left\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\neg\mathcal{C}} \begin{bmatrix}2n\\*\end{bmatrix}\right\} = \#\left\{\begin{bmatrix}0\\0\end{bmatrix} \to \begin{bmatrix}2n\\*\end{bmatrix}\right\}$$

which is the combinatorial meaning of (1).

To show this we claim that

$$\#\left\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\neg\mathcal{C}} \begin{bmatrix}2n\\*\end{bmatrix}\right\} = \binom{2n-2k}{n-k}. \qquad (2)$$

To prove this claim, assign the number $f(i,j) := \pm\#\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\neg\mathcal{C}} \begin{bmatrix}2k+i\\k+j\end{bmatrix}\}$ to vertex $\begin{bmatrix}2k+i\\k+j\end{bmatrix}$ for $0 \le j \le i \le 2n-2k$ [*i.e. for the vertices reachable from* $\begin{bmatrix}2k\\k\end{bmatrix}$ *up to row* $2n$] where the sign is + for $j \le i/2$ [*i.e. for vertices on the left of the central vertices*], and it is – for $j > i/2$, cf. Fig. 2.

We see the boundary conditions $f(i,0) = 1$ for any i, and $f(j,j) = -1$ for $j \ne 0$. Also $f(i+1, j+1) = f(i,j) + f(i,j+1)$ for any i and j because this recursion is inherited from the paths for $j+1 \ne (i+1)/2$, and for $j+1 = (i+1)/2$ it follows from symmetry and from the choice of the sign in f. Notice that this recursion with the boundary condition determines f uniquely. The binomial coefficients (and arbitrary linear combination of their shifted copies) satisfy this recursion, and it is easy to combine them to match the boundary conditions, the solution is $f(i,j) = \binom{i-1}{j} - \binom{i-1}{j-1}$, cf. Fig. 2. Hence the LHS of (2) is

$$\#\left\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\neg\mathcal{C}} \begin{bmatrix}2n\\*\end{bmatrix}\right\} = \sum_{j=0}^{2n-2k} |f(2n-2k, j)|,$$

which is a telescopic sum where all terms but two copies of $\binom{2n-2k-1}{n-k}$ cancel, so it is indeed $\binom{2n-2k}{n-k}$ as claimed. □

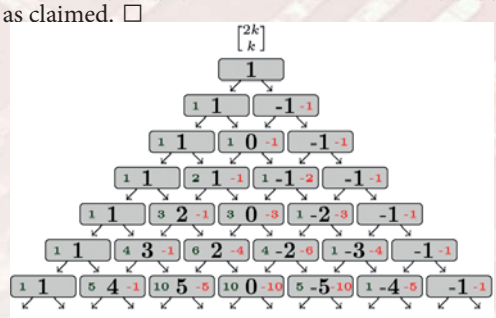


Figure 2 Smaller numbers on the sides are constant multiples of Pascal's triangle so that their sum is $f(i,j)$, the value in the centre.

The idea of this solution can be interpreted more elegantly purely combinatorially.

Solution II We will finish Solution I by showing $\#\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\neg\mathcal{C}} \begin{bmatrix}2n\\*\end{bmatrix}\} = \#\{\begin{bmatrix}2k\\k\end{bmatrix} \to \begin{bmatrix}2n\\n\end{bmatrix}\}$, from which (2) follows. We will do so by exhibiting a bijection

$$\varphi\colon \left\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\mathcal{C}} \begin{bmatrix}2n\\*\end{bmatrix}\right\} \longrightarrow \left\{\begin{bmatrix}2k\\k\end{bmatrix} \to \begin{bmatrix}2n\\\ne n\end{bmatrix}\right\}$$

between the complements in $\{\begin{bmatrix}2k\\k\end{bmatrix} \to \begin{bmatrix}2n\\*\end{bmatrix}\}$. Note that $\mathcal{C}$ is absent in the image.

For $k = n$ both the domain and the image are $\varnothing$. Else any path from the domain has a second vertex which is either $\begin{bmatrix}2k+1\\k\end{bmatrix}$ (call this condition $\mathcal{L}$) or $\begin{bmatrix}2k+1\\k+1\end{bmatrix}$ (condition $\mathcal{R}$). We also distinguish the cases in which the end vertex is in $\begin{bmatrix}2n\\\ge n\end{bmatrix}$ or in $\begin{bmatrix}2n\\<n\end{bmatrix}$. We will define φ separately on the resulting four partitions of the domain.

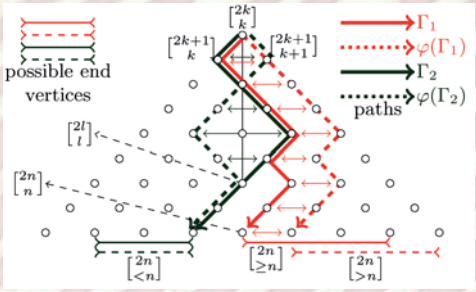


Figure 3 φ illustrated on the first two partitions

Pick $\Gamma_1 \in \{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{(\mathcal{C}\wedge)\mathcal{L}} \begin{bmatrix}2n\\\ge n\end{bmatrix}\}$ from the first partition. Now $\mathcal{C}$ is automatically satisfied as the second and end vertex are on different sides of the central vertices (in all such cases we put $\mathcal{C}$ into brackets). Define $\varphi(\Gamma_1) \in \{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\mathcal{R}} \begin{bmatrix}2n\\>n\end{bmatrix}\}$ by letting $\varphi(\Gamma_1)|_{2n}^{2k+1}$ be image of $\Gamma_1|_{2n}^{2k+1}$ under $[i_j] \to [i_{j+1}]$ [*i.e. we shift* Γ_1 *to the right by* 1 *but keep its start vertex fixed to get* $\varphi(\Gamma_1)$, *cf. Fig.* 3]. φ on this partition is a bijection. Pick $\Gamma_2 \in \{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\mathcal{C}\wedge\mathcal{L}} \begin{bmatrix}2n\\<n\end{bmatrix}\}$ from the second partition. $\mathcal{C}$ holds, so there is a maximum $l \ne k$ such that $\begin{bmatrix}2l\\l\end{bmatrix}$ is in Γ_2. Define $\varphi(\Gamma_2) \in \{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{(\mathcal{C}\wedge)\mathcal{R}} \begin{bmatrix}2n\\<n\end{bmatrix}\}$ by letting $\varphi(\Gamma_2)|_{2l}^{2k}$ be the image of $\Gamma_2|_{2l}^{2k}$ under $[i_j] \to [i_{i-j}]$ and $\varphi(\Gamma_2)|_{2n}^{2l} = \Gamma_2|_{2n}^{2l}$ [*i.e. we get* $\varphi(\Gamma_2)$ *from* Γ_2 *by reflecting its initial segment ending at* $\begin{bmatrix}2l\\l\end{bmatrix}$ *to the line of the central vertices, cf. Fig.* 3]. Note that Γ_2 satisfies $\mathcal{R}$ (as $l \ne k$) and $\mathcal{C}$ automatically, so φ on this partition is also a bijection from the unique choice of l.

So φ on these two partitions is a bijection from $\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\mathcal{C}\wedge\mathcal{L}} \begin{bmatrix}2n\\*\end{bmatrix}\}$ to $\{\begin{bmatrix}2k\\k\end{bmatrix} \xrightarrow{\mathcal{R}} \begin{bmatrix}2n\\\ne n\end{bmatrix}\}$. Analogously defining φ on the other two partitions gives the same result but with $\mathcal{L}$ and $\mathcal{R}$ swapped. Thus φ is indeed as claimed. □

Algebraic Solutions

Solution III Note that in $(2n)!$ we can separate even and odd numbers, hence observe that

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = \frac{1}{(n!)^2}\prod_{k=1}^{2}\prod_{i=0}^{n-1}\frac{-2}{-2}(k+2i)$$
$$= (-2)^{2n}\binom{-\frac{1}{2}}{n}\binom{-1}{n} = (-4)^n\binom{-\frac{1}{2}}{n}.$$

Using this observation we see that (1) is a special case of the identity $\sum_{k=0}^{n}\binom{\alpha_1}{k}\binom{\alpha_2}{n-k}=\binom{\alpha_1+\alpha_2}{n}$ (for $\alpha_1, \alpha \in \mathbb{R}$) when $\alpha_1=\alpha_2=-1/2$. We prove this identity in 3 steps.

For α_1, $\alpha_2 \in \mathbb{Z}_{\geq n}$, it is just $\sum_{k=0}^{n}\#\{\binom{0}{0}\to\binom{\alpha_1}{k}\}$. $\#\{\binom{\alpha_1}{k}\to\binom{\alpha_1+\alpha_2}{n}\}=\#\{\binom{0}{0}\to\binom{\alpha_1+\alpha_2}{n}\}$.

Next fix $\alpha_1 \in \mathbb{Z}_{\geq n}$. Now we have a polynomial in α_2 at both sides that coincide for $\alpha_2 \in \mathbb{Z}_{\geq n}$, so they must be the same, thus identity holds for $\alpha_2 \in \mathbb{R}$.

Finally fix $\alpha_2 \in \mathbb{R}$. Now the polynomials in α_1 coincide at $\alpha_1 \in \mathbb{Z}_{\geq n}$, so the identity is true for $\alpha_1 \in \mathbb{R}$ as required to finish the solution. □

Noting that the identity in this solution simply expresses the coefficient of x^n in the binomial expansion of $(1+x)^{\alpha_1}(1+x)^{\alpha_2}=(1+x)^{\alpha_1+\alpha_2}$, we can give a short version of this solution restricting to the $\alpha_1=\alpha_2=-1/2$ case only.

Solution IV Let $|4x|<1$, consider the square of the generating function $\sum_{n=0}^{\infty}\binom{2n}{n}x^n$, use the $\binom{2n}{n}=(-4)^n\binom{-1/2}{n}$ observation and the binomial theorem twice with $\alpha=-1/2$ and then with $\alpha=-1$ to prove (1) by comparing coefficients.

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n}\binom{2k}{k}\binom{2n-2k}{n-k}x^n=\left[\sum_{n=0}^{\infty}\binom{2n}{n}x^n\right]^2$$

$$\left[\sum_{n=0}^{\infty}\binom{-\frac{1}{2}}{n}(-4x)^n\right]^2=\left[(1-4x)^{-\frac{1}{2}}\right]^2=$$

$$(1-4x)^{-1}=\sum^{\infty}4^nx^n \quad \square$$

Generalisation

For $\alpha \in \mathbb{R}$, $k \in \mathbb{Z}$ call the real numbers $\left\langle{\alpha \atop k}\right\rangle$ *Pascal coefficients* if they satisfy $\left\langle{\alpha+1 \atop k+1}\right\rangle=\left\langle{\alpha \atop k}\right\rangle+\left\langle{\alpha+1 \atop k}\right\rangle$ (Pascal's recursion), i.e. binomial coefficients without specified boundary conditions.

In the combinatorial proofs, one key observation was to consider the sum of row $2n$ in the classical Pascal's triangle, now we will consider the arbitrary row segment $\left\langle{\alpha \atop m-k}\right\rangle$ (for $0\leq k\leq N$) of Pascal's coefficients. Note that $\left\langle{\alpha+N-i \atop m-j}\right\rangle=\sum_{k=0}^{N}\binom{N-i}{k-j}\left\langle{\alpha \atop m-k}\right\rangle$ for $0\leq i\leq j\leq N$, so this row segment determines the triangle below it completely. In particular, setting $i=2k$ and $j=k$ for $0\leq k\leq N/2$ (the resulting Pascal coefficients have similar form as $\binom{2n-2k}{n-k}$) in (1) in terms of k) we will have exactly enough independent equations to invert this linear system to express the sum of the row segment as a weighted sum of these resulting Pascal's coefficients as follows (cf. Fig. 4).

Lemma For $\alpha \in \mathbb{R}$, $m \in \mathbb{Z}$ and $N \in \mathbb{Z}_{\geq 0}$

$$\sum_{k=0}^{N}\binom{-N-1+2k}{k}\left\langle{\alpha+N-2k \atop m-k}\right\rangle=\sum_{k=0}^{N}\left\langle{\alpha \atop m-k}\right\rangle.$$

Proof (Split–Induce–Merge) Instead of the described method, use induction on N. The statement is true for $N=0, 1$. Let $N>1$.

Split the RHS of the statement as

$$\sum_{k=0}^{N}\left\langle{\alpha \atop m-k}\right\rangle=\sum_{k=0}^{N-1}\left\langle{\alpha+1 \atop m-k}\right\rangle-\sum_{k'=0}^{N-2}\left\langle{\alpha \atop m-1-k'}\right\rangle$$

using Pascal's recursion.
Now use *induction* on the two new summations and relabel the running variable $k'=k+1$ to get

$$\sum_{k=0}^{N}\left\langle{\alpha \atop m-k}\right\rangle=\sum_{k=0}^{N-1}\binom{-N+2k}{k}\left\langle{\alpha+N-2k \atop m-k}\right\rangle-$$
$$-\sum_{k=1}^{N-1}\binom{-N-1+2k}{k-1}\left\langle{\alpha+N-2k \atop m-k}\right\rangle$$

Notice that the binomial weight in the last summation for $k=0$ is 0, so we can add this term to that summation.

Collect $\left\langle{\alpha+N-2k \atop m-k}\right\rangle$ from the RHS and *merge* the difference of the binomial coefficients into one using Pascal's recursion. We are done after noting that for $k=N$ the resulting binomial coefficient is 0 (as $N\neq 0$). □

Corollary For $\alpha_1, \alpha_2 \in \mathbb{R}$

$$\sum_{k=0}^{n}\binom{\alpha_1+2k}{k}\binom{\alpha_2+2n-2k}{n-k}=$$
$$=\sum_{k=0}^{n}\binom{\alpha_1+\alpha_2+2n+1}{k}.$$

Proof Notice $\binom{-N-1+2k}{k}=0$ when $0\leq -N-1+2k<k$ (equivalently when $\frac{N+1}{2}\leq k\leq N$), so choosing N such that $\frac{N-1}{2}\leq n\leq N$ (denoted as †), we can let $0\leq k\leq n$ in the summation on the LHS in the lemma.

To match the corresponding terms, we apply the lemma with $\alpha:=\alpha_1+\alpha_2+2n+1 \in \mathbb{R}$, $m:=n \in \mathbb{Z}_{\geq 0}$ and $N:=-\alpha_1-1 \in \mathbb{Z}_{\geq 0}$

$$\sum_{k=0}^{n}\binom{\alpha_1+2k}{k}\left\langle{\alpha_2+2n-2k \atop n-k}\right\rangle=\sum_{k=0}^{N}\left\langle{\alpha_1+\alpha_2+2n+1 \atop n-k}\right\rangle$$

Let the Pascal coefficients be binomial coefficients in the natural way. Now the binomial coefficients on the RHS are 0 for $n-k<0$, but $n \le N$ from (†), so we can let $0 \le k \le n$ in this summation as well. (†) expressed with the new parameters is $-2n-2 \le \alpha_1 \le -n-1$, so we showed the statement for $n+2$ different integer α_1's (after reversing the order of the summation of the RHS). But in the statement there are two degree n polynomials in α_1, so they are the same. □

A notable special case is $\alpha_1 = -\alpha_2 =: \alpha \in \mathbb{R}$ (cf. Fig. 4). Now the RHS (being the half of row $2n+1$) simplifies to 4^n and we get

$$\sum_{k=0}^{n} \binom{\alpha+2k}{k} \binom{-\alpha+2n-2k}{n-k} = 4^n. \quad (3)$$

Binomial Series

This section sketches a deeper result giving a good insight to our main problem, (1).

The above corollary states that the polynomial in variables α_1, α_2 of the LHS is in fact a polynomial in the single variable $\alpha_1+\alpha_2$. This generalises to arbitrary linear terms in the binomial coefficients (see exercise 2) and is the main step in proving the following theorem.

As in Solution IV, for $\alpha, \beta \in \mathbb{R}$ define the generating function $\mathcal{G}_{\alpha,\beta}(x) := \sum_{n=0}^{\infty} \binom{\alpha+\beta n}{n} x^n$ which is convergent if $|x| < R_\beta := |\beta-1|^{\beta-1}/|\beta|^\beta$ (with $0^0 := 1$) independently of α.

Theorem For $\alpha, \beta \in \mathbb{R}$ and $|x| < R_\beta$ there is a function from $(-R_\beta, R_\beta)$ to $\mathbb{R}_{>0}$ given by $\xi_\beta(x) = 1 + \sum_{n=1}^{\infty} \frac{1}{n}\binom{\beta n}{n-1} x^n$ such that

$$\mathcal{G}_{\alpha,\beta} = \mathcal{G}_{0,\beta} \cdot (\xi_\beta)^\alpha = \frac{(\xi_\beta)^\alpha}{1-\beta(1-1/\xi_\beta)}.$$

This ξ_β also satisfies $x\xi_\beta(x)^\beta - \xi_\beta(x) + 1 = 0$ and $\xi_\beta(x)\,\xi_{1-\beta}(-x) = 1$.

Note that the exponential behaviour of $\mathcal{G}_{\alpha,0}$ (cf. the binomial theorem) remains true for any β, explaining (1) in depth.

The additional properties of ξ_β can be used to determine its closed form (and hence that of $\mathcal{G}_{\alpha,\beta}$) for some special values of β (and hence for $1-\beta$). It happens to be simpler to consider the scaled $\hat{\xi}_\beta(y) := \xi_\beta(R_\beta y)$ giving

$$\hat{\xi}_{-2}(\sinh^2 z) = \frac{\sinh z}{3\sinh(z/3)} \qquad \hat{\xi}_{-1}(y) = \frac{1+\sqrt{1+y}}{2}$$
$$\hat{\xi}_{-\frac{1}{2}}(\sin\theta) = \tfrac{4}{3}\cos^2\left(\tfrac{\pi}{6} - \tfrac{\theta}{3}\right) \qquad \hat{\xi}_0(y) = 1+y$$
$$\hat{\xi}_{\frac{1}{2}}(y) = (\sqrt{1+y^2}+y)^2.$$

The studied $\beta = 2$ case gives $\xi_2(x) = \frac{2}{1+\sqrt{1-4x}}$ and so

$$\sum_{n=0}^{\infty} \binom{\alpha+2n}{n} x^n = \frac{\xi_2(x)^\alpha}{\sqrt{1-4x}}.$$

From this, the crucial $\binom{2n}{n} = (-4)^n\binom{-\frac{1}{2}}{n}$ observation in the algebraic solutions simply follows from $\mathcal{G}_{0,2}(x) = \frac{1}{\sqrt{1-4x}} = \xi_0(-4x)^{-1/2} = \mathcal{G}_{-\frac{1}{2},0}(-4x)$. Finally we get another proof for generalisation (3) as a corollary of the theorem:

$$\sum_{n=0}^{\infty}\sum_{k=0}^{n} \binom{\alpha+2k}{k}\binom{-\alpha+2n-2k}{n-k} x^n = \frac{\xi_2(x)^\alpha}{\sqrt{1-4x}} \cdot \frac{\xi_2(x)^{-\alpha}}{\sqrt{1-4x}} = \sum_{n=0}^{\infty} 4^n x^n. \quad \square$$

We conclude that our main identity (1) is just the equality of two series expansions of $\mathcal{G}_{\alpha,\beta}(x)^2$ when $\alpha = 0$, $\beta = 2$, so using the closed forms for some β's above we can deduce similar identities.

Exercises

1. If $N > 0$ and $(-1)^k W_{N,k} = \binom{N-k}{k} + \binom{N-k-1}{k-1}$, then

$$x^N + y^N = \sum_{0 \le k \le N/2} W_{N,k} \cdot (x+y)^{N-2k}(xy)^k$$

[proving the fundamental theorem of symmetric polynomials in 2 variables constructively].

2. For fixed n and $\beta \in \mathbb{R}$

$$\sum_{k=0}^{n} \binom{\alpha_1+\beta k}{k}\binom{\alpha_2+\beta(n-k)}{n-k}$$

depends only on $\alpha_1+\alpha_2$.
[*Hint*: use induction on n, for the induction step use the polynomial argument and the Split–Induce–Merge technique for $\binom{\alpha_1+\beta k}{k} = \binom{\alpha_1-1+\beta k}{k} + \binom{\alpha_1+\beta+\beta(k-1)}{k-1}$ with induction on $\alpha_1 \in \mathbb{Z}_{\ge 0}$.]

Acknowledgement

I am grateful to my teachers of Fazekas Mihály Secondary School (Budapest, Hungary): to András Hraskó for his ideas and for the useful discussions, and to László Surányi for giving me (1) as a exercise.

Contemplation, by Paul Klee

1011101

Turing Instabilities

Diana Danciu
Part III Systems Biology Student, Murray Edwards

Alan Turing, a graduate of the University of Cambridge, was a true visionary and an important figure in the Mathematics of the 20th century. Mostly known as the "father of computer science and artificial intelligence" and for devising methods to break German ciphers during World War II (including the famous Enigma machine), he also had important contributions in the field of Mathematical Biology. In his thesis "The Chemical Basis of Morphogenesis" (1952) (see [1]), Turing explains how a system that is initially homogeneous and stable can later develop spatial instabilities – the so-called "Turing instabilities" – that lead to the formation of spatial patterns such as leopards' spots, zebras' stripes and even people's fingers.

Morphogenesis is the part of embryology that studies the formation of pattern and form. The embryo, initially homogeneous, contains two types of chemicals called morphogens (an inhibitor and an activator), which can be produced by cells in the embryo (reaction) or can diffuse into each other, giving the name of the following type of equations, a reaction-diffusion system

$$\frac{\partial u}{\partial t} = f(u,v) + D_1 \nabla^2 u$$
$$\frac{\partial v}{\partial t} = g(u,v) + D_2 \nabla^2 v$$

where u and v are the concentrations of two morphogens; D_1, D_2 are the two corresponding diffusion coefficients and f, g are the two functions describing the reaction kinetics. Turing's basic idea was that if, in the absence of diffusion (i.e. $D_1 = D_2 = 0$), u and v tend to a linearly stable uniform steady state, then spatially inhomogeneous patterns can evolve by diffusion driven instabilities (if $D_1 \neq D_2$) under conditions which we shall deduce. Thus, we start from a steady-state solution (u_0, v_0) of the homogeneous system ($D_1 = D_2 = 0$). In other words, we have $f(u_0, v_0) = g(u_0, v_0) = 0$. We consider small perturbations by setting $u = u_0 + \hat{u}(t)\cos(kx)$, $v = v_0 + \hat{v}(t)\cos(kx)$ and linearizing the system about the steady state, meaning that we expand in Taylor series about (u_0, v_0) up to linear order. We then get

$$\frac{\partial \hat{u}}{\partial t}\cos(kx) =$$
$$\hat{u}(t)\cos(kx)\frac{\partial g}{\partial u}(u_0,v_0) + \hat{v}(t)\cos(kx)\frac{\partial g}{\partial v}(u_0,v_0)$$
$$\frac{\partial \hat{v}}{\partial t}\cos(kx) =$$
$$\hat{u}(t)\cos(kx)\frac{\partial g}{\partial u}(u_0,v_0) + \hat{v}(t)\cos(kx)\frac{\partial g}{\partial v}(u_0,v_0)$$

Or, in matrix notation, by cancelling the $\cos(kx)$ factors

$$\begin{pmatrix} \dot{\hat{u}} \\ \dot{\hat{v}} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{pmatrix}_{(u_0,v_0)}}_{J} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix}$$

where J is the Jacobian matrix, whose eigenvalues tell us about the stability of the system. The system is stable if both eigenvalues are real and negative, which is equivalent to having $D = det(J) = f_u g_v - f_v g_u > 0$ and $T = trace(J) = f_u + g_v < 0$, with the functions calculated in the equilibrium point (u_0, v_0). This follows from the fact that the determinant and trace of a matrix are the product and the sum of its eigenvalues, respectively. If next we return to the inhomogeneous system, i.e. we add in the diffusion terms, then (by noting that $\nabla^2 u(t)\cos(kx) = -k^2 u(t)\cos(kx)$), we get a new so-called modified Jacobian

$$J_{mod} = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix} - k^2 \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

What we need to do now is to inspect the trace and determinant of this new Jacobian and set conditions such that the system is unstable. Indeed, the trace is still less than zero, but the determinant has a more interesting form:

$$det(J_{mod}) =$$

$$D_1D_2k^4 - (D_1g_v + D_2f_u)k^2 + (f_ug_v - f_vg_u)$$

$$= Ak^4 - Bk^2 + C,$$

having it written in the form of a quadratic function in k^2. In order for the new system to be unstable, having the trace of its Jacobian negative, we also need its determinant to be negative, and thus we look at the discriminant of this quadratic.

By inspecting the properties of the quadratic functions (see also Figure 1), we find $B > 0$ to be a necessary condition (i.e. some k gives instability) and $B^2 - 4AC > 0$ to be a sufficient one (i.e. all k give instability), giving us $B > \sqrt{AC}$. In conclusion, our condition for developing Turing instabilities is

$$D_1g_v + D_2f_u > 2\sqrt{D_1D_2(f_ug_v - f_vg_u)}$$

To give some examples of the theory and its generalizations, in 2D we can get regular planar tesselation patterns such as squares, hexagons, rhombi or triangles. These are solutions of

$$\nabla^2\psi + k^2\psi = 0, \quad (\mathbf{n}\cdot\nabla)\psi = 0 \quad \text{for} \quad \mathbf{r} \in \partial B,$$

where $r = (x, y)$, ∂B is the closed boundary of the reaction-diffusion domain B, and n is the unit outward normal to ∂B. The following functions

$$\psi(x,y) = \frac{\cos k\left(\frac{\sqrt{3}y}{2} + \frac{x}{2}\right) + \cos k\left(\frac{\sqrt{3}y}{2} - \frac{x}{2}\right) + \cos kx}{3},$$

$$\text{for } k = n\pi, n = \pm1, \pm2, \ldots$$

$$\psi(x,y) = \frac{\cos kx + \cos ky}{2}, \text{ for } k = \pm1, \pm2, \ldots$$

$$\psi(x,y) = \frac{\cos kx + \cos\{k(x\cos\phi + y\sin\phi)\}}{2},$$

where ϕ is the rhombus angle, $k = \pm1, \ldots$

$$\psi(x,y) = \cos kx, \text{ with } k = n\pi,\ n = \pm1, \pm2, \ldots$$

represent, respectively, the solutions for a hexagon, a square, a rhombus and a one dimensional version of the square and can be seen in Figure 2 (see [2] pp. 90-103). The beauty of Turing's theory lies in its simplicity and in the diversity of applications it may have: in addition to animal coat patterns, we can understand bacterial movements, cartilage condensation in limb morphogenesis, embryonic fingerprint formation, wound healing, or even growth of brain tumours – for a whole range of applications refer to [2].

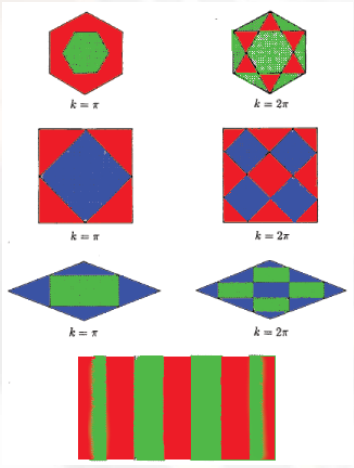


Figure 2 Patterns arising from Turing Instabilities

References

[1] Alan M. Turing; 1952; *The Chemical Basis of Morphogenesis*; Philosophical Transactions of the Royal Society of London 237 (641): 3772.

[2] James D. Murray; 2003; *Mathematical Biology II: Spatial Models and Biomedical Applications, Third Edition*; Springer.

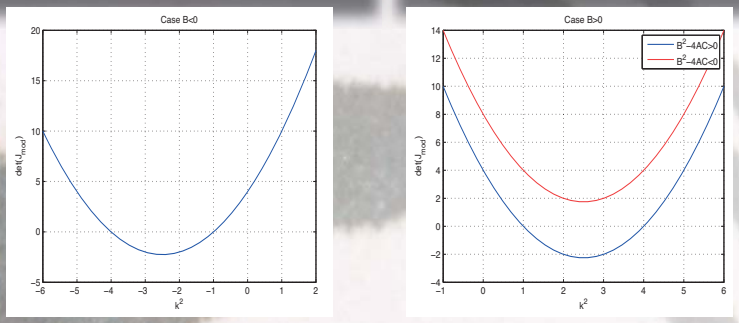


Figure 1 Plots of $det(J_{mod})$ against k^2 for various cases

A Nice Theorem in Multiplicative Functions

Masum Billal
Fourth Year Engineering Undergraduate, University of Dhaka

The theorem to be discussed in this article is a nice and powerful one involving multiplicative functions. We shall refer to it as the *Multiplicative Function Theorem* (MFT)

We define f: $\mathbb{N}\to\mathbb{N}$ to be a multiplicative function if $f(mn) = f(m)f(n)$ for any m coprime to n. Throughout this article, let f be a multiplicative function and the unique prime factorization of n be denoted by

$$n = p_1^{e_1}\cdots p_k^{e_k},$$

with $p_1, \ldots, p_k$ distinct primes.

Notation

1. $\tau(n)$ is the number of divisors of n.
2. $\omega(n)$ is the number of distinct prime factors of n.
3. For a multiplicative function f,

$$F(n) = \sum_{d|n} f(d).$$

Let's call F the *summation function* of f. If p is a prime, then we have

$$F(p^a) = \sum_{d|p^a} f(d) = \sum_{i=0}^{a} f\left(p^i\right).$$

4. For positive integers n and p, $p^\alpha \| n$, or alternatively $v_p(n) = \alpha$, means that $p^\alpha | n$ whereas $p^{\alpha+1} \nmid n$.

Our Core Theorem

Theorem (We use the notation F and f as defined above.) Let f be a multiplicative function.

If $F(n) = \sum_{d|n} f(d)$, then

$$\begin{aligned}F(n) &= (1+f(p_1)+\ldots+f(p_1^{e_1}))\ldots(1+f(p_k)+\ldots+f(p_k^{e_k}))\\ &= \prod_{i=1}^{k}\sum_{j=0}^{e_i} f\left(p_i^j\right)\\ &= \prod_{i=1}^{k} F\left(p_i^{e_i}\right).\end{aligned}$$

In other words, if f is multiplicative, then so is F.

Proof Let T be the expansion of the right side of the equation, and

$$S = \sum_{d|n} f(d)$$

If $d|n$ is a divisor of n, then $d = p_1^{w_1}\ldots p_k^{w_k}$, where $0 \le w_i \le e_i$ for $1 \le i \le k$. Then we have

$$\begin{aligned}f(d) &= f\left(p_1^{w_1}\right)\cdots f\left(p_k^{w_k}\right)\\ &= f\left(p_1^{w_1}\right)\cdots f\left(p_k^{w_k}\right),\end{aligned}$$

which is a term that is present in T. Thus, we conclude that each term of S is a term of T. Now we easily find that the converse is also true, since, after multiplying, we see that every term in T is of the form $f\left(p_1^{w_1}\right)\cdots f\left(p_k^{w_k}\right)$, which can be written as $f\left(p_1^{w_1}\ldots p_k^{w_k}\right)$ or $f(d)$. Therefore, every term of T is a term of S. Combining these two, $S = T$.

Problems

We see some applications of the theorem by solving some problems. The crucial fact deduced from the theorem is: if we can find the value of $F(p^a)$ for a positive integer a, we are done. And more satisfying, to do that we need to find the value of $f(p^a)$ only. Here are some examples. First we see a derivation of the number of divisors formula.

Example Problem 1 Find the number of divisors of n.

Solution Note that, if we set $f(n) = 1$ for all $n \in \mathbb{N}$, then f is multiplicative. Then it is obvious that,

$$\sum_{d|n} 1 = \tau(n)$$

Also, since f is multiplicative, we can invoke MFT. Using this,

$$\begin{aligned} F(p_i^{e_i}) &= \sum_{d|p_i^{e_i}} f(d) \\ &= f(1) + f(p_i) + \ldots + f(p_i^{e_i}) \\ &= 1 + \ldots + 1 = e_i + 1 \end{aligned}$$

Therefore, $\tau(n) = (e_1 + 1) \ldots (e_k + 1)$.

Example Problem 2 (Generalisation of sum of divisors) Let $\sigma(n)$ be the sum of divisors of n. Let $\sigma_r(n)$ be the sum of r^{th} powers of the divisors of n,. That is, if $\{d_1, d_2, \ldots, d_{\tau(n)}\}$ are divisors of n, then

$$\sigma_r(n) = d_1^r + \ldots + d_{\tau(n)}^r.$$

Prove that

$$\sigma_r(n) = \frac{p_1^{(e_1+1)r} - 1}{p_1 - 1} \cdots \frac{p_k^{(e_k+1)r} - 1}{p_k - 1}.$$

Solution Set $f(n) = n^r$. This is multiplicative, since $f(mn) = (mn)^r = m^r n^r = f(m)f(n)$ for any $m, n \in \mathbb{N}$ (and in particlular for m coprime to n). MFT gives

$$\sigma_r(n) = \sum_{d|n} d^r = f(d)$$

$$\begin{aligned} F(p_i) &= 1 + f(p_i) + \ldots + f(p_i^{e_i}) \\ &= 1 + p_i^r + \ldots + p_i^{e_i r} \\ &= \frac{p_i^{(e_i+1)r} - 1}{p_i - 1}. \end{aligned}$$

Therefore

$$\sigma_r(n) = \frac{p_1^{(e_1+1)r} - 1}{p_1 - 1} \cdots \frac{p_k^{(e_k+1)r} - 1}{p_k - 1}.$$

Note. The formula for the usual sum of divisors follows if we set $r = 1$.

Example Problem 3 Prove that

$$\sum_{d|n} \varphi(d) = n$$

where $\varphi(n)$ is the Euler function.

Solution

It is well known that φ is multiplicative (we don't prove it here), so we can invoke MFT here.

$$\begin{aligned} F(p) &= \sum_{i=0}^{e} \varphi(p^i) \\ &= 1 + (p-1) + p(p-1) + \ldots + p^{e-1}(p-1) \\ &= 1 + (p-1)\left(1 + p + \ldots + p^{e-1}\right) \\ &= 1 + (p-1)\left(\frac{p^e - 1}{p - 1}\right) \\ &= p^e \end{aligned}$$

Hence

$$\sum_{d|n} \varphi(d) = \prod_{p|n} p^e = n.$$

Example Problem 4 The *Möbius Function* $\mu(n)$ is defined by

$$n = \begin{cases} 1 \text{ if } n = 1 \\ (-1)^{\omega(n)} \text{ if } n \text{ is square-free} \\ 0 \text{ otherwise} \end{cases}$$

Prove that $\sum_{d|n} \mu(d) = 0$ for $n > 1$.

Solution First, note that, for a prime p, $\mu(p^a) = 0$ for $a > 1$, since it isn't square-free. Therefore

$$\begin{aligned} F(p^e) &= \mu(1) + \mu(p) + 0 + \ldots + 0 \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

since $\mu(p) = (-1)^1 = -1$. Therefore

$$\sum_{d|n} \mu(d) = \prod_{p|n} \sum_{i=0}^{e_i} \mu(p^i) = 0.$$

Exercise Prove that

$$\sum_{d|n} \mu(d) f(d) = \prod_{p|n} (1 - f(p)).$$

Exercise Prove that

$$\sum_{d|n} \tau(d) = n,$$

where $\tau(n)$ is the number of divisors of n.

The Disc Planimeter

Dr Gonzalo Gomez-Mataix
Lecturer of Civil Engineering, Miguel Hernandez University

Nowadays, when we need to know the area of an irregularly shaped figure, the easiest solution is to ask for the help of our everlasting work companion: the computer. But this was not always the case. In the last century brilliant mechanical devices provided great accuracy in the measuring of the area of irregular figures drawn over paper. The aim of this article is to show that it is possible to measure the area of an irregular figure without the need for a PC or any complex mechanical planimeter. The method put forward is a simple one (a geometrical tool) which obviates the need for other rough and obvious methods like counting squares over graded paper.

Mathematical Background

In order to achieve this aim, we firstly must consider that every planar figure limited by a closed curve may be expressed as a function $R = \rho(\theta)$ in polar coordinates, no matter where the point **O** – the origin of coordinates – is, external or internal to the curve.

Then, the area of this figure can be calculated by means of the following integral:

$$S = \frac{1}{2}\int_0^{2\pi} \rho^2(\theta)\ \mathrm{d}\theta. \qquad (1)$$

Note that $\mathrm{d}\theta$ is positive when the angle is counter clockwise, and negative when clockwise. This integral must in general be solved numerically. The rectangular method is the simplest of the eligible numerical methods, and has been chosen in this paper for this reason. In polar coordinates, the analogous approximation to that which would be performed in Cartesians by rectangles uses circular sectors of constant radius ρ_i and constant angular amplitude $\Delta\theta_i = \Delta\theta$. The integral that quantifies the area of the figure may then be approximated by the algebraic sum of the areas of the N circle sectors:

$$S \approx \frac{1}{2}\sum_{i=1}^{N} \rho_i^2 \Delta\theta_i = \\ = \frac{1}{2}|\Delta\theta| \sum_{i=1}^{N} (-1)^M \rho_i^2 = k_1 \sum_{i=1}^{N} (-1)^M \rho_i^2, \qquad (2)$$

where $N = 2\pi/\Delta\theta$ (the total number of sectors to take into account) and $M = 1$ if $\Delta\theta_i < 0$, or $M = -1$ if $\Delta\theta_i > 0$. The radius ρ_i represents the length of the intersections between the contour of the figure and the N rays equally spaced in angle that originate from **O**. A graphical explanation of this approximation can be seen in Figure 2.

In order to obtain this sum directly from the operation of a mechanical device, we need to linearise the sum – converting the squares into quantities capable of being added directly. This is possible using the Fermat spiral. In a Fermat spiral, the radius at any point is proportional to the root of the angle, as shown in Figure 1.

Consider now a double Fermat spiral, centred at the same point **O** and with the same origin of angles, defined as:

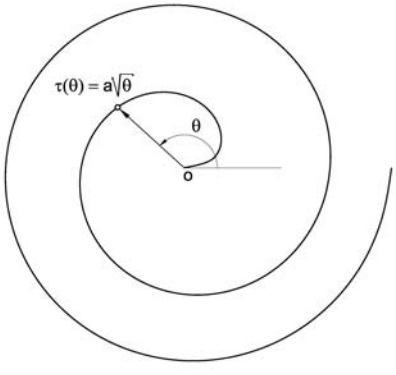


Figure 1 The Fermat spiral

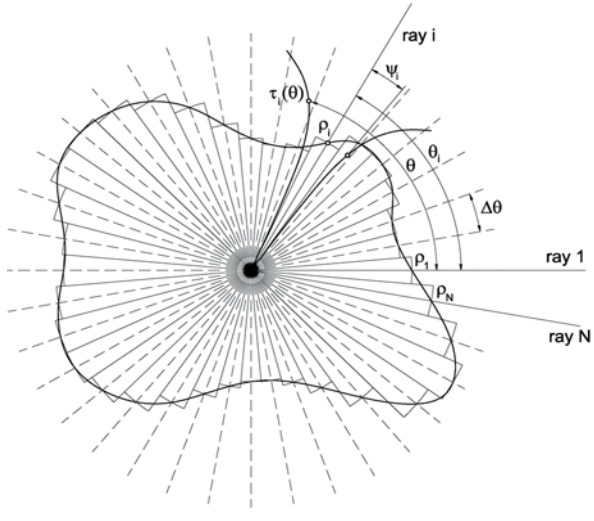


Figure 2 Approximation of a polar integration using sectors, showing the double Fermat spiral superimposed over the ray *i* and the rotation angle ψ_i.

$$\tau_i(\theta) = \sqrt{k_2(\theta - \theta_i)}, \quad -\alpha < \theta - \theta_i < \alpha, \tag{3}$$

where θ_i is the orientation of ray i in the discretization, with radius ρ_i. If we rotate $\tau_i(\theta)$ by ψ_i, of the same sign (and therefore of the same orientation) as $\Delta\theta_i$, so that the following equation is verified:

$$\tau_i(\theta_i - \psi_i) = \rho(\theta_i) = \rho_i, \tag{4}$$

then

$$\begin{gathered} \sqrt{k_2|(\theta_i - \psi_i) - \theta_i|} = \rho_i \\ \Leftrightarrow \\ k_2|\psi_i| = \rho_i^2. \end{gathered} \tag{5}$$

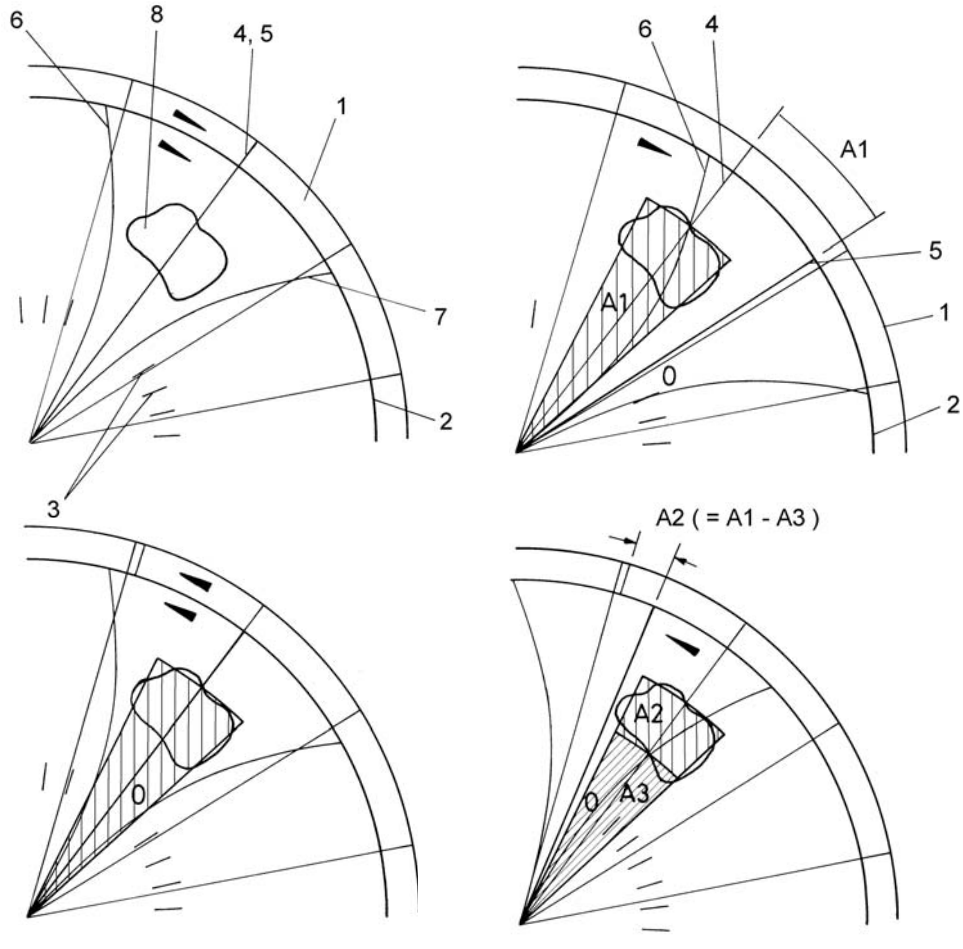
If we repeat this operation over all the rays of the approximation of the polar integral, and substitute in equation (2), we have:

$$S \approx k_1 k_2 \sum_{i=1}^{N} (-1)^M |\psi_i| = k_1 k_2 \sum_{i=1}^{N} \psi_i. \tag{6}$$

That is, we have stated that the unknown area is approximately equal to an algebraic sum of angles. Having reached this point the question is whether it is now possible to configure a new device to quantify this sum of angles.

The Device

The device consists of three transparent sheets, superimposed on top of each other. The three sheets are joined together so that they can be rotated with respect to each other. The shape of the sheets will be such that it is possible to rotate the third sheet without rotating the first and second. To facilitate this, the third sheet may have a smaller diameter. Furthermore, a rotation of the second sheet must imply that the third one rotates jointly with it. This can be achieved in a simple way by some sort of frictional contact between the two sheets.



Figures 4-7 Example of the measurement of a small figure. (1) Second sheet; (2) Third sheet; (3) Angular graded scale; (4) Origin of angular measurement (only shown outside the first sheet); (5) Ray acting as symmetry axis; (6) First branch of the Fermat spiral; (7) Second branch of the Fermat spiral; (8) Figure.

On the first sheet, a full set of equally spaced rays is plotted originating from the centre of the sheet. The angular amplitude between two consecutive rays corresponds to the value of $\Delta\theta$; the centre of the sheet (and indeed that of the other sheets) corresponds to the polar origin **O** (in the mathematical discussion above). On the second sheet, a single ray with the same centre plays the role of the origin of angular measurement. Finally, on the third sheet, the double Fermat spiral

$$\tau(\theta) = \sqrt{k_2\theta}, \qquad -\alpha < \theta < \alpha, \qquad |\alpha| = \frac{R^2}{k_2} \quad (7)$$

is plotted, with R being the radius of the sheet. On this sheet an angular graded scale, also centred on **O,** and a single ray that acts as the symmetry axis of the double Fermat spiral are also plotted. In Figure 3, a blown-up perspective of the device can be seen.

Using this device, the polar function defined by equation (3) may be formed over any figure, by means of the transparency of the sheets, simply by rotating the second and third sheets (while maintaining the first sheet stationary on the paper) until the ray of the third sheet overlaps ray ρ_i.

The angle ψ_i is then calculated by graphically solving equation (4). This can be achieved by rotating only the third sheet, with the same orientation of angular increase $\Delta\theta_i$, until the appropriate branch of the double Fermat spiral plotted on the sheet

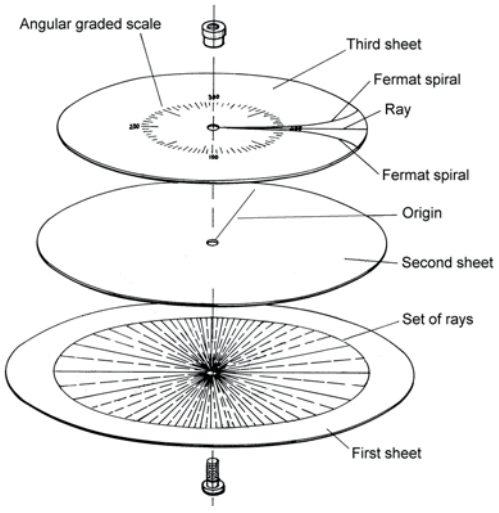


Figure 3 A blown-up perspective of the device

overlaps the point of intersection between the contour of the figure and the ray ρ_i from the full set of rays plotted on the first sheet.

The angle ψ_i is quantified by the angular amplitude between the origin of the angular measurements plotted in the second sheet and the ray of the third sheet.

When a new measurement is taken for ray ρ_{i+1} following this procedure, the first angle ψ_i will remain recorded, provided care is taken that the relative rotation between the second and third sheets is zero when overlapping the ray of the third sheet and the ray ρ_{i+1}. As mentioned above, some sort of frictional contact between the sheets will ensure this occurs.

The new angle ψ_{i+1} is then added to ψ_i. Therefore, at the end of the process, when all the rays that intersect the contour of the figure have been accounted for, the sum (6) will be shown in the angular graded scale of the third sheet as the total angular amplitude between the rays of the second and third sheet. If this angular scale is properly arranged, then the value that is read in the scale will be the unknown area of the object of interest.

In Figures 4-7 an example is shown of the measurement of a small figure, following the steps as explained above. In this case, a simple ray is used, and two angle measurements (one positive and the other negative) are carried out because the polar centre is located outside the figure.

The design parameters of the device are the following:

- R, proportional to the size of the device;
- $k_1 = \frac{1}{2}|\Delta\theta|$, proportional to the angular discretization plotted on the first sheet;
- k_2, the coefficient that affects the shape of the double Fermat spiral plotted on the third sheet.

The first parameter does not affect the quality of the measurement, but the accuracy of the device is inversely proportional to k_1 and k_2. The drawback is that the maximum number of rotations is inversely proportional to these parameters too. In order to avoid the implementation of a revolution counter, it is necessary to limit their value. This necessarily limits the accuracy of the device.

Conclusion

Through experimental research done with a number of different designs of the device, it has been demonstrated that it is relatively easy to obtain a measurement of an irregular planar figure to within a precision of 5%. This accuracy is achieved by arranging a set of rays on the first sheet of relatively low density (as small as 7.5º of angular amplitude, equivalent to a parameter k_1 = 0.06545 rad) and a parameter k_2 equal to 363 cm^2/rad, with a radius of the third sheet equal to R = 19.5 cm. The device, though somewhat complicated to explain mathematically, is actually extremely simple and easy to operate.

In conclusion, it has been demonstrated that, even today, there is still scope to add to the wide range of geometrical devices in existence. As has been seen, preparing a device for the manual measurement of areas is a relatively easy task which takes us back to an age prior to the dominance of CAD and GIS techniques.

References

[1] Abram M. Lopshitz; 1994; *Cálculo de las áreas de figuras orientadas*; Rubiños-1860; S.A.

Stochastic Modelling of Biological Systems

Michael Grayling
Statistics PhD Student, Sidney Sussex

Today, the use of mathematical modelling to help understand the intricacies of biological systems has become common place. Biologists now undoubtedly appreciate their multiplicative utility as a complement to wet-lab research. Traditionally, the approach to handling these dynamical systems has been based on the use of deterministic ordinary differential equations (ODEs). However, in recent years, it has become apparent that ODE models often fail to fully explain how complex biological systems truly work. Therefore, the increased deployment of stochastic models seems paramount. In this article I hope to briefly detail not only what a stochastic model is, but also why and when they should be used.

What is a Stochastic Model? Why do we Need Them?

Informally, a stochastic model is any model for which the solution's trajectory through time is not certain. That is, it is probabilistic in nature. This places them in clear contrast to the more familiar ODE, for which suitable initial conditions determine the solution for all time. But why do we need probabilistic representations? Why can't we just use deterministic models for biological processes? Well, the fact of the matter is that all such systems are under the influence of 'noise', affecting the accuracy of any model. This noise is generally divided into two categories. Extrinsic noise introduces uncertainty due to external environmental factors. For example, for cellular systems this could mean the cell cycle stage. In contrast, intrinsic noise is due to small numbers of molecules; this provides an uncertainty of knowing when a reaction will occur and which reaction it will be. You see, ODE models implicitly assume compartments are well stirred and that the abundance of all species is high enough to permit fluctuations to be ignored, but this is often not the case.

Now, stochastic models actually come in various forms, depending upon whether the dependent variable and the time variable are continuous or discrete. Here we consider a discrete dependent variable with continuous time.

The Degradation Model

As a most basic introduction to stochastic modelling we look to the 'degradation model', frequently used in chemistry to model the breakdown of a chemical species. Deterministically it takes the form:

$$\frac{dC}{dt} = -kC,$$

where C is the chemical of interest, k is a rate constant, and the initial condition $C(0) = C_0$ has been used.

For the stochastic modeller however, the formal definition of k is recalled: the probability a randomly chosen molecule of species C degrades in the interval $[t, t + \delta t)$ is given by $k\delta t$, with δt an infinitesimally small time step, and so the probability exactly one molecule degrades in this interval is given by $C(t)k\delta t$. This is actually all that is needed to design a stochastic simulation algorithm (SSA). The number of molecules of the chemical species, $C(t)$, at times $t = h\Delta t$, for some

pre-set small time step Δt and $h = 1, 2, 3, \ldots$, is found using the following algorithm:

1. Set $t = 0$ and $C(t) = C_0$. Choose Δt.
2. Generate a random number, r, according to the uniform distribution on (0, 1).
3. If $r < C(t)k\Delta t$ then $C(t + \Delta t) = C(t) - 1$, else $C(t + \Delta t) = C(t)$.
4. Repeat Steps 2 and 3 for $t = t + \Delta t$.

This works since $r \sim$ Unif(0,1), and therefore the probability that $r < C(t)k\Delta t$ is equal to $C(t)k\Delta t$. Thus Step 3 implies the probability of a single molecule degrading is $C(t)k\Delta t$, as required. Figure 1 depicts several realisations of the stochastic model as well as the deterministic. Each time the algorithm is run, a different result is achieved. One may therefore reasonably ask what useful information can be drawn from a stochastic model. In most cases this is done by computing the average and variance across a large number of simulations.

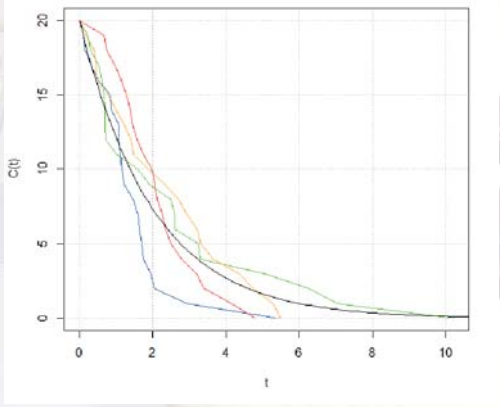


Figure 1 The stochastic and deterministic versions of the Degradation Model. Four realisations of the stochastic system are shown in various colours. In addition, the solution to the deterministic model is provided in black. Here, $C(0) = 20$ and $k = 0.5$.

For our case of the degradation model however, we can illustrate an interesting point by computing the stochastic mean across infinitely many realisations, as follows.

We define $\mathbb{P}(n,t)$ to be the probability that at time t there are n molecules of chemical species C. Then, we consider an infinitesimally small time step δt such that the probability more than one molecule degrades during $[t, t + \delta t)$ is negligible. Now there are two ways in which at time $t + \delta t$ there can be n molecules. Either at time t there were n molecules and no reaction took place, or at time t there were $n + 1$ molecules and in the interval $[t, t + \delta t)$ one molecule was degraded. Mathematically this can be formalised as:

$$\mathbb{P}(n,t+\delta t) = \mathbb{P}(n,t+\delta t)(1-kn\delta t) + \mathbb{P}(n+1,t)k(n+1)\delta t,$$

$$\Rightarrow \frac{d}{dt}\mathbb{P}(n,t) = k(n+1)\mathbb{P}(n+1,t) - kn\mathbb{P}(n,t),$$

where we take the limit as $\delta t \to 0$. This equation is usually called the *Chemical Master Equation* (CME). Numerical solution of a CME is often extremely computationally expensive. It is in this case though, possible to solve the above algebraically. Recalling that $C(0) = C_0$, and that the chemical species is only able to decay, it is easy to see that $\mathbb{P}(n,t) \equiv 0$ for $n > C_0$; leaving us with a system of $C_0 + 1$ coupled linear differential equations. Beginning with the ODE for $\mathbb{P}(C_0,t)$, then $\mathbb{P}(C_0 - 1,t)$, it is possible to formulate a hypothesis for the form of $\mathbb{P}(n,t)$ which can be solved inductively to find:

$$\mathbb{P}(n,t) = e^{-knt}\binom{C_0}{n}(1-e^{-kt})^{C_0-n}.$$

Principally, we are usually interested in the average number of molecules at time t, which is defined using the usual formula for expectation:

$$M(t) = \sum_{n=0}^{\infty} n\mathbb{P}(n,t).$$

In this case, having explicitly found $\mathbb{P}(n,t)$, it is possible to substitute in to find M. However we'll here illustrate a more general technique for analysing the CME. We multiply our CME across by n and sum to obtain:

$$\frac{d}{dt}\sum_{n=0}^{\infty} n\mathbb{P}(n,t) = k\left[\sum_{n=0}^{\infty} n(n+1)\mathbb{P}(n+1,t) - \sum_{n=0}^{\infty} n^2\mathbb{P}(n,t)\right]$$

$$= k\left[\sum_{n=0}^{\infty}(n-1)n\mathbb{P}(n,t) - \sum_{n=0}^{\infty} n^2\mathbb{P}(n,t)\right],$$

$$= -k\sum_{n=0}^{\infty} n\mathbb{P}(n,t),$$

i.e. we have:

$$\frac{d}{dt}M = -kM \Rightarrow M(t) = C_0 e^{-kt},$$

using the condition $M(0) = C_0$. Thus the mean of the stochastic system is equal to the solution of the corresponding deterministic model. This is actually true in general for linear systems, and for non-linear systems the ODE model can act as a first approximation to the mean of the system.

The SIR Epidemic Model

To really appreciate the utility of stochasic models, we now consider a more complex system: the SIR model for an epidemic. The SIR model has proven extremely useful since its creation, and has been successfully used to model the spread of numerous diseases, providing simple rules for how the number of Susceptibles (S), Infectious (I), and Recovereds (R) changes through time as a disease spreads. Its deterministic formulation is given by:

$$\frac{dS}{dt} = -\beta SI, \frac{dI}{dt} = \beta SI - \gamma I, \frac{dR}{dt} = \gamma I,$$

with $S(0), I(0), R(0) \geq 0$ and $S(0) + I(0) + R(0) = N$. Here, β represents the transmission rate of the disease, and γ the recovery rate.

The basic reproductive ratio, R_0, is roughly defined as the average number of secondary infections that occur when one infective is introduced into a completely susceptible population. For the SIR model it takes the form:

$$R_0 = \frac{\beta}{\gamma} N.$$

Its importance lies in the fact that if $R_0 < 1$ then no epidemic can occur; the solution for $I(t)$ decreases monotonically to zero. However, if $R_0 > 1$ then $I(t)$ first increases; i.e. an epidemic occurs.

To stochastically model this system, we again turn to the formal definitions of the parameters in the model, and make use of our SSA from earlier to produce Figure 2. The key point to notice is that for one of the stochastic realisations no epidemic has occurred. This is despite the fact that R_0 is greater than 1, where in a deterministic setting an epidemic would be guaranteed. It seems logical that an epidemic should never be certain just because a few simple conditions are met. This is therefore just one example of how incorporation of noise can help more accurately model biological systems.

Conclusion

What does this mean for the future? Should ODE models be abandoned entirely? Well obviously not. There are a few issues to consider: Firstly, the scale and stability of the system being modelled must be taken into account. Whilst stochastic effects are inherent in both the micro and macroscopic worlds, in reasonably stable environments, especially at the macroscopic level, ODE models do perform very well. Perhaps a greater issue though is that there are still many systems for which we simply do not know enough to utilise stochastic models, in which the simplified ODE version becomes a necessity. These two alternatives should therefore be used to assist each other in obtaining a greater understanding of biological mechanisms.

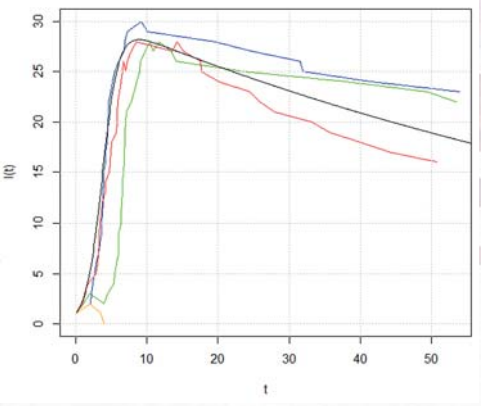


Figure 2 The stochastic and deterministic versions of the SIR Epidemic Model. Four realisations of the number of infecteds for the stochastic system are shown in various colours. In addition, the solution to the deterministic model is provided in black. Here, $\beta = 0.03$, $\gamma = 0.01$, $S(0) = 29$, $I(0) = 1$ and $R(0) = 0$.

References

[1] Leah Edelstein-Keshet; 2004; *Mathematical Models in Biology*; SIAM.

[2] Darren J. Wilkinson; 2011; *Stochastic Modelling for Systems Biology*; CRC Press.

[3] Yitzhak Pilpel; 2011; *Noise in biological systems: pros, cons and mechanisms of control*; Methods in Molecular Biology 759: 407-425.

[4] Matt J. Keeling, Pejman Rohani; 2007; *Modeling Infectious Diseases in Humans and Animals*; Princeton.

Metaswitch Networks

THE BRAINS OF THE NEW GLOBAL NETWORK

METASWITCH IS POWERING THE TRANSITION OF COMMUNICATION NETWORKS INTO A SOFTWARE-CENTRIC FUTURE. What does that mean to you? It means that, together, we can work at the centre of a major shift in the way the world communicates and consumes media. We're hiring the brightest minds, the most creative thinkers who are open to change, to help shape this dramatic new landscape. We're looking for people who value collaboration and innovation as much as we do.

COME BE A PART OF METASWITCH: THE BRAINS OF THE NEW GLOBAL NETWORK.

START YOUR JOURNEY AT **CAREERS.METASWITCH.COM**

Generalising the Division Algorithm

Samin Riasat
Former Mathematics Undergraduate (Graduated 2013), Queens' College

The division algorithm has many important consequences. For example, well-known proofs of:

1. The highest common factor of two natural numbers exist, and is their least positive linear combination
2. $\mathbb{Z}$ is a principal ideal domain
3. The minimal solution to Pell's equation generates all solutions

all follow the same pattern and use the same idea, that of the division algorithm. (Also note that 1 and 2 are essentially the same.) If you have noticed how the division algorithm applies in these cases, you might have wondered: in which other cases can we apply the division algorithm? What is the most general case where we can apply it? In this article we shall attempt to answer these questions using groups instead of the general ring theory approach.

The Set-up

Say S is a set where the division algorithm applies. We definitely need some sort of order in S to say that the 'remainder' must be 'less than' the 'divisor'. We might want S to be closed under some operation (so that we can repeatedly 'subtract' the 'divisor' from the 'dividend') and we also need an inverse operation (i.e. the 'subtraction').

Such an order on S needs to be a partial order. [Recall that a partial order on a set S is a binary relation $\leq$ on S that is (i) reflexive: $a \leq a$; (ii) antisymmetric: if $a \leq b$ and $b \leq a$ then $a = b$; and (iii) transitive: if $a \leq b$ and $b \leq c$ then $a \leq c$; for all $a, b, c \in S$.] In addition, the closure and inverse operations suggest that we take S to be a group. How should the order behave under the group operation? Clearly we want it to be compatible with the operation. We also probably want the inverse of a 'positive' element to be 'negative', and vice-versa. Do we need the group to be abelian? Maybe, but let's not impose all the conditions yet.

So, to put our ideas into action, let $(G, +)$ be a group with a partial order $\leq$ such that, for all g, $g_1, g_2 \in G$, $g_1 \leq g_2$ implies $g + g_1 \leq g + g_2$ and $g_1 + g \leq g_2 + g$. We say that an element $g \in G$ is *positive* if $0 \leq g$, where 0 is the identity (zero) element of G. Define the *positive cone* of G to be the set $G^+ := \{g \in G : 0 \leq g\}$ of all positive elements.

Moving on from Definitions

Now we hopefully have the necessary axioms in place. Let's see if we can prove anything from these. The first thing that we want is probably: if $0 \leq g$, then $-g \leq 0$. This follows easily: $0 \leq g$, add $-g$ to both sides and we are done. It works the other way around as well, so in fact we have proved:

Lemma 1 $0 \leq g$ iff $-g \leq 0$.

That was good. We wanted the inverse of positive elements to be negative and it just followed from the definition. But we want more! So take a non-zero element $g \in G$. By Lemma 1 we can take g to be positive without loss of generality. How about adding something to both sides of $0 \leq g$ again? Last time we added $-g$. We can add 0, but that doesn't

change anything. So let's add g: $g \leq g + g = 2g$. Now what? Let's add g again! $g + g \leq 2g + g$, i.e. $2g \leq 3g$. Combining the last two gives $0 \leq g \leq 2g \leq 3g$. It follows by induction that $mg \leq ng$ for all integers $0 \leq m \leq n$ (note: here $ng = g + g + \ldots + g$, n times). This looks promising.

What about 'negative' elements? Note that by Lemma 1, $-g \leq 0$. Adding $-g$ to both sides yields $-2g \leq -g$, and so on. So we have another nice result:

Lemma 2 $mg \leq ng$ for all integers $m \leq n$ and $g \in G^+$.

It seems that this is all we can derive from our first principles. So we want to apply more restrictions on G. Let $a, b \in G$ be positive with $b \leq a$. As in the division algorithm, let's look at $a - b$, $a - 2b$, ... etc. We want this sequence to stop as soon as $a - nb$ becomes negative. How do we do this? In other words, we want the set $\{a - nb : n \in \mathbb{Z}\}$ to have a least positive element. Did something just pop up in your mind? A set having a least element must have reminded you of something like... the well-ordering principle! (In case you don't know what it is, it is basically the statement that the natural numbers $\mathbb{N}$ are well-ordered; that is, every non-empty subset of $\mathbb{N}$ has a least element.) So how about we impose the extra condition that $\leq$ is a well-order on G? But hang on. This is clearly absurd if we think about it for a minute, as for any positive element g the set $\{ng : n \in \mathbb{Z}\}$ has no least element. How about the least *positive* element then? In other words, let's say G^+ is well-ordered under $\leq$.

Now G has quite a few nice properties: it is a group under $+$, $\leq$ is an order on G preserving $+$, and its positive cone is well-ordered. Let's see if our ideas work now!

The Generalisation

Let d be the least non-zero element in G^+ and $g \in G$ be any non-zero element. Without loss of generality, $0 < g$. Then $g \in G^+$ so $d \leq g$. Consider the elements nd for $n \in \mathbb{Z}$. We want $nd \leq g < (n + 1)d$ for some n. Can we achieve this? We certainly have $nd \leq g$ for some $n = 1$, so we need $g < n'd$ for some n'. By Lemma 2, n' must be greater than n. How do we know that n' exists?

Suppose it doesn't. Then $nd \leq g$ for all sufficiently large n, so $nd \leq g$ for all $n \in \mathbb{Z}$ by Lemma 2. Then $0 \leq g - nd$, $g - nd \in G^+$ for all integers n. Hence $\{g - nd : n \in \mathbb{Z}\} \subseteq G^+$, so it has a least element $g - md$. Then $g - md \leq g - nd$ for all n, which implies $0 \leq (m - n)d$ for all integers n, a contradiction.

Now we can take the maximal n such that $nd \leq g$. Then $g < (n + 1)d$, so $nd \leq g < (n + 1)d$. The left inequality says $g - nd \in G^+$, and the right inequality says $g - nd < d$. So $g - nd = 0$ and $g = nd$. This is exactly what we wanted.

We have shown that $G = \langle d \rangle$. In fact we can do more. Clearly G cannot be finite. Because otherwise d must have finite order, i.e. $kd = 0$ for some positive integer k. Then $0 \leq d \leq 2d \leq \ldots \leq kd = 0$ by Lemma 2. So all of these must be equalities (by antisymmetry), i.e. $d = 0$, a contradiction.

So our restrictions have not only worked, we've shown that all groups with these properties essentially have the same structure, that of the infinite cyclic group. Let's give G a name: we say that the group G is *well-ordered* if the set G^+ is well-ordered under $\leq$. We have thus proved:

Proposition 1 The only non-trivial well-ordered group is the group $(\mathbb{Z}, +)$ of integers (up to isomorphism).

Corollaries

Now we can give one-line proofs of the facts stated at the beginning using Proposition 1: (here any ordering is under the usual $\leq$ order in $\mathbb{R}$)

Corollary 1 The highest common factor of two natural numbers exists, and is their least positive linear combination.

Proof For $a, b \in \mathbb{N}$, the additive group $G = \{ax + by : x, y \in \mathbb{Z}\}$ is well-ordered, and so is equal to $\langle d \rangle$ for d the least positive element of G.

Corollary 2 $\mathbb{Z}$ is a principal ideal domain.

Proof Any ideal in $\mathbb{Z}$ is a well-ordered group, and so must be $\langle d \rangle$ for some d.

Corollary 3 If $x_0 + y_0\sqrt{d}$ is the least solution > 1 to Pell's equation $x^2 - dy^2 = 1$, then all solutions are given by $x_n + y_n\sqrt{d} = (x_0 + y_0\sqrt{d})^n$ for $n \in \mathbb{Z}$.

Proof The solutions $x_n + y_n\sqrt{d}$ to Pell's equation form a subgroup of the multiplicative group of units in the ring $\mathbb{Z}[\sqrt{d}]$, which is well-ordered.

Exercise Show that every discrete subgroup of $(\mathbb{R}, +)$ is infinite cyclic.

Minimum Clues: Sudoku and Sudokion

Stephen Jones
Co-founder, Muddled Puzzles

I make Sudokion, a large family of pure spatial-logic puzzles derived from Sudoku. The puzzles range in size from 25 cells (5×5 grid) to 369 cells (five interlocking and interdependent 9×9 puzzles). Sudokion's rules are the same as those governing Sudoku – every row, column and cluster must contain each of the numbers 1, . . ., *n*, for a grid measuring *n* cells by *n* cells.

Over the years I have developed an interest in the extent of complexity required to achieve *absolute economy* of clues in puzzles of increasing grid size. I define absolute economy as $n - 1$ clues, where n is the number of cells in any row, column or cluster of the puzzle.

On 1st January 2012 Gary McGuire, of University College, Dublin, announced that, after 7.1 million core-CPU hours on a supercomputer, using a hitting-set algorithm, his team had established that a traditional 9×9 Sudoku requires a minimum of 17 clues to guarantee a unique solution (see [1]). This reminded me that, even if they are not necessarily interested in the puzzles themselves, some mathematicians at least are interested in the mathematics behind the puzzles.

Very Small Puzzles

Puzzles of grid sizes 1×1, 2×2 and 3×3 permit absolute economy easily. Clearly, in the case of a 1×1 puzzle no clue is required as only one value (1) is possible. For a 2×2 puzzle, with two rectangular clusters, only one clue is ever required. The simplest grid shape for a 3×3 puzzle is three rectangles side-by-side. While the puzzle may have three clues that do not produce a unique solution, ideal selection of two (vital) clues allows a unique solution (see Figure 1).

Small Puzzles: Compensating Symmetry

In order to achieve absolute economy, ideal selection of vital clues is required for all puzzles of grid size greater than 3×3.

In the case of a traditional Sudoku, with four 2×2 boxes (Figure 2), three clues in any arrangement are insufficient to provide a unique solution. But when the square boxes are replaced by a Sudokion called *Logikion* (Figure 3) the irregularly-shaped clusters allow a 4×4 puzzle with only three clues,

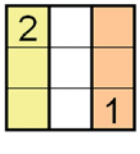


Figure 1 A 3×3 grid with absolute economy of clues

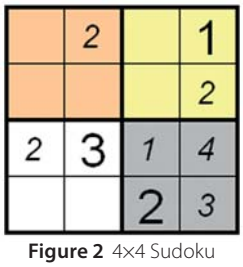


Figure 2 4×4 Sudoku

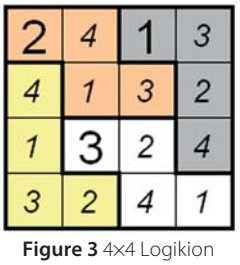


Figure 3 4×4 Logikion

provided that the shape of the clusters and the selection of clues are ideal. So, too, when the grid size is increased to 5×5 (Figure 4), a well-constructed Logikion will allow absolute economy of four clues.

The irregular shape of the Logikion's clusters tends to promote a feature that I call *compensating symmetry*. Figure 5 illustrates four compensating symmetries of the Logikion shown at Figure 4. Sometimes there is a direct correlation between two individual cells, as is the case with the two 1s in Figure 5.1. At other times a group of two or more cells are correlated in some combination.

Puzzles containing many compensating symmetries tend to allow greater economy of clues. A strong group of compensating symmetries is usually the basis for *eutaxy* – an ideal arrangement of irregularly-shaped clusters combined with ideal selection of clues – a combination most likely to produce as economic a puzzle as possible.

Medium-sized Puzzles: Fragmented Clusters

The achievement of a 6×6-grid puzzle with five clues requires more complexity than the Logikion can offer. Experience tells me that the 6×6 Logikion in Figure 6 is about as far as five clues will stretch.

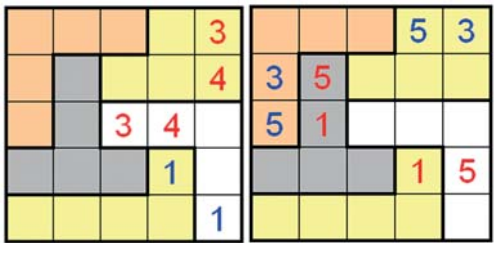


Figure 5.1 & 5.2 Compensating Symmetries

Figure 6 6×6 Logikion

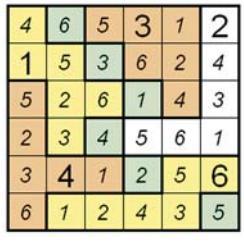


Figure 7 6×6 Pandemonion

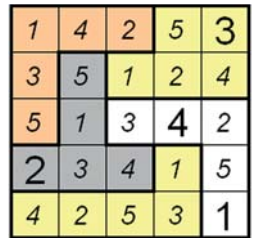


Figure 4 5×5 Logikion

Figure 8 7×7 Pandemonion

The *Pandemonion* (Figure 7), with one completely-fragmented cluster, is the least complex 6×6 pure spatial-logic puzzle that will allow five clues and a unique solution. Likewise, a 7×7 Pandemonion (Figure 8) is capable of as few as six clues.

The Pandemonion offers scope to distribute the fragmented cells anywhere in the grid, thus allowing a eutaxy greater than that available to Logikion and, therefore, more scope for absolute economy.

Introducing the Plus Factor

All the Logikion and Pandemonion puzzles illustrated thus far are presented in 'plain format' – the unadorned puzzle. I also create 'plus-format' Sudokion, puzzles upon which are superimposed a line or lines that must contain all the values of the puzzle.

My experience of making a very wide range of Sudokion has convinced me that seven clues for any plain-format 8×8 Sudokion will never produce a unique solution. So, it is to plus-format Sudokion that we resort in order to seek absolute

Figure 9 8×8 Diagonal Katastrophion: every row, column and cluster, including the fragmented green and pink clusters, and the red diagonal line must contain the numbers 1 to 8.

economy in puzzles with grid sizes 8×8 and 9×9.

In the Diagonal Katastrophion (Figure 9), the combination of the extra plane available to the values on the diagonal and the puzzle's essential eutaxy contribute to its absolute economy of seven clues.

The Holy Grail: 81 Cells, 8 Clues

Before beginning this article, the most economic 9×9 puzzle I had made was a Parallelogram Pandemonion (Figure 10). I had high hopes that this type of puzzle would be able to produce an example with just eight clues but, having made 35 examples with only one example having nine clues, I am almost certain that nine is as good as it gets.

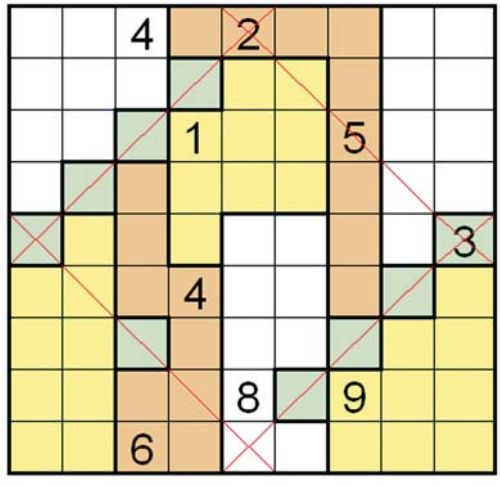


Figure 10 9×9 Parallelogram Pandemonion: every row, column, cluster, (including the fragmented green cluster) and both the upper and lower red 'V' lines of the parallelogram must contain the numbers 1 to 9.

When the opportunity to contribute to *Eureka* arose I decided to go for the Holy Grail, a 9×9 puzzle with only eight clues. The Para-X Pandemonion (Figure 11) is the result. The four planes created by the superimposed lines, the compensating symmetries, the nine linear intersections and ideal selection of clues combine to give the puzzle a unique solution with only eight clues. (To date I have made 18 Para-X Pandemonions, three of which contain only eight clues.)

I have almost run out of space but there is just enough to state that, for a number of reasons, a 10x10 plus-format Sudokion with nine clues is almost certainly impossible.

For more Sudokion puzzles and a brief explanation why a 10x10 Sudokion with nine clues is "impossible" please visit my website, *www.muddledpuzzles.com*.

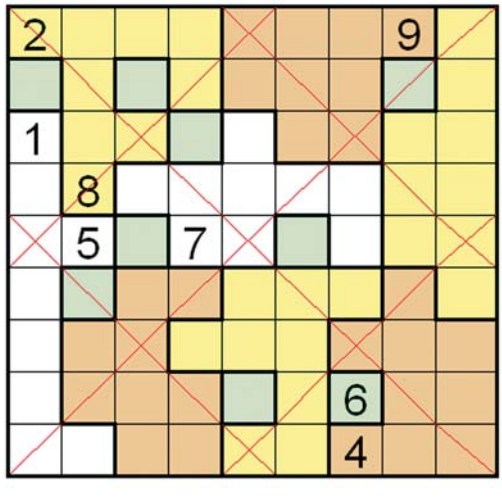


Figure 11 9×9 Para-X Pandemonion: every row, column, cluster (including the fragmented green cluster), both red diagonal lines and both the upper and lower red 'V' lines of the parallelogram must contain the numbers 1 to 9.

References, Further Reading

[1] Eugenie S. Reich; 2012; *Mathematician claims breakthrough in Sudoku puzzle*; Nature; *http://www.nature.com/news/mathematician-claims-breakthrough-in-sudoku-puzzle-1.9751.*
[2] Wikipedia; *Mathematics of Sudoku*; http://en.wikipedia.org/wiki/Mathematics_of_Sudoku.

Small Stellated Dodecahedron, by Vladimir Bulatov

A small stellated dodecahedron is a nonconvex polyhedron, existing in $\mathbb{C}^2$. It is formed of five pentagrams (stars) intersecting at vertices. The sculpture reflects the true form of the shape, with no extra intersections between faces.

See www.bulatov.org for more of Vladimir's pieces.

Get Your Geek On!

Some of the geekiest things you can ever buy...

...from www.uncommongoods.com

So Jar, So Good

This speaker-in-a-jar is completely self-contained, so you can bring it on-the-go with you for impromptu dance parties (if mathmos happen to evolve to be good at dancing of course), and can hook it up to pretty much anything. Apparently nothing says 'geeky' like using simple physics in everyday life.

The Opening Number

This mathematical bottle opener is sure to become an operational constant when you're seeking liquid inspiration. The hefty steel form is calculated for optimum leverage, solving stubborn bottle tops with near-infinite ease. When not in use, it becomes a smart, sculptural accent or paperweight. And we all know Pi can do everything.

The Gentleman Geek

An age-old problem that has plagued the mathematical community for years: how can I show off what a genius I am and still look devilishly handsome? Now we've all got the answer! Only a handful of people can find the phrase 'it's *e* o'clock' funny and we know they're currently reading this magazine.

Head of the Glass

Even though mathmos tend not to be the heaviest of drinkers, when we do it, we do it in style. The mathematical symbols on these glasses are bound to thrill fellow geeks and pique the interests of numerical novices. Whether they join you with water during class or with something a little stronger as you celebrate cracking your latest conundrum, you'll be glad you have your digits around.

Pascal found order in a game of dice.

Can you find the patterns in chance events?

If you can bring science to bear on the toughest challenges, apply here today for our Quantitative Analyst roles.

www.gresearch.co.uk
/predict-the-future

MACHINE LEARNING
FINANCIAL DATA
RESEARCH
LATENCY
EXECUTION
PLATFORM
FORECAST
LANGUAGE
MODELLING
RISK MODELS
TRADING ENGINE
SYSTEMATIC
BIG DATA
DATA STRUCTURES
MATHEMATICAL
MODELLING
DATABASE DESIGN
CODE
GLOBAL TRADING
PLATFORM
STATISTICAL
ARBITRAGE
NUMERICAL
PROGRAMMING
MARKET DATA
BAYESIAN
LANGUAGE
MODELLING
RISK MODELS
TRADING ENGINE
BIG DATA
CODE
SYSTEMATIC
ALGORITHMIC
TRADING PLATFORM
MACHINE LEARNING
FINANCIAL DATA
RESEARCH
LATENCY

G-RESEARCH

Funny Möbius strip geeky arts

Compiled by Diana Danciu

The Kiss Precise

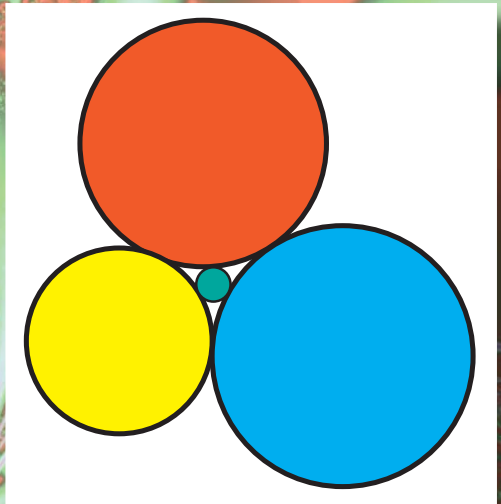
by Frederick Soddy

For pairs of lips to kiss maybe
Involves no trigonometry.
'Tis not so when four circles kiss
Each one the other three.
To bring this off the four must be
As three in one or one in three.
If one in three, beyond a doubt
Each gets three kisses from without.
If three in one, then is that one
Thrice kissed internally.

Four circles to the kissing come.
The smaller are the benter.
The bend is just the inverse of
The distance from the center.
Though their intrigue left Euclid dumb
There's now no need for rule of thumb.
Since zero bend's a dead straight line
And concave bends have minus sign,
The sum of the squares of all four bends
Is half the square of their sum.

To spy out spherical affairs
An oscular surveyor
Might find the task laborious,
The sphere is much the gayer,
And now besides the pair of pairs
A fifth sphere in the kissing shares.
Yet, signs and zero as before,
For each to kiss the other four
The square of the sum of all five bends
Is thrice the sum of their squares.

In *Nature*, June 20, 1936



Two mathmos were walking along the backs when one said, "Where did you get such a great bike?"

The second mathmo replied, "Well, I was walking along yesterday minding my own business when a beautiful woman rode up on this bike. She threw the bike to the ground, took off all her clothes and said, "Take what you want."

The first mathmo nodded approvingly, "Good choice; the clothes probably wouldn't have fit.

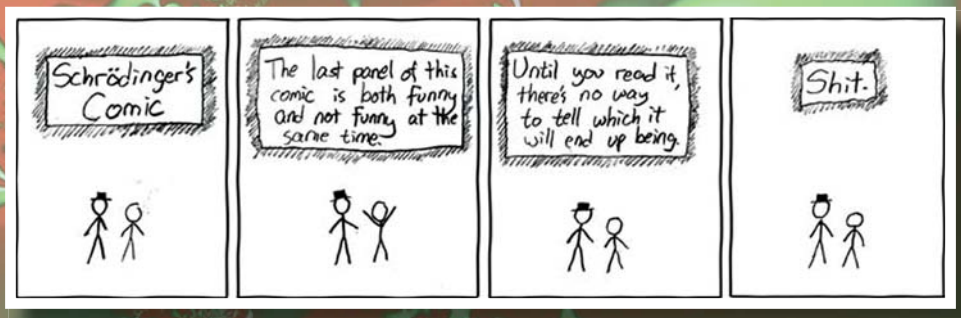
Q: Why did the chicken cross the road?
A: The answer is trivial and is left as an exercise for the reader!

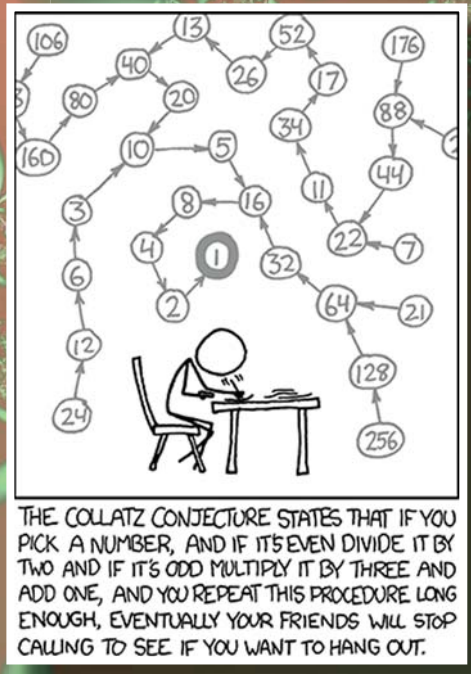
Let there be a spherical cow....

Two random variables were talking in a bar. They thought they were being discrete but I heard their chatter continuously.

What does the B in Benoît B. Mandelbrot stand for? Benoît B. Mandelbrot!

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random
}
```





An engineer, a physicist and a mathematician are staying in a hotel.

The engineer wakes up and smells smoke. He goes out into the hallway and sees a fire, so he fills a trash can from his room with water and douses the fire. He goes back to bed.

Later, the physicist wakes up and smells smoke. He opens his door and sees a fire in the hallway. He walks down the hall to a fire hose and after calculating the flame velocity, distance, water pressure, trajectory, etc. extinguishes the fire with the minimum amount of water and energy needed.

Later, the mathematician wakes up and smells smoke. He goes to the hall, sees the fire and then the fire hose. He thinks for a moment and then exclaims, "Ah, a solution exists!" and then goes back to bed.

Lecturer Reviews

Prof Imre Leader

★★★★★

Professor of Pure Mathematics, *DPMMS*

Any introduction about this gem of the DPMMS and Trinity College is, pause, take a deep breath, unnecessary. I think every student in the whole world would agree that his figure is best understood when seen in person, sipping coffee in a lecture and talking about induction. The fact that a thing such as The Imre Leader Appreciation Society exists speaks for itself and makes our argument utterly trivial, so we should maybe stop praising His Majesty and mention a couple of things about his lecturing that you might not have noticed:

Handwaving: (5/5) Mathematicians use the term 'handwaving' when they can't quite make their arguments rigorous (at which point one starts to wonder how applied mathematicians' arms haven't come off yet) but Prof. Leader takes this expression to a whole new level. "Mumble mumble, and we're done!"

Examples sheets: (5/5) No example sheet of Prof. Leader's is ever done by anyone completely. In fact, if someone did, I reckon that would earn him a Fields Medal or at least a substantial research grant.

So after we have some coffee and stare at the results for a while, we conclude that Prof. Leader deserves nothing less than 5/5 stars, after which we are done, aren't we? End of proof ✓ ✓ □

Prof Tom Körner

★★★★★

Professor of Fourier Analysis, *DPMMS*

There's one figure you can see at every happy hour, a person also known to students as 'The King of the CMS', and that's Prof Körner. He is the lecturer to go and see if you want to hear some jokes along with the statutory theorems, lemmas and propositions. If you listen carefully, you'll notice that he gives out life lessons on a regular basis, and that anybody who doesn't follow them without question may be considered a fool. You might have heard him say things like "trivial", "obvious" and "I'm going to do something clever" a lot, but that's only because his intelligence far transcends ours and we just have to live with this fact and drink from his infinite wisdom. For that very reason, he automatically gets 5/5 without any further discussion on the matter.

Dr Stephen Cowley ★★★★★

Chair of the Faculty of Mathematics, *DAMTP*

Dr Cowley is well known for his love for flying paper. In our first lecture he told us that we are allowed to throw paper planes at him with the proviso that if the planes flew for less than 10 seconds we had to go and pick them up. Of course, being such well-behaved students, lots of us listened to his advice and threw paper planes of increasing complexity, though I believe none ever flew for the required time period. Towards the end of term, it was Dr Cowley's turn: he would throw balls of paper at students who... let's say... were dreaming too loudly of mathematical equations.

Dr Cowley is also a great believer in auditory learning: each line on a graph is given its own specific sound: "Wee!", "Waa!", "Whoo!", and greek letters are always accompanied by an appropriate animal noise (μ even had it's own cow toy). After we *engage brain* for a second, and stop doing things the stupid way, we conclude immediately that Dr Cowley gets a 5/5.

Dr Piers Bursill-Hall ★★★★★

Researcher of the History of Mathematics, *DPMMS*

From here on in, Dr Bursill-Hall will be referred to by his favourite title, "Our Merciful and Glorious leader" (OMAGL). Perhaps OMAGL's strongest talent as a lecturer is his ability to convey information in as concise a manner as possible. Indeed, he often ends his lectures several hours early, having covered all of his intended material for the day. He is also highly respected for his strict adherence to the traditional rules of the lecture theatre. Food, even bottled water, is strictly forbidden under all circumstances, especially if the packets make rustling noises. Indeed, any noises from the audience are persecuted mercilessly; however, since such noises can be incredibly distracting, it is to our great delight when OMAGL gives an exclamation of "Oh *DO* Stop Coughing". It is absolutely vital to attend all History of Mathematics lectures every year, and also the supporting "History of Science for Mathmos" series. To help students realise the importance of this, OMAGL sends out thrice-weekly reminder emails, which usually go something like this: "HoM today, 4pm, usual place. Your Merciful and Glorious Leader". This is much appreciated by all of us, since a Cambridge maths student's favourite trick is to pretend that their eidetic memory obviates the need for a calendar. Oh Merciful and Glorious Leader, how we worship you and bask in your glory! 5/5.

Solutions to the Problems Drive

1 Infinite Sequences

yes

2 Isometries

e.g. $X = \{e^{in} : n \in \mathbb{N}\}$, $f(z) = ze^{i}$

3 Recursively Coloured Triangles

3/4

4 Primes

(c)

5 Subgroups

8

6 Random Numbers

(b)

7 Geometry

1/6

8 Spirals

2 seconds

11 Paths

120

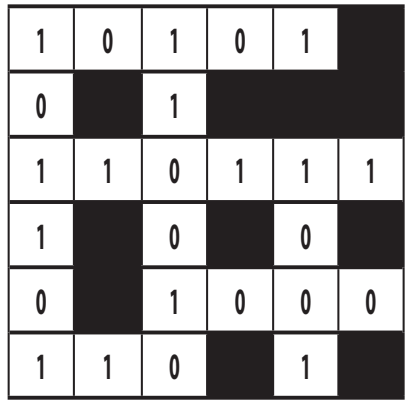
12 Binary Numbers

16383

13 Probabilities

http://en.wikipedia.org/wiki/Sleeping_Beauty_problem#Solutions

14 Time for a Crossword



14 Some Cake

16, 21, 26

15 And some more Cake

24

Copyright Notices

Illustrations

Pages 110-1000:
Background: MorgueFile Free / cohdra
Pen: MorgueFile Free / xolololounge

Page 1101:
Pisces: © Mike Naylor / Used With Permission

Pages 1110-10010:
Apple: MorgueFile Free / Alvimann

Pages 10100-10101:
Frank Ramsey: © Stephen Burch / Used With Permission

Pages 10110-11000:
Sphere: Creative Commons / xJaM

Page 11010-11110:
Spot It! artwork: © Blue Orange Games / Used With Permission

Pages 100000-100010:
Skeletons: Creative Commons / Tim Vickers

Pages 100100-100111:
Backgrounds: MorgueFile Free / pschubert

Page 101111:
Seahorse Valley: Creative Commons / Wolfgang Beyer

Pages 110000-110011:
Languages Map: Creative Commons / Urion Argador

Pages 110100-110110:
Coffee: Creative Commons / 3268zauber
Erdős: Creative Commons / kmhkmh

Page 110111:
Penrose Triangle: Creative Commons / chylld

Pages 111000-111010:
Background 1: Creative Commons / www.CGPGrey.com
Background 2: Creative Commons / amcilrick

Page 111011:
Pythagoras Tree 1: Creative Commons / Gjacquenot
Pythagoras Tree 2: Creative Commons / Atzecsse

Pages 1000010-1000100:
Backgrounds: freeimages.com

Page 1000101:
Triangular Wiggle: © Roberto Giardili / Used With Permission

Page 1000110:
Mandelbrot Set: Creative Commons / Wolfgang Beyer / Used With Permission

Page 1001001:
Julia Set: Creative Commons / Josep M Batlle I Ferrer

Pages 1001010-1001100:
Manifold: Creative Commons / Andrew J. Hanson / Used With Permission
Strings: Steuard Jensen / Used With Permission

Page 1001101:
Volvox: Creative Commons / Casey Dunn

Pages 1001110-1001111:
Background: Creative Commons / Ansgar Walk

Pages 1010000-1010001:
Ramanujan: Creative Commons / Konrad Jacobs
Higgs Boson: Creative Commons / Lucas Taylor
Homotopy: Creative Commons / Salix Alba

Pages 1010010-1010011:
Pointless Logo: © Endemol / Used With Permission

Pages 1011000-1011100:
Backgrounds: MorgueFile Free / pshubert

Pages 1011111-1100000:
Background: MorgueFile Free / carmanna

Pages 1100001-1100010:
Background: Morguefile Free / pshubert

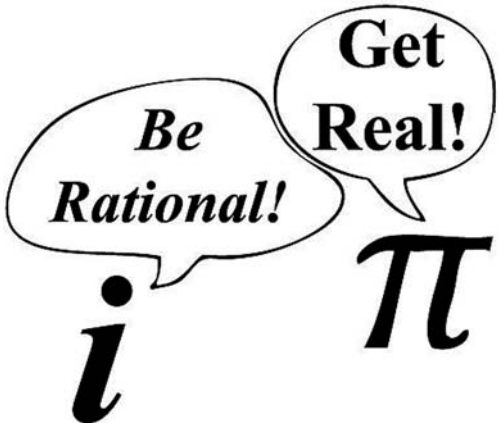
Pages 1100110-1101000:
Backgrounds: Morguefile Free / pshubert

Page 1101111:
Dodecahedron: Vladimir Bulatov /Used With Permission

Page 1110000:
Background: David Craig /Used With Permission

Pages 1110010-1110011:
Background: MorgueFile Free / supernewton2
Cartoons: xkcd.com

The Archimedeans would like to thank Cambridge Publishing Management for their help with the cover design and printing of Eureka 63.



Q[√2]ARCH

The Archimedeans' problems journal: restarting next year.

Join The Archimedeans

Join one of the oldest student societies and get free entrance to countless amazing talks, great social events, discounts in our bookshop and three free copies of Eureka!

Membership is only £5 per year or £10 for life.

Email *archim-eureka-secretary@srcf.ucam.org* for details, visit *www.archim.org.uk* or write to

The Archimedeans
Centre for Mathematical Sciences
Wilberforce Road
Cambridge, CB3 0WA
United Kingdom

Get Involved with Eureka

If you're interested in scientific publishing and want to get involved with Eureka, we'd love to have you. Email *archim-eureka-secretary@srcf.ucam.org.uk* for role descriptions.

Write for Eureka

If you want to contribute to future issues of Eureka, please email *archim-eureka-secretary@srcf.ucam.org*. Further details can be found on our website. Author guidelines are contained on http://www.archim.org.uk/eureka_author_guide.php.