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 64

Despite various organizational issues, which delayed the date of publication by a few months, it has been a pleasure to edit Eureka. It is a rewarding experience, not just for me, but also the entire editing team.

In July 2015, space probe New Horizons became the first spacecraft to fly by Pluto, with detailed measurements and observations made. This is a great discovery of our understanding to the Solar System. This resembles the meaning of ‘Eureka’ – ‘I have found it!’. We publish this magazine with the spirit of exploration, hoping that this can be an inspiration of further discoveries.

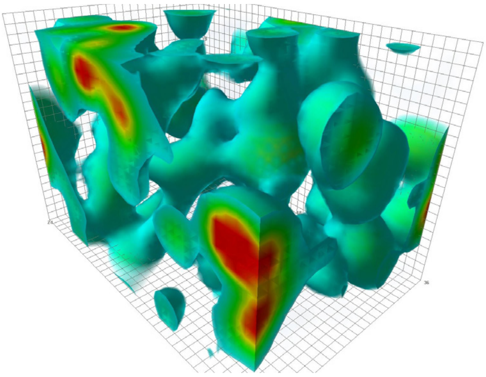
We continue the usual practice of publishing mathematical articles in various fields, such as *Bounded Gaps between Primes*, *Quantum Mechanics in the Sky* and *Puzzles, Prisoners and Probability*, so that readers who are interested in any part of mathematics will find something interesting.

This year, the publication of Qarch, our problems journal, has been resumed. I would like to thank Leo Lai, the Qarch Editor, and his publication team, for their great effort in making Qarch another success.

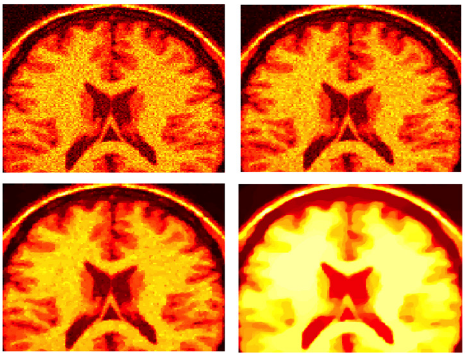
It is my privilege to work with a wonderful editorial team, whom I would like to thank for all of their hard work. I would also like to thank former editor Jasper Bird and Yanitsa Pehova for their invaluable advice and tremendous support, as well as our writers, our sponsors, the Archimedeans, and our readers. Without you none of this would have been possible. Happy reading!

Long Tin Chan
December, 2015

Quantum Mechanics in the Sky 8



Learning to Denoise 26



CONTENTS

The Archimedeans

Daochen Wang, President 2014 – 2015

2014-15 has been a great year for the Archimedeans, a year which marked 80 years since our society's founding.

In a spirit of celebration, we started our year by co-hosting the inaugural Maxwell Lecture with the Cambridge University Physics Society. The evening lecture, delivered by Professor Gerard t'Hooft on his recent proposal of a classical treatment of quantum laws, was enthusiastically received by a record-breaking audience of over 500. On the next morning was our Freshers Talk given by Professor Tim Gowers following which the society welcomed nearly 100 Freshers as new Life Members in the Small Examinations Room. No expenses were spared to ensure an abundance of pizza.

Going into Michaelmas term proper we hosted talks on a wide ranging variety of subjects. These went from mathematical biology to geometry to string theory. Personal highlights included "The Geometry of Speech" by Professor John Ashton where we saw entire languages being converted into manifolds and "The Mystery of Spinors" by Professor Michael Atiyah who never ceases to amaze all of us with his ability to explain apparently complicated concepts with so much clarity and ease.

Our energy and momentum did not let go in Lent. Thanks to Leo Lai, the term began by our relaunching the dormant QARCH problems which asks, among other things, for a proof of a disguised Riemann Hypothesis. The majority of the talks this term focused on combinatorics and the most memorable moment must be during Professor Imre Leader's tutorial on "Eating and Racing" when we the audience literally "laughed out loud" in response to the end of a proof using a method called strategy stealing. As tradition dictates, we also hosted the Annual Problems Drive and the Annual Dinner towards the end of Lent. Both events were capably organised by Andrew Yiu.

It has been a great honour to be involved with the Archimedeans. I thank all of this year's committee for their hard work and I am sure Andrew and his team for 2015-16 will do an even better job as we go into our 81st year.

The Committee 2014 – 2015

President
Daochen Wang (Sidney Sussex)

Vice-President
Emily Bain (Emmanuel)

Corporate Officer
Adam Goucher (Trinity)

External Secretary
Andrew Yiu (Christ's)

Internal Secretary
Ivan Loh (St Edmund's)

Treasurer
Sreya Saha (Murray Edwards)

Events Managers
Daan van de Weem (Homerton)

Publicity Officer
Eiki Norizuki (Jesus)

Webmaster
Joe Tomkinson (Trinity)

Using Asymptotic Methods to Investigate the Noise Generated by Aeroplanes

Dr. Lorna Ayton
Junior Research Fellow, DAMTP

Aeroacoustics is a branch of fluid dynamics which studies the noise generated by flow interactions with aerodynamic bodies. Arguably the most dominant studies focus on the noise generated by aeroengines, like those found on modern passenger aeroplanes, because, let's face it, they're pretty darn noisy! Aeroengines are complicated machines, with four key sources of noise, one of which is generated by the interaction of the blades within the engine with the air flow through the engine, and this is the noise discussed here.

The flow impinging on an initial layer of blades is assumed to be steady and uniform. The blades rotate and generate unsteady vorticity in the flow downstream, which goes on to interact with a second layer of blades and this generates noise. We study this noise and hope to relate it to parameters such as the blade geometry, the Mach number of the background steady flow, and the frequency of the unsteady vorticity, with the hope that this will allow us to find an optimal set of parameters that will reduce the overall unwanted noise, but not hinder the performance of the engine.

The noise of specific interest is that propagating away from the engine, because unless you're onboard the aeroplane, you'll only hear the noise from far away. A common approach to finding the noise is to ask a computer for help, however, because we want to compute the far-field noise, we need a very large computational domain. Further complications arise because high frequency noise is a key contributor. To retain accuracy as the frequency is increased, the grid resolution must also be increased. Overall this results in codes which have long runtimes, or lose accuracy at high frequencies. Analysing the effects of altering, for example, the blade geometry, would therefore take a very long time (assuming your code is accurate), and moreover would only illustrate a trend in behaviour as you alter the geometry, rather than yield a functional relationship between the noise generated and the geometry of the blades.

We therefore analytically model the far-field noise generated by so-called "blade-blade" interactions to obtain an asymptotic solution which depends on all the parameters of the problem. The solution allows us to quickly and efficiently discover the effects of altering any parameter on the far-field noise. Thankfully, a differential equation describing the interaction of unsteady vorticity with a thin aerofoil already exists [2], although it's not particularly pleasant:

$$\frac{\partial^2 h}{\partial \phi^2} + \frac{\partial^2 h}{\partial \psi^2} + k^2 w^2 (1 - 2\beta_\infty^2 \epsilon q) h \\ + \frac{(\gamma+1) M_\infty^4 \epsilon q}{\beta_\infty^2} \left(\frac{\partial^2 h}{\partial \psi^2} + 2ik\delta \frac{\partial h}{\partial \phi} + k^2 (w^2 + \delta^2) h \right) \\ - \frac{(\gamma+1) M_\infty^4}{\beta_\infty^2} \epsilon \frac{\partial q}{\partial \phi} \left(\frac{\partial h}{\partial \phi} - ik\delta h \right) = \epsilon S e^{i\Omega}$$

The coordinates, (φ, ψ), are defined as the velocity potential and the streamfunction of the background steady flow around the aerofoil (so that the solid aerofoil surface is defined as $\psi = 0$, $\phi \in [0, \phi_e]$, on which a zero normal flow boundary condition is imposed), and the function h describes the sound generated by the interaction. There are various constants, M, β_∞, γ, w, δ, which allow us to consider different background flow conditions.

The constant, $\epsilon \ll 1$, measures the relative size of the aerofoil thickness and camber to its length, whilst the constant, $k \gg 1$, is the (reduced) frequency of the incident vortical disturbance. The function $q(\varphi, \psi)$ depends on the exact geometry of the aerofoil, including effects of thickness, camber and angle of attack to the background steady flow. The source terms, $S(\varphi, \psi)$ and $\Omega(\varphi, \psi)$, describe the vortical disturbance.

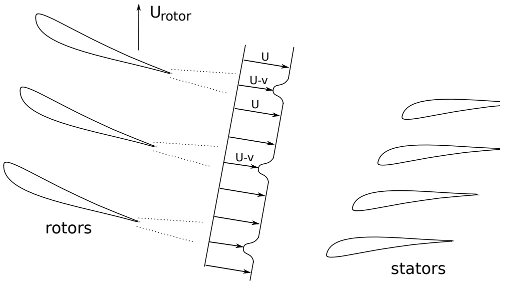


Figure 1: The model "blade-blade" interaction problem; rotating rotors create unsteady wakes in the steady background flow which interact with the stationary stators.

If the aerofoil is reduced to a flat plate ($\epsilon = 0$) we have only the Helmholtz equation to solve, however for any realistic geometry, all terms contribute. To analytically solve the equation we make use of the large and small parameters, k and , to construct an asymptotic solution,

$$h = h_0(\phi, \psi) + \epsilon\sqrt{k}h_1(\phi, \psi) + O(\epsilon, k^{-1})$$

where $\epsilon, k^{-1} \ll \epsilon\sqrt{k} \ll 1$. This process is illustrated by the following example.

Example

Consider the differential equation

$$\epsilon y''(x) + (1 + 2\epsilon)y'(x) + 2y(x) = 0$$

with boundary conditions, $y(0) = 0$ and $y(1) = 1$, where $\epsilon \ll 1$. If $x = O(1)$, then $y''(x), y'(x) = O(y)$, and the largest terms in the equation are *y'* and 2*y* which we suppose to be *O*(1) because of the boundary condition $y(1) = 1$. If $x = O(\epsilon)$, then $y''(x) = O(\epsilon^{-2}y)$ and the largest terms in the equation are *y"(x)* and *y'(x)* both of which are $O(\varepsilon^{-1}y)$. We see that for different scalings of x we obtain different dominant terms in the differential equation. To solve this we therefore construct two solutions, one which is valid for $x = O(1)$ and one when $x = O(\epsilon)$.

For $x = O(1)$, we propose a solution

$$y = y_0(x) + y_1(x) + O(\epsilon^2)$$

Substituting this into the equation and equating *O(1)* terms gives $y_0'(x) + 2y_0(x) = 0$, and the boundary condition is $y_0(1) = 1$. The solution is $y_0 = e^{2(1-x)}$. Considering the *O(ε)* terms, we require $y_1'(x) + 2y_1(x) = -y_0''(x) - 2y_0'(x)$, with boundary condition $y_1(1) = 0$ (since $y(1) = 1$ is satisfied at first order with y_0). This gives solution $y_1 = 0$, so $y(x) \approx e^{2(1-x)} + O(\epsilon^2)$

For *x* = *O(ε)*, we change variables, $y \to Y$, and $x \to X = x/\epsilon$. The equation becomes

$$Y''(X) + (1 + 2\epsilon)Y'(X) + 2\epsilon Y(X) = 0 \quad ,$$

with boundary condition *Y(0)* = 0. We suppose $Y(X) = Y_0(X) + \epsilon Y_1(X) + O(\epsilon^2)$, and solve for $Y_{0,1}$ similarly, giving

$$Y(X) \approx A(1 - e^{-X}) + \epsilon[B(1 - e^{-X}) - 2AX]$$

The constants *A* and *B* are unknown constants of integration; one arises from the second order differential equation for Y_0 and one for Y_1.

To determine these constants and finish the solution we use the principle of matched asymptotic expansions [3], which essentially says that as *X* becomes very large in *Y*, we must have the same solution as when *x* becomes very small in *y*. I.e. as we break out of the *x* = *O(ε)* region from the *Y* solution by making *X* large, we should get the same answer as when we break into the region from the other side by taking *x* to be small in the *y* solution.

For small *x*,

$$y(x) \approx e^2(1 - 2x + O(x^2)) \ ,$$

and for large *X*,

$$Y(X) \approx A - 2A\epsilon X + \epsilon B + O(\epsilon^2)$$

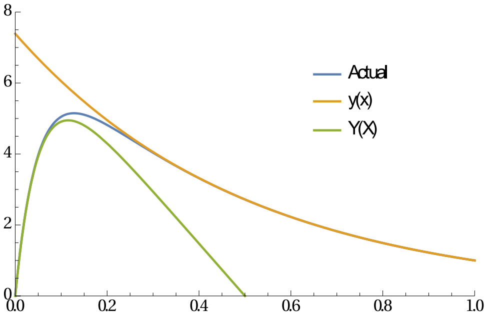


Figure 2: Comparisson of the approximate solutions, *y* and *Y*, against the actual solution.

Substituting *X* = *x/ε* into this second expansion and equating the coefficients of *x* and *ε* gives *A* = e^2 and *B* = 0. This completes the solution for *Y(X)*.

Figure 2 shows the two solutions, y and Y, plotted against the actual solution to the original differential equation, when ε = 0.05. We see that the approximations give very good agreement to the actual solution except in a small region near to *x* = ε. As long as we are only interested in the solution away from this erroneous region, we could use our approximation rather than the actual solution.

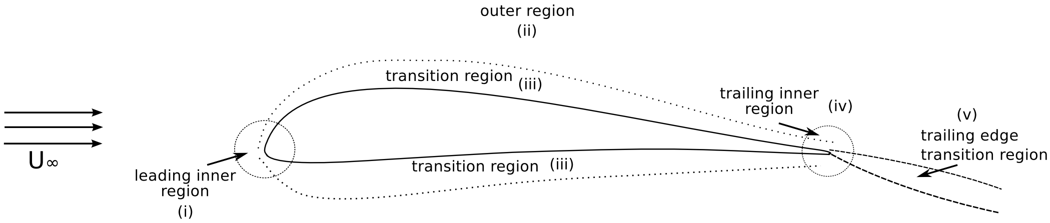


Figure 3: Asymptotic regions around an aerofoil required to obtain the far-field, "outer", solution. The coordinates in the inner regions (i) and (iv) scale as $O(k^{-1})$, in the transition regions (iii) and (v) scale as $O(k^{-1/2})$ and in the outer region (ii) scale as $O(1)$.

For the aeroacoustic question, we are only interested in the solution far away from the aerofoil. We obtain that solution by using the principle of matched asymptotic expansions for all of the regions illustrated in Figure 3. An example solution is plotted in Figure 4, which shows a polar plot of the magnitude of the far-field acoustic pressure as a function of observer angle. The aerofoil is located at the origin. We see a decrease in noise generated upstream (the left half plane) as we increase thickness, but altering thickness has much less of an effect downstream (the right half plane). The aerofoils used in Figure 4 have a small amount of camber, which results in the asymmetric plot - the noise generated above the aerofoil is not the same as that below, and we can see that at certain angles below the aerofoil (LHP) the zero thickness aerofoil generates less noise than those with thickness, however above the aerofoil, the noise decreases with thickness.

Further details of the method and solution to the aerofoil noise generation problem can be found in [1].

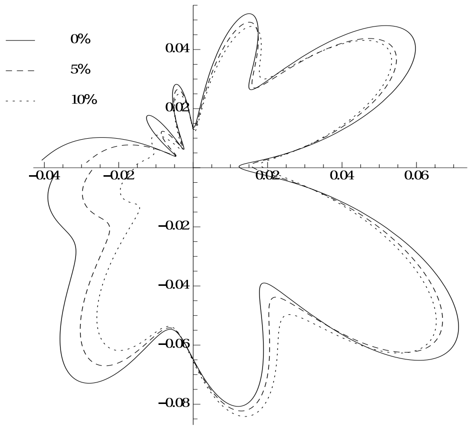


Figure 4: Far-field acoustic pressure magnitude generated by a vortical interaction in steady flow with a cambered NACA 4-digit aerofoil, for varying observer angle, θ. The legend denotes the percentage thickness of the aerofoil compared to its length.

References

[1] Ayton, L. J. & Peake, N. On high-frequency noise scattering by aerofoils in flow. Journal of Fluid Mechanics 734, 144–182, 2013.

[2] Kerschen, E. J. & Myers, M. R. 1987 Perfect gas effects in compressible rapid distortion theory. AIAA Journal 25, 504–507.

[3] Van Dyke, M. 1975 Perturbation methods in fluid mechanics . Parabolic Press.

Quantum Mechanics in the Sky

Dr. Daniel Baumann
Reader in Theoretical Physics, DAMTP

Imagine you look at the night sky. You randomly select a patch of the sky only a fraction of the size of the full moon. To the naked eye it will look pitch black. This doesn't mean that there is no light coming from that region. It just means that less than seven photons per second are entering your eyes, in which case your brain cannot form an image. A camera, however, can collect photons over a long period of time. This is what the camera on the Hubble Space Telescope (HST) has done. Instead of looking for just an instant, the HST collected light for more than 12 days. The result is one of the most stunning astronomical images ever produced: see fig. 1. Every object in this picture is an entire galaxy! We see a few thousand of them, each containing billions of stars. This becomes even more remarkable if we remember that this tiny patch of the sky was selected at random. Any other randomly selected region in the sky would look essentially the same. From this we can estimate that the observable universe contains some trillion galaxies and a few billion trillion stars. In this essay, I will describe our best answer to the question: Where did it all come from?

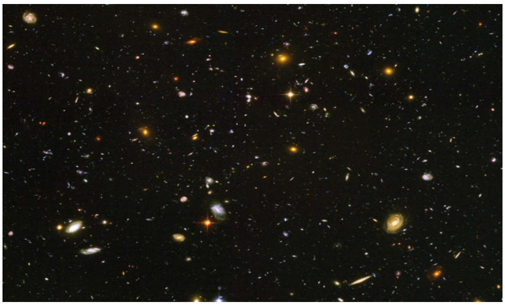


Figure 1: *The Hubble Ultra Deep Field. The image contains thousands of galaxies in an area that is only a fraction of the size of the moon.*

The Cosmic Microwave Background

Let us start 380,000 years after the Big Bang. The universe had just cooled enough for the first atoms to form. From this moment on, light was able to propagate freely. Today, 13.8 billion years later, we observe this afterglow of the Big Bang as the cosmic microwave background (CMB).

The fact that the intensity of the CMB varies across the sky (see fig. 2) means that the matter in the early universe wasn't distributed uniformly. Over time, and under the influence of gravity, these matter fluctuations grew. Regions of space that contained more than an average amount of matter, accreted matter from their surroundings and therefore got denser. Eventually, the local density became high enough for galaxies, stars and planets to form. This part of the story is well-understood; what needs explaining is where the small seed fluctuations came from.

On closer inspection, we realise that there is a problem with the map shown in fig. 2. The fluctuations in the map aren't just noise, but display non-random correlations across large regions of the sky. On the other hand, the universe was very young when the CMB was created. Even signals travelling at the speed of light wouldn't have gotten very far. In particular, regions separated by more than the size of the white circles in fig. 2 didn't have enough time to communicate with each other. We say that they were out of *causal contact* when the background radiation was created. Having anything but random noise over large patches of the sky therefore seems to violate causality.

The Causality Problem

It was this *causality problem* that Alan Guth was thinking about in the night of December 6, 1979. Guth is now a professor at MIT, but at the time he was a researcher at the Stanford Linear Accelerator Centre, struggling to find a permanent job. Having been trained in particle physics, Guth knew very little about cosmology. However, a year earlier he had learned about the conceptual problems of the standard Big Bang theory from a talk by Robert Dicke. In the night of December 6th, he was led to thinking about it again. What happened next has become history.

In a flash of insight, Guth understood that the problems of the standard cosmology would be resolved if the early universe went through a brief period of superluminal expansion. He imagined that, for a fraction of a second, space expanded nearly exponentially.

Points that were initially very close to each other got stretched apart at an enormous rate, faster than the speed of light. According to the theory of *cosmic inflation*, everything we see in the sky today was initially compressed into a tiny, causally-connected, region of space. Faster-than-light expansion of space then blows up such a microscopic region to enormous size and gives the illusion of a causality problem. In reality, the causality problem would just be an artefact of assuming that the expansion of space never exceeded the speed of light.

Although inflation solves the causality problem, for explaining the origin of structure it seems to be a disaster. Any seed fluctuations that might have existed before inflation are spread apart at such an enormous rate that nothing survives after inflation. Inflation seems to make the universe empty and featureless. This is clearly not the universe we live in. Fortunately, quantum mechanics comes to the rescue.

QM to the rescue!

In quantum mechanics, empty space is never completely empty, as this would violate the Heisenberg uncertainty principle. Instead, even the vacuum needs to fluctuate: energy and particles can appear and disappear spontaneously (see fig. 3). These quantum vacuum fluctuations have many important consequences for physics. Sometimes, they have observable effects like the *Lamb shift* and the *Casimir force*. The Lamb shift refers to a shift in the energy levels of hydrogen atoms, as the electrons in the atoms interact with the quantum vacuum. This effect was measured by Willis Lamb in 1947. Shortly after, Hans Bethe calculated the Lamb shift by showing how the fluctuations in the vacuum disturb the electron while it is revolving around the proton. Besides these observable effects, vacuum fluctuations are also crucial for the theoretical consistency of modern physics. Quantum field theory and the Standard Model of particle physics would make no sense without quantum vacuum fluctuations.

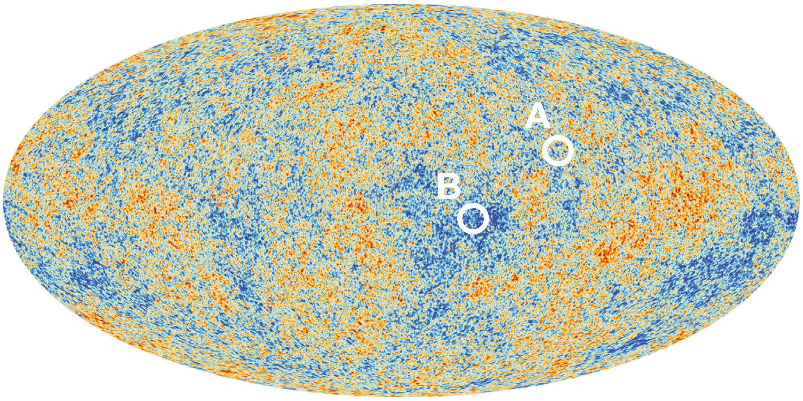


Figure 2: *Temperature variations in the cosmic microwave background as observed by the Planck satellite. Red (blue) spots are hotter (colder) than the average temperature, reflecting density variations in the primordial plasma. The white circles indicate the maximal size of regions that could have exchanged signals before the time that the CMB was created (according to the standard Big Bang theory). The points A and B naively were out of causal contact.*

There is therefore no doubt that quantum vacuum fluctuations really exist.

Under ordinary circumstances, vacuum fluctuations have tiny effects. However, during inflation they become dramatically important. Because of random quantum fluctuations, the inflationary energy density in some parts of the universe was slightly higher than average. In these regions, inflation lasted a bit longer, while in other regions inflation terminated earlier. This difference in the evolution amplifies the fluctuations, resulting in significant density fluctuations after inflation.
The exponential expansion of space during inflation stretches the vacuum fluctuations to enormous scales. Small-scale correlations in the quantum vacuum can therefore leave imprints in the large-scale density fluctuations. Moreover, the combination of inflation and quantum mechanics doesn't just make the apparently acausal correlations in the CMB possible, it also predicts the specific form that these correlations should take. Roughly speaking, inflation predicts the probability *P(θ)* that two points in the sky separated by an angle θ are both hotter (or colder) than average. The function *P(θ)* is expected to have a characteristic shape, i.e. the amount of correlation varies in a specific way as a function of angular separation. Amazingly, in recent years, it has become possible to measure the fluctuations in the CMB accurately enough to determine the function *P(θ)* and compare it to the prediction from inflation. The stunning result is shown in fig. 4.

What next?

So, are we done? Do we understand everything?

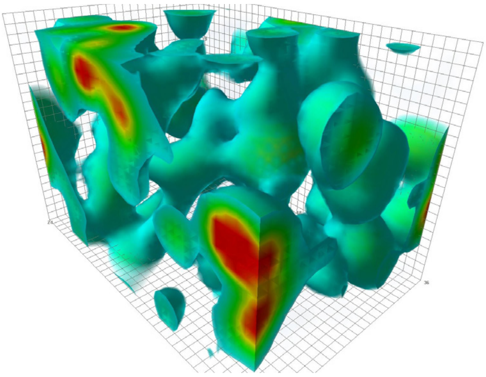


Figure 3: *Quantum fluctuations in quantum chromodynamics (QCD). The energy density associated with the gluon fields fluctuates spontaneously.*

Far from it. While fig. 4 is compelling evidence for inflation, it doesn't yet exclude alternative explanations for the origin of structure.
Moreover, the physics of inflation remains mysterious. We still don't know what caused the burst of inflationary expansion. In fact, the requirements for successful inflation are rather dramatic. The expansion has to be superluminal and must increase the size of the universe by more than a factor of 10^{26} in just 10^{-33} seconds! This can happen in General Relativity if the universe is filled with a source of negative pressure. However, ordinary matter doesn't behave in this way, so it is likely that inflation requires more exotic physics. We hope that future observations will us hints for what this physics might be. In the meantime, theorists like myself are working hard to discover mechanisms that could explain the inflationary epoch or find alternatives. At stake is nothing less than a complete understanding of the origin of all structure in the universe!

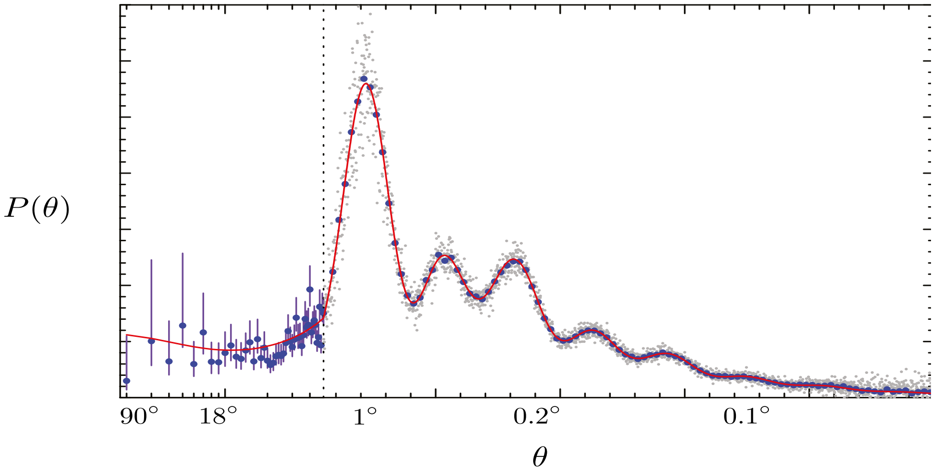


Figure 4: *Comparison of the latest measurements of the fluctuations in the CMB (bluepoints) with the prediction from inflation (red curve).*

Generating a Sequence of Pythagorean Triples Using Lucas Sequence

Chanchal Singh

Introduction:

A right angled triangle with sides as positive integers is also known as Pythagorean triangle. We will denote the length of the small side by s, the length of the large side by l and the hypotenuse length by h of a Pythagorean triangle. As is well known (Pythagoras Theorem),[1].

$$s^2 + l^2 = h^2 \quad (1)$$

Instead of working with Pythagorean triangles geometrically or working with equations like (1), it is more convenient to work with triples (s, l, h), known as Pythagorean triples.

It is well known that by using Euclid's formula (stated below) one can generate infinite number of Pythagorean triples (but not all of them). However generating sequences of Pythagorean triples help exhibit interesting relations within and among triples besides (hopefully) raising mathematical curiosity for the reader in establishing links between sequences of triples with other well known mathematical structures as shown in O' Shea [1] and [2].

To refresh reader's memory, we state **Euclid's Formula:** For any two positive integers m and n, with $m > n$, a Pythagorean triple is formed by $2mn$, $m^2 - n^2$, $m^2 + n^2$. Obviously $m^2 + n^2$ represents the hypotenuse length, the large (small) side is represented by larger (smaller) of the remaining two terms.

O'Shea's presentation is restricted to one sequence of Pythagorean triples, namely (3, 4, 5), (5, 12, 13), (7, 24, 25), (9, 40, 41), (11, 60, 61), (13, 84, 85), … and each triple in the sequence is such that
$h - l = 1$.

In our derivations we also restrict to one sequence of Pythagorean triples, generated in four different ways using Lucas sequence. In our sequence of Pythagorean triples $h - l$ increases as components of triples increase in magnitude.

Lucas sequence is specified by letting $L_1 = 1$, $L_2 = 3$, and $L_n = L_{n-1} + L_{n-2}$ for $n = 3, 4, 5, \ldots$ In the lengthy form it is: 1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199, 322, 521, …
Sometimes we will use symbols L_1, L_2, L_3, L_4, ... and identify the above string of numbers, also known as Lucas numbers, with symbols

Observation 1: Using any subsequence of natural numbers (not just Lucas sequence) and constructing fractions of the form of $(c/d + d/c)/2$ or $(c/d - d/c)/2$ with c and d positive integers and $c > d$ to generate a sequence of Pythagorean triples will always end up with components of triples satisfying the Eulicd's formula. An illustration of this observation is in Section (A) below.

Generating a Sequence of Pythagorean Triples in Different Ways.

Section (A): Select subsequences of size two using Lucas sequence, starting with number 3 as follows. So we have 3, 4; 4, 7; 7, 11; 11, 18; and so on. Then using each subsequence, construct fractions [4/3 + 3/4]/2; [7/4 + 4/7]/2; [11/7 + 7/11] /2 and so on After simplifying each product of fractions into a form a/b,
(i) The numerator is the hypotenuse length. (ii) The denominator is the large leg length.
(iii) the small leg length is computed as follows: suppose the sum of two fractions within
the brackets is $c/d + d/c$ with $c > d$. Then the small leg length is given by $(c^2 - d^2)$

$[4/3 + 3/4]/2 = 25/24 \rightarrow (4^2 - 3^2, 24, 25) \rightarrow (7, 24, 25)$

$[7/4+4/7]/2 = 65/56 \rightarrow (7^2 - 4^2, 56, 65) \rightarrow (33, 56, 65)$

$[11/7 + 7/11]/2 = 170/154$
$\rightarrow (11^2 - 7^2, 154, 170) \rightarrow (72, 154, 170)$

$[18/11 + 11/18]*(1/2) = 445/396$
$\rightarrow (18^2 - 11^2, 396, 445) \rightarrow (203, 396, 445)$

Note that in each of the above four triples,
$h = l + (c - d)^2$

The n^{th} term, for $n = 1, 2, 3, \ldots$ is given by
$[L_{n+2}/L_{n+1} + L_{n+1}/L_{n+2}]/2 = [(L_{n+2})^2 + (L_{n+1})^2] / [2L_{n+1} L_{n+2}] \rightarrow ((L_{n+2})^2 - (L_{n+1})^2, 2L_{n+1}L_{n+2}, (L_{n+2})^2 + (L_{n+1})^2)$

We see that the nth triple turns out to be the same as given by the Euclid's formula.

The reader may easily verify that the nth triple also satisfy the condition that $h = l + (L_{n+2} - L_{n+1})^2$.

Section (B): Select subsequences of size three from the Lucas sequence, starting with number 1 as follows. So we have: 1, 3, 4; 3, 4, 7; 4, 7, 11; 7, 11, 18; and so on. Then do the following with each subsequence:
(i) Multiply the first number by the sum of the middle and the last number. This gives the length of the small leg. (ii) Double the product of the middle number with the last number. This gives the length of the large leg. (iii) The hypotenuse length is the large leg length plus the square of the first term of the subsequence.

$1, 3, 4 \rightarrow 1\times(3+4),\ 2\times(3\times4),\ 2\times(3\times4)+1^2 \rightarrow (7, 24, 25)$

$3, 4, 7 \rightarrow 3\times(4+7),\ 2\times(4\times7),\ 2\times(4\times7)+3^2$
$\rightarrow (33, 56, 65)$

$4, 7, 11 \rightarrow 4\times(7+11),\ 2\times(7\times11),\ 2\times(7\times11)+4^2$
$\rightarrow (72, 154, 170)$

$7, 11, 18 \rightarrow 7\times(11+18),\ 2\times(11\times18),\ 2\times(11\times18)+ 7^2 \rightarrow$ (203, 396, 445) and so on

The n^{th} term, for $n = 1, 2, 3, \ldots$ is given by
$L_n, L_{n+1}, L_{n+2} \rightarrow (L_n (L_{n+1}+L_{n+2}), 2L_{n+1}L_{n+2}, 2L_{n+1}L_{n+2}+L_n{}^2)$

Now by definition of Lucas sequence, $L_{n+2} = L_{n+1} + L_n$ or $L_n = L_{n+2} - L_{n+1}$. So we may write the n^th triple as

L_n, L_{n+1}, L_{n+2}
$\rightarrow ((L_{n+2}-L_{n+1})(L_{n+2}+L_{n+1}),\ 2L_{n+2}L_{n+1},\ 2L_{n+2}L_{n+1} + (L_{n+2}-L_{n+1})^2)$
$\rightarrow (L_{n+2}{}^2 - L_{n+1}{}^2, 2L_{n+2}L_{n+1},\ L_{n+2}{}^2 + L_{n+1}{}^2)$

Section (C): Select subsequences of size three, skipping every third number in the Lucas sequence as follows. So we have 1, 3, 7; 3, 4, 11; 4. 7, 18; 7, 11, 29; and so on Next do the following for each subsequence:

(i) Multiply the first and the last numbers. This gives the length of the small leg.
(ii) Multiply the middle number by the sum of first and the last numbers. This gives the length of large leg.
(iii) The square of the first number in each subsequence plus the large leg length is the hypotenuse length.

$1, 3, 7 \rightarrow 1 \times 7, 3 \times (1 + 7), 3 \times (1 + 7) + 1^2$
$\rightarrow (7, 24, 25)$

$3, 4, 11 \rightarrow 3 \times 11, 4 \times (3 + 11), 4 \times (3 + 11) + 3^2$
$\rightarrow (33, 56, 65)$

$4, 7, 18 \rightarrow 4 \times 18, 7 \times (4 + 18), 7 \times (4 + 18) + 4^2$
$\rightarrow (72, 154, 170)$

$7, 11, 29 \rightarrow 7 \times 29, 11 \times (7 + 29), 11 \times (7 + 29) + 7^2 \rightarrow$ (203, 396, 445) and so on

The n^{th} term, for $n = 1, 2, 3, \ldots$, is given by
L_n, L_{n+1}, L_{n+3}
$\rightarrow (L_nL_{n+3}, L_{n+1}(L_n + \mathrm{L}_{\mathrm{n}\ +3}), L_{n+1}(L_n+\mathrm{L}_{\mathrm{n}\ +3})+L_n{}^2)$
(2)

Recall that by definition of Lucas sequence, $L_n = L_{n+2} - L_{n+1}$ and $L_{n+3} = L_{n+2} + L_{n+1}$. Replacing L_n and L_{n+3} in the Pythagorean triple in (2), we have

$L_n, L_{n+1}, \mathrm{L}_{\mathrm{n}\ +3}$
$\rightarrow ((L_{n+2} - L_{n+1})(L_{n+2} + L_{n+1}), L_{n+1} (L_{n+2} - L_{n+1} + L_{n+2} + L_{n+1}), L_{n+1} (L_{n+2} - L_{n+1} + L_{n+2} + L_{n+1}) + (L_{n+2} - L_{n+1})^2)$
$\rightarrow (L_{n+2}{}^2 - L_{n+1}{}^2,\ 2L_{n+1}L_{n+2},\ L_{n+2}{}^2 + L_{n+1}{}^2)$.

Section (D): Select subsequences of four consecutive Lucas numbers starting with 1 as shown below and perform the following operations with each subsequence:

(i) Multiply the end numbers of each subsequence. This gives the small leg length.
(ii) Double the product of two middle numbers. This is the large leg length.
(iii) The hypotenuse length is the large leg length plus square of the first number in each subsequence.

$1, 3, 4, 7 \rightarrow 1 \times 7, 2\ (3 \times 4),\ 2\ (3 \times 4) + 1^2$
$\rightarrow (7, 24, 25)$

$3, 4, 7, 11 \rightarrow 3 \times 11, 2\ (4 \times 7),\ 2\ (4 \times 7) + 3^2$
$\rightarrow (33, 56, 65)$

$4, 7, 11, 18 \rightarrow 4 \times 18, 2\ (7 \times 11),\ 2\ (7 \times 11) + 4^2$
$\rightarrow (72, 154, 170)$

7, 11, 18, 29 → 7 × 29, 2 (11 × 18), 2 (11 × 18) + 7^2 → (203, 396, 445)

The n^{th} term, for $n =$ 1, 2, 3,. . . is given by $L_n, L_{n+1}, L_{n+2}, L_{n+3} \to (L_nL_{n+3}, 2L_{n+1}L_{n+2}, 2L_{n+1}L_{n+2}+L_n^2)$ (3) Since $L_{n+3} = L_{n+2} + L_{n+1}$ and $L_n = L_{n+2} - L_{n+1}$, the triple in (3) reduces to L_n, L_{n+1}, L_{n+2}, $L_{n+3} \to$ $(L_{n+2}^2 - L_{n+1}^2, 2L_{n+1}L_{n+2}, L_{n+2}^2+L_{n+1}^2)$, as shown above.

Observation 2: We note an interesting happening in the generated sequence of Pythagorean triples. The sum of the lengths of the three sides of each Pythagorean triangle is equal to the length of the large leg of the next triangle. For instance 7 + 24 + 25 = 56, 33 + 56 + 65 = 154 and so on. This is true in general, as we show below. Sum of the three sides of the n^{th} Pythagorean triangle is

$$L_nL_{n+3} + 2L_{n+1}L_{n+2} + 2L_{n+1}L_{n+2} + L_n^2$$
$$= L_{n+2}^2 - L_{n+1}^2 + 2L_{n+1}L_{n+2} + L_{n+2}^2 + L_{n+1}^2,$$

as shown above,

$$= 2L_{n+2}^2 + 2L_{n+1}L_{n+2} = 2L_{n+2}(L_{n+2}+L_{n+1})$$

The $(n+1)^{\text{th}}$ triple is $(L_{n+1}L_{n+4}, 2L_{n+2}L_{n+3}, 2L_{n+2}L_{n+3} + L_{n+1}^2)$

The large leg length in the $(n+1)^{\text{th}}$ triple is $2L_{n+2}L_{n+3} = 2L_{n+2}(L_{n+2}+L_{n+1})$

Observation 3: Another interesting situation is the following. Let (s_n, l_n, h_n) be the n^{th} Pythagorean triple in the generated Pythagorean sequence of triples. We note that

$2s_1 + 10 = 2(7) + 10 = 24 = l_1$,
$2s_2 - 10 = 2(33) - 10 = 56 = l_2$,
$2s_3 + 10 = 2(72) + 10 = 154 = l_3$,
$2s_4 - 10 = 2(203) - 10 = 396 = l_4$ and so on.

In general one may conjecture, that in the n^{th} triple, $2s_n \pm 10 = l_n$ where one adds 10 if n is odd and subtract 10 if n is even. The conjecture is correct. A proof may be outlined as follows: One needs to show that $2(L_{n+2}^2 - L_{n+1}^2) \pm 10 = 2L_{n+1}L_{n+2}$ which is equivalent to (using some basic Lucas sequence identities) showing that $L_{n+1}(L_{n+1} - L_n) - L_n^2 = 5$ which can be shown to hold using mathematical induction.

So the components of triples of the generated Pythagorean sequence are interestingly tied together in more than one way as pointed out in Observations 2 and 3.

Observation 4: As seen in the first four triples of the generated Pythagorean sequence of triples, each hypotenuse length is a multiple of 5. This is no surprise since it is given by $L_{n+2}^2 + L_{n+1}^2$ for $n = 1, 2, 3, \ldots$ and the sum of squares of any two consecutive numbers of the Lucas sequence is a multiple of 5.

References

1. O, O'Shea, Simple – But Little Unknown– Methods of Generating Pythagorean Triples Journal of Recreational Mathematics, 37(1), pp. 1-8, 2008

2. O. O'Shea, More Simple – But Little Unknown– Methods of Generating Pythagorean Triples, Journal of Recreational Mathematics, 37(3), pp. 187-191, 2008

Bounded Gaps between Primes

Prof. Ben Green

Waynflete Professor of Pure Mathematics , University of Oxford

In May 2013, Yitang Zhang stunned the mathematical world by proving the following result, which we shall loosely refer to as “bounded gaps between primes”.

Theorem 1 (Zhang)

There is an absolute constant H such that there are infinitely many pairs of distinct primes differing by at most H.

The celebrated twin prime conjecture asserts that one may take $H = 2$. Zhang obtained $H = 70000000$, but subsequent feverish activity by a massively collaborative Polymath project reduced this to 4680. In late 2013, James Maynard and Terry Tao found a much simpler proof of Zhang’s result giving $H = 600$, and a further *Polymath* project based on this work has, at the time of writing, reduced H to 246.

Something I wish to emphasise in this article is that Zhang’s result should be thought of as the culmination of ideas about prime numbers developed by many of the great analytic number theorists of the 20th century. It melds two important bodies of work:

1. Ideas of Goldston, Pintz and Yldrm, building on work of Selberg and others, establishing a link between the distribution of primes in arithmetic progressions and small gaps between primes;

2. Ideas of Bombieri, Fouvry, Friedlander and Iwaniec, building on but going well beyond work of Bombieri-Vinogradov, pinning down strong results about how primes are distributed in arithmetic progressions.

Primes in progressions

The distribution of primes in progressions is central to the whole story, so let us begin with a brief tour of that subject. We begin by recalling the prime number theorem, which states that $\pi(X)$, the number of primes less than or equal to X, is roughly $X/\log X$. The fact that $\log X \to \infty$, which means the primes have density tending to zero, of course explains why Zhang’s theorem is not at all obvious: the *average* gap between primes less than X is about $\log X$. One might also remark that Zhang’s theorem is not at all *surprising*, either, since a random set of $X/\log X$ integers less than X will, for X large, have many pairs spaced by at most 2. In fact, it will have many pairs spaced by at most 1, something not true for the primes themselves of course - to model the primes by a random set one has to be more careful and take account, for example, of the fact that most primes are odd. It is convenient to state the prime number theorem a little differently by introducing the *von Mangoldt function* $\Lambda(n)$, which is basically defined to equal $\log n$ if n is a prime[1]. Then the prime number theorem is equivalent to saying that

$$\sum_{x \leq X} \Lambda(x) \approx X$$

How many primes x satisfy some additional congruence condition $x \equiv c \pmod{d}$? Clearly (apart from for very small primes $x \leq d$) we must have c coprime to d, but there is no obvious additional restriction. In fact, one expects the primes to be equally distributed amongst the $\phi(d)$ residue classes coprime to d. If this is the case, we have

$$\sum_{x \leq X, x \equiv c \pmod{d}} \Lambda(x) \approx \frac{X}{\phi(d)}$$

When this is true for a given value of d and for all c coprime to d then we shall say that the primes are *nicely distributed*[2] modulo d. Proving this statement is another matter. Think of X tending to infinity: then one would like to know that the primes are nicely distributed modulo d for all d up to some limit, which it would be nice to take as large as possible. This is, however, only known when d is less than a power of $\log X$ (in fact d can be less than any power of $\log X$, a statement known as the Siegel-Walfisz theorem).

It turns out that proving that the primes are nicely distributed for d up to about $X^{1/2}$ is equivalent to the Generalised Riemann Hypothesis, so one should not expect too much progress soon. Remarkably, one can do far better if one is prepared to know that the primes are nicely distributed only for[3] *almost all d*. The classic result in this vein is the Bombieri-Vinogradov theorem, which asserts that the primes are nicely distributed modulo d for almost all $d \leq X^{1/2}$. The Bombieri-Vinogradov theorem is often described as a kind of *Riemann Hypothesis on Average*.

[1] And $\log p$ if $n = p^k$ is a prime power.

[2] Of course, this is only an informal definition and not a rigorous one because we have not elaborated upon the meaning of $\approx$.

[3] Once again, this is an informal definition. To make it rigorous, one would need to elaborate upon the meaning of almost all as well as the $\approx$ notation from earlier.

It is suspected that even more is true, a conjecture known as the Elliott-Halberstam Conjecture. There is a different Elliott-Halberstam conjecture *EH(θ)* for each value of $\theta < 1$, and these conjectures get stronger as θ increases. Here is a rough statement:

Conjecture 2 (Elliott Halberstam conjecture EH(θ))

The primes are nicely distributed modulo d for most values of d up to X^θ. Currently, we cannot prove this statement for any value of $\theta > 1/2$.

GPY and BFFI

Goldston, Pintz and Yıldırım (henceforth referred to as GPY) established a remarkable link between the problem of finding bounded gaps between primes and the Elliott-Halberstam conjecture.

Theorem 3 (GPY, 2005)

Suppose the Elliott-Halberstam conjecture *EH(θ)* holds for some value of $\theta > 1/2$. That is, suppose the primes are nicely distributed modulo d for most values of d up to X^θ. Then we have bounded gaps between primes.

One should also say that GPY unconditionally proved some results about gaps between primes far superior to any that had appeared before their work, showing for example that there are always pairs of primes of size around X and separated by about $\sqrt{\log X}$, far less than the average spacing of log X for primes of this size.

About 20 years before that, in the 1980s, deep work of Bombieri, Fouvry, Friedlander and Iwaniec (in various combinations) had shown that a certain weak variant of the Elliott-Halberstam conjecture is true for some $\theta > 1/2$. In fact they obtained $\theta = 4/7$ in one of their results. Unfortunately, their result came with some technical restrictions which meant that it could apparently not be combined with the GPY method to prove bounded gaps between primes[4] . (As an aside, one of the main ingredients in these works of BFFI were certain complicated estimates for sums of Kloosterman sums due to Deshouillers and Iwaniec, and coming from the analytic theory of automorphic forms; some people apparently refer to this era as Kloostermania.)

What Zhang succeeded in doing is modifying, in a quite nontrivial way, both the GPY method and the BFFI ideas so that they meet in the middle. As it turns out, an equivalent modification of the GPY method had already been published by Motohashi and Pintz. They observed that in the Elliott Halberstam conjecture *EH(θ)* one does not need the primes to be nicely distributed modulo d for most d up to X^θ. Rather one only needs[5] this to be so for most smooth values of d, that is to say values of d with no prime factors bigger than X^δ for some very small δ. In particular, one does not need to know anything at all about the case when d is prime or almost prime. Thus Motohashi and Pintz, and independently Zhang, established a result of the following form.

Theorem 4 (Motohashi-Pintz, Zhang)

Suppose that the primes are satisfactorily distributed modulo d for most smooth values of d up to X^θ, for some value of $\theta > 1/2$. Then we have bounded gaps between primes.

We could, if we wanted, call the fact that the primes are satisfactorily distributed modulo d for most smooth values of d up to X^θ the weak Elliott-Halberstam conjecture, and denote it $EH_{weak}(\theta)$. To reiterate, then, Motohashi-Pintz-Zhang prove that if we have $EH_{weak}(\theta)$ for any $\theta > 1/2$ then we still get bounded gaps between primes.

Incidentally, now might be a good time to remark that the bound for the gap H depends on how close θ is to 1/2, the relation being very roughly of the form $H \sim (\theta - 1/2)^{-3/2}$ with a suitably optimised version of the argument, based on work of Conrey, Farkas, Pintz and Rev´esz.

The heart of Zhang's advance is the following result.

[4] There is a technical discussion of exactly why not in the book *Opera Cribro* by Friedlander and Iwaniec, pages 408–409.

[5] In fact one can get away with a still weaker property, in which one need only understand the number of primes congruent to *c mod d* for *c* varying in a small set of residue classes modulo *d* which varies in a multiplicative fashion with *d*.

Theorem 5 (Zhang)

We have the weak Elliott-Halberstam conjecture $EH_{weak}(\theta)$ for $\theta = 1/2 + 1/1168$. That is, the primes are satisfactorily distributed modulo d for most smooth values of d up to X^θ. Hence, we have bounded gaps between primes.

We now turn to a few more details. First we say something about the GPY method (we shan't say anything here about its modification due to Motohashi-Pintz and Zhang). This uses ideas related to the *Selberg sieve*. Then, we shall say a few words about the very technically demanding proof of Theorem 5. The key words here are bilinear forms, Kloosterman sums and deep bounds of Bombieri and Birch coming from Deligne's proof of the Riemann hypothesis over finite fields. There is no use of *automorphic form* bounds in Zhang's argument, and this is where he deviates somewhat from many of the papers of BFFI (though there are closely related ideas in a paper of Friedlander and Iwaniec on the divisor function).

The GPY method

What has the distribution of primes modulo d got to do with finding small gaps between primes? Exposing this hitherto unseen connection was the remarkable advance of Goldston, Pintz and Yıldırım.

GPY in fact prove a result that is strictly stronger than bounded gaps between primes. We say that a k-tuple of integers $h_1, \cdots, h_k$ is *admissible* if there is no obvious "congruence" or "local" reason why $n + h_1, \cdots, n + h_k$ cannot all be prime for infinitely many n. For example, $\{h_1,h_2,h_3\} = \{0,2,4\}$ is not admissible, because at least one of these numbers is divisible by 3, whereas $\{h_1,h_2,h_3\} = \{0,2,6\}$ is admissible (though no-one has the slightest idea how to *prove* that $n,n+2,n+6$ are all prime for infinitely many n). A moment's thought convinces one that the natural criterion for admissibility is that, for each prime p, the set $\{h_1, \cdots, h_k\}$ omits at least one residue class modulo p. If n_* is this class then $n + h_1, \cdots, n + h_k$ could perhaps all be prime when $n \equiv -n_* \pmod{p}$, because none of these numbers is divisible by p. Here is the stronger statement that GPY proved.

Theorem 6

Suppose we have the Elliott-Halberstam conjecture $EH(\theta)$ for some $\theta > 1/2$, that is to say the primes are nicely distributed modulo d for most values of d up to X^θ. Suppose that $k \geq k_0(\theta)$ is sufficiently large. Then for any admissible k-tuple $h_1, \cdots, h_k$, there are infinitely many n for which at least two of $n + h_1, \cdots, n + h_k$ are prime.

The modification of Motohashi-Pintz-Zhang is to show that this is still true if we instead assume just the weak Elliott-Halberstam conjecture $EH_{weak}(\theta)$, but we shall not be saying anyting further on the subject of this modification here. To see why Theorem 6 implies bounded gaps between primes, one need only note that there are admissible k-tuples for every k. Indeed the set of all prime numbers between M and $2M$ will be admissible for all sufficiently large M, and the number of primes in this range grows without bound by the prime number theorem or in fact by weaker statements. It should be pointed out that finding tightly packed admissible k tuples, which is necessary to elucidate the relationship between k and the prime gap H, brings one into contact with some thorny unsolved problems. To a large extent one must rely on computations to optimise this dependence for any particular value of k.

We'll sketch the very broad outline of the proof of Theorem 6. It uses sieve theory, the branch of analytic number theory that has ultimately grown out of a serious study of the Sieve of Eratosthenes. Something that has been learned, rather painfully, over the last 100 years is that

Almost primes are much easier to deal with than primes.

An r-almost prime is a product of at most r primes. Fix an admissible tuple $\{h_1, \cdots, h_k\}$. The rough idea of GPY is to choose an appropriate value of $r > k$ and try to compute the expected number of $n + h_1, \cdots, n + h_k$ that are prime, when n is selected at random from those n for which the product $(n + h_1) \cdots (n + h_k)$ is an r-almost prime. If we can show that this is > 1 then, for some value of n, there must be two primes amongst the $n+h_i$.

To be a little more formal about this, write $\nu(n) = 1$ if $(n + h_1) ... (n + h_k)$ is an almost prime and 0 otherwise. Let X be an arbitrary large quantity.

Then what we are interested in is the ratio

$$\frac{\sum_{X\leq n<2X}(\Lambda(n+h_1))+...(\Lambda(n+h_k))\nu(n)}{\log X\sum_{X\leq n<2X}\nu(n)} \quad (1)$$

(Recall that Λ is basically the characteristic function of the primes weighted by log.) If we can show that this is > 1, we will then know that for some $n\epsilon[X,2X)$ at least two of $n+h_1,\cdots,n+h_k$ are prime, and this will conclude the proof of Theorem 6. To elaborate this idea, one must be able to estimate the numerator and denominator of (1). Now we come to a completely crucial idea, invented by Selberg. The idea is that there are weight functions *ν(n)* which behave morally rather like the characteristic function of the almost primes (or the set of *n* for which $(n+h_1)\cdots(n+h_k)$ is almost-prime), but which are much easier to compute with. Let *D*, $1 < D < X$, be a parameter and consider the function

$$\nu(n)=(\sum_{d|n,d\leq D}\lambda_d)^2,$$

where at the moment $(\lambda_d)_{1\leq d\leq D}$ is any set of real numbers with $\lambda_1 = 1$. The weight *ν(n)* is always nonnegative, and furthermore if *n* is prime and between *D* and *X* then *ν(n)* = 1. The reason is that in this case, the only divisors *d* of *n* are 1 and *n*, and of these only *d* = 1 satisfies $d\leq D$. For this reason we'll call *ν* a *majorant for the primes.*

Let's see how we can use a majorant like this to study a classical problem called the Brun-Titchmarsh problem: that of estimating from above the number of primes in a range $[X_0, X_0+X)$. Now provided that[6] $D=o(X^{1/2})$, we can compute an asymptotic for the average value of *ν(n)* over $X_0\leq n<X_0+X$ and hence get an upper bound for the number of primes in this range. To see why the condition $D=o(X^{1/2})$ is critical, we need to do an actual calculation (though a very short one):

$$\begin{aligned}\sum_{X_0\leq n<X_0+X}\nu(n) &= \sum_{X_0\leq n<X_0+X}(\sum_{d|n,d\leq D}\lambda_d)^2\\ &= \sum_{d,d'\leq D}\lambda_d\lambda'_d\sum_{X_0\leq n<X_0+X;d,d'|n}1 \quad (2)\end{aligned}$$

Now the inner sum counts how many n there are in the range $X_0\leq n<X_0+X$ for which both *d* and d_0 divide *n*, or equivalently for which the lowest common multiple $[d,d_0]$ divides *n*. Note that $[d,d_0]\leq D^2$. Hence if $D=o(X^{1/2})$ then $[d,d_0]$ = *o*(*X*) and the answer is essentially $X/[d,d_0]$ (Imagine you were asked how many multiples of 2014 there are in the interval $[10^{10},10^{10}+10^5]$: since $10^5/2014$ = 49.65... it's either 49 or 50, but in either case 49.65 is a good approximation.) Therefore

$$\sum_{X_0\leq n<X_0+X}\nu(n)\approx\sum_{d,d'\leq D}\lambda_d\lambda'_d\frac{X}{[d,d']} \quad (3)$$

If, however, $D > X^{1/2}$, then we could have (indeed we will often have) $[d,d_0] > X$, so the answer might be 0, or it might be 1. (How many multiples of 2014 are there in the interval $[10^{10},10^{10}+10^2]$?) We cannot say which without carefully inspecting X_0, and getting a usable expression is impossible without further ideas. Let us say that $X^{1/2}$ is the sieving limit for this problem, and we call *D* the sieving level. Note that the larger we can take *D*, the more flexibility we have in choosing the weights λ_d. This ought to lead to the majorant *ν* being a better approximant to the characteristic function of the primes themselves. Miraculously, even though we are forced to take $D=o(X^{1/2})$ there are choices of the weights λ_d for which *ν* is a reasonably good approximation to the characteristic function of the primes. We can find such λ_d by minimising the quadratic form in (3) subject to $\lambda_1 = 1$, a routine exercise albeit one requiring some facility with Mobius inversion. When this is done, we find that in fact, for this choice of weights,

$$\sum_{X_0<X\leq X_0+X}\nu(n)\approx\frac{2X}{\log X}$$

When $X_0 = 0$ the majorant *ν* only overestimates the number of primes by a factor of 2. If one looks very carefully at *ν(n)* then one sees that it resembles a sort of characteristic function of almost primes, though it is best not to pursue this line of thought too far, leaving it perhaps as motivation for calculating somewhat blindly with *ν*. Note in particular that *ν* will not in general be {0,1}-valued. Note, by the way, that we have proven that the number of primes in $[X_0, X_0+X]$ is at most about 2*X*/log*X* for any X_0, but that is another story.

It turns out that even if *D* is a very small power of *X* we still get a majorant ν that overestimates the primes by a constant factor, although this constant gets worse as *D* becomes smaller.

[6] This should be read as "a bit smaller than $X^{1/2}$". In fact, one would require a condition like $D < X^{1/2-\varepsilon}$ for some $\varepsilon > 0$.

Returning to our main discussion, recall (1). In the light of the crash course on the Selberg sieve we have just given, it is natural to consider defining

$$\nu(n) = (\sum_{\substack{d|(n+h_1)\dots(n+h_k)\\ d\leq D}} \lambda_d)^2 \qquad (4)$$

for appropriate weights λ_d, where D is as big as possible. To recap, one should think of ν as telling us the extent to which $(n+h_1)\dots(n+h_k)$ is almost prime, though to attach any more precise meaning to such a statement one must be more precise about the nature of the λ_d, about which I shall not say any more.

By a small modification of the above reasoning, one can compute the denominator in (1) provided that the sieving level D is $o(X^{1/2})$. What, however, of the numerator? It may be split into terms of the form

$$\sum_{n\leq X} \Lambda(n+h)\nu(n)$$

for $h = h_1,\dots,h_k$. Trying to repeat the computation in (2) above, we instead arrive at the expression

$$\sum_{d_1,d_2\leq D} \lambda_{d_1}\lambda_{d_2} \sum_{\substack{n\leq X\\ d_1,d_2|(n+h_1)\dots(n+h_k)}} \Lambda(n) \qquad (5)$$

To understand this sum, we need to know how the primes (weighted using the von Mangoldt function Λ) behave modulo $d = [d_1,d_2]$, and this quantity may be as large as D^2. If we know the Elliott-Halberstam conjecture *EH*(θ) (primes are nicely distributed modulo d for most d up to X^θ) then we will be fine so long as $D = o(X^{\theta/2})$. Unconditionally, that is to say using just the Bombieri-Vinogradov theorem, we may only take the sieving level D to be about $X^{1/4}$, which means $\nu(n)$ gives a weaker notion of $(n+h_1)\dots(n+h_k)$ being almost prime.

At this point one must dirty the hands by doing an actual computation of the numerator and denominator of (1) with a judicious choice of the weights λ_d. Making a sensible (by which we mean more-or-less optimal) choice, and taking D to be almost $X^{\theta/2}$, one eventually computes that the ratio in (1) is

$$\frac{2\theta}{(1+k^{-1/2})^2}$$

Remember that k is the number of elements in our admissible tuple$\{h_1,\cdots,h_k\}$. I should say that I don't think I could motivate the result of this computation particularly well, if at all, even to an expert audience. I'm not such there even is a conceptual explanation of it – you just have to do it.

Recall that we wanted the ratio to be greater than 1: this would give us bounded gaps between primes. Even if k, the number of elements in our admissible tuple, is very large one does not achieve this if $\theta \leq 1/2$. This is pretty unfortunate, since we can only proceed unconditionally when $\theta \leq 1/2$. As soon as θ is even a tiny bit larger than 1/2, however, the ratio will indeed be larger than 1 provided that k is big enough, and we will get bounded gaps between primes as discussed above. The value $\theta = 1/2$ is thus a crucial barrier for the GPY method: with $\theta < 1/2$ one gets very little, whilst with $\theta > 1/2$ one obtains bounded gaps between primes.

This concludes our cursory discussion of the GPY method, which links bounded gaps between primes to the distribution of primes in progressions. We turn now to the other side of the story, in which the aim is to understand as much about the latter as possible.

Primes in arithmetic progression

We turn now to a description of Zhang's major advance, the proof of Theorem 5. Let us recall the statement.

Theorem 7 (Zhang)

We have the weak Elliott-Halberstam conjecture $EH_{weak}(\theta)$ for $\theta = 1/2 + 1/1168$. That is, the primes are satisfactorily distributed modulo d for most smooth values of d up to X^θ.

We did not say exactly what satisfactorily distributed means, but it basically means that we are interested in showing that[7] if $(c,d) = 1$ then

$$\sum_{\substack{x\leq X\\ x\equiv c \pmod d}} \Lambda(x) \approx \frac{X}{\phi(d)}$$

[7] In fact, this only needs to be shown when c belongs to a specific and fairly small set of residue classes varying in a multiplicative fashion with d, but will not mention this point again.

for most smooth $d < X^{\theta}$. Remember that by smooth we meant that all prime factors of d are at most X^{δ} for some very small δ. The way in which this is used is that it allows us to suppose that d is wellfactorable, which means that for any Q,R with $QR \sim d$ we can find a factorisation $d = qr$ with $q \approx Q$ and $r \approx R$. Exactly how this property is used is not very easy to explain properly, but bear with us. One can, however, immediately observe that if q,r are coprime then the condition $x \equiv c \pmod{d}$ is equivalent to conditions on x modulo q and modulo r, by the Chinese remainder theorem, so there is a sense in which we are reducing the size of the moduli and thereby making the problem simpler.

At this point in the exposition it is slightly convenient to work with averages rather than sums, which we notate using the probabilistic E notation, though there is nothing random in our discussion. Our task is more-or-less equivalent to estimating an average over primes of the form

$$\mathbb{E}_{x \leq X} \Lambda(x) \psi(x) = \frac{1}{X} \sum_{x \leq X} \Lambda(x) \psi(x)$$

where in this case

$$\psi(x) = 1_{x \equiv c \pmod{d}} - 1_{(x,d)=1} / \phi(d) \quad (6)$$

and the goal is to show that this average is appreciably smaller than the "trivial" bound of about 1 which comes from the prime number theorem.

To attack averages like this, we introduce a notion that is central to additive prime number theory: that of expanding in terms of bilinear forms. Suppose that instead of the above average we were instead asked to estimate

$$\mathbb{E}_{x \leq X} (\alpha * \beta)(x) \psi(x) \qquad (7)$$

where α, β are arithmetic functions with $|\alpha(m)|, |\beta(n)| \leq 1$, and with $\alpha(m)$ supported on the range $m \sim M$ and $\beta(n)$ on the range $n \sim N$, where $MN = X$. Here, $*$ denotes Dirichlet convolution, that is to say

$$(\alpha * \beta)(x) = \sum_{mn = x} \alpha(m) \beta(n)$$

The function ψ is completely arbitrary for the purposes of this discussion, except we assume that $|\psi(x)| \leq 1$ for all x. We even let ψ be complex-valued. The sum (7) can more-or-less be rewritten as

$$\mathbb{E}_{m \sim M} \mathbb{E}_{n \sim N} \alpha(m) \beta(n) \psi(mn)$$

Applying the Cauchy-Schwarz inequality, the square of this is bounded by

$$\begin{aligned} &\mathbb{E}_{m \sim M} \left| \mathbb{E}_{n \sim N} \beta(n) \psi(mn) \right|^2 \\ &= \mathbb{E}_{n,n' \sim N} \beta(n) \beta(n') \mathbb{E}_{m \sim M} \psi(mn) \overline{\psi(mn')} \end{aligned}$$

Applying Cauchy-Schwarz a second time, the square of this is bounded by

$$\mathbb{E}_{n,n' \sim N} \mathbb{E}_{m,m' \sim M} \psi(mn) \overline{\psi(mn') \psi(m'n)} \psi(m'n') \quad (8)$$

The key thing to note here is that the unspecified functions α, β have completely disappeared, and we are left staring at an expression involving only ψ. Perhaps we might hope to estimate it, the aim being to improve substantially on the trivial bound of 1. If we can do this, we have a kind of "certificate" which asserts that ψ always gives nontrivial cancellation in averages such as (7), no matter what α and β are.

There are two rather obvious barriers to this observation being at all useful, and they are the following.

(i) We in fact wish to estimate the average

$$\mathbb{E}_{x \leq X} \Lambda(x) \psi(x)$$

but we have said nothing about the extent to which Λ can be expressed in terms of Dirichlet convolutions $\alpha * \beta$, nor even offered any motivation for why this should be expected.

(ii) We are interested in a specific function ψ, given by (6). Why should this ψ allow us to provide a certificate by exhibiting nontrivial cancellation in (8)?

With regard to (i), many readers will know that $\Lambda = \mu * \log$, where μ is the Mobius function. However, this turns out not to help greatly in the above scheme. The reason is that in practice we will only be able to estimate expressions such as (8) for quite restricted ranges of M and N, usually with M and N close in size, and with the decomposition $\mu * \log$ there is no opportunity to seriously restrict these ranges.

A successful technique depends on a remarkable identity of Linnik, or rather on a kind of truncated variant of it due to Heath-Brown, which we shall not state. The identity states that

$$\frac{\Lambda(n)}{\log n} = \sum_k \frac{(-1)^k}{k} \tau'_k(n)$$

where $\tau'_k(n)$ is the number of ways to factor $n = n_1 \ldots n_k$ with $n_i > 1$ for all i. If one knows the definition and very basic properties of the ζ function, the proof is just a couple of lines long: observe that

$$\log \zeta(s) = \log(1 + (\zeta(s) - 1)) = \sum_{k=1}^{\infty} (\zeta(s) - 1)^k$$

and compare coefficients of n^{-s} on both sides. The right hand side is pretty obviously the right-hand side of Linnik's identity. As for the left-hand side, its derivative is $\zeta'(s)/\zeta(s)$, which is well-known to be $-\sum_n \Lambda(n) n^{-s}$. Indeed, it is precisely this relation that links primes and the ζ-function. Integrating with respect to s, we obtain Linnik's identity.

Note that $\tau'_k(n)$ is a Dirichlet convolution of k copies of the function which equals 0 at 1 and is 1 everywhere else. Chopping the domain of this function into various ranges, we can indeed write Λ as a sum of a number (not too large) of convolutions $\alpha * \beta$, with $\alpha(m)$ supported where $m \sim M$ and $\beta(n)$ where $n \sim N$, with considerable flexibility in arranging the ranges M and N we need to worry about. The precise arrangement of these ranges is a rather technical matter, and suffice it to say that Zhang classifies them into three different types (plus a somewhat trivial type): these are called Type I, II and III. Actually, the Type III sums in fact involve certain 4-fold convolutions $\alpha_1 * \alpha_2 * \alpha_3 * \alpha_4$.

What about (ii), that is to say the issue of obtaining a "certificate" for ψ which certifies that averages such as $\mathbb{E}_{x \leq X}(\alpha * \beta)(x)\psi(x)$ exhibit cancellation? One may note that (8) is certainly not *always* $o(1)$. Rather trivially, when ψ is the constant function 1 we get no cancellation. The same is true if $|\psi(x)| = 1$ and if ψ is multiplicative in the sense that $\phi(mn) = \phi(m)\phi(n)$, as can be easily checked. (One consequence of this observation is that the whole scheme we have just outlined is somewhat unhelpful for showing that Λ does not correlate with a single Dirichlet character χ.) However, our function

$$\psi(x) = 1_{x \equiv c \pmod d} - \frac{1_{(x,d)=1}}{\phi(d)}$$

does not obviously exhibit multiplicative behaviour and therefore one can hope to produce a certificate for this ψ, at least on average over d.

The reader will not be surprised to hear that the above discussion was an oversimplification, although it captures something of the key ideas. There are other types of "certificate" than (8). The key tools that are brought to bear on estimating averages such as $\mathbb{E}_{x \leq X}(\alpha * \beta)(x)\psi(x)$ with our particular choice of ψ are:

(i) Cauchy-Schwarz, similar to the above;

(ii) Fourier expansion, for example of ψ;

(iii) Shifting the range of summation, that is to say replacing the $\mathbb{E}_{n \sim N} F(n)$ by $\mathbb{E}_{n \sim N} \mathbb{E}_{|k| \leq K} F(n+k)$, which should be roughly equal to it if $K \ll N$;

(iv) Certain changes of variable and substitutions;

(v) Completion of sums, that is to say replacing an incomplete sum $\sum_{x \in I} f(x)$ by sums $\sum_{x \in \mathbb{Z}/d\mathbb{Z}} f(x) e_d(hx)$ where $I \subset \mathbb{Z}/d\mathbb{Z}$ is an interval, and $e_d(x) := e^{2\pi i x/d}$.

Considerable extra flexibility is available under the "well-factorable" assumption that $d = qr$; instead of averaging over $d \leq X^\theta$, one now averages over both q and r, and this affords still more opportunity to vary the application of the above four ingredients. The application of (i) to (v) above (in various combinations) throws up other sums that need to be estimated. The most interesting such case for Zhang occurs in his treatment of the so-called Type III sums, where expressions such as the sum

$$\sum_{n,n',l \pmod d} e_d\left(\frac{c_1}{ln} + \frac{c_2}{(l+k)n'} + h_1 n + h_2 n'\right) \quad (9)$$

are relevant. Here, h_1, h_2, k are parameters, and the sum over n, n', l is restricted somewhat, in particular to $n, n', l, l+k \neq 0$ so that it makes sense.

It is very hard to explain how an expression like this comes up without going through the applications of each of the techniques (i) to (v) (which all occur here) in turn. Suffice it to say that the k comes from shifting as in (iii), the h_1, h_2 come from (v) (followed by an application of Cauchy-Schwarz) and the instances of c/n ultimately come from an initial Fourier expansion of $1_{x \equiv c \pmod d}$ where $x = mn$.

Now it turns out that (9) is in fact bounded by (essentially) $d^{3/2}$, apart from in certain degenerate cases This is about as good an estimate as one should hope for, since it represents essentially square-root cancellation, as the number of things being summed over is d^3. This is by no means a trivial fact, and depends on cohomological ideas of Deligne, and in particular the Riemann hypothesis over finite fields. When $d = p$ is prime, the sum (9) may perhaps be written more suggestively, to an algebraic geometer, as

$$\sum_{(x_1,x_2,x_3,x_4)\in V} e_p(x_1+x_2+x_3+x_4)$$

where $V \subset F_p^4$ is a variety $\alpha x_1 x_2 + \beta x_3 x_4 = \gamma x_1 x_2 x_3 x_4$ In other cases (the analysis of the Type I and II sums) more standard sums such as Kloosterman sums $\sum_{x_1 x_2 = c} e_p(x_1 + x_2)$ come up. This particular sum is bounded by $2\sqrt{p}$, a famous bound of Weil which does admit an elementary (in the technical sense) proof due to Stepanov. This is not the case with the Bombieri-Birch bound,which needs the whole of the algebro-geometric machinery.

The great majority of these ideas can be found in the work of combinations of Bombieri, Fouvry, Friedlander and Iwaniec. The sad thing, however, is that this excellent and optimal bound of $d^{3/2}$ just fails to cancel out some losses coming from other manœuvres, particularly the completion of sums manoeuvre (v) which is rather costly if the length of I is much less than *d*. Instead of a valid certificate for ψ, one simply recovers, essentially, the trivial bound for (8) – not, perhaps, the most spectacular use of the deep machinery.

The crucial new innovation of Zhang is to exploit the presence of the factorisation *d=qr* (which, remember, can be selected quite flexibly). One might imagine that, by the Chinese remainder theorem, one obtains a product of sums similar to (9) modulo *q* and modulo *r*, leading to a bound of $(qr)^{3/2}$ and no eventual gain. What Zhang miraculously finds, however, is that by making sure the shift in (iii) is by a multiple of r, the sum modulo r is not of Bombieri-Birch type (9), but degenerates to something like

$$\sum_{\substack{s_1,s_2,n \pmod{r} \\ s_1,s_2,n\neq 0}} e_r\left(\frac{s_1-s_2}{s_1 s_2 n}\right)$$

This exhibits better than square root cancellation, being of size essentially *r*, as one can easily check. (In fact, this is a slight simplification: Zhang actually obtains a Ramanujan sum in which the variable *n* is constrained to be coprime to *r*, but the better-than-square-root cancellation still holds). Thus instead of $(qr)^{3/2}$ Zhang gets instead $q^{3/2}r$, and the factor of $r^{1/2}$ thus saved is crucial in making the argument work and establishing Zhang's remarkable theorem.

Polymath 8a

Shortly after Zhang's paper came out, Terence Tao orchestrated a collaborative project to reduce the value of *H* as far as possible from Zhang's value of 70000000, and more generally to understand all aspects of Zhang's work. One of the many great things about this project, in my view, was that without it there could well have been thousands of papers improving *H* in various different ways. One big achievement of this project was to increase Zhang's 1/1168 to 7/300, that is to say to prove $EH_{weak}(1/2 + 3/700)$, and to refine various other aspects of the argument, including the sieve theory and the computation determination of admissible tuples, so as to eventually reduce *H* to 4680.

Another significant achievement of the project was to remove the dependence on the deepest algebro-geometric results, although if this was one's concern then the value of H could only be taken to be 14950. At this point, the entire argument could reasonably be presented from first principles in an advanced graduate course. (Personally, I found it extraordinary that initially one could only get bounded gaps between primes using the full force of Deligne's machinery, and no finite bound without it.)

Maynard–Tao

As Polymath 8a was nearing completion, James Maynard and Terry Tao simultaneously made a dramatic advance which vastly simplifies the whole argument. Students wishing to study bounded gaps between primes may now read the 23-page paper of Maynard, freely available at

http://arxiv.org/abs/1311.4600.

In the Maynard–Tao argument, one still needs information about the distribution of primes in progressions, but things now work with any positive value of $\theta > 0$: the GPY barrier of $\theta = 1/2$ turned out to be somewhat illusory. The crucial new idea of Maynard and Tao is to consider a different weight function v.

A second Polymath project, again led by Terence Tao, has been working on the problem in the light of the Maynard–Tao development. When the first version of this paper was submitted on 20/2/14 the record value of *H* was 264, but when I came to correct some typos a few days later this value had already been reduced to *H* = 246.

Moonshine and the Meaning of Life

Prof. Yang-Hui He
Dept. of Mathematics, City University of London & Merton College, Oxford University
Prof. John McKay
Department of Mathematics and Statistics, Concordia University, Montreal, Canada

The elliptic modular function, j, invariant under PSL(2, $\mathbb{Z}$), has Fourier expansion

$$j(q) = \frac{E_4(q)^3}{\Delta(q)} = \sum_{m=-1}^{\infty} c_m q^m$$
$$= \frac{1}{q} + 744 + 196884q + 21493760q^2 + \ldots \qquad (1)$$

as $z \rightarrow i\infty$, where $q = e^{2\pi iz}$ is the nome for z;

$$E_4(z) = 1 + 240 \sum_{n=1}^{\infty} \sigma_3(n) q^n$$

is the theta series for the E_8 lattice, $\sigma_3(n) = \sum_{d|n} d^3$ and

$$\Delta(q) = q \prod_{n=1}^{\infty} (1 - q^n)^{24} = \sum_{m=1}^{\infty} \tau_m q^m$$
$$= q - 24q^2 + 252q^3 - 1472q^4 + 4830q^5 + \ldots \qquad (2)$$

is the modular discriminant [S]. There are two new congruences.

OBSERVATIONS:

- [JM] $\left(\sum_{m=1}^{24} c_m^2\right) \mod 70 \equiv 42$
- [YHH] $\left(\sum_{m=1}^{24} \tau_m^2\right) \mod 70 \equiv 42$

The vector ω = (0, 1, 2, ... , 24 , 70) lives in the Lorentzian lattice $II_{25,1}$ in 26 dimensions as an isotropic Weyl vector [C], allowing us to construct the Leech lattice as $\omega^{\perp}/\omega$. Watson's [L, W] unique

non-trivial solution to $\sum_{i=1}^{n} i^2 = m^2$ is (n, m) = (24, 70).

Indeed, the second author's observation 35 years ago that

$$196884 = 196883 + 1 \qquad (3)$$

sparked the field of "Monstrous Moonshine" [B, CN], underlying so much mathematics and physics, relating, inter alia, modular functions, finite groups, lattices, conformal field theory, string theory and gravity (see [G] for a review of some of the vast subjects encompassed) in which the j -invariant and the Leech lattice are central. As we ponder the meaning of life, we should be aware of the prescient remarks of the author [A], Douglas Adams:

> "The Answer to the Great Question . . . is . . . Forty-two," said Deep Thought, with infinite majesty and calm.

References

[A] Douglas Adams, "The Hitchhiker's Guide to the Galaxy", London, 1979.

[B] R. E. Borcherds, "Vertex algebras, Kac-Moody algebras, & the monster," Proc. Nat. Acad. Sci. 83, 3068 (1986); "Monstrous moonshine & monstrous Lie superalgebras", Invent. Math. 109 (1992) 405 - 444.

[C] J. H. Conway, "The automorphism group of the 26-dim even unimodular Lorentzian lattice", J. Alg. 80, Vol. 1, (1983), 159 - 163.

[CN] J. H. Conway and S. P. Norton, \Monstrous Moonshine," Bull. LMS, 11, (1979), 308 - 339.

[G] T. Gannon, "Moonshine beyond the Monster: The Bridge Connecting Algebra, Modular Forms and Physics", CUP, 2006, ISBN 0-521-83531-3.

[L] E. Lucas, "Question 1180," Nouvelles Annales de Mathematiques., ser. 14 (1875), 336.

[S] J-P. Serre, "A Course in Arithmetic", GTM 7, Presses Universitaires de France, (1970), xVII.3.3, 4.5

[W] G. N. Watson, "The problem of the square pyramid," Messenger of Mathematics, 48, (1918-19), 1-22

Learning to denoise

Luca Calatroni
Cambridge Centre for Analysis, University of Cambridge

1: Describing the noise

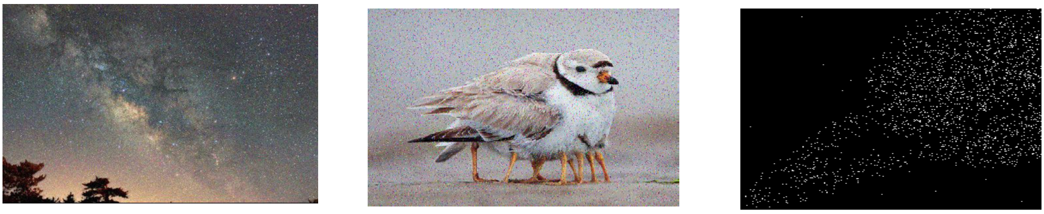
One of the key tasks in Image Processing consists in removing interferences in the image coming from acquisition problems due, for instance, to external factors, like temperature or illumination. We have all experienced similar issues when taking a picture of a very dark scene, for example in a clear and fresh summer night when looking at dazzling stars shining in the sky. Even with professional cameras, the result may look grainy, not clean nor clear. We say that the image is corrupted by what is called *noise*, the granular component we want to remove in order to get a clean version of the image. Depending on applications and on the physical properties characterising the transmission, acquisition and processing steps, different types of noise can be considered. In many cases, the noise is assumed to follow a Gaussian distribution, Fig. 1a. This assumption relies, mainly, on a probabilistic asymptotical property of noise distributions, called Central Limit Theorem. According to this result, the sum of independent random variables of any distribution converges, as the number of measurements goes to infinity, to a Gaussian-distributed random variable. Very often, though, this very reasonable assumption does not model realistically the actual physical source of noise corrupting the data. For instance, in the case when transmission problems `switch off’ just some of the pixels in the image, a different, not everywhere-spread noise distribution is preferred: the noise model that arises is called impulse noise, Fig. 1b. Finally, in astronomical Imaging applications, physical properties of the light reflecting its quantised nature have to be considered. This results in a photon-counting process which mathematically relates to a discrete Poisson probability distribution, Fig. 1c. Differently from the additive nature of the Gaussian-type noise, Poisson distribution is signal dependent, that is the intensity of the noise depends on the brightness of the regions in the image: brighter regions will present higher level of noise. In Fig. 1 we can observe some examples of the noise distributions described. Many more can be considered: normally they model signal-dependent noise distributions in the form of multiplicative noise (like speckle noise, Rician noise. . .) arising, for instance, in MRI applications. However, the mathematical modelling and analysis of these models is rather involved, so for what follows we will focus on the three noise models (Gaussian, impulse and Poisson) mentioned above.

2: Image denoising

The task of image denoising can be described mathematically as an inverse problem, that is given our noisy image f, we want to reconstruct the noise-free image u such that:

$$f = T(u) \tag{1}$$

In (1) the operator T models the degradation process u goes through: in our case this can be thought of as the operator that encodes the physical properties responsible of the noise present in the image. Other choices of T can be made: for instance, T can represent a blur operator or divide regions in the image where the information is available from others where the information has been lost (like, for example, for dis-occlusion problems).



(a) Gaussian noise.

(b) Impulse noise.

(c) Poisson noise.

Figure 1: Different noise distributions for different Imaging applications.

What is important to highlight here is that, in general, problem (1) is not easy to solve: uniqueness of the solution and/or its stability with respect to the initial data can fail. These problems are normally said to be *ill-posed*. In order to get a solution of (1) an alternative formulation has to be considered. A traditional approach consists in regularising the problem by adding some a-priori information describing the properties of the solution we are looking for. In our Imaging framework the question then is: what are the fundamental properties of an image? The answer is: the **edges**. Thanks to the edges in the image, objects can be identified from their background and details within objects are distinguishable: as such, every regularisation procedure used to reformulate (1) needs to encode such features.

2.1 Designing a tailored method

In order to derive a regularisation model suitable for our Imaging tasks, a careful mathematical modelling is needed. Images have to be interpreted as functions defined on an image domain (the bi-dimensional grid of pixels) and associated to a number or a set of numbers corresponding to either the greyscale or the RGB (red, green, blue) intensity values, respectively. The modelling of images in function spaces is rather a delicate point in the design of an optimal reconstruction method: the choice of the function space itself reflects in the expected regularity of the image we want to reconstruct. Function spaces which allow too much regularity have to be discarded as they will destroy the main structures in the image because of their strong smoothing properties, see Fig. 2. We refer the reader interested in knowing more about Image modelling and the choice of suitable function spaces for Imaging tasks to [1, 7].

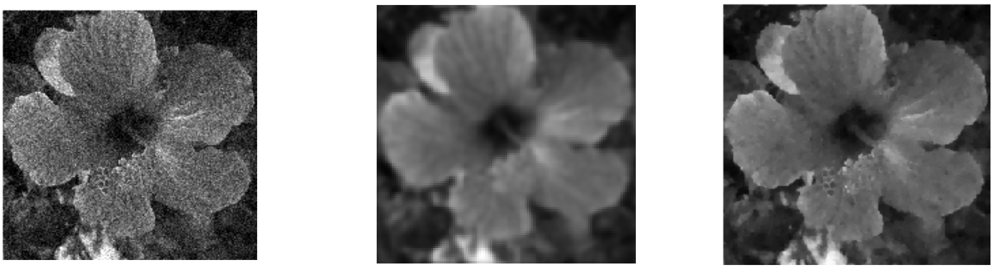
Over the last twenty years non-smooth regularisers have been studied. Namely, Imaging communities have focused their attention on *Total-Variation* (TV) regularisation metods. In words, TV can be thought of as a regularising term that while smoothing out the unwanted noise from the image, preserves its fundamental and geometrical structures, such as edges, compare Fig. (2c). In more mathematical words this corresponds to consider a weaker norm of the image gradient which does not enforce too much regularity, but identifies contours and edges in the image.

A general regularisation method of inverse problems like (1) has the form:

find v such that v minimises

$$J(v) := R(v) + \lambda\phi(f, v) \qquad (2)$$

The approach (2) is an example of energy-minimisation approach: we model our denoising problem by assuming that a scalar quantity J, the energy, is associated to our problem. Our solution u will be the function in correspondence of which the energy J achieves its minimum.



(a) Noisy Image.

(b) *Smooth* regularisation [10].

(c) *Non-smooth* regularisation [9].

Figure 2: The choice of the regularisation term and the function space affects the reconstruction: smooth regularisations remove noise though destroying fundamental structure; non-smooth regularisations (such as Total Variation) removes the noise, while preserving edges.

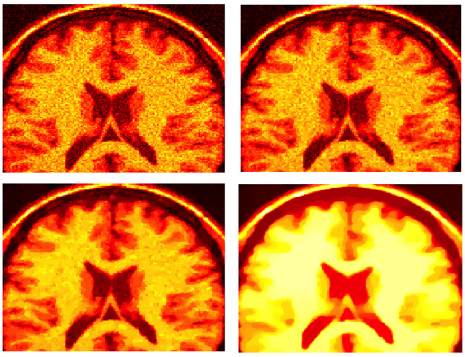


Figure 3: T.L.: Acquired noisy image. T.R.: Denoised image with large λ. The regularisation is poor and the noise is still present in the reconstruction. B.L.: Optimal denoised image. B.R.: Overregularised denoised image.Fundamental structures have been destroyed.

We model *J* as the sum of two different terms: the regularisation term *R(u)* which shall encode the properties listed above and the fidelity term $\phi(u, f)$which models in a mathematical way the statistics of the noise, i.e. its probability distribution. The parameter λ balances the effect of the regularisation against the fidelity term and, heuristically, can be thought of as a quantity that measures how much we trust the acquired data. In the following we want to focus on its optimal choice which is a fundamental issue in applications where normally the right value of this parameter is found by a trial-error method. For medical applications, the choice of such parameter is crucial. Figure 3 shows some different TV reconstructions of the noisy image of a brain.

For medical purposes we would not like a poor reconstruction of our measured noisy image, which would not remove noise that could possibly occlude some regions of interest. On the other hand, we would not like regularise our image too much and lose meaningful anatomical structures. Somewhere between these two extremes we look for the optimal balance between trust in the data and noise regularisation: as mathematicians we want to find it in a sensible, realistic and automatised way.

3: The training idea

Let us stick for a moment to the medical imaging framework. In such applications, like MRI for instance, the accuracy of the measurements can be tuned. In general, we can think the accuracy to be proportional to the image acquisition time. High-acquisition times will result in clean, almost noise-free images which will approximate well the ground-truth, while low-acquisition times will correspond to blurry and noisy images.

In clinical practice, ideally high acquisition times are not feasible due to the computational costs and the reluctance of patients to stay for too long in MRI scanners for examinations. On the other hand, standard clinical acquisition times do produce some noise in the image, so a pre-processing step is fundamental for any subsequent diagnostic image-based evaluation and therapy planning. For a realistic and tailor-made denoising model we will exploit in the following two main ideas: a learning approach to make the estimation of the optimal balance robust and a correct modelling of the different noise distributions in order to cover the different noise models presented in Section 1.

3.1 Learning the noise via training sets

The idea of using database of images is quite realistic in applications. In medical Imaging, specially designed objects called *phantoms* are used to analyse and tune scanning devices. These objects may resemble anatomical structures of the human body and because of their design they provide consistent and reliable results. Based on this, we describe in the following our learning idea:

1. We use fixed devices producing, at each scan, the same type of noise in the image, whose level (i.e. intensity) is unknown;

2. A previously-acquired (or simulated) database of images, typically of phantoms, is available. We assume we have two different versions of each image u_k *(k*=1,...,*N)* in the database: a clean, almost noise-free version u_k acquired with high acquisition times and a second version of it f_k, acquired within standard medical acquisition times and consequently corrupted by noise.

We aim to find the optimal λ for the problem (2) such that for every k, the TV reconstruction of f_k matches at best the corresponding noise-free u_k version of the same image [8]. Because of our assumption 1. above, we shall now use the computed λ as an optimal parameter in order to process a new, not-simulated image, such as a real MRI scan of the brain of a patient.

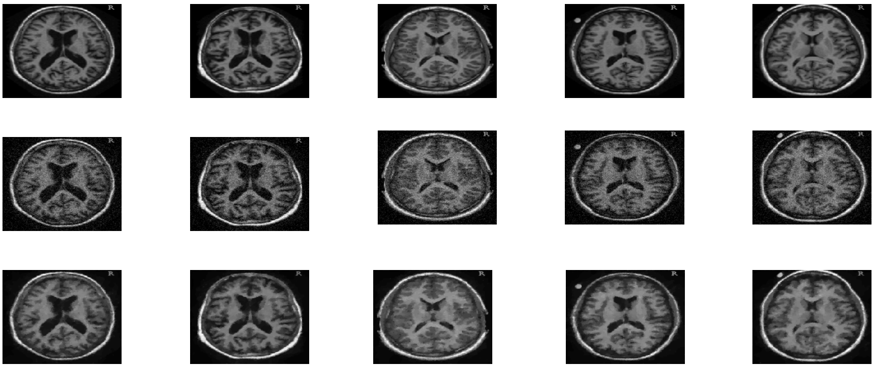


Figure 4: Sample of 5 images of OASIS MRI brain database: original images (upper row), noisy images (middle row) and optimal denoised images (bottom row), $\lambda_{opt} = 3280.5$.

In Figure 4 we show an example of a simulated database of MRI scans of brains: the first two rows represent the two noise-free and noisy versions of the images in the database, respectively, whereas the third row contains the denoised images obtained with the optimal parameter λ computed as described above. Further work has to be done in order to improve upon the characteristic `watercolor’ effect in the computed reconstruction: this, in fact, relates to the properties of the TV regularisation term used. More general regularising terms R(u) can improve upon this property, like, for instance, the Total Generalised Variation, [4].

For this example the noise has been assumed to be Gaussian-distributed, but in the following we will comment briefly on how to incorporate different noise models, as discussed in Section 1. From a computational point of view, solving non-smooth problems of the form (2) is very demanding, especially as the number of images in the database becomes very large, which is desirable in order to make the noise estimation robust. Due to our modelling assumptions 1. and 2., though, some sampling strategies can be used to improve upon efficiency, [6].

3.2 Optimal modelling

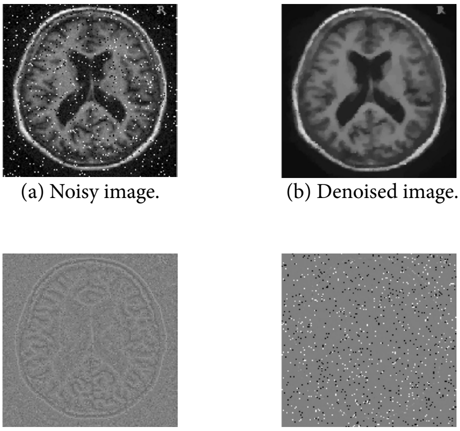
The regularisation approach (2) can accommodate easily different noise distributions like the ones corresponding to impulse or Poisson noise. Depending on the application considered, Heuristically, quadratic-type fidelity terms are generally considered for Gaussian noise distributions, whereas the modulus of the difference between u and f is preferred for impulse noise distributions. Finally, more sophisticated, logarithmic-type, fidelity terms are used for Poisson noise distributions [3]. When just one single noise distribution is assumed to corrupt the measurements, the approach described above can still be employed, by simply using the suitable fidelity term for describing the problem. But one can ask the question: what if multiple noise distributions are present in the image? As explained, each of them can be the result of different acquisition/transmission problems, so the combined presence of noise is perfectly reasonable in applications.

An immediate extension of (2) for the mixed-noise case would be considering a model that, for the very easy case of two noise distributions, reads as:

find v such that v minimises

$$J(v) := R(v) + \lambda_1 \phi_1(f, v) + \lambda_2 \phi_2(f, v) \tag{3}$$

that is a model that describes the joint presence of two noise distributions through the sum of the corresponding fidelity terms. The same strategy described in Section 3.1 would then, heuristically, determine the optimal size of the parameters λ_1 and λ_2 which on one side will resemble the data fitting with respect to the degradation due to each of the two noises present in the image and,



(a) Noisy image.

(b) Denoised image.

(c) Gaussian component.

(d) Impulse component.

Figure 5: Noise decomposition through infimal-convolution: the mixed-noise distribution is assumed to be a combination of Gaussian and impulse noise.

on the other side, would balance against the size of the regularisation, as before.

More complicated operations can be considered in order to solve this task. For instance, applying the discrete analogue of the convolution operator called infimal-convolution [2, Chapter 12] to ϕ_1 and ϕ_2 one can get from the model the additional property of noise decomposition into its components, compare Fig. 5.

4: Conclusions

The recipe for image denoising requires different ingredients. First of all, a careful understanding of the physical and statistical properties characterising the problem is needed in order to formulate an appropriate mathematical model. This reflects in the correct choice of the data fidelity term by which we can mimic the noise distribution corrupting the data. A fundamental aspect then is also the optimal choice of parameters which will result in an optimal image reconstruction. Training the model using database of images seems to be a promising and reliable strategy to design reliable and efficient image denoising methods, [5].

References

[1] Aubert, G., Kornprobst, P.: Mathematical problems in Image Processing: Partial Differential Equations and the Calculus of Variations, vol. 147, Springer Science and Business Media, Springer, 2006.

[2] Bauschke, H., Combettes, P.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, CMS Books in Mathematics, Springer, 2011.

[3] Benning, M., Burger, M.: Error Estimates for General Fidelities, Electronic Transactions on Numerical Analysis, vol. 38, pp. 44-68, 2011.

[4] Bredies, K., Kunisch, K., Pock, T.: Total generalised variation, SIAM Journal on Imaging Sciences, vol. 3(3), pp. 492-526, 2010.

[5] Calatroni, L. Chung, C., De Los Reyes J. C., Schönlieb C.-B., Valkonen T.: Bilevel approaches for learning of variational imaging models, survey paper, to appear in Radon book series, vol. 18, 2015.

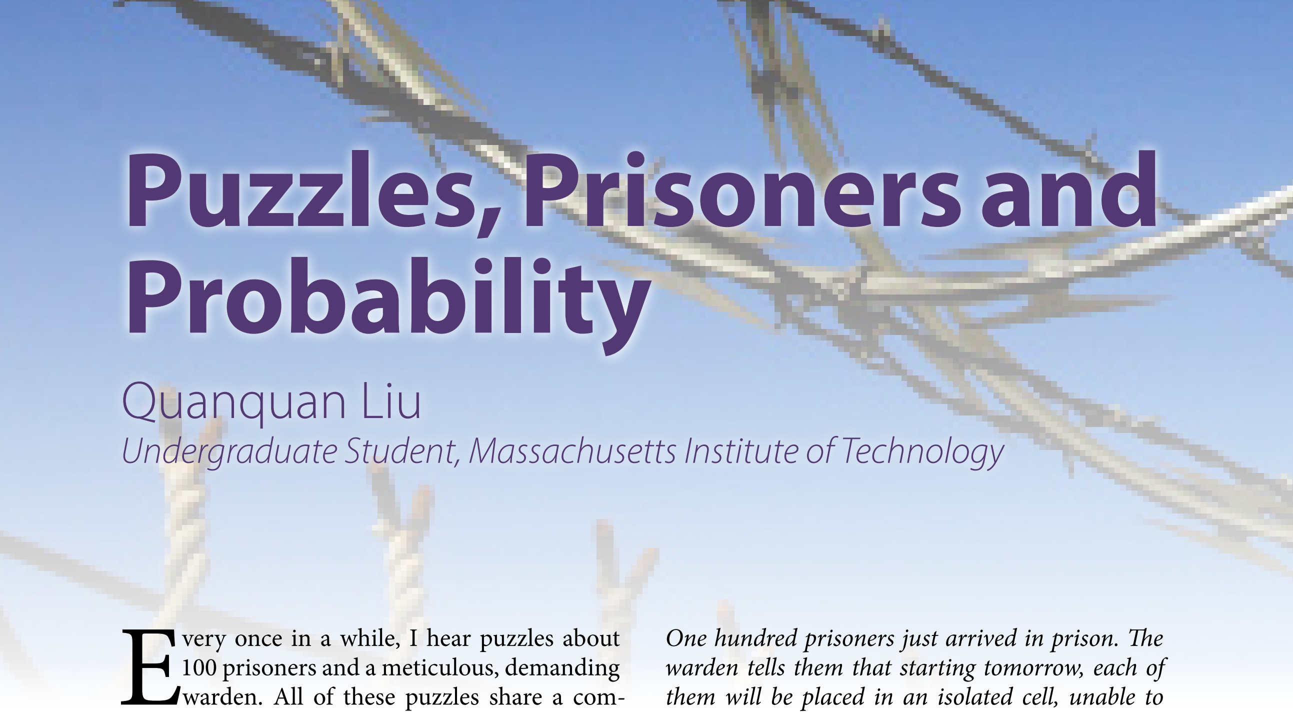
[6] Calatroni, L. De Los Reyes J. C., Schönlieb C.-B.: Dynamic sampling schemes for optimal noise learning under multiple nonsmooth constraints, IFIP TC7-2013 Proceedings, Springer, 2014.

[7] Chan, T., Shen, J.: Image processing and analysis: variational, PDE, wavelet, and stochastic methods, SIAM, 2005.

[8] De Los Reyes J. C., Schönlieb C.-B.:, Image denoising: Learning noise distribution via nonsmooth PDE-constrained optimisation, Inverse Problems and Imaging, vol. 7(4), 1183-1214, 2013.

[9] Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms, Physica D, vol. 60, pp. 259-268, 1992.

[10] Tikhonov, A. N., Arsenin V. Y.: Solutions of ill-posed problems. John Wiley & Sons, 1977.



Puzzles, Prisoners and Probability

Quanquan Liu
Undergraduate Student, Massachusetts Institute of Technology

Every once in a while, I hear puzzles about 100 prisoners and a meticulous, demanding warden. All of these puzzles share a common characteristic: a set of prisoners must work together to devise a clever scheme to thwart the warden. I hear these puzzles often enough that each time they reappear, I view them with an increased level of understanding corresponding to the stage of my mathematics education.

Any problem may have a solution, but, sometimes, that solution may not be the most efficient one possible. For example, suppose you want to find something in your room but don't remember where you put it. You can either search the room by yourself. Or you can call your (many) friends to help you. From a correctness standpoint, both solutions are correct. You'll find what you're looking for eventually. But from an algorithmic standpoint, the second solution where you search in parallel with your friends is better because it has a shorter runtime. The same holds for solutions to the 100 prisoners puzzle. Some solutions may be theoretically correct answers to the puzzle but may have expected runtimes that exceed the lifespan of an average person and are, thus, practically undesirable. Now, through this lens, the lens of a theoretical computer science student, I would like to present to you the 100 prisoners puzzle and its variants.

100 Prisoners and a Light Bulb

The original, very famous puzzle involving an interrogation room, a light bulb, and 100 prisoners is the following (paraphrased from Wu in [3]):

One hundred prisoners just arrived in prison. The warden tells them that starting tomorrow, each of them will be placed in an isolated cell, unable to communicate amongst themselves. Each cell has a window so the prisoners will be able to count the days. Each day, the warden will choose one of the prisoners uniformly at random with replacement, and place him in a central interrogation room containing only a light bulb with a toggle switch. The light bulb is initially switched off. The prisoner may observe the current state of the light bulb. If he wishes, he may toggle the light bulb. He also has the option of announcing that he believes all prisoners have visited the interrogation room at some point in time. If this announcement is true, then all prisoners are set free, but if it is false, all prisoners are executed. The warden leaves, and the prisoners huddle together to discuss their fate. Can they agree on a strategy that will guarantee their freedom? [3]

One common solution to the puzzle is to divide the days into 100-day blocks and instruct any prisoner to toggle the light off if he is interrogated twice within the same block. The first prisoner of each block turns the light on and the last prisoner checks whether the light is still on when he enters the interrogation room at the end of the 100-day block. If the light is still on and he did not enter the room on any previous day within the block, he declares that all prisoners have visited the interrogation room [3]. This solution is technically correct because it guarantees the prisoners their freedom, but the prisoners are expected to be freed after 1.072×10^{44} days, in years $\approx 10^{31}$ times the age of the universe. From an algorithmic standpoint, this solution is rather poor because it has

an expected runtime of $O(n^{1/2}e^{n})$ [3] for n prisoners.

The challenge now is to find a solution that is correct and also has an optimal runtime. Such a solution is more likely to guarantee that the prisoners are freed while they are still alive.

The canonical (better) solution is to designate a "leader" to be the person who counts the number of unique prisoners who have been interrogated. The leader may do so by counting the number of times the light bulb has been switched on. A prisoner who has not yet toggled the light switch will turn the light on if it is currently off. A prisoner will do nothing if he enters the room when the light is currently on. The leader turns the light off each time she leaves the room and increases her counter when she sees a light that is on. Thus, after counting 99, the leader may declare that all the prisoners have been interrogated at least once. (It is sufficient to count to 99 because the leader herself counts as the last prisoner.)

How long are the prisoners expected to wait? Suppose that T represents a counter for the number of times the bulb has been switched on. We may count the expected number of days until $T = 99$. Let X_i denote the number of days that pass between an increment of the counter from when $T = i$ until $T = i + 1$. Let Y_i denote the number of days from when a leader turns off a light bulb until a prisoner turns on the light. Let Z_i denote the number of days from when a prisoner turns on the light until the leader enters the room to see the newly turned on light bulb. Thus, $X_i=Y_i+Z_i$. Let X be the number of days the strategy requires in total before the prisoners are freed. Given n prisoners, the probability of turning on the i^{th} light is $\frac{n-i}{n}$. The probability that the leader enters the room on any day is $\frac{1}{n}$. By linearity of expectation,

$$E[X] = \sum_{i=1}^{n-1} E[X_i] = \sum_{i=1}^{n-1} (E[Y_i] + E[Z_i])$$
$$= \sum_{i=1}^{n-1} \left(\frac{n}{n-i} + n \right) = n^2 - n + nH_{n-1}$$

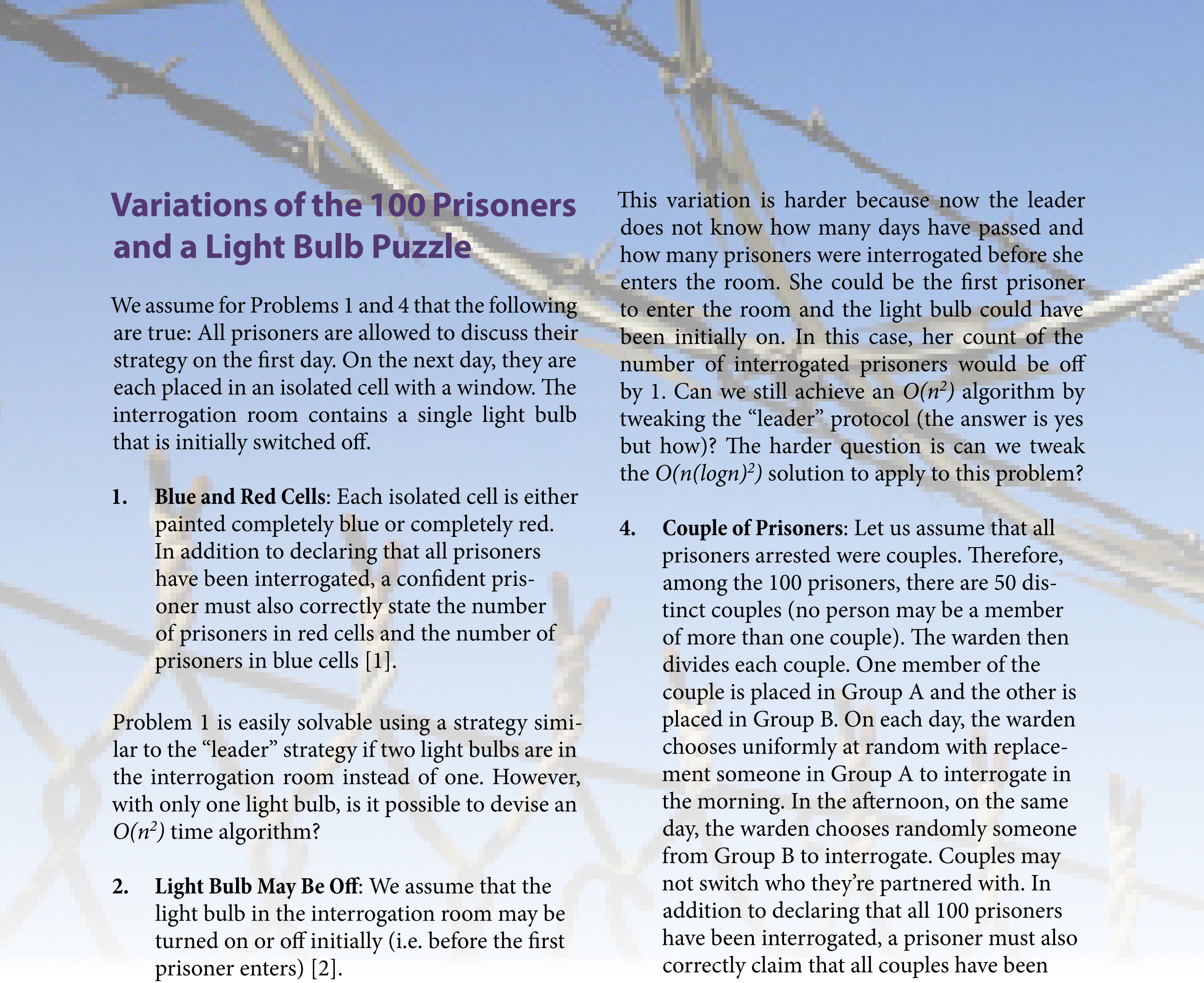
In asymptotic notation, the "leader" algorithm has an expected runtime of $O(n^2)$ days [3]. When there are 100 prisoners, the expected wait time is 10417.7 days or approximately 29 years [3]. Though still a long time, it is within the prisoners' lifespans.

Wu [3] further summarized some strategies that may lead to even shorter wait times. One such strategy achieves an expected runtime of $O(n(logn)^2)$. The key insight behind this algorithm is to allow "assistant" leaders to help the leader by doing some of the counting. Then, the leader would sum together the totals of all the "assistant" counts to determine if all prisoners have visited the interrogation room. To do this, we must be able to divide up the counting of the light bulbs into blocks of days. There must be a block for assistants to count the number of prisoners and a different stage for assistants to tell the leader their total [3]. See [3] for more details.

But can we achieve a solution with an even better runtime, for example, a solution with an $O(n)$ expected runtime? Turns out, the answer is no for the $O(n)$ runtime solution. It is a common joke among CS theoreticians that we hate lower bounds because it prevents us from making better algorithms. The reason why we can't create an $O(n)$ algorithm for the 100 prisoners problem is precisely that a lower bound prevents us from doing so. The expected number of days for all prisoners to enter the interrogation room at least once is $O(nlogn)$, therefore no strategy, no matter how clever, may achieve a better expected runtime than $O(nlogn)$, [1]. A simple calculation confirms this lower bound. Let the random variable X_i be the number of days until the i-th unique prisoner with probability of selection $\frac{n-i+1}{n}$ is picked.

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1}$$
$$= n \sum_{i=1}^{n} \frac{1}{i} = O(n \log n)$$

Naturally, when the original problem has been solved, we wonder if the solution still applies for variants of the problem. Some of these solutions, like the leader and the $O(n(logn)^2)$ solutions, depend on certain characteristics of the problem like the ability to tell time. What if we took away these abilities? Below, I present some harder instances of the 100 Prisoners puzzle and challenge you to find more efficient solutions for them.



Variations of the 100 Prisoners and a Light Bulb Puzzle

We assume for Problems 1 and 4 that the following are true: All prisoners are allowed to discuss their strategy on the first day. On the next day, they are each placed in an isolated cell with a window. The interrogation room contains a single light bulb that is initially switched off.

1. **Blue and Red Cells**: Each isolated cell is either painted completely blue or completely red. In addition to declaring that all prisoners have been interrogated, a confident prisoner must also correctly state the number of prisoners in red cells and the number of prisoners in blue cells [1].

Problem 1 is easily solvable using a strategy similar to the "leader" strategy if two light bulbs are in the interrogation room instead of one. However, with only one light bulb, is it possible to devise an $O(n^2)$ time algorithm?

2. **Light Bulb May Be Off**: We assume that the light bulb in the interrogation room may be turned on or off initially (i.e. before the first prisoner enters) [2].

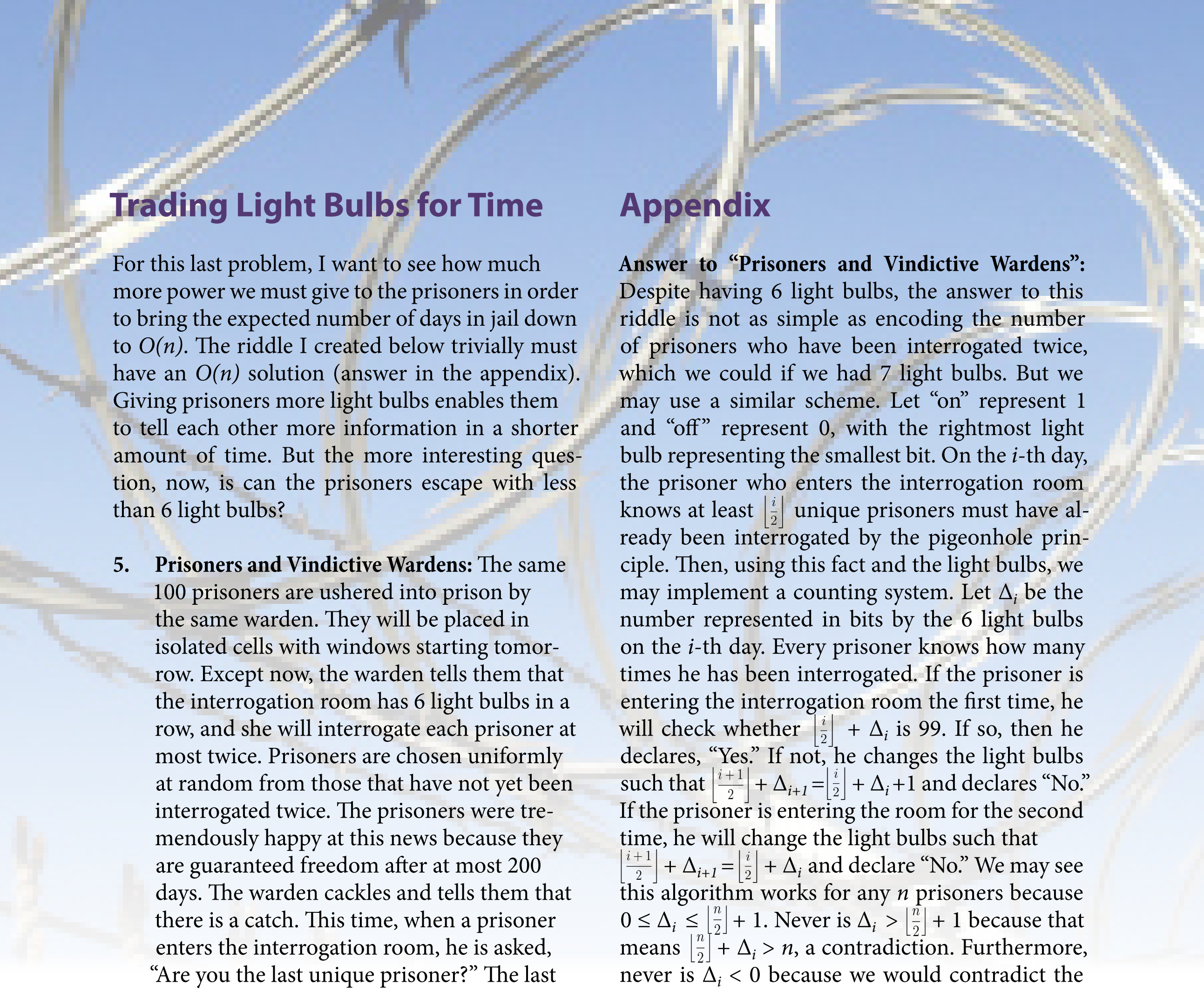
If prisoners still have windows in their rooms, then the "leader" algorithm still provides an $O(n^2)$ solution to Problem 2 because the leader can just record all the times the light is on starting from the second day. The first non-leader prisoner to enter the interrogation room must be unique; therefore, on the first day, he can simply leave the light on if it is on or turn it on if it is off. All other prisoners behave as before. However, this problem becomes trickier if prisoners do not have windows in their individual cells because the prisoners have just lost their ability to keep track of time.

3. No Windows: We keep the condition presented in Problem 2. Now, prisoners may no longer keep track of how much time has passed because they are placed in isolated cells with no windows and no way to keep time [2].

This variation is harder because now the leader does not know how many days have passed and how many prisoners were interrogated before she enters the room. She could be the first prisoner to enter the room and the light bulb could have been initially on. In this case, her count of the number of interrogated prisoners would be off by 1. Can we still achieve an $O(n^2)$ algorithm by tweaking the "leader" protocol (the answer is yes but how)? The harder question is can we tweak the $O(n(logn)^2)$ solution to apply to this problem?

4. **Couple of Prisoners**: Let us assume that all prisoners arrested were couples. Therefore, among the 100 prisoners, there are 50 distinct couples (no person may be a member of more than one couple). The warden then divides each couple. One member of the couple is placed in Group A and the other is placed in Group B. On each day, the warden chooses uniformly at random with replacement someone in Group A to interrogate in the morning. In the afternoon, on the same day, the warden chooses randomly someone from Group B to interrogate. Couples may not switch who they're partnered with. In addition to declaring that all 100 prisoners have been interrogated, a prisoner must also correctly claim that all couples have been interrogated (at least once) on the same day [4].

There exists a solution that assigns each couple to a particular day. The person from Group A may only turn the light on when they are called on their assigned day. Otherwise, they turn the light off. If the person from Group B is also called on their assigned day, they will leave the light on if it is on from the morning. If the person from Group B is called on any other day, they will turn the light off. A leader chosen from Group A counts the number of unique days she sees a light on when she enters the room. This indicates that both members of a couple were interrogated on their assigned day (the previous day). What is the expected runtime of this solution? Does this problem still have a solution if the prisoners are placed in isolated cells without windows?



Trading Light Bulbs for Time

For this last problem, I want to see how much more power we must give to the prisoners in order to bring the expected number of days in jail down to $O(n)$. The riddle I created below trivially must have an $O(n)$ solution (answer in the appendix). Giving prisoners more light bulbs enables them to tell each other more information in a shorter amount of time. But the more interesting question, now, is can the prisoners escape with less than 6 light bulbs?

5. **Prisoners and Vindictive Wardens:** The same 100 prisoners are ushered into prison by the same warden. They will be placed in isolated cells with windows starting tomorrow. Except now, the warden tells them that the interrogation room has 6 light bulbs in a row, and she will interrogate each prisoner at most twice. Prisoners are chosen uniformly at random from those that have not yet been interrogated twice. The prisoners were tremendously happy at this news because they are guaranteed freedom after at most 200 days. The warden cackles and tells them that there is a catch. This time, when a prisoner enters the interrogation room, he is asked, "Are you the last unique prisoner?" The last unique prisoner must declare, "Yes." Every other prisoner must declare, "No." Once a "Yes" is correctly declared, everyone is immediately freed. If someone declares incorrectly, everyone will be executed. How can they guarantee their freedom?

As a hint, the solution to this problem critically depends on the prisoners being able to tell time. If the isolated cells do not contain windows, what, then, is the minimum number of light bulbs needed in order to guarantee the prisoners' freedom?

Is the Solution Optimal?

I hope that you will take what I have written here to heart so that the next time you look at a puzzle, don't just find a right solution; find the optimal solution.

Appendix

Answer to "Prisoners and Vindictive Wardens": Despite having 6 light bulbs, the answer to this riddle is not as simple as encoding the number of prisoners who have been interrogated twice, which we could if we had 7 light bulbs. But we may use a similar scheme. Let "on" represent 1 and "off" represent 0, with the rightmost light bulb representing the smallest bit. On the i-th day, the prisoner who enters the interrogation room knows at least $\left\lfloor \frac{i}{2} \right\rfloor$ unique prisoners must have already been interrogated by the pigeonhole principle. Then, using this fact and the light bulbs, we may implement a counting system. Let Δ_i be the number represented in bits by the 6 light bulbs on the i-th day. Every prisoner knows how many times he has been interrogated. If the prisoner is entering the interrogation room the first time, he will check whether $\left\lfloor \frac{i}{2} \right\rfloor + \Delta_i$ is 99. If so, then he declares, "Yes." If not, he changes the light bulbs such that $\left\lfloor \frac{i+1}{2} \right\rfloor + \Delta_{i+1} = \left\lfloor \frac{i}{2} \right\rfloor + \Delta_i + 1$ and declares "No." If the prisoner is entering the room for the second time, he will change the light bulbs such that $\left\lfloor \frac{i+1}{2} \right\rfloor + \Delta_{i+1} = \left\lfloor \frac{i}{2} \right\rfloor + \Delta_i$ and declare "No." We may see this algorithm works for any n prisoners because $0 \le \Delta_i \le \left\lfloor \frac{n}{2} \right\rfloor + 1$. Never is $\Delta_i > \left\lfloor \frac{n}{2} \right\rfloor + 1$ because that means $\left\lfloor \frac{n}{2} \right\rfloor + \Delta_i > n$, a contradiction. Furthermore, never is $\Delta_i < 0$ because we would contradict the pigeonhole principle. For any n prisoners, this scheme would work given $\left\lceil \log_2 \left(\left\lfloor \frac{n}{2} \right\rfloor + 1 \right) \right\rceil$ light bulbs. May we achieve a better scheme using fewer light bulbs?

References

1. P.O. Dehaye, D. Ford, and H. Segerman, One hundred prisoners and a lightbulb, Mathematical Intelligencer (2003).

2. H. van Ditmarsch, J. van Eijck, and W. Wu, Verifying One Hundred Prisoners and a Lightbulb, Journal of Applied Non-Classical Logics (2007).

3. W. Wu, 100 prisoners and a lightbulb, https://www.ocf.berkeley.edu/~wwu/papers/100prisonersLightBulb.pdf (2002).

4. http://wiki.scn.sap.com/wiki/pages/viewpage.action?pageId=360972998.

Genes, Chinese Restauarants & Textual Analysis

Dr. Todd L. Parsons
Researcher, Centre National de la Recherche Scientifique (CNRS)

From genes to Chinese Restuarants...

In [6], Warren Ewens introduced what would become a foundational tool in the analysis of genetic data, a probability distribution that is now known as Ewens' Sampling Formula. The nearly two decades that had passed since the 1953 discovery of the double-helix structure of DNA had seen the confirmation the mechanism of DNA replication and the discovery of the coding relation between DNA, RNA, and proteins - in particular the system of triplets of nucleotide pairs, called *codons*, and their relationship to amino acids - and thus the discovery of the genetic code. By the end of the 1960's, new techniques of protein sequencing made it possible to confront the mathematical models of gene frequencies developed by R. A. Fisher, Sewall Wright, J. B. S. Haldane, and their successors with data: now, given a set of sample of n proteins from a population, it was possible to count how many distinct alleles appeared, and at what frequencies. This data was encoded in the form of an *allelic partition*, an n-tuple of non-negative integers $(a_1,...,a_n)$, where a_j is the number of distinct alleles that appear exactly j times in the sample, so that

$$\sum_{j=1}^{n} ja_j = n,$$

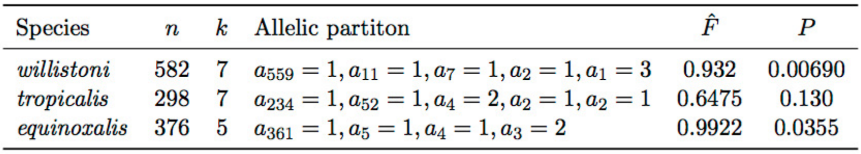
whereas

$$k = \sum_{j=1}^{n} a_j$$

is the number of distinct alleles observed. Some examples of allelic partitions, from population geneticists' favorite organism, the fruit fly *Drosophila*, appear in the table below. Thus, for example, for *Drosophila willistoni*, in a sample of $n = 582$ individuals, there were $k = 7$ alleles, of which the most frequent appeared 559 times, the next most frequent alleles appeared in 11 and 7 individuals, respectively, one allele was observed in two individuals, and 3 alleles appeared in exactly one individuals.

Looking at amino-acid differences (which, due to the transcription of proteins from DNA, correspond to different alleles) in haemoglobin molecules across a variety of mammals [9], Mootoo Kimura argued that the rate of *substitution* - the rate at which new alleles arising by mutation in a single individual replace the ancestral allele across the entire population - was much higher than was consistent with the neo-Darwinist theory that had emerged over the past few decades: Kimura's calculations showed that a nucleotide was substituted roughly every two years, much faster than the once every 300 years predicted by Haldane [8] a decade earlier! This discovery led Kimura to propose his *neutral theory of molecular evolution*, which argued that "non-Darwinian" forces (the randomness in birth and death in small populations composed of individuals whose total lifetime reproductive output is on average equal) as opposed to natural selection (some types have a higher average total reproduction rate) was the primary cause of genetic diversity, sparking a fierce debate that would rage throughout the following decades.

Kimura complemented his neutral theory with what we now call the *infinite alleles model*: mutations happen at approximately a constant rate, and given the typical length of genes, e.g. from an average 1600 nucleotide pairs per gene for yeast, up to an average of 16,600 for mammals [13], the probability of seeing the same mutation twice is very small. As a first approximation, we can thus assume that individuals acquire mutations at a constant per capita rate μ, and with every



Species	n	k	Allelic partiton	$\hat{F}$	P
willistoni	582	7	$a_{559} = 1, a_{11} = 1, a_7 = 1, a_2 = 1, a_1 = 3$	0.932	0.00690
tropicalis	298	7	$a_{234} = 1, a_{52} = 1, a_4 = 2, a_2 = 1, a_2 = 1$	0.6475	0.130
equinoxalis	376	5	$a_{361} = 1, a_5 = 1, a_4 = 1, a_3 = 2$	0.9922	0.0355

Table 1: Sample allelic partitions, sample heterozygosity, and significance values for three species of *Drosophila*, from [15,16] (n.b. unless a value is given, $a_j = 0$).

new mutation, the mutant individual has new allelic type that has never appeared previously. To complete the description of the model, we need to describe the population dynamics: we'll assume a population of fixed size *N*, and assume that each individual has a "reproductive clock" that rings at constant rate 1. When that clock rings, the individual produces an offspring of the same type, which replaces another individual chosen uniformly at random from the whole population. If this all seems a bit artificial, don't worry: although to prove this -- or even be precise about the necessary and sufficient conditions -- is a bit beyond of the scope of this article, the results below continue to hold provided the population is more or less of fixed size, everybody interacts with everybody, individuals give birth to relatively few offspring at any given time, and, essentially, when no type has a selective advantage.

Ewens devised his sampling formula as a robust statistical test of Kimura's infinite alleles model, which, given an allelic partition $(a_1,...,a_n)$ in a sample of size *n*, would give the maximum likelihood estimate of the total population mutation rate, $\theta = N\mu$, and the probability of seeing that particular partition under Kimura's neutral hypothesis:

> **Theorem 1** [Ewens' sampling formula] The probability of observing the allelic partition $(a_1,...,a_n)$ in a population of size *N* that evolves according to Kimura's infinite alleles model with a mutation rate of μ is

$$\frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\left(\frac{\theta}{j}\right)^{a_j}}{a_j!}. \qquad (1)$$

We'll give a proof of this, but instead of Ewens' original, we'll give an easier one that takes advantage of more recent perspectives and thus makes explicit the broad connections of the result (many different proofs have been given since Ewens' original paper e.g. [4] and [5] give two other approaches to the one taken here). In fact, we'll actually prove

> **Theorem 1'** The probability of a sample of *n* individuals containing of *k* distinct allelic types which occur in $n_i > 0$ individuals ($i=1,...,k$, $n_1+...+n_k=n$) is

$$\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{k} (n_i - 1)!. \qquad (2)$$

Theorem 1 then follows by a bit of combinatorics: an allelic partition $(a_1,...,a_n)$ corresponds to the situation where $\sum_{j=1}^{n} a_j = k$ and exactly a_j of the values n_i are equal to *j*; all such choices for the n_i are equally probable (with the value given by (2)), so to compute the probability of the allelic partition, we need only count the number of possible sequences, and multiply by (2), which, under the assumed allelic partition, becomes

$$\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{i=1}^{n} ((j-1)!)^{a_j}. \qquad (3)$$

To count the number of ways of partitioning *n* items into groups of size $n_1,...,n_k$, such that there are a_j groups of size *j*, we could start by imagining listing the *n* items, and then taking the first n_1 and assigning them to the first group, the next n_2 to the second group, etc. There are *n!* ways of forming our initial list, but this over-counts the number of partitions: for example, no matter how many ways we arrange the first n_1 items (which can be done in $n_1!$ ways), they are in the same group, and similarly for each of the n_i. Thus, we should divide *n!* by $\prod_{i=1}^{k} n_i! = \prod_{j=1}^{n} (j!)^{a_j}$ to avoid this over counting. Further, if, say $n_i = n_j$, we could swap the corresponding sets of elements from our list, and still have a partition into groups of size $n_1,...,n_k$. By the same reasoning, if we rearranged any of the a_j groups of size *j*, which can be done in $a_j!$ ways, we would have the same partition. Thus, we must further divide by $\prod_{j=1}^{n} a_j!$ to get the actual number of partitions:

$$\frac{n!}{\prod_{j=1}^{n} (j!)^{a_j} a_j!}.$$

Multiplying this by (3) gives (1).

Before turning to the proof of Theorem 1', let's look at some of the implications (and applications) of the sampling formula: first of all, let K_n be the (random) number of allelic types in a sample of size *n*. Looking at (2), we see that the probability of having *k* allelic types in a sample of size *n* is

$$\mathbb{P}\left(K_n = k\right) = \frac{S(n,k)\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

where

$$S(n,k) := \sum_{\{(n_1,\ldots,n_k):n_1+\cdots+n_k=n\}} \prod_{i=1}^{\kappa} (n_i - 1)!$$

is a well-known quantity, an *unsigned Stirling number of the first kind*. Named after 18th century Scottish mathematician James Stirling, they count the number of distinct permutations of *n* objects with *k* cycles. Using them, we can partition Ewens' sampling formula into two components,

$$\begin{aligned}\mathbb{P}\left((a_1,\ldots,a_n)\right) &= \mathbb{P}\left(K_n = k\right)\mathbb{P}\left((a_1,\ldots,a_n)|K_n = k\right)\\ &= \frac{S(n,k)\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \times \frac{1}{S(n,k)} \prod_{j=1}^{n} \frac{\left(\frac{1}{j}\right)^{a_j}}{a_j!}\end{aligned}$$

where **only the first component depends on** θ. In particular, we can use the first piece as a maximum likelihood estimator for θ that depends only on the size of the sample and the number of distinct alleles in the sample:

$$\frac{d}{d\theta}\log\mathbb{P}\left(K_n = k\right) = \frac{k}{\theta} - \sum_{i=1}^{n} \frac{1}{\theta + i - 1},$$

so $\mathbb{P}\left(K_n = k\right)$ takes its maximum at the unique value of $\hat{\theta}$ satisfying

$$k = \sum_{i=1}^{n} \frac{\hat{\theta}}{\hat{\theta} + i - 1}.$$

By the usual integral comparison argument (see appendix), we observe that $k \sim \hat{\theta} \ln n$, so that for a sufficiently large sample (too large to be very useful in practice, unfortunately), we can make the approximation $\hat{\theta} \approx \frac{k}{\ln n}$.

Note also that

$$\begin{aligned}\mathbb{E}\left[K_n\right] &= \frac{\sum_{k=1}^{n} kS(n,k)\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}\\ &= \frac{\theta\frac{d}{d\theta}\theta(\theta+1)\cdots(\theta+n-1)}{\theta(\theta+1)\cdots(\theta+n-1)} = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1},\end{aligned}$$

so $\hat{\theta}$ is the value for the (population) mutation rate such that the expected number of allelic types, $\mathbb{E}[K_n]$, is equal to the observed number of types, *k*.

Given the number of alleles in the sample, we can use the latter component,

$$\mathbb{P}\left((a_1,\ldots,a_n)|K_n = k\right) = \frac{1}{S(n,k)} \prod_{j=1}^{n} \frac{\left(\frac{1}{j}\right)^{a_j}}{a_j!}, \quad (4)$$

as a test of neutrality **independent of the mutation rate** [16]: we can do a significance test by determining the probability of seeing the observed allelic partition given the observed number of alleles - if it's too low, we can safely reject the neutral hypothesis. More precisely, one can use the Ewens' sampling formula to determine the probability *P* of observing the sample *heterozygosity*,

$$\hat{F} := \sum_{j=1}^{n} a_j \left(\frac{j}{n}\right)^2,$$

that is, the probability that two individuals, drawn uniformly at random **with replacement**, have the same allelic type. For example, in a sample of size 50 that contains 3 allelic types, the heterozygosity can be shown to be always greater than 0.33; (4) tells us that the probability that 0.33 ≤ $\hat{F}$ ≤ 0.37 is less than 5% i.e. if we observed a sample with heterozygosity less than 0.37, we would reject the neutral model, and assume that selection was at work (values from [16]). Table 1 shows values of $\hat{F}$ and *P* for the *Drosophila* samples: only for the species *tropicalis*, with *P = 0.130*, would we fail to reject the neutral hypothesis (when working with data, it's important to think like a scientist - we don't prove our hypotheses true or find counterexamples, but must content ourselves with rejecting -- or failing to reject - them on the basis of statistics).

Note also, that once we've accepted the neutral hypothesis and inferred the value of the population mutation rate, θ, we have a complete understanding of not just the sample, but of the relative abundance of all the allelic types in the population from which the sample came. In particular, the long-run stationary behaviour of the

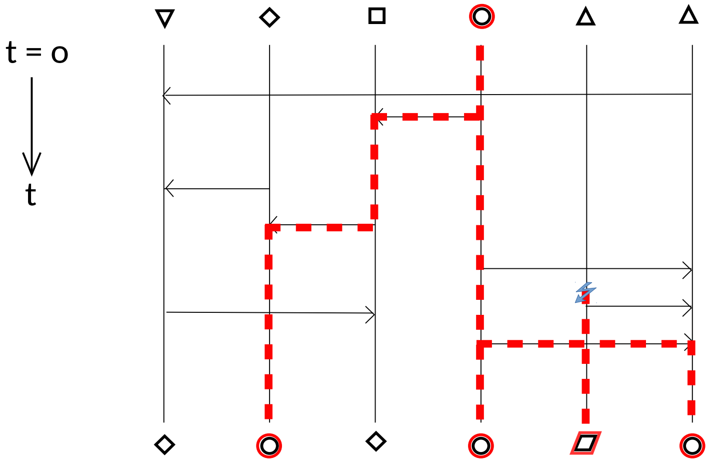


Figure 1: A graphical representation of the infinite alleles model, with the coalescent process superimposed (red). Arrows represent birth/replacement events, whereas one of the "triangular" species experiences a mutation, causing it to become a new "parallelogram" type.

population from which the sample came. In particular, the long-run stationary behaviour of the whole population can be described by a Poisson-Dirichlet process with parameter θ, or by the closely related GEM (Griffiths-Engen-McCloskey) distribution, which gives the proportions of each type in decreasing order (the most frequent type becomes allele A_1, the next most frequent A_2, and so-on). A full discussion of these is beyond the scope of this article, but we refer the interested reader to the extremely good presentations in [7,5].

Now that we're (hopefully) convinced that Ewens' sampling formula is useful, let's turn to its proof. To start, let's describe Kimura's model in a precise manner: let $X_i(t)$ be the number of individuals of allelic type A_i, $i=1,2,\ldots$, so

$$\sum_{i=1}^{\infty} X_i(t) = N,$$

and assume that $(X_1(t), X_2(t), \ldots)$ is a continuous time Markov process, such that in the time interval $[t,t+\Delta t)$, we can have two possible types of events: either some individual of type i gives birth, and its offspring replaces an individual of type j (leaving the population size N unchanged), or an individual of type i incurs a mutation, which causes it to become a new, previously unseen type. The transition probabilities for these events are

$$\begin{aligned}&\mathbb{P}\left(X_i(t+\Delta t)=X_i(t)+1, X_j(t+\Delta t)=X_i(t)-1\right)\\ &\quad=\frac{X_i(t)X_j(t)}{N}\Delta t+o(\Delta t), \quad \text{or}\\ &\mathbb{P}\left(X_i(t+\Delta t)=X_i(t)-1, X_{i^\star}(t+\Delta t)=1\right)\\ &\quad=\mu X_i(t)\Delta t+o(\Delta t),\end{aligned}$$

where i^* is the smallest value of i such that no individual of type A_i has appeared previously, and $o(\Delta t)$ represents any quantity such that

$$\lim_{\Delta t\downarrow 0}\frac{o(\Delta t)}{\Delta t}=0.$$

Note that every type i has equal probability of giving birth, and each type j has equal chance of being replaced, which is what makes this a *neutral* model.

In Figure 1, we give a graphical representation of the infinite alleles model for a population of size $N = 6$. We imagine lining up the individuals according to some arbitrary (but fixed) order, with lines below them on which we will track the birth, death, and mutation events. When an individual's reproduction clock rings, we draw an arrow from their line, pointing at another uniformly randomly chosen line; the individual at the tip of the arrow is replaced by the offspring of the individual at the tail of the arrow. When a mutation occurs (represented in Figure 1 by a small bolt of lightning), the individual at that position becomes a new type. Such figures are especially useful, as they indicate how we can change our perspective: instead of thinking about individuals, we can think about the lines of ancestry that trace back individuals to their parents, giving us a genealogical tree e.g. in Figure 1, the red lines trace the genealogy the sample consisting of the three "circle" types (who descend from a common ancestor at time *t = 0*) and the "parallelogram" (who arises from a mutation to a "triangle", but, if we only have access to the sampled "parallelogram", we can't infer this, so the genealogy stops at the mutation). In population genetics, this ancestral process is called the *coalescent*. It was introduced by J. F. C. Kingman in three papers published in 1982 ([10, 12, 11]) that revolutionised the mathematical approach to population genetics: instead of looking at the forward time evolution of the entire population, one could trace back the ancestry of small samples, an approach that for neutral models is equally powerful, but mathematically quite simple. Coalescents for models with natural selection, however, remain an active area of research!

We'll use a coalescent approach to prove (2). Recall that we started with a sample of *n* individuals, which contains representatives of *k* allelic types, such that there are n_i individuals with allele A_i, $i=1,...,k$. As we look backwards in time, two types of events can occur: two lines with the same allelic type can merge into a common ancestor (i.e., one is the parent of the other and gave birth at that moment), or an allele **that occurs exactly once** disappears - when a mutation occurs, we get exactly one representative of the new type, we but loose all information about their ancestry (if we only have the parallelogram, we have no way of knowing that it was previously a triangle...) Unlike the forward-time population dynamics, both types of events reduce the number of lines in our sample by one, so we only need to trace back at *n* such events to determine the ancestry of our entire sample.

We can obtain a particularly simple representation of our process by ignoring the exact times in the past at which events occurred, and only consider events that change the genealogy of our sample. In such an event, that occurred in the time interval $[t,t+\Delta t)$, when there were *m* ancestors (m_i will allele A_i, $i=1,...,j$) to the lines in our sample, either

1. one of the m_i individuals ancestral to the sample gave birth, and replaced an individual not currently in the sample, an event with probability
$$m_i\left(1-\frac{m}{N}\right)\Delta t + o(\Delta t),$$
or,
2. one of the *N-m* lines not in the sample was hit by a mutation, with probability $(N-m)\mu\Delta t + o(\Delta t) = (N-m)\,\theta\Delta t/N + o(\Delta t)$, and gave rise to a new allelic type that is in our sample.

If we condition on an event happening in $[t,t+\Delta t)$, it is of the first type with probability
$$\frac{m_i\frac{N-m}{N}\Delta t + o(\Delta t)}{\theta\frac{N-m}{N}\Delta t + \sum_{i=1}^{j} m_i\frac{N-m}{N}\Delta t + o(\Delta t)},$$
and of the second type with probability
$$\frac{\theta\frac{N-m}{N}\Delta t + o(\Delta t)}{\theta\frac{N-m}{N}\Delta t + \sum_{i=1}^{j} m_i\frac{N-m}{N}\Delta t + o(\Delta t)}.$$
Dividing the numerators and denominators by $(N-m)\Delta t/N$ and passing to the limit as $\Delta t \to 0$, we arrive at a discrete time Markov chain in which, when we have *m* lines such that m_i carry allele A_i, the number of lines increase by one with probabilities
$$\frac{m_i}{\theta+m} \quad\text{and}\quad \frac{\theta}{\theta+m} \qquad (5)$$
for mergers and mutations, respectively.

With this reduced Markov chain, the proof of Ewens' sampling formula is quite easy:

Proof [of (2)]: If our sample contains n_i individuals of type $i=1,...,k$, we must have first had the first individual of that type appear, with probability proportional to *θ*, and then, that individual gave birth with probability proportional to 1, then, one of the two individuals with allele A_i gave birth with probability proportional to 2, etc. Continuing inductively, we arrive at n_i individuals when one of the n_i - *1* individuals with allele A_i gives birth. That gives us the numerator of (2),
$$\prod_{i=1}^{k}\theta(n_i-1)!$$

To complete the proof, we note that independent of the type of event, when there are m individuals ancestral to the sample, the denominators of the probabilities in (5) are all $(\theta+m)$. As we go from having no individuals to having $n-1$ individuals (the last event we record is the arrival of the n^{th} line in our sample), the denominator is always $\theta(\theta+1)...(\theta+n-1)$, which gives us (2). □

In [1], David Aldous gave an amusing interpretation of the simplified Markov chain given by (5), which he called the *Chinese Restaurant Process*: Aldous imagined an idealised Chinese restaurant, with infinitely many tables, each with an unlimited supply of chairs, and a never-ending supply of hungry diners who arrive, but never leave. When the diners arrive, they either choose to sit at one of the currently occupied tables, with a probability proportional to the number of people already seated ("the more, the merrier"), or with probability θ, they start a new group by sitting at a currently unoccupied table. By contrast, we could imagine an English Restaurant Process, where everybody entering sits down at a vacant table, but that would be much less interesting.

...and beyond

Although we've looked at Ewens' sampling formula in the context of population genetics, it's use extends far beyond. Indeed, (2) was derived independently, in the context of abstract nonparametric Bayesian mixture models by Charles Antoniak in [2]. The recent rediscovery of Antoniak's paper by people working in the field of machine learning probably affects you on a daily basis -- if you do keyword searches on the internet - via an algorithm known as *latent Dirichlet allocation*. When you do a keyword search, you're usually looking for articles on a given topic. We choose keywords because we expect them to occur with high probability within documents on the desired topic. The power of latent Dirichlet analysis is that it's often possible to identify distinct topics without knowing what the topic is, nor knowing the meaning of the keywords, something that is increasingly valuable as the number of pages of text available online grow far faster than humans can read them and assign them to topics!

To understand how this works, lets make an analogy with genetics. We'll imagine that all the words in each topic is a distinct population, and each keyword is an allele. The relative abundance of the keywords is what defines a topic (e.g. an article about genetics is much more likely to contain words like "allele", "nucleotide", "protein", or "*Drosophila*", and in higher abundance than an article on maths, which will have words like "lemma", "theorem", and "proof". Articles like this one, on mathematical population genetics might confuse things a little, but that's a problem easily solved by letting them define a new topic). When we take a document from a topic, we get a sample of alleles from that topic. Much as we could use Ewens' sampling formula, together with small samples of individuals, to reconstruct the population abundances, we can use it, along with sample texts, to reconstruct the abundances of keywords that define the topic, and even determine the most probable assignment of a text to topics -- for example, this article might be assigned 60% genetics, and 40% maths, or *vice versa* - without ever having thought about the content at all! See [3] for a readable survey of such applications.

In addition to further refinements and extensions of the sampling formula for applications in genetics and machine learning, Ewens' work has also inspired a growing subfield that investigates random partitions and other objects at the interface of probability and combinatorics, a wonderful demonstration of the fruitful interaction between mathematics and applications, in which the benefits run in both directions. A comprehensive list of references is impossible, but an excellent point of departure is [14]. Perhaps the reader will be inspired to make their own contribution!

About the author

Dr. Todd L. Parsons is a permanent researcher with the Centre national de la recherche scientifique (CNRS) in France, who divides his time between the Laboratoire de Probabilités et Modéles Aléatoires at Université Pierre Marie Curie (Paris 6), where he belongs to the Équipe Probabilités, Statistiques & Biologie, and the Centre for Interdisciplinary Biology at the Collège de France, where he belongs to the Stochastic Models for the Inference of Life Evolution (SMILE) team. Todd gives his deepest gratitude to Prof. Warren J. Ewens, whose inspiration and mentorship helped him define his own research career.

References

[1] D. J. Aldous. Exchangeability and related topics. In P.L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour, XIII-1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1-198. Springer Berlin Heidelberg, 1985.

[2] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.*, pages 1152-1174, 1974.

[3] D. M. Blei. Probabilistic topic models. *Commun.* ACM., 55(4): 77-84, 2012.

[4] R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, New York, 2nd edition, 2009.

[5] A. M. Etheridge. Some mathematical models from population genetics. In J. Picard, editor, *École d'Été de Probabilités de Saint-Flour, XXXIX-2009*, volume 2012 of *Lecture Notes in Mathematics*, pages 1-119, Berlin Heidelberg, 2011. Springer.

[6] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112, 1972.

[7] W. J. Ewens. *Mathematical Population Genetics*. Springer, Berlin, 1979.

[8] J. B. S. Haldane. The cost of natural selection. *J. Genetics*, 55(3): 511–524, 1957.

[9] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129): 624–6, 1968.

[10] J. F. C. Kingman. The coalesent. *Stoch. Proc. Appl.*, 13: 235–248, 1982.

[11] J. F. C. Kingman. Exchangeability and the Evolution of Large Populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97-112. North- Holland, Amsterdam, 1982.

[12] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A: 27–43, 1982.

[13] B. Lewin. *Genes V*. Oxford University Press Oxford, 1994.

[14] J. Pitman. Combinatorial stochastic processes. In J. Picard, editor, *École d'Été de Probabilités de Saint-Flour, XXXII-2002*, volume 1875, pages 1 – 256, Berlin Heidelberg, 2006. Springer.

[15] G. A. Watterson. Heterosis or neutrality? *Genetics*, 85: 789–814, 1977.

[16] G. A. Watterson. The homozygosity test of neutrality. *Genetics*, 88(2): 405–417, 1978.

Appendix

To show that $k \sim \hat{\theta} \ln n$ in the above, our aim is to bound the sum by an integral which we know how to evaluate. Noting that

$$\int_1^{n+1} \frac{\hat{\theta}}{\hat{\theta} + x - 1}\, dx \leq \sum_{i=1}^{n} \frac{\hat{\theta}}{\hat{\theta} + i - 1} \leq 1 + \int_1^{n} \frac{\hat{\theta}}{\hat{\theta} + x - 1}\, dx,$$

whereas

$$\int_1^{n+1} \frac{\hat{\theta}}{\hat{\theta} + x - 1}\, dx = \hat{\theta}\left(\ln(\hat{\theta} + n) - \ln(\hat{\theta})\right)$$

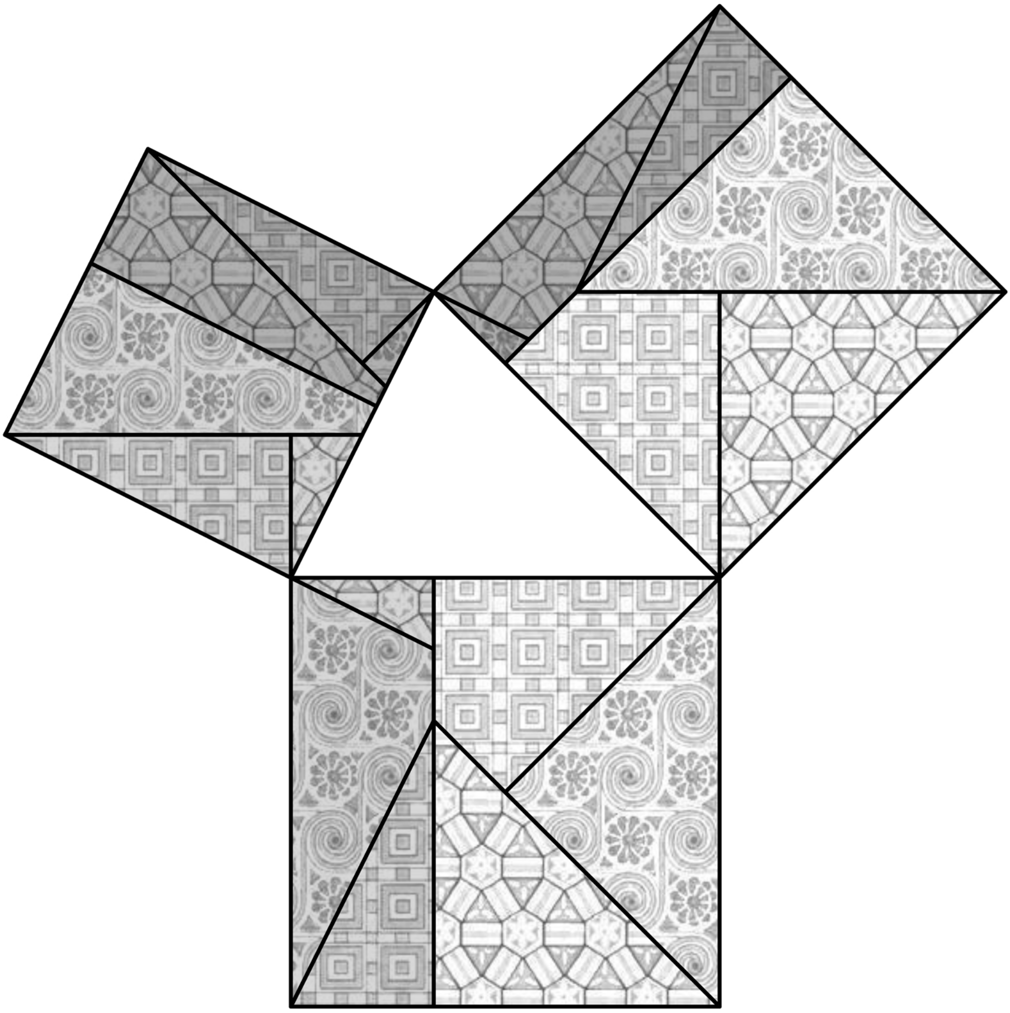
and

$$\int_1^{n} \frac{\hat{\theta}}{\hat{\theta} + x - 1}\, dx = \hat{\theta}\left(\ln(\hat{\theta} + n - 1) - \ln(\hat{\theta})\right)$$

we get the required bounds.

A proof without words of the Law of Cosines

David Brink
Universidade de Cabo Verde, Campus do Palmarejo, Praia, Cape Verde



The three patterns are from the ceiling of an Egyptian tomb (volutes), an Assyrian ornamental painting (squares), and a Persian glazed tile (hexagons), found in Owen Jones, The Grammar of Ornament, Day and Son, Lithographers to the Queen, London, 1856.
E-mail: david.brink@live.dk.

Split and crack the quartic

Dr. Raghavendra G. Kulkarni,
Department of Electronics & Communication Engineering
PES University

The solutions to cubic and quartic equations were found by Cardano and Ferrari respectively in the sixteenth century. Cardano solved the depressed cubic equation, $x^3 = ax + b$, by splitting the variable x as, $x = u + v$, and then expanding the resulting expression; while Ferrari solved the depressed quartic equation, $x^4 + ax^2 + bx + c = 0$, by rearranging the terms in the quartic equation on either side of equality sign and adding some terms on both sides, such that each side becomes a perfect square. Since these methods are well known, we don't discuss them further and readers are advised to see the literature [1].

However we are curious to know whether the quartic equation can be solved in a manner Cardano solved the cubic, i.e., by splitting the variable x as, $x = u + v$, and then expanding the resulting expression. Let us make an attempt here.
Consider the depressed quartic equation,

$$x^4 + ax^2 + bx + c = 0 \tag{1}$$

where a, b, and c are coefficients in (1). Substituting $x = u + v$ in (1) and expanding it we obtain,

$$\begin{aligned} u^4 + 4u^3v + 6u^2v^2 + 4uv^3 + v^4 + au^2 \\ + 2auv + av^2 + bu + bv + c = 0 \end{aligned}$$

Notice that the above equation can be rearranged such that the left-hand-side is made a perfect square as,

$$(u^2 + 2uv)^2 = -(2v^2 + a)\left(u^2 + \frac{4v^3 + 2av + b}{2v^2 + a}u + \frac{v^4 + av^2 + bv + c}{2v^2 + a}\right) \tag{2}$$

The quadratic term in u in the right-hand-side of (2) also can be made a perfect square as:

$$\left[u + \frac{4v^3 + 2av + b}{2(2v^2 + a)}\right]^2$$

if the condition,

$$\frac{(4v^3 + 2av + b)^2}{4(2v^2 + a)} - (v^4 + av^2 + bv + c) = 0 \tag{3}$$

is satisfied. Further simplification of this condition (3) leads to the following cubic equation in v^2 as,

$$v^6 + (a/2)v^4 - cv^2 + [(b^2 - 4ac)/8] = 0 \quad (4)$$

The cubic (4) in v^2 is known as resolvent cubic equation, and solving it results in three values of v^2 or six values of v. Now equation (2) becomes a perfect square as,

$$(u^2 + 2uv)^2 = -(2v^2 + a)[u + \frac{4v^3 + 2av + b}{2(2v^2 + a)}]^2 \quad (5)$$

Taking square root of (5) and rearranging the terms results in the following two quadratic equations.

$$u^2 + [2v \mp \sqrt{-(2v^2 + a)}]u \mp \sqrt{-(2v^2 + a)}\left[\frac{4v^3 + 2av + b}{2(2v^2 + a)}\right] = 0 \tag{6}$$

The two equations in (6) contain ∓ signs at two places; the first equation has – sign at both places, and the second one has + sign at both places. Solving these quadratic equations, we obtain four values of u for each value of v. Four solutions of quartic equation (1) are then obtained using $x = u + v$.

Let us solve one numerical example. Consider the following depressed quartic equation.

$$x^4 + 3x^2 - 6x + 10 = 0$$

The resolvent cubic (4) in v^2 is obtained as,

$$v^6 + 1.5v^4 - 10v^2 - 10.5 = 0;$$

and after solving it, we get three values of v^2 as, 3, −1, −3.5, and six values of v as, ±1.732050807568, $\pm i$, ±1.870828693387, where $i^2 = -1$. Choosing $v = i$, the two quadratic equations in (6) are obtained as:

$$u^2 + iu + 1 + 3i = 0; u^2 + 3iu - 1 - 3i = 0.$$

Solving above equations, we get four values of u as: $1 - 2i$, $-1 + i$, 1, and $-1 - 3i$. Using the relation, $x = u + v$, we obtain four solutions of the quartic equation as: $1 - i$, $-1 + 2i$, $1 + i$, and $-1 - 2i$. Notice that one can choose any of the six values of v to solve the quartic equation. Interested readers may verify by using other values of v to obtain the solutions.

ACKNOWLEDGEMENTS: The author thanks the management of PES University for supporting this work.

AUTHOR CONTACT INFORMATION: Department of Electronics & Communication Engineering, PES University, 100 Feet Ring Road, BSK III Stage, Bengaluru - 560085, INDIA.
Email: dr_rgkulkarni@yahoo.com; raghavendrakulkarni@pes.edu

References

[1] G. Birkhoff and S. MacLane, ”A survey of modern algebra”, Macmillan, 5th edition, New York, 1996.

A Few Interesting Sequences

A.Delgado, B.Goodman, M.Lewinter, B.Phillips
City College of New York

0. Introduction

Sequences are important in many branches of mathematics. The sequence of partial sums, for example, determines the convergence of an infinite series. Analysis courses abound with sequences of functions and the various ways in which they converge. Number theorists study the Fibonacci and Lucas sequences among many others. While the recursive definition of the Fibonacci sequence involves two prior terms to define the n-th term, you will see that a slightly altered version only requires one prior term – an esthetically pleasing and surprising fact! Guest appearances by π and e in the article will be appreciated by lover of mathematics at all levels.

In this article, we present several recursively defined sequences and obtain interesting results that involve relatively simple tools such as limits and asymptotic behavior. Students encounter asymptotic behavior in a first year calculus course when they take ratios of dominant terms to determine the limit of a rational function when x approaches infinity.

The only tool in this paper unfamiliar to math freshmen is Stirling's Approximation which can be explained quite readily.

If any readers sponsor undergraduate theses, this article should be a source of interesting problems for future projects with students.

I. Definition

Let the sequence $\{x_n\}$ be defined by the seed $x_1 = x_2 = 1$, and the recursive relation

$$x_{n+2} = x_n + \frac{1}{x_{n+1}} \qquad (1)$$

The first ten terms are

$$1, 1, \frac{2}{1}, \frac{3}{2}, \frac{8}{3}, \frac{15}{8}, \frac{48}{15}, \frac{105}{48}, \frac{384}{105}, \frac{945}{384} \qquad (2)$$

For reasons that will be apparent later, we do not reduce the fractions in (2) to lowest terms. Now multiplying both sides of (1) by x_{n+1} yields

$$x_{n+1}x_{n+2} = x_n x_{n+1} + 1 \qquad (3)$$

Letting $f(n) = x_n x_{n+1}$, (3) becomes $f(n+1) = f(n) + 1$. Since $f(1) = x_1 x_2 = 1$, we find that $f(n) = n$, implying that $x_n x_{n+1} = n$. It follows that

$$x_{n+1} = \frac{n}{x_n} \qquad (4)$$

In other words, the sequence defined by (1) has a simpler recursive relation (4) in which each term depends solely on the preceding term. Note that the first ten terms satisfy the new recursion. We use (4) to write another recursive formula from which the sequence may be obtained (with the seed $x_1 = x_2 = 1$). We have, using (4)

$$x_{n+2} = \frac{n+1}{x_{n+1}} = \left(\frac{n+1}{n}\right)x_n \text{ , or } x_{n+2} = \left(\frac{n+1}{n}\right)x_n \qquad (5)$$

II. Several Theorems

As a consequence of (5), we have the following theorem.

Theorem 1: $\lim_{x\to\infty}\left(\frac{x_{n+2}}{x_n}\right)^n = e$

Proof: By (5), we have $\left(\frac{x_{n+2}}{x_n}\right)^n = \left(1+\frac{1}{n}\right)^n$, which approaches e as n goes to infinity. ▪

We now find the closed form for the sequence, using (4). The two seed terms will be omitted, but they follow the pattern to be established.

$$x_3 = \frac{2}{x_2} = \frac{2}{1} = \frac{(1!)^2 2^2}{2!}$$

$$x_4 = \frac{3}{x_3} = \frac{1\cdot3}{2} = \frac{1\cdot2\cdot3}{2^2} = \frac{3!}{(1!)^2 2^2}$$

$$x_5 = \frac{4}{x_4} = \frac{2\cdot4}{1\cdot3} = \frac{(2\cdot4)^2}{1\cdot2\cdot3\cdot4} = \frac{(2!)^2 2^4}{4!}$$

$$x_6 = \frac{5}{x_5} = \frac{1\cdot3\cdot5}{2\cdot4} = \frac{1\cdot2\cdot3\cdot4\cdot5}{(2\cdot4)^2} = \frac{5!}{(2!)^2 2^4}$$

$$x_7 = \frac{6}{x_6} = \frac{2\cdot4\cdot6}{1\cdot3\cdot5} = \frac{(2\cdot4\cdot6)^2}{1\cdot2\cdot3\cdot4\cdot5\cdot6} = \frac{(3!)^2 2^6}{6!}$$

$$x_8 = \frac{7}{x_7} = \frac{1\cdot3\cdot5\cdot7}{2\cdot4\cdot6} = \frac{1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7}{(2\cdot4\cdot6)^2} = \frac{7!}{(3!)^2 2^6}$$

$$x_9 = \frac{8}{x_8} = \frac{2\cdot4\cdot6\cdot8}{1\cdot3\cdot5\cdot7} = \frac{(2\cdot4\cdot6\cdot8)^2}{1\cdot2\cdot3\cdot4\cdot5\cdot6\cdot7\cdot8} = \frac{(4!)^2 2^8}{8!}$$

Table 1

Clearly, the closed form of x_n will depend on the parity of n.

Case 1: *n* = 2*k* + 1. Then $x_{2k+1} = \frac{(k!)^2 2^{2k}}{(2k)!}$ (6)

Case 2: *n* = 2*k*. Then $x_{2k} = \frac{(2k-1)!}{[(k-1)!]^2 2^{2k-2}}$ (7)

The reader is reminded that two functions *f(x)* and *g(x)* are called asymptotic if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$ This does not imply that |*f(x)* – *g(x)*| is bounded, as can be seen by the pair of asymptotic functions *f(x)* = x^2+*x* and *g(x)* = x^2 whose absolute difference goes to infinity. We write *f(x)* ~ *g(x)* to denote that f and g are asymptotic. The reader is also reminded of Stirling's beautiful relation [1]

$$k! \sim \left(\frac{k}{e}\right)^k \sqrt{2\pi k} \tag{8}$$

which we will use to obtain asymptotic approximations for x_{2k} and x_{2k+1}. We require the easily verified asymptotic relations (9) and (10).

$$(k!)^2 \sim \left(\frac{k}{e}\right)^{2k} 2\pi k \quad (9) \qquad (2k)! \sim \left(\frac{2k}{e}\right)^{2k} 2\sqrt{\pi k} \quad (10)$$

Using (9) and (10) have

$$x_{2k+1} = \frac{(k!)^2 2^{2k}}{(2k)!} \sim \frac{(k/e)^{2k} 2\pi k 2^{2k}}{(2k/e)^{2k} 2\sqrt{\pi k}} = \sqrt{\pi k} \tag{11}$$

Obtaining an asymptotic approximation for x_{2k} will be more difficult.

$$x_{2k} = \frac{(2k-1)!}{[(k-1)!]^2 2^{2k-2}} \sim \frac{(\frac{2k-1}{e})^{2k-1}\sqrt{2\pi(2k-1)}}{(\frac{k-1}{e})^{2k-2} 2\pi (k-1)^{2k-2}} \sim \frac{1}{e}(\frac{2k-1}{k-1})^{2k-1}\frac{\sqrt{4\pi k}}{2^{2k-1}\pi}$$

$$= \frac{1}{e}(\frac{2k-1}{2k-2})^{2k-1}\frac{2\sqrt{k}}{\sqrt{\pi}} = \frac{1}{e}(\frac{2k-1}{2k-2})(1+\frac{1}{2k-2})^{2k-2}\frac{2\sqrt{k}}{\sqrt{\pi}} \sim 2\sqrt{\frac{k}{\pi}}$$

The last asymptotic approximation made use of the fact that $\lim_{x\to\infty}\left(1+\frac{1}{n}\right)^n = e$

On the basis of these calculations, we have the following theorem.

Theorem 2: $x_{2k} \sim 2\sqrt{\frac{k}{\pi}}$ and $x_{2k+1} \sim \sqrt{\pi k}$

Letting *n* = 2*k* in (4), we have $x_{2k}x_{2k+1}$=2*k*, which is consistent with the above theorem. Furthermore, Theorem 2 implies that $\lim_{x\to\infty} x_n = \infty$. The following corollaries of Theorem 2 will be useful.

Corollary 1: $\lim_{x\to\infty}\frac{x_{2k+1}}{x_{2k}} = \frac{\pi}{2}$

Proof: By Theorem 2, we have $\frac{x_{2k+1}}{x_{2k}} \sim \frac{\sqrt{\pi k}}{2\sqrt{\frac{k}{\pi}}} = \frac{\pi}{2}$. ■

Corollary 2: $\lim_{x\to\infty}\frac{x_{2k+2}}{x_{2k+1}} = \frac{2}{\pi}$

Proof: $x_{2k} \sim 2\sqrt{\frac{k}{\pi}}$ implies, upon replacing *k* by *k* + 1, that $x_{2k+2} \sim 2\sqrt{\frac{k+1}{\pi}}$. Then

$$\frac{x_{2k+2}}{x_{2k+1}} \sim \frac{2\sqrt{\frac{k+1}{\pi}}}{\sqrt{\pi k}} = \frac{2}{\pi}\sqrt{\frac{k+1}{k}} \sim \frac{2}{\pi} \;. ■$$

As a consequence of Corollaries 1 and 2, note that $\lim_{x\to\infty}\frac{x_{n+1}}{x_n}$ fails to exist. Moreover, by these corollaries, there exists a positive integer, *K*, such that for all *k* > *K*, one has $x_{2k+1} > x_{2k}$, while $x_{2k+2} < x_{2k+1}$. In other words, from some point on, terms with odd index are greater than the terms immediately before them, while the reverse is true for terms with even index. Note that this strict alternation between increasing and decreasing behavior appears to be the case for the terms in (2) with the exception of the equality of the first two terms.

The next theorem requires the following Lemma which we state without proof.

Lemma 1: Let a and b be positive integers, and let *p* be a prime that divides *b* but does not divide *a*. Then *a/b* is not an integer.

The reader is reminded that Bertrand's Postulate [2] says that for *n* > 1, there is at least one prime *p* such that *n* < *p* < 2*n*.

Theorem 3: Let *n* > 3. Then x_n is not an integer.

Proof: We have two cases depending on the parity of *n*.

Case 1: *n* = 2*k* + 1, where *k* > 1. Then by (6), $x_{2k+1} = \frac{(k!)^2 2^{2k}}{(2k)!}$. By Bertrand's Postulate, there exists a prime *p* such that *k* < *p* < 2*k*. Then *p* divides that denominator, but not the numerator, of x_{2k+1}. Then by Lemma 1, x_{2k+1} is not an integer.

Case 2: $n = 2k$, where $k > 1$. From Table 1, one has $x_{2k} = \frac{1 \cdot 3 \cdots (2k-1)}{2 \cdot 4 \cdots (2k-2)}$. Then 2 divides the denominator, but not the numerator, of x_{2k}. Then by Lemma 1, x_{2k} is not an integer, and the theorem is proven. ■

Let $P_n = x_1 x_2 x_3 \cdots x_n$, from which $x_{n+1}(P_n)^2 = x_1 x_1 x_2 x_2 x_3 x_3 \cdots x_n x_n x_{n+1}$. Recall that (4) can be written as $x_n x_{n+1} = n$. Then we have $x_{n+1}(P_n)^2 = n!$. We have proven the following theorem

Theorem 4: $P_n = \sqrt{\frac{n!}{x_{n+1}}}$ ■ (11)

The equations of Table 1, which were obtained from (4), imply that

$$x_{2k} = \frac{N_{2k}}{D_{2k}} = \frac{1 \cdot 3 \cdots (2k-1)}{2 \cdot 4 \cdots (2k-2)} \tag{12}$$

$$x_{2k+1} = \frac{N_{2k+1}}{D_{2k+1}} = \frac{2 \cdot 4 \cdots (2k)}{1 \cdot 3 \cdots (2k-1)} \tag{13}$$

In light of (12) and (13), we turn our attention to the sequence of numerators $\{N_n\}$, whose first few terms are 1, 1, 2, 3, 8, 15, 48, 105, 384, 945. Since $x_n = N_n/D_n$ and $x_{n+1} = n/x_n$, we have $N_{n+1}/D_{n+1} = nD_n/N_n$. As we are not reducing these fractions, it follows that

$$N_{n+1} = nD_n \text{ and } D_{n+1} = N_n \tag{14}$$

Note that the terms 8, 15, 48, 105 and 384 of the sequence $\{N_n\}$ can be rewritten $1(3^2 – 1)$, $1(4^2 – 1)$, $2(5^2 – 1)$, $3(6^2 – 1)$ and $8(7^2 – 1)$, where the factors before the parentheses are consecutive members of the same sequence. This is not a coincidence. By (14), $N_{n+2} = (n+1)D_{n+1} = (n+1)N_n$, implying that $N_{n+4} = (n+3)N_{n+2} = (n+3)(n+1)N_n = [(n+2)^2 – 1]N_n$.
We have proven the following theorem.

Theorem 5: $N_{n+4} = [(n+2)^2 - 1]N_n$. ■

By the second equation in (13), a similar identity holds for the sequence, $\{D_n\}$, of denominators.

We turn our attention to the sequence $\sum_{n=1}^{\infty} \frac{1}{x_n}$. Replacing n by $n – 1$ in (1) yields $x_{n+1} = x_{n-1} + 1/x_n$ in which case $1/x_n = x_{n+1} - x_{n-1}$. Defining $x_0 = 0$ enables us to sum both sides of this last equation from $n = 1$ to $n = m$.
Note that $1/x_1 = x_2 = 1$. We obtain $\sum_{n=1}^{m} \frac{1}{x_n} = \sum_{n=1}^{m} (x_{n+1} - x_{n-1})$

Telescoping the sum on the right gives $x_m + x_{m+1} - 1$. Thus we have proven the following theorem.

Theorem 6: $\sum_{n=1}^{m} \frac{1}{x_n} = x_m + x_{m+1} - 1$ ■

Corollary 3: $\sum_{n=1}^{\infty} \frac{1}{x_n}$ diverges.

Proof: Let m go to infinity in Theorem 6. Then use $\lim_{n \to \infty} x_n = \infty$ which was established using Theorem 2. ■

III. A related sequence

Let a new sequence $\{y_n\}$ be defined by the seed $y_1 = y_2 = 1$, and the recursive relation

$$y_{n+2} = y_{n+1} + \frac{1}{y_n} \tag{15}$$

The first few terms are $1, 1, 2, 3, \frac{7}{2}, \frac{23}{6}, \frac{173}{42}$.

Question 1: Find a recursive relation of the form $y_{n+1} = g(y_n, n)$ as was the case for $\{x_n\}$ in section I, that is, a relation such as (4).

By (14), one sees that $\{y_n\}$ is strictly increasing, unlike the sequence $\{x_n\}$. As a consequence, we have the following theorem.

Theorem 7: $\lim_{x \to \infty} y_n = \infty$.

Proof: Since $\{y_n\}$ is strictly increasing, it must either approach infinity or have a finite limit, say L. We show the latter is impossible by contradiction. Assume that $\lim_{x \to \infty} y_n = L < \infty$. Then taking the limit of both sides of (14) as n goes to infinity yields the absurd equation $L = L + 1/L$. and we are done. ■

The next theorem yields a result similar to Theorem 6.

Theorem 8: $\sum_{n=1}^{m} \frac{1}{y_n} = y_{m+2} - 1$

Proof: By (14), $1/y_n = y_{n+2} - y_{n+1}$. Then $\sum_{n=1}^{m} \frac{1}{y_n} = \sum_{n=1}^{m} (y_{n+2} - y_{n+1})$

which telescopes down to y_{m+2} - 1. ▪

Corollary 4: $\sum_{n=1}^{\infty} \frac{1}{y_n}$diverges.

IV. Another sequence whose recursive definition can be simplified

Let the sequence $\{x_n\}$ be defined by the seed $x_1 = x_2 = 1$, and the recursive relation

$$x_{n+2} = x_{n+1} + 2x_n \qquad (15)$$

which is a variant on the recursive relation of the Fibonacci sequence. The first few terms are

$$\{1, 1, 3, 5, 11, 21, 43, 85, ...\} \qquad (16)$$

We have the following interesting theorem which, by the way, is the reason the above sequence is included in this paper.

Theorem 9: The sequence defined by (15) and the seed $x_1 = x_2 = 1$ satisfies the recursion $x_{n+1} = 2x_n + (-1)^n$.

Proof: We use a variant of induction. The theorem is clearly true for $n = 1$ and $n = 2$. Now assume it is true for $n = k$ and $n = k + 1$. Then we have $x_{k+1} = 2x_k + (-1)^k$ and $x_{k+2} = 2x_{k+1} + (-1)^{k+1}$.

Using (15) and the preceding two equations, we have $x_{k+3} = x_{k+2} + 2x_{k+1}$

$= 2x_{k+1} + (-1)^{k+1} + 2[2x_k + (-1)^k]$

$= 2(x_{k+1} + 2x_k) + [(-1)^{k+1} + 2(-1)^k]$

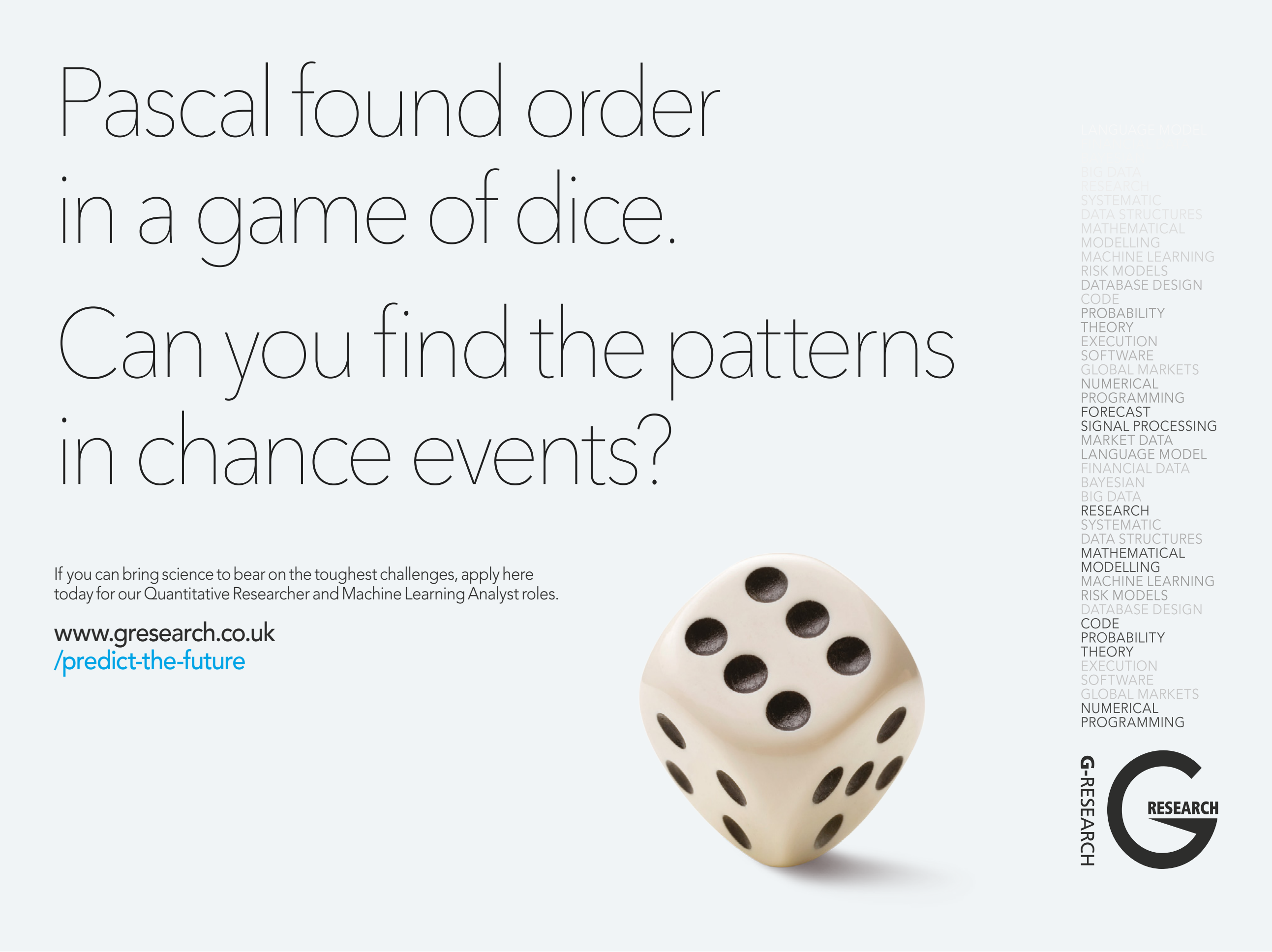
$= 2x_{k+2} + (-1)^k[-1+2] = 2x_{k+2} + (-1)^k$

$= 2x_{k+2} + (-1)^{k+2}$. ▪

References

[1] A. Taylor and W. Mann, Advanced Calculus. 3rd ed., Wiley, NYC, 1983.

[2] David M. Burton, Elementary Number Theory. 4th ed., McGraw-Hill, NYC, 1998.



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.